Gene 559 (2015) 52–61

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

# Comparative genomic analysis of aspartic proteases in eight parasitic platyhelminths: Insights into functions and evolution

Shuai Wang [a,1], Wei Wei [a,1], Xuenong Luo [a], Sen Wang [b], Songnian Hu [b], Xuepeng Cai [a,*]

[a] State Key Laboratory of Veterinary Etiological Biology, Key Laboratory of Veterinary Parasitology of Gansu Province, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou, Gansu, China
[b] Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

ARTICLE INFO

*Article history:*
Received 30 October 2014
Received in revised form 8 January 2015
Accepted 9 January 2015
Available online 14 January 2015

*Keywords:*
Aspartic protease
Platyhelminth
Gene loss
Gene expansion
Drug target

ABSTRACT

We performed genome-wide identifications and comparative genomic analyses of the predicted aspartic proteases (APs) from eight parasitic flatworms, focusing on their evolution, potentials as drug targets and expression patterns. The results revealed that: i) More members of family A01 were identified from the schistosomes than from the cestodes; some evidence implied gene loss events along the class Cestoda, which may be related to the different ways to ingest host nutrition; ii) members in family A22 were evolutionarily highly conserved among all the parasites; iii) one retroviral-like AP in family A28 shared a highly similar predicted 3D structure with the HIV protease, implying its potential to be inhibited by HIV inhibitor-like molecules; and iiii) retrotransposon-associated APs were extensively expanded among these parasites. These results implied that the evolutionary histories of some APs in these parasites might relate to adaptations to their parasitism and some APs might have potential serving as intervention targets.



## 1. Introduction

Parasitic platyhelminths are parasites of great medical and agricultural importance. They are responsible for many severe parasitic diseases that are neglected in humans and livestock (Garcia et al., 2007). Especially, the schistosomiasis and cysticercosis infections are highly prevalent worldwide and can cause a spectrum of debilitating pathologies and death (Tsai et al., 2013). However, these parasitic diseases are difficult to treat and control, because of the lack of efficient drugs and vaccines. Efficient targets for drugs or vaccines can be probably found in the future by referencing on proteomic and genomic studies of these parasites. Proteases, proteolytic enzymes responsible for many fundamental cellular activities in almost all organisms, are important targets valuable to investigate in these parasites.

This study focuses on aspartic proteases (APs) which have been characterized in a wide range of organisms, including vertebrates, plants, fungi, protozoa and in some retroviruses (Santos et al., 2013). Generally, most of the APs are characterized by two aspartate (Asp) residues as key catalytic groups at their active site and by optimal activity at acidic pH. For this type of APs (two Asp residues), they are monomeric proteins formed from a single polypeptide chain with two similar domains, each of which provides an aspartic acid carboxyl group to form the active site. By contrast, retroviral or retroviral-like aspartic proteases are homodimers that consist of two identical subunits, each roughly equivalent to one domain of former monomeric APs that contains only one Asp residue (Koelsch et al., 1994). In both the two types, the catalytic center is formed by the two Asp residues activating a water molecule that mediates the nucleophilic attack on the peptide bonds of its substrates. At present, the MEROPS database groups all the existing APs into 16 protein families, belonging to 5 clans (AA, AC, AD, AE and AF) which represent one or more families that have arisen from a single evolutionary origin of proteases (Rawlings et al., 2012).

APs serve important functions in some important metabolic processes and in many diseases. In vertebrates, APs participate in various physiological and pathological processes, such as the rennin in hypertension and beta-secretase in Alzheimer's disease (Tu et al., 2006). In plants, APs have roles in senescence, stress responses and fertilization as well as in pathogen defense (Simoes and Faro, 2004). Reciprocally, APs are deployed by some pathogens to facilitate infections, such as HIV/AIDS and *Candida albicans* (Santos and Braga-Silva, 2013). Recent studies have shown that one AP of *Plasmodium falciparum* is responsible for processing effector proteins in a step that is crucial for their export into

*Abbreviations:* AP, aspartic protease; Asp, aspartate; HMM, Hidden Markov Model; RVP, retroviral protease; API, aspartic peptidase inhibitor; TSA, *Taenia saginata*; TAS, *Taenia asiatica*; HmN, *Hymenolepis microstoma*; EmuJ, *Echinococcus multilocularis*; Egr, *Echinococcus granulosus*; TsM, *Taenia solium*; Smp, *Schistosoma mansoni*; Sjc, *Schistosoma japonicum*.

* Corresponding author at: Lanzhou Veterinary Research Institute, CAAS, Xujiaping 1, Yanchangbu, Lanzhou 730046, Gansu, China.
*E-mail addresses:* wslvri@163.com (S. Wang), xyzdf23@163.com (W. Wei), luoxuenong@caas.cn (X. Luo), wangsen@big.ac.cn (S. Wang), husn@big.ac.cn (S. Hu), wslanzhou@126.com (X. Cai).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gene.2015.01.020
0378-1119/

the host red blood cell, thereby reprogramming the host cell to meet the demands of the parasite (Boddey et al., 2010). However, to date, APs remain poorly identified and characterized in platyhelminth parasites, particularly in tapeworms. For the blood flukes, several members of APs have been found in *Schistosoma mansoni* on a genome-wide scale (Bos et al., 2009). The cathepsin D in flukes is the only gene that has been systematically studied, which is a member from family A01, clan AA (Brindley et al., 2001). Several studies have demonstrated that cathepsin D is expressed in the gut of the flukes and plays an apical role in the digestion of host hemoglobin and other serum components in a host-specific manner, suggesting that it may be an essential enzyme for parasite nutrition in the mammal-parasitic stages of schistosomes (Brindley et al., 2001; Morales et al., 2008). Given the important roles in many organisms, APs are considered as potential therapeutic targets for many important diseases. Many small-molecule inhibitors of APs have been developed for use in the current chemotherapy against some diseases. The best known drugs (e.g. amprenavir, indinavir, and saquinavir) involve AP inhibitors against the HIV protease, applied to the therapy for AIDS (Wensing et al., 2010). Intriguingly, some of these drugs targeted for HIV exert inhibitory effects on the APs of several other pathogens and may be applied as the drugs against the corresponding diseases (Santos et al., 2013). For instance, the APs in *Leishmania* spp. and *C. albicans* are experimentally evidenced to be inhibited by HIV PIs and are potential targets against these pathogens (Santos et al., 2013; Santos and Braga-Silva, 2013).

Compared with the wide investigations of other proteases in parasitic flatworms, a limited number of studies have been performed to identify and characterize the aspartic proteases in these species, particularly in tapeworms, although potential as the targets for novel anti-parasite interventions have been suggested (Morales et al., 2008). And little is known about the evolution of APs among these parasites. Therefore, it was of considerable interest to establish the complement of APs encoded within parasitic flatworm genomes by comparative genomic analyses, which can provide valuable insights into the functions and evolution of these protein families.

Herein, we performed systematical identifications and a comprehensive comparative study of the AP families in eight parasitic flatworm species with great medical importance, whose genomes were recently completed (and unpublished), including *Schistosoma japonicum*, *S. mansoni*, *Hymenolepis microstoma*, *Echinococcus multilocularis*, *Echinococcus granulosus*, *Taenia solium*, *Taenia saginata* and *Taenia asiatica* (genomes of the latter two are unpublished). We identified a repertoire of APs that constitutes one of the most complete genome-wide sets of APs in these parasitic flatworms to date. In particular, we i) performed analyses of evolutionary histories of APs in these parasites, ii) assessed potentials of APs in the family A22 as targets to be inhibited by HIV inhibitor-like molecules, iii) explored the expression patterns of different life stages and body parts (for *S. mansoni* and *H. microstoma*). Some of these APs may have important functions of catalytic processes in their parasitism and show potentials serving as targets of effective chemotherapeutic or immunological treatments.

## 2. Materials and methods

### 2.1. Sequence searches of parasitic platyhelminth genomes

Genomic sequences and predicted proteomes of the two flukes and six tapeworms (http://www.genedb.org/Homepage for *H. microstoma*, *E. granulosus*, *E. multilocularis*, *T. solium*, *S. mansoni* and *S. japonicum*; the data on *T. saginata* and *T. asiatica* was obtained from unpublished internal data) were retrieved. The complete set of core protease sequences (library for protease units of holotypes) from the MEROPS database (http://merops.sanger.ac.uk/) was searched by the MEROPS batch BLAST server, using the predicted proteomes as queries. These core sequences comprise a non-redundant library of the catalytic unit of a protease and exclude all other functional units, such as ATP-binding, or ubiquitin-associated domains. Therefore, such BLAST searches avoid the false positive identification of proteins as proteases due to the high homology in non-catalytic parts of the sequence. Only the sequences matching the corresponding holotypes with e-value < $1e^{-04}$ were retained as putative protease candidates. The HMMER package (Mutter et al., 2008) was also used to search the predicted proteomes against the corresponding HMMs of proteases and Pfam-A model from the Pfam database (http://pfam.janelia.org/). The query sequences that matched the HMMs of proteases with e-value < $10^{-4}$ were retained. If a sequence matched more than one model, we considered the model with the highest score and lowest e-value as the best. Then, the two sets of results (results from BLAST search and HMMER searches) for all the genomes were merged as one set of query sequences to perform a second search against each proteome by the BLAST (e-value < $10^{-4}$). Active aspartic residues in the sequences were detected by comparison with the protease units (domains) of holotypes in the MEROPS database using BLAST and then manual check. Furthermore, all protease candidates were further conformed for the protease domains, signal peptides and transmembrane regions by the InterProScan program (Jones et al., 2014). All the identified AP candidates were used as BLAST queries against the NR database to rule out the possibility of genome contamination for these parasites.

Several steps were further performed to confirm the evolutionary history of orthologous genes in a specific family, if the number of AP orthologs in this family or the existence of some members were dubious. First, all family members identified from these parasites and other evolutionarily closely related species were used as queries to search their genome assemblies with genBlastG (She et al., 2011) (e-value < 10, with other parameters default). Second, transcripts were de novo constructed by the program Trinity (Haas et al., 2013) without reference genomes, using available RNA-seq data (http://www.ebi.ac.uk/arrayexpress/; accession number: E-ERAD-56 for *H. microstoma*; E-MTAB-451 for *S. mansoni*; internal data for *T. solium* and *T. saginata*), which could provide transcript evidence and further confirmation free from the genome assemblies. The constructed transcriptomes were used as a database to be searched by tBLASTn using the queries described above (e-value < 10).

### 2.2. Phylogenetic analysis

All the alignments were performed by MAFFT (Katoh and Standley, 2013) (version 7) with default parameters (-auto) and then checked by manual edit. For the small set of conserved sequences, phylogenetic trees were constructed by the neighbor-joining method in MEGA (version 5) (Tamura et al., 2011) with default parameters, using a bootstrap test with 1000 replications. For the large set of the sequences that were highly divergent, several refining steps were performed to obtain a reliable tree. First, MaxAlign (Gouveia-Oliveira et al., 2007) was used to refine the datasets by removing the sequences with extensive gaps in the alignments; second, trimAl (version 1.30) (Capella-Gutierrez et al., 2009) was used to assist in the alignment trimming by removing the gapped regions (sites with 30% or more gaps) in the multi-alignments; third, PROTTEST (Darriba et al., 2011) was employed to select the best model of protein evolution for each alignment. The phylogenetic relationships were reconstructed by the Maximum Likelihood method implemented in PhyML (Guindon et al., 2009) with the selected evolution models (1000 bootstrap replicates) and by Maximum Parsimony method in MEGA (version 5) with CNI search method (1000 bootstrap replicates). The combined programs InParanoid and MultiParanoid (Alexeyenko et al., 2006) were used to provide supplementary evidence to confirm the orthologous relationships of some genes, if ambiguous evolutionary relationships were involved (e-value < $10^{-4}$; using the proteome of *Caenorhabditis elegans* as control).