Journal of Computational Biology and Bioinformatics Research Vol. 5(1), pp. 6-14, April 2013
Available online at http://www.academicjournals.org/JCBBR
DOI: 10.5897/JCBBR12.013
ISSN 2141-2227 ©2013 Academic Journals

Full Length Research Paper

Annotation of virulence factors in schistosomes for the development of a SchistoVir database

Adewale S. Adebayo[1] and Chiaka I. Anumudu[2]*

[1]Cell Biology and Genetics Unit, Department of Zoology, University of Ibadan, Oyo State, Nigeria
[2]Cellular Parasitology Programme, Department of Zoology, University of Ibadan, Oyo State, Nigeria.

Accepted 8 March, 2013

Scientific efforts in the eradication of neglected tropical diseases, such as those caused by the parasitic helminthes, can be improved if a database of key virulence factors directly implicated in pathogenesis is available. As a first step towards creating SchistoVir, a database of virulence protein factors in schistosomes, in this study, we curated, annotated and aligned sequences of twenty virulence factors identified from the literature, using several bioinformatics tools including UniProtKB, SchistoDB, VirulentPred, InterProScan, ProtScale, MotifScan, TDRtarget, SignalP, MODBASE, PDB and MUSCLE*.* Among the protein entries, the most frequently occurring amino acid residues were lysine, serine, leucine, glutamine, glycine and cysteine in order of magnitude. Although sequence repeat regions (SRRs) of significant value were identified manually in fifty percent of the proteins (while dipeptide repeats (DiPs) and single amino acid repeats (SAARs) were not), nevertheless, seventy-two percent of the protein entries were classified as virulent by the prediction model, VirulentPred. Most of the entries (eighty percent) did not have target compounds based on the database of available chemical compounds at TDRtargets. Fourteen of the twenty entries (seventy percent) had more than 30 consecutively negative amino acid residues based on the ProtScale's Kyte and Doolittle hydrophobicity plot. Hence, they would be hydrophobic enough to be transmembrane in location or secretory in nature. Only 7 (tyrosinase, serine protease1, Tspan-1, VAL4, cathepsin b and L and calreticulin) had cleavage sites and signal peptides, while none had a significant signal anchor probability. The annotations and characterization provided by this work and the development of a SchistoVir database will aid in further research of schistosome pathogenesis and control.

Key words: Protein database, bioinformatics tools, virulence proteins/factors, annotation, schistosomes.

INTRODUCTION

Schistosomes are pathogenic helminthes, a group of parasites which constitute important sources of morbidity and mortality in several parts of the world, with 2 billion persons affected (Fumagalli, 2010). Of the several species of schistosomes known, *Schistosoma mansoni*, *Schistosoma haematobium* and *Schistosoma japonicum* are important in the spread of morbidity. Among the tropical diseases caused by parasites, schistosomiasis ranks second only to malaria as a cause of catastrophic worldwide morbidity and mortality. Besides, it is believed to infect 200 million persons (Dvorak et al., 2008; Chalmers et al., 2008). The control of schistosomiasis involves population-based chemotherapy with the use of praziquantel (PZQ) and metrifonate drugs. PZQ increa-ses antigen exposure, induces adenosine receptor bloc-kade and calcium influx, causes paralysis and distorts the parasite's morphology. Snail control (through habitat modification, environmental planning and molluscicides) and immunological control (vaccination) are also components of the disease control strategy. The identification and development of vaccine candidates (usually protein antigens) will be useful in reducing morbidity drastically. Yet, a single actual potent vaccine is not within reach (WHO, 2010).

*Corresponding author. E-mail: chiaka.anumudu@mail.ui.edu.ng or walsaks002@yahoo.com.

Generally, parasites (including helminthes) use several mechanisms to attack hosts, while deriving nutrients and scheming for their own continued survival, and the success of parasites can be highly correlated with their ability to evolve a sophisticated immune evasion strategy (Matisz et al., 2011). Therefore, the key parasite proteins involved in these actions are critical to survival and they have been classified as virulence proteins (Fankhauser et al., 2007; Ramana and Gupta, 2009). These virulence proteins may be involved in attachment or adhesion to host membrane receptor cells, establishment and penetration within host cells, cleaving of host proteins and invasion (Ramana and Gupta, 2009). These proteins are strictly regulated and have also been identified as therapeutic agents; some of them are already at clinical trial stages (Gomez et al., 2010). A function-based classification of these virulence proteins was done by Ramana and Gupta (2009). The major classes identified were invasion, establishment, adhesion, proteases and others with unknown or putative function, although there is sometimes no sharp delineation amongst the different categories.

Schistosome virulence proteins have been evaluated for use as vaccine or drug targets. Chalmers et al. (2008), MacDonald et al. (2002) and Lopez-Quezada and McKerrow (2011) affirmed that SmVAL and serpins (serine protease inhibitors) are involved in immune system modulation. The SmVAL (*S. mansoni* venom allergen -like) have a conserved SCP/TAPS domain that possesses envenomation and larval penetration activity. The serpins are able to distort host proteases. Furthermore, proteomic studies have also shown that the Sm29, gluthatione-S-transferase, thioredoxin, tetraspanins, triose phosphate isomerase, Sm32 among others in schistosoma, are critical for host haemoglobin degradation, reduction of Th2 immune response and tissue invasion (Braschi et al., 2006; Hansell et al., 2008; Cardoso et al., 2008; Verjovski-Almeida and DeMarco, 2008; Sharma et al., 2009). There are also studies to indicate the importance of some of these proteins in treatment strategies. WHO /USAID partnership led to the establishment of the Schistosomiasis Vaccine Development Programme (SVDP) which identified epitopes of the triose phosphate isomerase (TPI), Sm14 and GST as virulence proteins with key vaccine candidates (WHO, 2010). Braschi et al. (2006), Reis et al. (2008) and Aslam et al. (2008) also highlighted the roles of tetraspanins, TPI and serpins in the preparation of efficacious vaccines.

In silico approaches have enhanced research in parasite pathogenesis and efforts in this direction continue till today (Devor, 2005; Hogeweg, 2011). This falls under the categorization of functional genomics (Garg and Gupta, 2008). Available databases of bacterial virulent proteins and toxins include VFDB (Virulence Factors Database), PRINTS (Protein Family fingerprints) and MVirDB (Microbial database of protein toxins and virulence factors) (Zhou et al., 2007; Tsai et al., 2009). These are available at http://prediction centre. llnl. gov, provided by the Lawrence Livermore National Laboratory, US. ProtVirDB (http://bioinfo.icgeb.res.in/protvirdb) is a database for virulence factors in protozoans. The databases mentioned provide unified information portals for researchers interested in a panoramic or in-depth view of the virulent proteins in a parasite of interest or in comparison with other parasites.

In fact, many works available have indicated the roles of several virulence proteins in schistosomes pathogennesis, the full genome database of the *S. mansoni* (SchistoDB, www.schistodb.net/schistodb) has been made publicly available. Also, there are a number of public protein databases which offer information on the proteome of *S. mansoni*, *S. haematobium or S. japonicum,* for example, UniProt (http://uniprot.org). Nevertheless, there is no specialized and simplified public information portal for the virulence factors identified so far in the schistosomes, either in *S. mansoni, S. haematobium or S. japonicum*.

To the best of our knowledge, no database or classification to specifically annotate virulence proteins in parasitic worms exists, although substantial research has been done in characterizing proteins in different helminth species (Caprona et al., 2005; Braschi et al., 2006; Curwen et al., 2006; Cardoso et al., 2008; Aslam et al., 2008; Bos et al., 2009; Boumis et al., 2011). Such a database will give a simplified information portal for the researchers interested in a panoramic or in-depth view of the virulent proteins in a parasite or in comparison with other parasites. It will also facilitate sequence retrieval and analysis and will be a useful tool for the research community in the study of schistosome pathogenesis.

METHODOLOGY

In order to construct and develop a preliminary secondary database of schistosome virulence proteins (SchistoVir), annotation of selected proteins were done using some tools.

Catalogue of protein entries

Protein entries were curated from nucleotide sequences, genomic sequences and literature available in NCBI's RefSeq and PubMed (http://ncbi.nlm.nih)[1];GeneDB[2](http://genedb.org/genedb/smansoni); SchistoDB Release 2.0[3] ((http://schistodb.net) and UniProt KB[4] (http://www.uniprot.org/search). The criteria for choice of entry will be essentiality, decisiveness or crucial nature of the protein for survival in the host. The use of multiple databases is to ensure that all possible entries are obtained. The databases will provide literature sources, genomic sequences, contigs and protein sequences of schistosome entries.

[1] National Centre for Biotechnology Information. It has a large depository of biomedical literature and genomic information. The RefSeq provides updated, universally confirmed genomic sequences and PubMed provides literature.

[2] GeneDB provides genomic and proteomic data on species which have been completely sequenced

[3] SchistoDB provides protein and genomic information on the Schistosoma mansoni genome

[4] UniProtKnowledgebase is a mega database on proteomic data from hundreds of species

Coding and protein sequences, status and amino acid length were obtained from SchistoDB or UniProt queries with the use of keywords or gene names. Ontogenic expression which gives a measure of the protein's expression at different stages of the life cycle of schistosomes was obtained from SchistoDB using a number of ESTs (expressed sequence tags) in adults per total number of ESTs for all stages. The results were saved as complete html pages. Blastp searches were conducted using the UniProt Blast tool with default parameters.

Phylogenetic and secretory/transmembrane analysis

Homologous protein sequences to each of the virulence factors from selected species are obtained from NCBI and alignments were made using the MUSCLE multiple alignment tool (at http://ebi.ac.uk/tools/muscle) (Edgar, 2004) with default parameters and results are displayed in ClustalW format. Muscle is hosted by European Bioinformatics Institute-EMBL and provides for automated sequence analysis with multiple alignment.

Signal sequences in the proteins were recognized with SignalP 3.0 (http://www.cbs.dtu.dk/services/SignalP) hosted by the Technical University of Denmark, using both the neural network and Hidden Markov Model (HMM) methods (Nielsen et al., 1997; Bendtsen et al., 2004). In order to increase accuracy, presence or absence of signal peptides was defined by the default Neural Network *D* score thresholds. This is in accordance with the method of Bos et al. (2009). SignalP results present d, s and y scores. The D score is the average of the maximal Y-score (the most likely location of the cleavage site of the signal sequence) and the mean S-score, and is the best way to discriminate true signal sequences in proteins (Emanuelsson et al., 2007). In responding to query sequences, proteins with a D score greater than 55 and HMM greater than 90% were scored as having an N-terminal signal sequence (hence, transmembrane) and presented in our results. The default eukaryote setting of SignalP, with each sequence truncated after 70 residues, was used to avoid false positive detection of signal sequence outside the N-terminus. SignalP results were read off the query results directly from both the HMM and neural networks.

Domain, model and transmembrane predictions

Model search or prediction was done using protein data bank (PDB or MODBASE; http://rscb.pdb.org and http://salilab.org/modbase respectively), with the protein sequence as a query with default settings. The model predictions are presented as 3D structures obtained from MODBASE[1] query searches. Helices, beta sheets, turns or coils and loops are easily visible this way. PDB (protein data bank) 3D structures are displayed when available.

ProtScale[2] from ExPasy[3] (http://web.expasy.org/cgi-bin/protscale) was used to generate a hydropathy plot based on the calculated hydrophobicity of constituent amino acids. Interpretation is based on the fact that twenty consecutive hydrophobic amino acids are needed for a peptide/protein to be transmembrane. The Kyte-Doolittle hydrophobicity plot method was adopted while using ProtScale. ProtScale results from query sequence were generated in a numerical verbose format so as to be able to obtain numbers of consecutive hydrophobic residues. Prediction of motifs and domains was done using Motif Scan (http://myhits.isb-sib.ch/cgi-bin/motif_scan) (Sigrist et al., 2010) of the Swiss Institute of Bioinformatics which makes use of the PROSITE database. Also, the InterPro Scan version 33.0 (Mulder et al., 2007) (http://www.ebi.ac.uk/Tools/services/web_iprscan) from European Bioinformatics Institute, an arm of European Molecular Biology Laboratory (EMBL) was used to discover conserved protein signatures and domains of individual entries.

Query sequences in plain format used in MotifScan and InterProScan returned results in html and SVG formats. Signatures, profiles and domains recognized are derived from these results and summarized. The relative hydrophobicity of a protein and the absence or presence of signal peptides with cleavage sites are important in determining if it is transmembrane or not.

Other tools in the annotation and characterization of entries

Protein entries were also characterized using the following tools:

1. VirulentPred (http:bioinfo.icgeb.res.in/virpred): a bacterial virulence factor prediction server, which relies upon amino acid composition, dipeptide composition, similarity search of known virulence factors in bacteria, and cascade support vector machine algorithms to predict likelihood of virulence.
2. TDRtargets version 4.0 (Crowther et al., 2010): Using homology to druggable proteins with specified modifiable criteria, the server generates possible drug targets. In the results, associated drugs or compounds represent the results of searches conducted on the TDR targets database for each entry. The GO terms display generally accepted terms of the function of a protein, the cellular component of which the protein is part and the interaction of the protein with other substances which is its molecular function.

RESULTS AND DISCUSSION

An initial twenty proteins derived from the literature were annotated in a simplified format using bioinformatics tools. Due to the large nature of the results or documentation generated, a summary of protein documentation and analysis are presented in Tables 1 and 2, respectively. The proposed schema is shown in Figure 1. Simply put, a database schema is a graphical depiction of the structure of the database and a similar pattern has been adopted by Ramana and Gupta (2009). A sample of the annotation pages for one of the protein entries is shown in Figure 2. The proteins included are discussed under the following headings.

Protein entries

Inclusion of the protein entries, data generated for the amino acid sequence, molecular weight, domains and signatures for the entries agree with the works of Chacon et al. (2003), Herve et al. (2003) for 28GST, Ramos et al. (2003, 2009) and Rabia et al. (2010) for Sm14, Fitzpatrick et al. (2007) for tyrosinase, Chalmers et al. (2008) for venom allergen like VAL, Kane et al. (2004) for major egg antigen p40, Berriman et al. (2009), Boumis et al. (2011) for thioredoxin, Lopez-Quezada and McKerrow (2011) for serpin and Wu et al. (2011).

Proteases, proteinase inhibitors and binding proteins

[1] MODBASE predicts a protein's 3D structure, when it is not available in its data bank (Pieper et al., 2011). It is a tool for comparative protein structure modeling.
[2] ProtScale analyses the profile of a query sequence using amino acid residues present
[3] Protein analysis software from Swiss Bioinformatics

Table 1. The database proteins with their basic features and classification.

Systematic name	Name	Feature	Functional classification
Smp_155560	Serpin	50% (germball), 33% (cercariae), 43.62 kD, 4 paralogs Smp_062080,062120,155530,155550)	Establishment (Protease inhibitor)
Smp_000022	Tyrosinase	Female adults only, 56.3kD	Invasion (egg migration/formation
Smp_075800.1	Sm32	2 paralogs (Smp_075790,179170)	Protease
Smp_054470	Thioredoxin	85.7% adult, 7.1% schistosomula,1 paralog (Smp_008070),11.2kD	Establishment (redox homeostasis)
Smp_155310	Tetraspanin/Tspan-1	Adults only, 24.05kD, 8 paralogs	Establishment (Protective)
Smp_157090	Cathepsin L	2.8% (adult) 1.9% (schistosomula), 168.16kD, 3 paralogs (Smp_034410.1-3)	Protease
Smp_158420	Cathespsin B	36.6kD, 4 paralogs (Smp_067060, 103610,141610,179980)	Protease
Smp_072190	Sm29	58% (adult), 21.2kD, no paralogs	Establishment
Smp_095360	Fatty acid binding protein or Sm14 (fabp)	85% (adult), 9.8% (schistosomula), 14.9kD, 3 paralogs (Smp_174440.4,095360.1-2)	Establishment (Uptake of host fatty acid)
Smp_030350	Serine Protease 1 (Sp1)	54.28kD, no paralogs	Protease
Smp_003990	Triose phosphate isomerase	81.8% adult,12.7% expression (cercariae), 28.09kD	Establishment
Smp_054160	28GST	85% (adult), 23.82kD, no paralogs	Invasion
Smp_030370	Calreticulin	50.3% (adult), 45.4kD,no paralogs	Establishment (protein folding)
Smp_183000	Major egg antigen p40	21.07kD, 17 paralogs	Establishment (immune evasion)
Smp_018890	Phosphoglycerate kinase (Pgk)	75% (adult),18.47kD,1 paralog (Smp_187370)	Establishment
Smp_042160.2	Fructose 1,6 biphosphate aldolase	80.8% adult, 39.7kD,1 paralog (Smp_042160.1)	Establishment
Smp_002070	Venom allergen like (VAL)	28 paralogs, confirmed only as transcripts	Establishment

*All systematic names are derived from the SchistoDB (Berriman et al., 2009); *Paralogs are homologous or similar sequences found in same species; *All percentages are calculated from number of EST per total EST.

Table 2. Classification of database proteins based on different bioinformatics tools.

Parameter	Number of entries	Entry protein
Virulence using bacterial parameters and score (ViruPred)	13	Tyrosinase (0.6586); fabp1-97aa (0.6656); Sm14 (0.7938); Sp1 (0.6586); thioredoxin 2 (0.6586); Tspan-1 (1.2771); egg antigen p40 (1.1121); serpin (1.0287); cathepsin B(0.2498); cathepsin L(1.3629); calreticulin (0.2028); Sm29 (0.9471); Sm32
Non virulence according to Virupred	5	Triose p.isom (-1.742); thioredoxin!; Pgk (-0.816); 28GST (-0.572); fructose bp aldolase (-1.367)
Having known targets / compounds	4	Tyrosinase, t.p.isomearse, fabp3, 28GST
No known target compounds [Human ortholog target available*]	16	Fabp1, SP1*, calreticulin, thioredoxin1 and 2*, Tspan-1, p40, cathepsin b*, cathepsin L*, serpin* PGK*, Sm29, Sm23, fructose biphosphate aldolase*, Sm32*
MODBASE/PDB data available [both available*, MODBASE only^]	13	Fabp3*, 28GST*, Sm14*, aldolase^, Sm29^, pgk^, thioredoxin1&2^, sp1, cathepsin L and b^, tpi^, p40^

include Sm32, serine protease and cathepsins. 58 kD serine protease has been documented to be capable of mitigating the effects of IgE immune response through cleavage (Aslam et al., 2008). Involvement of protease inhibitors or proteinase inhibitors such as serpin 1 in direct virulence or pathogenesis has also been recognised by Dalton et al.

Table 2. Contd.

No PDB data/chain No MODBASE prediction model	3	Tspan-1, Sm23, Sm32
Hydrophobicity: Entries with highest number of consecutive negative residues		Calreticulin (130), cathepsin L (79), VAL 4 (65), cathepsin B (49), p40 (42), Sm32(40)
Largest number of Epitopes		Cathepsin L (55), Sm32 (40), Sm23 (35), 28GST (30), aldolase (26), SP1(20), Sm32 (20), serpin1 (17), Pgk 23

!: Predicted virulent based on dipeptide composition but non virulent based on other parameters, ^ no PDB structural data available.

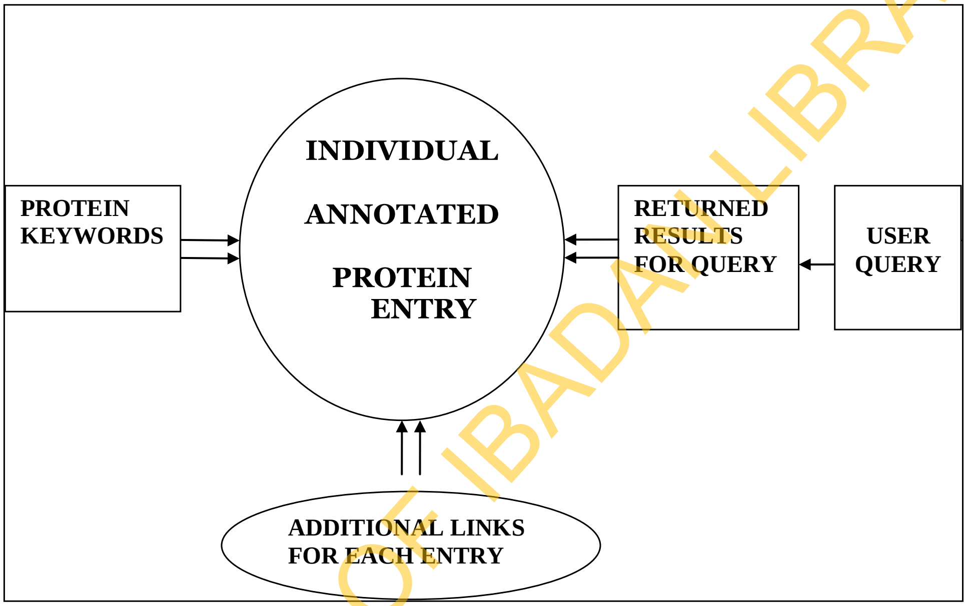

PROTEIN KEYWORDS
INDIVIDUAL ANNOTATED PROTEIN ENTRY
RETURNED RESULTS FOR QUERY
USER QUERY
ADDITIONAL LINKS FOR EACH ENTRY

Figure 1. Schema (graphical structure) for the database.

al. (1997), Lin and He (2006) and Lopez-Quezada and McKerrow (2011).

Triose phosphate isomerase, aldolase, calreticulin and glutathione s-transferase (GST) are entries which have been identified as excretory-secretory (ES) products in *Schistosoma* by previous studies and these ES proteins have been implicated in invasion of host tissues by sporocysts, haemocyte encapsulation and eventual immunosuppression and immune evasion (Guillou et al., 2007; Reis et al., 2008).

None of the entries had a confirmed myristoylation site, an amino acid side chain to which a lipid group, myristic acid (14C saturated fatty acid) can be added to aid the localization of a cytosolic protein to a membrane. Although one entry, the major egg antigen p40 had probable sites. Nevertheless, this did not connote the absence of transmembrane proteins among the entries.

It is also noteworthy that some of the entries: Sm14, Sm23, 28GST, triose phosphate isomerase and Sm32 have already been identified as vaccine candidates (Cardoso et al., 2008; WHO, 2010). This is due to their high level of expression and antigenicity. Also, each protein entry had at least two GO ID and term/names to represent the function term (F) and the process terms (P). The GO terms aptly describe entries based on their status as a cellular component; in biological process and molecular function.

Amino acid sequence

The amino acid length of the protein entries and coding sequence varied from 97 to 1471 aa and 231 to 4413 bp. Protein length is apparently diverse although long-chained proteins are not in abundance among the entries. The more frequently occurring residues among the protein sequences were lysine, serine, leucine, glycine, cysteine, glutamine and aspartic acid. The first 5 residues have also been found to occur frequently among eukaryotic virulence proteins as reported by Garg and Gupta (2008). Quezada et al. (2009) also showed that conserved cysteine residues and leucine rich domains are key to virulence protein function.

SAMPLE ANNOTATION PAGE

Sm 14 Fatty acid binding protein (FaBP)

Status: putative, conserved in few eukaryotes

Protein/ coding sequence

MSSFLGKWKLSESHNFDAVMSKLGVSWATRQIGNTVTPTVTFTMDG
DKMTMLTESTFKNLSCTFKFGEEFDEKTSDGRNVKSVVEKNSESKLT
QTQVDPKNTTVIVREVDGDTMKTTVTVGDVTAIRNYKRLS 133aa

Coding sequence

ATGTCTAGTTTCTTGGGAAAGTGGAAACTTAGCGAGTCACACAAC
TTCGATGCTGTCATGTCAAAGCTAGGTGTCTCATGGGCAACTCGAC
AGATTGGGAACACAGTGACCCCAACTGTAACCTTCACAATGGATG
GGGATAAAATGACTATGTTAACAGAGTCAACTTTCAAAAATCTTTC
TTGTACGTTCAAGTTCGGCGAGGAATTCGATGAAAAAACAAGTGA
CGGCAGAAATGTCAAGTCAGTTGTTGAAAAAAATTCCGAGTCGAA
GTTAACGCAAACTCAAGTAGATCCCAAAAACACAACTGTAATCGT
TCGTGAAGTGGATGGTGATACTATGAAAACGACTGTGACTGTTGG
TGACGTTACTGCCATTCGCAATTATAAACGACTATCCTAA 402bp

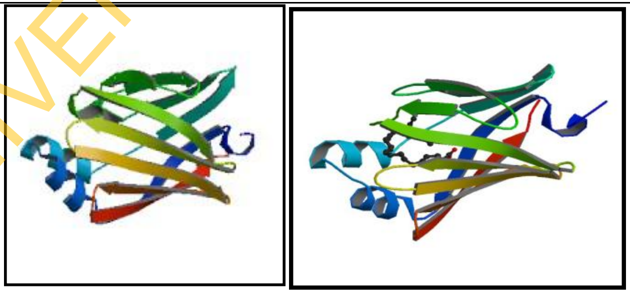
Suggested Virulence category: Establishment

Associated compound (TDR): none human ortholog cmpd: 3-carbazol-9-ylpropanoic acid, 2 hexyldecanoic acid

Ontogenic/stage specific exp.(schistodb)

(no of EST/total ESTs) Adult (139/163), schistosomula (16/163)

Model Prediction (A:MODBASE;B: PDB structure bound to oleic acid)



Homology **3 paralogs**

Best blastp hits (% identity): *Homo sapiens* fabp, brain (48%); *Bos taurus* heart fabp(44%); *Mus musculus* adipocyte binding protein (43%)

Figure 2. Sample annotation page for one of the protein entries.

None of the entries appeared to have single amino acid repeats (SAARS) or Di-peptide repeats (DiPs). On the other hand, sequence repeat regions (SRRs) of two amino acids in length were found in triose phosphate isomerase, tyrosinase, egg antigen p40, sp1, tspan-1, serpin, cathepsin L, calreticulin, 28GST and Sm14. These 2aa repeats were both heteropeptide and homopeptide repeats. The numbers of repeats found in the rest of the entries were not considered to be significant enough. The occurrence of repeats in many of the virulence proteins of microbial and protozoan pathogens has been reported by Gravekamp et al. (1998), Karlin et al. (2002), Fankhauser et al. (2007) and Ramana and Gupta (2009). None of the public repeat searching tools: *ProtrepeatsDB* and *RepSeq* (Kalita et al., 2006; Depledge et al., 2007) were available online for use. Hence, entries were examined manually for repeats, though the note was taken for natural probability of amino acid repeat occurring for any given sequence. Availability of repeat searching tools/servers would have provided a means of detecting the more complicated mismatch repeats that may be present.

Similarly, there were no notable conserved sequences observed in the alignment of the protein entries generated from MUSCLE. This might have been expected due to a wide range of protein families.

Virulence

The twenty protein entries were used as query on VirulentPred which identifies partial sequence of a protein that could aid virulence according to bacterial models with default settings. The results of these VirulentPred predictions (Table 1) may serve as a form of validation for some of the entries. 13 of the 20 entries queried on the server were declared virulent by all parameters (amino acid composition, dipeptide composition, similarity search and cascade support vector machine algorithms).

According to Garg and Gupta (2008), VirulentPred is highly sensitive even for eukaryotic sequences, but may produce false positives in eukaryotic sequences due to compositional differences.

Although schistosomes are eukaryotic and more highly evolved than bacteria pathogens, certain features may be conserved in virulence proteins. In fact, such conservation is seen in the occurrence of repeats and residues (previous page) in these pathogens. Lysine and serine, two of the most frequent residues in the protein entries (and in eukaryotic virulence proteins) also frequently occur in bacterial virulence proteins. Hence, the VirulentPred results cannot be totally discarded.

Other functional predictions

In the Kyte-Doolittle hydropathy plot, a large number of consecutive negative amino acid residues (20 to 30) are highly indicative of the hydrophobicity of a protein, its ability to be inserted into the internal hydrophobic environment of the membrane and its likelihood of being transmembrane. Most of the protein entries possess a significant number of hydrophobic amino acids, with calreticulin having 130 consecutive residues (the highest) and thioredoxin 2 having 14 residues (lowest). Fatty acid binding protein, thioredoxin 1 and 2, and Sm29 were the only entries with less than 20 consecutive residues. These are all evident from ProtScale results generated numerically (verbose format). The algorithm used by ProtScale takes no cognizance of the first 4 amino acids of the protein sequence used as query for such and are usually involved in signal peptide and cleavages. Hence, high number of entries (16 of 20) had strong hydrophobic portions with at least 20 to 30 consecutive negative residues. It has been postulated from previous studies that hydrophobic residues of virulence proteins would aid their integration of membrane, adhesion to host cells or stimulate binding of the proteins to targets in the hosts (Katsir et al., 2008; Quezada et al., 2009; Blanco et al., 2010).

Sites for cleavage of proteins involved in the conventional secretory pathway in cells were identified in 7 of the entries (35%). Such entries have signal sequences and a high D score, and it predicts that the entries are integral to the membrane and non-cytoplasmic. The usage of eukaryotic parameters and truncation of each query sequences to 70 terminal residues (in the signalP queries) increased reliability of the result generated (Bos et al, 2009). Virulence-related outer membrane, integral membrane or transmembrane proteins have been identified in prokaryotic and eukaryotic pathogens (Schulz and Vogt, 1999; Kim et al., 2009).

Associated compounds and drugs

Diverse compounds were found to be associated with different protein entries even among paralogs. As our queries of TDRtargets showed, only 4 of the protein entries including triose phosphate isomerase, 28GST and tyrosinase, had already named or known compounds that could be used to target them. There are still no known compounds to directly target 16 of the protein entries, although 9 of the 16 have known targets for their human orthologs. Hence some key protein factors involved in virulence of schistosomes cannot yet be targeted with any of the synthetic/natural compounds known to humans, at least within the TDRtarget database, a huge database formed by a collaboration of several universities and the WHO TDR drug target network (Crowther et al., 2010).

Ontogenic/stage specific expression

It is pertinent to note that most of the protein entries including tyrosinase, fabp, Sm14, thioredoxin, Tspan-1, cathepsin L, phosphoglycerate kinase, Sm29, calreticulin, 28GST, aldolase and Sm32 have high expression levels in adult schistosomes as results from SchistoDB showed it. Serpin is one exception of the entries (high expression in germball and cercariae). Adult pathogens are highly evolved and would be capable of producing proteins directly involving pathogenesis.

Model search and prediction

Surprisingly, a significant number of the protein entries had no structural data available in the protein data bank (PDB) and MODBASE was relied on for prediction of the structures of most of the entries. 28GST and Sm14 had 3D model data from both databases, 13 entries had model predictions generated from MODBASE. 3 entries: sm23, Sm32 and tspan-1 had no models from both databases. In terms of online availability and accessibility of 3D structural data, research on many proteins directly involved in schistosome pathogenesis may be slow.

Suggested virulence category

The protein entries had specific functions such as evasion of host immune responses, penetration of host barriers, degradation of host protective proteins and establishment in the host, all of which contribute to virulence, the relative ability of parasites to induce pathogenesis. Each of these functions was used to categorize the entries into a suggested virulence category. Hence 4 categories were identified: Proteases, Establishment (uptake of host nutrients, redox homeostasis, etc), proteases and invasion (proteinase inhibitors, host penetration) (Table 1)

Antigenicity

It is worthy of note that all protein entries had at least 5 putative antigenic epitopes with the highest being that of cathepsin L, SmSP1 with 55 putative epitopes. Epitopes or better still, the antigenic determinants are discrete sites on a protein or antigen which B and T lymphocytes recognize. Epitopes are the immunologically active regions in a complex antigen that actually bind to B-cell or T-cell receptors. Such information on epitopes is useful when considering adaptive immunity response to parasites and possible therapeutic targets. Recognition of these epitopes depended on evidence provided by TDR targets v4.0.

Algorithm [Structured English Text]

```
If a protein name is entered into the query search,
1. [Start] Copy alphabets of a query
2. Compare alphabets of the query with the stored template protein keywords 1 to 20
3. Search for a match
a. There is a match if all the alphabets are the same
b. There is also a match if 3 to 5 alphabets are the same
5. Display the matched protein keyword. If there is no match, display ‘no result’
6. Search the annotated protein entry for matched keyword
7. Display the annotation page[End].
8. If additional link search is performed, go to http://schistodb.net/schistodb[End].
```

Summary, conclusion and recommendation

Studies on molecular aspects of parasitic helminthes can be improved especially in the tropics if a database of key virulence factors which are directly implicated in pathogenesis is developed. This work has laid a foundation for us to develop such a database. It would be useful to the research community at a time when the search for vaccines for several helminth diseases is on the increase. It also provides grounds for further studies related to the significance of proteins involved in virulence of the parasitic helminthes. The database will be made public (depending on funds availability), updated regularly and additional tools for virulence prediction incorporated. It is proposed to expand the database to include other pathogenic helminthes so that users are provided with more possibilities in terms of species coverage.

REFERENCES

Aslam A, Quinn P, McIntosh RS, Shi J, Ghumra A, McKerrow JH, Bunting KA, Dunne DW, Doenhoff MJ, Sherie LM, Ke Z, Richard JP (2008). Proteases from Schistosoma mansoni cercariae cleave IgE at solvent exposed interdomain region. Mol. Immunol. 45(2):567-574.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004). Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340:783-795.

Berriman M, Haas BJ,LoVerdo PT, Wilson RA, Dillon GP, Cerquiera GC, El Sayed NM (2009). The genome of the blood fluke *Schistosoma mansoni*. Nature 460:352-358

Bos DH, Mayfield C, Minchella DJ (2009). Analysis of regulatory protease sequences identified through bioinformatic data mining of the *Schistosoma mansoni* genome. BMC Genomics 10: 488-492.

Boumis G, Angelucci F, Bellelli A, Brunori M, Dimastrogiovanni D, Miele AE (2011). Structural and functional characterization of *Schistosoma mansoni* Thioredoxin. Protein Sci. 20(6):1069-1076.

Blanco MT, Sacristán B, Lucio L, Blanco J, Pérez-Giraldo C, Gómez-García AC (2010). Cell surface hydrophobicity as an indicator of other virulence factors in *Candida albicans*. Rev. Iberoam Micol. 27(4):195-199.

Braschi S, Borges WC, Wilson RA (2006). Proteomic analysis of the schistosome tegument and its surface membranes. Mem Inst Oswaldo Cruz. 101(I): 205-212.

Caprona A, Riveaua G, Caprona M, Trottein F (2005). Schistosomes: the road from host–parasite interactions to vaccines in clinical trials. Trends Parasitol. 21(3): 143-149.

Cardoso FC, Macedo GC, Gava E, Kitten GT, Mati VL (2008). *Schistosoma mansoni* Tegument Protein Sm29 Is Able to Induce a Th1-Type of Immune Response and Protection against Parasite Infection. PLoS Negl Trop Dis. 2(10): e308.

Chalmers IW, McArdle AJ, Coulson RM, Wagner MA, Schmid R, Hirai H, Hoffmann KF (2008). Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the *Schistosoma mansoni* venom allergen-like (SmVAL) gene family. BMC Genomics 9:89.

Crowther GJ, Shanmugam D, Carmona SJ, Doyle MA, Hertz-Fowler C, Berriman M, Nwaka S, Ralph SA, Roos DS, Van Voorhis WC, Agüero F (2010). Identification of Attractive Drug Targets in Neglected-Disease Pathogens Using an *In Silico* Approach. PLoS Negl Trop Dis. 4(8): e804.

Curwen RS, Ashton PD, Sundaralingam S, and Wilson RA (2006). Identification of Novel Proteases and Immunomodulators in the

Secretions of Schistosome Cercariae That Facilitate Host Entry. Mol. Cell. Proteomics 5(5):835-844.

Dalton JP, Clough FA, Jones MK, Brindley PJ (1997). The cysteine proteinases of *Schistosoma mansoni* cercariae. Parasitology 114: 105-112.

Depledge DP, Lower RP, Smith DF (2007). RepSeq – A database of amino acid repeats present in lower eukaryotic pathogens. BMC Bioinformatics 8:122.

Dvorak J, Mashiyama ST, Braschi S, Sajid M, Knudsen GM, Hansell E, Lim KC, Hsieh I, Bahgat M, Mackenzie B, Medzihradszky KF, Babbitt PC, Caffrey CF and McKerrow JH (2008). Differential use of protease families for invasion by schistosome cercariae. Biochimie 90: 345-358.

Edgar RC (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Emanuelsson O, Brunak S, von Heijne G and Nielsen H (2007). Locating proteins in the cell using TargetP, SignalP and related tools. Nat. Protoc. 2:953-971.

Fankhauser N, Nguyen-Ha T, Adler J, Mäse P (2007). Surface antigens and potential virulence factors from parasites detected by comparative genomics of perfect amino acid repeats. Proteome Sci. 5: 20.

Fitzpatrick JM, Hirai YHH, Hoffmann KF (2007). Schistosome egg production is dependent upon the activities of two developmentally regulated tyrosinases. FASEB J. 21: 823-835.

Fumagalli M, Pozzoli U, Cagliani R, Comi GP, Bresolin N, Clerici M, Sironi M (2010). The landscape of human genes involved in the immune response to parasitic worms. BMC Evol. Biol. 10:264.

Garg A, Gupta D (2008). VirulentPred: a SVM based prediction method for virulent proteins in bacterial pathogens. BMC Bioinformatics 9:62.

Gomez C, Ramirez ME, Calixto-Galvez M, Medel O and Rodríguez MA (2010). Regulation of Gene Expression in Protozoa Parasites. J Biomed. Biotechnol. 2010: 726045.

Gravekamp C, Rosner B, Madoff LC (1998). Deletion of repeats in the alpha C protein enhances the pathogenicity of group B streptococci in immune mice. Infect. Immun. 66:4347-4354.

Guillou F, Roger E, Moné Y, Rognon A, Grunau C, Théron A, Mitta G, Coustau C, Gourbal BE (2007). Excretory–secretory proteome of larval *Schistosoma mansoni* and *Echinostoma caproni*, two parasites of *Biomphalaria glabrata*. Mol. Biochem. Parasitol. 155 (1):45-56.

Hansell E, Braschi S, Medzhiradszky KF, Sajid M, Debnath M (2008). Proteomic Analysis of Skin invasion by blood fluke larvae. PLoS Negl. Trop. Dis. 2(7):e262.

Herve M, Angeli V, Pinzar E, Wintjens R, Faveeuw C, Narumiya S, Capron A (2003). Pivotal roles of the parasite PGD2 synthase and of the host D prostanoid receptor 1 in schistosome immune evasion. Eur. J. Immunol. 33: 2764–2772.

Hogeweg P (2011). The Roots of Bioinformatics in Theoretical Biology. PLoS Comput. Biol. 7(3): e1002021.

Kalita MK, Ramasamy G, Duraisamy S, Chauhan VS and Gupta D (2006). ProtRepeatsDB: a database of amino acid repeats in genomes. BMC Bioinformatics (database) 7:336.

Kane CM, Cervi L, Sun J, McKee AS, Katherine SM, Sagi S, Christopher AH, Edward JP (2004). Helminth Antigens Modulate TLR-Initiated Dendritic Cell Activation. J. Immunol. 173(12):7454-61.

Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ (2002). Amino acid runs in eukaryotic proteomes, disease associations. Proc. Natl. Acad. Sci. USA 99:333-338.

Katsir LE, Schilmiller AL, Staswick PE, He SY, Howe GA (2008). COI1 is a critical component of a receptor for jasmonate and the bacterial virulence factor coronatine. Proc. Natl. Acad. Sci. 105(19): 7100-7105

Kim KH, Willger SD, Park SW, Puttikamonkul S, Grahl N (2009). TmpL, a Transmembrane Protein Required for Intracellular Redox Homeostasis and Virulence in a Plant and an Animal Fungal Pathogen. PLoS Pathog 5(11): e1000653.

Lin YL, He S (2006). Sm22.6 antigen is an inhibitor to human thrombin. Mol. Biochem. Parasitol. 147(1):95-100.

Lopez Quezada LA, McKerrow JH (2011). Schistosome serine protease inhibitors: parasite defense or homeostasis. Anais da Academia Brasileira de Ciências (Annals of the Brazilian Academy of Sciences) 83(2): 663-672.

MacDonald AS, Araujo MI, Pearce EJ (2002). Immunology of Parasitic Helminth Infections. Infect. Immun. 70(2):427–433.

Matisz CE, McDougall JJ, Sharkey KA, McKay DM (2011). Helminth Parasites and the Modulation of Joint Inflammation. J. Parasitol. Res. 2011:942616.

Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L (2007). New developments in the InterPro database. Nucleic Acids Res. 35: D224-D228.

Nielsen H, Engelbrecht J, Brunak S and von Heijne G (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10:1-6.

Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H et al (2011). MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res. 39:465-474.

Quezada CM, Hicks SW, Galán JE, Stebbins CE (2009). A family of *Salmonella* virulence factors functions as a distinct class of autoregulated E3 ubiquitin ligases. Proc. Natl. Acad. Sci. 106(12): 4864-4869.

Rabia I, El-Ahwany E, El-Komy W, Nagy F (2010). Immunomodulation of Hepatic Morbidity in Murine *Schistosoma mansoni* Using Fatty Acid Binding Protein. J. Am. Sci. 6(7):170-176.

Ramana J, Gupta D (2009). ProtVirDB: a database of protozoan virulent proteins. Bioinformatics 25 (12):1568-1569.

Ramos CR, Figueredo RC, Pertinhez TA, Vilar MM, Nascimento AL et al (2003). Gene structure and M20T polymorphism of the *Schistosoma mansoni* Sm14 fatty acid-binding protein: structural, functional and immunoprotection analysis. J. Biol. Chem. 278:12745-12751.

Ramos CR, Spisni A, Oyama S Jr, Sforca ML, Ramos HR, Vilar MM et al (2009). Stability Improvement of the fatty acid binding protein Sm14 from S mansoni by Cys rep: Structural and functional characterization of a vaccine candidate. J. Biochim. Biophys. Acta 1794(4):655-662.

Reis EAG, Mauadi Carmo TA, Athanazio R, Reis MG, Harn DA Jr (2008). *Schistosoma mansoni* triose phosphate isomerase peptide MAP4 is able to trigger naıve donor immune response towards a type-1 cytokine profile. Scand. J. Immunol. (Clinical Immunology) 68:169–176.

Sharma M, Khanna S, Bulusu G, Mitra A (2009). Comparative modeling of thioredoxin reductase from *Schistosoma mansoni*: a multifunctional target for antischistosomal therapy. J. Mol. Graph Model 27(6):665-675.

Schulz GE, Vogt J (1999). The structure of the outer membrane protein OmpX from *Escherichia coli* reveals possible mechanisms of virulence. Structure 7 (10): 1301–1309.

Sigrist CJA, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N (2010). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. (Database) 38: 161–166.

Tsai CT, Huang WL, Ho SJ, Shu LS, Ho SY (2009). Virulent-GO: Prediction of Virulent Proteins in Bacterial Pathogens Utilizing Gene Ontology Terms. Int. J. Biol. Life Sci. 5(4):2009

Verjovski-Almeida S, DeMarco R (2008). Current developments on *Schistosoma* proteomics. Acta Tropica 108:183-185.

World Health Organisation (WHO) Document (2010). Parasitic Diseases-Schistosomiasis. Available at http://who.int/vaccine_research/diseases/soa_parasitic/en/index5.html.Accessed 13 July 2011.

Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T (2007). MvirDB—a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res. (database) 35:391–394.