Insect Biochemistry and Molecular Biology 88 (2017) 48–62

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

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

Serine protease-related proteins in the malaria mosquito, *Anopheles gambiae*

Xiaolong Cao, Mansi Gulati, Haobo Jiang*

Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA

ARTICLE INFO

Article history:
Received 3 May 2017
Received in revised form
25 July 2017
Accepted 29 July 2017
Available online 2 August 2017

Keywords:
Phylogenetic analysis
Gene duplication
Chromosomal location
Insect immunity
Expression profiling
Hemolymph protein
Clip domain
Serine protease cascade

ABSTRACT

Insect serine proteases (SPs) and serine protease homologs (SPHs) participate in digestion, defense, development, and other physiological processes. In mosquitoes, some clip-domain SPs and SPHs (*i.e.* CLIPs) have been investigated for possible roles in antiparasitic responses. In a recent test aimed at improving quality of gene models in the *Anopheles gambiae* genome using RNA-seq data, we observed various discrepancies between gene models in AgamP4.5 and corresponding sequences selected from those modeled by Cufflinks, Trinity and Bridger. Here we report a comparative analysis of the 337 SP-related proteins in *A. gambiae* by examining their domain structures, sequence diversity, chromosomal locations, and expression patterns. One hundred and ten CLIPs contain 1 to 5 clip domains in addition to their protease domains (PDs) or non-catalytic, protease-like domains (PLDs). They are divided into five subgroups: CLIPAs (22) are clip$_{1-5}$-PLD; CLIPBs (29), CLIPCs (12) and CLIPDs (14) are mainly clip-PD; most CLIPEs (33) have a domain structure of PD/PLD-PLD-clip-PLD$_{0-1}$. While expression of the CLIP genes in group-1 is generally low and detected in various tissue- and stage-specific RNA-seq libraries, some putative GPs/GPHs (*i.e.* single domain gut SPs/SPHs) in group-2 are highly expressed in midgut, whole larva or whole adult libraries. In comparison, 46 SPs, 26 SPHs, and 37 multi-domain SPs/SPHs (*i.e.* PD/PLD-PLD$_{\geq 1}$) in group-3 do not seem to be specifically expressed in digestive tract. There are 16 SPs and 2 SPH containing other types of putative regulatory domains (*e.g.* LDLa, CUB, Gd). Of the 337 SP and SPH genes, 159 were sorted into 46 groups (2–8 members/group) based on similar phylogenetic tree position, chromosomal location, and expression profile. This information and analysis, including improved gene models and protein sequences, constitute a solid foundation for functional analysis of the SP-related proteins in *A. gambiae*.

1. Introduction

Chymotrypsin-related serine proteases (SPs) form a large family of enzymes that hydrolyze peptide bonds at different rates and with various degrees of specificity (Rawlings and Barrett, 1993). For instance, trypsin cleaves at a high rate, specifically after most Lys and Arg residues, consistent with its role in protein digestion; pancreatic elastase cuts efficiently after any accessible small nonpolar residues (*e.g.* Ala) in many proteins, whereas human coagulation factor Xa cleaves only few protein substrates in plasma, after certain Arg residues and at a low rate (k_{cat}). The S1 pocket of a protease interacts with the P1 residue of a protein substrate, governing its primary specificity (Schechter and Berger, 1967). Regulatory domains or regions in some nondigestive SPs provide additional specificity by localizing enzyme catalysis through specific interactions with activators, substrates, cofactors, and inhibitors (Kanost and Jiang, 2015; Krem and Di Cera, 2002). His, Asp and Ser residues in the active site of SPs are responsible for the acyl transfer mechanism of catalysis, with well-formed substrate binding clefts defining their specificities. SPs often contain a signal peptide guiding them to extracellular or granular locations, where they persist as inactive zymogens and then become activated by proteolytic cleavage at a particular peptide bond. In extracellular spaces, several SPs can constitute a cascade pathway in which one

Abbreviations: SP, serine protease; SPH, (non-catalytic) serine protease homolog; PD, SP catalytic domain; PLD, protease-like domain in SPH; LDLa, low-density lipoprotein receptor class A repeat; SR, scavenger receptor; TSP, thrombospondin; CUB, C1r/C1s, Uegf, Bmp1; MSP, modular serine protease; CLIP, clip-domain SP or SPH; GP and GPH, gut serine protease and gut serine protease homolog; PO and proPO, phenoloxidase and its precursor; PAP, proPO activating protease.

* Corresponding author. 127 Noble Research Center, Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA.
E-mail address: haobo.jiang@okstate.edu (H. Jiang).

http://dx.doi.org/10.1016/j.ibmb.2017.07.008
0965-1748/

SP activates the zymogen of another in each step to trigger a rapid local response, such as blood coagulation or the complement system in mammals. In addition to active proteases, related serine protease homolog (SPH) genes encode SP-like sequences lacking one or more of the catalytic triad residues and, thus, proteolytic activity. Some cleaved SPHs are active as modulators of interacting SPs (Jiang et al., 2010; Park et al., 2010). While molecular mechanisms for such modulation are unclear, the SP-like fold and associated structural unit (*e.g.* clip domain) of SPHs are likely essential for the interactions that determine their biochemical functions.

SP-related proteins mediate insect immune responses (*e.g.* melanotic encapsulation, cytokine activation, and antimicrobial peptide induction) (Jiang et al., 2010). Like human clotting factors, insect SPs and SPHs form complex networks to stop bleeding and fight infection. In each insect species with a known genome, SP-related proteins form a large family with 60–400 members (Cao et al., 2015; Christophides et al., 2002; Ross et al., 2003; Waterhouse et al., 2007; Zhao et al., 2010; Zou et al., 2006, 2007). Their roles in defense and development have been explored in *Drosophila melanogaster*, *Manduca sexta*, *Tenebrio molitor*, and other insects (Kanost and Jiang, 2015; Park et al., 2010; Veillard et al., 2016). In mosquitoes, clip-domain SPs/SPHs have been named CLIPs (Waterhouse et al., 2007). As summarized by Cao et al. (2015), numbers of the clip-domain SP/SPH genes identified in genomes are 63 in *Aedes aegypti*, 55 in *A. gambiae*, 45 in *D. melanogaster*, 42 in *M. sexta*, and 49 in *Tribolium castaneum.*

Accurate gene models form a solid base for protein identification and elucidation of biochemical functions. Continuous efforts have been made to improve quality of the predicted genes after the initial genomes of *D. melanogaster*, *A. gambiae*, *Apis mellifera*, and other insects were published. The *M. sexta* genome project greatly benefited from next-generation sequencing, which provided RNA-seq data for the genome assembly, gene modeling and expression profiling (Kanost et al., 2016). We developed a method to select the best of the models from **M**AKER, **C**ufflink, **O**ases and **T**rinity programs (*i.e.* MCOT model) (Cao and Jiang, 2015). As this method has been automated and successfully applied in other insect genome projects (Cao and Jiang, 2017), we thought it would be interesting to test whether our method can further improve the latest release of *A. gambiae* gene models using the available RNA-seq data, with a focus on SP-like genes. Numerous discrepancies were identified between AgamP4.5 and corresponding AgMCOT models. To substantiate the observations and promote research on SP-related proteins in this species, we examined and improved the models in the official protein set (OPS), studied their domain organization and sequence diversity, classified them into the groups of CLIPs, GP(H)s and SP(H)s, and established an information system that contains systematic names, putative activation sites, predicted enzyme specificity, genomic locations, expression patterns, and phylogenetic relationships. Through further studies, we hope to establish a platform for comparing SP-related sequences from various insects and suggest functions for orthologs based on genetic and biochemical analyses in a few model species.

2. Materials and methods

2.1. Identification of A. gambiae SP-related proteins

OPS AgamP4.4 was downloaded from VectorBase (https://www.vectorbase.org/). Protein-coding genes were modeled using the MCOT pipeline (Cao and Jiang, 2015) by selecting the best for each gene from the OPS, TopHat-Cufflinks (Kim et al., 2013; Trapnell et al., 2012), Trinity (Haas et al., 2013) and Bridger (Chang et al., 2015) assemblies to constitute an AgMCOT protein set (unpublished data). Domains in the AgamP4.4 and AgMCOT sequences were identified by InterProScan5 v5.17 (Jones et al., 2014) in a local supercomputer. Proteins containing a chymotrypsin-like (*i.e.* S1 family), SP-related domain were extracted and pooled. After removal of redundant, alternatively spliced, and severely incomplete genes, the sequences were manually examined and improved according to characteristic features of the S1 SPs, such as signal peptide and conserved regions.

2.2. Properties of A. gambiae SP and SPH sequences

Sequences were separated into SPs or SPHs by examining the presence of a His-Asp-Ser catalytic triad as described before (Cao et al., 2015). Signal peptides were predicted using SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP/) (Petersen et al., 2011) and Signal-3L (http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/) (Shen and Chou, 2007). Some clip domains were identified by InterProScan5 and others by manual inspection of the sequences for a Cys doublet in the region close to the protease or protease-like domain (PD or PLD). SPs and SPHs with four additional Cys residues at particular locations (Cao et al., 2015; Jiang and Kanost, 2000) upstream of the doublet were designated CLIPs to indicate the presence of a clip domain (Kanost and Jiang, 2015). Residues 190, 216 and 226 (chymotrypsin numbering) (Perona and Craik, 1995) that form the primary substrate-binding pocket of PD were identified in the aligned sequences for predicting their substrate specificity (Cao et al., 2015).

2.3. Multiple sequence alignment and phylogenetic analysis

Multiple sequence alignments of the entire sequences in the CLIP, GP(H), and SP(H) groups were performed using MUSCLE (Edgar, 2004), one module of MEGA 7.0 (Kumar et al., 2016), under the default setting with maximum iterations changed to 1000. The classification and naming were based on: 1) clip domain presence or absence, 2) position in a phylogenetic tree of non-CLIP SPs or SPHs, and 3) expression patterns. Neighbor-joining trees were constructed in a preliminary analysis of the SP-related sequences, and reliability of the trees was tested using a bootstrap method with 1000 trials. Alignments of the three individual groups were converted to NEXUS format by MEGA, and phylogenetic analyses were conducted using MrBayes v3.2.6 (Ronquist et al., 2012) under the default model with the setting "nchains = 12". MCMC (Markov chain Monte Carlo) analyses were terminated after the standard deviations of two independent analyses were <0.01 for GP(H)s and CLIPs, and <0.02 for SP(H)s. FigTree 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) was used to display the phylogenetic trees.

2.4. Chromosomal locations of the SP and SPH genes

For most of the SP-related genes, their genomic locations were available in the information lists of the AgamP4.4 or Cufflinks models. Retrieved position data were plotted using ArkMAP 2.0 (http://www.bioinformatics.roslin.ed.ac.uk/arkmap/) and improved using Adobe Illustrator.

2.5. Expression profiling of the SP-related genes

The 113 RNA-seq data sets of *A. gambiae* from previous research (Bonizzoni et al., 2012; Mead et al., 2012; Pinheiro-Silva et al., 2015; Rinker et al., 2013; Vannini et al., 2014) were downloaded from NCBI Sequence Read Archive (SRA) and converted to fastq format using SRA Toolkit. Reads were first trimmed with Trimmomatic (Bolger et al., 2014) to remove adaptors and low quality bases with the setting "SLIDINGWINDOW:4:30 LEADING:20 TRAILING:20 MINLEN:50". Transcript sequences of the SP-like proteins in