12-15-2014

Understanding the Evolutionary History of Biochemical Innovation

Madeline Opal St. Julien
University of South Carolina - Columbia

UNDERSTANDING THE EVOLUTIONARY HISTORY OF BIOCHEMICAL INNOVATION

by

Madeline Opal St. Julien

Bachelor of Science
University of South Carolina, 2011

Submitted in Partial Fulfillment of the Requirements

For the Degree of Master of Science in

Biological Sciences

College of Arts and Sciences

University of South Carolina

2014

Accepted by:

Jeffry L. Dudycha, Director of Thesis

Bert Ely, Reader

Bob Friedman, Reader

Lacy Ford, Vice Provost and Dean of Graduate Studies

DEDICATION

I dedicate this thesis to my mother, who has passed down her innate infatuation of nature to me.

ACKNOWLEDGEMENTS

I am incredibly grateful for the patience, advisement, and funding of my professor, Jeffry L. Dudycha. The professional support and advice from Bob Friedman, Jijun Tang, and Bert Ely has aided the progression of my research. Chris Brandon, Matt Greenwold, and Claire Hann are reputable colleagues who have contributed an immense amount of intellectual support and comradery.

ABSTRACT

The serine protease (SP) gene family is an ecologically important gene family because of observed involvement in innate immunity, digestive processes, and embryological development of arthropods. In the past decade, all genes of the serine protease family have been classified in a number of arthropods, with the exception of crustacean. Possible evolutionary mechanisms have been observed based off of varying selectional pressures acting on recent SP expansions in respect to varying diets. *Daphnia* is the first crustacean to have its genome sequenced, and their genomes were analyzed in this study to elucidate the expansion and divergence of the SP gene family across arthropods in respect to similar diet. In this study, all SP-like genes were extracted from the *D. pulex* and *D. magna* genomes. Multiple bioinformatic approaches were used to catalogue the structural and biochemical properties of functional serine proteases in both *Daphnia* genomes. Phylogenetic analysis and selection tests, within and between both species of *Daphnia*, showed purifying selection reinforced the role of basal digestive proteases within *Daphnia* before and after divergence in respect to similar diet preferences.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Ψ	Genes hypothesized to be pseudo genes based off of lack of gene expression from Expression Sequence Tags.
*	Presence of a signal peptide

List of Abbreviations

C ..... Clip-Domain

CBD2 ..... Chitin-binding domain type 2

CBD4 ..... Chitin-binding domain type 4

CHY ..... Chymotrypsin

CLECT ..... carbohydrate-recognition domain

ELA ..... Elastase

H-SP ..... Serine Protease Homolog

KR ..... Kringle domain

LC ..... Region of Low compositional complexity

LDLa ..... Cysteine-rich Low-density lipoprotein receptor domain class A

PAN_AP ..... APPLE domains

SEA ..... Domain found in sea urchin

SERP ..... Serine Protease with unknown specificity

SP ..... Serine Protease

SR ..... Scavenger receptor Cys-rich

SRCR ..... Egg peptide speract receptor

TM ..... Transmembrane

TRY ..... Trypsin

TSP1 ..... Thrombospondin type 1 repeats

CHAPTER 1

EVOLUTIONARY DIVERSIFICATION OF SERINE PROTEASES IN THE CRUSTACEAN DAPHNIA PULEX

1.1 INTRODUCTION

Serine proteases (SP) are enzymes that hydrolyze peptide bonds to break down proteins. Few individual serine protease genes are widely conserved across taxa, however the gene family is found in all taxa, and thus its diversification may be important in adaptation. Members of the family are known to have multiple functional roles, including digestion, embryonic development, and innate immunity (Rawlings and Barrett 1993). Over the past 20 years, research in digestive physiology have shown members of the gene family in the mid gut of arthropods to develop resistance against serine protease inhibitors from their resources (Casaretto and Corcuera 1995). The family's role in innate immunity and embryonic development has been extensively studied in *Drosophila*. Serine proteases occur in pathways such as the antimicrobial peptide producing Toll pathway (Jang et al 2008) and in the pathway for dorso-ventral polarization during embryonic development (Hong and Hashimoto 1996; Lemosy et al 1998).

All serine proteases have an eponymous serine residue (Ser-195) that is critical for catalyzing the hydrolysis reaction. Serine proteases are endopeptidases and on the basis of substrate specificity have been classified into three subfamilies: trypsins, chymotrypsins, and elastases. The SP domain structure starts with a cleavage site that

may or may not be downstream of a signal peptide (Ross et al 2003). This cleavage site, with the conserved motif R^IVGG, is crucial for turning the inactive zymogen into its catalytically-active primary structure (Hedstrom et al 1996). Once active, this enzymatic structure has three amino acids which comprise a catalytic triad that hydrolyze peptide bonds of a peptide chain targeted for degradation.

The amino acids of the catalytic triad and their respective motifs, TAAHC, DIAL, and GDSGGP, are highly conserved across many genes and species(Greer 1990). The histidine (His-57) in the TAAHC motif is the residue that attracts a proton from the serine hydroxyl side chain to allow for nucleophilic attack on the protein substrate in the catalytic cleft (Kraut 1977). The aspartate (Asp-102) in the DIAL motif is critical for stabilizing the protonated histidine in the TAAHC motif. The serine residue (Ser-195) in the GDSGGP motif then hydrolyzes the scissile peptide bond of the substrate by an acylation-deacylation mechanism(Kraut 1977). Additional residues surround the GDSGGP motif to define the substrate specificity of the enzyme: Asp-189, Gly-216, and Gly-226 define the trypsin subfamily; Gly-189, Gly-216, and Gly-226 define chymotrypsins; Ser-189, Val-216, and Ala-226 define elastases (Perona and Craik 1995). Standard residue numbering for serine proteases is based on early studies of the structural properties of Bovine chymotrypsin-A(Hartley 1964) and we adopt that numbering throughout this study. In addition, conserved cysteine residues that form three or four disulfide bridges are found in the SP domain and play a role in the structural integrity of the enzyme (Greer 1990).

At the organismal level, evolution of the serine protease family has been associated with adaptation to specific resource diet. For example, in the analysis of the SP family in

Anopheles gambiae, a relationship between adaptation to blood meal and recent expansions of the gene family was observed (Wu et al 2009). The gene family was also analyzed in twelve species of *Drosophila* in the context of food preference (Li et al 2012). Both dipteran studies revealed positive selection within the SP gene family, suggesting a relationship between gene expansion and adaptation to meal preference. This suggests that expansion of the gene family may permit adaptation to use novel resources. Under this hypothesis, negative selection may maintain the function of ancestral proteases, while novel proteases experience positive selection.

In ecology, serine proteases have been important in understanding the mechanisms of consumer-resource interactions. In particular, serine proteases are known to mediate the consumption of algae by the zooplankter *Daphnia*, the dominant herbivore in lakes around the world (Sarnelle 2005). Observations in *Daphnia magna*, a fresh water crustacean, showed the SP family to make up 75-83% of the catalytic activity in the gut (Elert et al 2003).

Experiments focusing on resource exploitation have shown Daphnia-phytoplankton interactions affect life history traits and cause differential gene expression across the *Daphnia* genomes (Gliwicz and Boavida 1993; Tessier et al 2000; Dudycha et al 2012). The *Daphnia pulex* genome shows a high gene count, hypothesized to be the result of an elevated rate of tandem gene duplications (Colbourne et al 2011). This preliminary study will test the relationship between elevated rate of gene duplications in the S1 protease family, one of ecological importance, and resource exploitation in *Daphnia*.

In this study, we describe the evolution of the SP gene family found in *Daphnia pulex*, and evaluate the potential for functional evolution with respect to resource

exploitation. All SPs were identified, manually curated, and catalogued on the basis of structural features of the SP domain and any accessory functional domains. In particular, we sought to determine whether functional groupings, such as substrate specificity, were monophyletic or evolutionarily labile. Furthermore, we sought to quantify selection within the gene family to test whether gene duplication was associated with adaptive processes that may influence the evolution of resource exploitation.

1.2 METHODS AND MATERIALS

Database searching, sequence retrieval and annotation of active SPs and SP homologs

We began our search with TRY4B, a serine protease previously identified in the *Daphnia* genome that is known to be expressed (Schwerin et al 2009). TRY4B (Schwerin et al 2009) contains all structural features necessary for serine protease function. We initially used its protein sequence to BLAST the *Daphnia pulex* genome using at NCBI. However, to ensure all candidate serine proteases were identified, we constructed a search parameter for PHI-Blast (Zhang et al 1998), an algorithm that allows you to identify genes that share conserved motifs. To determine the search parameter characteristics, we used Prosite (http://prosite.expasy.org/), a database of conserved motifs, amino-acid sequence patterns, functional protein domains, families, and functional sites (Sigrist et al 2002). The serine protease search parameter we constructed is shown below; the amino acid residues critical for an active catalytic triad are highlighted:

[LIVM]-[ST]-A-[STAG]-H-C-X(10,500)-[NSHY]-D-[IVL]-X(10,500)-[DNSTAGC]-[GSTAPIMVQH]-X(2)-G-[DE]-S-G-[GS]-[SAPHV]-[LIVMFYWH]-[LIVMFYSTANQH]

We then used this parameter in PHI-Blast, with BLOSUM62 scoring matrix, and retrieved an output of 99 putative serine protease genes.

Genes retrieved from NCBI PHI-Blast search were easily discriminated between genes with high similarity to the search parameter, and those with low similarity. Therefore, sequences with an E-value > 0.0005 were discarded from this study. ScanProsite (http://prosite.expasy.org/scanprosite/) (de Castro et al 2006) surveyed each gene from the output with E-value < 0.0005 to determine whether all conserved structural components of an active-SP were present, including the catalytic triad, the cysteine-cysteine disulphide bridges, and the cleavage site. If at least one of these structural elements were missing, the gene was catalogued as a homolog (H-SP). All SPs containing all three structural elements were then used to re-query the *Daphnia pulex* genome at NCBI with BLASTP (States and Gish 1994). This procedure was repeated until no more novel SP or H-SPs were found in the output from the *D. pulex* genome.

All SPs and H-SPs were manually curated in the JGI database (Colbourne et al 2011) of the *Daphnia pulex* genome. Sequences were examined individually to confirm predicted start/stop codons and intron-exon boundaries. Where necessary predicted gene models were corrected. Genes were located on WFleabase's (http://wfleabase.org) GBrowse Maps, and tracks for environment-specific expression and expressed sequence tags were examined to determine expression. SPs and H-SPs with the absence or lack of gene expression were cataloged as pseudogenes in (Table 1.3 & 1.4) (Gilbert and Singan, V.R., Colbourne 2005). Following phylogenetic reconstruction (see below) genes were named in the JGI database according to *Daphnia* gene nomenclature standards. Existing

gene names were retained, except in a few instances where structural features indicated that existing names were functionally misleading.

Identifying Functional Motifs

As mentioned before, genes containing all three amino-acid residues of the catalytic triad were catalogued as SPs. If at least one amino-acid residue was missing from the triad, the gene was catalogued as an H-SP. ScanProsite also identified the three putative motifs that contain the three critical amino-acid residues of the catalytic triad (ie., TAAHC, DIAL, and GDSGGP)(de Castro et al 2006). The amino acid sequences of the three putative motifs are essential in the formation of the catalytic cleft in serine proteases. However, it is not known how much variation around the critical residues is tolerated. To quantify the frequency of variants in the catalytic triad and determine if any amino acid substitutions retained biochemical properties of the canonical residue, amino acid sequences of the SP domain in each SP were analyzed with the Multiple EM for Motif Elicitation interface (MEME; version 4.9.0 http:// http://meme.nbcr.net) (Bailey et al 2006).

Three additional residues at amino acid location 189, 216, and 226 determine the substrate specific-binding pocket of a serine protease. To determine the position and presence of the residues involved in substrate specificity (Schwerin et al 2009), we aligned all SPs and H-SPs using Muscle (Edgar 2004) multiple alignment We then cataloged SP and H-SP subfamily based on the following substrate specific residues: Asp-189, Gly-216, and Gly-226 in trypsin-like SPs; Gly-189, Gly-216, and Gly-226 in chymotrypsin-like SPs; Ser-189, Val-216, and Ala-226 in elastase-like SPs (Perona and Craik 1995). If the residues at the substrate specificity locations did not identify known

specificity, the gene was catalogued as Serine Protease-like (SERP) or Serine Protease-like homolog (SERP-H). Alignments were manually examined to confirm that SERP classification did not occur due to misalignment.

We scanned the amino acid sequence of each SP and H-SP using SMART (Onting 1998) to identify signal peptides and additional functional domains. The subcellular localization of each SP and H-SP was predicted using pTARGET, a prediction server for protein subcellular localization (http://golgi.unmc.edu/ptarget/) (Guda 2006).

A BLASTP off all SPs and H-SPs in the *Daphnia pulex* genome against the *Drosophila melanogaster* genome at NCBI database retrieved predicted orthologs for genes conserved in arthropods. This output contained additional functional information about the sequence properties in a select number of SPs and H-SPs.

Sequence alignments and Phylogenetic analysis

Attempts to align all the genes in the dataset based on their full length failed due to high levels of variation outside of the SP domain. Therefore, we focused on constructing an alignment of the SP domain itself for phylogenetic analysis. Domains were aligned via Muscle with a -2.9 open gap penalty using MEGA5.10 (Tamura et al 2011). The alignment output was manually examined to ensure all amino acid residues comprising the critical structural elements of serine proteases were aligned.

To build a phylogenetic reconstruction of serine protease evolution, we used RAxML. In RAxML, we applied the maximum likelihood method using a General Time Reversal nucleotide substitution model with four discrete GAMMA rate categories and the estimated proportion of invariable sites (Stamatakis 2006). An additional test ran

1000 bootstrap replicates to assign confidence to nodes. Differences between the synonymous and nonsynonymous distances per site for each SP domain were calculated to test for non-neutral selection in strongly supported clades of gene duplicates. This analysis was done in MEGA5.10 with the Nei-Gojobori model (Tamura et al 2011).

1.3 RESULTS

Overview and classification of SPs and H-SPs

A total of 211 serine protease (SP) genes and homologs (H-SP) were identified, classified, and cataloged based on of the presence of functional elements characteristic of serine proteases (Greer 1990; Perona and Craik 1995). The 106 genes containing all characteristic elements were classified as SP. An additional 105 genes were missing one or more functional elements and were classified as homologs. Though these homologs are missing elements thought to be required for proteolytic function, they share enough elements to show they are evolutionarily members of the serine protease gene family.

All SPs and H-SPs were classified subfamilies as Trypsin-like (TRY), Chymotrypsin-like (CHY), or Elastase-like (ELA) based on substrate-specificity residues. These subfamilies were determined on the basis of amino acid residues at position 189, 216, and 226 of the SP domain (Perona and Craik 1995). As a result, we identified 73 Trypsin-like, 14 Chymotrypsin-like, and only one elastase-like SPs (Table 1.1). The 18 remaining SP genes with alternate substrate specificity residues were classified as serine proteases for which the substrate is unknown (SERPs). The H-SPs showed a markedly different distribution, with 92 being SERPs, and only 10 that were trypsin-like, 2 that were chymotrypsin-like and one that was elastase-like. Unknown substrate specificity of

the H-SERPs and SERPs may represent novel substrate specificities in *Daphnia pulex* (Table 1.3, Table 1.4).

Evaluation of expression maps and expression sequence tags (ESTs) on GBrowse Maps at wFleabase (Gilbert and Singan, V.R., Colbourne 2005), showed that all but three SPs were expressed. Most H-SPs also had evidence of expression. The eleven H-SPs for which there was no evidence of expression were labeled as putative pseudogenes (Vanin 1985) ($^{\Psi}$) in Table 1.3 and 1.4.

Motif conservation within the catalytic cleft of active-SPs

In the MEME analysis of motif conservation in the SP domains, 84.3% of the SPs contain the conserved TAAHC motif; the remaining SPs contained a variety of substitutions around the required histidine. DASHC, HAAHG, SAGHC, TACHC, and TASHC were variants that occurred once, whereas NAAHC(3), DAAHC(4), and TAGHC(4) occurred multiple times within SPs. The underlined residues in TAAHC are highly conserved and hydrophobic (Table 1.4). Additional residues around the TAAHC motif, ILTAAHCV undergo substitution, but the hydrophobic properties are still highly conserved to insure conservation of the structure of the catalytic cleft (Figure 1.2).

The DIAL showed much greater variation that the other components of the catalytic triad. Of the 106 SPs, only 29.6% contain the conserved DIAL motif without any substitutions. The non-underlined residues in the motif DIAL undergo substitution, but the hydrophobicity of the motif remains conserved in all SPs (Figure 1.2). The most common alternate motifs were DIAI and DVAL, each found in eleven genes. Other substitutions seen multiple times include DICL (2 genes), DLAI (4), DIAV (5), DISL (5), DVAV (6), and DLAL (8). Several substitutions were seen in only a single gene: DIAM,

DIGI, DIGL, DISI, DITL, DLAV, DLGL, DLGV, DMAI, DMAL, DVAI, DVAM, and DVGM (Table 1.4). As stated before, hydrophobic residues in the DIAL motif ensure conservation of the structure of the active site in the enzyme.

Of the 106 SPs, 88.2% contain the GDSGGP motif without substitutions around the critical residue Ser195 (Table 1.4). NDSGGP and YDSGGP were observed five times and twice in the genome, respectively. Observed substitutions occurring once include GDSGGP, GDSGDP, GDSGGA, GDSGGG, GDSGGQ, GDSGSA, GDSGSP, HDSGGP, SDSGGP, and LISGGP. Both residues adjacent to the the critical Ser195 residue in GDSGGP are as highly conserved as the serine itself is. These two residues may be important in conserving the structural stability of the serine in the catalytic cleft (Figure 1.2).

Analysis of the H-SPs

105 H-SPs, which are genes of the family missing at least one of the structural elements necessary to function as a serine protease, are classified as a homolog due to possible structural restraints and unknown function. However, they may still be functional genes as they contain start and stop codons, intron-exon boundaries, and are expressed (http://www.wfleabase.org). Analysis of the H-SPs showed that frequency of substitutions of the catalytic residue His-57 was more common than deletion of the residue or deletion of the whole motif. Deletion of His57, occurred more frequently than Asp102 and Ser195 downstream. Asp102 was conserved in 81.5% of the homologs, His57 was conserved in 27.8% and Ser195 was conserved in 14.8% of homologs. 72.2% of homologs had a substitution at Ser195 where as 13% of homologs had a deletion at this site (Table 1.4).

Analysis of SPs and H-SPs with single SP domains

Approximately 87% of the *D. pulex* active-SPs and H-SPs range from 200-500 amino acid residues in length (Figure 1.1). We began our study with a particular interest in serine proteases likely to function in food digestion. Digestive SPs are expected to contain only the serine protease domain with a signal peptide and to have a total length ~300 amino acid residues (Ross et al 2003).

We identified 53 SPs and H-SPs that match all of the expected characteristics of digestive enzymes, including 22 trypsin-like SPs (TRYs 1, 2, 3, 4A, 4B, 5B, 5C, 5D, 5F, 5G, 5I, 5J, 5K, 10A, 10B, 10C, 14, 32A, 32B, 32D, 43, and 48), 7 chymotrypsin-like SPs (CHYs 1B, 1D, 1E, 1F, 1G, 1H, and 7B), 1 elastase-like SP (ELA 1), 5 SERPs (SERPs 6, 15, 18, 17, and 18) and 18 H-SERPs (H-SERPs 009, 011, 028, 029, 031, 032, 041, 050, 051, 055, 062, 067, 072, 078, 085, 101, 102, and 106) (Table 1.3).

Analysis of Active-SPs with a Clip-domain

Six conserved cysteine residues that are upstream of the SP domain form a disulfide-bridged structure known as a clip domain. Clip domain serine proteases have been observed to be involved in embryonic development, innate immunity of arthropods, and may aid in shielding the catalytic site of the serine protease zymogen (Jiang and Kanost 2000). Ross et al. (2003) identified 37 clip-domain active SPs and H-SPs in the *Drosophila melanogaster* genome (Ross et al 2003). Only 5 SPs were found to have the clip-domain, ranging between 45-55 amino acids in length: SERP11, TRY15A, TRY15B, TRY18, and TRY336. TRY15A, TRY18, and TRY36 have a signal peptide. No homologs in the *D. pulex* genome with clip-domains were found (Table 1.3).

We aligned each clip domain using Multiple Muscle Alignment and phylogenetic reconstruction for analysis. All clip domains contain 9 amino acid residues between Cys_1 and Cys_2 and 5 amino acid residues between Cys_2 and Cys_3. Clip domain serine proteases in *D. pulex* were divided into 2 categories: category 1 contains 13 amino acid residues between Cys_4 and Cys_5 while category 2 contains 9 amino acid residues between Cys_4 and Cys_5. Category 2 was divided in two 2 subcategories based on the length of the peptide between Cys_3 and Cys_4. Category 1 and 2.1 contain 22 amino acid residues between Cys_3 and Cys_4 whereas category 2.2 contains 16 residues (Figure 1.3.A). Comparison of the two domains showed variable patterns of missense mutations, resulting in the clip domain undergoing similar rates of substitution as the SP domains on the respective genes (Figure 1.3.B). Our phylogeny and alignments suggest TRY18 is an active trypsin that may have been the first to have a deletion in the clip domain. After duplication into TRY15A and TRY15B, another deletion resulting in shorter clip-domain peptides may have occurred.

Analysis of multiple domain SPs

Of the 211 SP's extracted from the *D. pulex* genome, 45 SPs and H-SPs are multi-domain SPs. SPs with only one additional domain either carry a transmembrane domain, CBD2 domain, or a CBD4 domain. In order to understand the history of gene duplication in the SP gene family, we compared patterns of duplication events in the CBD2 domains and CDB4 domains with respect to their SP domains.

SPs only containing the CBD2 domain, a chitin binding domain involved in chitin metabolic processes (Suetake et al 2000), are found in 1 chymotrypsin-like SP (CHY2), and 6 H-SERPs (H-SERPs 007, 033, 084, 054, 077, and 005). This domain ranges

between 49-55 amino acid residues in length. Muscle multiple alignment algorithm aligned all CBD2 domains and their respective SP domains with -2.9 open gap penalty using MEGA5.10 (Tamura et al 2011). Phylogenetic analysis of the CBD2 domains was done in RAxML using the Maximum Likelihood method and GTR (General Time Reversal) nucleotide substitution model with 4 discreet GAMMA rate categories and estimated proportion of invariable sites (Stamatakis 2006). An additional test ran 1000 bootstrap replicates. The CBD2 domain is divided in to two categories. Category 2 (CHY2, H-SERP077, and H-SERP005; Figure 1.4.A) shows a deletion at Asp37 of the alignment in Figure 1.4.A. Category 2 is subdivided into category 2.1 and 2.2. Category 2.1 only contains the deletion at Asp37 whereas category 2.2 (CHY2) also contains deletions at sites Tyr4, Gly5, Glu17, and Cys18. Category 1 (H-SERPs 007, 033, 084, 054, and 083) does not show any deletion and the domain is 60 amino acid residues in length. Phylogenetic comparison of the two domains showed variable substitution and deletion patterns across the domain sequences (Figure 1.4.B).

SPs containing only one additional CBD4 domain, an insect cuticle protein domain found in early stages of development (Rebers and Willis 2001), are found in 4 trypsin-like SPs (TRYs 46, 47A, 47B, and 47C), 2 H-SERPs (H-SERP 001 and 002) and 1 SERP (SERP14). This domain ranges between 57-58 amino acid residues in length. Muscle multiple alignment algorithm aligned all CBD4 domains and their respective SP domains with -2.9 open gap penalty using MEGA5.10 (Tamura et al 2011). Phylogenetic analysis of the CBD4 domains was done in RAxML using the Maximum Likelihood method and GTR (General Time Reversal) nucleotide substitution model with 4 discreet GAMMA rate categories and estimated proportion of invariable sites (Stamatakis 2006).

An additional test ran 1000 bootstrap replicates. The CBD4 domain was divided into two categories. Category 2 (H-SERP001 and TRY46 in Figure 1.5.A) shows a deletion at Glu-56 of the alignment in Figure 1.5.A, along with amino acid substitutions at residues Iso-46, Gly-53, and Glu-55. Category 1 contains no deletion at this site. However, TRY46 and 47A show substitutions at sites Tyr-3, Leu-5, Glu-7, Gly-9, and Asp-11 in Figure 1.5.A. Comparison of the two domains showed variable substitution and deletion patterns across the domain sequences, resulting in in the CBD4 domain containing more conserved residues than the SP domain (Figure 1.5.B).

Comparison against the Drosophila melanogaster genome

Disulfide stabilized domains like the LDLa, SRCR, KH, KR, and Pan/apple are found among a few SPs that contain multiple additional domains (Table 1.4.3). Multiple-domain SPs containing LDLa repeats are said to be involved in molecular recognition and possible cholesterol metabolism (Brown and Goldstein 1986). LDLa domains are observed on TRY6, TRY7, TRY20, TRY21, and H-SERP004. The SRCR observed on TRY6 and TRY20 are proposed to be involved in protein-protein interactions and ligand binding for endocytosis if LDLa repeats are present (Resnick et al 1994). KH (K-homology) domain observed on H-SERP034 has been observed in RNA binding to function in RNA recognition and degredation (García-Mayoral et al 2007). The Kringle (KR) domain found on TRY6 is proposed to be a binding mediator and aids in regulating proteolysis (Patthy et al 1984). The Pan/apple domain has also been proposed in protein recognition or carbohydrate recognition and is also found on TRY6 (McMullen et al 1991).

211 SPs and H-SPs were used as a query against the *Drosophila melanogaster* database at NCBI database. The top 10 hits returned with orthologs with hypothetical functions based on significant E-values < 0.0005. Among these, TRY6 in *D. pulex* is most similar to Tequila in *D. melanogaster's* genome, more specifically the splice variant isoform D (E-value = 2.00E-148). Tequila in *D. melanogaster* contains 15 chitin-binding domains, 2 scavenger receptor domains, 2 LDL domains and one SP. The splice variant isoform D contains only 2 CBD2 domains, 2 scavenger receptor domains, 2 LDL domains and one SP domain. *D. pulex*'s Tequila-like SP contains 2 chitin-binding domains, 3 SRCR domains, 3 LDLa domains, 1 KR domain, 1 CLECT domain, and one SP domain. Tequila in *D. melanogaster* was hypothesized to be an ortholog of the Human Neurotrypsin which regulates long term memory formation in humans (Didelot et al 2006). Furthermore, this *Drosophila* ortholog indicates that the Tequila domain may be important in information processing in arthropods (Didelot et al 2006)

TRY20 is an LDLa and SRCR rich gene that is similar to a gene that encodes Nudel in *D. melanogaster* (E-value = 4.00E-62). The protein that encodes Nudel has been observed in *Drosophila* to be important in regulating the protease cascade for dorsal and ventral polarity of the embryo and stability of the egg (Hong and Hashimoto 1996; Lemosy et al 1998). TRY21 is also a multi-domain SP rich in LDLa domains and contains a single SRCR domain. This gene is similar to Corin in *D. melanogaster* (E-value = 6.00E-63) and found to aid in the regulation of blood circulation and coagulation in mammals (Rao et al 2001).

Try22 contains two TSP-like domains that are involved in formation of the extracellular matrix site (Bark 1993). TRY23 contains an extracellular CUB domain

involved in developmentally regulated proteins (Bork and Beckman 1993). CHY2 contains a peritrophin A-type chitin-binding domain found in proteins that line the midgut of insects and assist in digestion as well as protection (Suetake et al 2000). Genes containing this domain is usually expressed only during feeding stages (Elvin et al 1996).

Phylogenetic and Evolutionary Analysis of all SP and H-SP domains

We estimated the phylogenetic history of the SP gene family in *D. pulex* by phylogenetic reconstruction using the Maximum Likelihood method. Only clades with bootstrap values >79 are shown to observe ancestors of gene duplicates in Figure 1.6.

Group A, Clade B, and Clade C are representative of CBD4 (Chitin binding domain-4) carrying SPs. This relationship may represent a basal clade of SP genes. Group A is the unresolved relationship between domains from the homologs H-SERP001 and H-SERP002. However, in Figure 1.4.B.2, the SP domain on H-SERP001 and H-SERP002 are closely related duplicates of each other, the CBD4 domain on these genes are also closely related duplicates of each other. The CBD4 domain on TRY46 is a closely related duplicate of H-SERP001. However, the SP domain is a close duplicate of TRY47C. This may be a result of different duplication and/or selection mechanisms for each domain of the gene. A closer relationship between TRY47A, B, & C is observed in Clade C when compared to all the SP domains in the *D. pulex* genome (Figure 1.6).

Clade D shows tandem duplicates of trypsins with the loss of the CBD4 domain, along Scaffold 6 at wFleabase Gbrowse Maps (http://wfleabase.org/gbrowse/). Genes within this clade have >300 amino acid residues and all genes except for TRY44A contain a signal peptide. Clade D represents a series of tandem duplication events

resulting in 3 active-SPs that are trypsin-like (TRYs 44A, B, and C) as well as their 3 homologs (H-SERPs 086, 087, and 076) (Figure 1.6).

Chymotrypsin-like genes that form Clade E are located on Scaffold 36 (http://wfleabase.org/gbrowse/) and may be the result of the divergence of a trypsin-like SP (TRY043) and a primitive chymotrypsin-like active-SP. The primitive chymotrypsin may have underwent tandem duplication events resulting in 3 chymotrypsin-like SPs (CHY7A, CHY7B, and their homolog H-SERP057) and 5 serine protease-like genes and their homologs (SERPs 12, 13, and H-SERPs 040, 072, and 071). CHY7B is the only SP in this clade that shows all attributes of a digestive protease. CHY7A and H-SERP057 are missing the signal peptide for subcellular localization (Figure 1.6).

F is the largest clade with a bootstrap value of 96 and may represent the origin of a putative expansion of SP domains resulting in novel SPs functioning in various biological processes other than regulation of development and digestion. Clade F.1 represents the duplication events of trypsin-like SPs (TRY32 A, B, C, and D) and meet the criteria of being a digestive serine protease with the exception of TRY32 showing the loss of the signal peptide. The duplication events spanned across 4 scaffolds (http://wfleabase.org/gbrowse/).

Within clade F.2, is the only Elastase-like active-SP, ELA1, with substrate specificity Ser-189, Val-216, and Ala-226 found in the *D. pulex* genome which is located on scaffold 452 (http://wfleabase.org/gbrowse/). ELA1 has a homolog, H-SERP090, on scaffold 6, (http://wfleabase.org/gbrowse/) and also contains the specificity of Ser-189, Val-216, and Ala-226 but does not have the required signal peptide or the TAAHC motif required for structural stability of the catalytic triad. Also within this clade is

Chymotrypsin-like active-SPs (CHYs 5A & 5B) along with their serine-protease like homologs (H-SERPs 050 and 073) (Figure 1.6).

Clade F.3 represents the divergence of Trypsin-like SPs (Figure 1.6 Clade F.3A) and their homologs (Figure 1.6. Clade F.3B). The duplicates within Clade F.3A conserved the digestive Trypsin-like genes (TRY33A, TRY33B, and TRY34). These genes along with the remainder of the genes within this clade are along scaffolds 25 and 29 (http://wfleabase.org/gbrowse/). The largest clade of recent duplicates of homologs is observed on Clade F.3B. Within this clade, partial domain deletion and partial gene deletion may have resulted in the duplication events of H-SPs. Although some genes have dispersed across the genome, the more recent tandem duplications within this clade occurred on scaffold 36 (http://wfleabase.org/gbrowse/).

Clade F.4 shows many SP-like genes with inherited complex domain architectures, or having multiple accessory domains in addition to the SP domain. Many of these SPs and their relationships with one another are left unresolved. However, three specific expansions are evident as seen in clade F.4A, F.4B and F.4C. Clade F.4A is the second largest expansion of homologs within the phylogeny. Clade F.4B contains active-SPs that may be highly conserved tandem duplicates: TRY2, TRY3, TRY4A, TRY4B and TRY5B-5K. Expansion in Clade F.4B may be primitive with the exception of the recent duplication events between pairs TRY5F and TRY5L as well as TRY5B and TRY5K. Clade F.4C shows the expansion of Chymotrypsin-like SPs and their homologs. CHY1A-H expanded along scaffold 29 suggesting tandem duplications. H-SERP061, H-SERP033, H-SERP007, H-SERP084, H-SERP054, H-SERP083, H-SERP035, H-SERP077, and H-SERP006 expanded along scaffold 18 suggesting another occurrence of

tandem duplication events within this expansion. The CBD2 domain has been observed across this clade. Figure 1.3.B shows that the CBD2 domain duplicated separately from the SP domain. The CBD2 domain on CHY2 is much more conserved than that SP domain on CHY2, which is a closely related duplicate to H-SERP005 (Figure 1.3.B.2). The CBD2 Domain on H-SERP005 is closely related to the duplicates of H-SERP033 and H-SERP077 (Figure 1.3.B.1).

Subcellular Localization

The presence of a cleavage site for subcellular localization was hypothesized for each gene and indicated as part of the domain architecture in Figure 1.6. To further elucidate subcellular localization of each gene, the presence and probability of subcellular localization was estimated using pTARGET web interface (Guda 2006). Each clade shows variability in the localization within recent gene duplications, with the exception of the recent duplicates observed in Figure 1.6 Clade F.3.A. SERP7 (93.90%), SERP8 (75.10%), SERP9 (93.90%), H-SERP038 (100.00%), H-SERP076 (93.90%), TRY34 (93.90%), TRY33B (100.00%), and TRY33A (75.10%) show the probability of being localized in the lysosome. SERP10 (93.90%), however, is shown to be a duplicate localized on the Plasma membrane, most likely the result of a deletion resulting in the absence of the signal peptide for lysosome localization.

The five clip-domain active-SPs show variable subcellular localization. TRY15A (81.40%), TRY15B (87.60%), and TRY18 (87.60%) may be localized as an extracellular/secretory protease whereas TRY36 may be a plasma membrane Trypsin and SERP11 may be a lysosome serine protease.

Selection on SPs and H-SPs

Clades F.4A, F.4B, and F.4C in Figure 1.5 were chosen for selection analysis because of 1.) gene similarity, 2.) Confidence that these genes are the product of tandem duplication of one ancestral gene, and 3.) monophyletic patterns of the SP subfamilies within Clade F.4 of Figure 1.6. Selection tests using the Nei-Geobori substitution model retrieved values suggesting significant evidence for purifying selection (*P*-value < 0.0001) (Table 1.2). Positive selection was not observed within these clades. Analysis between the clades was not possible due to the dilution of amino acid substitution, suggesting that the divergence of Clades F.4A, F.4B, and F.4C was primitive and then underwent tandem duplication.

1.4 DISCUSSION

The serine protease gene family makes up approximately 73-85% of the enzymatic activity in the gut of *Daphnia* (Elert et al 2003). Because *Daphnia* are a model organism in studying mechanisms of development, cell function, immune response, disease, and the genetic basis of phenotypic patterns, their genome was used as an additional model in studying the evolution of the Serine Protease gene family in arthropods. More specifically, the large protease gene pool, across taxa of arthropods, is exposed to natural selection or alternative expression, which may quickly adapt to SP inhibiters, Serpins, in the control of agricultural pests, land (Ross et al 2003) or water.

We found 211 serine protease-like genes, including their homologs, in the *Daphnia pulex* genome. In other arthropods, previous studies found 57 SP-like genes in the *Apis mellifera* genome, 305 SPs in *A. gambiae*, and 206 SPs have been found in *Drosophila melanogaster* (Ross et al 2003; Zou et al 2006; Wu et al 2009), suggesting that duplication events in arthropods after divergence is variable. Evidence of positive

selection was found on sites located in the binding region of the serine protease genes in the *A. gambaiae* genome and may suggest adaptive evolution for the process of digestion of food (Wu et al 2009).

The structural integrity of the catalytic cleft is influenced by the interaction of the amino acid residues within the primary structure of the translated peptide. Events that eliminate or change the structural integrity will render the digestive enzyme inactive. These events favor the conservation of the DIAL motif, containing the catalytic residue Aspartate, as well as conservation of the surrounding hydrophobic residues. However conservation of the other two motifs containing the other two residues of the catalytic triad, TAAHC and GDSGGP, were susceptible to events that rendered the motif inactive for digestive function. Overall changes in the redundant copies of serine proteases still preserve the biochemical composition of the motifs in the catalytic cleft of active serine proteases as well as the size of the overall protease, 200-300 amino acid residues in length. Subcellular localization patterns are variable within the gene family, even the digestive serine proteases. The expansion of digestive SPs is observed across taxa of arthropod to be common rather than abundant and could reinforce the assumption that expansion within this gene family is a neutral process. Analysis of the serine protease gene family in the more closely related species of *Daphnia magna* will aid in isolating orthologous serine proteases which could be specific to crustacea. Possible orthologs within the 53 hypothesized digestive serine proteases found in *Daphnia pulex* could aid in isolating targeted digestive proteases for protease inhibitors, specifically serpins, found in algae.

A possible mechanism of evolution in the SP gene family involves unequal crossing over, which was observed in *Drosophila melanogaster* genome, and may have increased the chance of yielding large expansions of SPs and SP-Hs (Ross et al 2003). Large expansions of Homologs, Trypsin SPs, and Chymotrypsin SPs (Figure 1.6 in Clades F.4a, F.4b, and F.4c respectively), were observed in the *Daphnia* genome to be novel expansions that may be the result of sequence divergence following gene duplication. Negative selection along with variable subcellular localization within these expansions were observed, suggesting the SPs within these clades did not evolve for the beneficial increase in dosage, but rather reinforced the original copy to maintain its original function. The CBD2 domain on this chymotrypsin (CHY2) within clade F.4C of Figure 1.6 is a part of the peritrophic matrix proteins of chitinases and is found on the plasma membrane (Shen 1998; Suetake et al 2000). We hypothesize CHY2 to be the reinforced basal copy within this expansion and to be a gut-specific chitinase involved in food digestion.

Clip-domain serine proteases are proposed to be involved in the innate immunity of arthropods (Jiang and Kanost 2000). In the *A. gambai* 41 clip-domain SPs were observed, 18 clip-domain SPs were found in *A. mellifera*, and 37 were found *in Drosophila melanogaster* (Rawlings and Barrett 1993; Hong and Hashimoto 1996; Lemosy et al 1998; Jang et al 2008). However, only 5 active clip-domain serine proteases were found in the *Daphnia pulex*, genome (TRYs 36, 18, 15B, 15A, and SERP 11). These genes are widely dispersed across the *Daphnia pulex* genome and their phylogenetic position suggests the expansion of the Clip-domain serine proteases to be basal whole gene duplicates of each other. Clip-domain serine protease expansions became more prominent

in the class Hexapoda than what was observed in Crustacea after the divergence within Arthropoda.

Based on scaffold positioning of each gene, we hypothesize homologous recombination to be involved in the mechanism of expansion of not only the CBD2 carrying SPs, but also the CBD4 carrying SPs (TRYs 46, 47A, 47B, and 47C; SERP 14; H-SERPs 001 and 002). The presence of the CBD4 domain on serine protease like genes was not observed in other taxa of arthropoda, when compared to *Drosophila melanogaster*. We observe more conserved sites within the CBD4 domain than the SP domain of these genes; either suggesting that this domain is a recent insertion or conservation of the CBD4 domain is putative for the genes function. Scaffold 6, in the sequenced *Daphnia pulex* genome on WFleabase (http://www.wfleabase.org), shows the CBD4 carrying SPs to be within close proximity to each other and highly conserved in relation to the more ancesteral neighboring SP domains within the Serine protease gene family. CBD4 is found in the cuticular proteins of some invertebrates during embryological (Rebers and Willis 2001). In addition, 11 more SP-like genes are found to be located on Scaffold 6 and are also observed to be clustered near the CBD4 carrying SPs in the phylogeny (Figure 1.6). Transposable elements along this scaffold, along with other scaffolds containing SP domain expansions (Scaffolds 18, 25, 52, 42, 72, 36, and 29) would be helpful in isolating an ancestral transposable elements affecting the expansion of serine protease domains and the Chitin-binding 4 domains.

Three serine protease-like genes of interest were found to have essential functions and to be conserved within arthropods. Tequila (TRY6) could be studied for its involvement in information processing in *Daphnia* (Didelot et al 2006). Nudel (TRY20)

could be involved in regulating the protease cascade for dorsal and ventral polarity of the embryo and stability of the egg (Hong and Hashimoto 1996; Lemosy et al 1998). Corin (TRY21) is hypothesized to be involved in blood circulation and coagulation (Rao et al 2001). These genes were annotated for gene expression analysis in *Daphnia pulex.*

Because of saturation, the SPs do not resolve into strongly supported monophyletic clades on the basis of the serine protease domain itself. However, some clades exhibit clustering of subfamilies and may indicate recent duplication events. We tested for the presence of non-neutral selection within these clades. The subfamilies in Clades F.4A, F.4B, and F.4C are observed to be under purifying selection to reinforce the ancestral gene's function. We propose this to also be present in a closely related species *Daphnia magna* because of exposure to similar diets of phytoplankton.

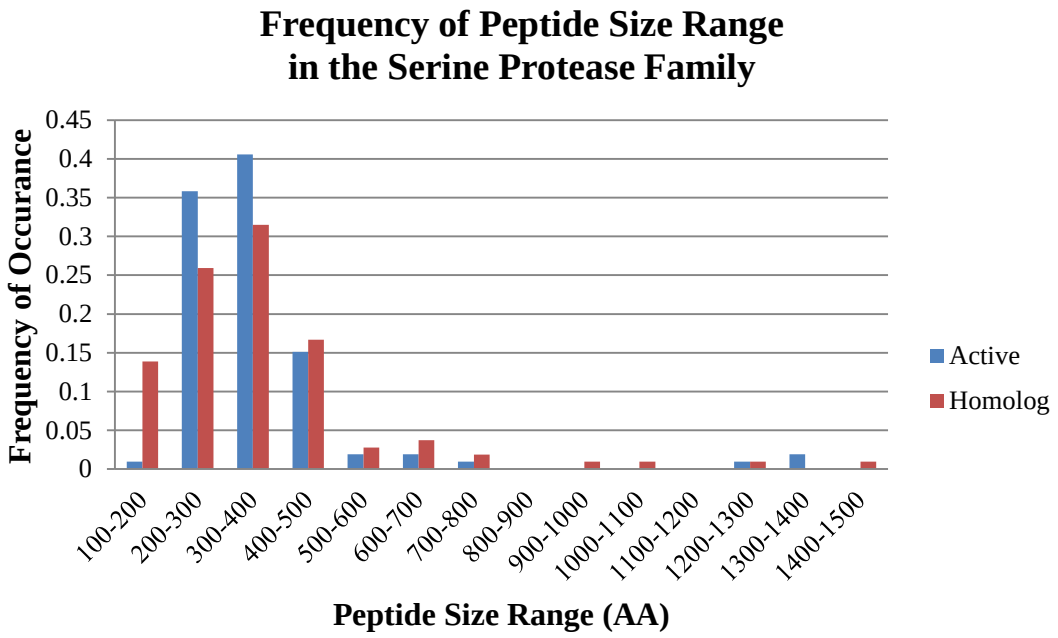


Figure 1.1. **Frequency of peptide size in amino acids in the Serine Protease family.** Range of length is in amino acids. Blue bars represent the frequency of serine proteases (SPs) found in the *Daphnia pulex* genome. Red bars represent the homologs of the serine protease genes (H-SPs).

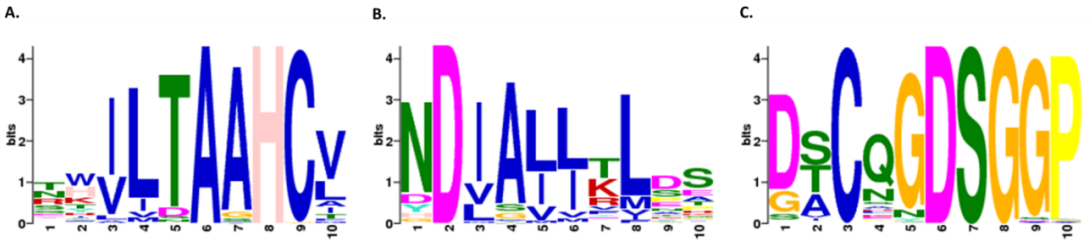


Figure 1.2. Features of the Motifs in the Catalytic Triad of Complete SPs. The residues involved in peptide chain hydrolysis are embedded in the motifs A, B, and C. Height of the logo, in bits, represents the probability of that residue occurring in that position multiplied by the total amount of information in that position. The colors of each residue represent the following: Blue: most hydrophobic; Green: Polar, non-charged, non-aliphatic; Magenta: Acidic; Red: Positively Charged(Bailey et al 2006).

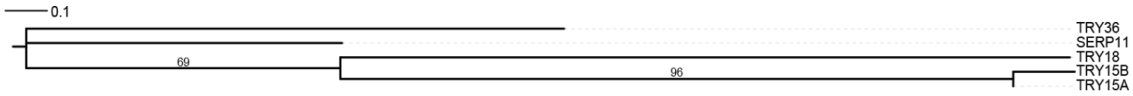
A

Clip Domain

```
                |         |         |         |         |         |       Cat.
TRY36           CLTREGNIGYCTSIRSCYPRLNKFHHFNFESRTLAIRGACIYHRADDRQVYGICCP 1
SERP11          CWMSDGKSGLCGPVRSCHPHDELQEPLNPESRMLPSRTLCGYVNKNGKQDTGVCCP 1
TRY18           CQTPEGVVGTCTPLTNCPHLADMLSVPSPAILNFLRQSICGY----EGYDPKVCCS 2.1
TRY15A          CLTPISQSGRCRFVQHCA-LPEII-----VTLNAFVTYACSI----GSDYMGVCCP 2.2
TRY15B          CSTPLSQSGRCRFVQHCA-RQEII-----ATLNAFVSYACPI----GSDYMGVCCP 2.2
                *    .  * *  :  *    :               *           :**.
```

B.1

CLIP domains on the CLIP domain SPs



B.2

SP domains on the CLIP domain SPs

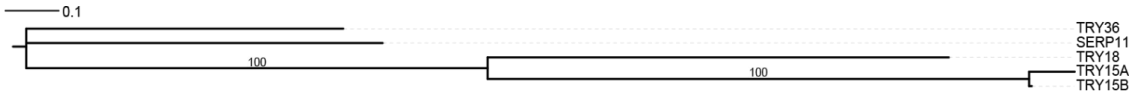


Figure 1.3. Sequence comparison and phylogenetic relationships among the *Daphnia pulex* clip-domain SPs and H-SPs. A. alignment of the clip domain sequences. Six conserved Cys residues form 3 disulphide bonds. B.1 Phylogenetic tree based on an alignment of the CLIP domains. B.2 Phylogenetic tree based on an alignment of the SP domains. Category number indicates genes sharing similar in-del patterns.

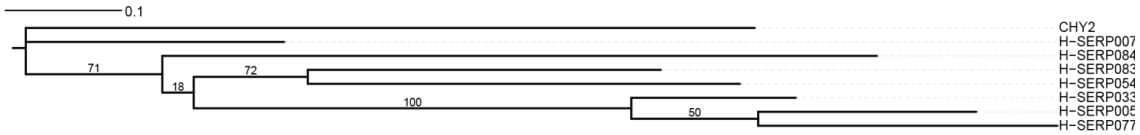
A

CBD2 Domain

```
                |         |         |         |         |         | Cat.
CHY2            FSC--QSDGIKSNPND--CNSFYMCSNGTPYLFNCP-GGLVFNPQLQQCDYRQNVPQCNY 2.2
H-SERP007       FTCEGKPSGIYPNPACDCCTTFYKCSNGYAYLYDCPDAGTVFDPQISVCVYPGNLPACGG 1
H-SERP033       FDCTNKVDGNYPNPASTCSATFYMCSNGDAYLFTCAQAGTVYRPDIYACDWPSNVAGCAX 1
H-SERP084       FSCKNRENGLYPYPDLECTKYFYYCSNGMAYLYDCPVAGTIFYYAMCNCEFPGNVPGCED 1
H-SERP054       FSCRNKPDGIYANPFDDCSIIFYMCFNSNKYEYTCPDAGTVFNPQICACDFPYNVPACGV 1
H-SERP083       FDCKGKPNGVYPNPWNDCSRTFFYCSNGYSYEYICPDAGTVFNEFICDCDYPSNVAGCLD 1
H-SERP077       FTCTGKTDGNYPNPASSCSANFYTCSPGNASLFACP-SGLVYHAEIGVCDWPFNVAGCKK 2.1
H-SERP005       FSCTGKPNGNYPNPESNCSNTFYTCSNGNSYLFNCA-SDLVYREEIGVCDFPSNVAGCHX 2.1
                * *  : .*  . *       *: *  .    : *. .. ::   :  * :  *:. *
```

B.1

CBD2 domains on CBD2 carrying SPs



B.2

SP domains on CBD2 Carrying SPs

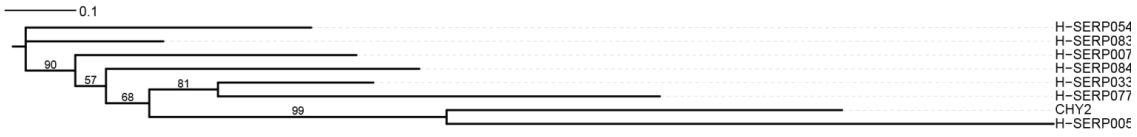


Figure 1.4. Sequence comparison and phylogenetic relationships among the *Daphnia pulex* CBD2 SPs and H-SPs. A. alignment of the CBD2 domain sequences. B.1 Phylogenetic tree based on an alignment of the CBD2 domains. B.2 Phylogenetic tree based on an alignment of the SP domains. Category number indicates genes sharing similar residues.

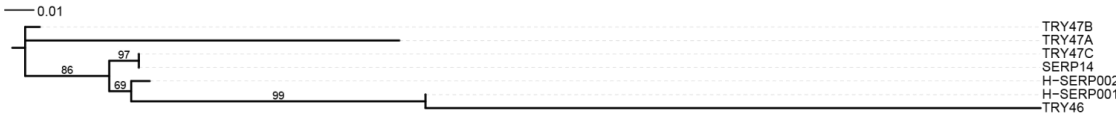
A

CBD4 Domain

```
                |         |         |         |         |         |          Cat.
H-SERP001       QWYTLDEQRRANFGYAYPGQAASNIRDADGNMAGSWSYVDADGNLVRATYTADKR-GF 1
H-SERP002       QWYTLDEQRRANFGYAYPGQAASNIRDADGNMAGSWSYVDADGNLIRATYTAGREQGF 2
TRY46           QWHTQDGQGRASFGYSYSGQAAATIRDPDGNMAGSWSYIDLDGNLVRATYTADER-GF 1
SERP14          QWYTLDEQRRANFGYAYPGQAASNIRDADGNMAGSWSYVDADGNLIRATYTAGREQGF 2
TRY47C          QWYTLDEQRRANFGYAYPGQAASNIRDADGNMAGSWSYVDADGNLIRATYTAGREQGF 2
TRY47B          QWYTLDGQGQANFGYAYPGQAASNIRDANGNMAGSWSYVDADGNLIRATYTAGREQGF 2
TRY47A          QWHAQNGQGEASFGYAYPGQAASNIRDANGNMAGSWAFVDADGNLIRATYTAGREQGF 2
                **:: : * .*.***:*.****:.***.:********:::* ****:******... **
```

B.1

CBD4 domains on the CBD4 Carrying SPs



B.2

SP domains on the CBD4 Carrying SPs

Figure 1.5. Sequence comparison and phylogenetic relationships among the *Daphnia pulex* CBD4 SPs and H-SPs. A. alignment of the CBD4 domain sequences. B.1 Phylogenetic tree based on an alignment of the CBD4 domains. B.2 Phylogenetic tree based on an alignment of the SP domains. Category number indicates genes sharing similar residue.

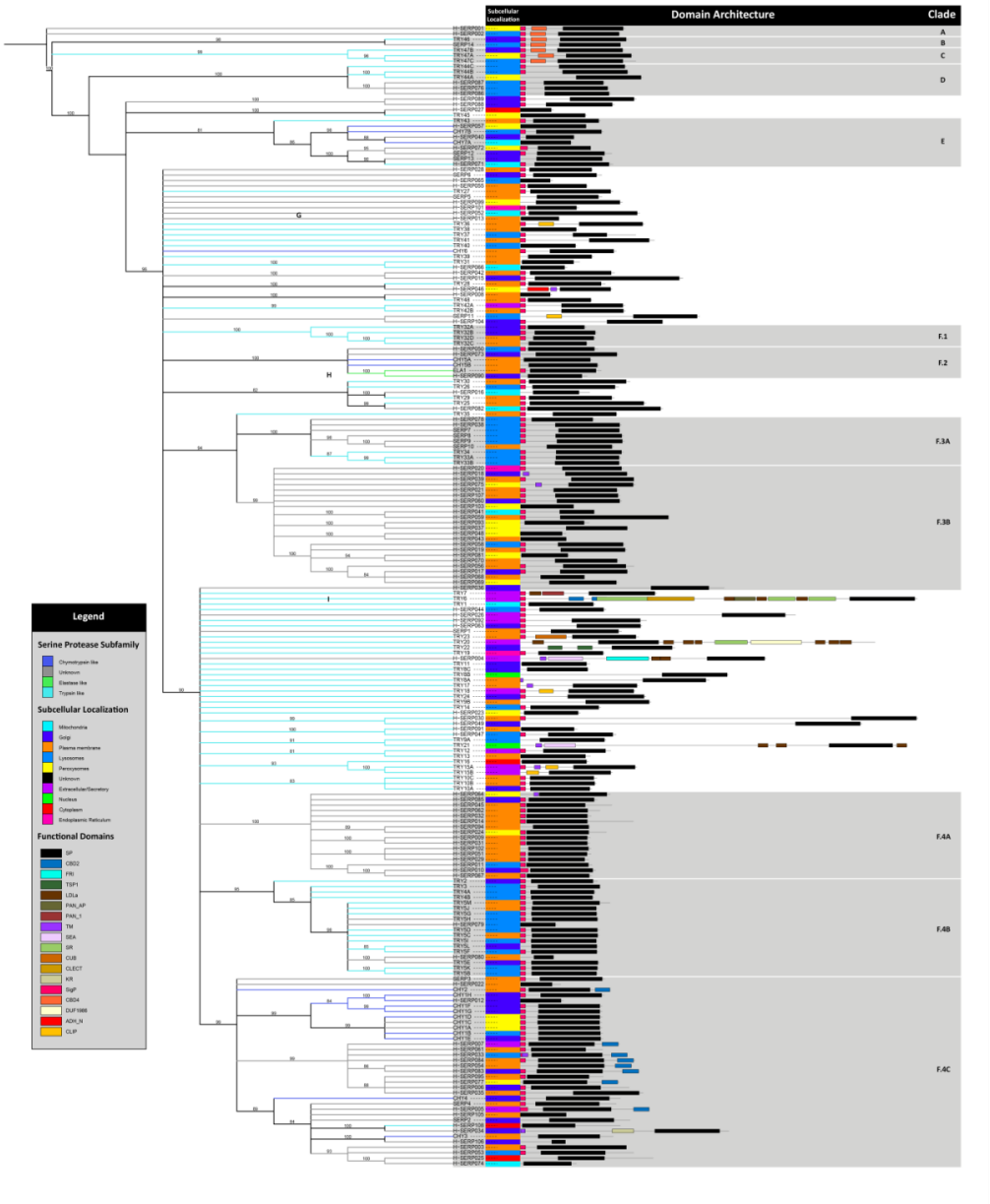


Figure 1.6. Phylogenetic relationship of all SP domains found in the *Daphnia pulex* genome. Phylogenetic analysis of the catalytic SP domains was performed as described in Section 1.2 using RAxML. Branch colors represent the subfamily classification of each serine protease which is dependent on the substrate specificity of the amino acid residues. The colored bars indicate the hypothesized subcellular localization of each gene. The colored domain architecture represents additional functional domains that may be present on each SP containing gene. Putative gene clusters are highlighted and labeled for analysis. Original gene names from previous studies are included and not fixed (see appendix).

Table 1.1. Frequency of genes in each subfamily of Serine Protease gene family. Subfamily classification was dependent on the substrate specificity of the active site and quantified in both the active serine proteases (SPs) and in the inactive homologs (H-SPs). Substrate specificity is governed by three residues surrounding the GDSGGP motif. Residues for each subfamily are as follows: DGG for trypsins, SGG or GGG for chymotrypsins, SVA or SAA for elastases and XXX for residues of unknown substrate specificity (SERP) (Perona and Craik 1995).

Serine Protease Subfamily	Active SPs	Inactive H-SPs
Trypsin (TRY)	73	13
Chymotrypsin (CHY)	14	2
Elastase (ELA)	1	1
Serine Protease-like (SERP)	18	89

Table 1.2. Estimates of Codon-based Evolutionary Divergence between Sequences The mean difference between the nonsynonymous and synonymous distances per site from averaging over all sequences from clades F.4A, F.4B, and then F.4C. of Figure 1.6. Standard error estimate(s) are shown in the last column. Analyses were conducted using the Nei-Gojobori model (Nei and Gojoborit 1986). The analysis involved 18 nucleotide sequences from domains of clade F.4.A, 16 nucleotide sequences from domains of clade F.4.B, and 35 nucleotide sequences from the domains of clade F.4C. Evolutionary analysis were conducted in MEGA 5 (Tamura et al 2011). Hypothesis testing against the null ($H_0 = 0$) was done using the one-sample T-test. We reject the null for overall mean distance ($H_0 = 0$) if *P*-value > 0.05.

Overall Mean Distance	dN-dS	StdDev	*P*-value
Homolog Clade (Figure 1.6.F.4.A)	**-1.461**	**0.088**	**<0.0001**
Trypsin Clade (Figure 1.6.F.4.B)	**-1.415**	**0.076**	**<0.0001**
Chymotrypsin Clade (Figure 1.6.F.4.C)	**-0.850**	**0.108**	**<0.0001**

Table 1.3. Characteristics of each Serine Protease gene in the *Daphnia pulex* genome.
Gene names, and fixed gene names[a] (see appendix) for each serine protease are catalogued. Signal peptide prediction from smart.embl-heidelberg.de is indicated by Y (yes) or N (no) (Onting 1998). This prediction method indicates possible cleavage sites of secretory proteins for movement across the Endoplasmic Reticulum and is not related to subcellular localization predictions. Probability of the Subcellular Localization predictions are calculated by pTARGET (Guda 2006). Superscript $^{\Psi}$ indicates whether the gene is a pseudogene or not, predicted from wFLEABASE ESTs and Expression maps (Gilbert and Singan, V.R., Colbourne 2005). Prediction of accessory domains from smart.embl-heidelberg.de (Onting 1998) and acronyms stand for the following:

SP Trypsin-like serine protease
CBD2 Chitin-binding domain type 2
LC Region of low compositional complexity
* Signal peptide
TM Transmembrane Domain
C Clip-Domain
RPT Internal repeat
CBD4 Chitin-binding domain type 4
LDLa Cysteine-rich Low-density lipoprotein receptor domain class A
SRCR Egg peptide speract receptor
DUF Function unknown
SEA Domain found in sea urchin proposed to regulate or bind carbohydrate sidechains
TSP1 Thrombospondin type 1 repeats
CUB Domain commonly present in developmentally-regulated proteins.
SR Scavenger receptor Cys-rich
CLECT C-type lectin (CTL) or carbohydrate-recognition domain (CRD)
KR Kringle domain
PAN_AP From a subfamily of APPLE domains

Gene Name	Scaffold	Clade	Length (aa)	Signal Peptide	Subcellular Localization	Probability	Domain Architecture

SERP15[a]	scaffold_29:1208117-1209646	F.4C	302	Y	Peroxysomes	81.40%	SP
CHY1B	scaffold_29:1214169-1215591	F.4C	304	Y	Lysosomes	81.40%	SP
SERP16[a]	scaffold_29:1215928-1217469	F.4C	301	Y	Peroxysomes	81.40%	SP
CHY1D	scaffold_29:1218557-1220207	F.4C	302	Y	Peroxysomes	81.40%	SP
CHY1E	scaffold_29:1221007-1222706	F.4C	305	Y	Golgi	75.10%	SP
CHY1F	scaffold_29:1223853-1225266	F.4C	302	Y	Golgi	81.40%	SP
CHY1G	scaffold_29:1229475-1230909	F.4C	302	Y	Golgi	81.40%	SP
CHY1H	scaffold_29:1237677-1239164	F.4C	309	Y	Golgi	81.40%	SP
CHY2	scaffold_18:1292962-1294526	F.4C	331	Y	Plasma membrane	87.60%	SP/CBD2
CHY3	scaffold_18:1131261-1132510	F.4C	343	N	Plasma membrane	81.40%	LC/SP/LC
CHY4	scaffold_18:969980-971068	F.4C	369	Y	Golgi	75.10%	LC/SP/LC
CHY5A	scaffold_4:567291-568494	F.2	265	N	Plasma membrane	75.10%	SP
CHY5B	scaffold_86:71267-72947	F.2	291	N	Plasma membrane	81.40%	SP
CHY6	scaffold_28:307795-309564	F	355	Y	Plasma membrane	75.10%	SP
CHY7A	scaffold_36:1006339-1007494	E	193	N	Mitochondria	62.60%	SP

CHY7B	scaffold_36:1000228-1001580	E	306	Y	Lysosomes	93.90%	SP
ELA1	scaffold_452:10676-12417	F.2	299	Y	Plasma membrane	75.10%	SP
SERP1	scaffold_143:93263-99290	F.4	374	Y	Plasma membrane	75.10%	TM/LC/SP
SERP10	scaffold_29:660492-662257	F.4A	345	N	Plasma membrane	93.90%	LC/SP
SERP11	scaffold_52:625674-626930	F	661	N	Lysosomes	87.60%	C/2(RPT)/4(LC)/SP
SERP12	scaffold_36:1003646-1005373	E	338	Y	Golgi	81.40%	SP/LC
SERP13	scaffold_6:2268254-2269246	E	304	N	Golgi	87.60%	SP
SERP14	scaffold_18:1186790-1187973	C	370	Y	Lysosomes	93.90%	CBD4/SP
SERP2	scaffold_18:1186365-1187973	F.4C	405	N	Golgi	75.10%	SP/LC
SERP3	scaffold_2:2462158-2463978	F.4C	311	Y	Plasma membrane	81.40%	SP
SERP4	scaffold_75:298879-300199	F.4C	354	Y	Plasma membrane	93.90%	SP/LC
SERP5	scaffold_72:606543-609411	F	302	N	Plasma membrane	75.10%	SP
ΨSERP6	scaffold_1432:2645-4172	F	301	Y	Golgi	68.90%	SP
SERP7	scaffold_29:640538-643027	F.3A	368	Y	Lysosomes	93.90%	LC/SP
SERP8	scaffold_29:669392-671400	F.3A	380	Y	Lysosomes	75.10%	SP

SERP9	scaffold_29:649138-651042	F.3A	383	Y	Lysosomes	93.90%	SP
TRY1	scaffold_15:1352090-1354801	F.4	277	Y	Mitochondria	81.40%	SP
TRY10A	scaffold_58:679350-680463	H	264	Y	Golgi	75.10%	SP
TRY10B	scaffold_58:682435-683648	F.4	281	Y	Plasma membrane	87.60%	SP
TRY10C	scaffold_58:685294-686510	F.4	283	Y	Plasma membrane	81.40%	SP
TRY11	scaffold_78:508276-510423	F.4	257	N	Golgi	81.40%	SP
TRY12	scaffold_94:349792-350856	F.4	334	Y	Extracellular/Secretory	75.10%	SP
TRY13	scaffold_25:597273-598595	F.4	259	N	Plasma membrane	93.90%	SP
TRY14	scaffold_25:1166998-1169145	F.4	297	Y	Lysosomes	93.90%	SP
TRY15A	scaffold_66:471634-472661	F.4	430	Y	Extracellular/Secretory	81.40%	TM/C/LC/SP
TRY15B	scaffold_42:952193-954218	F.4	340	N	Extracellular/Secretory	87.60%	LC/C/SP
TRY16	scaffold_42:937598-938714	F.4	252	N	cytoplasm	75.10%	SP
TRY17	scaffold_59:167076-168982	F.4	437	N	Plasma membrane	93.90%	TM/SP
TRY18	scaffold_59:163533-165576	F.4	424	Y	Extracellular/Secretory	87.60%	LC/C/LC/SP
TRY19	scaffold_20:712645-716498	F.4	311	Y	Endoplasmic Reticulum	87.60%	LC/SP

TRY2	scaffold_53:657364-658975	F.4B	285	Y	Golgi	81.40%	SP
TRY20	scaffold_19:349775-355175	F.4	1308	N	Extracellular/Secretory	93.90%	LDLa/SP/LDLa/LC/2(LDLa)/SRCR/DUF1986/3(LDLa)
TRY21	scaffold_64:579258-580418	F.4	1428	N	Nucleus	100.00%	LC/TM/SEA/8(LC)/2(LDLa)/2(LC)/SP/LDLa
TRY22	scaffold_146:67633-69229	F.4	571	N	Golgi	75.10%	LC/TSP1/LC/TSP1/LC/SP
TRY23	scaffold_83:15235-17228	F.4	437	Y	Plasma membrane	75.10%	CUB/SP
TRY24	scaffold_11:763004-764198	F.4	463	Y	Golgi	81.40%	SP
TRY25	scaffold_17:1293795-1296035	F	464	Y	Plasma membrane	93.90%	SP
TRY26	scaffold_72:487642-489410	F	363	Y	Lysosomes	81.40%	SP
TRY27	scaffold_173:149713-151403	F	342	Y	Plasma membrane	87.60%	SP
TRY28	scaffold_72:452716-454469	F	314	Y	Plasma membrane	93.90%	SP
TRY29	scaffold_72:449027-451070	F	347	Y	Plasma membrane	75.10%	SP
TRY3	scaffold_61:647364-649170	F.4B	280	Y	Lysosomes	87.60%	SP
TRY30	scaffold_17:184128-185905	F	405	Y	Plasma membrane	75.10%	SP
TRY31	scaffold_16:1540112-1542257	F	219	N	Plasma membrane	93.90%	SP
TRY32A	scaffold_57:415280-416377	F	244	Y	Golgi	75.10%	SP

TRY32B	scaffold_79:359735-361603	F	284	Y	Golgi	75.10%	SP
ΨTRY32C	scaffold_818:7683-9783	F	249	N	Plasma membrane	93.90%	SP
TRY32D	scaffold_167:38937-40314	F	279	Y	Plasma membrane	87.60%	SP
TRY33A	scaffold_29:674151-676178	F.3A	371	Y	Lysosomes	75.10%	SP
TRY33B	scaffold_29:665417-667351	F.3A	371	Y	Lysosomes	100.00%	SP
TRY34	scaffold_29:626410-628274	F.3A	380	Y	Lysosomes	93.90%	TM/SP
TRY35	scaffold_36:675750-677100	F.3	359	Y	Plasma membrane	81.40%	SP
TRY36	scaffold_25:1160431-1161768	F	458	Y	Plasma membrane	87.60%	C/SP
TRY37	scaffold_25:1141432-1143073	F	426	Y	Lysosomes	100.00%	LC/SP
TRY38	scaffold_25:1147223-1149730	F	213	N	Plasma membrane	75.10%	SP
TRY39	scaffold_52:639171-640479	F	270	N	Plasma membrane	93.90%	LC/SP
TRY40	scaffold_52:644871-646202	F	209	N	Lysosomes	93.90%	SP
TRY41	scaffold_52:631054-636117	F	495	Y	Plasma membrane	87.60%	LC/RPT/LC/RPT/SP
TRY42A	scaffold_7:1455592-1456662	F	386	Y	Extracellular/Secretory	100.00%	TM/SP
TRY42B	scaffold_52:706562-707563	F	388	Y	Plasma membrane	87.60%	SP

TRY43	scaffold_6:2215769-2217039	F.3A	296	Y	Plasma membrane	93.90%	SP
ΨTRY44A	scaffold_6:2169329-2171542	D	451	N	Peroxysomes	81.40%	2(LC)/SP
TRY44B	scaffold_6:2160303-2162026	D	401	Y	Lysosomes	87.60%	SP
TRY44C	scaffold_36:1164924-1166613	D	391	Y	Lysosomes	81.40%	SP
TRY45	scaffold_6:2293387-2294870		245	N	Peroxysomes	81.40%	SP
TRY46	scaffold_6:2256296-2257239	B	397	Y	Golgi	81.40%	CBD4/SP
TRY47A	scaffold_6:2288318-2289461	C	414	Y	Peroxysomes	87.60%	CBD4/LC/SP
TRY47B	scaffold_6:2279707-2280707	C	382	Y	Golgi	81.40%	CBD4/SP
TRY47C	scaffold_6:2275329-2277465	C	426	Y	Lysosomes	93.90%	CBD4/LC/SP/LC
TRY48	scaffold_105:40545-42304	F	266	Y	Plasma membrane	93.90%	SP
TRY4A	scaffold_23:1042595-1044361	F.4B	278	Y	Lysosomes	93.90%	SP
TRY4B	scaffold_23:1034539-1036426	F.4B	272	Y	Lysosomes	100.00%	SP
TRY5B	scaffold_42:910015-911951	F.4B	287	Y	Lysosomes	81.40%	SP
TRY5C	scaffold_42:912602-914102	F.4B	291	Y	Plasma membrane	75.10%	SP
TRY5D	scaffold_42:914582-916190	F.4B	292	Y	Lysosomes	87.60%	SP

TRY5E SERP17[a]	scaffold_42:916833-918416	F.4B	292	Y	Golgi	75.10%	LC/SP
TRY5F	scaffold_42:922639-924365	F.4B	288	Y	Lysosomes	81.40%	SP
TRY5G	scaffold_42:925469-927146	F.4B	287	Y	Lysosomes	100.00%	LC/SP
SERP18[a]	scaffold_42:928294-929760	F.4B	290	Y	Lysosomes	93.90%	SP
TRY5I	scaffold_42:930019-931606	F.4B	286	Y	Lysosomes	100.00%	SP
TRY5J	scaffold_42:932263-933996	F.4B	286	Y	Plasma membrane	75.10%	SP
TRY5K	scaffold_85:91288-92886	F.4B	290	Y	Lysosomes	87.60%	SP
TRY5L	scaffold_85:93151-94845	F.4B	288	N	Golgi	75.10%	SP
TRY5M	scaffold_245:57840-59550	F.4B	331	Y	Plasma membrane	81.40%	SP
TRY6	scaffold_17:78402-84229	F.4	1464	Y	Extracellular/Secretory	100.00%	3(LC)/CBD2/LC/CBD2/SR/CLECT/KR/LDLa/PAN_AP/LDLa/SR/LDLa/SR/SP
TRY7	scaffold_17:880316-881427	F.4	504	Y	Extracellular/Secretory	100.00%	LDLa/PAN_1/LC/SP
TRY8A	scaffold_23:783527-785611	F.4	697	N	Plasma membrane	93.90%	TM/5(LC)/SP
TRY8B	scaffold_52:148746-150662	F.4	768	N	Nucleus	81.40%	7(LC)/SP
TRY8C	scaffold_1310:1964-3049	F.4	255	N	Golgi	93.90%	SP

TRY9A	scaffold_38:319591-323523	F.4	317	N	Lysosomes	81.40%	LC/SP
TRY9B	scaffold_20:346803-349301	F.4	482	N	Plasma membrane	81.40%	LC/SP
H-SERP001	scaffold_6:2270795-2272425	A	387	Y	Peroxysomes	81.40%	CBD4/SP
H-SERP002	scaffold_6:2283673-2285099	A	367	Y	Lysosomes	75.10%	CB4/SP
H-SERP003	scaffold_14:1091567-1093468	F.4C	370	Y	Plasma membrane	87.60%	SP/LC
H-SERP004	scaffold_166:40679-45446	F.4	907	Y	Extracellular/Secretory	93.90%	TM/SEA/FRI/2(LDLa)/SP
H-SERP005	scaffold_18:1019629-1021535	F.4C	476	N	Extracellular/Secretory	81.40%	TM/SP/LC/CBD2
H-SERP006	scaffold_18:1022014-1023630	F.4C	401	Y	Golgi	75.10%	SP/LC
H-SERP007	scaffold_18:1281298-1282713	F.4C	365	Y	Extracellular/Secretory	100.00%	SP/LC/CBD2
$^{\Psi}$H-SERP008	scaffold_1911:1839-2336	F.4A	116	N	Plasma membrane	93.90%	SP
H-SERP009	scaffold_21:967316-968647	F.4A	258	Y	Plasma membrane	81.40%	SP
H-SERP010	scaffold_21:969637-970913	F.4A	278	N	Golgi	75.10%	TM/SP
H-SERP011	scaffold_21:976176-977463	F.4A	265	Y	Lysosomes	93.90%	SP
$^{\Psi}$H-SERP012	scaffold_2471:6550-7260	F.4C	159	N	Golgi	75.10%	SP
H-SERP013	scaffold_25:1139392-1140102	F	149	N	Plasma membrane	75.10%	SP

H-SERP014	scaffold_26:174232-176079	F.4A	417	Y	Plasma membrane	87.60%	SP/SP
H-SERP015	scaffold_28:128279-130687	F	600	Y	Golgi	68.90%	SP
H-SERP016	scaffold_34:527714-529694	F	256	N	Mitochondria	87.60%	TM/SP
H-SERP017	scaffold_36:548669-56579	F.4A	400	Y	Golgi	75.10%	LC/SP
H-SERP018	scaffold_36:662746-664543	F.4A	399	N	Golgi	68.90%	TM/2(LC)/SP
H-SERP019	scaffold_36:62765-64497	F.4A	393	Y	Plasma membrane	81.40%	SP
H-SERP020	scaffold_36:415789-417463	F.4A	379	Y	Endoplasmic Reticulum	68.90%	2(LC)/SP
H-SERP021	scaffold_4680:594-2226	F.4A	364	Y	Plasma membrane	75.10%	LC/SP
H-SERP022	scaffold_59:692256-693767	F.4C	150	N	Plasma membrane	81.40%	SP
H-SERP023	scaffold_90:82572-84188	F.4	222	N	Peroxysomes	75.10%	SP
H-SERP024	scaffold_91:113735-115012	F.4A	318	Y	Peroxysomes	81.40%	SP/LC
ΨH-SERP025	scaffold_99:63389-67318	F.4C	490	N	cytoplasm	75.10%	SP/LC
H-SERP026	scaffold_10:278365-282303	F.4	1015	Y	Extracellular/Secretory	100.00%	7(LC)/RPT/LC/RPT/LC/RPT/LC/SP/LC
H-SERP027	scaffold_6:2303719-2304218		123	N	cytoplasm	93.90%	SP
H-SERP028	scaffold_120:305280-307111	F	269	Y	Plasma membrane	81.40%	SP

H-SERP029	scaffold_132:189622-190695	F.4A	252	Y	Plasma membrane	93.90%	SP
H-SERP030	scaffold_13:1593414-1601879	F.4	1467	Y	Plasma membrane	81.40%	9(LC)/2(RPT)/2(LC)/SP
H-SERP031	scaffold_178:151103-152218	F.4A	256	Y	Plasma membrane	93.90%	SP
H-SERP032	scaffold_178:147398-148552	F.4A	262	Y	Plasma membrane	87.60%	SP
H-SERP033	scaffold_18:1025212-1027599	F.4C	395	Y	Lysosomes	87.60%	SP/LC/CBD2
H-SERP034	scaffold_18:1135695-1139304	F.4C	769	N	Golgi	81.40%	TM/LC/COIL/LC/KH/SP
H-SERP035	scaffold_18:1270324-1272178	F.4C	453	Y	Plasma membrane	81.40%	SP
H-SERP036	scaffold_23:443758-446647	F.4	752	N	Golgi	81.40%	LC/SP/LC
H-SERP037	scaffold_249:43763-45565	F.3B	400	N	Peroxysomes	75.10%	LC/SP
H-SERP038	scaffold_29:607832-609815	F.3B	373	Y	Lysosomes	100.00%	SP
H-SERP039	scaffold_36:679188-680956	F.3B	418	Y	Plasma membrane	81.40%	2(LC)/SP
H-SERP040	scaffold_36:1014389-1019516	E	200	N	Golgi	93.90%	SP
H-SERP041	scaffold_36:79941-81127	F.3B	274	Y	Mitochondria	68.90%	SP
H-SERP042	scaffold_36:135566-15246	F	348	Y	Plasma membrane	68.90%	SP
$^{\Psi}$H-SERP043	scaffold_549:7348-12232	F.3B	172	N	Plasma membrane	87.60%	SP

H-SERP044	scaffold_6158:64-1416	F.4	317	Y	Lysosomes	87.60%	LC/SP
H-SERP045	scaffold_62:585964-587560	F.4A	338	Y	Plasma membrane	75.10%	SP/4(LC)
H-SERP046	scaffold_72:455597-458084	F	339	N	Peroxysomes	81.40%	ADH_N/TM/SP
H-SERP047	scaffold_4:2861345-2865046	F.4	354	Y	Lysosomes	93.90%	LC/SP
ΨH-SERP048	scaffold_100:21328-23305	F.3B	169	N	Peroxysomes	75.10%	SP
H-SERP049	scaffold_13:424477-432710	F.4	1260	N	Golgi	81.40%	12(LC)/SP
H-SERP050	scaffold_145:109047-110465	F	286	Y	Lysosomes	75.10%	SP
ΨH-SERP051	scaffold_1698:1969-3028	F.4A	261	Y	Plasma membrane	87.60%	SP
H-SERP052	scaffold_17:1342061-1344071	F	440	Y	Mitochondria	75.10%	SP
H-SERP053	scaffold_18:1133138-1135194	F.4C	418	Y	Lysosomes	87.60%	SP/LC/RPT/LC
H-SERP054	scaffold_18:1283623-1285459	F.4C	413	N	Plasma membrane	68.90%	SP/CBD2
H-SERP055	scaffold_18:441263-442416	F	249	Y	Plasma membrane	81.40%	SP
H-SERP056	scaffold_36:48798-51548	F.3B	419	Y	Plasma membrane	87.60%	LC/SP
H-SERP057	scaffold_36:1024048-1025194	E	247	N	Peroxysomes	68.90%	SP
H-SERP058	scaffold_36:748455-76541	F.3B	386	Y	Lysosomes	81.40%	2(LC)/SP

H-SERP059	scaffold_36:83614-85636	F.3B	525	Y	Plasma membrane	75.10%	SP
H-SERP060	scaffold_36:62094848-622610	F.3B	374	Y	Golgi	68.90%	SP
H-SERP061	scaffold_49:714173-715476	F.4C	304	Y	Plasma membrane	87.60%	SP/LC
H-SERP062	scaffold_62:548986-550271	F.4A	294	Y	Plasma membrane	87.60%	SP/LC
H-SERP063	scaffold_65:213377-215310	F.4	449	Y	Golgi	68.90%	2(LC)/SP
H-SERP064	scaffold_91:124189-125727	F.4A	332	N	Peroxysomes	81.40%	TM/SP
ΨH-SERP065	scaffold_98:232068-233510	F	120	N	Lysosomes	75.10%	SP
H-SERP066	scaffold_16:1543607-1544903	F	173	N	Mitochondria	87.60%	SP
H-SERP067	scaffold_21:972059-973315	F.4A	265	Y	Plasma membrane	81.40%	SP
H-SERP068	scaffold_36:32214-33141	F.3B	237	N	Plasma membrane	87.60%	LC/SP
H-SERP069	scaffold_36:38487-40231	F.3B	360	N	Peroxysomes	93.90%	LC/SP
H-SERP070	scaffold_36:42488-44031	F.3B	364	N	Plasma membrane	75.10%	LC/SP
H-SERP071	scaffold_36:998300-999702	E	342	Y	Mitochondria	93.90%	SP
H-SERP072	scaffold_36:1021896-1023167	E	260	Y	Peroxysomes	81.40%	SP
H-SERP073	scaffold_87:330777-332210	F	362	N	Golgi	81.40%	SP

ΨH-SERP074	scaffold_99:48146-49140	F.4C	207	N	Mitochondria	93.90%	SP
H-SERP075	scaffold_36:957804-959588	F.3B	424	N	Peroxysomes	81.40%	TM/SP
H-SERP076	scaffold_6:2144266-2145723	D	328	Y	Lysosomes	93.90%	LC/SP
H-SERP077	scaffold_18:1008178-1009478	F.4C	365	N	Peroxysomes	62.60%	SP/LC/CBD2
H-SERP078	scaffold_29:644864-646833	F.3B	274	Y	Lysosomes	81.40%	SP
H-SERP079	scaffold_42:907067-907918	F.4B	135	N	Lysosomes	100.00%	SP
H-SERP080	scaffold_85:89648-90551	F.4B	125	N	Plasma membrane	87.60%	LC/TRY
ΨH-SERP081	scaffold_762:7605-8441	F	182	N	Peroxysomes	75.10%	SP
H-SERP082	scaffold_17:1357116-1359392	F	526	Y	Mitochondria	75.10%	SP
H-SERP083	scaffold_18:1286578-1288289	F.4C	440	Y	Golgi	68.90%	SP/LC/CBD2
H-SERP084	scaffold_18:1289072-1290798	F.4C	423	Y	Plasma membrane	81.40%	SP/LC/CBD2
H-SERP085	scaffold_91:118370-119733	F.4A	284	Y	Golgi	75.10%	SP
H-SERP086	scaffold_6:2156065-2157557	D	333	Y	Lysosomes	93.90%	SP
H-SERP087	scaffold_6:2165080-2166566	D	312	Y	Lysosomes	87.60%	2(LC)/SP
H-SERP088	scaffold_6:2290419-2291974		347	Y	Golgi	75.10%	LC/SP

H-SERP089	scaffold_6:2299797-2301556		428	Y	Golgi	81.40%	LC/CBD4/LC/SP
H-SERP090	scaffold_6:2316878-2317960	F	247	N	Golgi	75.10%	SP
H-SERP091	scaffold_4:2868260-2870542	F.4	211	N	Plasma membrane	81.40%	SP
H-SERP092	scaffold_65:395468-397135	F.4	466	Y	Extracellular/Secretory	87.60%	SP
H-SERP093	scaffold_178:145723-146956	F.4A	254	N	Peroxysomes	81.40%	SP
H-SERP094	scaffold_29:657908-659160	F.3B	256	N	Plasma membrane	81.40%	SP
H-SERP095	scaffold_91:116028-117736	F.4A	306	Y	Plasma membrane	93.90%	SP/LC
H-SERP096	scaffold_40:64742-68164	F	694	N	N/A	N/A	SP
H-SERP099	scaffold_72:446418-448164	F	377	N	Peroxysomes	87.60%	SP
H-SERP101	scaffold_98:203152-208679	F	216	Y	Endoplasmic Reticulum	87.60%	SP
ΨH-SERP102	scaffold_178:144327-145385	F.3B	261	Y	Plasma membrane	87.60%	SP
H-SERP103	scaffold_36:346492-347371	F	186	N	Peroxysomes	87.60%	SP
H-SERP104	scaffold_52:620327-624121	F.4C	526	Y	Golgi	81.40%	4(LC)/SP
H-SERP105	scaffold_18:1010333-1011073	F.4C	170	N	Plasma membrane	93.90%	SP
H-SERP106	scaffold_18:1132154-1132657	F.4C	167	Y	Golgi	75.10%	LC/TRY

H-SERP107	scaffold_36:954343-955902	F.3B	368	Y	Plasma membrane	75.10%	LC/SP
H-SERP108	scaffold_190:73552-76305	F.4C	369	N	cytoplasm	87.60%	SP

Table 1.4. Characteristics of each Serine Protease domain in the *Daphnia pulex* genome. Below are fixed[a] gene names and their domain positions (see appendix). Predicted location of the cleavage sites for activation of the zymogen and conserved motifs of the catalytic triad are catalogued. Fields left blank indicate that the domain has either the full TAAHC, DIAL, or GDSGGP motif. Motifs in **bold** indicate that the putative residue for the catalytic triad is either substituted or missing. Motif predictions were made using the database from smart.embl-heidelberg.de and http://prosite.expasy.org/ as well as multiple alignments in MEGA 5.10 (Onting 1998; de Castro et al 2006; Tamura et al 2011). Predicted substrate specificity using the multiple alignment algorithm in MEGA 5.10 (Perona and Craik 1995; Tamura et al 2011) are also catalogued for each Serine Protease.

Gene Name	Position	Activation Site	TAAHC	DIAL	GDSGGP	Substrate Specificity
SERP15[a]	66-294	R^IVGG		DVAL		?(SGA)
CHY1B	68-296	R^IVGG				C(GGG)
SERP16[a]	65-293	R^IVGG		DVAL		?(SGA)
CHY1D	66-294	R^IVGG		DVAL		C(GGG)
CHY1E	68-297	R^IVGG		DLAL		C(SGG)
CHY1F	68-294	R^IVGG		DLAL		C(SGG)
CHY1G	68-294	R^IVGG		DVAL		C(SGG)
CHY1H	76-301	R^IVGG		DVAL		C(GGG)
CHY2	32-258	R^IVGG		DVAL		C(GGG)
CHY3	67-294	R^IVGG				C(GGG)
CHY4	96-329	R^IVGG		DIGL		C(GGG)
CHY5A	14-259	T^IIGG		DIAI		C(GGG)
CHY5B	32-285	Y^IIGG		DIAI		C(GGG)
CHY6	110-345	N^IMEG				C(GGG)
CHY7A	2-187	V^GSDV		DIAI		C(GGG)
CHY7B	60-300	R^MVGS		DIAI		C(GGG)
ELA1	37-280	E^IIGG				E(SVA)
SERP1	114-363	R^IING				?(GGD)

SERP10	99-339	G^IVGG		DVAI		?(GGS)
SERP11	418-653	R^IVGG		DIAI		?(DGS)
SERP12	63-311	R^MINS		DIAI	GDSGSP	?(HGD)
SERP13	61-303	R^RMTD				?(DG-)
SERP14	145-370	R^MVGS		DIAV		?(DG-)
SERP2	108-346	K^IVGG			SDSGGP	?(GGS)
SERP3	73-303	R^IING		DVAL	GDSGSA	?(GVD)
SERP4	57-292	K^IVGG		DMAL		?(GGN)
SERP5	62-289	S^IYGG		DIAI		?(NGS)
$^{\Psi}$SERP6	49-282	R^IVGG		DLGV		?(SAA)
SERP7	142-366	R^IVGG		DIAV		?(DS-)
SERP8	131-375	G^IAGG		DVAL		?(DGS)
SERP9	136-377	S^IVGG		DIAV		?(DGS)
TRY1	32-270	R^IVNG		DLAL		T(DGG)
TRY10A	38-257	K^IVNG				T(DGG)
TRY10B	36-275	R^IVNG				T(DGG)
TRY10C	37-272	K^IVNG				T(DGG)
TRY11	6-245	R^IVGG		DLAI		T(DGG)
TRY12	86-315	R^IVGG				T(DGG)
TRY13	2-241	G^GAST		DLAI		T(DGG)
TRY14	39-290	R^IVGG		DIAI		T(DGG)
TRY15A	197-424	R^IVGG	TACHC			T(DGG)
TRY15B	107-334	R^IVGG	TASHC			T(DGG)
TRY16	6-245	R^DEGK		DIAV		T(DGG)
TRY17	195-431	R^IAGG		DIAI	GDSGAP	T(DGG)
TRY18	179-419	R^IVGG		DIAI		T(DGG)
TRY19	68-306	R^VVGG		DLAL		T(DGG)
TRY2	41-269	R^IIGG		DIGI	GDSGGQ	T(DGG)
TRY20	289-511	R^IVGG	TAGHC	DITL		T(DGG)
TRY21	1139-1373	R^IVGG	SAGHC	DISI		T(DGG)

TRY22	339-561	R^IIGG		DVAL		T(DGG)
TRY23	193-427	R^VVGG				T(DGG)
TRY24	176-457	R^IMGG				T(DGG)
TRY25	36-456	K^IVNG		DVGM		T(DGG)
TRY26	46-350	K^IVNG		DIAM		T(DGG)
TRY27	38-334	S^IVGG		DVAM		T(DGG)
TRY28	31-299	S^VVGG	TAGHC	DLGL		T(DGG)
TRY29	33-338	S^IVGG		DVAL		T(DGG)
TRY3	66-294	R^IVGG		DVAL		T(DGG)
TRY30	33-391	S^VVGG		DVAV	GDSGGA	T(DGG)
TRY31	7-197	Q^VFGL	NAAHC	DIAV		T(DGG)
TRY32A	29-237	E^NVGG			GDSGDP	T(DGG)
TRY32B	54-277	Q^IVGG				T(DGG)
ΨTRY32C	32-245	H^IVGG		DLAL		T(DGG)
TRY32D	49-272	H^IVGG		DLAL		T(DGG)
TRY33A	132-365	A^IVGG		DVAV		T(DGG)
TRY33B	129-365	G^IVGG		DVAV		T(DGG)
TRY34	133-374	G^IVGG		DVAV		T(DGG)
TRY35	121-355	G^IVGG				T(DGG)
TRY36	218-452	R^IVGG		DIAI		T(DGG)
TRY37	195-420	R^IVGG				T(DGG)
TRY38	3-207	G^RFFC		DIAI		T(DGG)
TRY39	30-264	R^IVGG		DIAI		T(DGG)
TRY40	2-204	S^PTHU		DIAI		T(DGG)
TRY41	255-476	W^LVAI		DIAI		T(DGG)
TRY42A	153-379	A^VDIN				T(DGG)
TRY42B	153-381	R^IVGG				T(DGG)
TRY43	49-290	R^MVGG		DIAI	NDSGGP	T(DGG)
ΨTRY44A	207-446	Y^MVAS		DIAI	YDSGGP	T(DGG)

TRY44B	157-396	Y^MVAS		DIAI	YDSGGP	T(DGG)
TRY44C	144-386	Y^NVES		DIAI	NDSGGP	T(DGG)
TRY45	2-240	V^ASKE		DMAI		T(DGG)
TRY46	148-391	R^MVES			NDSGGP	T(DGG)
TRY47A	170-410	R^MVGS		DLAI	NDSGGP	T(DGG)
TRY47B	138-378	R^MVGS		DLAI	NDSGGP	T(DGG)
TRY47C	139-379	R^MVGS		DLAV	HDSGGP	T(DGG)
TRY48	29-261	H^IVGG		DIAI		T(DGG)
TRY4A	39-273	K^IVGG	DASHC			T(DGG)
TRY4B	41-267	R^IVGG		DICL		T(DGG)
TRY5B	43-282	K^IVGG	DAAHC			T(DGG)
TRY5C	44-286	K^IVGG	DAAHC			T(DGG)
TRY5D	42-285	K^IVGG				T(DGG)
SERP17[a]	42-287	K^IVGG	NAAHC	DISL	LISGGP	?(FGG)
TRY5F	42-283	K^IVGG		DISL		T(DGG)
TRY5G	42-282	K^IVGG	DAAHC	DISL		T(DGG)
SERP18[a]	44-285	R^ILSG	HAAHG	DISL		?(PGG)
TRY5I	41-281	K^IIGG				T(DGG)
TRY5J	44-281	K^IVGG		DISL		T(DGG)
TRY5K	43-285	K^IVGG	NAAHC			T(DGG)
TRY5L	42-283	R^IVGG	DAAHC	DICL		T(DGG)
TRY5M	47-293	K^IVGG				T(DGG)
TRY6	1215-1455	K^VVKG			GDSGGG	T(DGG)
TRY7	254-497	R^VVNG				T(DGG)
TRY8A	453-685	K^IVSG				T(DGG)
TRY8B	524-763	R^IVGG	TAGHC	DLAL		T(DGG)
TRY8C	26-249	R^IVGG		DLAL		T(DGG)
TRY9A	73-311	R^IIGG		DVAV		T(DGG)
TRY9B	242-476	R^IVGG	TAGHC	DVAV		T(DGG)
H-SERP001	158-380	R^LAKL	**TAAYC**	DIAI	**DYGGP**	?(GGS)

H-SERP002	141-365	K^SSKE	**TSARC**	DIAV	**EKGGP**	?(IGS)
H-SERP003	63-392	K^SSVE	**TTASC**	DIAL	**N/A**	?(VVG)
H-SERP004	690-901	P^SAHG	**TASSC**	**QLVL**	**EFAGSP**	?(DNR)
H-SERP005	85-336	K^IVGG	**TAAAC**	**NIAL**	**GDNGGP**	?(GSS)
H-SERP006	41-297	G^RPNL		DIAV	**GDDGGP**	?(RNN)
H-SERP007	32-274	R^IVGG			**GDDGGP**	?(HGN)
ΨH-SERP008	1-111	MWATV	**N/A**	**N/A**		T(DGG)
H-SERP009	24-248		**TAASC**	DIAM	**YDEGSP**	?(SNT)
H-SERP010	38-268	R^IIGG	**TAAEC**	**NIAL**	**YDEGSP**	?(SNT)
H-SERP011	24-252	R^LVGG	**TAASC**		**GDEGDP**	?(ANS)
ΨH-SERP012	1-151	M^HPKW	**N/A**	DVAL		C(GGG)
H-SERP013	3-143	V^SEHD	**N/A**	**N/A**		T(DGG)
H-SERP014	24-248	R^IIGG	**TTAAC**	DIAM	**GDAGTP**	?(DTT)
H-SERP015	55-587	H^IIVI		DVAV	**DDEGGP**	?(SFA)
H-SERP016	48-217	S^IVGG	VAAHC	DVAL	**N/A**	?(P-A)
H-SERP017	148-395	K^ILGS	**LAATC**	DIAI	**EDVGGP**	?(FIS)
H-SERP018	167-394	R^ISGG	**LAAQC**	DIAI	**GDVGGP**	?(YTG)
H-SERP019	150-387	R^VAGS	**LAANC**	DIAI	**DDVGGP**	?(FTS)
H-SERP020	156-374	R^ITTG	**TAAQC**	DIAI	**GDVGGP**	?(FTG)
H-SERP021	124-357	R^IAGG	**LAAQC**	DIAI	**IDVGGP**	?(FTG)
H-SERP022	2-118	Q^DRHE	**N/A**	**N/A**	**SDNGGP**	?(GFS)
H-SERP023	15-214	A^IAGS	**N/A**	DLAL		T(DGG)
H-SERP024	24-253	R^IVGG	**TTASC**		**YDEGSP**	?(ANS)
ΨH-SERP025	140-311	R^RSGI	**TSARC**		**GDNGGP**	?(-VG)
H-SERP026	744-977	L^GTGE	TSFHC	DTAV	**VDGGSP**	?(NSG)
H-SERP027	1-115	M^DVNF	**N/A**	**N/A**		?(GTS)
H-SERP028	35-264	E^SLVG	**TGADC**	DIAV	**DDSNGGP**	?(IGG)
H-SERP029	30-238	R^ILGG	**TTVTC**	**NLAV**	**YDEGSP**	?(SNT)
H-SERP030	1221-1461	P^YADG		DIAI	**GDGGGP**	T(DGG)

H-SERP031	24-248	R^IYNG	**TVASC**	DIAM	**HFDGSP**	?(SNT)
H-SERP032	24-248	R^MTNG	**TSASC**	DIAM	**YDEGSP**	?(SNT)
H-SERP033	43-303	R^IAGG			**GDDGGP**	?(RGN)
H-SERP034	497-735	R^IVGG	**TAARC**	DVAL		?(GGN)
H-SERP035	196-439	R^ILGG	**TAMSC**	DWAI	**GEKGGP**	?(RDN)
H-SERP036	483-696	R^VLSG	TVAHC		**GDGGA**	?(DGS)
H-SERP037	170-395	R^LAGG	**LAAQC**	DIAI	**GDIGGP**	?(FTG)
H-SERP038	130-367	G^VIGG		DVAV	**KDGGP**	?(DAA)
H-SERP039	189-419	A^VVKA	**LAAQC**	DIAI	**GDIGGP**	?(FTG)
H-SERP040	22-198	F^IWNT	N/A	**TEHI**	GDSGGT	?(GN-)
H-SERP041	146-273	K^IAGG	**TAAQC**	DIAM	**N/A**	?(--G)
H-SERP042	37-336	R^IIGG		DIAI	**DDEGGP**	?(SFA)
$^{\Psi}$H-SERP043	2-170	N^GVTL	**N/A**	DIAI	**IDVGGP**	?(YTK)
H-SERP044	72-308	R^VVGG			**GDGGGP**	T(DGG)
H-SERP045	24-240	R^IAGG	**TAASC**		**YDEGSP**	?(SNS)
H-SERP046	174-334	S^IMGG	TAKHC	**N/A**	**SDYGQR**	T(DGG)
H-SERP047	104-342	R^SDGL			**GDGGGP**	T(DGG)
$^{\Psi}$H-SERP048	1-155	M^TRRI	**N/A**	DIAI	**IDVGGP**	?(YT-)
H-SERP049	1015-1254	H^SDGE	TVAHC	DIAV	**GDGGGP**	T(DGG)
H-SERP050	30-274	S^IIGG		DIAI	**GDDGGP**	?(ASS)
$^{\Psi}$H-SERP051	30-274	R^ILGG	**TTVTC**	**NLAV**	**YDEGSP**	?(SNT)
H-SERP052	32-432	F^NVSG	**TAAQC**	DVAL	**TDIGGP**	?(NGG)
H-SERP053	77-314	K^IIGG	**TSASC**		VDEGNP	?(EY-)
H-SERP054	67-310	R^IVGG		DMAL	**GDDGGP**	?(RGN)
H-SERP055	28-245	S^IIGG	**TTAWC**	DIAV	**LDAGSP**	?(GAV)
H-SERP056	159-390	R^IAGS	**LAASC**	DIAI	**EDVGGP**	?(IID)
H-SERP057	2-243	R^MVGS	**N/A**	DIAV		C(GGG)
H-SERP058	140-380	R^VSGG	**LAASC**	DIAI	**GDVGGP**	?(FTS)
H-SERP059			**TAAQC**	DIAI	**GDVGGP**	?(FT-)

H-SERP060	136-367	R^ISGG	**LAAQC**	DIAI	**TDVGGP**	?(YTG)
H-SERP061	41-242	R^VVGG			**GDDGGP**	?(-GN)
H-SERP062	26-249	R^VMGG	**TAASC**		**YDEGSP**	?(SNS)
H-SERP063	209-444	T^QAKF	TAAHK	DIAV	**GDGGSP**	T(DGG)
H-SERP064	71-320	R^IIGG	**TTASC**		**LDEGSP**	?(TSS)
ΨH-SERP065	1-111	MATIR	**N/A**	**N/A**	DISGGP	?(HGG)
H-SERP066	2-164	SLRQS	**N/A**	DIAI	GDSGGP	T(DGG)
H-SERP067	25-255	R^IMGG	**TTAKC**		**YDEGGP**	?(SNT)
H-SERP068	76-237	R^IFGP	**LAATC**	DIAI	**N/A**	?(---)
H-SERP069	108-355	R^ILGS	**LAATC**	DIAI	**EDVGGP**	?(YIG)
H-SERP070	129-358	R^VAGV	**LAASC**	DVAV	**GDIGGP**	?(FTS)
H-SERP071	54-328			DIAI		T(DGG)
H-SERP072	66-260	R^MVDG	**TSANC**	**N/A**	**GDAGGP**	?(HGA)
H-SERP073	58-357	N^IVGG		DIAI		?(HSS)
ΨH-SERP074	13-189	R^RSGI	**TSARC**		**GDNGGP**	?(QVG)
H-SERP075	179-417	R^IAGG	**LAAQC**	DIAI	**VDVGGP**	?(FTG)
H-SERP076	91-323	T^SIGM	**TFDSC**	DILI	**TYRGGP**	?(D-N)
H-SERP077	12-247	R^IVGG	TTTHC	DIAV	**ADDGGP**	?(QGN)
H-SERP078	43-268	T^IVGL	**TAASC**	DIAI	**NDMGGP**	?(DSA)
H-SERP079	1-130	M^QLST	**N/A**	**N/A**	**RDLGGP**	?(DGD)
H-SERP080	51-123	T^SLLF	**N/A**	**N/A**	**N/A**	?(-MT)
ΨH-SERP081	5-176	T^ILVG	**N/A**	DVAV	**GDIGGP**	?(FT-)
H-SERP082	27-518	R^IVMM	**TAAQC**	DVAL	**IDIGGP**	?(TGG)
H-SERP083	62-304	R^IVAG			**GDDGGP**	?(RGN)
H-SERP084	67-309	R^IVGG		DISL	**GDDGGP**	?(RGN)
H-SERP085	32-273	R^ITGG	**TAASC**	DVAL	**YDEGSP**	?(STS)
H-SERP086	91-328	S^SVEI	TESHC	**ELAI**	**N/A**	?(DYG)
H-SERP087	88-307	S^AVEM	**TVSGC**	DVLI	**YYSGGP**	?(D-G)
H-SERP088	96-340	E^AGLN		DIAI	**N/A**	?(-AD)

H-SERP089	184-420	D^ETGR		DIAI	**YNDGGP**	?(EAS)
H-SERP090	29-228	F^YFGC	**N/A**			E(SVA)
H-SERP091	4-199	L^IINL	**N/A**		**GDGGGP**	T(DGG)
H-SERP092	191-443	R^ITNF		DVAL	**GDGGSP**	T(DGG)
H-SERP093	17-236	A^IYQG	**TAASC**	DLAL	**FDQGSP**	?(SNT)
H-SERP094	48-253	R^IIGG	**LAAQC**	DIAI	**YDMGGP**	?(FVK)
H-SERP095	26-254	R^ITGG	**TTASC**	DVAL	**YDEGSP**	?(ANS)
ΨH-SERP096	1-141	M^TLKD	**N/A**	**N/A**	**N/A**	?(FT-)
H-SERP099	59-368	S^IAGG	**TAAYC**	DLAI	YDSGGP	?(N-G)
H-SERP101	26-208	K^FVGN	**PAASG**	DIAI	DISGGP	?(SNT)
ΨH-SERP102	30-247	R^ILGG	**TTVTC**	**NLAV**	**YDEGSP**	?(YNG)
H-SERP103	2-197	S^LFEF	**N/A**	DIAI	**SDKGGP**	?(D--)
H-SERP104	321-525	R^IVGG		DVAI	**N/A**	?(G--)
H-SERP105	1-170	MTSGA	**TSANC**	DIAM	**N/A**	?(G-G)
H-SERP106	117-166	IVGG		**N/A**	**N/A**	?(GA-)
H-SERP107	130-361	R^IAGG	**LAAQC**	DIAI	**DDIGGP**	?(GVE)
H-SERP108	8-203	R^IVGG	**TAARC**		**N/A**	T(DGG)

CHAPTER 2

EVOLUTIONARY DIVERSIFICATION OF SERINE PROTEASES IN THE CRUSTACEAN *DAPHNIA MAGNA*

2.1 INTRODUCTION

The serine protease gene family, or the SP family, assists in multiple functional roles including digestion, embryonic development, and innate immunity (Rawlings and Barrett 1993). Observations in *Daphnia magna,* a fresh water crustacean, showed the SP family to make up 75-83% of the catalytic activity in the gut (Elert et al 2003). The SP family involvement in innate immunity and embryonic development has been extensively studied in *Drosophila*. For example, serine proteases (SPs) are observed in the antimicrobial peptide producing Toll pathway in *Drosophila* (Jang et al 2008) as well as in the pathway for dorso-ventral polarization during embryonic development of *Drosophila* (Hong and Hashimoto 1996; Lemosy et al 1998). Few genes of this gene family have expanded across taxa and has been proposed that this gene family evolved from two ancestral proteases to obtain the analogous features of the active site putative for peptide chain hydrolysis (Brenner 1988).

All peptidases of the serine protease (SP) gene family have a Ser-195 residue putative for catalysis, and shows strong sequence similarity to the Bovine chymotrypsin-A (Hartley 1964), which was one of the first serine proteases to be studied. The SP family is characterized as only containing serine endopeptidases, which encompasses all subfamilies of the SP family: trypsins, chymotrypsins, and elastases. The SP domain

structure within the SP family starts with a cleavage site at the start of the domain and lies downstream of a signal peptide (Ross et al 2003). This cleavage site, R^IVGG, is crucial in turning an inactive enzyme, a zymogen, into its catalytically-active primary structure (Hedstrom et al 1996). The active serine protease has three amino acid residues that hydrolyze peptide bonds of a peptide chain targeted for degradation. The motifs containing the catalytic residues are well conserved across observed taxa, and they are TAAHC, DIAL, and GDSGGP (Greer 1990). The histidine (His-57) in the TAAHC motif is the catalytic residue that attracts a proton from the serine hydroxyl side chain to allow for nucleophilic attack on the protein substrate in the catalytic cleft. The aspartate (Asp-102) in the DIAL motif is critical for stabilizing the protonated histidine in the TAAHC motif. The serine residue (Ser-195) in the GDSGGP motif then hydrolyzes the scissile peptide bond of the substrate by an acylation-deacylation mechanism (Kraut 1977). Three additional residues surrounding the GDSGGP motif discriminate the substrate specificity of the serine protease and they are as follows: Asp-189, Gly-216, and Gly-226 in Trypsin-like SPs; Gly-189, Gly-216, and Gly-226 in Chymotrypsin-like SPs; Ser-189, Val-216, and Ala-226 in Elastase-like SPs (Perona and Craik 1995). Three to four disulfide bridges are also found on the domain and play a role in the structural integrity of the protease (Greer 1990).

In the analysis of the SP family in *Anopheles gambai* genome, a relationship between adaptation to blood meal and recent duplicates of the gene family were observed (Wu et al 2009). In *Drosophila*, the gene family was extracted from the genomes of 12 species and compared to food preference (Li et al 2012). Both dipteran studies reveal positive selection within their SP gene families, suggesting a relationship between novel

genes and adaptation to meal preference. We hypothesize purifying selection to be acting on recent duplicates of SPs in order for the gene family to maintain the function of ancestral proteases that have expanded across species of *Daphnia* who feed on only phytoplankton.

The zooplankton *Daphnia* are a micro-crustacean that act as a keystone species in freshwater ecosystems (Sarnelle 2005). Experiments focusing on resource exploitation have shown Daphnia-phytoplankton interactions affect life history traits and cause differential gene expression across the *Daphnia* genome (Tessier, Leibold, & Tsao, 2000, Gliwicz & Boavida, 1993, Dudycha, Brandon, & Deitz, 2012). This preliminary study will introduce the relationship between elevated rate of gene duplicates in a specific gene family, one of ecological importance, and resource exploitation in *Daphnia pulex* and *Daphnia magna.* This study focuses on finding all genes and their homologs of the SP family in *Daphnia magna* and compares them to the SP family in *Daphnia pulex* to assess how the gene family has evolved before and after divergence of the two *Daphnia* species.

In this study, we sought to understand evolutionary patterns within the SP family. To do so, all peptidases of the SP family are identified within the *Daphnia magna* genome and compared to the *Daphnia pulex* genome. A phylogenetic analysis of the SP family will aid in investigating possible monophyletic patterns and the functional history of each SP in freshwater crustacea. This study will identify orthologs likely to be ecologically significant for digestive function and analysis of selection within the SP family will serve as an initial platform in understanding the relationship between molecular evolution and resource exploitation in arthropods.

2.2 METHODS AND MATERIALS

Database searching, sequence retrieval and annotation of active SPs and SP homologs

Standalone Blast-2.2.27+ from NCBI was downloaded to a Microsoft Windows operating system. The program allowed their users to conduct various algorithms of BLAST using the command prompt. The *Daphnia magna*, specifically the 2012 version of the genome, was downloaded from wfleabase.org. The version trall7set9rbest dataset included the translated amino acid sequence and the transcript sequences of the genome.

All 211 SPs and H-SPs from the *Daphnia pulex* genome was the query for a stand-alone Blast against the *Daphnia magna* translated sequence data. The stand-alone blast retrieved an output that was then surveyed for conserved motifs, amino-acid sequence patterns, found only in serine protease domains, families, and functional sites at Prosite (http://prosite.expasy.org/) (Sigrist et al 2002).

Genes retrieved from the Stand-alone Blast search with an E-value < 0.0005 were discarded from this study. ScanProsite (http://prosite.expasy.org/scanprosite/) (de Castro et al 2006) surveyed each gene from the output with E-value < 0.0005 to ensure the presence of all conserved structural components of an SP. The following are the conserved structural elements: 1.) The presence of the residues His-57, Asp-102, and Ser-195 of the catalytic triad; 2.) Contains either three or four cysteine-cysteine disulfide bridges that control the conformation of the resulting protein structure; 3.) The presence of an activation site that indicates the cleavage site of the SP domain (Greer 1990; Perona and Craik 1995). If at least one of these structural elements were missing, the gene was catalogued as a homolog (H-SP). The SPs containing all three structural elements were used as a query for follow-up stand-alone BLASTPs (States and Gish 1994) against the

Daphnia magna genome. The output with an E-value < 0.0005 was again surveyed by ScanProsite and catalogued as either SP or H-SP. This procedure repeated until no more novel SP or H-SPs were found in the output from the *D. magna* genome.

Searching for Sequence Properties of Serine Protease Gene Family in D. magna.

As mentioned before, genes containing all three amino-acid residues of the catalytic triad were catalogued as SPs. If at least one amino-acid residue was missing from the triad, the gene was catalogued as an H-SP. ScanProsite also identified the three putative motifs that contain the three amino-acid residues of the catalytic triad (ie.. TAAHC, DIAL, and GDSGGP) (de Castro et al 2006). The amino acid sequences of the three putative motifs are essential in the formation of the catalytic cleft in all SPs and H-SPs. To measure the probability of the presence of specific residues, and their biochemical composition, the amino-acid sequences of the SP and H-SP domains were extracted from each catalogued gene. A motif search for the SP domains used the Multiple EM for Motif Elicitation (MEME; version 4.9.0 http:// http://meme.nbcr.net)(Bailey et al 2006). The parameters were adjusted to retrieve at most 10 motifs ranging from 6 (minimum width) to 10 (maximum width) amino acids.

Along with identifying the three putative motifs containing the residues of the catalytic triad for hydrolysis, there are additional specific residues at location 189, 216, and 226 that determine the substrate-binding pocket of the enzyme. A Muscle Multiple alignment of all active-SPs and H-SPs against TRY4B, a gene already annotated and observed to contain residues involved in substrate specificity, aided in determining the position and presence of the residues involved in substrate specificity (Schwerin et al 2009). SP and H-SPs were catalogued based off of the following substrate specific

residues: Asp-189, Gly-216, and Gly-226 in Trypsin-like SPs; Gly-189, Gly-216, and Gly-226 in Chymotrypsin-like SPs; Ser-189, Val-216, and Ala-226 in Elastase-like SPs (Perona and Craik 1995). If the residues at the substrate specificity locations did not identify known specificity, the gene was catalogued as Serine Protease-like (SERP) or Serine Protease-like homolog (SERP-H).

The amino acid sequence for each SP and H-SP was scanned using SMART (Onting 1998) for presence of a signal peptide and the presence of additional functional domains. These characteristics were catalogued. For each SP domain on SPs and H-SPs, the conserved amino acid sequence of the cleavage site (i.e., R^IVGG) was catalogued as well as the domain length.

Sequence alignments and Phylogenetic analysis

The translated amino acid sequence of each SP and H-SP domain was isolated for multiple sequence alignment and then phylogenetic analysis. A Muscle multiple alignment algorithm aligned all SP and H-SP domains with -2.9 open gap penalty using MEGA6 (Tamura et al 2011). The alignment output was manually observed to ensure all of the amino acid residues making up the critical structural elements of all SP domains were aligned. The structural elements include: 1.) Three or four disulfide bridges; 2.) motifs containing the amino acid residues of the catalytic triad; 3.) amino acid residues for substrate specificity; 4.) conserved amino acid residues of the cleavage site. Phylogenetic analysis of the resulting multiple sequence alignment of the SP and H-SP domains was done in RAxML. RAxML used the Maximum Likelihood method and the GTR (General Time Reversal) nucleotide substitution model with 4 discreet GAMMA

rate categories and estimated proportion of invariable sites (Stamatakis 2006). An additional test ran 1000 bootstrap replicates.

A phylogenetic analysis of all SPs and H-SPs in the *Daphnia magna* genome with the *Daphnia pulex* retrieved predicted orthologs. RAxML used the Maximum Likelihood method and the GTR (General Time Reversal) amino acid substitution model with 4 discreet GAMMA rate categories and estimated proportion of invariable sites (Stamatakis 2006). An additional test ran 1000 bootstrap replicates. Output contained additional functional information about the sequence properties in a select number of SPs and H-SPs.

2.3 RESULTS

Classification of SPs and H-SPs in D. magna

The output from a stand-alone blast of the 211 SPs and H-SPs found in the *Daphnia pulex* genome against the *Daphnia magna* genome was surveyed for sequence properties of an SP. Then, follow up BLASTP of this output against the *Daphnia magna* genome until no more novel SPs or H-SPS were found. All serine protease genes were identified, classified, and cataloged based on of the presence of the following conserved regions: 1.) three or four disulfide bridges; 2.) motifs containing the amino acid residues of the catalytic triad; 3.) the presence of amino acid residues for substrate specificity; 4.) conserved amino acid residues of the cleavage site.

This process yielded 71 SPs and H-SPs from Daphnia magna genome. SP and H-SPs were then characterized by the following substrate specific residues: Asp-189, Gly-216, and Gly-226 in Trypsin-like SPs; Gly-189, Gly-216, and Gly-226 in Chymotrypsin-like SPs; Ser-189, Val-216, and Ala-226 in elastase-like SPs. Observed in the Daphnia

magna genome were 50 trypsin-like serine proteases, 4 chymotrypsins-like serine proteases, and zero elastase-like serine proteases. If the gene had all properties of a serine protease, but lacked substrate specificity, it was catalogued as a SERP, serine-protease like gene. Of the 71 SPs and H-SPs, 9 were catalogued as SERP and 8 were found to be homologs, missing at least one part of the catalytic triad or disulphide bridges.

8 H-SPs are missing at least one of the conserved residues within the catalytic triad or one of the conserved disulphide bridges, classifying it as a homolog due to possible structural restraints and unknown function. Analysis of the H-SPs showed that the loss of at least one disulphide bridge was more common than deletion of a catalytic residue or deletion of a whole motif. Of the 8 homologs, H-SERPs 001, 002, 003, 004, 005, 006, and 007 are missing at least one disulphide bridge for structural stability of the activated enzyme. (Table 2.1).

Motif conservation within the catalytic cleft of active-SPs

The catalytic function of the Serine Protease gene family largely depends on the structure of the active site, a catalytic cleft, within the enzyme. The catalytic cleft contains three conserved residues, His57, Asp102, and Ser195 (Greer 1990). Each residue is embedded in the following conserved motifs unique to serine proteases: TAAHC, DIAL, and GDSGGP. Multiple EM for Motif Elicitation (MEME) of active SP domains compared the frequency of residue substitution in relation to the conservation of the catalytic residues in each motif of a complete SP.

Of the 71 SPs and H-SPs, 83.1% contain the conserved TAAHC motif; the remaining genes contained a variety of substitutions. SAGHC, SAAHC, SASHC, and TASHC were variants that occurred once, whereas NAAHC(2), DAAHC(4), and

TAGHC(2) occurred multiple times. The underlined residues in TAAHC are highly conserved and hydrophobic (Table 2.1). Additional residues around the TAAHC motif, ILTAAHCV undergo substitution, but the hydrophobic properties are still highly conserved to insure conservation of the structure of the catalytic cleft (Figure 2.1).

Of the 71 SPs and H-SPs, 29.5% contain the conserved DIAL motif. Observed substitutions occurring once in the genome are observed among, DIGL, DISI, DIVL, DLAI, DLGV, and DMAL. DIAI (18), DIAV (2), DISL, (5), DLAL (5), DVAI (5), DVAL (6), and DVAV (3) were other substitutions observed in all SPs and H-SPs. The underlined residues in the motif DIAL undergo substitution, but the hydrophobicity of the motif remains conserved in all active-SPs. As stated before, hydrophobic residues in the DIAL motif ensure conservation of the structure of the active site in the enzyme.

Of the 71 SPs and H-SPs 90.1% contain the conserved GDSGGP motif. Observed substitutions occurring once in this motif, but conserving the catalytic residue Ser195, are GDSGGG, GDSGSA, GVSGGP, NDSGGP, and NESGGP. The underlined residues in GDSGGP are observed to be as highly conserved as the catalytic Ser195. These two residues may be important in conserving the structural stability of the serine in the catalytic cleft (Figure 2.1).

Analysis of Active-SPs and H-SPs with single SP domains

We began our study with a particular interest in single domain serine proteases likely to function in food digestion. Digestive SPs are expected to contain only the serine protease domain with a signal peptide and to have a total length ~300 amino acid residues (Ross et al 2003).

We identified 19 SPs out of the total 71 SPs and H-SPs that match these characteristics, including 10 trypsin-like SPs (TRYs 14, 16, 18, 19, 20, 24, 28, 32, 34, and39), zero chymotrypsin-like SPs, zero elastase-like SP, 5 SERPs (SERPs 02, 03, 04, 07, and 08) and 4 H-SERPs (H-SERPs 004, 005, 006, and 008).

Of the listed digestive serine proteases that contained the signal peptide and the SP domain without any additional functional domains, 2 were longer than 300 residues: H-SERP004 and TRY24. These may be "long" digestive serine proteases (Table 2.1).

Phylogenetic analysis of SPs and H-SPs in both D. magna and D. pulex

D. manga and *D. pulex* were compared to one another using Maximum Likelihood estimation model with general time reversal model of amino acid substations. RAxML was used for phylogenetic construction of the SP gene family in *Daphnia magna and Daphnia pulex.* This model's parameters were 4 discrete GAMMA rate categories with an estimate of proportion of invariable sites. An additional test ran 1000 bootstrap replicates. Only clades with bootstrap values >60 are shown to observe possible monophyletic patterns and orthologs that have expanded across *Daphnia* genomes (Figure 2.2).

Clades in figure 2.2 are not monophyletic; substrate specificity is variable across both *Daphnia* genomes. Observed are 31 pairs orthologs with conserved Trypsin-like substrate specificity, 3 pairs of orthologs with chymotrypsin-like specificity, and 16 pairs of orthologs that show substrate specificity of unknown origin. No orthologs with Elastase-like substrate specificity were observed in the phylogeny.

Clade A in figure 2.2 shows the most ancestral orthologs in both genomes, though it is observed that TRY03 and 06 in *D. magna* have lost the signal peptide after

divergence. Clade B shows a clip-domain serine protease expansion across both species of *Daphnia*, though H-SERP104 has lost the clip domain after divergence. Clade C in Figure 2.3 shows a trypsin expansion that is unresolved across both genomes, where then Trypsin and their homologs duplicated within each genome.

Group D in figure 2.2 contains the most orthologs; the relationship between the duplicates within species of both species is unresolved, however the relationship between the orthologs across species show confidence that the duplication events occurred before divergence. Within this group are several orthologs shared across species of arthropods. TRY20 in *D. pulex* and TRY27 in *Daphnia* magna are Nudel-like orthologs, sharing multiple LDLa domains. TRY21 in *pulex* and TRY40 in *D. magna* are Corin-like. However, TRY40 in D. magna do not share the SEA domain or the transmembrane domain with two LDLa domains. Instead, TRY40 only contains an SP-like domain, which shows high sequence similarity to the Corin-like orthologs in *D. pulex*. TRY6 in *D. pulex* and TRY38 in *D.magna* exhibit similar domain architecture patterns and sequence similarity to the ortholog Tequila, a neurotrypsin. Expansions of clip domain SPs is also observed to be randomly distributed in this unresolved group.

Clade E in figure 2.4 shows recent duplicates of homologs only within *D. pulex*. Clades showing similar patterns of duplicates of homologs are not observed in D. magna. Clade F in figure 2.5 shows the expansion of Trypsin and their homologs in *Daphnia*. Eight ancesteral nodes are observed within this clade to have gone under further duplication after divergence.

Clade G in figure 2.6 shows the expansion of the CBD2 domain across *daphnia*. However, only *Daphnia pulex* show further duplications of this CBD2 carrying SP. It

was hypothesized that CHY2 in *pulex* was the most SP domain of the clade within the *D. pulex* genome. CHY02 is *D. magna'*s ortholog of CHY2 in *D.pulex* and contains a transmembrane domain.

Analysis of CBD2-SPs and Clip-domain SPs across species of Daphnia

It was observed that only 1 CBD2 carrying domain exists in *D. manga* (CHY02) whereas there are 8 CBD2 carrying domain in *pulex*. Phylogenetic analysis of both genomes show that CHY2 in *D. pulex* and CHY02 in *D. magna* are orthologs. The remaining CBD2 carrying domains in *pulex* are duplicates that resulted after divergence (Figure 2.6).

Five clip-domain SPs were observed in *D. magna* as well in *D. pulex*. Phylogenetic analysis shows the distribution of the Clip-domain SPs to be variable rather than orthologous. Sequence comparison of the 10 clip-domain SPs using muscle alignment algorithm with -2.9 gap penalty in MEGA 6, and Maximum Likelihood method in RAxML for phylogentic reconstruction, to further investigate orthologous relationships across *Daphnia* species. In figure 2.7, we observe the relationship between SERP01 in *D. magna* and TRY36 in *D. pulex* is unresolved. TRY13 in *D.magna* is shown to be an orthologs of TRY18 in *D. pulex*. TRY 15B and TRY15A are duplicates that only occurred in *D. pulex* after divergence. H-SERP001, TRY12, and TRY37 are duplicates that occurred only in *D. magna* after divergence.

Selection on SPs and H-SPs

Four orthologs pairs from clade G in figure 2.6 were chosen for selection analysis because of 1.) strong sequence similarity, 2.) Confidence that specific pairs of genes are orthologs. We used the transcript sequences from both genomes *D.magna* and *D. pulex.*

Selection tests using the Nei-Gojobori substitution model retrieved negative values for all orthologs pairs, thus the overall mean distance across the orthologs (dN-dS = -1.009; S.E. 0.093) exhibited purifying selection. Hypothesis testing for was done using the one-sample T-test. When tested against the null ($H_0 = 0$) using once sample t-test, we found the overall mean distance across the orthologs (dN-dS = -1.009; S.E. 0.093) to be significantly different than the null ($P < 0.0001$). Positive selection was not observed within this clade. Selection analysis between species across the all orthologs within the phylogeny was not possible due to nucleotide saturation.

2.4 DISCUSSION

The purpose of this study is to catalogue all serine protease like genes in the crustacea *Daphnia magna*. We expected to find monophyletic patterns of substrate specificity and non-neutral selection within orthologous pairs of SPs and H-SPs that have expanded across genomes of crustacea. In this study, both the *Daphnia pulex* and the *Daphnia magna* genomes are used to model the evolution of the Serine protease gene family, which responds and adapt to SP inhibiters called serpins found in algae and plants (Potempa et al 1994). This preliminary work will contribute to convey possible mechanisms in the evolution of gene expression among serine protease gene duplicates, observed to also be involved in extracellular digestion, embryonic development, innate immunity, and the nervous system of arthropods.

The serine protease gene family makes up approximately 73-85% of the enzymatic activity in the gut of *Daphnia* (Elert et al 2003). *Daphnia* are a model organism in observing immune response (McTaggart et al 2009) and the genetic

expression brought on by resource allocation (Schwarzenberger et al 2010; Dudycha et al 2012), characteristics of the role of the SP gene family (Rawlings and Barrett 1993).

Seventy one SPs and H-SPs were found in the *Daphnia magna* genome, this number is low in comparison to the 211 genes found in *Daphnia pulex*. Among the 71 SPs and H-SPs, we observed conservation of the biochemical properties in each motif involved in the formation of the catalytic triad. Residues in both the TAAHC and the DIAL motif conserved the hydrophobicity where as GDSGGP residues remained as highly conserved as the Ser-195. Substrate specificity residues of the SP subfamilies were not monophyletic, but instead showed varying points of origin throughout the phylogeny. Varying number of SPs and H-SPs are found across other species of arthropods: 57 in *Apis Mellifera* (Zou et al 2006), 305 in *A. gambiae* (Wu et al 2009), and 206 in *D. melanogaster* (Ross et al 2003). Although the number of gene duplicates across species is variable, a number of orthologs across all arthropods with known function remain conserved and have expanded across all arthropods. For example, Tequila is a conserved ortholog involved in information processing (Didelot et al 2006). Nudel, also found in all arthropods, is important in regulating the protease cascade for dorsal and ventral polarity of the embryo and stability of the egg (Hong and Hashimoto 1996; Lemosy et al 1998). Also conserved across taxa of arthropods is Corin, which aids in regulation of blood circulation and coagulation in mammals (Rao et al 2001).

Clip domains are involved in the innate immunity of arthropods (Jiang and Kanost 2000). In the *Daphnia pulex* and *Daphnia magna* genome, 5 clip domain SPs were found in each. Of these 5, SERP01 in *D. magna* and TRY36 in *D. pulex* are proposed to be orthologous clip-domain SPs. TRY 13 in *D. magna* and TRY18 and *D. pulex* are also

proposed to be orthologous clip-domain SPs. The remaining clip-domain SPs may not have expanded across both genomes, but instead, after divergence, underwent duplication either by unequal crossing over, homologous recombination, or transposon involvement within the species genome. Still, this number of clip-domain SPs is small relative to the 41 found in *A. gambai* (Wu et al 2009), 18 in *A. mellifera* (Zou et al 2006), and 37 in *D. melanogaster* (Ross et al 2003). The duplication events in clip-domain SPs are more prominent in hexapoda than crustacea.

Unequal crossing over may have occurred in recent expansions of trypsins, observed in Clade C (Figure 2.3) and Clade F (Figure 2.5), and chymotrypsins, observed in Clade G (Figure 2.6) before divergence of the two *Daphnia* species. This is conveyed by the 50 total orthologs that were found to be shared between both genomes. Positive selection was observed in the catalytic sites of SPs in *A. gambie* and proposed to be a result of adaptive evolution for the process of digestion of food (Wu et al 2009). However, in both *D. pulex* and *D. magna*, strong evidence of purifying selection was observed. This non-neutral selection of SPs and H-SPs that have expanded across genomes may reinforce the conservation of the most basal chymotrypsin (CHY02 in *D. magna*, CHY2 in *D. pulex*) that has expanded across all *Daphnia*.

Overall, no monophyletic clades were observed within the phylogeny of *D. magna* and *D. pulex*. However, comparison of the two species does exhibit small clustering of subfamilies. The nucleotide sequences of orthologs within Clade G (Figure 2.6) showed purifying selection to reinforce the basal digestive function of chymotrypsins during expansion. This could be because of the similar diet preference. The expansion of the SP serine protease gene family in arthropods is large, and varying selectional patterns

have been observed within and across species of arthropod based on resource preference (Wu et al 2009; Li et al 2012). This framework of genomic information across all species of arthropods reveals interesting selectional pressures that could be further investigated by observing the effects of resource allocation, immune response to serpins (serine protease inhibitors), and embryological development on gene expression.

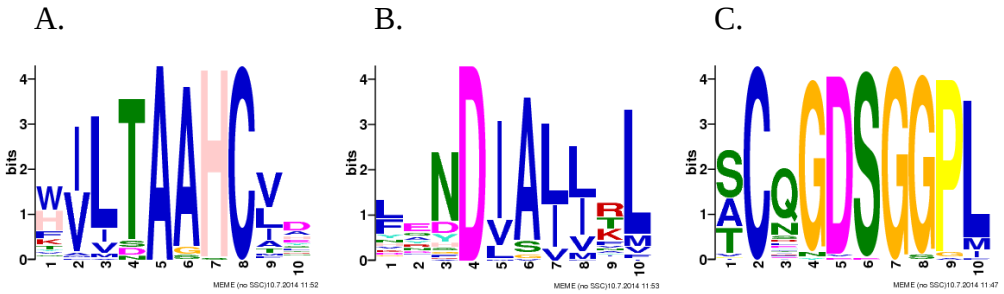


Figure 2.1 Features of the Motifs in the Catalytic Triad of Complete SPs. The residues involved in peptide chain hydrolysis are embedded in the motifs A, B, and C. Height of the logo, bits, represents the probability of that residue occurring in that position multiplied by the total amount of information in that position. The colors of each residue represent the following: Blue: most hydrophobic; Green: Polar, non-charged, non-aliphatic; Magenta: Acidic; Red: Positively Charged(Bailey et al 2006).

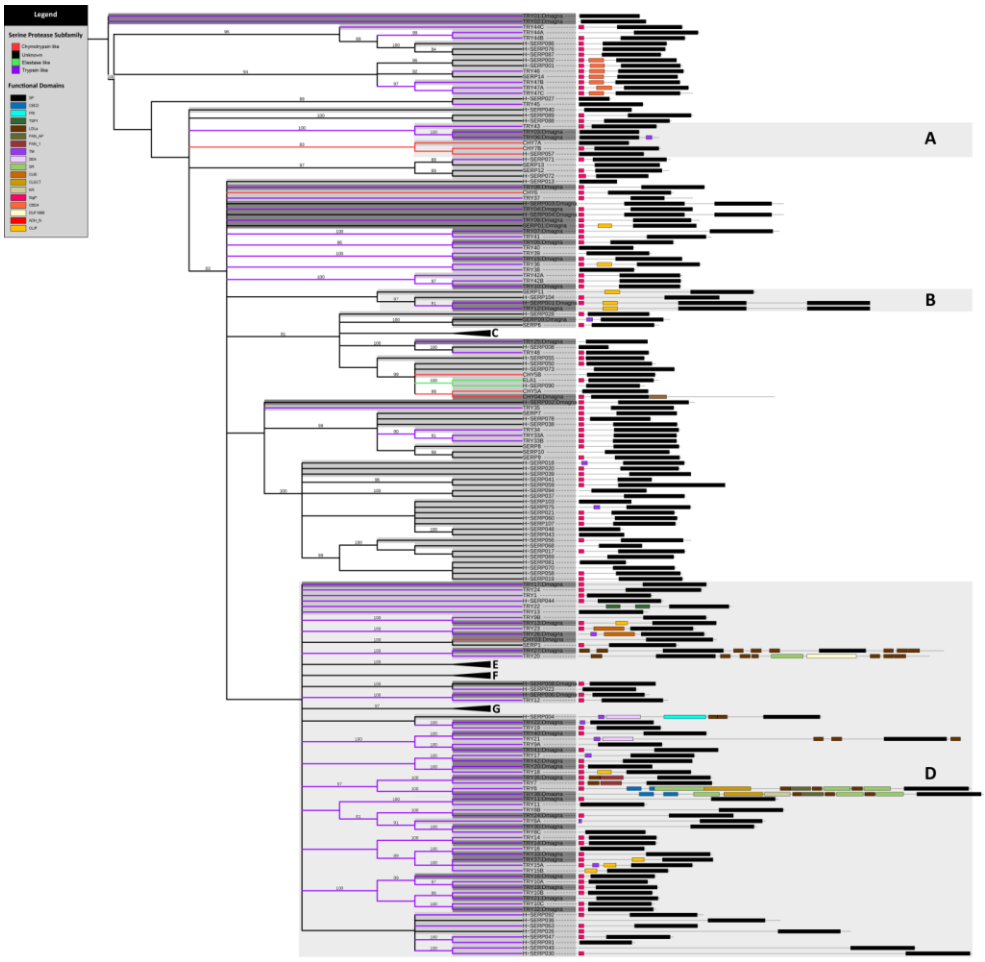


Figure 2.2. Phylogenetic relationship of all serine protease domains found in the *Daphnia magna* and *D. pulex* genome. Phylogenetic analysis was done in RAxML, as mentioned in Section 2.2. Collapsed clades C, E, F, and G are expanded in Figure 2.4.2, Branch colors represent the subfamily classification of each serine protease which is dependent on the substrate specificity of the amino acid residues. The colored domain architecture represents additional functional domains that may be present on each SP containing gene. Vertical lines represent putative gene clusters labeled for analysis.

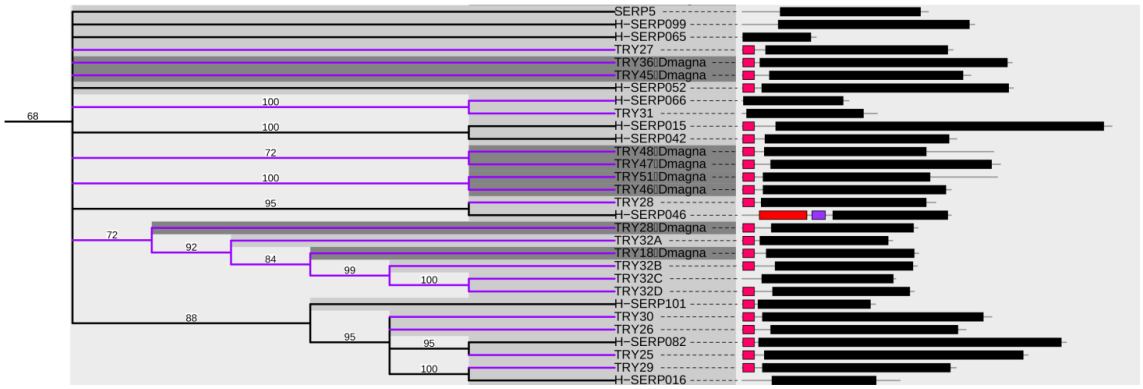


Figure 2.3. Phylogenetic relationship of all serine protease domains found from Clade C in the *Daphnia magna* and *D. pulex* genome. Phylogenetic analysis was done in RAxML, as mentioned in Section 2.2. Branch colors represent the subfamily classification of each serine protease which is dependent on the substrate specificity of the amino acid residues. The colored domain architecture represents additional functional domains that may be present on each SP containing gene. Vertical lines represent putative gene clusters labeled for analysis.

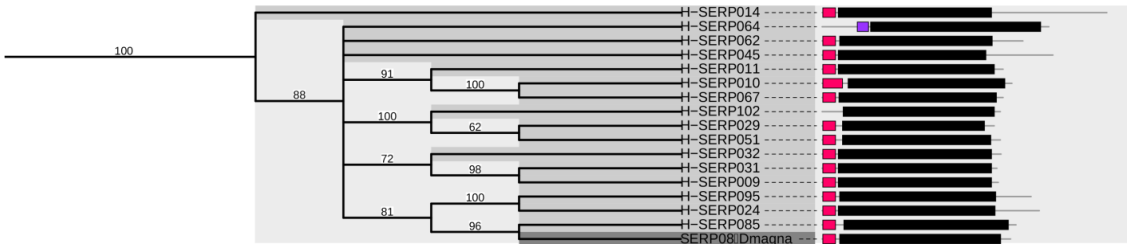


Figure 2.4. Phylogenetic relationship of all serine protease domains found in Clade E in the *Daphnia magna* and *D. pulex* genome. Phylogenetic analysis was done in RAxML, as mentioned in Section 2.2. Branch colors represent the subfamily classification of each serine protease which is dependent on the substrate specificity of the amino acid residues. The colored domain architecture represents additional functional domains that may be present on each SP containing gene. Vertical lines represent putative gene clusters labeled for analysis.

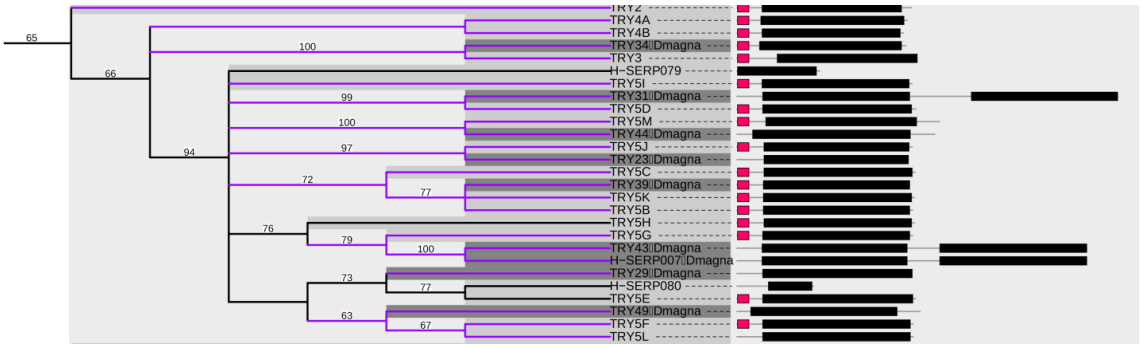


Figure 2.5. Phylogenetic relationship of all serine protease domains found in Clade F in the *Daphnia magna* and *D. pulex* genome. Phylogenetic analysis was done in RAxML, as mentioned in Section 2.2. Branch colors represent the subfamily classification of each serine protease which is dependent on the substrate specificity of the amino acid residues. The colored domain architecture represents additional functional domains that may be present on each SP containing gene. Vertical lines represent putative gene clusters labeled for analysis.

Figure 2.6. **Phylogenetic relationship of all serine protease domains found in Clade G in the *Daphnia magna* and *D. pulex* genome.** Phylogenetic analysis was done in RAxML, as mentioned in Section 2.2. Branch colors represent the subfamily classification of each serine protease which is dependent on the substrate specificity of the amino acid residues. The colored domain architecture represents additional functional domains that may be present on each SP containing gene. Vertical lines represent putative gene clusters labeled for analysis.

A.

```
Group
SERP01_Dmagna              CYTAEGRFGSCMSFRSCYPTS---ELSYLQTWSVIMQIACSYTVKDGRQMHGICCP 4
TRY36                      CLTREGNIGYCTSIRSCYPRLNKFHHFNFESRTLAIRGACIYHRADDRQVYGICCP 1
SERP11                     CWMSDGKSGLCGPVRSCHPHDELQEPLNPESRMLPSRTLCGYVNKNGKQDTGVCCP 1
TRY13_Dmagna               CTTPDGNRGQCRDLGSCPAL--------LLQLDSLRKSICFQSL----FVPGVCCP 2.2
TRY18                      CQTPEGVVGTCTPLTNCPHLADMLSVPSPAILNFLRQSICGYEG----YDPKVCCS 2.1
TRY15A                     CLTPISQSGRCRFVQHCALPEIIVTLNAFVT----------YACSIGSDYMGVCCP 3
TRY15B                     CSTPLSQSGRCRFVQHCARQEIIATLNAFVS----------YACPIGSDYMGVCCP 3
H-SERP001_Dmagna           CWMPDGTYGVCNSALSCNLRDKLHEARYPLSSMFSSRNTCRYIT-NGKEETGVCCP 2
TRY12_Dmagna               CWMPDGTYGVCNSALSCNLRDKLHEARYPLSSMFSSRNTCRYIT-NGKEETGVCCP 2
TRY37_Dmagna               CWMPDGTYGVCNSALSCNLRDKLHEARYPLSSMFSSRNTCRYIT-NGKEETGVCCP 2
                           *    .  * *    *                                  :**.
```

B.

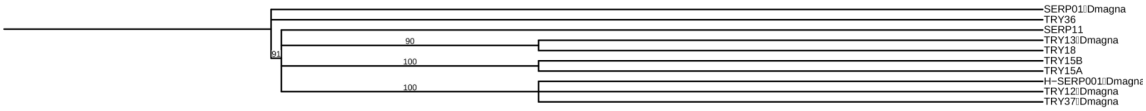


Figure 2.7. Sequence comparison and phylogenetic relationships among the *Daphnia pulex* and *Daphnia magna* clip-domain SPs and H-SPs. A. alignment of the clip domain sequences. Six conserved Cys residues form 3 disulphide bonds. B Phylogenetic tree based on an alignment of the CLIP domains. Group number indicates genes sharing similar residues.

Table 2.1 Characteristics of each Serine Protease domain in the *Daphnia magna* genome. Superscript [d] indicates the predicted location of the cleavage site for activation of the zymogen, (Onting 1998). Superscript [e,f,g] indicates the conserved motifs of the catalytic triad. Fields left blank indicate that the domain has either the full TAAHC, DIAL, or GDSGGP motif. Motifs in red indicate that the putative residue for the catalytic triad is either substituted or missing. Motif predictions were made using the database from smart.embl-heidelberg.de and http://prosite.expasy.org/ as well as multiple alignments in MEGA 5.10 (Onting 1998; de Castro et al 2006; Tamura et al 2011). Superscript [f] indicates the predicted substrate specificity using the multiple alignment algorithm in MEGA 5.10 (Perona and Craik 1995; Tamura et al 2011).

Magna 2012 ID	Name	Active Site	TAAHC	DIAL	GDSGGP	Signal Peptide	Substrate Specificity	Length (aa)
m8AUGepir7p2s01581g44t1	TRY01_Dmagna	SLATG	TAAHC	DIAI	GDSGGP	N	DGG	230
m8AUGepir7p2s01581g41t1	TRY02_Dmagna	RMTES	TAAHC	DIAI	GDSGGP	N	DGG	255
m8AUGep24b_p2s01581g41t1	TRY03_Dmagna	--MKR	TAAHC	DIAI	NDSGGP	N	DGG	230
m8AUGep24bs01253g60t1	TRY04_Dmagna	RIVGG	TAAHC	DIAI	GDSGGP	Y	DGG	427
m8AUGepir7s00872g334t1	TRY05_Dmagna	RIVGG	TAAHC	DVAI	GDSGGP	Y	DGG	358
m8AUGep24bs01285g298t1	TRY06_Dmagna	KR---	TAAHC	DIAI	NESGGP	N	DGG	300
m8PASAgasmbl_36231	TRY07_Dmagna	RIIGG	TAAHC	DIAI	GDSGGP	Y	DGG	749
m8AUGepir7s00872g333t1	TRY08_Dmagna	RIVGG	TAAHC	DIAI	GDSGGP	Y	DGG	473
m8AUGepir7s01253g127t1	TRY09_Dmagna	RIVGG	TAAHC	DVAI	GDSGGP	Y	DGQ	452
m8PASAgasmbl_36302	TRY10_Dmagna	RIVGG	TAAHC	DIAL	GDSGGP	Y	DGG	386
m8AUGapi5s02489g294t1	TRY11_Dmagna	RIVGG	TAAHC	DLAI	GDSGGP	Y	DGG	747
m8AUGep24bs00872g271t1	H-SERP001_Dmagna	RIVGG	TAAHC	DVAI	GDSGGP	Y	DGG	1087
m8AUGep24bs00872g271t1	TRY12_Dmagna	RIVGG	TAAHC	DVAI	GDSGGP	Y	DGG	1087
m8PASAgasmbl_70821	SERP01_Dmagna	RIVGG	TAAHC	DIAI	GDSGGP	Y	SGV	439
m8PASAgasmbl_13592	TRY13_Dmagna	RVVGG	TAAHC	DLAL	GDSGGP	Y	DGG	514
m8PASAgasmbl_48424	TRY14_Dmagna	RIVGG	TAAHC	DIAI	GDSGGP	Y	DGG	295
m8AUGep24bs00872g275t1	TRY15_Dmagna	RIVGG	TAAHC	DIAI	GDSGGP	Y	DGG	368
m8PASAgasmbl_39448	SERP02_Dmagna	RIVGG	TAAHC	DLAL	GDSGGP	Y	GGA	303
m8AUGapi5p1s00944g3t1	SERP03_Dmagna	RIVGG	TAAHC	DVAL	GDSGGP	Y	GGT	309

m8AUGepir7s02545g145t1	TRY16_Dmagna	KIVNG	TAAHC	DIAL	GDSGGP	Y	DGG	291
m8AUGepir2s02140g119t1	H-SERP002_Dmagna	QIVSG	TAAHC	DIAI	GDSGGP	Y	SGA	433
m8PASAgasmbl_73275	TRY17_Dmagna	RIVGG	TAAHC	DIAL	GDSGGP	Y	DGG	477
m8AUGepir7p1s00944g9t1	SERP04_Dmagna	RIVGG	TAAHC	DIAL	GDSGGP	Y	GGA	274
m8AUGapi5p1s01581g54t1	TRY18_Dmagna	KIVGG	TAAHC	DIAL	GDSGGP	Y	DGG	286
m8AUGepir2s02545g132t1	TRY19_Dmagna	KIVNG	TAAHC	DIAL	GDSGGP	Y	DGG	301
m8PASAgasmbl_39465	SERP05_Dmagna	RIVGG	TAAHC	DVAL	GDSGGP	Y	SGA	415
m8PASAgasmbl_44254	TRY20_Dmagna	RIVGG	TAAHC	DIAI	GDSGGP	Y	DGG	275
m8PASAgasmbl_79498	TRY21_Dmagna	RIVNG	TAAHC	DIAL	GDSGGP	N	DGG	304
m8PASAgasmbl_39453	SERP06_Dmagna	RIVGG	TAAHC	DMAL	GDSGGP	N	SGA	1446
m8AUGepir6s00311g147t1	TRY22_Dmagna	RIVGG	SAGHC	DISI	GDSGGP	N	DGG	280
m8AUGapi5s00868g254t1	TRY23_Dmagna	KIVGG	NAAHC	DISL	GDSGGP	N	DGG	280
m8AUGep24bs01253g122t1	H-SERP003_Dmagna	RLFGP	TAAHC	DIAI	GDSGGP	N	DG-	765
m8AUGep24bs01253g122t1	H-SERP004_Dmagna	RLFGP	TAAHC	DIAI	GDSGGP	Y	DG-	765
m8AUGepir3s00872g288t1	TRY24_Dmagna	RIVGG	TAGHC	DLAL	GDSGGP	Y	DGG	577
m8AUGapi5s00311g135t1	TRY25_Dmagna	QIVGG	TAAHC	DIAI	GDSGGP	N	DGG	257
m8AUGepir7s00084g86t1	TRY26_Dmagna	RVVGG	TAAHC	DIAL	GDSGGP	N	DGG	478
m8AUGepir7s00915g51t1	TRY27_Dmagna	RVVGG	SAAHC	DIAL	GDSGGP	N	DGG	1362
m8AUGep24bs01117g8t1	CHY01_Dmagna	KIVEG	TAAHC	DIAL	GDSGGP	Y	GGG	440
m8AUGapi5p1s00944g362t1	TRY28_Dmagna	QIVGG	TAAHC	DVAI	GDSGGP	Y	DGG	285
m8PASAgasmbl_35335	TRY29_Dmagna	RIVGG	DAAHC	DISL	GDSGGP	N	DGG	286
m8AUGepir3p2s00024g202t1	TRY30_Dmagna	RIVGG	TAAHC	DLAL	GDSGGP	N	DGG	665
m8AUGepir7s00868g265t1	TRY31_Dmagna	KIVGG	TAAHC	DIAL	GDSGGP	N	DGG	622
m8AUGepir3s02545g136t1	TRY32_Dmagna	KIVNG	TAAHC	DIAL	GDSGGP	Y	DGG	289
m8AUGepir7s03102g104t1	SERP07_Dmagna	RIING	TAAHC	DVAL	GDSGSA	Y	GVG	310
m8PASAgasmbl_27533	CHY02_Dmagna	RIVSG	TAAHC	DIGL	GDSGGP	Y	GGG	435
m8AUGepir7p1s00944g16t1	H-SERP005_Dmagna	RIVGG	TAAHC	DVAL	GDSGGP	Y	SGA	272
m8PASAgasmbl_35325	TRY33_Dmagna	KLSQA	TAAHC	DIAV	GDSGGP	Y	DGG	498

m8PASAgasmbl_87235	TRY34_Dmagna	-IVGG	SASHC	DIVL	GDSGGP	Y	DGG	276
m8PASAgasmbl_40324	TRY35_Dmagna	RVVNG	TAAHC	DIAL	GDSGGP	Y	DGG	499
m8AUGep24bs02837g9t1	H-SERP006_Dmagna	RIVGG	TAAHC	DIAI	GDSGGP	Y	DGG	265
m8AUGepir7p1s00944g445t1	TRY36_Dmagna	KIVNG	TAAHC	DVAL	GDSGGP	Y	DGG	438
m8AUGapi5s00868g251t1	TRY37_Dmagna	RIVGG	TASHC	DIAL	GDSGGP	Y	DGG	501
m8PASAgasmbl_13216	SERP08_Dmagna	RIIGG	TAASC	DIAL	YDEGSP	Y	TSI	276
m8AUGep24bs00626g76t1	TRY38_Dmagna	KIVKG	TAAHC	DIAL	GDSGGG	N	DGG	1511
m8AUGepir7s00868g268t1	TRY39_Dmagna	RIVGG	NAAHC	DIAL	GDSGGP	Y	DGG	282
m8AUGepir7s02076g49t1	TRY40_Dmagna	RIVGG	TAGHC	DVAV	GDSGGP	Y	DGG	477
m8AUGep24bs00005g95t1	CHY03_Dmagna	RIING	TAAHC	DIAL	GDSGGP	N	GGD	515
m8PASAgasmbl_68665	TRY41_Dmagna	RIIGG	TAAHC	DVAV	GDSGGP	Y	DGG	521
m8AUGep24b_p1s01361g366t1	CHY04_Dmanga	EIIGG	TAAHC	DIAI	GDSGGP	Y	GGG	731
m8AUGepir7p2s00024g219t1	SERP09_Dmagna	RIVGG	TAAHC	DLGV	GDSGGP	N	SAA	341
m8AUGepir3s01005g231t1	TRY42_Dmagna	RIAGG	TAAHC	DIAI	GDSGGP	Y	DGG	424
m8AUGepir7s00868g262t1	H-SERP007_Dmagna	RIVGG	DAAHC	DISL	GDSGGP	N	DGG	571
m8AUGepir7s00868g262t1	TRY43_Dmagna	RIVGG	DAAHC	DISL	GDSGGP	N	DGG	571
m8AUGepir7s01764g47t1	TRY44_Dmagna	KIVGG	TAAHC	DIAL	GDSGGP	N	DGG	323
m8AUGapi5p2s00024g124t1	TRY45_Dmagna	SIVGG	TAAHC	DVAL	GDSGGP	Y	DGG	371
m8AUGepir7p1s00944g397t1	TRY46_Dmagna	SIVGG	TAAHC	DIAV	GDSGGP	Y	DGG	339
m8AUGep24b_p1s00944g332t1	TRY47_Dmagna	SIVGG	TAAHC	DIAL	GVSGGP	Y	DGG	419
m8PASAgasmbl_41322	TRY48_Dmagna	SIVGG	TAAHC	DIAL	GDSGGP	Y	DGG	408
m8AUGepir7s00868g263t1	TRY49_Dmagna	KIVGG	DAAHC	DISL	GDSGGP	N	DGG	299
m8AUGepir2s02066g6t1	H-SERP008_Dmagna	KIVGG	TAAHC	DLAL	------	Y	D--	287
m8AUGepir7p1s00944g398t1	TRY51_Dmagna	SIVGG	TAAHC	DVAV	GDSGGP	Y	DGG	414

REFERENCES

Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34:W369–373.

Bark P (1993) The modular architecture of a new family of growth regulators tissue growth factor. 327:125–130.

Bork P, Beckman G (1993) The CUB Domain: A Widespread Module in Developmentally Regulated Proteins. J Mol Biol 231:539–545.

Brenner S (1988) The molecular evolution of genes and proteins: a tale of two serines. Nature 334:528–530.

Brown M, Goldstein J (1986) A receptor-mediated pathway for cholesterol homeostasis. Science (80-) 232:34–47.

Casaretto J, Corcuera L (1995) Plant proteinase inhibitors: a defensive response against insects. (Review). Biol Res 28:239–49.

Colbourne JK, Pfrender ME, Gilbert D, et al (2011) The ecoresponsive genome of Daphnia pulex. Science 331:555–61. doi: 10.1126/science.1197761

De Castro E, Sigrist CJ a, Gattiker A, et al (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34:W362–5. doi: 10.1093/nar/gkl124

Didelot G, Molinari F, Tchenio P, et al (2006) Tequila, a Neurotrypsin Ortholog, Regulates Long-Term Memory Formation in Drosophila. Science (80-) 313:851–853.

Dudycha JL, Brandon CS, Deitz KC (2012) Population genomics of resource exploitation: insights from gene expression profiles of two Daphnia ecotypes fed alternate resources. Ecol Evol 2:329–40. doi: 10.1002/ece3.30

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7. doi: 10.1093/nar/gkh340

Elert E von, Agrawal MK, Gebauer C, et al (2003) Protease activity in gut of Daphnia magna: evidence for trypsin and chymotrypsin enzymes. Comp Biochem Physiol Part B Biochem Mol Biol 137:287–296.

Elvin CM, Vuocolo T, Pearson RD, et al (1996) Characterization of a Major Peritrophic Membrane Protein , Peritrophin-44 , from the Larvae of Lucilia cuprina. 271:8925–8935.

García-Mayoral MF, Hollingworth D, Masino L, et al (2007) The structure of the C-terminal KH domains of KSRP reveals a noncanonical motif important for mRNA degradation. Structure 15:485–98. doi: 10.1016/j.str.2007.03.006

Gilbert D, Singan, V.R., Colbourne JK (2005) wFleaBase: the Daphnia genomics information system. BMC Bioinformatics 6:45. doi: 10.1186/1471-2105/6/45

Gliwicz ZM, Boavida MJ (1993) Clutch size and body size at first reproduction in Daphnia pulicaria at different levels of food and predation. 18:863–880.

Greer J (1990) Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins 7:317–34. doi: 10.1002/prot.340070404

Guda C (2006) pTARGET: a web server for predicting protein subcellular localization. Nucleic Acids Res 34:W210–3. doi: 10.1093/nar/gkl093

Hartley BS (1964) Amino-acid sequence of bovine chymotrypsinogen-A. Nature 201:1284–1287.

Hedstrom L, Lin T-Y, Fast W (1996) Hydrophobic Interactions Control Zymogen Activation in the Trypsin Family of Serine Proteases. Biochemistry 35:4515–4523.

Hong CC, Hashimoto C (1996) The maternal nudel protein of Drosophila has two distinct roles important for embryogenesis. Genetics 143:1653–61.

Jang I-H, Nam H-J, Lee W-J (2008) CLIP-domain serine proteases in Drosophila innate immunity. BMB Rep 41:102–7.

Jiang H, Kanost MR (2000) The clip-domain family of serine proteinases in arthropods. Insect Biochem Mol Biol 30:95–105.

Kraut J (1977) Serine proteases: structure and mechanism of catalysis. Annu Rev Biochem 46:331–58. doi: 10.1146/annurev.bi.46.070177.001555

Lemosy EK, Kemler D, Hashimoto C (1998) Role of Nudel protease activation in triggering dorsoventral polarization of the Drosophila embryo. Development 125:4045–4053.

Li L, Memon S, Fan Y, et al (2012) Recent duplications drive rapid diversification of trypsin genes in 12 Drosophila. Genetica 140:297–305. doi: 10.1007/s10709-012-9682-5

McMullen B, Fujikawa K, Davie E (1991) Location of the disulfide bonds in human plasma prekallikrein: the presence of four novel apple domains in the amino-terminal portion of the molecule. Biochemistry 30:2050–6.

McTaggart SJ, Conlon C, Colbourne JK, et al (2009) The components of the Daphnia pulex immune system as revealed by complete genome sequencing. BMC Genomics 10:175. doi: 10.1186/1471-2164-10-175

Nei M, Gojoborit T (1986) Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions ’. Mol Biol Evol 3:418–426.

Onting CHPP (1998) SMART , a simple modular architecture research tool : Identification of signaling domains. 95:5857–5864.

Patthy L, Trexler M, Vbli Z, et al (1984) Kringles : modules specialized for protein binding Homology of the gelatin-binding region of fibronectin structures of proteases with the kringle. FEBS Lett 171:131–136.

Perona JJ, Craik CS (1995) Structural basis of substrate specificity in the serine proteases. Protein Sci 4:337–60. doi: 10.1002/pro.5560040301

Potempa J, Korzus E, Travis J (1994) The Serpin Superfamily of Proteinase Inhibitors: Structure, Function, and Regulation. J Biol Chem 269:15957–15960.

Rao S, Lang C, S. Levitan E, L. Deitcher D (2001) Visualization of neuropeptide expression, transport, and exocytosis in Drosophila melanogaster. J Neurobiol 49:159–172.

Rawlings ND, Barrett AJ (1993) Evolutionary families of peptidases. Biochem J 290:205–218.

Rebers JE, Willis JH (2001) A conserved domain in arthropod cuticular proteins binds chitin. Insect Biochem Mol Biol 31:1083–93.

Resnick D, Pearson A, Krieger M (1994) The SRCR superfamily: a family reminiscent of the lg superfamily. Trends Biochem Sci 19:5–8.

Ross J, Jiang H, Kanost MR, Wang Y (2003) Serine proteases and their homologs in the Drosophila melanogaster genome: an initial analysis of sequence conservation and phylogenetic relationships. Gene 304:117–31.

Sarnelle O (2005) Daphnia as keystone predators: effects on phytoplankton diversity and grazing resistance. J Plankton Res 27:1229–1238. doi: 10.1093/plankt/fbi086

Schwarzenberger A, Zitt A, Kroth P, et al (2010) Gene expression and activity of digestive proteases in Daphnia: effects of cyanobacterial protease inhibitors. BMC Physiol 10:6. doi: 10.1186/1472-6793-10-6

Schwerin S, Zeis B, Lamkemeyer T, et al (2009) Acclimatory responses of the Daphnia pulex proteome to environmental changes. II. Chronic exposure to different temperatures (10 and 20 degrees C) mainly affects protein metabolism. BMC Physiol 9:8. doi: 10.1186/1472-6793-9-8

Shen Z (1998) A Type I Peritrophic Matrix Protein from the Malaria Vector Anopheles gambiae Binds to Chitin. CLONING, EXPRESSION, AND CHARACTERIZATION. J Biol Chem 273:17665–17670. doi: 10.1074/jbc.273.28.17665

Sigrist CJA, Cerutti L, Hulo N, et al (2002) PROSITE : A documented database using patterns and profiles as motif descriptors. 3:265–274.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–90. doi: 10.1093/bioinformatics/btl446

States D, Gish W (1994) Combined use of sequence similarity and codon bias for coding region identification. J Comput Biol 1:39–50.

Suetake T, Tsuda S, Kawabata S, et al (2000) Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif. J Biol Chem 275:17929–32. doi: 10.1074/jbc.C000184200

Tamura K, Peterson D, Peterson N, et al (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–9. doi: 10.1093/molbev/msr121

Tessier A, Leibold M, Tsao J (2000) A FUNDAMENTAL TRADE-OFF IN RESOURCE EXPLOITATION BY DAPHNIA AND CONSEQUENCES TO PLANKTON COMMUNITIES. Ecology 81:826–841.

Vanin EF (1985) Processed Pseudogenes Characteristics and evolution. Annu Rev Genet 19:253–272.

Wu D-D, Wang G-D, Irwin DM, Zhang Y-P (2009) A profound role for the expansion of trypsin-like serine protease family in the evolution of hematophagy in mosquito. Mol Biol Evol 26:2333–41. doi: 10.1093/molbev/msp139

Zhang Z, Schäffer AA, Miller W, et al (1998) Protein sequence similarity searches using patterns as seeds. 26:3986–3990.

Zou Z, Lopez DL, Kanost MR, et al (2006) Comparative analysis of serine protease-related genes in the honey bee genome : possible involvement in. 15:603–614.

APPENDIX A: FIXED SERINE PROTEASE NAMES

Table A.1. Fixed SP Names. Fixed names for serine proteases in the *Daphnia pulex* genome. Each name presented below is changed based on further analysis of substrate specificity residues. Original names were from a 2009 study by Schwerin, et. al.

Original Name	Fixed Name
CHY1A	SERP15
CHY1C	SERP16
TRY5E	SERP17
TRY5H	SERP18