Schistosoma mansoni is one of the three main causative agents of human schistosomiasis, a major health problem with a vast socio-economic impact. Recent advances in the proteomic analysis of schistosomes have revealed that peptidases are the main virulence factors involved in the pathogenesis of this disease. In this context, evolutionary studies can be applied to identify peptidase families that have been expanded in genomes over time in response to different selection pressures. Using a phylogenomic approach, we searched for expanded endopeptidase families in the S. mansoni predicted proteome with the aim of contributing to the knowledge of such enzymes as potential therapeutic targets. We found three endopeptidase families that comprise leishmanolysins (metallopeptidase M8 family), cercarial elastases (serine peptidase SI family) and cathepsin D proteins (aspartic peptidase A1 family). Our results suggest that the Schistosoma members of these families originated from successive gene duplication events in the parasite lineage after its diversification from other metazoans. Overall, critical residues are conserved among the duplicated genes/proteins. Furthermore, each protein family displays a distinct evolutionary history. Altogether, this work provides an evolutionary view of three S. mansoni peptidase families, which allows for a deeper understanding of the genomic complexity and lineage-specific adaptations potentially related to the parasitic lifestyle.

Key words: phylogenomics - maximum likelihood analysis - homology prediction - functional annotation - proteases - paralogous families - parasite genomics

Schistosomiasis, which is caused by different species from the Schistosoma genus, remains one of the most prevalent tropical neglected diseases, affects 210 million people worldwide, and is responsible for at least 280,000 deaths every year (van der Werf et al. 2003, Steinmann et al. 2006, Han et al. 2009). Schistosoma mansoni is one of the three major species that infect humans and is the causative agent of intestinal and hepatic schistosomiasis mainly in Africa and South America (Han et al. 2009). Measures to control schistosomiasis rely almost entirely on praziquantel®, which is the only drug available for mass chemotherapy. Despite the effectiveness of this treatment, re-infection is common and drug-resistant parasites have been found in the laboratory and in the field, which demonstrate the urgent need to develop additional chemotherapeutic agents and effective vaccines (Liang et al. 2003, Pica-Mattoccia & Cioli 2004, Botros & Bennett 2007, Melman et al. 2009).

Over the past several years, advances in the molecular analysis of major parasites have identified some key factors involved in parasitic diseases and peptidases as one of the major factors of pathogenicity (McKerrow et al. 2006, Kasný et al. 2009). These enzymes have been implicated in processes that are crucial to the development and survival of helminth parasites, including digestion, invasion from host tissues, activation of inflammation and evasion of the host immune system (McKerrow et al. 2006, Kasný et al. 2009).

Peptidases (also termed proteases, proteinases or proteolytic enzymes) are hydrolytic enzymes that cleave peptide bonds in proteins. Endopeptidases cleave internal peptide bonds, whereas exopeptidases hydrolyse the amino terminus (aminopeptidases) or carboxy terminus (carboxypeptidases) of different proteins. Enzymatic specificity is determined based on the chemical groups responsible for catalysis in the peptide’s active site. Thus, peptidases are classified into one of the following classes: asparagine, aspartic, cysteine, glutamic, metallo, serine, threonine and unknown peptidases (Rawlings & Barrett 1993, Rawlings et al. 2010).

Asparagine peptidases are enzymes that have active sites composed of an aspartic acid and an asparagine, the latter being the P1 residue, the amino acid or molecule, which can be found at a specific location in the cleavage site (Rawlings et al. 2010). In turn, aspartic peptidases have their catalytic centres formed by two aspartate residues that activate a water molecule that mediates the nucleophilic attack on the peptide bond (James 2004, Rawlings...
In general, cysteine peptidases have cysteine and histidine residues forming their “catalytic dyad”. Meanwhile, other active site residues have been found. Glutamic peptidases have glutamic acid residues as their primary catalytic residues, which are probably the nucleophilic attack mediators involved in the catalysis (Fujinaga et al. 2004, Rawlings et al. 2010). In metallopeptidases, the catalytic mechanism usually involves a single catalytic zinc ion tetrahedrally coordinated by one glutamate and two histidine residues (Rawlings et al. 2010). Serine peptidases have serine residues at their active sites, which together with two other variable amino acids constitute the “catalytic triad” (Hedstrom 2002, Rawlings et al. 2010). Threonine peptidases have threonine residues as their nucleophiles during catalysis. For unknown peptidases, the active site residues have not yet been determined.

Evolutionary analyses have been applied to a broad range of studies, which include the identification of gene/protein families that have expanded in a specific lineage over evolutionary time and possibly indicate the existence of selective pressure (Irving et al. 2003, Sargeant et al. 2006, Nahum & Pereira 2008, Robinson et al. 2008, Wu et al. 2009, Huzurbazar et al. 2010). The availability of faster and more powerful computers combined with the development of automated pipelines has enabled the investigation of such evolutionary processes through the reconstruction of phylogenetic trees for the complete set of proteins encoded in a genome (known as phylome). The results obtained by this analysis provide a broad view of the evolution of an organism’s genome and proteome, which allows for a deeper understanding of genomic complexity and lineage-specific adaptations (Huerta-Cepas et al. 2007, 2010b).

In a previous study, we described the reconstruction of the S. mansoni phylome to improve gene/protein functional annotation and provide insights into parasite’s biology (phylomedb.org). By applying an automated pipeline, we also identified lineage-specific gene duplications, which may have led to a potential diversification of several protein families that are relevant for host-parasite interactions, such as tetraspanins, fucosyltransferases and sperm-coating protein-like proteins. Here, we explore the S. mansoni phylome data to analyse three endopeptidase families that expanded in this lineage since its diversification from 15 other metazoan species with the aim of contributing to the available knowledge of parasite biology and host-parasite interactions from an evolutionary perspective. The members of these families include leishmanolysins (metallopeptidase M8 family), cercarial elastases (serine peptidase S1 family) and cathepsin D proteins (aspartic peptidase A1 family).

The present paper is centred on two main research questions: (i) Did any peptidase families expand in the S. mansoni genome/proteome and if so, which ones? (ii) What are the evolutionary histories of these peptidase families? To address these questions, we used a so-called species-overlap algorithm (Huerta-Cepas et al. 2007) to detect lineage-specific duplications that occurred during the evolution of the parasite’s genome. We also integrated information on sequence alignments, phylogenetic trees, protein architecture and the conservation of critical residues to characterise these proteins. Our results indicate that each peptidase family has a unique evolutionary history within/across the analysed species. Furthermore, our data support the hypothesis that gene duplication events followed by divergence is the main mechanism shaping the evolution of S. mansoni-specific paralogous groups.

The analysis of the evolutionary histories of these three S. mansoni families is relevant to functional genomics, evolutionary biology, medicine and biotechnology, especially taking into account the importance of S. mansoni peptidases in the development of schistosomiasis and that they have been described as promising vaccine and drug targets (McKerrow et al. 2006, Abdulla et al. 2007, Kasný et al. 2009).

**MATERIALS AND METHODS**

**Organisms and sequence data** - The dataset of species selected for analysis includes eight invertebrates (Nematostella vectensis, Caenorhabditis elegans, Caenorhabditis briggsae, S. mansoni, Drosophila melanogaster, Anopheles gambiæae, Bombyx mori and Strongylocentrotus purpuratus), one tunicate (Ciona intestinalis), one cephalochordate (Branchiostoma floridæae), three vertebrates (Danio rerio, Mus musculus and Homo sapiens), three fungi (Neurospora crassa, Saccharomyces cerevisiae and Ustilago maydis) and one plant (Arabidopsis thaliana). Information on the selected taxa is provided as Supplementary data.

This dataset is particularly rich in metazoans (76% of the selected species) that cover important evolutionary innovations, for example, the origin of bilateral symmetry, the third germ layer, the development of organs, systems, complex patterns of communication and the emergence of the adaptive immune system, which makes it especially suitable for addressing the evolutionary innovations in S. mansoni in comparison with other metazoan species (phylomedb.org).

The S. mansoni predicted proteome dataset was downloaded from SchistoDB version 2.0 (schistodb.net) (Zerlotini et al. 2009). Proteomes derived from the 16 fully sequenced genomes were downloaded from the Broad Institute Ustilago maydis Database, Ensembl, Integr8, JGI Genome Projects, National Center for Biotechnology Information Genome Database and SilkDB, which can be collectively accessed through the Genomes OnLine Database (genomesonline.org).

**Endopeptidase protein families** - Peptidases are hydrolases that act on peptide bonds [Enzyme Commission (EC) 3.4]. Three endopeptidase families were selected and analysed in detail in the present work. They include the metallopeptidase M8 family (EC 3.4.24.-), serine peptidase S1 family (EC 3.4.21.-) and aspartic peptidase A1 family (EC 3.4.23.-) members and belong to three peptidase clans (MA, PA and AA, respectively), as described in the MEROPS database (Rawlings et al. 2010).

Information on enzymes was collected from the literature and database references and included in the Supplementary data. The EC numbers were collected from the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology database, which is available online (chem.qmul.ac.uk/iubmb/enzyme/).
**Alignments and phylogenetic trees** - Sequence alignments and phylogenetic trees of the endopeptidase families selected for analysis were retrieved from the *S. mansoni* phylome data, which were reconstructed through a comparative analysis among all proteins encoded by the parasite genome and their potential homologs in 16 other eukaryotic species (phylomedb.org) (Huerta-Cepas et al. 2011).

Briefly, the *S. mansoni* phylome was reconstructed using each protein encoded in the *S. mansoni* genome (“seed” proteins) and the potential homologs identified through similarity-based searches (Smith & Waterman 1981) against the dataset of selected proteome data described above. The groups of homologous sequences were aligned using MUSCLE v3.6 (Edgar 2004) and gap-rich columns were filtered using trimAl (Capella-Gutierrez et al. 2009). Phylogenetic analyses were performed using the neighbour-joining and maximum likelihood (ML) methods, as implemented in PhyML (Guindon & Gascuel 2003).

For the phylogenetic reconstruction of each “seed protein”, we tested four different evolutionary models (JTT, WAG, BLOSUM62 and VT). In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was assumed with the gamma parameter and the fraction of invariant positions estimated from the data. Tree support values were computed using the approximate likelihood ratio test as implemented in PhyML (Guindon & Gascuel 2003, Anisimova & Gascuel 2006). The evolutionary model best fitting the data was determined by comparing the likelihood of the used models according to the Akaike Information Criterion (Akaike 1973). The resulting alignments, phylogenies and homology prediction can be accessed at PhylomeDB (phylomedb.org) (Huerta-Cepas et al. 2011) through protein sequence identifiers (e.g., UniProt: C4PZH6; SchistoDB: Smp_127030; PhylomeDB: Phy000V7EC_SCHMA).

To integrate information from SchistoDB (Zerlotini et al. 2009) and PhylomeDB (Huerta-Cepas et al. 2011), we built a local relational database, named SchistoPhylomeSQL, which allowed us to extract and interpret the large amount of data in this work (Fig. 1). Access to this local database was implemented using DbVisualizer version 7.0.5 (dbvis.com). The SchistoPhylomeSQL database was the main resource for data mining in this work. In-house Perl scripts and Structured Query Language queries were used to parse data files during the database building and searching processes.

**Paralogy and orthology relationships** - To derive a complete catalogue of the paralogy and orthology relationships between *S. mansoni* proteins and those from other eukaryotic proteomes, we applied a “species-overlap” algorithm, as previously described (Huerta-Cepas et al. 2007). This algorithm uses the level of species overlap between the two daughter partitions of a given node to define it as a duplication or speciation event, which give rise to paralogs and orthologs, respectively. Once all the nodes have been classified, the algorithm establishes the paralogy and orthology relationships between the “seed protein” and other proteins included in the phylogenetic tree, according to the original definition of these terms (Fitch 1970, Gabaldón 2008).

**Lineage-specific duplications** - Using a python Environment for Tree Exploration (Huerta-Cepas et al. 2010a), we analysed the *S. mansoni* phylome data (phylomedb.org) to identify protein families that were specifically expanded in the *S. mansoni* lineage since its diversification from the other selected taxa (Supplementary data). The duplication events defined by the “species-overlap” algorithm that only comprised paralogs from *S. mansoni* were considered lineage-specific duplications. In cases where more than one phylogenetic tree contained the same paralogous proteins, by changing only the “seed” protein position, the data were filtered to obtain a non-redundant list of in-paralogs.

**Protein architecture and critical residues** - In this study, we used the Pfam database (Finn et al. 2010) to identify the presence and organisation of protein sequence domains as well as critical residues present in the three *S. mansoni* endopeptidase families. Pfam is a large and widely used database of protein domains families. This database contains multiple sequence alignments and profile hidden Markov models (profile HMMs) for each protein family. Pfam-A entries are derived from the underlying sequence database, which is termed Pfam-seq. This database is built from the most recent release of UniProtKB at a given time point (Finn et al. 2010, Apweiler et al. 2011). To predict active sites in new sequences, Pfam uses the information available in UniProtKB for homologous proteins, whose catalytic residues have been experimentally characterized (Mistry et al. 2007). Based on Pfam information, the illustrations of the *S. mansoni* protein domain architectures were generated using DOG 2.0 (Ren et al. 2009).

![Fig. 1: flowchart of the applied methodology. The Schistosoma mansoni proteome data was retrieved from SchistoDB and each protein was used as “seed” to reconstruct the S. mansoni phylome. The resulting alignments, phylogenies, and homology predictions are available at PhylomeDB. To integrate information from SchistoDB and PhylomeDB, we built the SchistoPhylomeSQL, a local relational database as the main resource for data mining in this work.](image-url)
RESULTS

Comparative genomics has revealed a great deal of sequence and/or functional diversity within and across organisms with respect to gene/protein family size, composition and the relatedness of their members (Huerta-Cepas et al. 2007, 2010b, Nahum et al. 2009, Andrade et al. 2011, Avelar et al. 2011). The rationale underlying the present work is that lineage-specific duplications may reflect molecular biodiversity and that the adaptation of organisms to different environments may ultimately help to identify potential therapeutic targets against parasitic diseases.

Our previous work identified lineage-specific gene duplications that led to the diversification of several families in *S. mansoni* (phylocedb.org). Furthermore, recent advances in proteomic analyses of schistosomes have revealed that peptidases are one of the main virulence factors involved in the pathogenesis of schistosomiasis (McKerrow et al. 2006, Kasný et al. 2009). In this work, we performed a phylogenomic analysis to address the two main questions of (i) whether peptidase families are expanded in the *S. mansoni* proteome and (ii) whether they share similar evolutionary histories.

Endopeptidase family members are duplicated in *S. mansoni* - To investigate which peptidase families are expanded in the *S. mansoni* genome, we explored the parasite phylome data available at PhylomeDB (Huerta-Cepas et al. 2011). Phylogenetic analyses were performed using an automated pipeline and a complete list of the paralogy relationships among the *S. mansoni* proteins was retrieved using a “species-overlap” algorithm that identifies family members originated by lineage-specific duplication events (Huerta-Cepas et al. 2007).

Based on the functional annotation available from the SchistoDB (Zerlotini et al. 2009) and UniProt (Apweiler et al. 2011) databases, the results revealed that the most significant peptidase expansions in the *S. mansoni* proteome corresponded to endopeptidases such as leishmanolysins, cercarial elastases and cathepsin D proteins. These enzymes belong to three distinct endopeptidase families, metalloelastase M8 family (EC 3.4.24.-), serine peptidase S1 family (EC 3.4.21.-) and aspartic peptidase A1 family (EC 3.4.23.-), as described in the MEROPS database (Rawlings et al. 2010) and represent promising targets for vaccine and drug development.

In total, we identified 12 leishmanolysins, 13 cercarial elastases (Supplementary data) and 11 cathepsin D proteins (Supplementary data) in the predicted *S. mansoni* proteome. These proteins vary in length and sequence composition, but they are highly conserved with respect to the presence of a conserved sequence domain, which is distinct for each protein family as defined by the Pfam database (see details below). Currently, no crystal structure has been obtained for the *S. mansoni* peptidases described here.

Leishmanolysin (also called invadolysin) is a major surface peptidase member of the metalloelastase M8 family. Leishmanolysins are believed to share the same mechanism used by the other zinc metalloproteinases, such as thermolysin. The conserved glutamate residue in the catalytic site acts in conjunction with a zinc ion to deprotonate and activate a water molecule. In turn, the activated water molecule acts as a nucleophile to attack the carbonyl of the peptide bond of a variety of substrates (Macdonald et al. 1995, Schlagenhauf et al. 1998). In *Leishmania*, these proteins are involved in different types of processes, such as the inhibition or perturbations of host cell interactions and the degradation of the extracellular matrix (Fitzpatrick et al. 2009). These proteins may have similar activities in schistosomes. Indeed, the *S. mansoni* protein, SmPepM8 (Smp_090100), is the second most abundant constituent in cercarial secretions, which provides insight on how it may contribute to tissue invasion by schistosomes and suggests this protein as a potential anti-parasitic target (Curwen et al. 2006, Fitzpatrick et al. 2009).

The catalytic triad of serine, histidine and aspartate residues is conserved in members of the serine protease family (Wilmouth et al. 2001, Hajjar et al. 2010). In elastases, this triad and an essential water molecule are involved in the catalysis. The peptide to be cleaved is bound noncovalently in the enzyme near the catalytic triad. In the first reaction step, the hydroxyl of the serine residue performs a nucleophilic attack on the substrate amide bond to form an ester. The amino terminus of the substrate is then covalently bound to the enzyme. The histidine residue abstracts a proton from a water molecule, which then attaches to the ester carbon to give rise to an oxoanion intermediate. Cercarial elastases play a key role in the penetration by the cercariae of mammalian skin to initiate infection and recent studies have revealed that these peptidases are also employed by the schistosomes to overcome or evade the host immune response (Salter et al. 2002, Aslam et al. 2008).

Cathepsin D is a member of the aspartic protease family. The active site of cathepsin D contains two aspartate residues, which perform an acid-base catalysis. This enzymatic mechanism involves the deprotonation of water by an ionised aspartate residue. This water molecule attacks the peptide carbonyl and there is a simultaneous protonation of the carbonyl oxygen by the other aspartate residue (e.g., Northrop 2001). Schistosome cathepsin D is involved in haemoglobin digestion, a process that provides the parasite with its main source of amino acid nutrients and that is essential for its development, growth and reproduction (Brindley et al. 2001, Caffrey et al. 2004, Delcroix et al. 2006). Given the essential function of cathepsin D in parasite nutrition and the ability of recombinant forms to cleave human immunoglobulin G, this protein is considered a potential target for novel anti-parasitic interventions (Verity et al. 2001, Morales et al. 2008).

The phylogenetic relationships of each endopeptidase family (Figs 2-4) are shown with protein sequences represented by identifiers in PhylomeDB (phylocedb.org) (Huerta-Cepas et al. 2011), UniProt (uniprot.org) (Apweiler et al. 2011) and/or SchistoDB (schistodb.net) (Zerlotini et al. 2009). In each phylogenetic tree, the *S. mansoni* endopeptidases form a well-supported clade of closely related proteins.

Together, the analysis of the *S. mansoni* proteome through an evolutionary approach identified endopeptidase family members that arose by gene duplication after
the divergence of this parasite from the other eukaryotic species studied in this work. These lineage-specific duplications are related to the parasite’s biology and evolution.

**Leishmanolysins (metallopeptidase M8 family)** - Our pipeline identified 12 *S. mansoni* leishmanolysins (Supplementary data). Proteins Smp_171330 and Smp_171340 are located in the same genomic region of Smp_090100 and Smp_090110, respectively, and could not be retrieved from the UniProt (Apweiler et al. 2011) and GeneDB (genedb.org) databases, which suggests that these genes were incorrectly annotated and probably deleted from these databases. Similar findings were obtained in two previous studies (Berriman et al. 2009, Bos et al. 2009).

To reconstruct the evolutionary history of *S. mansoni* leishmanolysins and their homologs in selected taxa, we performed a sequence alignment of 32 protein sequences identified as potential homologs by our pipeline. The trimmed alignment contained 1,822 sites, which cover most of the conserved protein domain identified in these proteins.

By analysing the phylogenetic tree (Fig. 2), it is possible to demonstrate that *S. mansoni* leishmanolysins have homologs in most species analysed in the present work, with the exception of *C. intestinalis* (tunicata) and fungi. However, this result does not completely discard the presence of homologous proteins in other organisms because they may be very divergent from the others in the database and therefore be missed by the pipeline search. The same is true for the other protein families mentioned in this paper.

Based on the information available in the literature and curated databases, three leishmanolysin homologs have been experimentally confirmed in *D. melanogaster*, *M. musculus* and *H. sapiens* and their function is related to the coordination of mitotic progression and cell migration (for details see Supplementary data). Although predicted functions or experimental evidence are not yet available, the metallopeptidase M8 family is also expanded in the sea anemone (*N. vectensis*) and sea urchin (*S. purpuratus*). The metallopeptidase M8 family also has more paralogs in the schistosomes (12 proteins).

Fig. 2: phylogenetic relationships of the *S. mansoni* leishmanolysins and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.
compared to *H. sapiens* (4 proteins), which is in contrast with what is normally observed for the human peptidase families (Berriman et al. 2009).

A conserved protein domain (Pfam: PF01457), which characterises members of the metallopeptidase M8 family, was identified in all *S. mansoni* proteins analysed here (Fig. 5). Length variation and conservation of active sites were also observed. According to the Pfam profile HMMs, truncated domains were identified in all proteins, which possibly reflects the presence of different protein isoforms, as has been described elsewhere (Floris et al. 2008). The truncated domains could also indicate that parts of the sequences are missing at the N-terminal, C-terminal regions, or both due to annotation issues.

The data also reveals that the protein domain is duplicated in Smp_167090, Smp_167120 and Smp_135530. Seven *S. mansoni* proteins (Smp_090100, Smp_090110, Smp_127030, Smp_135530, Smp_153930, Smp_167090 and Smp_173070) were identified as active due to the presence of expected active site residues and metal ligand sites in the correct positions based on alignments with reference sequences, as previously described (Berriman et al. 2009).

**Cercarial elastases (serine peptidase S1 family)** - Our analysis identified a total of 13 cercarial elastases encoded in the *S. mansoni* genome (Supplementary data). Similar results were obtained by Berriman et al. (2009). However, with TreeFam (Ruan et al. 2008), Berriman et al. (2009) also...
identified the Smp_192850 protein, which is annotated as a hypothetical protein and only contains 69 amino acids.

Two proteins, Smp_152560.2 and Smp_056680.2, are encoded in the same genomic location and could not be recovered in UniProt (Apweiler et al. 2011). Searches for the former protein in GeneDB (genedb.org) retrieved only the latter (Smp_056680), which indicated that the Smp_152560.2 gene was improperly annotated and thus was eliminated from both databases. In the original version of the S. mansoni genome, some sequences were interpreted as isoforms and different gene models were constructed. However, further studies indicated that these were actually mistakes in the genome assembly/annotation due to low sequence coverage. In the new version of the parasite genome, which is to be released by the Wellcome Trust Sanger Institute (sanger.ac.uk), many of these sequences have been collapsed.

Whole amino acid sequences from 35 proteins were aligned and filtered to remove gap-rich columns as previously described. The trimmed alignment contains 583 sites, which cover the conserved protein domain.

The phylogenetic analysis of the S. mansoni elastases and their homologs in the other species included in this work was performed as already described. The parasite elastases form a well-supported monophyletic clade, which suggests that these proteins originated from a common ancestor by gene duplication events followed by divergence in the Schistosoma lineage.

In observing the resulting phylogeny (Fig. 3), it is possible to demonstrate that S. mansoni elastases have homologs in six of the 16 other species considered in this analysis (N. vectensis, D. melanogaster, An. gambiae, B. floridae, M. musculus and H. sapiens). The serine peptidase S1 family is also expanded in all of these species except for one, D. melanogaster. According to the information available in UniProt (Apweiler et al. 2011), seven homologs have been experimentally confirmed in D. melanogaster, M. musculus and H. sapiens, and their function is related to a digestive process and immune response (Supplementary data). It is believed that similar activities are performed by elastases in schistosomes (Salter et al. 2002, Aslam et al. 2008).

Fig. 4: phylogenetic relationships of the Schistosoma mansoni elastases and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.
A conserved protein sequence domain (Pfam: PF00089), which is found in all characterised members of the serine peptidase S1 family, was identified in the *S. mansoni* elastases and ranges in length from 141-265 amino acids (Fig. 6). The catalytic triad of histidine, aspartate and serine residues is present in most of these proteins. Based on profile HMMs available in Pfam, truncated regions were assigned to all 12 of these elastases, perhaps reflecting their degree of divergence in relation to other proteins in the database. Meanwhile, it is important to emphasise that protein databases do not cover all of the existing diversity in nature.

Together, these results indicate that the correct number of cercarial elastases encoded in the *S. mansoni* genome is 12 and not 13 as described before. However, only Smp_006510, Smp_006520 and Smp_141450 were previously predicted as active proteins (Berriman et al. 2009). Smp_194800 has a much shorter domain compared to others. This difference could reflect either the presence of an elastase pseudogene in the parasite genome or that the sequence was incorrectly annotated due to an error in the gene model. Considering that the first-pass annotation of the *S. mansoni* genome was produced by a combination of gene-finding algorithms (Augustus,

**Fig. 5**: conserved protein domain architecture of the *Schistosoma mansoni* elastases. Protein identifiers were assigned in SchistoDB. The conserved protein domain according to Pfam (PF00089) is present in all proteins. Truncated regions (yellow block) are indicated. Sequence length, domain limits, and the catalytic triad of histidine (H), aspartate (D), and serine (S) are also shown.
Fig. 6: phylogenetic relationships of the *Schistosoma mansoni* cathepsins D and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.
Twinscan and GlimmerHMM) (Berriman et al. 2009), this genome has not received extensive manual curation and therefore, many gene models will be refined in the future. Furthermore, EVidenceModeler (Haas et al. 2008) has also been used to incorporate expressed sequence tag (EST) evidence into the data.

**Cathepsin D proteins (aspartic peptidase A1 family)** - Our pipeline identified 11 S. mansoni cathepsin D proteins (Supplementary data) that were duplicated after the divergence of S. mansoni from the other metazoa analysed here. The evolutionary history of cathepsin D proteins was reconstructed from the sequence alignment of 111 protein sequences from S. mansoni and the selected taxa. The final trimmed alignment contained 1,676 sites, which covered most of the conserved protein domain (Pfam: PF00026). Two S. mansoni proteins corresponded to alternative splicing products (Smp_136830.2 and Smp_013040.2). Similar results were found by Berriman et al. (2009).

The phylogenetic tree indicates that the S. mansoni cathepsin D proteins have homologs in all but one species (S. purpuratus) analysed in this work (Fig. 4). The aspartic peptidase A1 family has also been expanded in 12 of the 15 species in which homologous proteins were identified (A. thaliana, U. maydis, S. cerevisiae, N. crassa, C. elegans, C. briggsae, D. melanogaster, C. intestinalis, B. floridiae, D. rerio, M. musculus and H. sapiens). The number of paralogous proteins ranges from two-17 and includes different aspartic peptidases, such as pep- sins, renins, gastricsin and cathepsin D proteins. Based on the information available in the literature and curated databases, these homologous proteins are involved in digestion and protein degradation (Supplementary data). In schistosomes, cathepsin D proteins play an integral role in haemoglobin proteolysis (Brindley et al. 2001, Caffrey et al. 2004, Delcroix et al. 2006).

To predict the protein domain architecture of S. mansoni cathepsin D proteins, we applied the same methodology as previously described. The conserved domain (Pfam: PF00026), which has been found in all characterised aspartic peptidase A1 family members, was also identified in the S. mansoni proteins with sequence lengths ranging from 94-430 amino acids (Fig. 7). Active sites are also indicated. Based on the profile HMMs available in Pfam, truncated regions were observed in the N-terminal, C-terminal or both regions. The data also indicate that an additional short sequence domain (Pfam: PF07966), which is known as the A1 propeptide domain, is present at the N-terminal region of two S. mansoni proteins, Smp_013040.1 and Smp_013040.2. Smp_136840 has a much shorter domain compared to other proteins in the same family.

In a previous study, four S. mansoni cathepsin D proteins (Smp_013040.1, Smp_013040.2, Smp_136730 and Smp_136830.2) were identified as active proteins (Berriman et al. 2009), but the variation in the domain architecture and its implications in functional complexity were not investigated. One interesting study would be to analyse the functional properties of Smp_013040.1 and Smp_013040.2, which contain the A1 propeptide domain (PF07966).

**DISCUSSION**

We found that three endopeptidase families are expanded in the helminth parasite S. mansoni, which include members of the metallopeptidases (M8 family), serine peptidases (S1 family) and aspartic peptidases (A1 family). In this work, a comparative analysis of these three protein families in S. mansoni and 16 other eukaryotic proteomes revealed their distinct evolutionary histories and provided further information with respect to the sequence and functional features of the parasite family members.

Based on the S. mansoni genomic data, 335 peptidases were identified, which comprise 2.5% of the predicted proteome (Berriman et al. 2009). They include members of five major classes of peptidases (aspartic, cysteine, metallo, serine and threonine). Of the 61 peptidase families, 44 are expanded in this parasite and the number of paralogous proteins range from two-26.

Using a computational approach, Bos et al. (2009) analysed all putative peptidases encoded in the parasite’s genome in addition to using EST data, which is similar to work by Berriman et al. (2009). After removing redundant sequences, inactive homologs, likely pseudogenes and sequences smaller than 100 amino acids from the dataset, they identified a total of 255 peptidase sequences from the five catalytic classes.

Our results are not fully comparable to those obtained by Bos et al. (2009) with respect to elastases and cathepsin D proteins. However, it is worth noting that the phylogenetic analysis of the serine peptidase S1 family performed by these authors also indicated a well-supported clade of four S. mansoni elastases, which are corroborated by our findings. The other homologs with high similarities to the cercarial elastases were likely pseudogenes and, for this reason, they were excluded from the analysis by Bos et al. (2009).

Our results suggest that Schistosoma members of these endopeptidase families originated from successive gene duplication events in the parasite lineage after its diversification from the other metazoa analysed here. These results were corroborated by previous proteomic and phylogenetic analyses on Fasciola hepatica peptidases, which showed that the repertoire of virulence-associated cathepsin L proteins was established by a series of gene duplication events (Irving et al. 2003, Robinson et al. 2008). These studies also indicate that the gene duplications were followed by active site residue refinements, which interfere with the substrate specificity of the F. hepatica cathepsin L proteins. Whether the S. mansoni proteins share a similar refinement remains to be established.

Gene duplication followed by divergence is known to be the most predominant mechanism of molecular evolution and represents the main source of raw material for the generation of new genes and proteins through the processes of neo and sub-functionalisation (Ohno 1970, Conant & Wolfe 2008, Nahum & Pereira 2008, Hamilton et al. 2009). Although in some cases sequences have diverged to the extent that it is impossible to recognise homologous relationships, different proteins that arose by gene duplication may be related at distinct levels, such as sequence, structure, function or a combination of these features and can be grouped into families and superfamilies (Nahum & Pereira 2008).
Gene fusion, gene fission and domain shuffling were not observed as mechanisms shaping the evolution of the *S. mansoni* endopeptidase families analysed in this work. Whether gene fusion/fission also plays a role in the evolution of the *S. mansoni* genome will be a subject of a future work. Our previous study indicated that domain shuffling is one of the main evolutionary forces driving the sequence and functional diversification of the protein kinases of this parasite (Andrade et al. 2011, Avelar et al. 2011).

Peptidases have been implicated in various processes that are crucial to the development and survival of parasites, including host invasion, degradation of haemoglobin in blood feeding, immune evasion and activation of inflammation (McKerrow et al. 2006, Kasný et al. 2009).

Experimental work suggests that the SmPepM8 metallopeptidase (leishmanolysin) may contribute to tissue invasion by schistosome cercariae. This peptidase was the second most abundant protein released during the transformation of *S. mansoni* cercariae into schistosomes (Curwen et al. 2006). Leishmanolysins are a major surface peptidase member of the metallopeptidase M8 family, which in leishmaniasis are involved in different types of processes, such as the inhibition or perturbation of host cell interactions and the degradation of the extracellular matrix (Fitzpatrick et al. 2009). It is speculated that these proteins could perform similar activities in schistosomes during host-parasite interactions (Curwen et al. 2006, Fitzpatrick et al. 2009).

**Fig. 7:** conserved protein domain architecture of the *S. mansoni* cathepsin Ds. Protein identifiers were assigned in SchistoDB. The conserved protein domain according to Pfam (PF00026) is present in all proteins. The PF07966 additional N-terminal domain was identified in two proteins. Truncated regions (yellow block) are indicated. Sequence length, domain limits, and active sites are also shown.
Invasion of host skin is the initial event in establishing an infection in mammalian hosts. Considering the complexity of host skin barriers that the cercariae must go through during invasion, it has been suggested that multiple enzyme activities are required for this process (Salter et al. 2002). However, only one peptidase (cercarial elastase) has been identified as a major secretory product released during skin penetration (Knudsen et al. 2005, Hansell et al. 2008). These proteins may also be involved in eliminating the outer layer of the cercariae during transformation. Although cercarial elastases were named based on their ability to degrade insoluble elastin, numerous substrates for these enzymes have been identified, which include collagen, keratin and extracellular matrix proteins (Salter et al. 2002, Mckerror 2003, Knudsen et al. 2005).

Orthologous genes encoding elastase proteins were found in Schistosoma haematobium, Schistosoma japonicum and Schistosoma douhitti (Salter et al. 2002, Zhou et al. 2009). The expression of S. japonicum cercarial elastases was confirmed in both the sporocyst and cercarial stages and evidence that this peptidase is released by the parasite during the invasion of mammalian skin was obtained by anti-recombinant SjCE antibodies in infected mouse skin (Zhou et al. 2009). However, orthologous peptidases to S. mansoni cercarial elastases were not detected in the acutabular secretions of S. japonicum (Dvorák et al. 2008). Furthermore, the faster penetration by S. japonicum into the host skin may reflect the differential use of proteolytic enzymes in addition to those characterised in S. mansoni or even involve new peptidases not yet characterised (Chlichlia et al. 2005, He et al. 2005). Recent studies have also demonstrated that S. mansoni elastases are capable of cleaving IgE molecules from human, mouse and rat, indicating that the parasite may be able to overcome or evade the IgE response (Aslam et al. 2008). However, this subject remains controversial.

The biological complexity of S. mansoni is related to evolutionary innovations that took place before and after its diversification from other metazoans. Because duplicated genes are important substrates for improving an organism’s adaptation to its environment, understanding how members of protein families evolved may link evolutionary studies to parasite biology. In turn, this knowledge will provide insights into host-parasite relationships and accelerate the identification of novel vaccine and drug targets aimed at the treatment and eradication of schistosomiasis.

In conclusion, this paper provides an evolutionary view of three S. mansoni peptidase families, thus allowing for a deeper understanding of the genomic complexity and lineage-specific adaptations potentially related to the parasitic lifestyle. In the future, our results obtained using a systemic approach (proteome-wide analyses) may accelerate the understanding of schistosomiasis, its etiologic agents and host-parasite interactions and optimise the discovery of therapeutic targets for the development of new drugs and vaccines.

ACKNOWLEDGEMENTS

To the use of the computing resources of CRG (Spain) and CEBio-Fiocruz (Brazil), to Jaime Huerta-Cepas (CRG), Eric Aguiar and Francislon Silva (CEBio), for bioinformatics technical support, to Mariana de Oliveira (CEBio), for help with illustrations, and to the two anonymous reviewers, for the valuable suggestions that improved this paper.

REFERENCES


