Identification and in silico structural and functional analysis of a trypsin-like protease from shrimp *Macrobrachium carcinus*

José M. Viader-Salvadó[1], José Alberto Aguilar Briseño[1],
Juan A. Gallegos-López[1], José A. Fuentes-Garibay[1],
Carlos Alfonso Alvarez-González[2] and Martha Guerrero-Olazarán[1]

[1] Instituto de Biotecnología, Facultad de Ciencias Biológicas, Universidad Autónoma de Nuevo León, San Nicolás de los Garza, Nuevo León, Mexico
[2] Laboratorio de Acuicultura Tropical, División Académica de Ciencias Biológicas, Universidad Juárez Autónoma de Tabasco, Villahermosa, Tabasco, Mexico

ABSTRACT

Macrobrachium carcinus (Linnaeus, 1758) is a species of freshwater shrimp widely distributed from Florida southwards to southern Brazil, including southeast of Mexico. In the present work, we identified a putative trypsin-like protease cDNA fragment of 736 nucleotides from *M. carcinus* hepatopancreas tissue by the 3′RACE technique and compared the deduced amino acid sequence to other trypsin-related proteases to describe its structure and function relationship. The bioinformatics analyses showed that the deduced amino acid sequence likely corresponds to a trypsin-like protease closely related to brachyurins, which comprise a subset of serine proteases with collagenolytic activity found in crabs and other crustacea. The *M. carcinus* trypsin-like protease sequence showed a global sequence identity of 94% with an unpublished trypsin from *Macrobrachium rosenbergii* (GenBank accession no. AMQ98968), and only 57% with *Penaeus vannamei* trypsin (GenBank accession no. CAA60129). A detailed analysis of the amino acid sequence revealed specific differences with crustacean trypsins, such as the sequence motif at the beginning of the mature protein, activation mechanism of the corresponding zymogen, amino acid residues of the catalytic triad and residues responsible for substrate specificity.

INTRODUCTION

In crustacea, the organ in the digestive tract called the hepatopancreas or midgut gland is known to have high proteolytic activity and it is responsible for assimilating ingested proteins. The proteolytic activity is mainly due to the presence of serine proteases, with trypsins being the main digestive endopeptidases (*Muhlia-Almazán, Sánchez-Paz & García-Carreño, 2008*; *Linton et al., 2014*; *Perera et al., 2015*). Trypsins (EC 3.4.21.4) are

Corresponding authors
José M. Viader-Salvadó,
jose.viadersl@uanl.edu.mx
Martha Guerrero-Olazarán,
martha.guerreroool@uanl.edu.mx

serine proteases found in the digestive systems of many vertebrates and invertebrates that cleave peptides at the carboxyl side of Lys or Arg residues. Because trypsins can cause tissue damage due to autologous protein hydrolysis, they are synthesized as inactive precursors (proenzymes or zymogens) called trypsinogens. The importance of trypsins in protein digestion is given by its role in the activation of other digestive proenzymes, and trypsinogen itself. The conversion of trypsinogen to active trypsin is initiated by the specific cleavage of a small N-terminal peptide by an autoactivation process or in mammals, by the action of enterokinase. In the shrimp *Penaeus vannamei*, trypsin synthesis from trypsinogen has been proposed from sequence analysis (*Klein et al., 1996*), the demonstration of trypsinogen storage in the midgut gland has been proved (*Sainz et al., 2004*), and recently we demonstrated a rapid trypsinogen autoactivation process (*Guerrero-Olazarán et al., 2019*).

Trypsin sequences from different species have been used as a model for studies of evolution. A continuous evolutionary divergence of trypsins from a common ancestor has been proposed (*Rypniewski et al., 1994*). Comparisons of activation peptide sequences have also been used to study the evolution and function of trypsinogen activation (*Chen et al., 2003*).

Macrobrachium carcinus (Linnaeus, 1758) is a species of freshwater shrimp widely distributed from Florida southwards to southern Brazil, including southeast of Mexico. This species has great aquaculture potential due to its large size, high fertility in captivity and resistance to handling and stress conditions. Furthermore, it has a short larval period, is omnivorous and its meat is of good quality and widely accepted. Another freshwater crustacean of interest to aquaculture, especially in Europe, is the crayfish *Astacus leptodactylus* that resembles a small lobster. Nevertheless, little is known about the digestive physiology of these species.

Rapid amplification of cDNA ends (RACE) is a widely used technique for obtaining a cDNA copy of a specific RNA transcript from a cell. In the 3′RACE version of this technique, cDNAs are synthesized in a reverse transcription reaction using an oligo-dT-adaptor primer directed to the natural polyA tail of eukaryotic mRNAs. In the second step, specific cDNA is amplified by a polymerase chain reaction (PCR) using a sense gene-specific primer and an anti-sense primer that is complementary to the adaptor sequence of the primer used in the first step (*Frohman, Dush & Martin, 1988*).

In this work, the cDNA sequence of a putative trypsin-like protease from *M. carcinus* hepatopancreas tissue was identified by the 3′RACE technique and the deduced amino acid sequence was compared to other trypsin-related proteases to describe structure and function relationship of the enzyme. Our findings contribute to the understanding of the digestive physiology of this species and the molecular mechanism of crustacean trypsins.

MATERIALS AND METHODS

Specimens, plasmids, medium composition, chemicals and enzymes

Macrobrachium carcinus specimens were from the Grijalva River, Centla, Tabasco, Mexico (latitude 18°14′11.9″ N, longitude 92°39′49.4″). All oligonucleotides were purchased from

Integrated (DNA Technologies, Inc., Coralville, IA, USA). *Escherichia coli* DH5α, pGEM-T easy vector, RQ1 RNase-free DNase, SV Total RNA Isolation System, and GoTaq DNA polymerase were purchased from Promega (Madison, WI, USA). Luria-Bertani (LB) agar plates (1% tryptone, 0.5% yeast extract, 1% NaCl, 15 g/L agar, pH 7.0) with 100 μg/mL ampicillin was used for *E. coli* transformants selection. *Pfu*Ultra II Fusion HotStart DNA polymerase and AccuScript High-Fidelity Reverse Transcriptase were from Agilent Technologies (Santa Clara, CA, USA). RNAlater was from Life Technologies (Gaithersburg, MD, USA). All chemicals were of analytical grade and purchased from Sigma–Aldrich Co. (St. Louis, MO, USA) or from Productos Químicos Monterrey (Monterrey, Nuevo León, Mexico).

3′RACE assay

Total RNA from *M. carcinus* hepatopancreas tissue was isolated using SV Total RNA Isolation System, treated with RQ1 RNase-free DNase and used to synthesize cDNA by reverse transcription using 14.7 μM of T17AP primer (5′-GACTCGAGTCGACATC GAT$_{17}$-3′) (*Frohman, Dush & Martin, 1988*) and 2 μl of AccuScript High-Fidelity Reverse Transcriptase in a final reaction volume of 20 μl, according to the manufacturer's recommendations. The cDNA was amplified by PCR using the RACEAP primer (5′-GAC TCGAGTCGACATCG-3′) (*Frohman, Dush & Martin, 1988*) and the consensus primer Pig1 (5′-CACTTCTGCGGCGCCTCCAT-3′) designed from a highly-conserved region of crustacean trypsin nucleotide sequences. The PCR was performed in a PCR Multigene Mini thermal cycler (Labnet International Inc., Edison, NJ, USA) in a 25 μl reaction volume containing 0.5 μM of each primer, 0.24 mM dNTP's each, 1× buffer, 1.25 U *Pfu*Ultra II Fusion HotStart DNA polymerase, and 2 μl of primary cDNA, with a 40-cycle amplification program under the following conditions: 95 °C for 20 s, 60 °C for 20 s and 72 °C for 30 s, with a first denaturation step at 95 °C for 1 min and a final extension step at 72 °C for 3 min.

Cloning and sequencing

The amplified cDNAs were adenylated and cloned into the pGEM-T easy vector according to the manufacturer's instructions and using *E. coli* DH5α as the host strain. Transformants were selected for their ability to grow on LB-agar plates with 100 μg/mL ampicillin at 37 °C. Colonies were then randomly selected and plasmids were isolated by alkaline lysis extraction. The presence of *M. carcinus* cDNAs in the plasmids was confirmed by PCR using the Pig1 and RACEAP primers, as described above for cDNA amplification but using GoTaq DNA polymerase. Positive plasmids, as determined by PCR analysis from the different colonies of *E. coli*, were sequenced using T7 and SP6 universal primers and an ABI Prism 310 sequencer (Applied Biosystems, Foster City, CA, USA) at the Molecular Biology Unit, Institute of Cellular Physiology, Universidad Nacional Autónoma de México. The sequences obtained with the T7 and SP6 primers were aligned using the Contig Assembly Program (CAP) module from the BioEdit 7.2.6 program (*Hall, 1999*). The consensus sequences generated by CAP were compared to sequences reported in the databases (non-redundant GenBank coding sequence

translations, Protein Data Bank, SwissProt, Protein Information Resource and Protein Research Foundation excluding environmental samples from whole genome shotgun projects) using the Blastx tool (*Gish & States, 1993*; *Altschul et al., 1997*) of the National Center for Biotechnology Information (Bethesda, MD, USA).

Bioinformatics analysis of *M. carcinus* trypsin-like protease

The putative signal peptide cleavage site prediction was performed with the SignalP 4.0 Server (*Petersen et al., 2011*). The deduced amino acid sequence was compared to the Pfam protein families database (*Finn et al., 2014*) and the Prosite database of annotated motif descriptors (*Sigrist et al., 2013*) for functional domain analysis, analyzed in the CATH database (*Orengo et al., 1997*; *Dawson et al., 2017*) for protein structure classification and in the MEROPS database (*Rawlings, Barrett & Bateman, 2010*) for classifying in a peptidase family. The protein model was constructed by protein threading using the Phyre2 server (*Kelley et al., 2015*) and visualized using the Swiss-PdbViewer/DeepView 4.1 (*Guex & Peitsch, 1997*), which was also used to perform a structural superposition of the protein model with the crayfish trypsin structure (PDB code: 2F91), and to determine residues that are more than 30% surface accessible. Multiple sequences were aligned with the Clustal Omega program (*Sievers et al., 2011*), which was also used to construct the dendrogram plot by the neighbor-joining method. The unrooted radial tree was visualized with MEGA 7 (*Kumar, Stecher & Tamura, 2016*). Potential glycosylation sites were predicted with the Glycosylation Predictor server (*Hamby & Hirst, 2008*). Unless otherwise stated, the percentages of sequence identity were calculated by the global Needleman–Wunsch pairwise alignments using Blast tools or by multiple sequence alignments using the Clustal Omega program. The residue position identification in the *M. carcinus* trypsin-like (Mc-TryL) protease is denoted according to the first amino acid residue of the preproenzyme. Thus, the first residue of the mature Mc-TryL protease started at position 31. The chymotrypsin-based conventional numbering system (chymo#) is also given for reference (*Hartley, 1970*).

RESULTS

3′RACE assay

The 3′RACE assay for RNA from *M. carcinus* hepatopancreas rendered mainly an 800-bp band on agarose gels. PCR analysis of the isolated pGEM-T plasmids from different colonies showed six products of 800, 754, 500, 345, 305 and 200 bp. After sequencing, all products showed the sequence of the Pig1 and RACEAP primers, however, only the 800-bp product harbored a poly(A) tail.

Bioinformatics analysis of *M. carcinus* trypsin-like protease

The longer cDNA with the poly(A) tail obtained by sequencing consisted of 736 nucleotides (GenBank accession no. MH900228). Blastx analysis showed an ORF of 627 nucleotides that encodes for 208 amino acid residues with a sequence identity of 94% with an unpublished trypsin from the Malaysian shrimp *Macrobrachium rosenbergii* (GenBank accession no. AMQ98968) and 57% sequence identity with Pacific white shrimp

(*Penaeus vannamei*) trypsin (GenBank accession no. CAA60129). The sequenced cDNA was thus assigned to a nucleotide sequence that encodes for a trypsin-like protease fragment of *M. carcinus*. A more detailed analysis showed that the sequenced cDNA fragment had an untranslated region (UTR) at the 3′ end of 91 nucleotides while the same region in *M. rosenbergii* has only 64 nucleotides. The 27 extra nucleotides of the *M. carcinus* 3′UTR shared no identity at all with the *M. rosenbergii* 3′UTR.

Since the deduced amino acid sequence had height identity with the *M. rosenbergii* trypsin (94%), we continued the trypsin-like protease sequence analysis with a hybrid sequence; that is, 16 amino acid residues corresponding to the signal peptide (predicted by the SignalP 4.0 server), 14 amino acid residues corresponding to the propeptide and the first 28 amino acid residues of mature trypsin, all from the *M. rosenbergii* trypsin sequence. This was followed by the 208 amino acid residues from the trypsin-like protease sequence identified from *M. carcinus*. The first 28 amino acid residues from mature *M. rosenbergii* trypsin would likely differ in no more than 2 residues from the first residues sequence of Mc-TryL protease. According to the Pfam protein families database, the Mc-TryL protease contains a functional domain from residue 31 to 261 of the Trypsin PF00089 family. Other proteases, including trypsins, belong to this protein family. In addition, the CATH protein structure classification database classifies *M. carcinus* trypsin as 2.40.10.10 that corresponds to a mainly beta structure with a beta-barrel architecture, and thrombin topology, belonging to the serine protease homologous superfamily. The Phyre2 server constructed a full molecular model of the mature Mc-TryL protease (Fig. 1A) using the d2f91a1 fold from the 2F91 structure, corresponding to the *A. leptodactylus* trypsin (Fodor et al., 2005). This trypsin has a 53% sequence identity with the Mc-TryL protease and Phyre2 considered as a true homolog of Mc-TryL protease with 100% confidence. The Mc-TryL protease model showed a typical trypsin fold with 14 beta-strands and 2 alpha-helixes, covering 39% and 8% of the protein, respectively, and with the characteristic four loops that have been described for crayfish trypsin (Fodor et al., 2005).

The Prosite protein domain database and comparisons with other trypsins (Fig. 2) predicted four disulfide bonds for the Mc-TryL protease at Cys60 and Cys76, Cys158 and Cys224, Cys189 and Cys203 and Cys214 and Cys242. Nevertheless, only Cys60 and Cys76 remain close enough to form a disulfide bond according to the Mc-TryL protease molecular model and its superposition with the 2F91 structure; therefore, the Mc-TryL protease structure probably has only one disulfide bond (Fig. 1A). The Glycosylation Predictor server identified 14 potential glycosylation sites (2 N-linked and 12 O-linked glycosylations), 8 of which were likely to be glycosylated since they are at the protein surface. None of these are close to the active site. The MEROPS peptidases database classifies Mc-TryL protease in the Clan PA, Subclan PA(S), Peptidase family S1 (chymotrypsin family), Subfamily A, and Subfamily type S01.035, with the holotype brachyurin-T from *Astacus astacus* in a similar manner than crustacean trypsins, but not like mammalian trypsins that are classified in the Subfamily type S01.127.

A further analysis of the amino acid sequence revealed the absence of some typical trypsin sequence motifs. Most trypsin sequences, including crustacean trypsins, start with
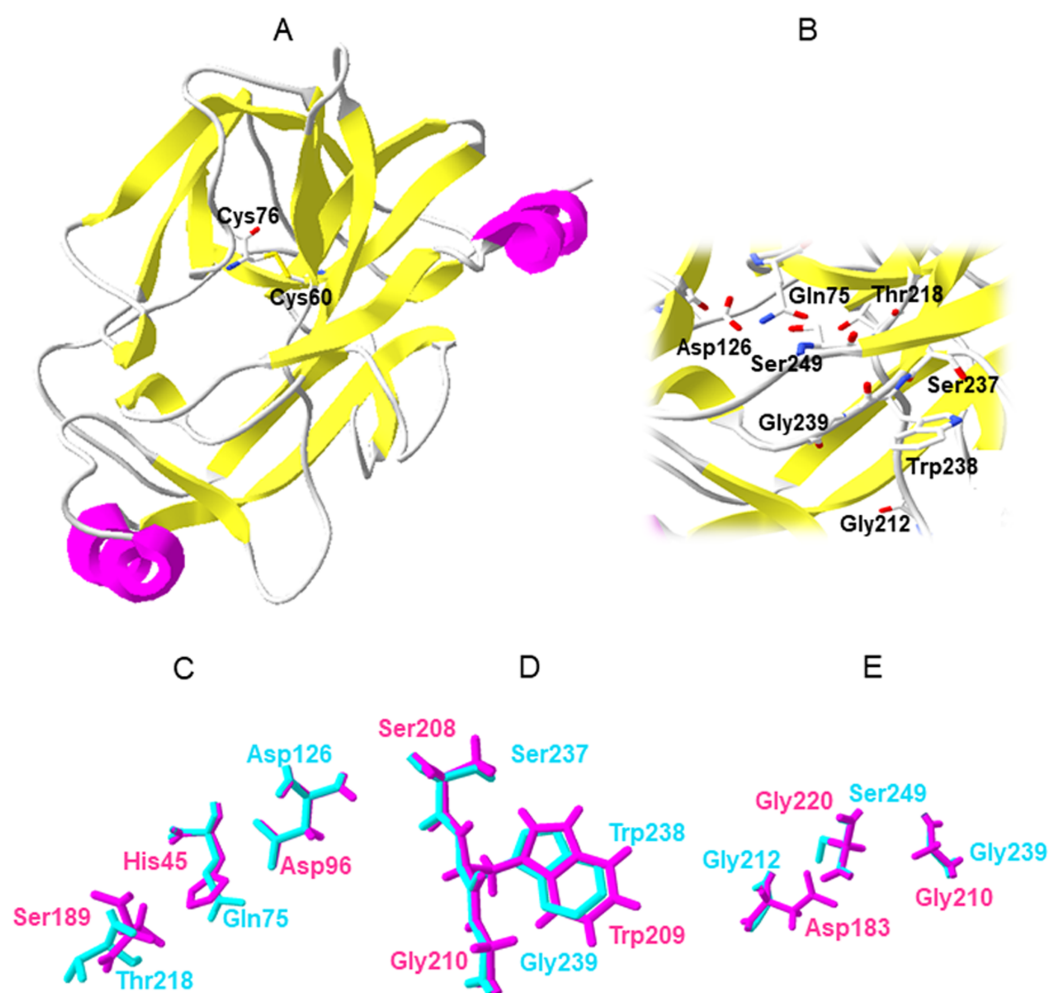
Figure 1 Molecular model of the Mc-TryL protease, and specific amino acid superposition of the Mc-TryL protease (cyan) and *A. leptodactylus* trypsin (magenta). (A) Molecular model of the Mc-TryL protease showing secondary structure elements and the disulfide bond Cys60–Cys76. (B) Enlarged view of the active site. (C) Catalytic triad. (D) Substrate binding residues. (E) Residues that confer the peptide-bond specificity. The numbers for Mc-TryL protease are according to the first amino acid residue of the preproenzyme, while for *A. leptodactylus* trypsin are based on mature protein.

the IVGG motif, whereas the Mc-TryL protease sequence starts with the IIGG motif. In addition, the multiple sequence alignment analysis (Fig. 2) showed that for most trypsin sequences, the last amino acid residue of the propeptide is a Lys or Arg, which corresponds to the recognition site for trypsins, so that trypsins are involved in the activation process from trypsinogen to trypsin (*Rypniewski et al., 1994*). Nevertheless, Mc-TryL protease contains a Gly at the end of the propeptide and an Arg at the −4 position from the trypsin N-terminus. Therefore, the Mc-TryL protease cannot activate its zymogen in the same form like most other trypsins. The multiple sequence alignment (Fig. 2), Prosite analysis and molecular model showed that at the site of the typical catalytic triad Ser/His/ Asp of serine proteases, the Mc-TryL protease contains Thr/Gln/Asp triad (218/75/126) (Figs. 1B and 1C). In addition, the multiple sequence alignment showed that the
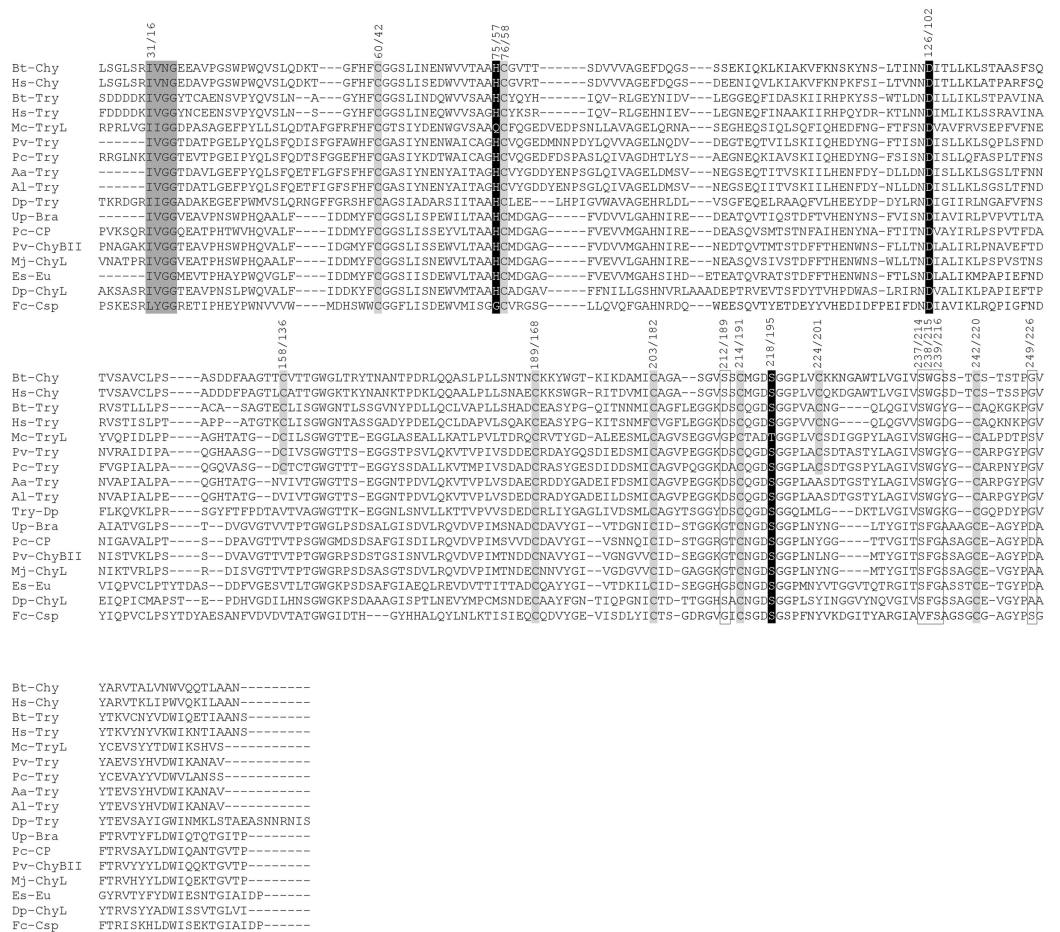
Figure 2 Multiple sequence alignment of the trypsin-like protease from *M. carcinus* (Mc-TryL), 12 mature brachyurins and 2 mammalian trypsins and 2 mammalian chymotrypsins as the reference. Grey shadows show sequence motif at the beginning of the mature protein, or conserved cysteines. The catalytic triad is denoted by black boxes, while substrate binding residues and residues that confer the peptide-bond specificity are shown by white boxes. The numbers above the sequences correspond to Mc-TryL protease position according to the first amino acid residue of the preproenzyme and second numbers are the chymotrypsin-based conventional numbering system. Five serine proteases classified by the MEROPS database as brachyurin-T: Pv-Try, trypsin from the Pacific white shrimp *Penaeus vannamei* (CAA60129); Pc-Try, trypsin from the red king crab *Paralithodes camtschaticus* (AAL67442); Aa-Try, trypsin from the noble crayfish *Astacus astacus* (P00765); Al-Try, trypsin from the narrow-clawed crayfish *Astacus leptodactylus* (Protein Data Bank (PDB) code 2F91); Dp-Try, trypsin from the water flea *Daphnia pulex* (EFX75427). Four serine proteases classified by the MEROPS database as brachyurin-C: Up-Bra, brachyurin (collagenolytic protease) from the Atlantic sand fiddler crab *Uca pugilator* (P00771), Pc-CP, collagenolytic serine protease from the red king crab *Paralithodes camtschaticus* (AAL67441), Pv-ChyBII, chymotrypsin BII from the Pacific white shrimp *Penaeus vannamei* (CAA71673), Mj-ChyL, chymotrypsin-like proteinase from the Japanese tiger prawn *Marsupenaeus japonicus* (BAI49929). Three serine proteases classified by the MEROPS database as euphauserase: Es-Eu, euphauserase from the antarctic krill *Euphausia superba* (MEROPS number MER0097318); Dp-ChyL, chymotrypsin-like protein from the water flea *Daphnia pulex* (EFX79603); Fc-Csp, collagenolytic serine protease from the Chinese white shrimp *Fenneropenaeus chinensis* (ACV97157). Mammalian trypsins and chymotrypsins: Hs-Try, trypsin from *Homo sapiens* (P07477); Bt-Try, trypsin from cattle *Bos taurus* (Q29463); Hs-Chy, chymotrypsin from *Homo sapiens* (P17538); Bt-Chy, chymotrypsin from cattle *Bos taurus* (P00766). Unless stated, GenBank accession numbers are given in parentheses.
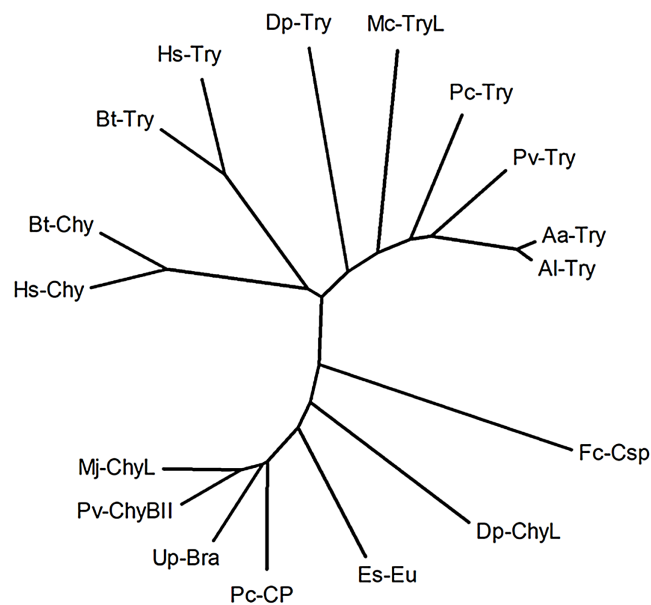
Figure 3 Neighbor-joining radial dendrogram from amino acid sequence alignment of the Mc-TryL protease, 12 mature brachyurins and two mammalian trypsins and two mammalian chymotrypsins as the reference. Abbreviations are as in Fig. 2.

substrate binding residues are Ser237, Trp238 and Gly239 (Figs. 1D and 2), and Gly212, Gly239 and Ser249 are residues that confer the peptide-bond specificity to the Mc-TryL protease (Figs. 1E and 2).

Figure 3 shows a neighbor-joining radial dendrogram from the amino acid sequence alignment of 13 mature brachyurins, including the Mc-TryL protease and two mammalian trypsins and chymotrypsins as the reference. Every brachyurin-type protease forms a cluster, with the Mc-TryL protease being close to the brachyurin-T proteases, but only with a 57% sequence identity with the more related member of the cluster.

DISCUSSION

Macrobrachium carcinus is an endemic freshwater shrimp from southeast Mexico with potential use in aquaculture. Because proteases are the most important digestive enzymes in crustacea, in the present work we describe structural and functional implications of a Mc-TryL protease.

In our initial 3′RACE experiments, we used as the gene-specific primer a 5′-primer designed from a consensus sequence from the beginning of the crustacean trypsin coding sequences. Nevertheless, a sequence coding for a peptide with high identity to reported trypsins was not amplified. Instead, four sequences were amplified and after sequencing they showed a query coverage and local identity that was greater than 72% and 58%, respectively, with crustapain from the shrimp *Pandalus borealis* (GenBank accession no. AB091669). Crustapains are papain-like cysteine proteinases from crustacea species (*Aoki, Ahsan & Watabe, 2003*). For this reason, a new consensus 5′-primer (Pig1) was designed from crustacean trypsin sequences directed to a highly conserved internal region of the mature proteins. This new 3′RACE experiment led us to deduce an amino acid

sequence of 208 residues with identity to the reported crustacean trypsins. Following our results, the trypsin sequence (cDNA and protein) of *M. rosenbergii* was reported at the GenBank database. This sequence explained why our initial 3′RACE design for fishing the *M. carcinus* trypsin mRNA failed, since the 5′ end of the *M. rosenbergii* trypsin cDNA has only 57% identity with the *P. vannamei* trypsin cDNA. Furthermore, the 3′ end of the consensus primer from the beginning of the crustacean trypsin coding sequences lacked complementarity at the position that would hybridize to *M. rosenbergii* trypsin cDNA sequence. With the amino acid sequence (208 residues) obtained in the present work and the *M. rosenbergii* trypsin sequence reported in the GenBank database, the most probable sequence of the Mc-TryL protease can be inferred.

The bioinformatics analyses showed that the deduced amino acid sequence likely corresponds to a trypsin-like protease that is closely related to brachyurins, which are a subset of serine proteases with collagenolytic activity found in crabs and other crustacea. These proteases possess a specificity that resembles a combination of trypsins, chymotrypsins and elastases, with activity (type Ia or brachyurin-C) or limited activity (type Ib or euphauserase) towards Arg substrates or with a strictly trypsin-like protease specificity (type II or brachyurins-T) (*Rudenskaya, Kislitsin & Rebrikov, 2004*; *Page & Craik, 2013*). The Mc-TryL protease sequence showed a global sequence identity of only 57% with the *P. vannamei* trypsin and a lower identity with other crustacean trypsins, which places it in the dendrogram somewhat separated from the group of crustacean trypsins. A detailed analysis of the amino acid sequence revealed specific differences from crustacean trypsins, such as the sequence motif at the beginning of the mature protein, activation mechanism of the corresponding zymogen, amino acid residues of the catalytic triad and residues responsible for substrate specificity.

Although most of the trypsin sequences start with the IVGG motif, including a low molecular weight trypsin from the hepatopancreas of *M. rosenbergii* (*Sriket et al., 2012*), the IIGG motif is also present in other trypsins, with the sequence of *Daphnia pulex* trypsin (GenBank accession no. EFX75427) being the most related to the Mc-TryL protease with the IIGG motif (44% sequence identity). After cleavage of the propeptide for trypsinogen activation, the N-terminus of trypsins is buried in the C-terminal domain forming an ion pair with Asp194 (chymo#) making several hydrogen bonds (*Rypniewski et al., 1994*). Since the V to I change is conservative, both N-terminal sequences (IVGG and IIGG) would likely have the same function. Furthermore, the multiple sequence alignment of Mc-TryL protease with other trypsin sequences (Fig. 2) showed that Mc-TryL protease contains Asp217, the equivalent position 194-chymo#.

As other crustacean trypsins, Mc-TryL protease does not contain a tetra-Asp sequence in the propeptide, neither Asp217 (chymo#), as containing other most vertebrate trypsinogens. This tetra-Asp sequence together with Asp217 (chymo#) has been involved in an inhibitory function for trypsin-mediated trypsinogen activation (autoactivation), by an electrostatic repulsion between Asp217 (chymo#) and the tetra-aspartate (*Nemoda & Sahin-Tóth, 2007*). Therefore, the corresponding trypsinogen of the Mc-TryL protease likely has an increased autoactivation rate in comparison with other non-crustacean trypsins in a similar manner as we have recently described for the *P. vannamei* trypsinogen

(*Guerrero-Olazarán et al., 2019*), although with slight differences due to the lack of Arg or Lys at the end of the propeptide. Since the storage of active trypsin is a risk of tissue damage through hydrolysis of autologous protein, *M. carcinus* must possess a tightly regulated mechanism for controlling activation until needed for a precise function similar to other crustaceans. The control of trypsin activity in crustacea is a fine-tuning mechanism that involves the continuous regulation of trypsin activity through zymogen storage, secretion and activation like frequent feeder species (*Sainz et al., 2004*; *Guerrero-Olazarán et al., 2019*), with a main role of trypsin inhibitors for controlling trypsin activity and/or trypsinogen activation (*García-Carreño, Carrillo & Navarrete del Toro, 1999*; *De Albuquerque-Cavalcanti, García-Carreño & Navarrete del Toro, 2002*; *Guerrero-Olazarán et al., 2019*).

Other trypsins that are without an Arg or Lys at the end of the propeptide have already been described. For instance, *Musca domestica* (housefly) and *Aedes albopictus* (tiger mosquito) trypsins contain an Arg or Lys, respectively, at the −4 position from the trypsin N-terminus, like the Mc-TryL protease. For this type of propeptide, an activation mechanism through cleavage of the carboxy terminal of the Arg or Lys, followed by removal of some amino acid residues by an aminopeptidase has been initially suggested. Now, however, a propeptide cleavage by the action of a cathepsin has been proposed (*Le et al., 2011*). For the *M. carcinus* trypsin, the cathepsin involved in the zymogen activation could be cathepsin B since this protease mainly cleaves a peptide bond at the carboxy side of a Gly, if another Gly is located at the +4 position from the protease N-terminus (*Rawlings, Barrett & Bateman, 2010*), as happens with the Mc-TryL protease sequence.

Variations of the catalytic Ser/His/Asp triad configuration have been described for unconventional serine proteases (*Ekici, Paetzel & Dalbey, 2008*); however, the Mc-TryL protease contains Thr/Gln/Asp as the catalytic triad, which has not been previously described. Although Thr is as reactive as a nucleophile as Ser, Thr is rarely used in a similar functional context as is Ser. Nevertheless, a Ser195Thr (chymo#) substitution in human thrombin serine protease reduces but does not nullify activity (*Pelc et al., 2015*). This was explained by biases Thr mobility within the active site and stabilization of rotamers incompatible with substrate binding due to the additional methyl group on the side chain (*Pelc et al., 2015*). In any case, this substitution generated Thr/His/Asp triad still different from that of Mc-TryL protease, where His was changed by Gln, which has more mobility than His, furthermore, clash analysis by Phyre2 showed no clashes for Thr218. In the Mc-TryL protease catalytic triad, Thr would be the nucleophile, Gln would be the base catalyst through the imidic acid tautomer and Asp would help orient the Gln residue and be the base that catalyzes the Gln tautomerization via N-deprotonation. Amide-imidic acid tautomerization has been proposed for other enzyme mechanism (*Nakamura et al., 2015*; *Grigorenko, Khrenova & Nemukhin, 2018*). Thus, the sequence identified as the Mc-TryL protease is actually not a serine protease, but instead, it is a threonine protease. To the best of our knowledge, the sequence identified as the Mc-TryL protease is the first protease described with a trypsin domain with a Thr/Gln/Asp catalytic triad.

In serine proteases, the presence of Ser214, Trp215 and Gly216 (chymo#) are critical for efficient substrate hydrolysis (*Hedstrom, 2002*) since they form hydrogen bonds with P1 or P3 residues, Schechter and Berger nomenclature (*Schechter & Berger, 1968*). The Mc-TryL protease contains the same residues at equivalent positions (237, 238 and 239) ensuring the efficient protein binding of the substrate to the Mc-TryL protease.

The specificity of serine proteases is usually defined by the residues at positions 189, 216 and 226 (chymo#), which are in the S1 pocket, Schechter and Berger nomenclature (*Schechter & Berger, 1968*). This pocket is adjacent to Ser195 (chymo#) and is formed by residues 189–192, 214–216 and 224–228 (chymo#) (*Hedstrom, 2002*). For example, Ser189, Gly216 and Gly226 for chymotrypsins, Asp189, Gly216 and Gly226 for trypsins and Val190, Val216 and Thr226 for elastases (chymo#). In the case of the Mc-TryL protease, Gly, Gly, Ser are at these conserved positions (212, 239 and 249, respectively). Substitutions involving Gly189 (chymo#) have been described for chymotrypsin from the fire ant (*Solenopsis invicta*) (*Botos et al., 2000*), and for the collagenolytic protease from the hepatopancreas of crabs (*Uca pugilator* and *Paralithodes camtschaticus*) (*Grant et al., 1980*; *Rudenskaya, Kislitsin & Rebrikov, 2004*). Nevertheless, these proteases contain an Asp at position 226 (chymo#) resembling a 189 to 226 (chymo#) switch from the trypsin specificity site, while the Mc-TryL protease contains a Ser at this position, resembling a switch from the chymotrypsin specificity. The Gly189, Gly216 and Ser226 (chymo#) pattern has been previously described for unknown substrate specificity serine proteases by a bioinformatic analysis of putative serine proteases from the *Daphnia pulex* genome, the first crustacean genome that was sequenced (*Julien, 2014*).

The specific differences between the Mc-TryL protease sequence and crustacea trypsins approximate the Mc-TryL protease to crustacean collagenolytic proteases of type Ia and Ib brachyurins. The differences in the catalytic triad and specificity of residues likely confer a broad substrate specificity (including collagen) to the Mc-TryL protease. The ability to hydrolyze collagen reflects the diets of crustaceans, which include animal detritus that would contain a large amount of collagen. The global structure of trypsin (type II brachyurins) and the specific properties of type I brachyurins would place the Mc-TryL protease among these groups of proteases. Interestingly, *D. pulex* trypsin is also separated from the crustacea trypsins in the dendrogram.

Moreover, the high number of potential glycosylation sites, none of which close to the active site, could provide stability to the Mc-TryL protease, even stability from proteolysis by autolysis or by the action of other proteases since the oligosaccharides that are bound to this protease may be able to protect the proteolysis susceptible sites by steric hindrance.

CONCLUSIONS

We identified a putative trypsin-like protease cDNA from *M. carcinus* hepatopancreas, deduced the amino acid sequence and described its structure and function relationship. The bioinformatics analyses showed a mainly beta structure for the Mc-TryL protease with a beta-barrel architecture and thrombin topology, belonging to the serine protease homologous superfamily. The amino acid sequence is closely related to brachyurins, but somewhat separated from other crustacean trypsins. The Mc-TryL protease sequence

has important specific differences with this group, such as the sequence motif at the beginning of the mature protein, activation mechanism of the corresponding zymogen, amino acid residues of the catalytic triad and residues responsible for substrate specificity. Although eight serine proteases have recently been detected in the hepatopancreas of *M. carcinus* (*Manríquez-Santos et al., 2018*), the specificity of the Mc-TryL protease described in this work, which is also present in *M. rosenbergii*, may be broader than what is currently believed for other brachyurins. Therefore, further studies are needed to better understand the digestive physiology in this species.

ACKNOWLEDGEMENTS

We thank Glen D. Wheeler for his stylistic suggestions in the preparation of the manuscript.

Funding

The authors received no funding for this work.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- José M. Viader-Salvadó conceived and designed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- José Alberto Aguilar Briseño performed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.
- Juan A. Gallegos-López analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- José A. Fuentes-Garibay performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Carlos Alfonso Alvarez-González conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Martha Guerrero-Olazarán conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the paper, and approved the final draft.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

Trypsin-like protease cDNA (MH900228) and amino acid sequences from *Macrobrachium carcinus* (AYN64566) are available at GenBank.

Data Availability

The raw data is available at GenBank and in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.9030#supplemental-information.

REFERENCES

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25(17)**:3389–3402 DOI 10.1093/nar/25.17.3389.

Aoki H, Ahsan MN, Watabe S. 2003. Molecular cloning and functional characterization of crustapain: a distinct cysteine proteinase with unique substrate specificity from northern shrimp *Pandalus borealis*. *Journal of Biochemistry* **133(6)**:799–810 DOI 10.1093/jb/mvg102.

Botos I, Meyer E, Nguyen M, Swanson SM, Koomen JM, Russell DH, Meyer EF. 2000. The structure of an insect chymotrypsin. *Journal of Molecular Biology* **298(5)**:895–901 DOI 10.1006/jmbi.2000.3699.

Chen JM, Kukor Z, Le Maréchal C, Tóth M, Tsakiris L, Raguénès O, Férec C, Sahin-Tóth M. 2003. Evolution of trypsinogen activation peptides. *Molecular Biology and Evolution* **20(11)**:1767–1777 DOI 10.1093/molbev/msg183.

Dawson NL, Lewis TE, Das S, Lees JG, Lee D, Ashford P, Orengo CA, Sillitoe I. 2017. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research* **45(D1)**:D289–D295 DOI 10.1093/nar/gkw1098.

De Albuquerque-Cavalcanti C, García-Carreño FL, Del Toro MA. 2002. Trypsin and trypsin inhibitors from penaeid shrimp. *Journal of Food Biochemistry* **26(3)**:233–251 DOI 10.1111/j.1745-4514.2002.tb00854.x.

Ekici ÖD, Paetzel M, Dalbey R. 2008. Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. *Protein Science* **17(12)**:2023–2037 DOI 10.1110/ps.035436.108.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. *Nucleic Acids Research* **42(D1)**:D222–D230 DOI 10.1093/nar/gkt1223.

Fodor K, Harmat V, Hetenyi C, Kardos J, Antal J, Perczel A, Patthy A, Katona G, Graf L. 2005. Extended intermolecular interactions in a serine proteinase-canonical inhibitor complex account for strong and highly specific inhibition. *Journal of Molecular Biology* **350(1)**:156–169 DOI 10.1016/j.jmb.2005.04.039.

Frohman MA, Dush MK, Martin GR. 1988. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of the National Academy of Sciences of the United States of America* **85(23)**:8998–9002 DOI 10.1073/pnas.85.23.8998.

García-Carreño FL, Carrillo O, Navarrete del Toro MA. 1999. Control of digestive functions in shrimp: I. Inhibitor of measured trypsin activity in the hepatopancreas. In: Schram FR, Von Vaupelklein JC, eds. *Crustaceans and the Biodiversity Crisis*. Amsterdam: Proceedings of the Fourth International Crustacean Congress, 915–922.

Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. *Nature Genetics* **3(3)**:266–272 DOI 10.1038/ng0393-266.

Grant GA, Henderson KO, Eisen AZ, Bradshaw RA. 1980. Amino acid sequence of a collagenolytic protease from the hepatopancreas of the fiddler crab, *Uca pugilator*. *Biochemistry* **19(20)**:4653–4659 DOI 10.1021/bi00561a018.

Grigorenko BL, Khrenova MG, Nemukhin AV. 2018. Amide–imide tautomerization in the glutamine side chain in enzymatic and photochemical reactions in proteins. *Physical Chemistry Chemical Physics* **20(37)**:23827–23836 DOI 10.1039/C8CP04817G.

Guerrero-Olazarán M, Castillo-Galván M, Gallegos-López JA, Fuentes-Garibay JA, Viader-Salvadó JM. 2019. Biochemical characterization of recombinant *Penaeus vannamei* trypsinogen. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **238**:110337 DOI 10.1016/j.cbpb.2019.110337.

Guex N, Peitsch MC. 1997. SWISS-MODEL and the Swiss-Pdb viewer: an environment for comparative protein modeling. *Electrophoresis* **18(15)**:2714–2723 DOI 10.1002/elps.1150181505.

Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/NT. *Nucleic Acids Symposium Series* **41**:95–98.

Hamby SE, Hirst JD. 2008. Prediction of glycosylation sites using random forests. *BMC Bioinformatics* **9(1)**:500 DOI 10.1186/1471-2105-9-500.

Hartley BS. 1970. Homologies in serine proteinases. *Philosophical Transactions of the Royal Society B* **257**:77–87.

Hedstrom L. 2002. Serine protease mechanism and specificity. *Chemical Reviews* **102(12)**:4501–4524 DOI 10.1021/cr000033x.

Julien MO. 2014. Understanding the evolutionary history of biochemical innovation. Unpublished Master's thesis, College of Arts and Sciences, University of South Carolina, Columbia, SC.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10(6)**:845–858 DOI 10.1038/nprot.2015.053.

Klein B, Le Moullac G, Sellos D, Van Wormhoudt A. 1996. Molecular cloning and sequencing of trypsin cDNAs from *Penaeus vannamei* (Crustacea, Decapoda): use in assessing gene expression during the moult cycle. *The International Journal of Biochemistry & Cell Biology* **28(5)**:551–563 DOI 10.1016/1357-2725(95)00169-7.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33(7)**:1870–1874 DOI 10.1093/molbev/msw054.

Le QT, Min H-K, Xia H-Z, Fukuoka Y, Katunuma N, Schwartz LB. 2011. Promiscuous processing of human α/β-Protryptases by cathepsins L, B, and C. *Journal of Immunology* **186(12)**:7136–7143 DOI 10.4049/jimmunol.1001804.

Linton SM, Saborowski R, Shirley AJ, Penny JA. 2014. Digestive enzymes of two brachyuran and two anomuran land crabs from Christmas Island, Indian Ocean. *Journal of Comparative Physiology B-biochemical Systemic and Environmental Physiology* **184(4)**:449–468 DOI 10.1007/s00360-014-0815-2.

Manríquez-Santos TDJ, Álvarez-González CA, Peña E, Camarillo-Coop S, Martínez-García R, Vega-Villasante F. 2018. Partial characterization of digestive proteases in adults of bigclaw river shrimp *Macrobrachium carcinus*. *Latin American Journal of Aquatic Research* **46(3)**:525–533 DOI 10.3856/vol46-issue3-fulltext-5.

Muhlia-Almazán AM, Sánchez-Paz A, García-Carreño FL. 2008. Invertebrate trypsins: a review. *Journal of Comparative Physiology B* **178**:665–672.

Nakamura A, Ishida T, Kusaka K, Yamada T, Fushinobu S, Tanaka I, Kaneko S, Ohta K, Tanaka H, Inaka K, Higuchi Y, Niimura N, Samejima M, Igarashi K. 2015. Newton's cradle proton relay with amide–imidic acid tautomerization in inverting cellulase visualized by neutron crystallography. *Science Advances* **1(7)**:e1500263 DOI 10.1126/sciadv.1500263.

Nemoda Z, Sahin-Tóth M. 2007. The tetra-aspartate motif in the activation peptide of human cationic trypsinogen is essential for autoactivation control but not for enteropeptidase recognition. *Journal of Biological Chemistry* **280(33)**:29645–29652 DOI 10.1074/jbc.M505661200.

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH—a hierarchic classification of protein domain structures. *Structure* **5(8)**:1093–1109 DOI 10.1016/S0969-2126(97)00260-8.

Page MJ, Craik CS. 2013. Chapter 669: Brachyurins. In: Rawlings ND, Salvesen G, eds. *Handbook of Proteolytic Enzymes.* Vol. 3. Third Edition. London: Academic Press, 3049–3052.

Pelc LA, Chen Z, Gohara DW, Vogt AD, Pozzi N, Di Cera E. 2015. Why Ser and not Thr brokers catalysis in the trypsin fold. *Biochemistry* **54(7)**:1457–1464 DOI 10.1021/acs.biochem.5b00014.

Perera E, Rodriguez-Viera L, Perdomo-Morales R, Montero-Alejo V, Moyano FJ, Martinez-Rodriguez G, Mancera JM. 2015. Trypsin isozymes in the lobster *Panulirus argus* (Latreille, 1804): from molecules to physiology. *Journal of Comparative Physiology B* **185(1)**:17–35 DOI 10.1007/s00360-014-0851-y.

Petersen TN, Brunak S, Von Heijen G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8(10)**:785–786 DOI 10.1038/nmeth.1701.

Rawlings ND, Barrett AJ, Bateman A. 2010. MEROPS: the peptidase database. *Nucleic Acids Research* **38(Suppl._1)**:D227–D233 DOI 10.1093/nar/gkp971.

Rudenskaya GN, Kislitsin YA, Rebrikov DV. 2004. Collagenolytic serine protease PC and trypsin PC from king crab *Paralithodes camtschaticus*: cDNA cloning and primary structure of the enzymes. *BMC Structural Biology* **4(1)**:2 DOI 10.1186/1472-6807-4-2.

Rypniewski WR, Perrakis A, Vorgias CE, Wilson KS. 1994. Evolutionary divergence and conservation of trypsin. *Protein Engineering, Design and Selection* **7(1)**:57–64 DOI 10.1093/protein/7.1.57.

Sainz JC, García-Carreño F, Sierra-Beltrán A, Hernández-Cortés P. 2004. Trypsin synthesis and storage as zymogen in the midgut gland of the shrimp Litopenaeus vannamei. *Journal of Crustacean Biology* **24(2)**:266–273 DOI 10.1651/C-2423.

Schechter I, Berger A. 1968. On the active site of proteases. III. Mapping the active site of papain; specific peptide inhibitors of papain. *Biochemical and Biophysical Research Communications* **32(5)**:898–902 DOI 10.1016/0006-291X(68)90326-4.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgiens DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Molecular Systems Biology* **7(1)**:539 DOI 10.1038/msb.2011.75.

Sigrist CJ, De Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. 2013. New and continuing developments at PROSITE. *Nucleic Acids Research* **41(D1)**:D344–D347 DOI 10.1093/nar/gks1067.

Sriket C, Benjakul S, Visessanguan W, Hara K, Yoshida A, Liang X. 2012. Low molecular weight trypsin from hepatopancreas of freshwater prawn (*Macrobrachium rosenbergii*): characteristics and biochemical properties. *Food Chemistry* **134(1)**:351–358 DOI 10.1016/j.foodchem.2012.02.173.