Inferences on the biochemical and environmental regulation of universal stress proteins from Schistosomiasis parasites

Andreas N Mbah1,2
Ousman Mahmud1
Omotayo R Awofolu2
Raphael D Isokpehi1
1Center for Bioinformatics and Computational Biology, Department of Biology, Jackson State University, Jackson, MS, USA; 2Department of Environmental Sciences, College of Agriculture and Environmental Sciences, University of South Africa, Pretoria, South Africa

Background: Human schistosomiasis is a freshwater snail-transmitted disease caused by parasitic flatworms of the Schistosoma genus. Schistosoma haematobium, Schistosoma mansoni, and Schistosoma japonicum are the three major species infecting humans. These parasites undergo a complex developmental life cycle, in which they encounter a plethora of environmental signals. The presence of genes encoding the universal stress protein (USP) domain in the genomes of Schistosoma spp. suggests these flatworms are equipped to respond to unfavorable conditions. Though data on gene expression is available for USP genes, their biochemical and environmental regulation are incompletely understood. The identification of additional regulatory molecules for Schistosoma. USPs, which may be present in the human, snail, or water environments, could also be useful for schistosomiasis interventions.

Methods: We developed a protocol that includes a visual analytics stage to facilitate integration, visualization, and decision making, from the results of sequence analyses and data collection on a set of 13 USPs from S. mansoni and S. japonicum.

Results: Multiple sequence alignment identified conserved sites that could be key residues regulating the function of USPs of the Schistosoma spp. Based on the consistency and completeness of sequence annotation, we prioritized for further research the gene for a 184-amino-acid-long USP that is present in the genomes of the three human-infecting Schistosoma spp. Calcium, zinc, and magnesium ions were predicted to interact with the protein product of the gene.

Conclusion: Given that the initial effects of praziquantel on schistosomes include the influx of calcium ions, additional investigations are required to (1) functionally characterize the interactions of calcium ions with the amino acid residues of Schistosoma USPs; and (2) determine the transcriptional response of Schistosoma. USP genes to praziquantel. The data sets produced, and the visual analytics views that were developed, can be easily reused to develop new hypotheses.

Keywords: ATP binding protein, calcium, functional sites, praziquantel, Schistosoma, schistosomiasis

Introduction

Human schistosomiasis is a freshwater snail-transmitted disease caused by parasitic flatworms of the Schistosoma genus.1 Schistosoma haematobium, Schistosoma mansoni, and Schistosoma japonicum are the three major species infecting humans. Schistosomiasis has been designated as one of the “neglected tropical diseases” of poverty and is the second most significant tropical disease, after malaria, in public health significance.2 The large majority of human schistosomiasis and most of the severest disease states are now concentrated in the relatively resource-poor countries of sub-Saharan Africa, contributing to approximately 280,000 deaths per annum.3 Schistosoma-
miasis is also among the severest parasitic diseases targeted, in terms of morbidity and mortality, and has been highlighted for control by the World Health Organization (WHO), with the urinary form highly associated with increased risks for bladder cancer.24 The drug of choice for treatment of schistosomiasis is praziquantel (PZQ), but there is great concern regarding effective treatment in affected communities, due to the potential for parasite resistance to PZQ.4–6

The genomes of *S. haematobium*, *S. mansoni*, and *S. japonicum* encode proteins with the universal stress protein (USP) domain (Pfam Identifier: PF00582).7–9 The USPs are known to function during unfavorable environmental conditions, including the life cycle developmental stages in *Schistosoma* spp.10–12 Though data on gene expression is available for genes encoding USPs (USP genes), their biochemical and environmental regulation are incompletely understood.10–19 The identification of additional regulatory molecules for *Schistosoma* USPs, which may be present in the human, snail, or water environments, could also be useful for schistosomiasis interventions. Our hypothesis is that the USPs of *Schistosoma* spp. have shared protein sequence features that can help us infer their biochemical and environmental regulation. Further, in the context of an infectious disease caused by geographically dispersed species of the same genus, proteins that have shared features could be targets for intervention.

The USPs are found in a diverse group of organisms, including archaea, bacteria, yeast, fungi, and plants, and encompass a conserved group of proteins whose expressions are triggered by a variety of environmental insults, including toxic chemicals, drought, and extreme temperature.20 Genes encoding USPs have not been identified in the human genome, thus making them attractive as drug targets.21 In a previous report,22 we analyzed the developmental expression of eight USP genes predicted from the *S. mansoni* genome. In the case of *S. japonicum*, multiple research investigations have detected developmental stage expression of USP genes.10,11,16

The *Schistosoma* spp. undergo a complex developmental life cycle that includes multiple morphological stages and transition between hosts. The life cycle developmental forms of *Schistosoma* spp. include egg, miracidium, sporocyst, cercaria, schistosomulum, and adult (male and female). These stages must survive diverse stress conditions. The eggs, miracidia, and cercariae are found outside the human and snail hosts and are thus exposed to the stress conditions associated with the freshwater environment of the snail vectors. The cercariae and sporocysts are found in the snail host and must respond to toxic substances, causing oxidative and nitrosative stresses in the snail hemocytes.22,23 In the human host, to develop to the adult form, the schistosomula migrate through multiple organs, including lungs, heart, and liver. These organ systems have defense mechanisms, including production of nitric oxide and hydrogen peroxide, designed to kill the parasite stages.24,25 In summary, all the developmental stages of the *Schistosoma* spp. are exposed to various stresses. A common aspect of these environment-inducing stresses is that they result in proteins with nonnative conformations.26 Inducible stress tolerance has increasingly been understood to result from numerous molecular mechanisms, such as heat shock proteins (Hsps) and USPs.12,27–31

Adenosine triphosphate (ATP) binding is a biochemical mechanism that regulates the function of USPs through phosphorylation.32 Members of the USP family can be categorized into two groups based on the presence or absence of the ATP-binding motif G-2xG-9xG(S/T) in their amino acid sequence.20,33 USPs can be phosphorylated on serine and threonine residues by phosphate donors ATP and guanosine triphosphate (GTP), in the absence of other proteins, coupled with an upregulation response to stressors.34 This observation indicates that in addition to environmental-stressor-mediated regulation, other cellular factors could modulate the activity of USPs by controlling their phosphorylated state.35 Given the broad range of resistance functions conferred by the USPs, it is not surprising that they are encoded in the genomes of a variety of both pathogens and nonpathogens.29,36–38

We have determined protein sequence length, protein domain length, ligand-binding sites, biologically relevant chemical ligands, enzymatic regulation, developmental regulation, and subcellular localization for 13 *Schistosoma* USP sequences (five *S. mansoni* and eight *S. japonicum* sequences). Multiple sequence alignment and phylogenetic analysis provided the evolutionary groupings to allow inferences on biochemical and environmental regulation of the proteins. The results from the analyses were integrated and visualized with visual analytics software. Multiple sequence alignment identified conserved sites of aspartate (Asp), glycine (Gly), histidine (His), leucine (Leu), and proline (Pro) residues in all the sequences. These residues could be key residues regulating the function of the USPs of *Schistosoma* spp. We prioritized a group of two 184-amino-acid-long USP sequences (Q86DW2 [S. japonicum] and G4LZ13 [S. mansoni]) because they had identical values for multiple annotation features. Data visualization revealed the two proteins have identical values for subcellular localization, ligand-binding sites, chemical ligands, and enzymatic
regulation. Specifically, calcium, zinc, and magnesium ions were predicted to interact with the two proteins. Given that the initial effects of PZQ on schistosomes include the influx of calcium ions, additional investigations are required to (1) functionally characterize the interactions of calcium ions with the amino acid residues of S. USPs; and to (2) determine the transcriptional response of Schistosoma USP genes to PZQ.

Methods
Overview of bioinformatics and visual analytics methods
A variety of limitations, including costs, preclude the functional characterization of all predicted proteins from a genome sequencing project. The selection of proteins for further research is a decision-making process by a researcher or research team. Thus, we developed a protocol that integrates the visual analytics stages, to facilitate the interaction with the results from sequence analysis and data collection, on a set of USPs from *S. mansoni* and *S. japonicum*. Visual analytics is an iterative process conducted via visual interfaces that involves collecting information, data preprocessing, knowledge representation, interaction, and decision making.

The overview of the bioinformatics and visual analytics methods is summarized in Figure 1. The protocol consists of five stages that start with the protein sequences to be investigated (Stage 1). Two sets of bioinformatics analyses are performed (Stage 2 and Stage 3). Stage 2 consists of analyses done on each sequence, while, in Stage 3, all the sequences are used for multiple sequence alignment and to construct phylogenetic trees. The bioinformatics sequence analyses in Stage 2 determine (1) the protein sequence length; (2) the protein domain length; (3) the ligand-binding sites; (4) chemical ligand binding; (5) kinase binding; and (6) subcellular localization. These analyses can be particularly useful for prioritizing sequences for research on the biochemical and environmental regulation of proteins. Information on the developmental expression of gene transcripts can assist in deciding the choice of life cycle parasite form to investigate. The data on the developmental expression of gene transcripts were obtained from publications. Multiple sequence alignment and phylogenetic trees provided the statistically and evolutionary support for groupings of the sequences.

The prioritization process was done in Stage 4, with the criterion determined by the researchers. In this report, the criterion was to identify pairs of protein sequences (one from *S. mansoni* and another from *S. japonicum*) that share identical annotations, from the following analyses: protein sequence length, ligand-binding sites (amino acid type and amino acid position), and chemical ligands that are predicted to bind. Stage 5 was the product of the prioritization process. In this report, we expect (i) orthologous pairs of protein sequences and (ii) a visualization that provides an integrated view of the shared annotations. We used a visual analytics software package (Tableau 7.0, Tableau Software Inc, Seattle, WA, USA) to perform several visual analytics tasks, including interaction, computing, analysis, integration, and visualization.

**Retrieval of protein sequences**
Proteins annotated with the USP domain (PF00582) from the *S. mansoni* and *S. japonicum* genomes were identified in Universal Protein Resource (UniProt release 2011_11: http://www.uniprot.org/). Predicted protein sequences were retrieved. The final list of protein sequences for comparative sequence analysis was determined by the revision history of the sequence in the UniProt as well as by entries in GeneDB and SchistoDB.
Conserved domain search for functional sites

The search for amino acid residues that are functionally important was performed using two public servers. A single-sequence input server (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and a multiple-sequence input or batch server (http://www.ncbi.nlm.nih.gov/Structure/bwrpsb/bwrpsb.cgi) were used in finding conserved domains for the sequences. The ATP-binding motif residues and other ligand-binding residues were identified and documented for all the USP sequences, including their domain architecture. To facilitate comparison of the functional sites, we constructed a functional site signature for each sequence. The signature is a string of the amino acid letters. In a case where no site is predicted in the 12-letter signature, the position was assigned "X." Therefore, for protein sequence MJ0577, the template for the ligand-binding sites, the signature is PTDVMGHGGGSVT. This approach of constructing functional sites has been implemented in previous research on functional sites.31

Prediction of chemical ligand and enzymatic regulation

The three-dimensional (3D) chemical ligands were predicted using the 3DLigandSite server (http://www.sbg.bio.ic.ac.uk/3dligandsite).52 These biologically relevant chemical ligands are potential regulators of the function of the USPs. The 3DLigandSite is a top-performing web server for chemical ligand prediction and provides structural models for unsolved proteins, using protein-structure prediction. The specific kinases were predicted using the NetPhosK 1.0 tool (http://www.cbs.dtu.dk/services/NetPhosK),33 with a stringent threshold value set at 0.65. The kinase with the highest threshold value was selected.

Prediction of subcellular location

The subcellular locations of all the genes were retrieved from literature, databases or predicted if possible using the server Euk-mPLoc 2.0 (http://www.csbio.sjtu.edu.cn/bioinf/euk-multi-2/).54 The prediction by Euk-mPLoc2.0 is based on integrating information from gene ontology, functional domain, and evolutionary relationships.

Compilation of developmental expression of genes

The developmental stage expression profiles of the selected Schistosoma USP genes were extracted from our previous publication22 for S. mansoni. In the case of S. japonicum, expression profiles were from the SjTPdb, an integrated transcriptome and proteome database and analysis platform for S. japonicum.53 The database contains expressed sequence tags (ESTs), EST clusters, and the proteomic dataset for S. japonicum.

Prediction of evolutionary relatedness of sequences

Evolutionary relatedness of the selected USPs, from both S. mansoni and S. japonicum, was determined to ascertain their functional and evolutionary relationship. The sequences were aligned using the ClustalW tool (http://www.ch.embnet.org/software/ClustalW.html),56 applying the default settings. The evolutionary relationship of the sequences was inferred using the maximum likelihood method,57 based on the JTT matrix-based model58 at 1000 bootstrap, with MEGA software version 5 (http://www.megasoftware.net/ [Center for Evolutionary Medicine and Informatics, Tempe, AZ, USA]).59 The bootstrap test indicated above, at 1000 replications, was used to determine the percentage of the replicate trees, in which the genes clustered together.60,61

Visual analytics of datasets

A purpose of Stage 4 of the protocol was to provide an integration and visualization portal for the results of the bioinformatics analysis in Stage 2 and Stage 3 (Figure 1). The visual analytics tasks to be performed on the data sets can be influenced by how the data is organized in the data records (rows) and data fields (columns) in the data source (eg, spreadsheet file and comma delimited file). For Stage 2, each data record had the following data fields: (1) organism; (2) locus tag; (3) UniProt ID; (4) feature; and (5) feature value. The feature field had the following types: protein domain length, protein length, protein domain start position, protein domain end position, ATP-binding motif, kinase type, kinase type score, 3D chemical ligand, ligand-binding amino acid, amino acid and sequence position, developmental expression, and subcellular localization. In the case of the functional site signature data set, each record consisted of the UniProt ID and 12 fields for each of the 12-letter signatures for the USP ligand-binding sites. For the Stage 3 data set (phylogenetic tree groupings), each data record consisted of data fields for (1) organism; (2) UniProt ID; and (3) phylogenetic group. The data sources (in this case, spreadsheet files) were loaded to the visual analytics software for the visual analytics tasks, including the design of the data integration and visualization.
Table 1: Annotation features for universal stress proteins of *Schistosoma mansoni* and *Schistosoma japonicum*

<table>
<thead>
<tr>
<th>Organism</th>
<th>UniProt ID</th>
<th>Locus tag</th>
<th>Frequency per protein</th>
<th>Predicted subcellular localization</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. japonicum</em></td>
<td>QSDDH7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QSDED2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QSDG9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QSDH4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QSDHK1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>QSDD16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q86DW2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Q86DX1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>S. mansoni</em></td>
<td>C1MOQ2</td>
<td>Smp_097930</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G4LZ21</td>
<td>Smp_076400</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G4V5S2</td>
<td>Smp_001000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G4VW9</td>
<td>Smp_043120</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>G4VPM6</td>
<td>Smp_031300</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Results

**Data set for visual analytics**

The dataset analyzed consisted of 13 USP sequences from the two species (*S. japonicum* and *S. mansoni*). The sequences were grouped by protein length and domain length. For the protein sequence grouping, eight distinct protein sequences from both species are observed. These groups are based on the evolutionary relatedness of the sequences. The additional bioinformatics predictions were performed using multiple sequence alignment.

### The 5 annotation features with variable frequency per protein sequence

- **3D chemical ligand**
- **Ligand binding motif**
- **Amino acid and sequence position**
- **Developmental expression**
- **Subcellular localization**

These features were shared by the two species. The amino acid sequence position, developmental expression, and subcellular localization were particularly conserved in shared annotations that can help us infer the joint biochemical and environmental regulation of the USPs from the two pathogenic *Schistosoma* spp.

### Notes

- **Organism:** *S. japonicum* and *S. mansoni*
- **Feature value:** Protein domain length (aa), Protein length (aa), Protein domain start position, Protein domain end position, ATP-binding motif, Kinase type, Kinase type score, 3D chemical ligand, Ligand binding amino acid, Amino acid and sequence position, Developmental expression, Subcellular localization
- **Frequency per protein:** 5, 12, 12, 5, 3, 1, 4
- **Predicted subcellular localization:** 1, 2, 6, 1, 1, 3, 2

Note: Sequences of *S. mansoni* have "Smp" in the sequence identifier.

Abbreviations: aa, amino acid; PKA, protein kinase A; PKC, protein kinase C; UniProt, Universal Protein Resource (Apweiler et al.).
length of 1 aa, as in the 184 aa USPs (Q86DW2 and G4LZI3), or 2 aa, as in the 159 aa USPs (Q5DG19 and C1M0Q2).

Functional site signatures of Schistosoma USP sequences

The Conserved Domain Search tool at the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) uses the fragment A (UniProt Identifiers: Q57997; Y577_METJA) of the USP MJ0577 of Methanocaldococcus jannaschii as the template to align the Schistosoma USPs for the conserved protein domain search. The search reports the ligand-binding sites (functional sites) including the ATP-binding motif [G-2X-G-9X-G(S/T)] present in the query sequence. A total of 12 ligand-binding sites are predicted for MJ0577. The amino acids and their position are Pro11, Tyr12, Asp13, Val41, Met126, Gly127, His129, Gly130, Gly140, Ser141, Val142, and Thr143. The Asp13 and Val41 are binding sites for adenosine nucleoside. The amino acid sequences from met 126 to Thr143 contain the motif G2xG9xG(S/T), which includes binding sites for the phosphoryl and ribosyl groups of ATP. Among the 13 Schistosoma sequences with ATP-binding motif, eight functional site signatures (AIDA1GRGGGSVS, AIDA5GRGGGSVS, PIDV1GRGGGSVS, PIDVMGRGGGSVS, PVDI1GRGGGSVS, PVDVMGRGGGSVS, and XXXVMGRGGGSVS) were observed (Figure 3). The last seven letters (GRGGGSVS) of the signature, which corresponds to the ATP-binding motif, were identical for all the signatures of the Schistosoma USPs. Three shared signatures were observed for the two species. However, only one signature (PVDI1GRGGGSVS) had the same members as in the protein length grouping (184 aa: Q86DW2, G4LZI3).

Figure 3 Grouping of 13 Schistosoma universal stress proteins by functional site signature.

Notes: The functional site signature is constructed by joining the twelve ligand binding sites known for the ATP-binding USP from Methanocaldococcus jannaschii (UniProt [Apweiler et al] ID: Y577_METJA). The image provides a visual comparison of the functional site signatures for 13 Schistosoma USPs. A visual analytics resource that can be used for interacting with the data is available at http://public.tableausoftware.com/views/schisto_features_usp/groupbylength. Sequences of Schistosoma mansoni have “Smp” in the sequence identifier.

Abbreviations: ATP, adenosine triphosphate; UniProt, Universal Protein Resource; USP, universal stress protein.
The grouping of sequences by the maximum parsimony was 100% bootstrap statistical support value for the branch of Group A and E contained multiple five (A to E) of sequences were observed with each of the and 176 (Figure 4). The relationship of the sequences was based on the primary amino acid sequences of the USPs. To facilitate dynamic integration and updates of the data.
A visual analytics resource that can be used to view the image, with other associated annotations when compared with Group C (Figure 8). The biologically relevant chemical ligands predicted to bind to the Group C proteins include four phosphate-containing ligands (adenosine diphosphate [ADP], adenosine monophosphate [AMP], adenosine triphosphate [ATP] and guanosine triphosphate [GTP]) and three metallic ion ligands (calcium [Ca$^{2+}$], magnesium [Mg$^{2+}$], and zinc [Zn$^{2+}$]). The two proteins in the group were also predicted to be (1) localized in the cytoplasm and (2) capable of phosphorylation by phosphokinase C, with a value of 0.93. In the Group C USP genes, there was evidence of gene expression in all the stages by at least one of the genes. The schistosomulum stage had the only identical annotation for the developmental gene expression for the two Group C USP genes. A screenshot showing a design that provides an integrated view of the chemical ligands, ligand-binding sites, functional site signature, the presence of ATP-binding motif, kinase type, and kinase score is presented in Figure 9.

### Discussion

*S. haematobium*, *S. japonicum*, and *S. mansoni* are the major human schistosomiasis parasites. These parasites undergo a complex developmental life cycle, in which they encounter a plethora of environmental stressors, such as transition from aerobic to anaerobic environment during the cercarial penetration of the human skin. The presence of genes encoding the USP domain in the genomes of *Schistosoma* spp. suggests these flatworms are equipped to respond to unfavorable conditions that induce USP function. The bioinformatics-based predictions generated a variety of data types, including amino acid functional site, multiple sequence alignment, prediction score, protein domain organization, phylogenetic tree, and sequence length. We used a visual analytics approach...
### Figure 7 Integration and visualization of the data on the sequence features, evolutionary relatedness, and developmental expression of Schistosoma universal stress proteins (Q86DW2 and G4LZI3).

**Notes:** The integration and visualization design was implemented in the visual analytics software environment (Tableau Software Inc, Seattle, WA, USA). Among the 13 sequences compared, the two 184-amino-acid-long sequences Q86DW2 (Sjp_0058490) and G4LZI3 (Smp_076400) were prioritized for further research. The decision was based on statistical support from the phylogenetic analysis as well as the relatively complete and consistent annotations in the protein sequence length, biologically relevant chemical ligands, and ligand-binding amino acids (amino acid type and amino acid position). A visual analytics resource that can be used to interact with the view is available at [http://public.tableausoftware.com/views/schisto_features_usp/phylo_group](http://public.tableausoftware.com/views/schisto_features_usp/phylo_group). Sequences of *S. mansoni* have “Smp” in the sequence identifier.

**Abbreviations:** ADP, adenosine diphosphate; AMP, adenosine monophosphate; ATP, adenosine triphosphate; CA, calcium; D, aspartate; G, glycine; GTP, guanosine triphosphate; I, isoleucine; Mg, magnesium; PKC, protein kinase C; P, proline; R, arginine; S, serine; UniProt, Universal Protein Resource (Apweiler et al); V, valine; Zn, Zinc.

### Table 1: Phylogenetic group and features of Schistosoma universal stress proteins

<table>
<thead>
<tr>
<th>Phylogenetic group</th>
<th>Organism</th>
<th>UniProt ID</th>
<th>Locus tag</th>
<th>Protein length (aa)</th>
<th>Domain length (aa)</th>
<th>Domain start</th>
<th>Domain end</th>
<th>Developmental expression</th>
<th>Predicted subcellular localization</th>
<th>3D chemical ligand</th>
<th>Ligand binding amino acid (amino acid type and position)</th>
<th>ATP binding motif</th>
<th>Kinase type</th>
<th>Kinase score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group C</td>
<td>Schistosoma japonicum Q86DW2</td>
<td></td>
<td></td>
<td>155</td>
<td>147</td>
<td>7</td>
<td>148</td>
<td>Adult</td>
<td>Cytoplasm</td>
<td>ADP, AMP, ATP, MG, ZN, D, G, I, M, P, R, S, T, V</td>
<td></td>
<td></td>
<td></td>
<td>PKC</td>
</tr>
<tr>
<td></td>
<td>Schistosoma mansoni C4LZI3 Smp_076400</td>
<td></td>
<td></td>
<td></td>
<td>149</td>
<td>132</td>
<td>51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group D</td>
<td>Schistosoma japonicum Q86DX1</td>
<td></td>
<td></td>
<td>159</td>
<td>147</td>
<td>7</td>
<td>154</td>
<td>Adult</td>
<td>Female</td>
<td>ADP, AMP, ATP, MG, ZN, D, G, I, M, P, R, S, T, V</td>
<td></td>
<td></td>
<td></td>
<td>PKA, PKC</td>
</tr>
<tr>
<td></td>
<td>Schistosoma mansoni C1M0Q2 Smp_097930</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Figure 8 Integration and visualization of the data on the sequence features, evolutionary relatedness, and developmental expression of Schistosoma universal stress proteins (Q86DX1 and C1M0Q2).

**Notes:** This figure illustrates the decision-making process. In comparison with Q86DW2 and G4LZI3 (Figure 7), the annotations for the protein sequence length, biologically relevant chemical ligands, and ligand-binding amino acids (type and position) were not identical. A visual analytics resource that can be used to interact with the view is available at [http://public.tableausoftware.com/views/schisto_features_usp/phylo_group](http://public.tableausoftware.com/views/schisto_features_usp/phylo_group). Sequences of *S. mansoni* have “Smp” in the sequence identifier.

**Abbreviations:** ADP, adenosine diphosphate; AMP, adenosine monophosphate; ATP, adenosine triphosphate; D, aspartate; G, glycine; GTP, guanosine triphosphate; I, isoleucine; M, methionine; Mg, magnesium; PKA, protein kinase A; PKC, protein kinase C; P, proline; R, arginine; S, serine; T, threonine; UniProt, Universal Protein Resource (Apweiler et al); V, valine; Zn, Zinc.
Figure 9 Design layout and visualization of data sets from the sequence analysis, evolutionary relatedness, and developmental expression of 13 Schistosoma universal stress proteins.

Notes: The details of the annotation features are available in the Methods section. The views constructed and data are available for download from an Internet website: http://public.tableausoftware.com/views/schisto_features_usp/integrated_view. The free software Tableau Reader (http://www.tableausoftware.com/products/reader) (Tableau Software Inc) can be used for offline access to the downloaded views and data.

Abbreviations: ADP, adenosine diphosphate; AMP, adenosine monophosphate; ATP, adenosine triphosphate; CA, calcium; D, aspartate; g, glycine; gTP, guanosine triphosphate; I, isoleucine; M, methionine; Mg, Magnesium; PKA, protein kinase A; PKC, protein kinase C; P, proline; R, arginine; S, serine; T, threonine; UniProt, Universal Protein Resource (Apweiler et al); V, valine; Zn, Zinc.

to integrate these data types and to identify orthologous pairs of protein sequences with a protein length of 184 aa USP in S. mansoni (Smp_076400) and S. japonicum (Locus Tag: Sjp_0058490; UniProt ID: Q86DW2). Gene synteny, obtained from SchistoDB and evolutionary genomics analysis called the S. mansoni phylome, indicated that an ortholog (Sha_107834) is encoded in the S. haematobium genome. Thus, the genomes of the three major human schistosomiasis parasites encode the 184 aa USP.

Since inferences on chemical and environmental regulation are our interest, we focus the discussion of the results on the findings on the five conserved residues and the chemical ligands predicted to bind to the prioritized protein. All the 13 protein sequences have conserved sites for Asp, Leu, Gly, His, and Pro at positions 57, 101, 127, 166, and 176, using Smp_076400 from S. mansoni as a reference sequence (Figure 4). These conserved residues did not coincide with any of the predicted ligand-binding sites and could be common functional sites for regulating Schistosoma USPs.

The predicted 3D chemical ligands for Smp_076400 and Sjp_0058490 (Q86DW2) included three metal ions Ca\(^{2+}\), Mg\(^{2+}\), and Zn\(^{2+}\) (Figure 7). Metal ions are involved in many diverse biochemical reactions, including cellular cofactors for phosphorylation. The UspA protein of Escherichia coli undergoes phosphorylation in vitro with its phosphate donors ATP and/or GTP, in the absence of other proteins. The ATP molecule and metallic chemical ligands, such as Mg\(^{2+}\) ion, might bind together at the Mg-ATP-binding groove during phosphorylation or ATP-dependent stress-response mechanism. The presence of Mg\(^{2+}\) ion suggests that it can be an integral and critical component in the reaction. This result could be affected if there is any structural conformation in the binding site residues that prevents the Mg\(^{2+}\) ion from binding to the ATP molecule at the active groove. The resultant effect might be translated to compromised functional efficiency in binding ATP during phosphorylation and also in keeping the metallic Mg ions unstable in the active groove while it is in contact with the ATP molecule.

Ca\(^{2+}\) was predicted to bind to proteins in Group C of the phylogenetic tree (Figures 6 and 7). In S. mansoni, Ca\(^{2+}\) is considered vital for regulated motor-related activities and also critical for the egg hatching process in fresh water.
In the tegument fraction of *S. mansoni*, Ca\(^{2+}\) simulated the activity of ATPase in the absence of Mg\(^{2+}\).\(^{73}\) Further, cyclic adenosine monophosphate and Ca\(^{2+}\) work in synergy to regulate the transformation of miracidial to sporocysts.\(^{76}\) The protein kinase C and Ca\(^{2+}\) metabolism regulate the induction of proteolytic enzyme from cercariae, which is vital for modulating the musculature activity of the schistosome.\(^{77,78}\) A key mechanism for the action of PZQ has been proposed to be the disruption of the Ca\(^{2+}\) homeostasis in schistosomes, leading to the large, rapid influx of Ca\(^{2+}\) ions into the worm and quick muscular contractions.\(^{41,79,81}\) Microarray-based transcriptome analysis of the response of the *S. mansoni* PR-1 strain to PZQ has identified genes for cytosolic Ca\(^{2+}\) regulation.\(^{82}\)

**Conclusion**

*S. haematobium*, *S. mansoni*, and *S. japonicum* are human parasites that undergo a complex developmental life cycle, in which they encounter a plethora of environmental stressors. Though there are multiple research reports on the developmental regulation of genes encoding USPs in *Schistosoma* spp., knowledge of their biochemical and environmental regulation is still limited. The draft status of the genome sequences of *Schistosoma* spp. also provides possibilities that future revisions could be made to gene prediction and protein annotations. We have used a decision-making strategy, facilitated by visual analytics, to identify USPs in two *Schistosoma* species with shared sequence features and when compared with the other sequences they have relatively complete and consistent annotations. These findings further enabled us to make inferences about the biochemical and environmental regulation of *Schistosoma* USPs. Future research directions could (1) functionally characterize the interactions of Ca\(^{2+}\) ions with the amino acid residues of *Schistosoma* USPs; and (2) determine the transcriptional response of *Schistosoma* USP genes to PZQ. The data sets produced, and the visual analytics views developed, can be easily reused to develop new hypotheses.

**Acknowledgments**

National Institutes of Health: Research Centers in Minority Institutions (RCMI) – Center for Environmental Health at Jackson State University (NIH-NCCR 2G12RR013459); Mississippi Idea Network for Biomedical Research Excellence (NIH-NCCR-P20RR016476 and NIH-NIGMS-8P20GM103476); Arkansas Idea Network for Biomedical Research Excellence (NIH-NCCR-P20RR016460); Bioinformatics Programs in Minority Institutions (1T36GM095353); and National Center for Integrative Biomedical Informatics (U54DA021519). National Science Foundation: Mississippi NSF-EPSCoR Grant Awards (EPS-0903787 and EPS 1006883); Undergraduate Research and Mentoring Program (DBI-0958179); and Visual Analytics in Biology Curriculum Network (DBI-1062057). US Department of Homeland Security Science and Technology Directorate (2009-ST-062-000014; 2011-ST-062-000048). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agencies.

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


