FELIPE JUN FUZITA

MOLECULAR PHYSIOLOGY OF DIGESTION IN ARACHNIDA: FUNCTIONAL AND COMPARATIVE-EVOLUTIONARY APPROACHES

Thesis presented to the Programa de Pós-Graduação Interunidades em Biotecnologia USP/Instituto Butantan/IPT, to obtain the Title of Doctor in Biotechnology.

São Paulo
2014

FELIPE JUN FUZITA

MOLECULAR PHYSIOLOGY OF DIGESTION IN ARACHNIDA: FUNCTIONAL AND COMPARATIVE-EVOLUTIONARY APPROACHES

Thesis presented to the Programa de Pós-Graduação Interunidades em Biotecnologia USP/Instituto Butantan/IPT, to obtain the Title of Doctor in Biotechnology.

Concentration area: Biotechnology

Advisor: Dr. Adriana Rios Lopes Rocha

Corrected version. The original electronic version is available either in the library of the Institute of Biomedical Sciences and in the Digital Library of Theses and Dissertations of the University of Sao Paulo (BDTD).

São Paulo
2014

DADOS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
Serviço de Biblioteca e Informação Biomédica do
Instituto de Ciências Biomédicas da Universidade de São Paulo

Fuzita, Felipe Jun.
Molecular physiology of digestion in Arachnida: functional and comparative-evolutionary approaches / Felipe Jun Fuzita. -- São Paulo, 2014.

Orientador: Profa. Dra. Adriana Rios Lopes Rocha.

Tese (Doutorado) – Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação Interunidades em Biotecnologia USP/IPT/Instituto Butantan. Área de concentração: Biotecnologia. Linha de pesquisa: Bioquímica, biologia molecular, espectrometria de massa.

Versão do título para o português: Fisiologia molecular da digestão em Arachnida: abordagens funcional e comparativo-evolutiva.

1. Digestão 2. Aranha 3. Escorpião 4. Enzimologia 5. Proteoma 6. Transcriptoma I. Rocha, Profa. Dra. Adriana Rios Lopes I. Universidade de São Paulo. Instituto de Ciências Biomédicas. Programa de Pós-Graduação Interunidades em Biotecnologia USP/IPT/Instituto Butantan III. Título.

ICB/SBIB07/2014

UNIVERSIDADE DE SÃO PAULO
Programa de Pós-Graduação Interunidades em Biotecnologia
Universidade de São Paulo, Instituto Butantan, Instituto de Pesquisas Tecnológicas

---

Candidato(a): Felipe Jun Fuzita.

Título da Tese: Molecular physiology of digestion in Arachnida: functional and comparative-evolutionary approaches.

Orientador(a): Profa. Dra. Adriana Rios Lopes Rocha.

A Comissão Julgadora dos trabalhos de Defesa da Tese de Doutorado, em sessão pública realizada a ...../...../....., considerou

( ) Aprovado(a) **( ) Reprovado(a)**

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Examinador(a): Assinatura: .....
Nome: .....
Instituição: .....

Presidente: Assinatura: .....
Nome: .....
Instituição: .....

À memoria do meu pai, Ernesto Junji Fuzita, que desde cedo me ensinou a importância de ser útil ao próximo e à sociedade

ACKNOWLEDGMENTS

I am grateful to many people who helped me to produce this thesis:

My mother Sonia and my sister Yumi who provided me the basis for becoming a scientist in many aspects. Together we could overcome the natural obstacles of life.

My girlfriend Anna Lindinha for supporting me in the last years of this doctorate.

My advisor, Dr. Adriana Lopes, which is an open-minded person who not only taught me the laboratory work but also heard my opinions about how could we test our hypothesis.

Carlos, Nati, Antônio, Ciça, Pedro, Aline, Rodrigo, Carla, Angela and recently Mariana. It is very nice to be part of a united research group that it is worried about helping each other. Also for the nice moments at SBBq.

The Analytical Biotechnology Group from the Delft University of Technology at The Netherlands for receiving me as an integrant of their group. Dr. Peter Verhaert for giving to me the opportunity of working there and Dr. Martijn Pinkse for daily assisting me with the sample analysis. Yuanjie Yu for the chats, technical support, meeting company and coffee-breaks. Mervin Pieterse and Johan for the nice chats and technical support. Also Matthijs for the bioinformatics support and good talking. I felt glad for being part of a group with such nice people.

All the researches from the Laboratório de Bioquímica e Biofísica from Instituto Butantan, Drs. Marilene Demasi, Daniel Pimenta, Ivo Lebrun, Ana Chudzinski-Tavassi, Ana Olívia de Souza, Isabel Correia, Janice Onuki, Fernanda Faria, Luziane Chaguri and Sonia Aparecida Andrade. In many aspects, including equipment usage and protocol discussion, these researchers contributed a lot for my thesis.

Also the employees for technical support and funny moments, Pati, Silvana, Toninha, Bia, Marcia, Valdeli, Rafael, Júlio, João and Heleusa.

Drs. Walter Terra and Clélia Ferreira from the Departamento de Bioquímica at the Universidade de São Paulo for opening the doors of their laboratory always that I needed. Their help was fundamental for developing this work. To the students from this group for assisting me with the equipment usage and for the friendship, Dani, Thaís, Marcelo, Vanvan, André, Tici, Nati, Pererê, Val, Rebola and Fabi. Also to the

employees Cris, Nilde, Giliard and Luísa for the technical support and good conversation.

Dr Pedro Ismael from Instituto Butantan for equipment usage and nice talks about arachnid biology.

Dr. Rogério Bertani and Dr. Irene Knysak from Instituto Butantan for providing me with *Tityus serrulatus*.

All my friends from the biology bachelor course of Universidade de São Paulo. At the student´s “living space” I could meet all kind of different people and opinions, in discussions ranging from biology or musical taste to politics and religion. I can say that I learned a lot of cultural things there.

CAPES, CNPq (process number 237706/2012-1) and FAPESP for the financial support.

“I believe it is worthwhile trying to discover more about the world, even if this only teaches us how little we know.”

Karl Popper

ABSTRACT

Fuzita FJ. Molecular physiology of digestion in Arachnida: functional and comparative-evolutionary approaches. [Ph. D. thesis (Biotechnology)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2014.

Spiders and scorpions are efficient predator arachnids which can consume preys larger than themselves. This is only feasible due to a particular mechanism which combines extra-oral and intracellular digestion associated to extremely branched midgut glands and efficient digestive enzymes. Chelicerata is a basal arthropod group and scorpions passed to few changes during millions of years. Thus scorpions and other Chelicerata, such as spiders, are interesting animals for comparative-evolutionary studies. Few studies reported the molecular mechanisms of digestion in predator arachnids and until now there wasn´t a single complete sequence of digestive enzymes from this group of animals in the public databases. Hence, this work describes a biochemical, transcriptomic and proteomic characterization of the midgut and midgut glands (MMG) from the spider *Nephilengys cruentata* and from the scorpion *Tityus serrulatus.* Animals under two different physiological conditions, fasting and fed, were dissected and their MMG removed, homogenized and used as sample source. The digestive juice (DJ) of spiders was also collected and analyzed. Enzymatic activity assays focused on endopeptidases identification using different substrates and inhibitors and purification of some peptidases involved in protein digestion in scorpions and spiders. A transcriptome followed by a shotgun proteome was performed using the MMG of fasting and fed animals as well as the spider digestive juice proteome. The enzymological data showed, for the first time, that predator arachnids possess cysteine cathepsins (L, B and F) as digestive enzymes acting mainly intracellularly under acidic conditions and can be found as zymogens, which can be activated *in vitro* after acidification. In the proteomic label-free quantitative analysis cathepsin L is one of the most abundant enzymes in the MMG of both animals, but cathepsins B, D and F (only in the scorpion), legumain, trypsins, astacins, carbohydrases and lipases were also identified. Astacins are very representative in the spider DJ with 26 different isoforms identified, but trypsins (CUB and LDL domain-containing) are also frequent (9 isoforms). Other enzymes like carboxypeptidase B, alpha-amylase and triacylglycerol lipase are also present in DJ. Quantitatively, chitotriosidase constitutes about one fifth of the digestive enzymes composition in the DJ. Proteins so far exclusively associated to the venom glands, as peptide isomerase and ctenitoxins are expressed and translated in the MMG, being part of the DJ composition, indicating that these genes in the venom glands could be evolutionary originated from a MMG gene. In summary, predator arachnids relies in a multiproteolytic system mainly constituted of astacins ad trypsins for prey liquefying and cathepsin L for intracellular digestion. Carbohydrases and lipases are also present. Evolutionarily, many gene duplication events led to a large diversity of astacins present in *Nephilengys cruentata* in contrast to *Tityus serrulatus.* On the other hand, cathepsin L sequences are more conserved through Arachnida. Moreover, the feeding habits are intimately correlated with the evolutionary history of the digestive peptidases since blood-sucker arachnids as ticks presented more divergent cathepsin sequences from predator arachnids.

Keywords: Digestion. Spider. Scorpion. Enzymology. Proteome. Transcriptome.

RESUMO

Fuzita FJ. Fisiologia molecular da digestão em Arachnida: abordagens funcional e comparativo-evolutiva. [Ph. D. thesis (Biotechnology)]. São Paulo: Instituto de Ciências Biomédicas, Universidade de São Paulo; 2014.

Aranhas e escorpiões são eficientes aracnídeos predadores que podem consumir presas maiores que eles mesmos. Isso é possível devido a um mecanismo que combina digestão extraoral e intracelular associado às glândulas digestivas extremamente ramificadas e enzimas digestivas eficazes. Chelicerata é um grupo basal dentre os artrópodes sendo que escorpiões passaram por poucas mudanças durante milhões de anos, o que os torna bons animais para estudos comparativo-evolutivos assim como outros grupos de Chelicerata, por exemplo as aranhas. Poucos estudos reportaram a fisiologia molecular da digestão em aracnídeos predadores e até o presente momento não há sequências de enzimas digestivas deste grupo de animais disponíveis nos bancos de dados públicos. Portanto, este trabalho apresenta uma caracterização bioquímica, transcriptômica e proteômica do intestino e glândulas digestivas (IGD) da aranha *Nephilengys cruentata* e do escorpião *Tityus serrulatus*. Animais sob duas condições fisiológicas diferentes, jejum e alimentados, foram dissecados e seu IGD removido, homogeneizado e utilizado como amostra biológica. O suco digestivo (SD) da aranha também foi coletado e analisado bioquimicamente. Os ensaios bioquímicos concentraram-se na caracterização de endopeptidases com o uso de diferentes substratos e inibidores e com a purificação de algumas enzimas envolvidas na digestão de proteínas. Um transcriptoma seguido de um proteoma foram realizados utilizando-se o IGD de animais em jejum e alimentados assim como o SD da aranha. Os dados enzimológicos mostram, pela primeira vez, que aracnídeos predadores possuem cisteíno-catepsinas (L, B e F) digestivas atuando em condições ácidas e na forma de zimógeno, o qual pode ser ativado *in vitro* após acidificação. Na análise proteômica quantitativa sem marcação catepsina L é uma das enzimas mais abundantes no IGD, mas também são identificadas catepsinas B, D e F (somente em escorpião), legumaína, tripsinas, astacinas, carboidrases e lipases. Astacinas são bastante representadas no SD com 26 isoformas identificadas, mas tripsinas (contendo os domínios CUB e LDL) com 9 isoformas também são frequentes. Outras enzimas como carboxipeptidase B, alfa-amilase e triacilglicerol lipase também estão presentes nessa secreção. Quantitativamente quiitotriosidase constitui um quinto da composição das enzimas digestivas no SD. Proteínas até o momento exclusivamente associadas com as glândulas de veneno, como peptídeo isomerase e ctenitoxinas são expressas e traduzidas no IGD, sendo parte da composição do SD, indicando que os genes da glândula de veneno podem ao longo da evolução serem originários do IGD. Em síntese, aracnídeos predadores contam com um sistema multi-proteolítico constituído principalmente de astacinas e tripsinas para liquefação da presa extracelularmente e catepsina L para digestão intracelular. Carboidrases e lipases também estão presentes. Evolutivamente, muitos eventos de duplicação de genes levaram à grande diversidade de astacinas presente in *Nephilengys cruentata* em contraste com *Tityus serrulatus*. Por outro lado, as sequências de catepsina L são mais conservadas e menos diversificadas nestes aracnídeos. Ademais, os hábitos alimentares estão intrinsecamente associados com a história evolutiva das peptidases digestivas uma vez que aracnídeos que se

alimentam de sangue como carrapatos mostraram sequências de catepsinas mais divergentes do que os aracnídeos predadores.

Palavras-chave: Digestão. Aranha. Escorpião. Enzimologia. Proteoma. Transcriptoma.

ABBREVIATION LIST

Abz - *ortho*-Aminobenzoic acid

bp – base pairs

CA-074 - (L-3-*trans*-(Propylcarbamyl)oxirane-2-carbonyl)-L-isoleucyl-L-proline

cDNA – complementary DNA

DJ – digestive juice

Dnp - 2,4-dinitrophenyl

DMSO - dimethyl sulfoxide

EDDnp – (2,4-dinitrophenyl)ethylenediamine

E-64 - trans-epoxysuccinyl-l-leucyl-amido (4-guanidino butane)

EDTA - ethylenediaminetetracetic acid

IGD – intestino e glândulas digestivas

MCA - 7-methyl-coumarin amide

MMG – midgut and midgut glands

NSC – normalized spectra counting

PMSF - phenylmethanesulfonyl fluoride

Suc - succinyl

Tris - tris(hydroxymethyl)aminomethane)

U – International unit of enzymatic activity

Z - carbobenzoxy

LIST OF FIGURES

LIST OF TABLES

CONTENTS

CHAPTER 1 – GENERAL INTRODUCTION

1.1 General considerations about the digestive process

Locating, selecting and capturing of food, its ingestion, and the subsequent digestion and assimilation of nutrients, are essential aspects of 'Life', and, therefore, common and fundamental to all living organisms. This holds true for life forms at the very base of the evolutionary scale, like Bacteria and heterotrophic Protista, and remains so throughout the entire Metazoa group with all its diversity.

Within this general biological phenomenon of nutrient processing, digestion is an indispensable process to homeostasis. Hydrolytic enzymes are employed to reduce the complexity of macromolecules from food into biochemically simpler monomers, which will subsequently be recycled as biochemical building blocks. As such maintenance and/or development of a particular organism are ensured, and metabolic reactions can be fueled, including those for energy generation.

Digestion can be extracellular, i.e. occurring outside of the cell(s), such as in a typical metazoan intestine ('gut chamber'). A specific type of extracellular animal digestion is extracorporeal or extra-oral digestion (EOD), and this refers to digestion which takes place outside the organism's body (1, 2). Additionally and/or alternatively, part of the digestion happens intracellularly, i.e. inside the cell.

Different strategies are employed by heterotrophic organisms to capture food. Animals can be classified according to their alimentary habits as carnivore, herbivore or omnivore. Among the carnivore predators one distinguishes motile stalkers, lurking predators, sessile opportunists or grazers (2). Spiders and scorpions are typically lurking predators or motile stalkers.

The predation event in early stage life forms probably started together with the eukaryogenesis process, beginning with external (i.e. extracellular) prey digestion (in unicellular organisms extracellular is synonym with extra-corporeal). This is then presumably followed by intracellular digestion after the evolutionary acquisition by the cells of all requirements to perform endocytosis (3). In a brief overview through Metazoa diversity these two kinds of digestion combined or alone can be observed. Intracellular digestion is a common feature to most representatives of the

invertebrate phyla and the basal chordates. The taxa Placozoa, Porifera, Lophophorata (2) and the non vertebrate Chordata (4) rely exclusively on intracellular digestion. Platyhelminths, Nemertea, Annelida, Mollusca (2), Chelicerata (5) and Crustacea (6, 7) perform both intra and extracellular digestion. Ctenophora, Onycophora, Tardigrada (2), Myriapoda (8, 9), Hexapoda (10) and Vertebrata (4) digest the food primarily extracellularly.

Comparative investigations of the digestive systems are important for evolutionary studies and the discovery of novel enzymes with different specificities from selected animal groups can be used for human benefit. The digestive physiology plays a central role in homeostasis, understanding all aspects of digestion provide the tools to control poisonous animals, agriculture pests and vectors of animal, plants and human diseases. Metazoan diversity is enormous (about 1,200,000 described species (11), and this may be one of the reasons that some fascinating animals and efficient predators like spiders and scorpions still have not been thoroughly studied in terms of their particular way of feeding. Therefore, we selected two chelicerate species present in Brazil, i.e. the yellow scorpion *Tityus serrulatus* and the hermit spider *Nephilengys cruentata*, to further characterize the molecular physiology of their digestion.

1.2 The Chelicerata clade

Although the Chelicerate clade has its origins in ancient Cambrian seas, nowadays most of these animals successfully conquered the land. This basal group in Arthropoda phylogeny is characterized by the presence of six pairs of appendages (pre-oral chelicerae, pedipalp and four locomotory members), twelve opisthosomal segments and a post-anal telson. Traditionally, chelicerates are divided in the aquatic Merostomata (Xiphosura and Eurypterida) and the terrestrial Arachnida (12).

The arachnids are the second biggest group in Metazoa diversity with 93,000 species described including spiders (Araneae), scorpions (Scorpiones), harvestmen (Opiliones), ticks and mites (Acari), false scorpions (Pseudoscorpiones), windscorpions (Solifugae) and vinegaroons (Uropigy) (13). The phylogeny of the arachnid taxon is still under discussion in the literature. Some authors consider them a paraphyletic group since most part of the characters that sustains its monophyly

are regarded to terrestrial adaption (14). Monophyletic or not, these animals are adapted to virtually every imaginable situation and lifestyle.

Arachnids have an extensively branched digestive system and efficiently combine extracorporeal/extra-oral digestion (EOD) with intracellular digestion. This not seldom allows the consumption of a prey bigger than themselves. EOD is divided in two types: type I, where the chemical liquefying occurs totally inside the prey´s body and type II, in which prey 'milling' and chemical digestion occur in the pre-oral chamber. EOD Type I can be further subdivided in two groups, refluxers and non-refluxers. Non-refluxers secrete the enzymes only once whereas refluxers repeatedly secrete and pump digestive juice into the prey (1).

1.2.1 Overview of the digestion in arachnids

Arachnids typically have an extremely branched system of midgut glands, i.e. diverticula connected to the midgut with the bigger part of their volume in the opisthosomal region. Basically, these diverticula consist of a simple dimorphic epithelium with digestive and secretory cells. Only in Acari some major differences are observed (5). In (Acari-) Anactinotrichida, for example the predatory tick *Pergamasus longicornus* and the blood-sucking ticks (15), the secretory cells are absent whereas in other cases more cell types are present like in *Acarus siro* (16). The diverticula are linked by an intermediate tissue that can be highly developed as in scorpions or smaller as in Ricinulei. "Finger-like" connections between these tissues are observed in Uropygi, Amblypygi, Araneae, Palpigradi, Pseudoscorpiones and Acari-Actinotrichida (5). Most part of predator arachnids present EOD type II (Scorpiones, Pseudoscorpiones, Solifugae, Uropygi and Amblypygi). In Araneae and Acari families both types of EOD can be observed, however Araneae presenting EOD type I are refluxers and Acari EOD type I are non refluxers. Opiliones are the only arachnid group that does not do any EOD type (1) and together with some Acari (Oribatida, Acaridida) are the only arachnids that ingest solid food (16, 17). Probably due to this fact peritrophic membranes can be found in these two groups (16, 18). Nevertheless peritrophic membranes were also found in the midgut of spiders that only ingest liquefied food (19) and in the feces of Solifugae (20).

In a general way the digestive process in Arachnida by the observation in spiders can be summarized as follows: the digestion of the food starts externally with

regurgitation of the digestive juice, which is composed by the contents of secretory granules produced in the secretory cells. After 20 minutes the partially digested food is filtered and passes through the digestive system reaching the midgut glands. Strong pinocytic activity starts in the apices of the digestive cells. After one hour, all mature granules from the secretory cells are already discharged into the lumen and new ones are being synthesized by the rER. The nutritional vacuoles derived from the previous meal are moved to the basal parts of the cell while the new ones are getting bigger because more pinocytic vesicles are aggregating. Small lipid droplets that often unite are formed inside the nutritional vacuoles in different regions together with glycogen. Thereafter some of the vacuoles remain the same with storage function but in others the lipids and glycogen are extruded from the degenerating vacuoles to the cytosol. After 48 hours, the cytosol from the digestive cells is filled with fat and glycogen which then will be largely stored in the intermediate tissue. Seven days after re-feeding, excretory vacuoles can be abundantly observed in the lumen (5, 21).

1.2.2 Scorpiones and their digestive system

Scorpions are the oldest known arachnids with a fossil register in the middle Silurian from 428 millions of years ago (14). Their first existence is believed to be aquatic and some authors consider that they evolved independently from other arachnids being more related to Eurypterida (14). They definitely conquered land about 325-350 Ma with their bauplan practically remaining the same, which makes them good animals for comparative-evolutionary studies. The yellow scorpion *Tityus serrulatus* used in this study belongs to the family Buthidae, which is widely spread in the world occupying all six faunal regions. Therefore it is considered as the plesiomorphic sister group of the other recent scorpions. All deadly scorpions can be found in this family and some of them in the genus *Tityus* (22). *Tityus serrulatus* is the main responsible for accidents with deaths in Brazil. Due to the fact that this species reproduces parthenogenetically and adapts well to urban environments, its distribution, which originally was restricted to Minas Gerais State, now is expanded to Bahia, Ceará, Mato Grosso do Sul, Espírito Santo, Rio de Janeiro, São Paulo, Paraná, Pernambuco, Sergipe, Piauí, Rio Grande do Norte, Goiás, Distrito Federal and Santa Catarina. There is no pesticide available on the market which can kill

scorpions and the ones used to other pests such as cockroaches and rats are not efficient against scorpions because of their particular biology (23). The lack of pesticides to control these animals is certainly a consequence of the scarce literature about their physiology, in particular the digestive physiology of scorpions in comparison to other animals such as insects.

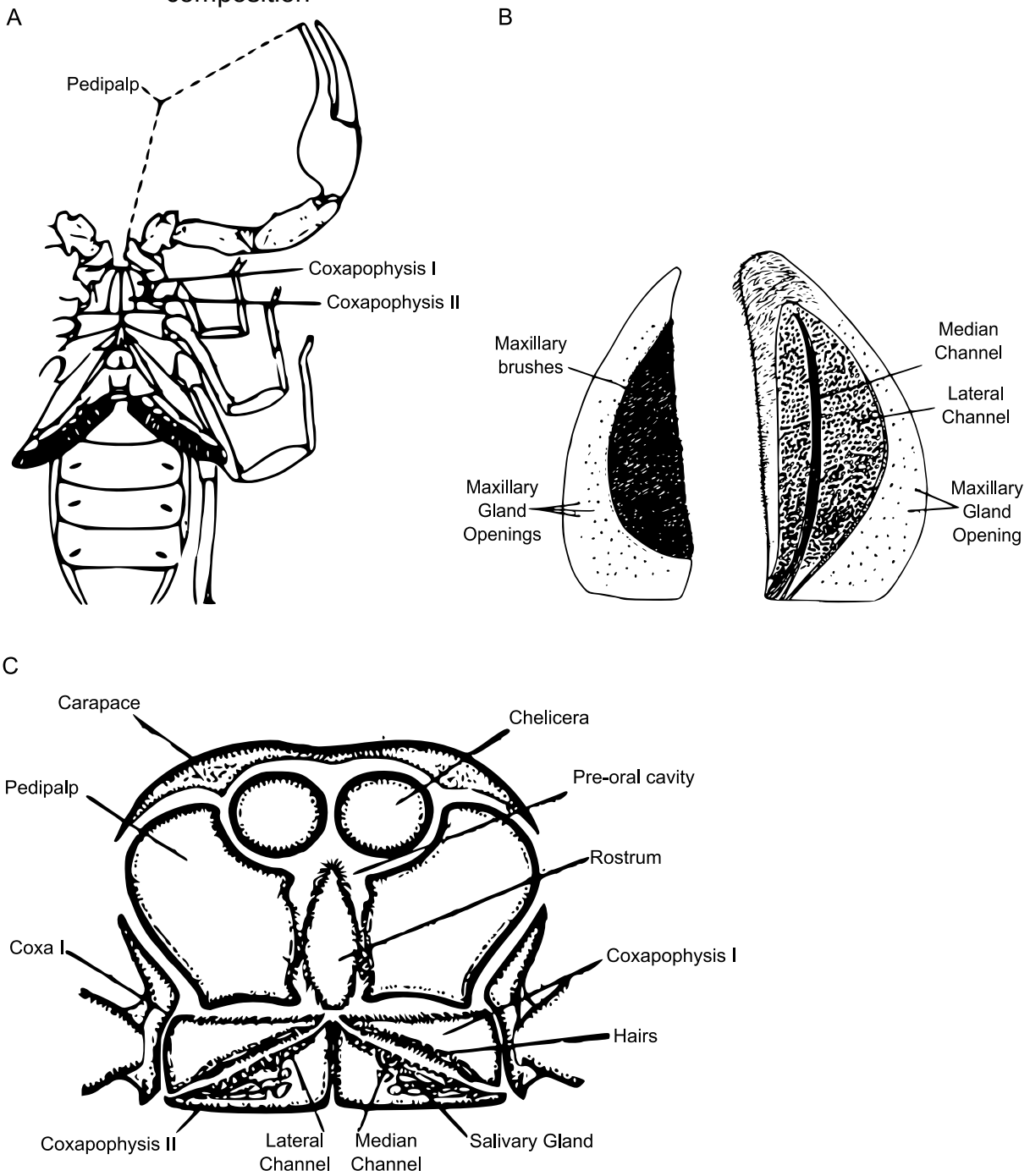
It is believed that, in scorpions, the digestive process could start with the secretions from the first segment of legs I and II, respectively denominate coxapophysis I and coxapophysis II (Figure 1.1A). These structures delimitate the basis of the pre-oral chamber and contain simple alveolar glands that release their content during feeding and hairs that are responsible for the filtration of solid particles remaining behind after prey liquefying in the pre-oral cavity (Figures 1.1B and 1.1C) (24). The presence of activities such as lipase, amylase and peptidase were verified by Auber (25) in *Butus occitanus* saliva. However; by the method utilized, the collected saliva was certainly contaminated with digestive juice. So far, it has not been possible to confirm if these glands have a digestive role or if they act primarily as a lubricant and/or mucus producer that helps to stick the solid particles in the hairs.

Although the first enzymological article about the digestive system of scorpions dates from 1922 (26), subsequent studies have only been published very sporadically using different terminologies to these organs. In his textbook, Polis (22) compiled in a clear way the different nomenclatures so far present in the literature that were used by different authors. In 1999, Farley (27) made some considerations about the wrong use of some terms, for example “stomach” as a reference to the prosomal midgut. In the description below it will be used the one of the nomenclatures used by Farley.

The scorpion digestive system has three major subdivisions: foregut, midgut and hindgut (Figure 1.2). The foregut is composed of mouth, pharynx and esophagus. Scorpions present EOD type II, with pharynx and esophagus acting together as sucking apparatus of the liquefied food. A muscular valve at the esophageal-midgut junction prevents the digestive juice being regurgitated at the time of swallowing. The midgut starts in the prosoma extending to the last metasomal segment where it joints with a short hindgut lined with cuticle. The midgut is divided in prosomal midgut, anterior intestine and posterior intestine (Figure 1.2A). The midgut glands are all laterally connected to the midgut with the sole exception of the

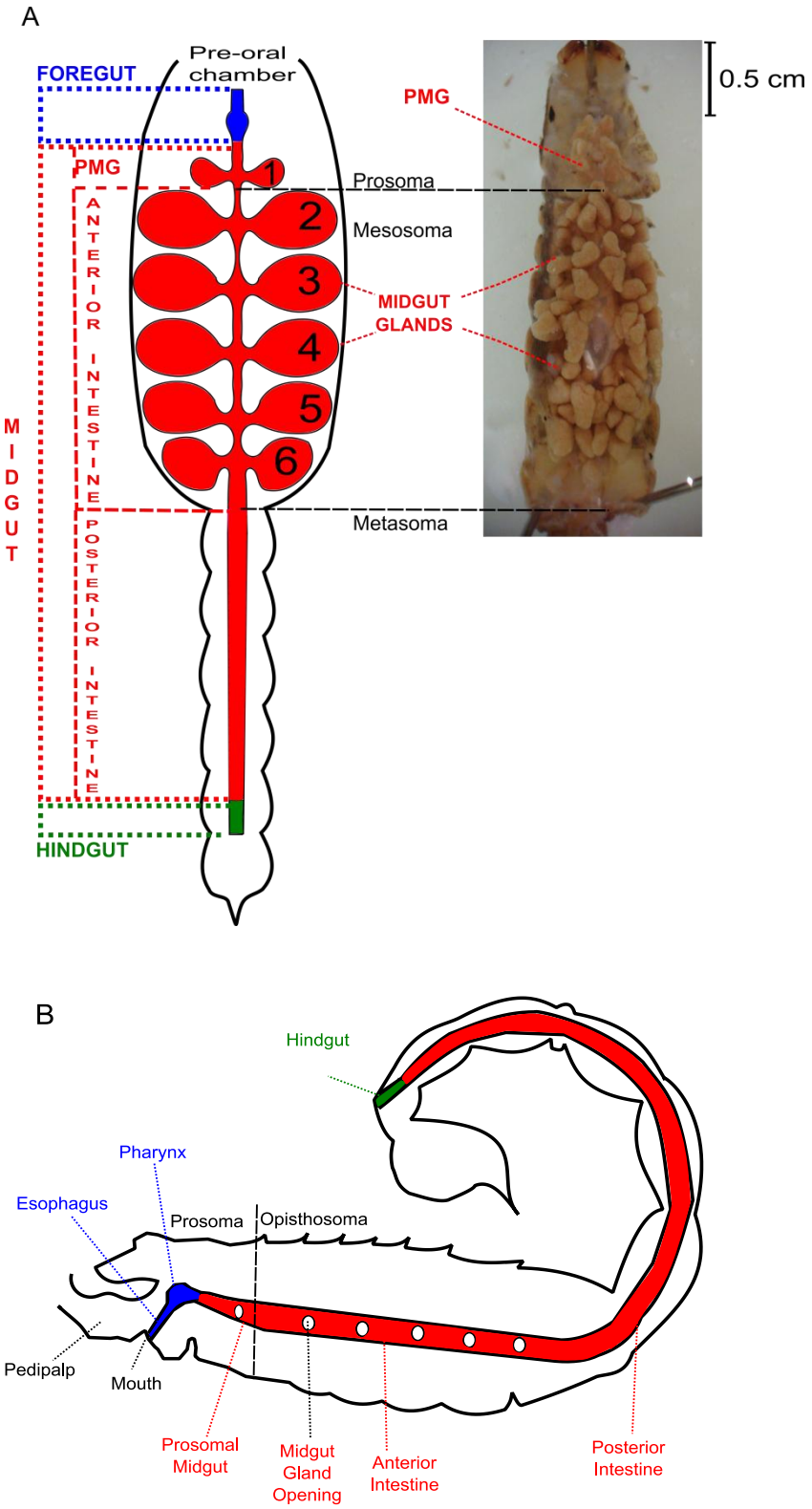
first pair of glands which is localized in the prosomal region and also has a dorso-medial connection (not shown). The other five pairs are localized in the mesosoma. The midgut in this region is called anterior intestine (Figure 1.2A). For simplification

Figure 1.1 - Location and anatomy of coxapophyses I and II and the pre-oral cavity composition



A) Ventral view of the scorpion *Androctonus australis* evidencing the location of the coxapophyses I and II. B) Ventromedial surface of coxapophysis I (left) and dorsolateral surface of coxapophysis II (right) showing their anatomy and the presence of the maxillary brushes, the gland opening and the channels that are filled with the gland secretion. C) Diagrammatic transverse section from the pre-oral cavity in order to show the different parts that compose it.
Sources: A) Modified from Polis (1990). B and C, *Heterometrus bengalensis*, modified from Srivastava (1955).

Figure 1.2 - General morphology of scorpion digestive system and its location



Schematic ventral (A) and lateral (B) views of scorpion digestive system and its divisions. On the right (A) a ventral picture from *Tityus serrulatus* MMG. PMG, prosomal midgut [source: modified from Polis (1990) (B)].

reasons in this work, the prosomal midgut and the anterior intestine with their respective digestive glands are referred as midgut and midgut glands (MMG), unless when previously specified.

Midgut glands are composed by the midgut diverticula surrounded by an intermediate tissue, which is believed to have a metabolic and storage role analogous to the fat body of insects (28). Due to the multi-functional nature of this tissue some authors call it hepatopancreas. Nevertheless, in view of the arguments presented by Van Weel (29) about the proper use of terms to invertebrates, this work will always refer to it as midgut glands/ digestive glands. The digestive glands contain secretory and digestive cells. The former secretes its contents into the lumen to start EOD and the latter absorbs the partially digested food through pinocytosis, initiating the slow process of intracellular digestion. Glycogen and lipids are stored at digestive and intermediate tissue cells (28). The metasomal part of the midgut, or posterior intestine, probably is not associated with digestion since it has a muscular appearance without secretory cells (30).

Scorpions feed on immensely varied diets, the most common are insects and other arachnids (spiders, solifuges and other scorpions) but they also eat isopods, gastropods and vertebrates such as lizards, snakes and rodents. A scorpion can gain one-third or more of its own weight in one single meal with little waste (22) and can survive up to one year of starvation (31). Despite these interesting and efficient mechanisms of feeding, few studies so far exploited the molecular properties of digestion in this group. Sarin (26), Pavlovsky and Zarin (30) as well as Bard and George (32) found activities such as chymosin, pepsin and trypsin. Said (33) observed an acid cysteine endopeptidase activity (pH 3), a trypsin-like activity at pH 6, cysteine and serine carboxypeptidases most active in pHs 4.6 and 5.2-7.2, respectively, and finally an aminopeptidase and a dipeptidase with optimum pH at 8. Louati and collaborators (34) purified an endopeptidase with higher hydrolytic activity at pHs between 7 and 8 and classified it as a chymotrypsin-like activity. Our group studying the endopeptidases from *Tityus serrulatus* purified cysteine endopeptidases with an optimum at pH 3 and 5.5, as well as alkaline peptidases active at pH 8 probably belonging to serine peptidase and/or metallopeptidase families (Chapter 2). Other articles about digestive enzymes different from peptidases in scorpions are: Vijayalakshmi and Kurup (35, 36); Zouari et al., (37-40).

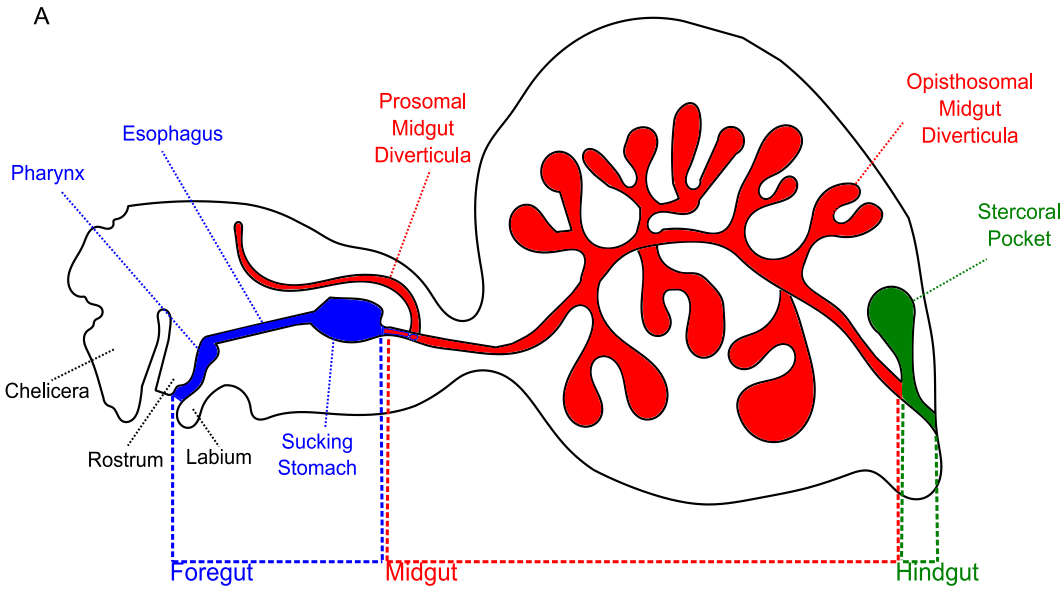
1.2.3 Araneae and their digestive system

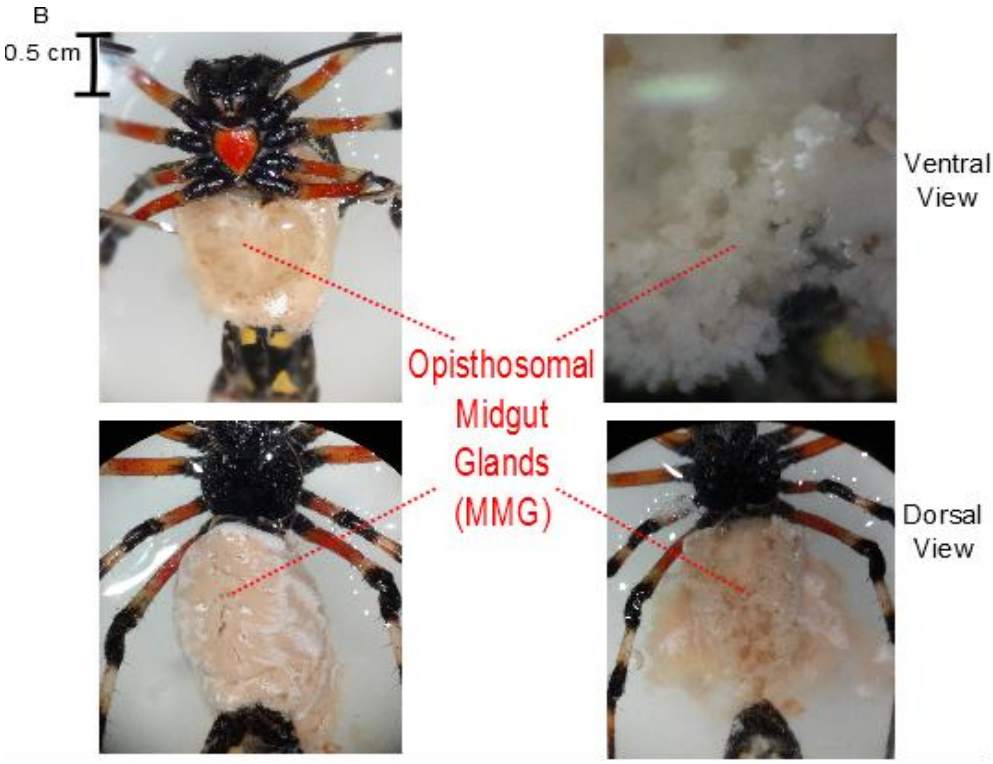
The Araneae group is the second most diversified clade of Arachnida with nearly 39,000 described species distributed in 110 families. Some of the major anatomical features, including cheliceral venom glands, male pedipalpi modified for sperm transfer, abdominal spinnerets and silk glands, strongly support the monophyly of the group (13). Despite remaining debates in Arachnida phylogeny, it is well accepted that the closest relatives to the spiders are the Amblypgi, Schizomida and Uropygi (41). The plesiomorphic spider suborder, Mesothelae, has a fossil dating from the late Carboniferous and in these spiders it is possible to see substantial traces of segmentation (42). The other suborder is Opisthelae and comprises the groups Mygalomorphae and Araneomorphae (the “true” spiders). The latter is the most diversified group with about 90% of all known spider species, including the spiders that spin orb webs (Orbiculariae) (13).

Nephilengys cruentata, an orbicularid spider from the Nephilidae family, is highly synanthropic and known for constructing more than 1 meter high semiorb webs with tubular recoil attached to the axis. It has been reported in tropical and subtropical Africa, Colombia, Paraguay and Brazil´s southeast being commonly found in and surrounding human constructions (43). The females are relatively big with a body size ranging from 17.4 to 28 mm (43) and they choose sites to construct their webs where other conspecifics are present (44). We observed that the digestive juice from this spider can be obtained by mechanical or electrical stimulus. Thus, because of its easy obtainment and the possibility of collecting digestive juice, females of this species were chosen as the object for studying the poorly understood and very efficient molecular mechanisms of digestion in spiders, which use EOD type I with refluxes allowing the consumption of big preys including birds (45) and bats (46) in addition to other arthropods.

Differently from scorpions the mouth parts from the spiders do not contain alveolar glands. However the entire digestive system is more complex with the presence of a specialized sucking stomach and filtering pharynx. The foregut is covered with chitin and is localized in the prosomal region comprising the mouth, pharynx, esophagus and sucking stomach (Figure 1.3A). The oral cavity is frontally and posteriorly delimitated by the pharynx articulations rostrum and labium, respectively, and laterally by the maxillas, which contain hairs that assist the filtration

Figure 1.3 – General morphology from the digestive system of a spider and its location





A) Diagrammatic lateral view of the digestive system of a spider showing its components and location. B) *Nephilengys cruentata* opisthosomal midgut diverticula. On the right, it is possible to see the diverticula ramifications after decompression.
Source – A) Modified from Foelix (47).

of solid particles and in some cases have saws that help the triturating process. The pharynx starts to suck the liquefied food and also filters the solid particles. The sucking stomach acts as a pump doing the food pass through the esophagus

reaching this specialized apparatus to be then delivered to the midgut. The midgut starts next to the sucking stomach still in the prosomal region, it is branched in diverticula and not covered with chitin. In some species the prosomal diverticula can reach the beginning of the legs (47). The midgut pass through the pedicel and in the opisthosomal region is extremely branched in diverticula that occupy most part of the internal volume in this region (Figure 1.3A). The opisthosomal midgut with their midgut glands will be referred in this work as MMG (Figure 1.3B) as a matter of simplifying, unless when previously specified.

As in scorpions the diverticula is surrounded by an intermediate tissue with storage and metabolic roles, both tissues together are the midgut glands (Figure 1.3B). At the midgut end after the insertion of the Malpighian tubules starts the hindgut, which contains a stercoral pocket before the anus (47). Two types of cells compose the midgut epithelia as in scorpions, the secretory and digestive cells, the latter one being more common. These two cell types are connected with intermediate tissue through finger-like protrusions. The digestive cells are club-shaped and possess a brush border that protrudes into the lumen with regularly arranged microvilli. It contains nutritional/digestive vacuoles that vary in size according the feeding condition, they are formed by the pinocytic activity and fusion of small vesicles during the eating process. The pinocytic vesicles are formed by an apical tubular system present in the cell. In some cases are also present excretory vacuoles and frequently spherites. Secretory cells are smaller and their microvillous border is not as regular as in digestive cells and the apical tubular system is not present. They have an elliptical shape and a big amount of concentric layers of rough endoplasmatic reticulum (rER) surrounding secretory granules which contain digestive enzymes. Secretory cells appearance do not change during starvation and 20 minutes after start feeding their contents already were released in the lumen by eccrine extrusion, within one hour these granules are rebuilt and they are complete mature after 30 hours (21).

The molecular characterization of the digestive process in spiders was poorly exploited so far. Mommsen (48) found activities such as trypsin, chymotrypsin, carboxypeptidase, aminopeptidase and an unidentified endopeptidase in *Tegenaria atrica*. Two low molecular mass metallo peptidases with optimum pH at 7.8 were isolated by Kavanagh and Tillinghast (49). Attkinson and Wright (50) verified colagenolytic activity in the digestive juice from different species of spiders. Foradori

and collaborators showed that the digestive juice from *Argiope aurantia* is capable of hydrolyzing different proteins from the connective tissue and thereafter isolated two metallo peptidases from the astacin family in the same kind of sample (51, 52). In our group we identified two metallo peptidases from the astacin family acting at basic pHs (Fuzita et al, unpublished) and at least two cysteine peptidases with acidic optimum pH in the digestive system of the spider *Nephilengys cruentata* (Chapter 4).

1.2.4 The digestion process in Ixodida

The suborder Ixodida from the Parasitiformes group is composed by the hard and soft ticks and belongs to the Acari subclass from the Arachnida group. These blood-sucking ectoparasites are responsible for transmitting human and animal diseases caused by different pathogens as viruses, bacteria and protozoa (53). The vectorial capacity of Ixodida is the reason why this group is, by far, the best arachnids studied nowadays.

The protein digestion in ticks relies in a multiproteolytic system composed by cysteine (legumain and cathepsins B, C and L,) and aspartic peptidases (cathepsin D) as described by Sojka (54). Lots of different authors separately studied these proteins and/or genes (15, 55-60). Franta (61) compiled all this knowledge and performed a detailed study about the hemoglobinolytic pathway in the gut of *Ixodes ricinus*. This article showed that all the digestion takes place inside the midgut cells and that the initial cleavage process is started by the legumain, cathepsins L and D followed by cathepsin B activity which is the most abundant protein in this pathway. The complete digestion is obtained by the aminopeptidase activity from cathepsin C and the carboxy-dipeptidase activitiy of cathepsin B.

1.3 Final considerations

The digestive process is deeply associated with the evolutionary history and feeding habits from a determined group. The radiation of arachnids started in the Silurian and these animals are commonly predators. Their branched digestive system coupled with the storage/metabolic role played by the intermediate tissue it was probably present in their ancestor since the same can be observed in xiphorusans. EOD is likely a terrestrial adaptation and arachnids efficiently combine it with

intracellular digestion. This harmonic arrangement allows predators arachnids to consume and store a big amount of nutrients from a single prey, which in some cases are bigger than the own predators. Despite this so far only few studies attempted doing the molecular and functional characterization of the process in spiders and scorpions. Understanding it is helpful not only to sum knowledge about Arachnida group and its relation with other arthropods, but also the screening of new molecules from non-studied groups is a source of biological tools that can be used in human benefit.

1.4 Aim of the thesis

The aim of this Ph.D thesis was to study the poorly understood molecular physiology of digestion in two arachnids species, the yellow scorpion *Tityus serrulatus* and the hermit spider *Nephilengys cruentata*. In order to achieve this, the enzymology studies focused on the characterization of the peptidase activities whereas the transcriptomic analysis, using a next generation sequencing platform, identified the main set of proteins in two different physiological conditions, fasting and fed animals (in the spider case the relative gene expression was obtained). A shotgun proteomics approach was used in MMG samples to confirm the presence of these transcripts as proteins at the Arachnida MMG and describe their subcellular location.

The chapters 2 and 4 of this thesis are about the enzymological characterization of the peptidases activities in the MMG of *Tityus serrulatus* and *Nephilengys cruentata*, respectively. A series of varied substrates was used in different assays conditions in the presence or absence of peptidase inhibitors in crude or chromatographically separated samples in order to identify peptidasic activity. Molecular biology techniques as RACE (rapid amplification of cDNA ends) were also used to obtain some enzyme sequences. In chapter 4 a recombinant enzyme from the MMG was functionally heterologously expressed in bacteria system.

The chapters 3 and 5 give the high throughput data obtained by the next generation sequencing and the shotgun proteomics experiments realized with the MMG from *Tityus serrulatus* and *Nephilengys cruentata*, respectively. A huge amount

of protein sequences was achieved and their subcellular location was determined after the digestive juice analysis or software prediction.

Lastly astacins and cathepsins L sequences obtained from the two species under investigation were phylogenetically analyzed in chapter 6, presenting the concluding remarks and future perspectives about this work.

CHAPTER 2 - ANALYSIS OF PROTEIN DIGESTION IN THE SCORPION *TITYUS SERRULATUS*: INSIGHTS INTO THE DIGESTIVE PHYSIOLOGY OF AN ANCIENT ARTHROPODA

2.1 Introduction

Scorpions are very efficient predators that have a varied diet (e.g., insects, spiders, solifugae, scorpions, isopods, gastropods, snakes, lizards, rodents) and exhibit a high rate of conversion for prey biomass relative to scorpion biomass (22), which suggests that scorpions must have efficient digestive enzymes and nutrient absorption systems. The scorpion body plan is considered successful because it has experienced very few changes during the past 400 million years (22). Due to these characteristics, scorpions are particularly attractive animals for ecological, physiological and evolutionary studies.

Prey capture and envenomation are well-studied processes (62-64). However, very few physiological processess related to digestion and digestive enzymes in scorpion species have been published. Sarin (26) and Pavlovsky and Zarin (30) identified the presence of the first peptidases: pepsin, trypsin and chymosin. Said found cysteine catheptic activity in *Buthus quinquestriatus* (33). The only recent studies about digestive enzymes in scorpions have described the characterisation of an amylase (34), a lipase (37) and a chymotrypsin from *Scorpio maurus* (65). No studies regarding the scorpion digestive physiology have been reported and there is not available any DNA nor protein complete sequences from this tissue. Studies of Arachnida digestive peptidases have concentrated on the Acari, mainly hard ticks, due to their vectorial competence. The major digestive endopeptidases in these animals are cathepsin L, B and D (61, 66, 67).

In this work, for the first time, we report the identification of 27 complete and 14 incomplete DNA sequences of endopeptidases identified at the transcriptional level in *Tityus serrulatus* midgut and midgut glands. Nine of these sequences were confirmed at the protein level by mass spectrometry. We also performed a biochemical characterization of the peptidase activities using different substrates and assay conditions with a further purification of two cysteine peptidases cysp1 and cysp 2 (cathepsins F and L, respectively). We show that cysteine, aspartic, serine and

metallopeptidases are present in both protein and transcript levels and that the activity over proteinaceous substrates ranges from pH 2.8 to 9. In addition to this, we proposed that the cysteine peptidases are likely acting intracellularly and at least one of them is in the zymogen form that can be activated under acidic conditions.

2.2 Materials and methods

2.2.1 Animals

Tityus serrulatus (Buthidae) were maintained under a natural photoregime at room temperature. The scorpions were maintained without food for at least two weeks and fed with *Gryllus* sp. adults. The fed animals were dissected less than 24 hours after the beginning of feeding.

2.2.2 Enzyme samples

Tityus serrulatus adult fed females were immobilised in a carbon dioxide chamber for 10 minutes. The telson was removed, and the animals remained in the carbon dioxide chamber for two more minutes. After that, the animals were laid down dorsally in a paraffin plate; the legs and pedipalps were removed and the females were dissected in order to isolate the prosomal midgut and anterior intestine with their respective digestive glands (Figure 1.2A). These tissues were stored at -20 °C until use. The midgut (prosomal and anterior intestine) with its prosomal and mesosomal glands are collectively referred as midgut and midgut glands (MMG). The MMG were homogenised in cold deionised water (Milli Q system) with the use of a Potter-Elvhejem homogeneizer. For the samples used in the purifications steps the deionised water contained 1.0 mM of methyl methanethiosulfonate (MMTS) (68).

2.2.3 Protein determination, hydrolase assays and peptidase classification

The protein concentration was determined according to the method of Smith et al. (69) using ovalbumin as a standard. The assays for the identification of peptidases were performed using different substrates. Substrates and assay conditions are listed on Table 2.1. All assays were performed at 30 °C and the

measured activity was proportional to the protein concentration and the incubation time. No-enzyme and no-substrate controls were included. A combination of substrates, assay conditions and specific inhibitors were used to classify the peptidase activities at chromatographic fractions from MMG (70, 71). Inhibitors used were: 10 µM E-64 (cysteine peptidase), 10 µM pepstatin (aspartic peptidase), 1 mM PMSF (serine peptidase), and 5 mM benzamidine (serine peptidases). Chicken cystatin (0.5, 50 and 500 nM) from eggs (Calbiochem) was tested with the purified samples (7.8 pmol solution of cysp1 and 23 pmol solution of cysp2). In the assays with inhibitors, under either control or experimental conditions, the substrates were added after a 30 minute pre-incubation with the inhibitor at 30 °C in the same buffers used for activity assays.

The samples were incubated at 30 °C in citrate-phosphate buffer containing 3 mM cysteine and 3 mM EDTA for 60 minutes (for the homogenate) or 10 minutes (for the partially purified samples) for zymogen activation experiments,

2.2.4 Isolation of cysteine peptidases

The samples from the homogenate of the MMG of *Tityus serrulatus* containing 1 mM MMTS were fractionated in 1.8 M ammonium sulfate for at least 16 hours at 4 °C. The samples were centrifuged for 20 minutes at 16,100 × *g* and 4 °C. The supernatant was applied to a hydrophobic column (Hitrap Butyl FF-GE) coupled to an ÄKTA-FPLC system (GE). Column was equilibrated in 50 mM phosphate buffer (pH 6) containing 1.7 M ammonium sulfate.and eluted with a 25 mL gradient of 1.7 – 0 M ammonium sulfate in 50 mM phosphate buffer (pH 6); fractions of 1 mL were collected. Active fractions on Z-FR-MCA were pooled, desalted (HiTrap desalting column, GE) and concentrated using a Vivaspin 6 membrane (GE). The samples were then applied to a cation-exchange column (Resource S-GE) equilibrated in 50 mM sodium acetate buffer (pH 5). The protein was eluted using a 40 mL gradient of 0 – 0.6 M NaCl in the equilibrating buffer, and fractions of 0.5 mL were collected and assayed using Z-FR-MCA. The two purified enzymes were visualised by SDS-PAGE and named cysp1 and cysp2.

A simpler partial purification was also used. The homogenised samples were applied to an ion-exchange column (HiTrap S, GE) equilibrated in 50 mM citrate

phosphate buffer. The protein was eluted in a 25 mL gradient of 0–1.0 M NaCl followed by 5 mL of 1 M NaCl, both in the equilibrating buffer.

2.2.5 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

Samples of the isolated cysteine peptidases were diluted in a sample buffer containing 60 mM Tris-HCl buffer (pH 6.8), 2.5% SDS, 0.36 mM β-mercaptoethanol, 10% (v/v) glycerol and 0.005% (w/v) bromophenol blue. The samples were heated for 5 minutes at 95 °C in a water bath and then loaded onto a 12% (w/v) polyacrylamide gel slab containing 0.1% SDS (72). The gels were run at a constant voltage of 200 V at room temperature and then silver-stained to identify the proteins (Blum et al., 1987). The molecular mass values were calculated according to the method of Shapiro et al. (1967) using the following standards (Da): lysozyme (14.4 kDa); trypsin inhibitor (21.5 kDa); carbonic anhydrase (31kDa); ovoalbumin (45 kDa); serum albumin (66.2 kDa200); phosphorylase b (97.4 kDa).

2.2.6 Acidic activation of cysteine peptidases

The homogenate samples and the samples that had been partially purified using hydrophobic interaction chromatography were diluted in 0.1 M citrate-phosphate buffer containing 3 mM cysteine and 3 mM EDTA at a range of pH values from 2.6 to 7.0 and incubated for 1 hour at 30 °C. The incubated samples were diluted in deionized water and the activity was then measured using a 1:100 ratio of the sample to 10 µM Z-FR-MCA in 0.1 M citrate-phosphate buffer (pH 5.5), which guaranteed no pH change in the assay buffer. The pH that resulted in the highest rate of hydrolysis was selected for an incubation time course to verify the length of time that was required for acidic activation *in vitro*. After this incubation, enzymatic assays using Z-FR-MCA were performed as described above. The controls were diluted in deionised water prior to the assay or incubated at 30 ºC for the same time the experimental ones.

2.2.7 ***pH stability***

The stability of the cysteine peptidases under different pH conditions was evaluated by incubating the activated enzyme samples from the MMG homogenates in buffers with different pH values at 30 °C for 3 h or at -20 °C for 24 h. The incubation buffers used were: 50 mM citrate-phosphate and 50 mM Tris-HCl. The samples were then diluted in deionized water to guarantee adequate pH for residual activity measurement. The assays were performed in a ratio of 1:100 of the incubated sample to 10 µM Z-FR-MCA in 0.1 M citrate-phosphate buffer pH 5.5.

2.2.8 Effect of pH on enzyme activity

The purified and partially purified samples described above were assayed with Z-FR-MCA 10 µM in a series of 0.1 M citrate-phosphate buffers with pH values ranging from 2.6 to 7.0 and containing 3.0 mM cysteine and 3.0 mM EDTA.

2.2.9 Migration on gel filtration chromatography

Gel filtration chromatography using a Superdex G75 column (GE) was performed in a FPLC system (GE) to determine the molecular masses of the cysteine peptidases. The molecular mass standards used were: ribonuclease A (13.7 kDa), soybean trypsin inhibitor (21.5 kDa), ovalbumin (45 kDa), bovine serum albumin (66 kDa), and thyroglobulin (660 kDa).

2.2.10 Thermal inactivation

The thermal inactivation of cysteine peptidases after hydrophobic separation was performed by incubating the activated samples at 60 °C in 100 mM citrate-phosphate buffer with 3 mM cysteine and 3 mM EDTA at pHs 3.0 and 5.5. The residual activity was measured at different time points as previously described.

2.2.11 Isoelectric focusing

Isoelectric focusing was performed as described by Terra et al. (1978) using the homogenate or partially purified samples after ion-exchange chromatography in f 7.5% polyacrylamide gels containing 10% ampholytes (pH 3–10) (Pharmalyte 3–10,

Pharmacia, Sweden). The samples were loaded onto the top of the alkaline side of the gels after polymerisation and pre-focusing (30 min at 31 $V.cm^{-1}$).

2.2.12 Effect of substrate concentration

The effect of substrate concentration on the activity of the purified cysteine peptidases was studied using at least 15 different substrate concentrations (Z-FR-MCA and Abz-FRQ-EDDnp). The K_m values (mean ± SEM) were determined from a weighted linear regression using EnzFitter software (Biosoft). These assays were also performed in the presence of 5 different concentrations of pepstatin ranging from 1 to 50 µM.

2.2.13 Titration of purified cysteine peptidases with E-64

The molar concentrations of the purified cysp1 and cysp2 solutions were determined by titration with at least fifteen distinct concentrations of E-64 prepared from a 1.0 mM stock solution in water without pre-incubation. Then, the samples were assayed for residual activity against the appropriate methylcoumarylamide substrate, as previously described, and the results were plotted against inhibitor concentration.

2.2.14 Analysis of the Abz-FRQ-EDDnp products after hydrolysis with purified cysp1 and cysp2

The substrate Abz-FRQ-EDDnp was diluted to a final concentration of 1 µM in 0.1 M citrate-phosphate buffer (pH 3.0 or 5.5) containing 3 mM cysteine and 3 mM EDTA and incubated with the purified cysp1 and cysp2 samples for 16 hours at 30 °C. Next, 100 µL of the hydrolysis product was applied to a C18 column (4.6 mm x 50 mm, Ace) coupled to an HPLC system (Shimadzu), and the products of interest were eluted using a linear gradient of 0-100% acetonitrile with 0.1% TFA as the polar solvent. The different fractions corresponding to the observed peaks were independently subjected to mass spectrometry using an MSQ-Surveyor instrument (Thermo) with electrospray ionisation, and the cleavage point was determined.

2.2.15 Transcriptomics and proteomics procedures

For the detailed methodology see chapter 3, items 3.2.2, 3.2.3 and 3.2.4.

2.3 Results

2.3.1 Molecular biology and mass spectrometry approaches

Transcriptomics and shotgun proteomics analysis of MMG intracellular and extracellular contents allowed the identification of cysteine, serine, metallo and aspartic peptidases (Table 2.2). All complete peptidases transcriptomically identified are true enzymes presenting the catalytic residues except three cysteine peptidases from the C54 family (Clan CA) which lack the catalytic Cys and Gln. A Neighbor-joining dendrogram of all the complete endopeptidases sequences illustrate the differences between them (Figure 2.1).

The bootstrap values strongly support the different groups of enzymes as expected, with values of 100, 100 and 99 for metallo-, serine and catalytic cysteine peptidases from clan CA, respectively. Aspartic peptidase (cathepsin D) and legumain did not form any group since only one sequence of each was obtained. The present data shows that cathepsin L-like cysteine peptidases and astacin-like metallopeptidases are the most diversified enzymes. Furthermore, three trypsin-like serine peptidases containing the CUB and LDL domains, a legumain and an exemplar of cathepsins D, F and O were also identified. Some of these sequences could be confirmed not only at the transcript level but also at the protein level by LC-MS/MS (Table 2.2). Incomplete sequences were also identified at the transcript level: 5 cathepsins L, 4 astacins, 3 trypsins, 1 cathepsin D and 1 cathepsin B.

2.3.2 Classification of the digestive enzymes from Tityus serrulatus MMG

In order to identify peptidasic activity involved in prey protein digestion in the scorpion *Tityus serrulatus* MMG a series of substrates for cysteine, serine, aspartic and metallopeptidases were tested under different assay conditions (Table 2.1). The data in Table 2.3 show activities on hydrolysed substrates. The optimum pH range to hemoglobin hydrolysis was 2.6 to 3.0 whereas to casein-FITC 8.0 to 9.0 (Figure 2.2).

Although hemoglobin hydrolysis could be observed in very acidic pHs, the activity measured below pH 2 was highly unstable.

Table 2.1 – Assay conditions and methods used in the determination of peptidase activities from *Tityus serrulatus* midgut and midgut glands

Enzyme Class	Enzyme	Substrate, concentration	Buffer (mM); pH	Reference
Cysteine peptidase*	**Cathepsin L**	Z-FR-MCA; 10 µM	Citrate-phosphate 100; 2.6-6 MES 100; 6-7 TRIS-HCl 100; 7-9	(70)
		Hemoglobin; 2%	Gly-HCl 100; 1,5-2 Citrate-phosphate 100; 2,6-3,8	(73, 74)
		Abz-FRQ-EDDnp; 0.1 µM	Citrate-phosphate 100; 3, 5.5	(75)
		Abz-GIVRAK-EDDnp; 0.1µM	Citrate-phosphate 100; 3, 5.5	(76)
		Abz-GIVRPK-EDDnp; 0.1µM	Citrate-phosphate 100; 3, 5.5	
	Cathepsin B	Z-RR-MCA; 10µM	Citrate-phosphate 100; 3, 5.5	(71)
		Abz-GIVRAK-OH; 0.1 µM	Citrate-phosphate 100; 3, 5.5	(76)
		Abz-GIVRPK-OH; 0.1 µM	Citrate-phosphate 100; 3, 5.5	
	Legumain	Z-AAN-MCA; 30 µM	Citrate-phosphate 100; 3, 5.5	(77)
		Z-VAN-MCA; 30 µM	Citrate-phosphate 100; 3, 5.5	
Aspartic peptidase		Hemoglobin; 2%	Gly-HCl 50; 1,5-2 Citrate-phosphate 100; 2,6-3,8	(73, 74)
		Abz-AIAFFSRQ-EDDnp; 0.1µM	Gly-HCl 100; 2.8	(78)
Serine peptidase**	**Trypsin**	Z-FR-MCA; 0.1µM	Citrate-phosphate 100; 2.6-6 MES 100; 6-7 TRIS-HCl 100; 7-9	(71)
		Casein-FITC; 0.2%	MES 100; 6-7 TRIS-HCl 100; 7-9 Gly-HCl 100; 9-10	(79)
		Z-GGR-MCA; 10µM	TRIS-HCl 100; 8	(80)
	Chymotrypsin	N-Suc-AAPF-MCA; 10µM	TRIS-HCl 100; 8	(81)
		Casein-FITC; 0.2%	MES 100; 6-7 TRIS-HCl 100; 7-9 Gly-HCl 100; 9-10	(79)
Metallopeptidase		Casein-FITC; 0.2%	MES 100; 6-7 TRIS-HCl 100; 7-9 Gly-HCl 100; 9-10	(79)
		Abz-GPKRAPWV-EDDnp; 0.1µM	TRIS-HCl 100; 8.5	--

Notes: *cysteine peptidases assay buffers contain 3 mM cysteine and 3 mM EDTA.
**serine peptidases assay buffers contain 10 mM $CaCl_2$

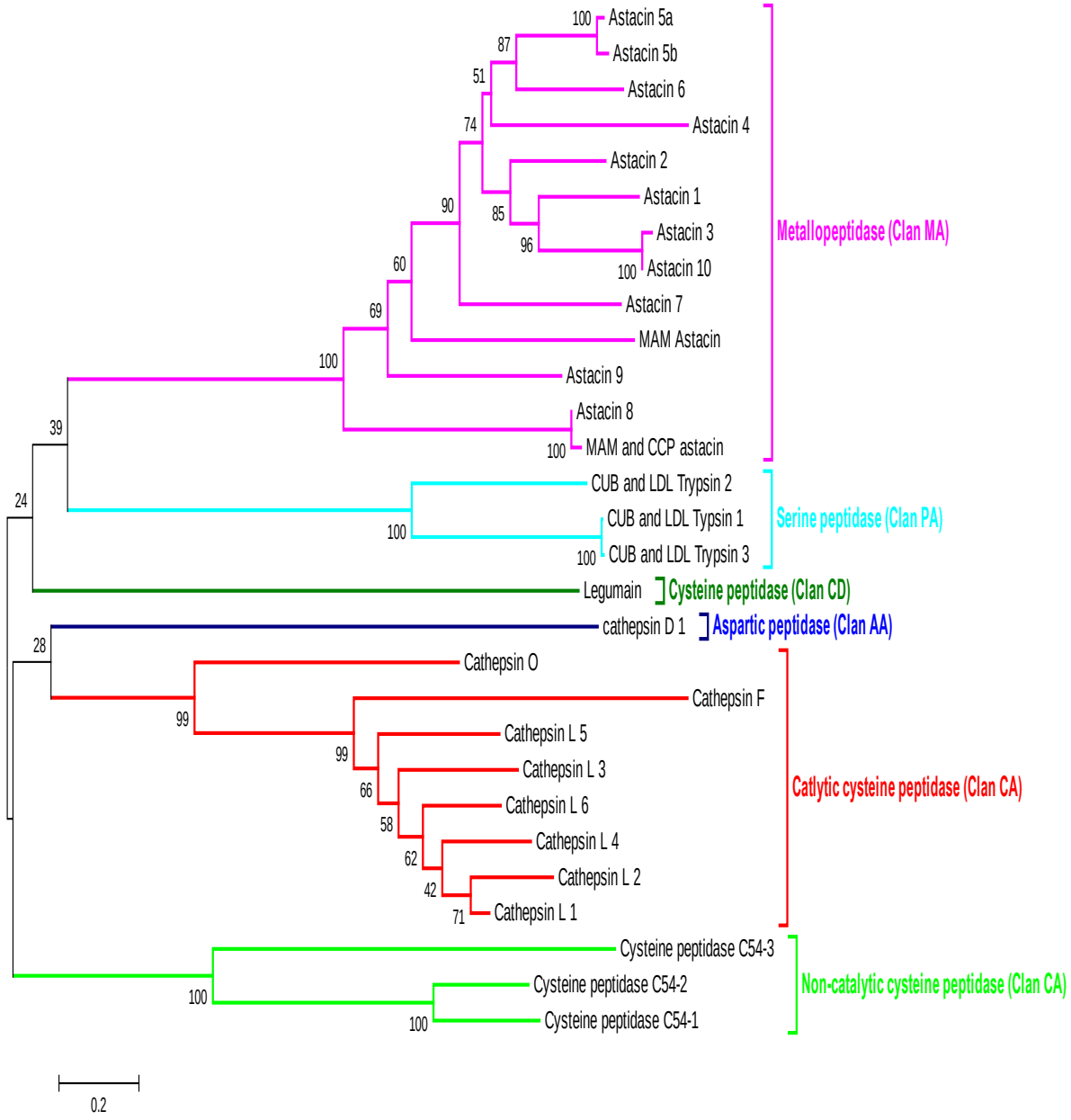
Table 2.2 - List of endopeptidases identified* in the MMG of *Tityus serrulatus*

Enzyme (Transcriptome)	Theoretical Molecular Mass (Da)	LC-MS/MS Identification**	Coverage (%)	Accession number
Cathepsin L 1	--	Yes	75	HG710154
Cathepsin L 2	37,091.84	Yes	12	HG710155
Cathepsin L 3	38,247.18	No	--	HG710156
Cathepsin L 4	37,612.66	No	--	HG710157
Cathepsin L 5	40,940.34	No	--	HG710158
Cathepsin L 6	36,867.36	No	--	HG710159
Cathepsin F	50,821.43	Yes	17	HG710161
Cathepsin O	39,300	No	--	HG710160
Legumain	37,048.43	Yes	5	HG710153
Cathepsin D	37,177.63	Yes	41	HG710162
Cysteine peptidase C54-1	46,758.64	No	--	HG710163
Cysteine peptidase C54-2	46,130.27	No	--	HG710164
Cysteine peptidase C54-3	69,934.01	No	--	HG710165
Astacin 1	27,314.88	No		HG710140
Astacin 2	23,605.33	Yes	49	HG710141
Astacin 3	26,6881.08	No	--	HG710142
Astacin 4	23,067.17	No	--	HG710143
Astacin 5	25,33.58	Yes	12	HG710144
Astacin 6	28,092.41	No	--	HG710145
Astacin 7	26,920.32	No	--	HG710146
Astacin 8	28,324.93	No	--	HG710147
Astacin 9	28,799.60	No	--	HG710148
Astacin 10	22,762.31	No	--	HG710149
MAM Astacin	47,652.83	No	--	HG710152
MAM and CCP Astacin	74,392.33	No	--	HG710151
CUB and LDL Trypsin 1	49,244.97	Yes	5	HG710166
CUB and LDL Trypsin 2	51,330.29	No	--	HG710167
CUB and LDL Trypsin 3	53,763.13	Yes	12	HG710168

Notes: *Only enzymes with complete DNA sequences (coding region) except cathepsin L1 and astacin 2.

**Only the proteins with at least 2 different peptides found by LC-MS/MS were considered as a positive identification.

Figure 2.1 - Dendrogram of the endopeptidases protein sequences using the Neighbor-joining algorithm (82)



The optimal tree with the sum of branch length = 16.1 is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches (83). The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (84) and are in the units of the number of amino acid substitutions per site. The analysis involved 29 amino acid sequences. All ambiguous positions were removed for each sequence pair. There were a total of 878 positions in the final dataset. The analyses were conducted in MEGA5 (85).

The alkaline hydrolysis of casein-FITC suggested the presence of serine and metallopeptidases. The former was corroborated by the hydrolysis of Z-FR-MCA and

N-Suc-AAPF-MCA at pH 8.0 (Table 2.3) and identification by mass spectrometry (Table 2.2). The activities of the chromatographic fractions against Z-FR-MCA at pH 8 were characterised .This activity is calcium dependent; no activity was observed in homogenate samples dialysed against EDTA in the absence of $CaCl_2$, while the absolute and specific activities were recovered in the presence of 10 mM $CaCl_2$. The hydrolysis of Z-FR-MCA was inhibited by at least 45% in the presence of benzamidine at pH 8 in a buffer containing 10 mM $CaCl_2$ (Figure 2.3A). This data is consistent with the identification of two CUB and LDL domain-containing trypsin-like serine peptidases identified by mass spectrometry, in which the LDL domain contain the motif DXSDE involved in calcium binding. The activity of astacin-like metallopeptidases could not be clearly distinguished from the serine peptidase activities. The observed activities on casein-FITC and Abz-GPKRAPWV-EDDnp seem to be result of a mixture of distinct enzymes such as metallo and serine peptidase (Table 2.2).

The hydrolysis of hemoglobin under acidic conditions indicated the presence of aspartic and cysteine peptidases. Despite both enzymes could be found by mass spectrometry (Table 2.2), the hydrolysis of hemoglobin was completely dependent of the cysteine and EDTA presence in the assay medium. In addition to that, the absence of hydrolysis of an aspartic peptidase substrate (Table 2.3) corroborates that, probably, hemoglobin hydrolysis is dependent on cysteine peptidases.

The analysis of hydrolysis ratio to distinct cysteine peptidase substrates in pH 3.0 and 5.5 showed that Z-FR-MCA presented the highest hydrolyse ratios. Based on these results Z-FR-MCA was chosen for subsequent MMG chromatographic fraction assays to separate distinct peptidases active on Z-FR-MCA and to classify them based on their susceptibility to specific inhibitors (Figure 2.3B). Assays of the fractions from hydrophobic chromatography separation with Z-FR-MCA at pH 5.5 showed two distinct activities, C1 and C2 (Figure 2.3B). A comparison of the elution profiles of cysteine peptidase and serine peptidase activities on Z-FR-MCA indicated that cysteine peptidases are more active than serine peptidases and that these enzymes presented distinct patterns of interaction and elution from the hydrophobic resin (Figures 2.3A and 2.3B).

The activities of cysteine peptidases and trypsin against Z-FR-MCA can be distinguished by the optimum pH values of these enzymes, which typically differ by 2

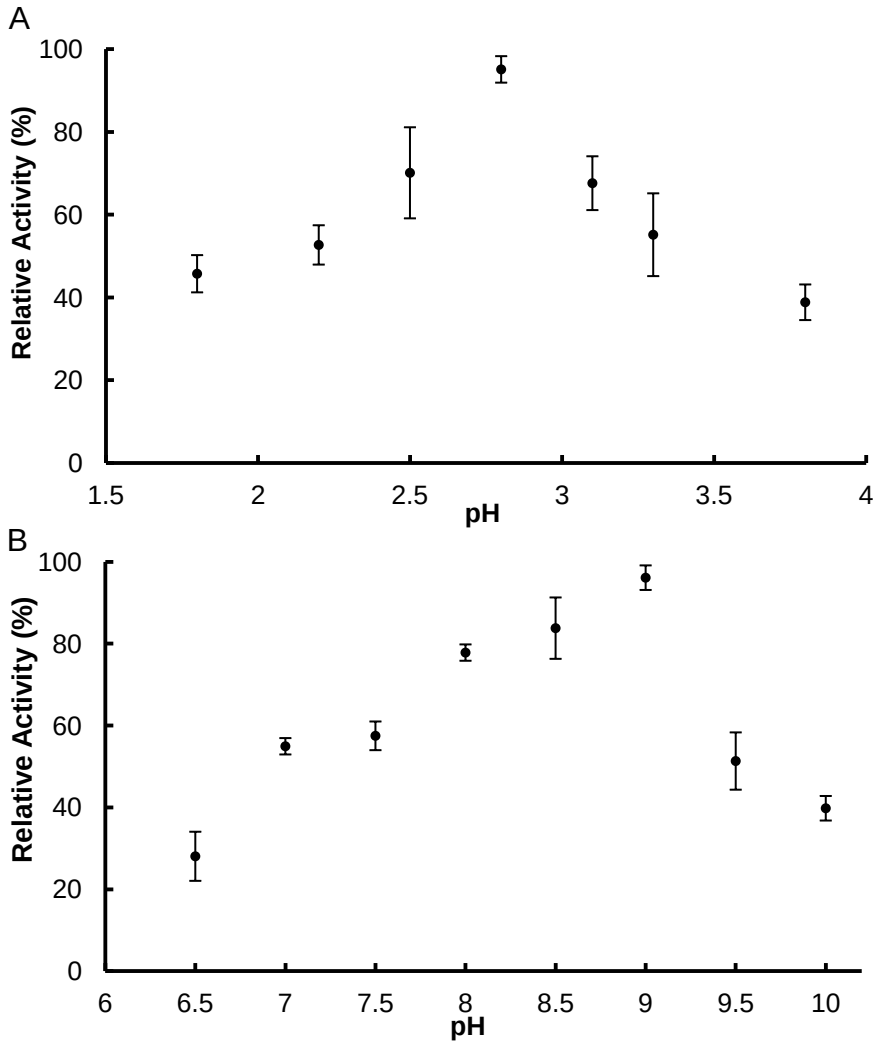
pH units, and by the fact that cysteine peptidases are activated by reducing agents such as cysteine (71).

Table 2.3 - Peptidase absolute and specific activities in the scorpion *Tityus serrulatus* MMG using different substrates*

Substrate; pH	Absolute Activity (U/MMG)	Specific Activity (U/mg)
Z-FR-MCA (3)	580 ± 8	16 ± 2
Z-FR-MCA (5.5)	700 ± 234	15 ± 5
Z-FR-MCA (8)	43 ± 2	1.2 ± 0.2
Z-RR-MCA (5.5)	81 ± 16	1.6 ± 0.2
N-Suc-AAPF-MCA (8)	4 ± 1	0.1 ± 0.04
Casein-FITC (8.5)**	2.1 ± 0.4	0.06 ± 0.03
Hemoglobin (2.8)**	28 ± 4	0.93 ± 0.04
Abz-FRQ-EDDnp (3)	35 ± 3	0.8 ± 0.2
Abz-FRQ-EDDnp (5.5)	2.2 ± 0.7	0.06 ± 0.02
Abz-GIVRAK-EDDnp (3)	0.42 ± 0.06	0.009 ± 0.001
Abz-GIVRAK-EDDnp (5.5)	0.16 ± 0.03	0.005 ± 0.001
Abz-GIVRPK-EDDnp (3)	0.18 ± 0.03	0.004 ± 0.001
Abz-GIVRPK-EDDnp (5.5)	0.2 ± 0.03	0.005 ± 0.001
Abz-GIVRAK-(Dnp)OH (3)	0.77 ± 0.06	0.018 ± 0.003
Abz-G-I-V-R-A-K-(Dnp)OH (5.5)	0.9 ± 0.3	0.025 ± 0.09
Abz-G-I-V-R-P-K-(Dnp)OH (3)	0.270 ± 0.004	0.006 ± 0.001
Abz-G-I-V-R-P-K-(Dnp)OH (5.5)	0.53 ± 0.04	0.015 ± 0.001
Abz-G-P-K-R-A-P-W-V-EDDnp (8)	0.9 ± 0.1	0.02 ± 0.004

Notes: *values are means ± SEM of cleaved substrates in at least three different biological samples from *Tityus serrulatus* MMG. Assay conditions are listed in Table 2.1. Substrates described in Table 2.1 and not discribed in this table were tested but were not hydrolysed.
** One U (unit of actitvity) to casein and hemoglobin hydrolysis are correspondent to a variation of 1,000 fluorescent units.

Figure 2.2 – Effect of the pH on activity using different substrates

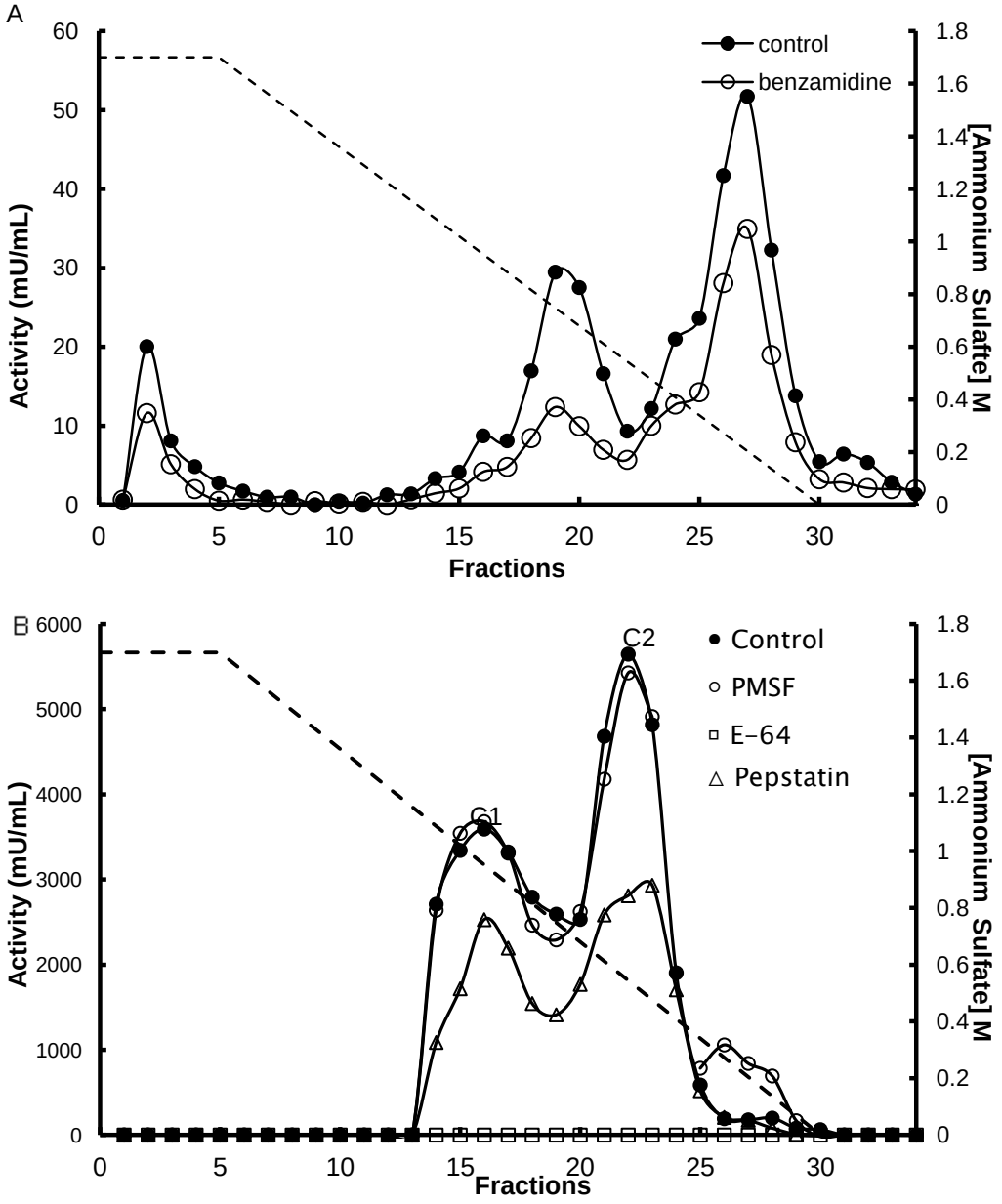


(A) Hemoglobin 2%. (B) Casein-FITC 0.2%. Buffers used (100 mM): Gly-HCl pHs 1.5 and 2; Citrate-phosphate pHs 2.6-3.8; MES pHs 6.5 and 7; TRIS-HCl pHs 7.5-9; Gly-HCl 9.5-10.

The acidic activity against Z-FR-MCA was characterised by the use of distinct inhibitors. Under these conditions, PMSF, a serine peptidase inhibitor, and the chelator EDTA, a metallopeptidase inhibitor, did not affect the activity. The only inhibitor that produced a 100% inhibition was E-64. Nevertheless, pepstatin, an aspartic peptidase inhibitor, inhibited 30% of the C1 activity and 45% of the C2 activity. However, as mentioned above, no activity was observed when Abz-AIAFFSRQ-EDDnp, a quenched fluorescent substrate specific for aspartic peptidases, was used (78), and there was no hemoglobin hydrolysis in the absence of cysteine and EDTA, suggesting that pepstatin inhibits cysteine peptidase activities. We were able to identify by proteomics analysis that the MMG possesses 4 cysteine peptidases. To confirm which enzymes are responsible for the activity in C1 and C2 fractions, LC-MS/MS experiments were performed. The same enzymes (cathepsin L

1 and 2 and the cathepsin F) were identified in these activity pools. Despite other proteins were also identified, cathepsin L 1 and cathepsin F were the most abundant

Figure 2.3 - Hydrophobic chromatographic fractioning of *Tityus serrulatus* MMG



The MMG was fractioned with 50% ammonium sulfate on a HiTrap Butyl column (GE) equilibrated in 50 mM phosphate buffer (pH 6.0). The elution was performed using a gradient of 1.7-0 M ammonium sulfate in the same buffer. (A) The activity of each fraction against 10 µM Z-FR-MCA was measured in 100 mM Tris-HCl buffer (pH 8.0) containing 10 mM $CaCl_2$ (●) or in the presence of 5.0 mM benzamidine (○). (B) The activity of each fraction against 10 µM Z-FR-MCA was measured in 100 mM CP-buffer (pH 5.5) containing 3.0 mM cysteine and 3.0 mM EDTA in the absence (●) and presence of different peptidase inhibitors: (□) 10 µM E-64; (○) 1.0 mM PMSF; (∆) 10 µM pepstatin.

ones according to the normalized spectra counting (data not shown). For the first time, we obtained evidence of other cysteine peptidase (distinct of legumains and calpains) inhibition by pepstatin. This interaction was further characterised following the isolation of the two cysteine peptidases.

Exopeptidase activity was observed in crude extracts using quenched fluorescent substrates with free C-terminus in assay conditions for cysteine peptidases (Table 2.3). Although the hydrolysis of these substrates and cleavage of Z-RR-MCA (Table 2.2) are indicative of a cathepsin B-like activity presence (71) we have only evidences of the presence of a small cathepsin B fragment at the transcriptomic level which would be hardly corroborated by proteomics data.

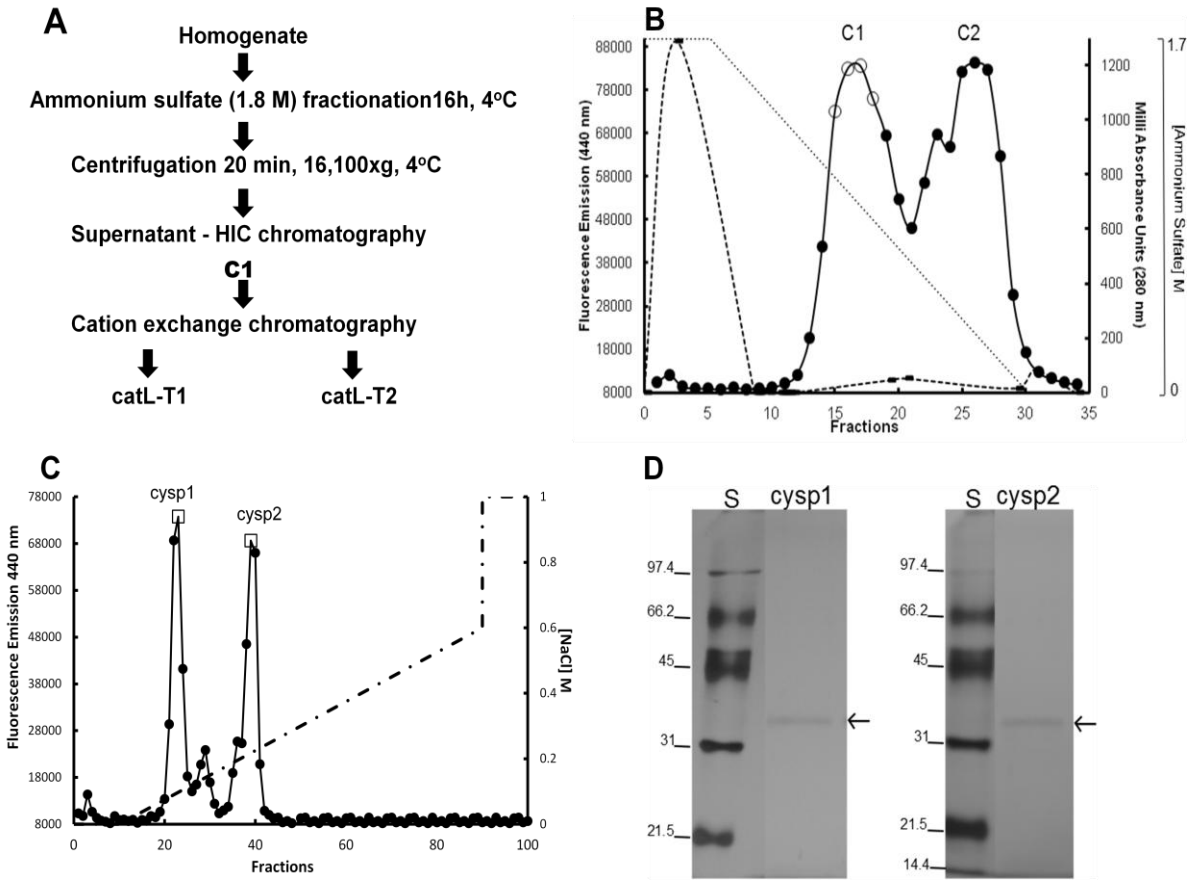
2.3.3 The purification of two cysteine peptidases found in the Tityus serrulatus midgut and midgut glands

The following sequence of steps was used to purify the two distinct cysteine peptidases: ammonium sulfate fractionation, hydrophobic chromatography and cation-exchange chromatography (Figure 2.4A). Two peaks of activity in the presence of Z-FR-MCA (C1 and C2) were observed during the hydrophobic separation (Figure 2.4B). When C1 was subjected to cation-exchange chromatography, two peaks of activity in the presence of Z-FR-MCA were observed (cysp1 and cysp2; Figure 2.4C). An SDS-PAGE analysis showed that both enzymes were effectively purified and exhibited molecular masses of 33 kDa (Figure 2.4D). The C2 fraction was also subjected to cation-exchange chromatography, but this fractionation did not successfully purify the enzymes. Table 2.4 shows the specific activity, yield and purification factor for each purification step at pH 5.5. Despite the recovery of the process is too low, this sequence of purification steps was the only one between different attempts in which the proteins were obtained successfully purified.

2.3.4 Acidic activation of cysteine peptidases

Cysteine peptidases were abundantly found in *Tityus serrulatus* MMG. As, in general, cysteine peptidases are synthesised as zymogens (86-88), activation experiments involving these enzymes under acidic conditions were performed. Figure 2.5A shows the activities of the homogenate samples after incubation for 1 hour at 30 °C in solutions with different acidic up to neutral pH values. The hydrolysis of substrate was assayed as previously described in item 2.7 and no differences were observed in incubated or not incubated controls. Activation pattern was obtained after

Figure 2.4 – Purification of two cysteine peptidases from *Tityus serrulatus* MMG



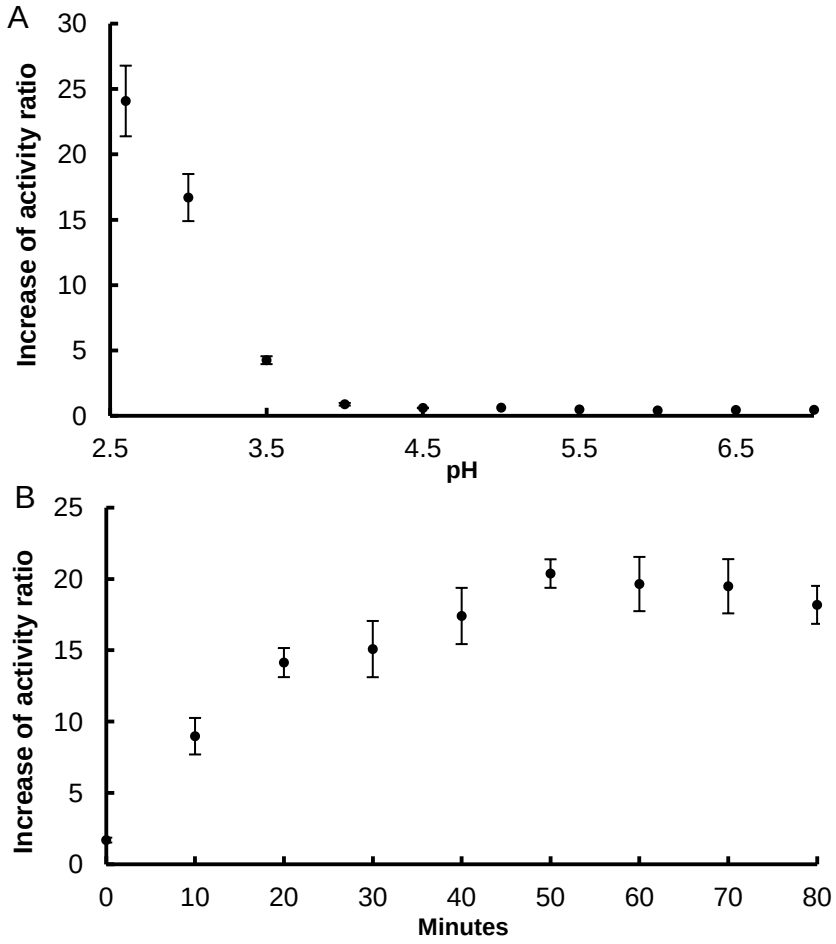
A) Schematic representation of the purification steps. (B) Chromatography of the supernatant from the ammonium sulfate fractionation on a HiTrap Butyl column equilibrated in 50 mM phosphate buffer (pH 6.0). The samples were eluted using a gradient of 1.7-0 M ammonium sulfate in the same buffer. (C) Chromatography of the active fractions from the previous chromatography step (after desalting), represented by open circles (○), on a Resource S column equilibrated with 50 mM citrate-phosphate buffer (pH 5.0) (C1). The samples were eluted in a gradient of 0–0.6 M sodium chloride in the same buffer. (D) SDS-PAGE of the samples exhibiting maximal activity, generated after cation-exchange chromatography, represented by open squares (□). The substrate used to follow the activity in all steps was 10 µM Z-FR-MCA in 0.1 M citrate-phosphate buffer (pH 5.5) containing 3 mM cysteine and 3 mM EDTA. MMTS was added to a final concentration of 1 mM to the fractions exhibiting activity. S, standard (kDa).

Table 2.4 - Purification cysteine endopeptidases from *Tityus serrulatus* MMG

Source	Total Activity (U)*	Total Protein (µg)	Specific Activity (U/µg)	Purification Factor	Yield (%)
Homogenate	566	43500	0.01	1	100
Sulfate Ammonium Precipitation	383	1545	0.24	19	67
Hydrophobic chromatography	164	23	7.1	548	6.7
Cation-exchange chromatography	3.2	0.03	106	8197	0.6

Note: * Substrate used 10 µM Z-FR-MCA diluted in 0.1M citrate-phosphate buffer containing 3.0 mM cysteine and 3.0 mM EDTA.

Figure 2.5 - Acid activation of digestive cysteine endopeptidases from *Tityus serrulatus* MMG



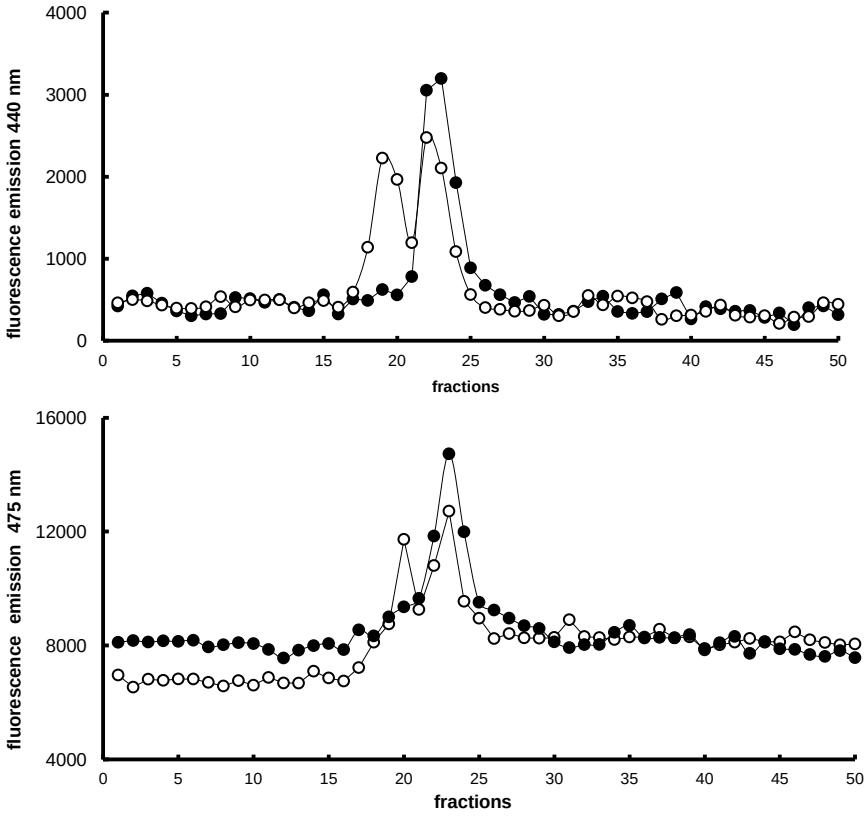
The effect of incubating the MMG homogenate (A) at 30 °C for 60 minutes under different pH conditions. (B) The effect of time on the acidic activation of cysteine peptidases from *Tityus serrulatus* MMG homogenate. After incubation in an acidic buffer (pH 2.6), 2 µl of each enzyme preparation was assayed in 200 µl of 0.1 M CP buffer (pH 5.5) with Z-FR-MCA to measure the activity at a constant pH. The activity increase was calculated as the ratio of the incubated enzyme activity over the non-incubated control activity. All buffers used for the activation (0.1 M CP, pH 2.6 - 7.0) and activity assays contained 3.0 mM cysteine and 3.0 mM EDTA.

incubation at pH 2.6 (Figure 2.5A). Figure 2.5B shows the activation rate indicating that the maximal activity was obtained after at least 50 minutes of incubation at 30 °C at pH 2.6. Loss of activity, most likely due to autolysis, was observed only after 70 minutes of incubation. The same experiment was performed with partially purified samples where, the optimum pH for activation was 3 with an incubation time of 10 minutes at 30 °C (data not shown).

Gel filtration of the activated and non-activated samples resulted in different elution patterns for the homogenate samples (Figure 2.6). The non-activated samples exhibited two activity peaks, at 66 kDa and 44 kDa, independently of the substrate used. The activated samples exhibited only the 44 kDa activity peak, suggesting that the 66 kDa activity peak observed in the non-activated samples

corresponds to the zymogen that was activated during the chromatographic process and acidic activity assay. The molecular mass differences between the active forms obtained using gel filtration (44 kDa) and electrophoresis may be a result of the different methodologies used.

Figure 2.6 – Gel filtration fractionation of the MMG homogenate from *Tityus serrulatus*

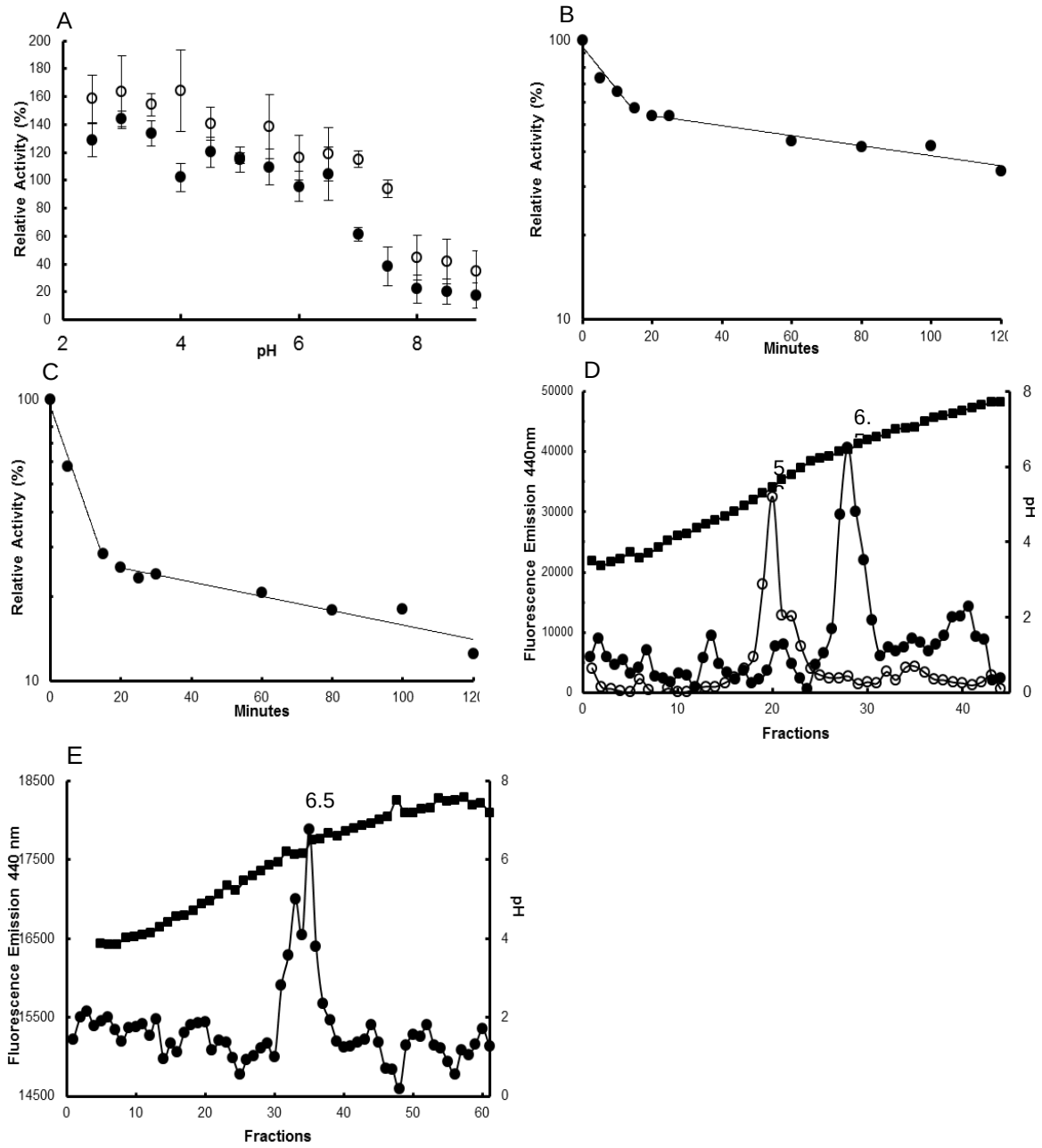


The Superdex G75 column was equilibrated with 20 mM Tris-HCl buffer (pH 7.0). The activated (●) or non-activated (○) fractions were assayed using different endopeptidase substrates to determine whether zymogens were present. (A) Z-FR-MCA, pH 3; (B) hemoglobin, pH 2.8. Buffers used: 0.1 M citrate-phosphate containing 3 mM cysteine and 3 mM EDTA.

2.3.5 Properties of cysteine peptidases

The stability of the activated crude homogenate samples under different pH conditions was tested over a wide range of pH values at 30 °C or -20 °C (Figure 2.7A). The enzymes presented a stability of approximately 100% between pH 3.0 and 6.5. Thermal stability of C1 sample was determined at 60 °C. The remaining activity was measured at pHs 3.0 (7B) and 5.5 (7C) to distinguish between the two activities. The plots shown in Figure 2.7B and 2.7C corroborated the presence of two distinct enzymes. The estimated half-lives of these enzymes were 17 and 167 minutes when the assays were performed at pH 3, whereas their half-lives were 7

Figure 2.7 – Properties of the cysteine peptidases from *Tityus serrulatus* MMG



(A) The effect of pH on the stability of cysteine peptidases from the MMG homogenate. The samples were incubated at 30 °C for 3 h (●) or at -20 °C for 24 h (○). (B and C) Thermal inactivation of C1 at 60 °C. The assays were performed at pH 3.0 (B) or 5.5 (C). (D) The activities of the activated (●) or non-activated (○) samples of partially purified cysteine peptidases against Z-FR-MCA after separation by isoelectric focusing in a polyacrylamide gel. (E) The activity of purified cysp2 at pH 3 against Z-FR-MCA after separation by isoelectric focusing in a polyacrylamide gel. Buffers used with pH 7 or below were 0.1 M citrate-phosphate and at pHs 0.1 M 8-9 Tris-HCl, all containing 3 mM cysteine and 3mM EDTA.

and 132 minutes when the assays were performed at pH 5.5.

The isoelectric points determined using the partially purified activated or non-activated samples demonstrated the presence of different patterns of separation. The isoelectric focusing of the non-activated protein showed a major activity peak with a

pI of 5.3, whereas the activated sample had a major peak with a pI of 6.5 (Figure 2.7D).

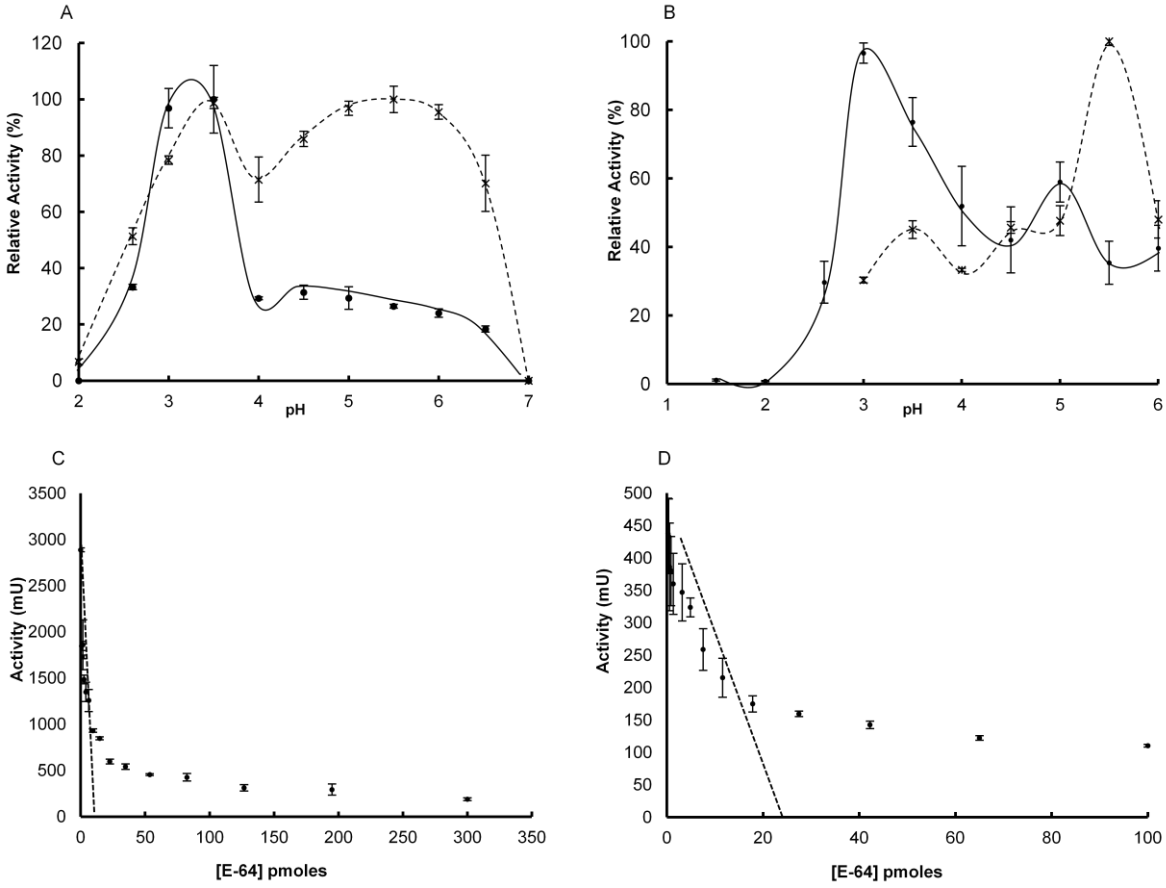
The purified cysp2 and cysp1 samples were also subjected to isoelectric focusing analysis. However, the activity of cysp1 was lost during this process. Cysp2 exhibited an isoelectric point of 6.5 (Figure 2.7E), which is similar to that of the second peak shown in Figure 2.7D.

The effect of pH on the activity of the C1 fraction was determined with Z-FR-MCA. The activity profile indicated the presence of two distinct activities, one at pH 3.0 and other at pH 5.5 (Figure 2.8A). The optimum pH profile of C1 suggests that the majority of the peptidase activity occurs at pH 3.0 if the sample is not activated. The activation of C1 produced one new activity peak at pH 5.5 (Figure 2.8A). The optimum pH (2.8) of hemoglobin hydrolysis (figure 2.2A) is similar to that of the acidic activity observed for the C1 samples (Figure 2.8A). These data were confirmed by determining the effect of pH on the isolated enzymes. Cysp1 exhibited an optimum pH of 5.5, and cysp2 exhibited an optimum pH of 3.0 (Figure 2.8B). Both enzymes were not stable in pHs above 7.

The K_m values (Table 2.5) that were obtained with Z-FR-MCA were 8.4 and 45 µM for cysp1 and cysp2, respectively, whereas K_m values of 0.02 and 0.06 µM were obtained when Abz-FRQ-EDDnp was used.cysp1cysp2 The V_{max}/K_m ratios that were determined with Z-FR-MCA were 390 for cysp1 and 13 for cysp2, whereas the V_{max}/K_m ratios determined when Abz-FRQ-EDDnp was used were 3790 and 660 (min^{-1}). These values indicate that cysp1 is catalytically more efficient than cysp2.

Classification assays using combinations of different substrates and inhibitors indicated that cysp1 and cysp2 could be inhibited by pepstatin. The resulting Lineweaver-Burk plots are shown in Figure 2.9A. The lines in these plots intersect the x-axis to the left of the origin as the pepstatin concentration increases, indicating that the K_{mapp} values increase with higher pepstatin concentrations. The V_{max} values were essentially equal to the control values when 1, 5 or 10 µM pepstatin was used. Nevertheless, the addition of 25 or 50 µM pepstatin resulted in a decrease in V_{max}, which can be observed as the lines crossing the y-axis at higher values (Figure 2.9A). A replot of the reciprocal plot versus the corresponding inhibitor concentration (Figure 2.9B) shows that pepstatin is a competitive inhibitor (Segel, 1975) of cysp1 with a K_i of 40 µM. Cysp2 was also inhibited by pepstatin, but the experiments did not provide a clear pattern for the inhibition in this case.

Figure 2.8 - Effect of the pH on the activity of C1, cysp1 and cysp2 and active site titration of cysp1 and cysp2 with E-64



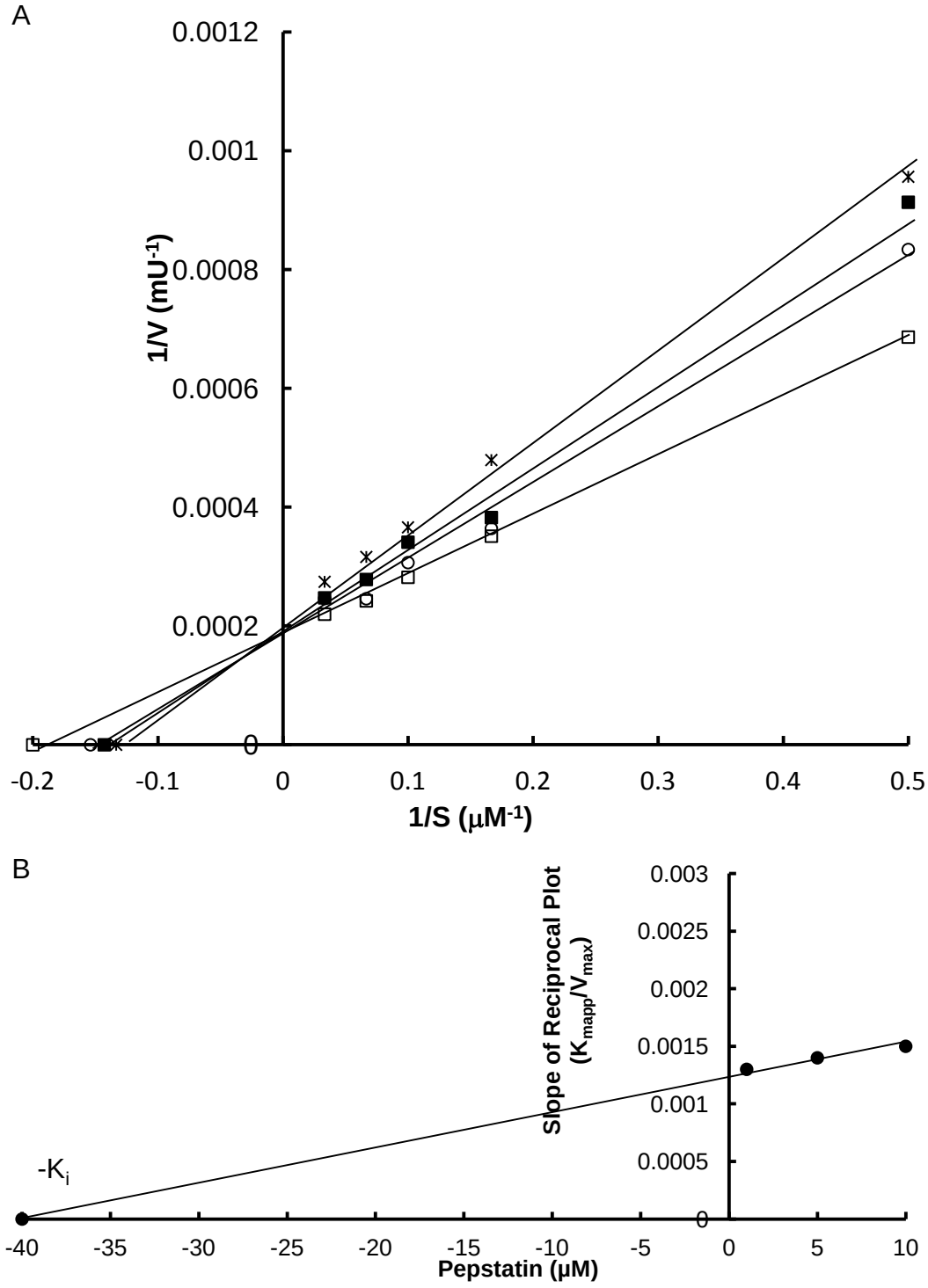
A) The effect of pH on C1 activated (∗) and non-activated (●) samples. (B) The effect of pH on isolated cysp1(∗) and cysp2(●) samples. (C and D) Titration of the purified cysp1 and purified cysp2 from *Tityus serrulatus* with different concentrations of E-64. The assays were performed using Z-FR-MCA in 100 mM citrate-phosphate buffer, pH 5.5 (cysp1) or 3.0 (cysp2), with 3.0 mM cysteine and 3.0 mM EDTA.

Table 2.5 – Kinetic parameters* of cysp1 and cysp2 using 2 different substrates

Substrate	**Cysp1**			**Cysp2**		
	K_m (µM)	**V_{max} (mU/min)**	**V_{max}/K_m**	**K_m (µM)**	**V_{max} (mU/min)**	**V_{max}/K_m**
Z-FR-MCA	8.4 ± 1.5	3,300 ± 240	390 ± 75	45 ± 5.7	614 ± 33.7	13 ± 2
Abz-FRQ-EDDnp	0.022 ± 0.002	83.4 ± 3.8	3790 ± 0.04	0.065 ± 0.005	42.7 ± 1.9	660 ± 1

Note: * Kinetic parameters were (Mean and S.E.M.) were calculated using Enzfitter.

Figure 2.9 – Cysp1 inhibition by pepstatin



(A) Lineweaver-Burk plots obtained with different pepstatin concentrations (Control (□); 1 µM (○), 5 µM (■), 10 µM (∗), 25 µM (●) and 50 µM (▲) pepstatin). The assays were performed using purified cysp1 in 0.1 M citrate-phosphate buffer (pH 5.5) with Z-FR-MCA. (B) Replot of the slopes of the curves obtained from the Lineweaver-Burk plots against pepstatin concentration, indicating a K_i value of 40 µM.

The cleavage point of Abz-FRQ-EDDnp by cysp1 and cysp2 was determined using HPLC and mass spectrometry analysis showing that a Phe residue occupies

the P2 position in both cases (data not shown). Moreover, cysp1 and cysp2 were completed inhibited by 500 nM egg cystatin.

2.4 Discussion

2.4.1 Tityus serrulatus and Arachnida protein digestion

A complete understanding of the physiology of Arachnida digestion has at least three important possible applications: 1) the comprehension of the evolution of the digestive system in Arthropoda and the enzymatic strategies of digestion; 2) the development of control strategies in Arachnida species that cause human health problems, such as ticks and venomous scorpions and spiders; and (3) the use of members of Arachnida, or the chemicals synthesised by arachnids, as insect biological control tools. Despite the potential usefulness of this data, few studies have exploited the Arachnida digestive system and, more specifically, the digestive peptidases involved.

This work reports the identification of cysteine, serine, aspartic and metallopeptidases in the MMG of a scorpion at both transcript and protein levels (Table 2.2). Also, enzyme activity could be measured in a pH range from 2.8 to 9 using different substrates (figures 2.2, 2.8A and 2.8B) and were associated to metallo-, serine and cysteine peptidases with a subsequent biochemical characterization of the latter ones. Our results are in accordance with the reports on literature about this topic. Mommsen (48) published an analysis of the peptidases that are present in the digestive juice of spiders. These results demonstrated the presence of distinct enzymes that are active on natural or synthetic substrates and that have an alkaline optimum pH. As in our study, activities over a trypsin substrate were activated by calcium and he also observed activities which were classified as chymotrypsin and “protease”, the latter probably is related to metallopeptidases. Subsequently, other authors (51, 52) identified metallopeptidases in the digestive juices of *Argiope aurantia*. Although the activity over casein was not clearly related to the astacins in *Tityus serrulatus* MMG, it is likely that these enzymes also cooperate for the hydrolysis of this substrate.

Franta and co-authors have shown, in a series of scientific articles, that protein digestion in ticks involves different endopeptidases, including aspartic peptidases (cathepsin D) and cysteine peptidases (cathepsins L and B and legumains) (61, 66). Cathepsin L and legumain were identified by mass spectrometry in *Tityus serrulatus* but not cathepsin B (Table 2.2). Cysteine peptidase activity was high in the MMG (Table 2.3) and one cathepsin L was purified (cysp 2) and characterized, the other purified protein was likely a cathepsin F (cysp1) as it will be further discussed. Activity over cathepsin B substrate was observed (Table 2.3) thus the fact that this protein was not identified by LC-MS/MS is probably related to the fact that the cathepsin B sequence in our database is just a small fragment. Moreover, activity over Z-FR-MCA in crude extract samples can also be attributed to cathepsin B. Even though cathepsin D was detected by mass spectrometry (Table 2.2) and is relatively abundant in *Tityus serrulatus* MMG (chapter 3) we were not able to detect its activity.

Recently, a digestive chymotrypsin was described in *Scorpio maurus* (65). We could not obtain any significant hit with the N-terminal sequence provided in this latter work in a search against our database but a chymotrypsin-like activity was also observed in the MMG of *Tityus serrulatus* (Table 2.3). Furthermore, it is possible that one or more proteins annotated as trypsins can present chymotrypsin-like activity since a detailed analysis should be still performed in the sequences of their subsites.

2.4.2 The cysteine peptidases

After scanning the enzymes present in *Tityus serrulatus* MMG using different approaches, we biochemically characterised the activity of the cysteine peptidases, since this is the first report of such kind of proteins related to the digestive process in scorpions and due to the high activity observed (Table 2.3). The *Tityus serrulatus* cysteine peptidases present a larger stability at acidic pH values ranging from 3 to 6.5 than other cathepsin L proteins described in the literature. Human recombinant cathepsin L, when incubated for 1 hour at 37 °C, exhibited approximately 80% activity at pH 4-6, whereas the same enzyme purified from the kidney is stable only in the pH range from 4 to (89). A *Tenebrio molitor* cathepsin L purified from the midgut content is stable from pH 4.5 to 6.5 (90).

Two cysteine peptidases were purified to homogeneity and characterised further. The activity peak C1 which originated the purified enzymes cysp1 and cysp2 contained 3 cysteine peptidases: cathepsin L1, cathepsin L2 and cathepsin F. It is still not clear which of these enzymes were purified since LC-MS/MS experiments failed in sequence achievement due to the low yield levels of the purification procedure (Table 2.4). However, due to the normalized spectra counting being higher for cathepsin L1 and cathepsin F, it is more likely that these enzymes were purified rather than cathepsin L2. Cysp1 and cysp2 exhibited a molecular mass of 33 kDa determined by SDS-PAGE. Both enzymes hydrolysed Z-FR-MCA and Abz-FRQ-EDDnp but not Z-RR-MCA. A similar intolerance to Arg at P2 position was also observed in tick cathepsin L (66) and *Manduca sexta* cathepsin F (91). Both cysp1 and cysp2 are stable under acidic conditions. Although they have some similar characteristics, cysp1 and cysp2 show some differences; for example, the optimum pH for cysp1 is 5.5 and while that of cysp2 is 3.0. Cysp1 has an optimum pH similar to those of cathepsin L enzymes from other arthropods that use the same substrate (57, 90, 92, 93) and that of the human cathepsin L and F (89, 94). Cysp2 (Figure 2.8B) exhibited maximal activity at pH 3, similarly to IrCL1 a digestive cathepsin L from the tick *Ixodes ricinus* - optimum pH: 3-4 (66). Said observed a “cysteine catheptic” activity in pH 3 that acts only intracellularly (33), this enzyme is likely cysp2. In both crude extracts and C1 samples the cysteine peptidases were found as zymogen(s) that can be activated under acidic conditions (Figures 2.5 and 2.8A). This is probably because these enzymes are acting in the intracellular digestion and will be activated in the digestive vacuoles after acidification and/or cleavage by other peptidase. Based on the C1 activation data cysp1 is the zymogen present in that sample once after acidification a new activity peak at pH 5.5 can be observed, which is the optimum pH of the purified cysp2.

Cysp1 and cysp2 exhibited K_m values (Table 2.5) similar to other cathepsin L using Z-FR-MCA such as human cathepsin L - 1 µM (95); cruzain - 10 µM (96); *Boophilus microplus* cathepsin L -18.8 µM (57); *Tenebrio molitor* cathepsins 1 (50 µM), 2 (9.6 µM) and 3 (16.4 µM) (90, 93). Human cathepsin F has a K_m of 0.44 µM using Z-FR-MCA (94), which is more similar to the K_m of cysp1 than cysp 2. *Tityus serrulatus* cysp1 and cysp2 presented high affinities to Abz-FRQ-EDDnp (Table 2.5), binding to this substrate even better than human cathepsin L and papain which presented K_m values of 0.3 and 1.7 µM, respectively (Melo *et al*., 2001). Because of

the acidic characteristics of cysp2 and a K_m value about 100 times greater than the one of the human cathepsin F using Z-FR-MCA we believe that this enzyme is one of the cathepsins L identified in C1, likely cathepsin L1. Thus, we believe that cysp1 is cathepsin L1 and cysp2 is cathepsin F, the former will act intracellularly due to its acidic characteristics and based on Said observation as above described (33). We were still not able to determine if cysp2 (cathepsin F) is secreted or act exclusively inside the cells. This enzyme was abundantly found as zymogen (Figure 2.8A) which indicates an intracellular role in the digestive vacuoles since this enzyme still is not activated during the feeding process. However, the hypothesis of resysnthesis to secretory granules can´t be discarded.

Although pepstatin is an irreversible inhibitor of aspartic peptidases with a K_i of 45 pM (97) some cysteine peptidase, calpains (clan CA, family C2) (98) and legumains (clan CD, family C13) (99) are inhibited by pepstatin. Pepstatin apparently inhibits cysp1 via a reversible competitive mechanism, with a K_i of 40 µM (Figure 2.9B). Cysp2 is also inhibited by pepstatin; however, it was not possible to determine the mechanism of this inhibition (data not shown). A reason for this competitive inhibition is the higher magnitude of the calculated K_i (40 µM) for cysp1 is contrast to the cathepsin D K_i (45pM). Nevertheless, the recommended use of pepstatin is in the micro molar range when screening for peptidase activity (70) and till now such kind of inhibition was not reported for the family C1.

5. Conclusions

The scorpion digestive system relies in a multi proteolytic system capable of protein digestion in a pH range from 2.8 to 9. Aspartic, serine, cysteine and metallopeptidases are components of this system and its presence could be confirmed at transcript and protein levels, with the latter three types being also observed by enzyme activity. These results show that scorpion proteolytic system is similar to other arachnids such as ticks and spiders. Cysteine peptidase activity was abundant and 2 enzymes that likely act intracellularly were purified. Cysp1 (cathepsin F) has an optimum pH of 5.5 and it was found as zymogen that can be activated under acidic conditions. Cysp 2 (cathepsin L1) has an optimum pH of 3 and acts intracellularly. This was the first report with protein and mRNA sequencing of cysteine cathepsins acting in the digestive system of a scorpion.

CHAPTER 3 – DIGESTIVE ENZYMES LOCATION REVEALED BY A DEEP ANALYSIS IN THE DIGESTIVE GLANDS OF THE SCORPION *TITYUS SERRULATUS*

3.1 Introduction

Scorpions are amazing ancestor animals with a well succeed evolutionary history over the past 450 million years. Nowadays, most part of the scorpion DNA and protein sequences available are associated to their venom. The large interest in venom knowledge is not only due to the fact that scorpionism is a health problem (63) but also because of the potentiality of venom use in medicines or pesticides (100). Consequently, only few molecular studies were done with other organs of these animals. Nevertheless, a better understanding of the animal molecular physiology could lead to directions of future researches in how to combat them when they became a problem to humans. Particularly, the digestive system is a good target for being a less protected interface between the animal and its environment.

The digestive system of scorpions is a very important organ to the animal physiology presenting a high capacity of nutrients digestion and storage with reports that they can survive until one year starvation (31). The nutrients hydrolysis is achieved with an efficient combination of extra-oral with intracellular digestion. The enzymes are secreted by the prosomal midgut, anterior intestine and its respective digestive glands (Figures 1.2A and 1.2B) to be then regurgitated into the pre-oral cavity starting to liquefy the chewed food. After being filtered by the coxapophyses, the liquefied nutrients will reach the prosomal midgut with the help of the musculature from the pharynx and esophagus. In the midgut and midgut glands the predigested food is absorbed by pinocytosis to then start the intracellular digestion (28). Anatomically, the midgut is divided in prosomal midgut, anterior (mesosomal) intestine and posterior (metasomal) intestine. In the prosomal midgut the digestive glands are trilobed with two lateral and one dorso-medial connections (27). Five pairs of midgut glands are laterally connected to the anterior intestine (Figures 1.2A and 1.2B).

The midgut glands digestive enzymes sequences (protein and DNA), composition (individual to each gland) and subcellular location have been

underexplored so far. The first authors tried to screen and compare the composition of the prosomal with the mesosomal midgut glands by enzyme activity assays (30, 32, 33). Despite some differences could be observed the lack of powerful analytical tools still left lots of questions opened. More modern studies were also made with techniques that only allow an indirect inference of the enzyme classes without obtaining their full sequences. Also they did not attempt to investigate the location of the enzymes in subcellular regions. The exception to that is the immunohistological location of a lipase in the digestive vesicles from *Scorpio maurus* (38). Partial sequences available are the amino-terminal from this same lipase (37); an amino-terminal sequence of a peptidase classified as chymotrypsin-like by Louati and collaborators (34) from the same animal and tissue. Currently there aren´t any full sequences of digestive enzymes from scorpions in the databases except the ones obtained in the previous chapter of this thesis.

The use of transcriptomics and proteomics techniques together showed to be a strong approach not only to identify and sequence DNA and proteins from non-sequenced organisms but also to do the correct assembly (101, 102). This work, for the first time, used a successful combination of next generation sequecing using Illumina® platform with shotgun proteomics in order to 1) acquire a large set of proteins expressed in the midgut and assembly them correctly; 2) quantify the digestive enzymes in each animals under different physiological condition using mass spectrometry; 3) Predict their subcellular location using the software WoLF PSORT and literature data.

3.2 Materials and Methods

3.2.1 Animals and sample obtaining

Adult *Tityus serrulatus* females were obtained from the Laboratory of Arthropods from Instituto Butantan (São Paulo, Brazil). The animals were starved for at least 8 days and then fed with *Acheta domesticus*. After 9 hours eating, the fed animals were dissected whereas the starved ones were dissected 8 days after start feeding.

After anesthetizing the animals in a CO_2 chamber the dissection as performed in a cold isotonic saline solution (300 mM KCl pH 7.0). The MMG (as previously

described in chapters 1 and 2-item 2.2.2) was used to the proteomics and transcriptomics experiments. MMG samples used to RNA extraction were dissected with sterilized instruments in autoclaved saline solution containing 0.1% (v/v) diethyl pirocarbonate (DEPC).

3.2.2 mRNA Library Preparation and Sequencing

All enzymes, primers and buffers cited in this section are from Illumina® unless otherwise specified. RNA extraction was done using TRIzol® reagent (Invitrogen) according the manufacturer instructions. The RNA amount was spectrophotometrically quantified at 260 nm and its purity evaluated by the absorbance ratio 260 nm and 280 nm. The RNA quality and integrity were analyzed submitting 250 ng of the samples to a run in the Agilent 2100 Bioanalyser (Agilent Technologies).

The mRNA was purified according to the protocol described by Illumina® (http:/grcf.jhmi.edu/hts/protocols/mRNA-Seq_SamplePrep_1004898_D.pdf) using magnetic microbeads containing oligo(dT) for rRNA separation. Thereafter, the mRNA was fragmented in the proper buffer and the first cDNA strand synthesis was made using the kit Superscript II® Reverse Transcriptase (Invitrogen). After cDNA first strand synthesis the sample as treated with RNaseH and the second cDNA strand was obtained using DNA polymerase I. The end of the molecules were phosphorylated and the 3´ terminal adenylated using the enzymes T4 PNK and Klenow exo, respectively. The adapters were then linked to the DNA fragments trough the enzyme T4 DNA ligase and the libraries amplified with specific primers to the adapters.

After the library construction, its quality was validated by confirming that most part of the fragments were with size around 260 base pairs (bp) using the Agilent 2100 Bioanalyzer (AgieInt Technologies) with the chip DNA 1000. The libraries were individually quantified through quantitative polymerase chain reaction (qPCR) with the kit KAPA Library Quantification (KAPA biosystems). The library was diluted to a final concentration of 20 pM and each one was clustered and amplified by using the TruSeq PE Cluster Kit v30cBot-HS. The next generation sequencing was performed in a HiScanSQ (Illumina®) using the TruSeq SBS Kit v3-HS (200 cycles) according to

the manufacturer instruction. Each lane had 6 samples and they were sequenced until 10 millions of reads be obtained from each sample.

3.2.3 Bioinformatic tools

The HiScanSq (Illumina®) data obtained was analyzed in four main steps. In the raw data obtainment step it was used the software CASAVA (2011) 1.8.2 provided by Illumina®, which makes the base call from raw data transforming them into fastq format reads followed by the phred´s quality scores. The reads were visualized with the program FastQC 0.10.1 (2012) and then the Agalma pipeline shuffles the reads and removes the ones with low quality (less than 30). Next, vectors, primers and ribosomal RNA sequences are withdrawn after comparison with the Univec and ribosomal RNA databases, both from NCBI (National Center for Biotechnology Information).

The *de novo* assembly was done by the programs Velvet/Oases incorporated to the Agalma pipeline (103, 104). It was done four assembling to all samples with kmers of 31, 41, 51 and 61 that thereafter were merged and the redundant contigs removed. A BLAST (basic local alignment search tool- (105)) was used to identify and annotate the name in assembled sequences using the SwissProt as a database with an e-value threshold of 10^{-10}. The fasta file was filtered with removal of transcripts smaller than 150 bp, splicing variants and low confidence contigs.

The gene ontology was obtained using the program BLAST2GO (106) with the non-redundant NCBI database. Subcellular location was predicted using the software WoLF PSORT (107).

The contig translation based on the DNA coding regions was performed using the software FrameDP v 1.2.0 (108). After using the BLASTX tool against the SwissProt database the program creates a training set to predict the more likely coding DNA sequence (CDS) based on the interpolated Markov models (IMMs). Contigs with less than 50 amino acids were removed. The databases from fed and fasting animals were combined for the MASCOT searches (below) and the redundancy of the possible digestive enzymes was manually removed comparing the sequences. For the rest of the sequences the redundancy was removed using the program BLASTClust with sequence length coverage of 90% and a percent identity threshold of 97% after the MASCOT searches with the partially redundant database.

This prevented discarding isoforms and partial sequences that contain an overlapping region but also different parts of the proteins.

3.2.4 Proteomics procedures

Prior to the proteomics experiment the MMG samples were centrifuged for 20 min at 1,000 x g and the supernatant was collected. The same volume of triplicate samples from supernatant were separated by SDS-PAGE on a 10 well PAGE® Novex 4-12% Bis-Tris Gel (Invitrogen, Bleiswijk, NL) for 30 min at a constant voltage of 200 V using MES-SDS as running buffer. Gel pieces were reduced with 10 mM dithiotreitol (30 min at room temperature), alkylated with 20 mM iodoacetamide (60 min at room temperature in the dark) and digested with sequencing-grade trypsin overnight at room temperature. After digestion, formic acid and DMSO were added (both 5% v/v) to increase peptide recovery. Protein digests were analysed on a reversed-phase nano-HPLC coupled to a LTQ-Orbitrap Velos (Thermo Fisher). An Agilent 1200 series HPLC system was equipped with an in-house packed capillary trapping column (100 µm ID x 20 mm length) and analytical column (50 µm ID and 300 mm length) filled with Reprosil Pur 120 C18-AQ (Dr. Maisch, Ammerbuch-Entringen). Trapping was performed at 5 µl min^{-1} for 10 min in solvent A (0.1 M acetic acid), and a linear gradient from 0 to 40% solvent B (0.1 M acetic acid in 8:2 v/v acetonitrile:water) for 240 min at a flow rate of 100-150 nl min^{-1} was used to elute the peptides. The column effluent was directly electro-sprayed in the ion source of a LTQ-Orbitrap Velos (Thermo Fisher), which was programmed to operate in data-dependent mode, automatically switching between MS and MS/MS. Survey full-scan MS spectra were acquired from m/z 400 to 1,500 in the Orbitrap analyser at a resolution of 30,000 at m/z 400 after accumulation of ions to a target value of 1×10^6. The twenty most intense multiply charged ions above a set threshold of 5,000 were fragmented in the linear ion trap using collision-induced dissociation (CID) after accumulation to a target value of 1×10^4. The isolation width was set to 2.5 amu, the normalized collision energy at 35% and dynamic exclusion was 90 sec. All raw data files were processed into peaklists using the software ReAdW 4.3.1 and then deconvoluted using the program MS-deconv (109). The files generated from MS-deconv were then analyzed by MASCOT (Matrix Sciences). After that the MASCOT searches were loaded in the software Scaffold 4 and analyzed with X!Tandem (110).

The false discovery rate (FDR) threshold was 0.1% using a random decoy database and identified proteins chosen had at least 2 peptides assigned. Label-free quantitative analysis was based in the normalized spectra counting using the software Scaffold 4.

3.3 Results

3.3.1 Transcriptome and proteome general analysis

The data of *de novo* assembly results from the RNA-seq of the midgut and midgut glands (MMG) is summarized in table 3.1. About 30 and 36% of the contigs from fasting and fed animals presented BLASTX hits (Table 3.1), respectively, and its relation to the sequence length distribution is depicted in figure 3.1A and 3.1B, which shows that sequences under 1,600 bp lengths are more suitable to do not have a BLAST hit. The gene ontology (GO) was studied using the software BLAST2GO (106). MMG from fed and fasting animals presented 7,250 and 6,350 sequences respectively. Figure 3.2 (A, B and C) shows the GO graphics related to the biological process, cellular component and molecular function from the transcriptomics data acquired from fasting conditions.

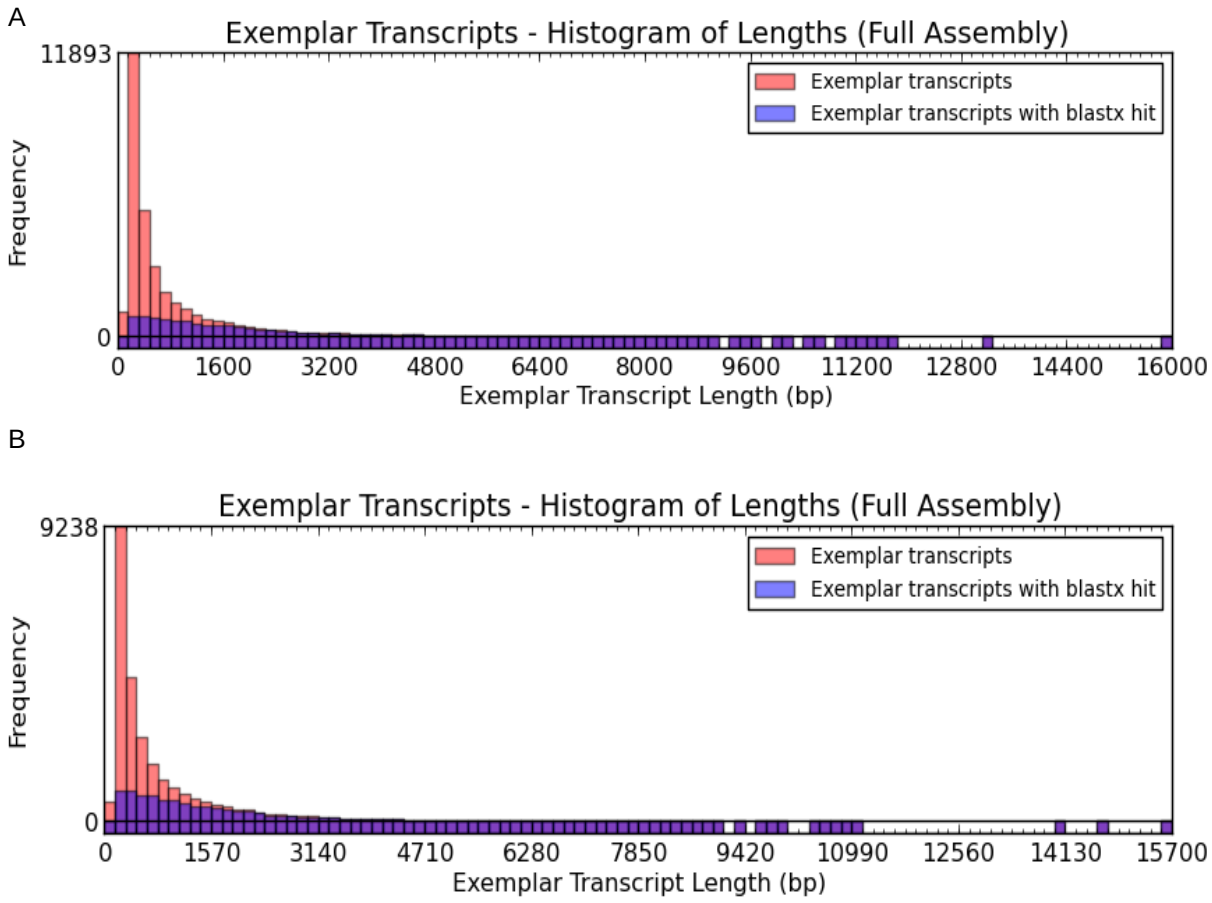
Table 3.1 - Summary of *de novo* assembly results

Condition	Read Pairs	Number of Contigs	Mean Length (bp)	N50 Length (bp)	BlastX Hits
Fasting	13,834,783	32,249	797.5	1,496	9,857
Fed	12,598,159	27,172	824.6	1,499	9,903

After obtaining the new database of the expressed sequences in the MMG of the scorpion *Tityus serrulatus* under two different physiological conditions, the proteomics experiments were performed using the translated nucleotide sequences for the searches as described in the methodology section. A total of 476 and 501 proteins were identified in respectively fed and starved animals with at least 2 peptides and 0.1% of false discovery rate. The GO was recovered from the molecules identified at the protein level in fed scorpions and it is exhibited in figure3.2 (D, E and F). The GO terms from the contigs and proteins identified in the MMG of

fed scorpions are omitted but they present a similar general pattern. The best BLAST hits results are related to the sequences of the tick *Ixodes scapularis* followed by other invertebrates, which is in accordance with the phylogenetic group, Arachnida, of the studied scorpion (data not shown). Sequences obtained in the proteomics analysis without BLAST hits summed 6.6 and 3% of all detected proteins in contrast to the 64 and 70% of unidentified contigs from MMG samples of fed and fasting scorpions, respectively.

Figure 3.1 - Histogram of lenghts and BLASTX hits of the transcriptome assembled contigs from the midgut and midgut glands of *Tityus serrulatus*



A) Fasting scorpion. B) Fed scorpion.

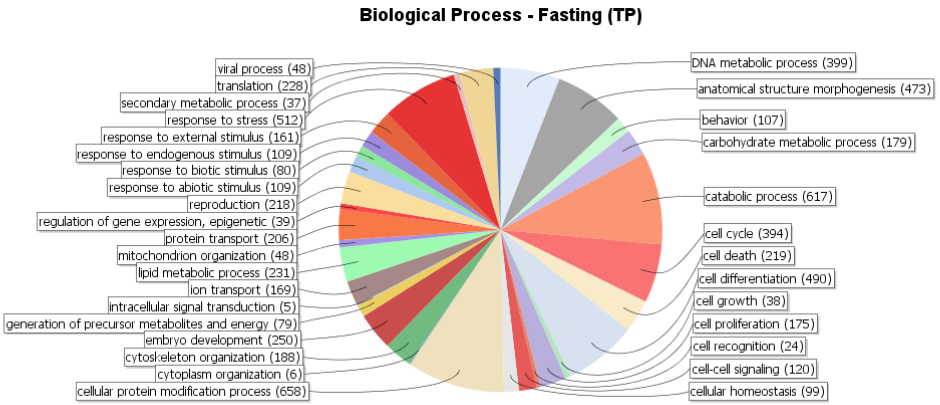
3.3.2 Possible digestive enzymes identification

A total of 235 different enzymes with a possible digestive role were found in the MMG of the scorpion *Tityus serrulatus* at the mRNA level (Table 3.2 and figure 3.3A) and 43 distinct enzymes were identified at the protein level (Figure 3.3B and

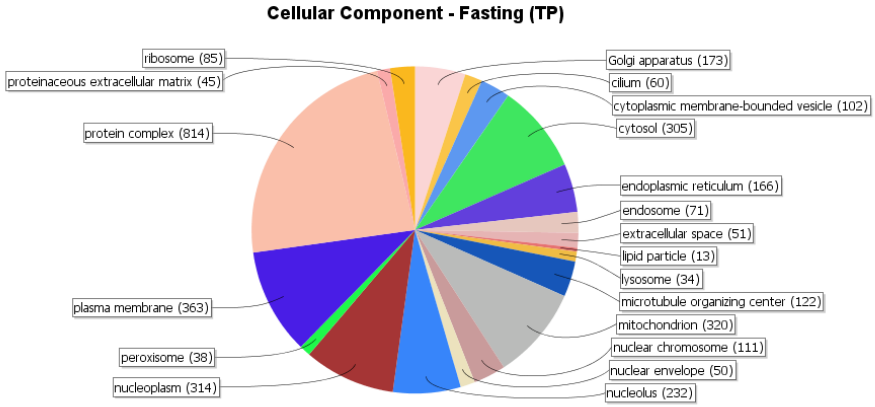
table 3.3). The transcriptome data shows a large variety of enzymes with 32% exopeptidases, 31% carbohydrases, 20% lipases and 17% endopeptidases.

Figure 3.2 - Gene ontology terms of biological process, molecular function and cellular component of the transcriptome (fasting) and proteome (fed) sequences obtained in the midgut and midgut glands of *Tityus serrulatus*

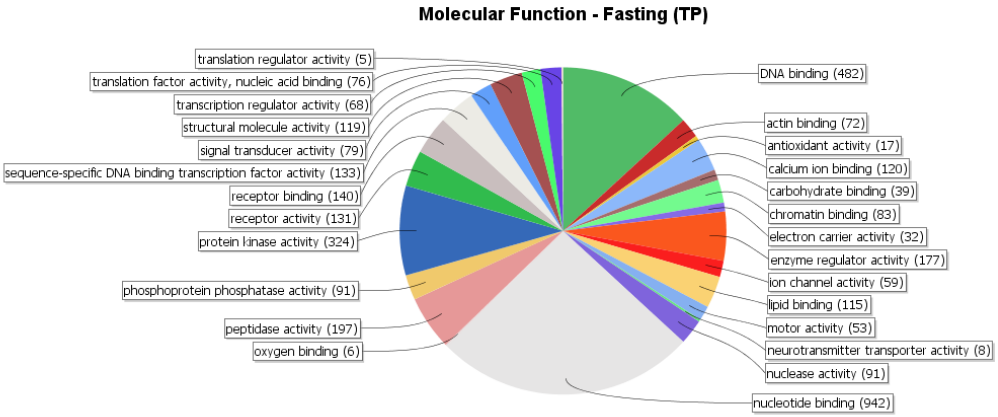
A



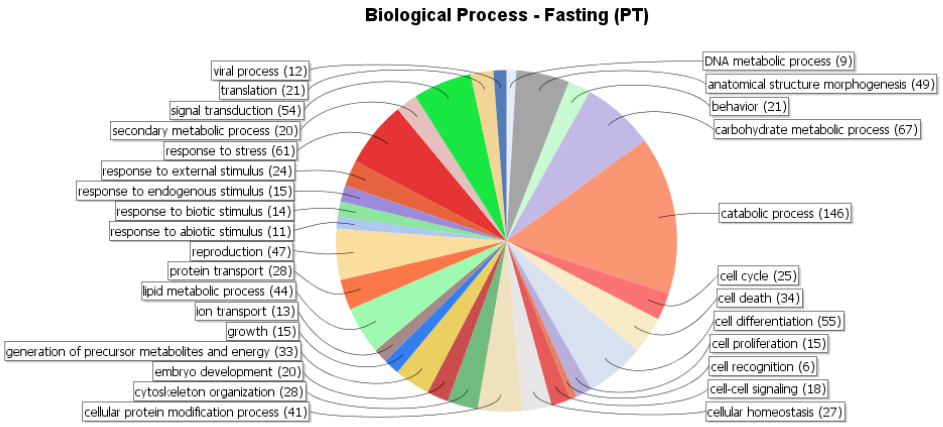
B



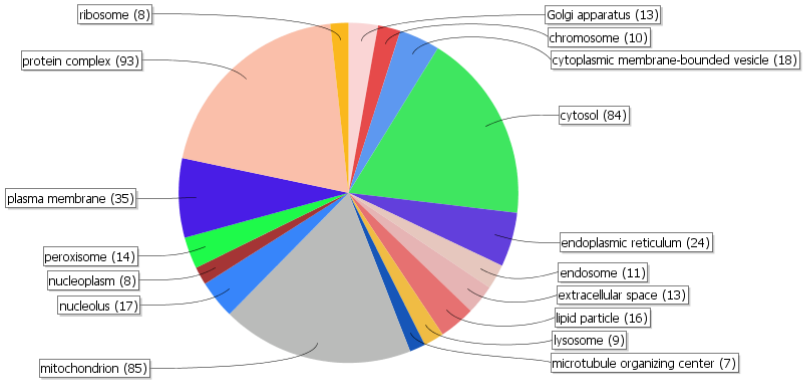
C



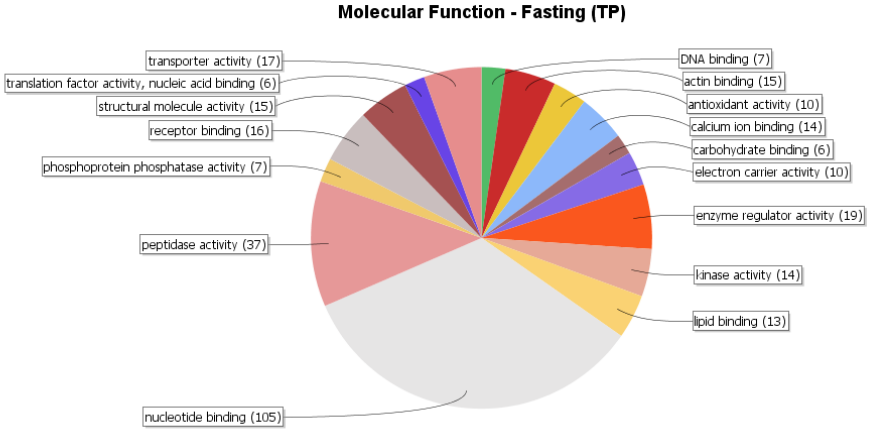
D



E



F



GO terms from the protein sequences obtained in the MMG of fasting animals. A, B and C – transcriptome (TP). D, E and F – proteome (PT).

For the initial protein digestion all the four groups of peptidases are represented (Figure 3.3A). Metallopeptidases are the most abundant enzymes with 17 sequences including16 astacins and one zinc metallopeptidase. One of these astacins contains a MAM domain whereas in another one MAM and CCP domains are present (Table 3.2). Cysteine peptidases are the second largest group with 16

Table 3.2 - Possible digestive enzymes identified after the transcriptomics experiment in the midgut and midgut glands of the scorpion *Tityus serrulatus*

Exopeptidases	Number of different copies
Aminopeptidase N	2
Xaa-Pro aminopeptidase	9
Methionine aminopeptidase	3
Dipeptidyl aminopeptidase	4
Glutamyl aminopeptidase	2
Aminopeptidase O	3
Leucyl-cystinil aminopeptidase	1
Aminopeptidase NPEPL1	1
Carboxypeptidase N subunit 2	9
Carboxypeptidase 1	3
Carboxypeptidase M	2
Zinc carboxypeptidase A 1	1
Glutamate carboxypeptidase 2	1
Carboxypeptidase Q	1
Carboxypeptidase B	1
Carboxypeptidase D	1
Carboxypeptidase E	1
Lysosomal Pro-X carboxypeptidase	1
Dipeptidyl peptidase 10	2
Dipeptidyl peptidase 9	1
Dipeptidyl peptidase 4	2
Dipeptidyl peptidase 3	2
Dipeptidyl peptidase 2	1
N-acetylated-alpha-linked acidic dipeptidase 2	15
Xaa-Pro dipeptidase	2
Alpha-aspartyl dipeptidase	1
Tripeptidyl-peptidase 2	3

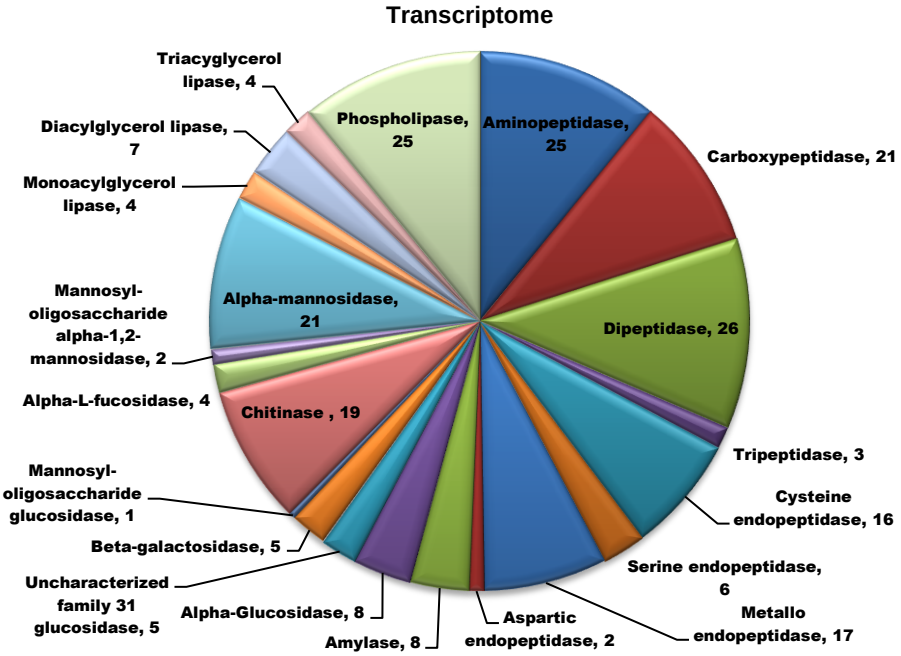
Endopeptidase	Number of different copies
Cathepsin L	11
Cathepsin B	1
Cathepsin F	1
Cathepsin O	2
Cathepsin D	2
Legumain	1
Astacin	14
MAM domain-containing astacin	1
MAM and CCP domains-containing astacin	1
Zinc metallopetidase	1
Chymotrypsin/Trypsin	6

Carbohydrases	Number of different copies
Alpha-amylase	8
Beta-galactosidase	5
Alpha-glucosidase	8
Uncharacterized family 31 glucosidase	5
Mannosyl-oligosaccharide glucosidase	1
Chitinase	19
Alpha-L-fucosidase	4
Mannosyl-oligosaccharide alpha-1,2-mannosidase	2
Alpha-mannosidase	21

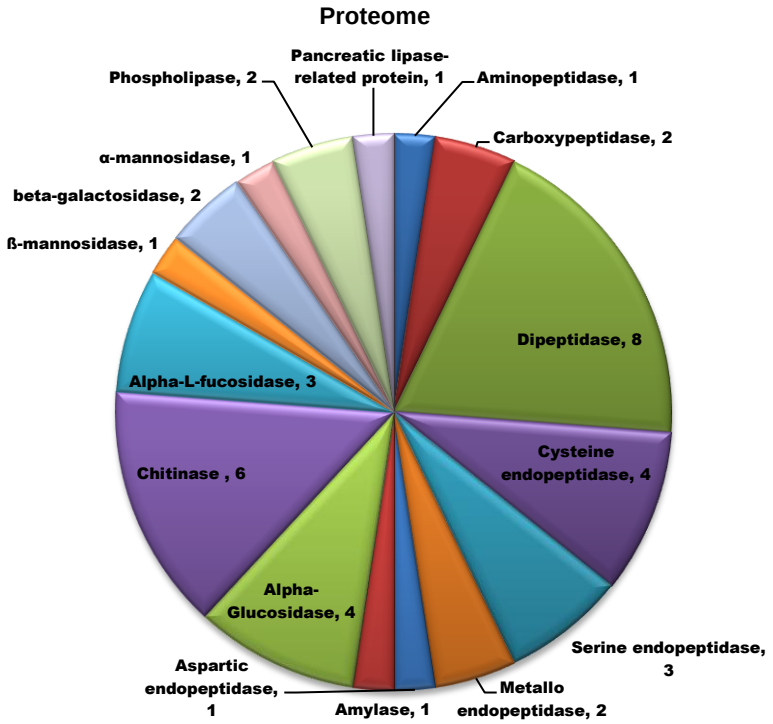
Lipases	Number of different copies
Monoacylglycerol lipase	4
Diacyglycerol lipase	7
Pancreatic-like triacylglycerol lipase	2
Gastric-like triacylglycerol lipase	2
Pancreatic lipase-related protein 2	6
Hormone-sensitive lipase	2
Acid lipase	1
Patatin-like phospholipase	2
Phospholipase A2	12
Phospholipase B-like 2	4
Phospholipase D1	4
Phospholipase D3	1
Other phospholipases	2

Figure 3.3 - Pie charts of the possible digestive enzymes identified in the transcriptome and proteome of the midgut and midgut glands of *Tityus serrulatus*

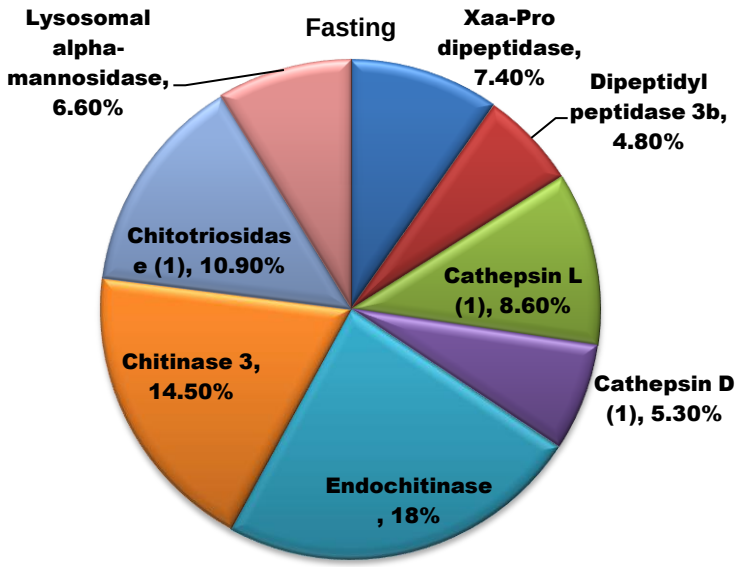
A



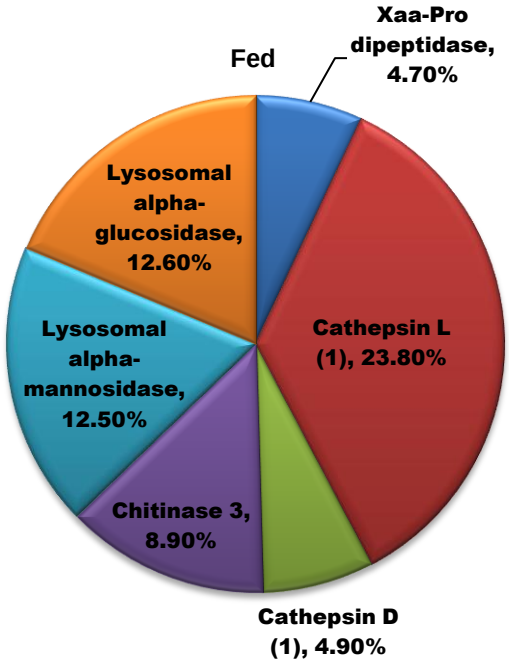
B



C



D



A and B) Respectively transcriptome and proteome general digestive enzymes number of different sequences. C and D) Respectively fasting and fed animals normalized spectra counting percentage of the possible digestive enzymes identified in the MMG. Only proteins with percentage higher than 4.5 are displayed.

sequences, in between then there are 11 cathepsins L, two cathepsins O, 1 legumain, 1 cathepsin B and 1 cathepsin F. Six serine peptidases with the catalytic residues from the trypsin family were found of which 3 contain the domains CUB and LDL. Lastly, 2 cathepsins D-like aspartic peptidases contigs were identified. For the proteome data table 3.3 exhibits the identified proteins, their normalized spectra counting and in which physiological condition at the mRNA and protein levels they were found. As for the transcriptome data, the proteins identified by mass

spectrometry are shown in a graphic with the number of each different enzyme group (Figure 3.3B). The proteomics experiment confirmed the presence of 9 endopeptidases previously described, with the identification of 2 cathepsins L, 2 astacins, 2 CUB and LDL domains-containing trypsin, 1 cathepsin F, 1 legumain and 1 cathepsin D (Table 3.3).

The exopeptidases, with a total of 75 proteins have almost twice more molecules than the endopeptidases. Twenty six dipeptidases, 25 aminopeptidases, 21 carboxypeptidases and 3 tripeptidases were detected at the mRNA level. Carboxypeptidase N and Xaa-Pro aminopeptidase presented the widest number of distinct sequences (Table 3.2, Figure 3.3A). Exopeptidases were also identified by proteomics including 8 dipeptidases, 2 carboxypeptidases and 1 aminopeptidase (Table 3.3, Figure 3.3B).

The mRNA sequences of carbohydrases comprise 73 different molecules (Table 3.2 and Figure 3.3A) which are mainly constituted of chitinases (19 sequences) and alpha-mannosidases (21 sequences). At the protein level, chitinase sequences (6 in total) were the most abundant carbohydrases identified followed by the presence of 4 alpha-glucosidases and 3 alpha-L-fucosidases (2 of them are isoforms with small differences), 2 beta-galactosidases, 1 alpha-mannosidase, 1 beta-mannosidase and 1 alpha-amylase (Figure 3.3B). The majority of lipolytic enzymes at the mRNA level are formed by 25 sequences of phospholipases but also monoacyl, diacyl- and triacylglycerol lipases were found with 4, 7 and 4 molecules each one, respectively (Figure 3.3A). In table 3.2 it is possible to see the diversity of the different phospholipases sequenced. Curiously, from a total of 46 sequenced lipases only 3 could be detected after the LC-MS/MS experiments (Table 3.3 and Figure 3.3B).

3.3.3 Label-free quantitative analysis

After the general identification of different enzymes found at the protein level, a label-free quantitative analysis was performed. The quantitation was done using the normalized spectra counting after submitting the samples from S1 to the LC-MS/MS experiments and the percentages are in relation only to the possible digestive enzymes found (Table 3.3). The quantitative graphics from figures 3.3C and 3.3D show all the enzymes with a percentage of spectra counting higher than 4.5%.

The most abundant enzymes in both, fasting and fed animals, were carbohydrases and exo/endopeptidases (Figure 3.3C and 3.3D). Forty three percent of the fasting animals MMG enzymes were chitinases with three different isoforms contributing to this sum (Figure 3.3C). Another carbohydrase found with high protein levels (6.6%) was alpha-mannosidase. The hydrolases Xaa-Pro dipeptidase and dipeptidyl peptidase with 7.4 and 4.8%, respectively, are the two exopeptidases in the top of the spectra counting for this type of molecule (Figure 3.3C). The percentage of cathepsins L1 and D1 are 8.6 and 5.3, respectively. After feeding cathepsin L1 is the most abundant enzyme with 23.8% of the total spectra, cathepsin D1 presented 4.9% (Figure 3.3D). Chitinases were no longer the most identified carbohydrases but still represented 8.9% of the enzymes, both alpha-mannosidase and alpha-glucosidase are about 12.5% each one of the total spectra counting (Figure 3.3D). With 4.7% Xaa-Pro dipeptidase was the exopeptidase most represented in the MMG of fed animals (Figure 3.3D).

3.3.4 Subcellular prediction

We were able to identify, by transcriptomics, 235 different hydrolases that could present a digestive role in the midgut and midgut glands of the scorpion *Tityus serrulatus* (Table 3.2) of which 43 were confirmed by mass spectrometry (Figure 3.3B). In order to do some prediction of secretion or cellular localization of these digestive enzymes, *in silico* analysis of subcellular localization prediction was done with the use of the program WoLF PSORT (107). Table 3.3 shows the scores calculated for the subcellular prediction. Additionally, sequence alignment and literature data will also be used for the analysis of *in silico* results and interpretation.
Based on the prior knowledge that scorpions present extra-oral digestion combined with an intracellular phase of digestion (28) it will be assumed that digestive enzymes are the ones with extracellular and lysosomal signals. Databases on molecular localization prediction are mainly based on mammalian and yeast data and probably present few arachnid sequences. Thereby, even low k-NN values can be a good evidence of protein location. In addition to that, GO terms from extracellular space and lysosomal sequences were used in order to corroborate WoLF PSORT data (Table 3.3).

Moreover we have tested WoLF PSORT locations for two known lysosomal enzymes: Human (95) and *Tenebrio molitor* (90) cathepsins L and both presented higher k-*NN* values for extracellular space than for lysosome (data not shown).

Table 3.3 – Possible digestive enzymes identified in the proteome and the physiological condition in which they were found at the mRNA and protein levels (continues)

	mRNA		NSC		WoLF PSORT (*k*-NN)									GO	
Exopeptidase	**Fed**	**Fasting**	**Fed**	**Fasting**	**Ex**	**Ly**	**E.R**	**Cy**	**Pl**	**Go**	**Mi**	**Nu**	**Pe**	**Ex**	**Ly**
Xaa-Pro dipeptidase	yes	yes	40	263	-	-	-	17.5	-	-	9	10.5	-	no	no
Dipeptidyl peptidase 3b	no	yes	16	170	-	-	-	13	-	-	-	17	-	no	no
Dipeptidyl peptidase 3a	yes	no	20	139	-	-	-	11	-	-	9	11	-	no	no
Alpha-aspartyl dipeptidase	yes	no	-	25	-	-	-	9.5	-	-	12	8	7	no	no
Xaa-Pro aminopeptidase	yes	yes	2	-	-	-	-	3	2	-	15	6	3	no	no
N-acetylated-alpha-linked acidic dipeptidase 2 (1)	yes	yes	8	-	5	-	-	24	-	-	-	15.5	-	no	no
N-acetylated-alpha-linked acidic dipeptidase 2 (2)	yes	yes	8	-	3	-	19	-	6	-	-	-	2	no	no
Carboxypeptidase Q	yes	no	6	-	8	-	20.5	-	-	12	-	-	-	yes	yes
Lysosomal Pro-X carboxypeptidase	yes	yes	4	2	24	-	3	-	2	-	-	-	-	no	yes
Dipeptidyl peptidase 9	yes	yes	-	3	-	-	-	15	-	-	-	18	-	no	no
N-acetylated-alpha-linked acidic dipeptidase 2	yes	yes	2	-	-	-	-	17	9.7	-	-	16.3	-	no	no

	mRNA		NSC		WoLF PSORT (*k*-NN)									GO	
Endopeptidase	**Fed**	**Fasting**	**Fed**	**Fasting**	**Ex**	**Ly**	**E.R**	**Cy**	**Pl**	**Go**	**Mi**	**Nu**	**Pe**	**Ex**	**Ly**
Cathepsin L1*	yes	no	203	305	--	--	--	--	--	--	--	--	--	no	no
Cathepsin D1	No	yes	42	187	25	-	4	-	-	-	-	-	-	no	yes
Astacin 2*	No	yes	13	122	--	--	--	--	--	--	--	--	--	yes	no
Cathepsin L 2	yes	Yes	16	62	30	-	-	-	-	-	-	-	-	no	no
Cathepsin F	yes	no	12	28	30	-	-	-	-	-	-	-	-	no	no
Astacin 5a*	yes	yes	-	32	--	--	--	--	--	--	--	--	--	no	no
CUB and LDL domains-containing Trypsin 1	yes	no	-	10	15	-	5	3	3	-	-	2	3.5	no	no
CUB and LDL domains-containing Trypsin 2	yes	yes	-	8	26	3	-	-	-	-	-	-	-	no	no
CUB and LDL domains-containing Trypsin 3	No	yes	7	-	20	-	-	3	-	-	4	-	3	no	no
Legumain	yes	no	6	-	26	5	-	-	-	-	-	-	-	no	yes

	mRNA		NSC		WoLF PSORT (*k*-NN)									GO	
Carbohydrase	**Fed**	**Fasting**	**Fed**	**Fasting**	**Ex**	**Ly**	**E.R**	**Cy**	**Pl**	**Go**	**Mi**	**Nu**	**Pe**	**Ex**	**Ly**
Endochitinase	yes	yes	15	635	23	3	-	2.5	-	-	-	-	2.5	yes	no
Chitinase 3	yes	yes	76	516	20	-	9	-	-	-	-	-	-	yes	no
Chitotriosidase 1	no	yes	9	385	25	-	4	-	-	-	-	-	-	yes	no
Lysosomal alpha-mannosidase	yes	yes	107	234	-	-	-	17	-	-	-	12	-	no	no

Table 3.3 – Possible digestive enzymes identified in the proteome and the physiological condition in which they were found at the mRNA and protein levels. (conclusion)

	mRNA		NSC		WoLF PSORT (*k*-NN)								GO	
Carbohydrase	**Fed**	**Fasting**	**Fed**	**Fasting**	**Ex**	**Ly**	**E.R**	**Cy**	**Pl**	**Go**	**Mi**	**Nu**	**Pe**	**Ex**	**Ly**
Lysosomal alpha-glucosidase	yes	yes	108	66	-	-	6	15.5	4	-	-	3.5	-	no	no
Beta-galactosidase 1	yes	Yes	20	76	7	-	-	6	7	-	10.5	-	4	no	yes
Alpha-L-fucosidase 1a	no	Yes	27	68	10	2	-	-	-	-	15	-	3	no	no
Chitotriosidase (2)	no	yes	-	37	12	-	11	-	-	-	-	-	5	yes	no
Alpha-L-fucosidase (1b)*	yes	No	23	10	-	-	-	-	-	-	-	-	-	yes	no
Chitotriosidase (3)	yes	yes	-	29	14	2	14	-	-	-	-	-	-	yes	no
Neutral aplha-glucosidase AB*	yes	no	-	22	25	-	2	-	2	-	-	1.5	-	no	no
Alpha-L-fucosidase (2a)	no	yes	20	-	14	3	3	-	-	-	8	-	3	yes	no
Beta-galactosidase (2)	yes	yes	-	18	7	5	5.5	-	7	5.5	-	-	-	no	yes
Glucosidase 2 subunit beta (1)	no	yes	17	-	16	-	-	9	-	-	4	-	-	no	no
Acidic chitinase (1)	no	yes	2	9	22	-	2	3	-	-	-	2	-	yes	no
Beta-mannosidase	no	yes	2	5	-	-	-	23.5	-	-	-	5.5	-	no	yes
Glucosidase 2 subunit beta (2)	no	yes	6	-	11	-	-	4.5	-	-	11	4.5	-	no	no
Acidic chitinase (2)	no	yes	-	4	-	-	-	13.5	-	-	14	-	4	yes	no
Alpha-amylase 1	yes	yes	2	-	8	-	16	-	-	-	2.5	2	2.5	yes	no
	mRNA		**NSC**		**WoLF PSORT (*k*-NN)**								**GO**	
Lipase	**Fed**	**Fasting**	**Fed**	**Fasting**	**Ex**	**Ly**	**E.R**	**Cy**	**Pl**	**Go**	**Mi**	**Nu**	**Pe**	**Ex**	**Ly**
Pancreatic lipase-related protein 2	no	yes	-	37	25	3	-	-	1	-	-	-	-	yes	no
Phospholipase B-like 2	yes	no	13	18	-	-	-	12	-	-	-	19	-	no	yes
Phospholipase D3	yes	yes	2	16	-	-	-	8	17	-	4	-	-	no	no

Notes: The mass spectrometry data is from 3 different biological samples. The normalized spectra counting (NSC) is shown as a quantitative value and it was done after putting all the MASCOT searches for each physiological condition together with a 0.1% FDR and at least 2 peptides identification. In the middle a prediction of the subcellular location using WoLF PSORT and in the right the presence of GO term related to extracellular space and lysosome is displayed. The numbers in the brackets are regarded to different isoforms. *k*-NN, *k*-nearest neighbor classifier from PSORT; Ex, extracellular space; Ly, lysosome; E.R, endoplasmatic reticulum; Cy, cytosol; Mi, mitochondria; Nu, nucleus; Pe, peroxisome, Pl, membrana plasmática.

The lysosomal Pro-X carboxypeptidase had a high score for secretion and none for lysosome using WoLF PSORT analysis. However, in the GO analysis, the sequence was associated with lysosome. BLAST analysis of this sequence against the Uniprot database resulted in a high identity (e-value $1x10^{-169}$) with the known human lysosomal Pro-X carboxypeptidase. Thus, it is more likely that the scorpion enzyme is also inside lysosomes. These analysis indicated that the *in silico* analysis

is just a first approach to digestive enzymes distribution which will have to be confirmed by imunocitolocalization studies.

All the complete endopeptidases identified by transcriptomics and proteomics analysis exhibited high *k*-NN values for extracellular location (Table 3.3). Only 2 enzymes CUB and LDL domains-containing trypsin 2 and legumain, also had scores for lysosome. Due to the usual alkaline characteristics of serine peptidases also confirmed to *Tityus serrulatus* trypsin activity (it was measured trypsin-like activity in pH 8, chapter 2), it is more likely that it is secreted and not present inside the lysosome. Furthermore, CUB domains are usually associated with extracellular proteins (111). Legumain also had the lysosome GO term associated to its sequence and it was shown that in the tick *Ixodes ricinus* this endopeptidase acts inside the digestive vacuoles (60). Hence *Tityus serrulatus* legumain is probably a lysosomal enzyme as well. Cathepsin D was predicted as a secreted molecule by WoLF PSORT and as lysosomal by GO term. This enzyme is commonly associated with intracellular digestion (61, 112) but it also can act extracellularly (112). Ticks present cathepsin D activity intracellular digestion role and then, due to phylogenetic proximity, it will be assumed that scorpion cathepsin D is also intracellular. CUB and LDL domains-containing trypsin 1 and 3 are likely secreted enzymes despite scores for other locations are also observed. Cathepsin F and cathepsin L2 presented score only for extracellular space, however activity assays demonstrated that these both enzymes do not hydrolyze the substrate in pH 7 with optimum ph 5.5 (Figure 2.8A). Thus it will be assumed that these enzymes could be secreted or lysosomal. Even though cathepsin L1 is incomplete in the N-terminal region, it will be considered lysosomal due to the discussion presented in chapter 2. Astacins 2 and 5a sequences are also incomplete. Nevertheless, these enzymes are normally active on alkaline pH and were found as secreted enzymes in the digestive juice of the spiders *Argiope aurantia* (52) and *Nephilengys cruentata* (chapter 5). Thereby *Tityus serrulatus* astacins will be considered as secreted enzymes.

Endochitinase, chitinase 3, acidic chitinase 1, chitotriosidase (1, 2 and 3), alpha-L-fucosidase 2a and neutral alpha-glucosidase presented high signals for extracellular space in WoLF PSORT and also at the GO term, so they are probably secreted enzymes. Acidic chitinase 1 and alpha-L-fucosidase 1b had high scores for extracellular space and it will be assumed their secretion although this cellular location is not corroborated by the GO term. Lysosomal alpha-mannosidase and

lysosomal alpha-glucosidase, as well as, beta-galactosidase 1 and 2 and beta-mannosidase are possibly lysosome enzymes (BLAST identity analysis) Despite the small *k*-NN value for secretion and high value for endoplasmatic reticulum, alpha-amylase unlikely belong to this organelle and, presented the GO term for extracellular space, which will be its supposed location. Spiders also present secreted alpha-amylases as observed in *Nephilengys cruentata* (chapter 5), *Tegenaria atrica* and *Cupiennius salei* (113, 114).

The pancreatic lipase-related protein score for extracellular space is 25 and the GO term confirm the same location, indicating a possible secretion. Also lysosomal score was observed for this same enzyme and between all lipase sequences identified in this work after the RNA-seq this is the most similar with the N-terminal fragment of the purified digestive lipase from *Scorpio maurus* (37), with 54% identity and 61% similarity (data not shown). In his study, the author found this enzyme only in the digestive vacuoles and not in the secretory granules (38) so it will be proposed that it also could be a lysosomal enzyme. Phospholipase B-like 2 is a lysosomal enzyme in humans (115) and it was mapped to the GO term lysosome, thus it is likely that it is a lysosomal enzyme.

3.3.5 Other molecules identified in the midgut and midgut glands

Regardless the molecules related to organism homeostasis and the possible digestive enzymes above described some proteins that are indirectly associated with digestion were also identified at the protein level. Proteins related to the vesicular trafficking such as clathrin (light and heavy chains), Rab (1a, 2, 5c, 11a and 14), sorting nexin (2, 6, 12 and 17) and proteins related to vesicular acidification (V-type proton ATPase subunits A and B) could be detected. Two MAM and LDL-receptor class A domain-containing were identified in the MMG probably related to endocytosis. The peptidase inhibitors cystatin and serpin as well as one beta-galactosidase activator (lysosomal protective protein) were also present. Moreover 3 different toxins (U_{24}-ctenitoxin-Pn1a) with similarity to cysteine peptidase inhibitors from the venom of the spider *Phoneutria nigriventer* were found expressed and as proteins in the midgut glands of the scorpion *Tityus serrulatus*.

3.4 Discussion

3.4.1 General analysis of the transcriptome and proteome results

The gene ontology graphics presented in figure 3.2 are the proof of concept that the combination of two high throughput techniques, such as next generation sequencing and shotgun proteomics, is very efficient to do a *de novo* assembly of the expressed sequences from a tissue of a non-sequenced organism. The GO multi level pie charts shown in figure 3.2 from both transcriptome and proteome data are important parameters about the quality of the experiments. For the three root terms (biological process, molecular function and cellular component) all the graphics from figure 3.2 contain many different GO terms associated to a large variety of roles. This is a clear evidence that both approaches were well succeed in doing a broad mRNA and protein sequencing contemplating molecules involved in the main functions of the tissue under investigation.

3.4.2 The identified enzymes involved in digestion

Although the limitations of using bioinformatics tools to predict subcellular location, the data obtained by these *in silico* analysis will be used for the proposal of a mechanism of digestion in the scorpion *Tityus serrulatus*. Mass spectrometry allowed the identification of 43 proteins. Among these, 30 may have a digestive role as showed before (section 3.3.2 of the results), which allowed the design of an improved sketch of the molecular model of digestion in scorpions. All the secreted proteins identified presented their respective catalytic residues with the only exception to chitinase 3. The lost of the catalytic residues does not prevent the binding to chitin, which is a feature of the chitolectin family. This protein also contains one peritrophin domain in its N-terminal region. It is too early to a deep speculation about the role of this protein but it is likely involved in the formation of a peritrophic gel- or membrane-like structure, a common structure present in the midgut of different arthropods (116), including other arachnids as spiders (19) and harvestmen (Fuzita et al., unpublished). Chitinase 3 was extremely abundant in fed (8.9%) and fasting animals (14.5%) so it may have an important function. Regardless the presence of peritrophic membranes in arachnid have being demonstrated to some species, this is not well studied for this group as for insects (116, 117). In figure 3.4

the enzymes that could be present in the secretory granules and in the lysosome-like vesicles are depicted.

Based on the bioinformatics analysis and in the literature data used for the prediction of the subcellular location of the digestive enzymes, the food processing in the scorpion *Tityus serrulatus* probably happens as described below. The secretory cells after few days of feeding are ready for a new digestive cycle filled with secretory vesicles, which will be released as soon as the prey is captured to compose the digestive juice (28). Up to now, it was not clearly shown that these secretory granules contain the digestive enzymes. In this work, RNA-seq followed by shotgun proteomics allowed the identification of 18 enzymes that could be present in these secreted vesicles (Figure 3.4). This is the first molecular evidence of enzymes present in the MMG at the protein level that could be secreted to form the digestive juice. It is important to note the quantitative values of the fasting animal's proteins because it shows the intracellular stock for the next predation event, differently from the fed ones where the dynamics of digestion can lead to misinterpretation.

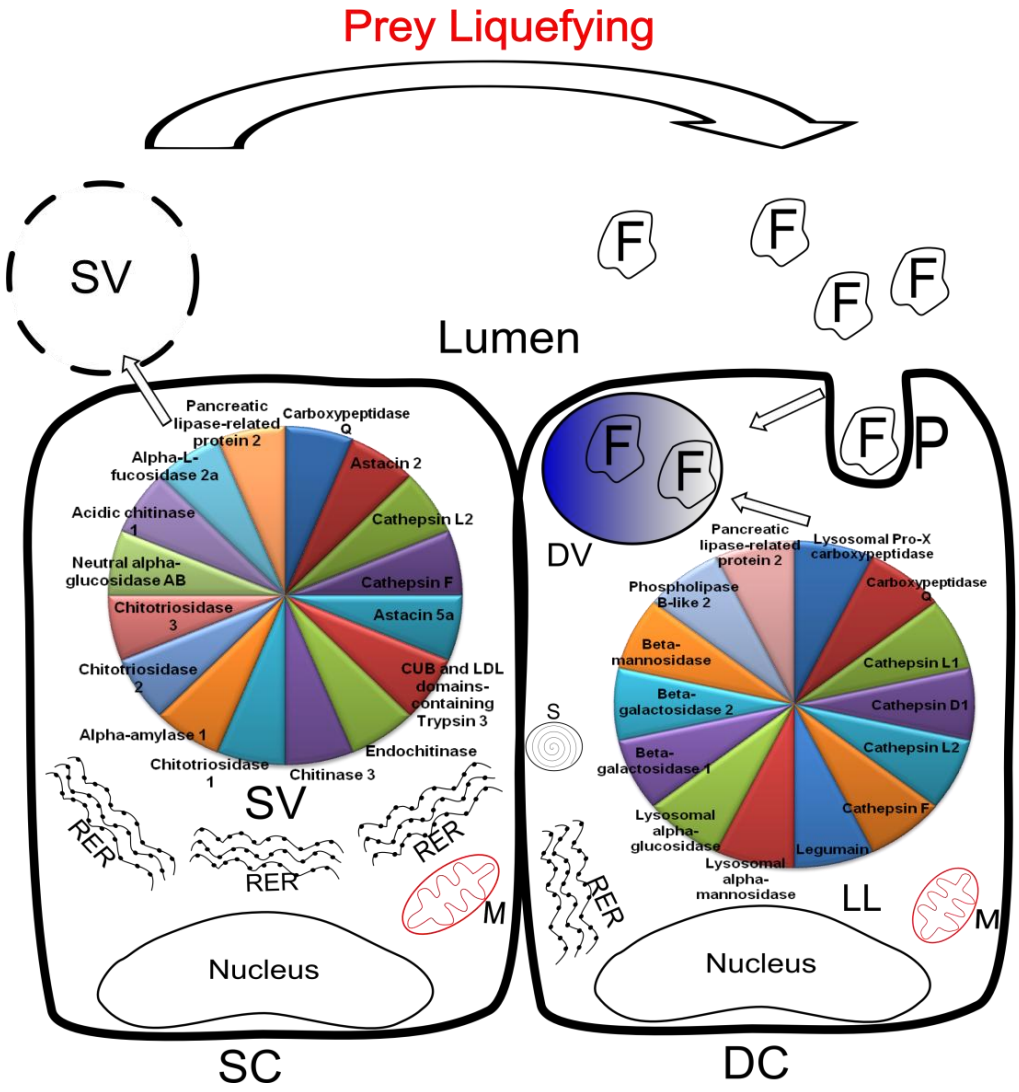
It was possible to observe that in some cases the intracellular enzymes are more representative during feeding and the extracellular enzymes in the fasting conditions, whereas other ones kept unchanged. The carbohydrases acting extracellularly are mainly constituted of endochitinase (18%) and chitotriosidase 1 (10.9%) in fasting scorpions. When the animals are feeding for 9 hours the percentages of the main represented carbohydrases change, with the ones responsible for intracellular digestion being the most abundant now (Figures 3.3C and 3.3D). Lysosomal alpha-mannosidase pass from 6.6% to 12.5% and lysosomal alpha-glucosidase from 1.9% to 12.6%. Endochitinase and chitotriosidase 1 decreased from 18% to 1.7% and 10.9% to 1%, respectively. The extracellular neutral alpha-glucosidase AB is only 0.62% of the spectra counting in fasting animals and it was not identified in the fed ones.

The beginning of the extracellular protein digestion will be performed by endopeptidases such as astacins, CUB and LDL domains-containing trypsin and cathepsins L2 and F. The intracellular phase will be done by cathepsin L1, cathepsin D, legumain and also maybe cathepsin L2 plus cathepsin F (Figure 3.4). The notable changes are of the intracellular cathepsin L1 and the extracellular astacin 2, the former passes from 8.6 to 23.6% and the latter from 3.5 to 1.5% in fasting and fed animals, respectively. Cathepsin D and L2 are practically constant around 5% and

1.8%, respectively. Astacin 5a is present only in fasting animals and cathepsin F changes less than 2 fold from fasting to fed conditions.

One lipase was identified at the protein level. This enzyme was found only in fasting animals and could be secreted (WoLF PSORT and GO term) or lysosomal (WoLF PSORT and literature). Between the 23 lipases identified after the RNA-seq (Table 3.2) the pancreatic lipase-related protein identified by mass spectrometry has the most similar N-terminal with the lipase identified in the midgut glands of *Scorpio*

Figure 3.4 - Schematic representation of the midgut and midgut glands secretory (SC) and digestive cells (DC)



The figure displays the enzymes present in the secretory vesicles (SV) and lysosome-like (LL) organelles. The lysosome-like vesicles probably fuse with pinocytic vesicles to originate the digestive vacuoles. DC: digestive cells, DV: digestive vacuoles, F: pre digested food, M: mitochondria, P: pinocytosis, RER: rough endoplasmic reticulum, S: spherites, SC: secretory cells.

maurus (37). This protein was further localized into the digestive vacuoles of the same animal and not in the secretory vesicles (38). However the optimum pH for this

lipase was 9 (37), which is not in accordance with an acidic intracellular digestion as expected for a digestive vacuole. Thus, the location of the *Tityus serrulatus* lipase still can be extracellular or lysosomal. The phospholipase B-like 2 was predicted to be cytosolic or nuclear according to WoLF PSORT (Table 3.3). Nevertheless, the GO term lysosome was mapped to this sequence and in humans this enzyme is lysosomal (115). After feeding there is a 3 times increase in the percentage of this enzyme, which passes from 0.5 to 1.5% of the total spectra counting. Carboxypeptidase Q and lysosomal Pro-X carboxypeptidase were identified with very low protein levels (Table 3.3) so it is hard to state something about a possible digestive role. The most abundant exopeptidase was Xaa-Pro dipeptidase in both physiological conditions and also dipeptidyl peptidase 3b in fasting animals. These enzymes are probably cytosolic (Table 3.3) and in humans Xaa-Pro dipeptidase is a cytosolic enzyme involved in protein turnover and food final digestion (118). Maybe these molecules even being cytosolic could also be involved in the final dietary protein digestion. Also other exopeptidases could still be translated in a posterior stage of digestion not represented in our data.

In summary, the secretory granules will be discharged in the lumen once the feeding process begins which probably contain the following enzymes responsible for extracellular digestion: chitinases, alpha-amylase, alpha-glucosidase, pancreatic lipase related-protein, alpha-L-fucosidase, cathepsins F and L, trypsins and astacins. The pinocytosis will rapidly start and the internalized vesicles will fuse with each other and lysosome-like vesicles to originate the big digestive vacuoles previously observed by other authors (28, 38). These lysosome-like vesicles probably are filled with phospholipase B, pancreatic lipase-related protein, cathepsins D, F and L, legumain, alpha and beta-mannosidase, alpha-glucosidase, beta-galactosidases, Pro-X carboxypeptidase and carboxypeptidase Q.

3.4.3 Molecules from the vesicular trafficking and inhibitors

The first goal of this study was to identify the enzymes with a possible digestive role. However, due to the nature of the deep analysis performed, proteins from the vesicular pathway and peptidase inhibitors were also identified by mass spectrometry, which leads to an expansion of the model above proposed. Two

different MAM and LDL-receptor class A domain-containing were identified in the MMG of fed animals. The MAM domain is known for protein-protein interactions with many different functions (119) and the LDL domain is also involved in a variety of roles, including receptor-mediated endocytosis (120). This multi domain protein from *Tityus serrulatus* contains 3 LDL-class A and 3 MAM domains and in this context and due to the nature of these domains, we will propose that these receptors are involved in a receptor-mediated endocytosis of proteins and lipids process. Based on this, it is possible that these receptors have as function the internalization of the pre-digested meal, which after 2 hours of feeding can be observed (28). Some proteins as chlatrin, Rab and nexins, are related to the vesicular trafficking after the endocytosis/pinocytosis and also before exocytosis. The light and heavy chains of clathrin, five different Rab proteins (1a, 2, 5c, 11a, 14) and 4 sorting nexins (2, 6, 12 and 17) were identified. The fact that the subunits A and B of V-type proton ATPase were identified reinforces the theory on an acidic phase of digestion, probably intracellular.

Other interesting aspect of the digestive process in predators is the control of prey´s peptidase ingested. One serpin that is likely secreted (WoLF PSORT, data not shown) was identified, possibly acting as an inhibitor of prey´s serine peptidases. A cystatin was also identified however, after both BLAST and WoLF PSORT analyses, it seems that is an intracellular inhibitor. Maybe the function of inhibition of the cysteine peptidases from the prey is done by the U24-ctenitoxins-Pn1a found in the *Tityus serrulatus* MMG. Despite this is a not biochemically characterized toxin, it contains the tyroglobulin domain, which is an inhibitor of cysteine peptidases. The three isoforms of U24-ctenitoxins-Pn1a found are secreted according to the prediction software, which corroborates the hypothesis of prey´s cysteine peptidase inhibition. Ctenitoxins were also found in the digestive juice of the spider *Nephilengys cruentata* (chapter 5). This is the first report of such kind of toxin in the digestive system of a scorpion. Although contamination might always be debatable, these proteins were found expressed (RNA-seq) and at the protein level in both, fed and fasting animals it is unlikely that these molecules are original from the venom.

3.5 Conclusions

A combination of high throughput techniques was applied for the first time to study the molecular physiology of digestion in a scorpion. We showed that the useof these two technologies together was very efficient to obtain a large number of protein sequences of an organism without complete genome. Endo- and exopeptidases, carbohydrases and lipases were transcriptomically and proteomically identified. Based on the sequences identified by shotgun proteomics and using subcellular prediction software bioinformatic tools, a new molecular mechanism of digestion in the scorpion *Tityus serrulatus* was proposed. Firstly, it was confirmed by mass spectrometry the hystological observation that fasting animals already contain the digestive enzymes needed for the next predation event. Some of the proteins involved in extracellular digestion (e.g. chitinases) are more representated is fasting animals whereas the ones involved in intracellular digestion are more abundant in fed animals (e.g. lysosomal alpha-glucosidase). Secondly, the secretory granules are composed of different hydrolases such as chitinases, astacins, alpha-L-fucosidase, pancreatic lipase-related protein, cathepsins L and F, trypsins and also peptidase inhibitors (serpin and U24-ctenitoxin-Pn1a) whereas the lysosome-like organelles contain legumain, alpha-mannosidase, alpha-glucosidase, beta-galactosidase, beta mannosidase, cathepsins D, F and L, Pro-X carboxypeptidase, carboxypeptidase Q, pancreatic lipase-related protein and phospholipase B. Multi-domain MAM and LDL receptors may be involved in the partially digested food uptake to the cells. A chitolectin with a peritrophin domain that possibly is involved in the formation of a peritrophic gel/ membrane was for the first time identified in a scorpion. The availability of these protein sequences opens the doors for future research with recombinant enzymes including the preparation of antibodies for *in situ* location.

CHAPTER 4 – CYSTEINE CATHEPSINS AS DIGESTIVE ENZYMES IN THE SPIDER *NEPHILENGYS CRUENTATA:* BIOCHEMICAL CHARACTERIZATION OF THE NATIVE AND RECOMBINANT FORMS

4.1 Introduction

Cysteine peptidases from C1 family can be found in virtually all living beings and it is likely that a peptidase from this family was present in the ancestor of all organisms (71). Cysteine cathepsins belong to C1A family and are commonly associated with protein degradation in lysosomes. However, recent studies have shown that these enzymes can also play an extracellular role in humans (87). In arthropods the cathepsins L and B play important roles in the digestive process. There are examples of digestive cathepsins L and B in all main Arthropoda taxa such as Crustacea (6, 121), Hexapoda (10) and Arachnida (61).

The presence of these enzymes in arachnids has been shown in ticks (61) and scorpions (Chapter 2) until now. Cathepsins L and B are important peptidases involved in the hemoglobinolytic pathway in *Ixodes ricinus* (61, 66) and the former was also located in digestive vesicles from *Rhipicephalus (Boophilus) microplus* (58). The enzymatic studies on digestive peptidases from spiders are still underexplored. The peptidase activities observed in the digestive juice (DJ) from *Tegenaria atrica* were classified as carboxipeptidase A, aryl aminopeptidase, glycylglycine dipeptidase, chymotrypsin and trypsin (48). Kavanagh and Tillinghast (49) and Foradori (51) provided strong biochemical evidences of the presence of metallo peptidases in the digestive juice from *Argiope aurantia*. Recently, this was confirmed by obtaining the amino-terminal sequence from two astacin-like metallo peptidases in the same kind of sample (52). The only study using the midgut as enzyme source identified a collagenase-like activity in thirteen spider species from Australia (50).

The studies cited above were mostly concerned in the characterization of the digestive juice. Nevertheless, it is known that arachnids, in general, perform extra-oral digestion (EOD) (1) in combination with the intracellular one (5). Thus it is also important the characterization of digestion in the midgut. The predigested food is absorbed by the digestive cells in the midgut and midgut glands (MMG) through pinocytosis and the final digestion takes place inside the digestive vacuoles (21).

Moreover, none study attempted to do enzymatic assays using the combination of substrates, assay conditions and specific inhibitors to investigate the presence of cysteine cathepsins. Thus, in the present work, enzymatic assays were done using properly conditions to detect the activity of cysteine cathepsins using as enzyme sources, not only the digestive juice but also, the midgut with its digestive glands. For the first time it was identified the presence of cathepsins L and B as digestive enzymes in the DJ and MMG from a spider.

4.2 Materials and Methods

4.2.1 Animals and sample obtaining

Adult *Nephilengys cruentata* females were collected in São Paulo city (Brazil), kept under natural photoregime and room temperature conditions with water spraying 4 times per week in their artificial environment. The animals were starved for at least one week and then fed with *Gryllus sp*. Fed spiders were dissected after 9 hours eating. The starved spiders were dissected two weeks after start feeding. The animals were immobilized in a CO_2 chamber and dissected in cold isotonic saline solution (300 mM KCl pH 7) and the opisthosomal midgut with its glands (MMG) (Figures 1.3A and 1.3B) were isolated. MMGs were stored at -20 ºC until use. Samples were homogeneized in a Potter-Elvehjem homogenizer in cold deionised water to a final volume of 1 ml per MMG. Digestive juice samples were collected by electrical or mechanical stimulus in two weeks fasting animals or after 1, 3, 9, 25 and 48 hours of feeding. Samples prepared for purification attempts contain 1 mM of methylmethane thiosulfonate (MMTS), a reversible cysteine peptidase inhibitor (68).

4.2.2 Protein determination and enzymatic assays

The protein concentration was determined according to the method of Smith et al. (69) using ovoalbumin as standard. The 4-methylcoumarin-7-amide (MCA) fluorescent substrates were purchased from Sigma. Stock solutions (1 mM) of Z-FR-MCA and Z-RR-MCA were prepared in dimethyl sulfoxide (DMSO) and diluted to a 10 µM final concentration in 0.1 M adequate assay buffer as listed below. Buffers used: pH ranges 2.6-7.0 citrate-phosphate (CP), 8.0-9.0 TRIS-HCl and 10.0 Gly-

NaOH. All buffers contained 3 mM cysteine and 3 mM EDTA to avoid oxidation of the catalytic cysteine from the enzymes (70, 71). All assays were performed such that the measured activity was proportional to the protein concentration and the incubation time. No-enzyme and no-substrate controls were included.

4.2.3 Partial isolation of cysteine peptidases and inhibition assays

MMG homogenate samples from *Nephilengys cruentata* were fractionated in 1.8 M ammonium sulfate for at least 16 hours at 4 °C. The samples were centrifuged for 20 minutes at 16,100 × *g* and 4 °C. The supernatant was applied to a hydrophobic column (Hitrap Butyl FF-GE) coupled to an ÄKTA-FPLC system (GE) that had been equilibrated in 50 mM phosphate buffer (pH 6) containing 1.7 M ammonium sulfate. The elution was performed using a 25 mL gradient of 1.7 – 0 M ammonium sulfate in 50 mM phosphate buffer (pH 6); fractions of 1 mL were collected. The fractions that exhibited activity with Z-FR-MCA as a substrate were pooled, desalted (HiTrap desalting column, GE) and concentrated using a Vivaspin 6 membrane (GE). The samples were then applied to a cation-exchange column (Resource S-GE) that had been equilibrated in 50 mM sodium acetate buffer (pH 5.0). The protein was eluted using a 40 mL gradient of 0 – 0.6 M NaCl in the equilibrating buffer, and fractions of 0.5 mL were collected and assayed using Z-FR-MCA. The fractions from the first chromatographic step were assayed in the presence and absence of the cathepsin B inhibitor CA-074, using the substrates Z-FR-MCA or Z-RR-MCA.

Alternatively, a simpler partial purification was used to the inhibitory assays. The homogenized samples were diluted in 50 mM citrate-phosphate buffer (pH 5.0). This diluted sample was applied to a cation-exchange column (HiTrap S, GE) that had been equilibrated in the same buffer. The protein was eluted in a 25 mL gradient of 0–1.0 M NaCl, in the equilibrating buffer. The fractions were assayed using 10 µM Z-FR-MCA in the presence and absence of the following peptidase inhibitors: 10 µM E-64 (cysteine peptidase), 10 µM pepstatin (aspartic peptidase) and 1 mM PMSF (serine peptidase).

In all inhibitory assays the incubation without the substrate was done for 30 minutes at 30 ºC and after this period the substrates were added to a final concentration of 10 µM.

4.2.4 Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

Samples of the cell lysates used for cathepsin L1 expression or from the purification attempt were diluted in a sample buffer containing 60 mM Tris-HCl buffer (pH 6.8), 2.5% SDS, 0.36 mM β-mercaptoethanol, 10% (v/v) glycerol and 0.005% (w/v) bromophenol blue. The samples were heated for 5 minutes at 95 °C in a water bath and then loaded onto a 12% (w/v) polyacrylamide gel slab containing 0.1% SDS (72). The gels were run at a constant voltage of 200 V at room temperature and then stained with Coomassie Blue (Colloidal Blue Staining Kit, Invitrogen).

4.2.5 Acidic activation of cysteine peptidases

Crude and recombinant samples were 10 times diluted and incubated in 0.1 M citrate-phosphate buffer containing 3 mM cysteine and 3 mM EDTA in a range of pH values from 2.6 to 7.0 for 60 min or 10 min at 30 °C, respectively. The incubated samples were then 5 times dilute in deionised water and activity was measured using a 1:100 ratio of activated enzymes to 10 µM Z-FR-MCA diluted in 0.1 M citrate-phosphate buffer (pH 5.0), this ratio guaranteed that the pH of the assay was always the same. Two controls were done: a) sample was 50 times diluted in deionised water and incubated at 30 ºC for 1 hour or b) the dilution was done prior to the activity assays. The pH that resulted in the highest rate of hydrolysis was selected for an incubation time course to verify the length of time that was required for acidic activation *in vitro*. After this incubation, enzymatic assays using 10 µM Z-FR-MCA were performed as described above. For standard activation, the crude samples were incubated at 30 °C in citrate phosphate buffer pH 2.6 for 120 min whereas recombinant samples were incubated for 60 min in pH 3 at the same temperature. Both buffers contain 3 mM cysteine and 3 mM EDTA.

4.2.6 pH stability

The stability of the cysteine cathepsins under different pH conditions was evaluated by incubating the standard activated enzyme samples from the MMG

homogenates or the recombinant one in a pH range from 5.0-10.0 at 30 °C for 3 h. The samples were then diluted in deionised water to guarantee adequate pH for residual activity measurement in 0.1 M citrate-phosphate buffer pH 5.0. The stability of the recombinant cathepsin zymogen was tested as above described with the difference that the enzyme was activated after the incubation. In all stability experiments the buffers concentration and enzyme dilution were adjusted to guarantee no pH changes in the activation, incubation and assay steps. The percentage of activity is relative to controls that were prepared prior to the assay. Buffers used for incubation: citrate-phosphate pHs 5.0-7.0; TRIS-HCl pHs 8.0-9.0. All assay and incubation buffers contained 3 mM cysteine and 3 mM EDTA.

4.2.7 The effect of pH on enzyme activity

MMG homogenate or recombinant samples (both activated) were assayed with 10 µM Z-FR-MCA in pHs ranging from 2.6-9.0 in the same conditions and buffers as above described in item 4.2.2.

4.2.8 The effect of substrate concentration

The effect of substrate concentration on the activity of the partially purified cysteine peptidases or the recombinant catLN1 was studied using at least 15 different Z-FR-MCA concentrations of ranging from 1-150 µM. The K_m values (mean ± SEM) were determined from a weighted linear regression using EnzFitter software (Biosoft).

4.2.9 Mass spectrometry procedures

For the detailed methodology see chapter 3 item 3.2.4.

4.2.10 Molecular cloning of digestive cysteine cathepsins from MMG

All nucleic acid manipulations were performed as previously described (122) or according to the manufacturer's protocol, unless otherwise specified. The RNA extraction from the MMG was performed using the TRIzol® (Invitrogen) reagent.

Complementary DNA (cDNA) was obtained using the Superscript III First-Strand Synthesis System for RT-PCR® kit (Invitrogen). The RACE technique - rapid amplification of cDNA ends, (123) - was applied as follows: in the first step to obtain the 3´region the cDNA was submitted to a polymerase chain reaction (PCR) using a degenerate primer A (forward) and primer B (reverse). Primer A, with the sequence 5'-GAMAVTGYGGWTCBTGYTGG-3' (where M = A or C; Y = C or T; W = A or T; and B = C, G or T), was designed based on the GQCGSCW sequence that is present in the reactive site of cathepsin L-like cysteine peptidases from different organisms. Primer B is a hybrid primer (Q_T) previously described (123). The PCR product was sequenced as described below and one gene-specific reverse primer C for catLN1 and D for catLN2 were designed (primer C=5´-GGTCGTTTAAACCAGAGGGTA-3´; primer D=5´-CACATTTTAAACTAATGGGTA-3´). In a second step, to obtain the 5`region and also the complete sequence from the mRNA, the cDNA was adenine-tailed using a terminal deoxynucleotidyl transferase (Fermentas). A new PCR was done but now primer B was used as a forward primer and primers C or D were separately used as reverse primers.

For the sequencing of the cloned DNA sequences the PCR products were loaded onto a 1% agarose gel. The band was extracted using a GeneJet Gel Extraction Kit® (Fermentas) and inserted into the pGEM-T Easy vector® (Promega). Thermally competent XL1-Blue cells were transformed and selected by ampicillin resistance. Single colonies were selected and grown in LB Amp (Luria-Bertani broth containing ampicillin) at 37 °C overnight, and the plasmids were isolated (GeneJET Plasmid Miniprep Kit®, Fermentas) and sequenced using a Big Dye Terminator Mix® (Applied Biosystems) in an ABI PRISM 3100 Genetic Analyzer (Applied Biosystems). The sequences obtained were submitted to a Basic Local Alignment Search Tool (BLAST) analysis using default parameters (available at www.ncbi.nih.gov/blast). Then, a multiple sequence alignment was constructed using ClustalW software (124).

Other sequences used to the mass spectrometry identification were obtained in a transcriptomics study that it will be further discussed in chapter 5.

4.2.11 Construction of the expression vectors

After removing the signal peptide using the online tool Signal P 4.0 (125) the two clones of cathepsin L-like cysteine peptidases obtained (catLN1 and catLN2) as

described above were separately inserted in a pAE expression vector (126). New primers were designed and catLN1 forward primer E contained a restriction site to the enzyme KPNI whereas reverse primer F presented a restriction site to HindIII. The forward primer catLN2 G and the reverse primer H contained restriction sites to the enzymes BamHI and BstBI, respectively. After the PCR amplification of the two sequences the products were purified by loading the samples in a 1% agarose gel and extracting the DNA from the single band observed to each sequence. The PCR products (overnight digestion) and pAE plasmid (3 hours digestion) were double digested with the respective restriction enzymes at 37 ºC. The plasmids pAE-catLN1 and pAE-catLN2 were obtained after inserting the digested PCR products in their respective linearized plasmids using the enzyme T4 DNA ligase. All enzymes used were from Fermentas. Primers: E= 5´-GGACAACACAAACTAAAGACC-3´; F=5´-GGTCGTTTAAACCAGAGGGTA-3´; G=5´-GAAGGTCAGCACGCAAAGAAG-3´; H=5´GATTTGAACATTGGAAAGAGG-3´.

4.2.12 Heterologous expression and purification of catLN1 and catLN2

Escherichia coli competent cells from the strain BL21 Star (DE3)pLysS were separately transformed using the constructions pAE-catLN1, pAE-catLN2 and only pAE as a control. The transfected cells were grown overnight at 37 ºC in LB medium containing 100 µg/ml of ampicillin and 34 µg/ml of chloramphenicol (LB_{ac}). The cultures were then diluted 20 times in LB_{ac} to a final volume of 50 ml with 1% glucose and grown until a 0.6 value of optical density at 600 nm be reached. Thereafter, the cells were collected by centrifugation at 5,000 x g and the medium was changed to an LB_{ac} without glucose. For the expression induction, isopropyl-ß-D-thiogalactoside (IPTG) was added to a final concentration of 1 mM and the culture was incubated for 15 hours at 20 ºC. The cells were centrifuged at 5,000 x g for 30 minutes at 4 ºC and the pellet suspended in 1 ml of lysis buffer [10 mM TRIS-HCl, pH 8.0; 100 mM NaCl; 20 mM imidazol; 1 mM phenylmethylsulfonyl fluoride (PMSF) and 10% glycerol (v/v)]. Cell lyses was achieved by sonication (100W potency) in the R1 cell disruptor (Unique) for 5 cycles of 1 min each with 3 min chilling in between. The lysate was centrifuged at 16,100 x g for 30 min at 4 ºC and the supernatant (soluble fraction) applied onto a Ni-NTA agarose column (Qiagen) previously equilibrated in the lysis

buffer. The procedure was followed according to the manufacturer´s instructions except that the buffers contained 10 mM TRIS instead of 50 mM monobasic phosphate. After the washing steps, the elution of the recombinant proteins was accomplished using elution buffer with 150 mM imidazol.

4.3. Results

4.3.1 Identification of the cysteine cathepsins by mass spectrometry and molecular biology techniques

The supernatant 1 and the digestive juice were submitted to a proteomics experiment as described in the methodology. Table 4.1 shows the cysteine cathepsins identified as proteins or mRNA and in which physiological condition.

Table 4.1- Cysteine cathepsins identified in the MMG and digestive juice from the spider *Nephilengys cruentata* under different physiological conditions

	MMG				Signal peptidade cleavage site
Enzyme	Protein		mRNA		
	Fed (9 hours)	Fasting	Fed (9 hours)	Fasting	
Cathepsin L 1	Yes	Yes	Yes	Yes	21-22
Cathepsin L 2	Yes	Yes	Yes	Yes	15-16
Cathepsin L 4	No	Yes	Yes	Yes	19-20
Cathepsin L 8	Yes	Yes	Yes	No	14-15
Cathepsin B 1	Yes	Yes	Yes	No	*

	Digestive Juice						
	Fasting	1 hour	3 hours	9 hours	25 hours	30 hours	48 hours
Cathepsin L 2	Yes	No	No	Yes	Yes	Yes	Yes
Cathepsin B 1	Yes	No	No	No	Yes	Yes	Yes

Note: * The sequence was incomplete at the N-terminal region

Cathepsins L 1 and 2 were identified in the MMG at both, protein and mRNA levels, in fasting or fed animals whereas the later could also be identified in the digestive juice under fasting conditions and after 9, 25, 30 and 48 hours of feeding. The mRNA of cathepsin L 4 was observed in the MMG of fasting and fed animals but the protein could only be observed in fasting animals. The presence of cathepsin L8 and cathepsin B as proteins were confirmed by LC-MS/MS experiments and their mRNA could be detected only in fed animals. Cathepsin B was also identified in the digestive juice but only after 25 hours of feeding and in fasting spiders.

All cathepsin L sequences identified in table 4.1 possess the four residues of the active site (Gln 152, Cys 158, His 292 and Asn 308, papain numbering), the signal peptide and the propeptide. The cathepsin B sequence is not complete and lacks the propeptide and the histidine residue of the active site but it contains the 3 other residues and it is complete in the C-terminal position. None peptide from this part of the sequence was identified thus with the present data it is not possible to confirm if the lack of the catalytic His is true or just an artifact of the contig assembly. Other cysteine cathepsins (5 cathepsins L, 1 cathepsin B and 1 cathepsin O) were identified only in the transcriptomics experiments and will be further discussed in chapter 5:

4.3.2 Characterization of the cysteine peptidases present in the MMG of the spider Nephilengys cruentata

Activity in the MMG samples was measured using the substrates Z-FR-MCA or Z-RR-MCA in citrate phosphate buffer pH 5.0 containing 3 mM cysteine and 3 mM EDTA of activated or no-activated samples (Table 4.2).

Table 4.2 – Absolute and specific activities in the MMG of the spider *Nephilengys cruentata*

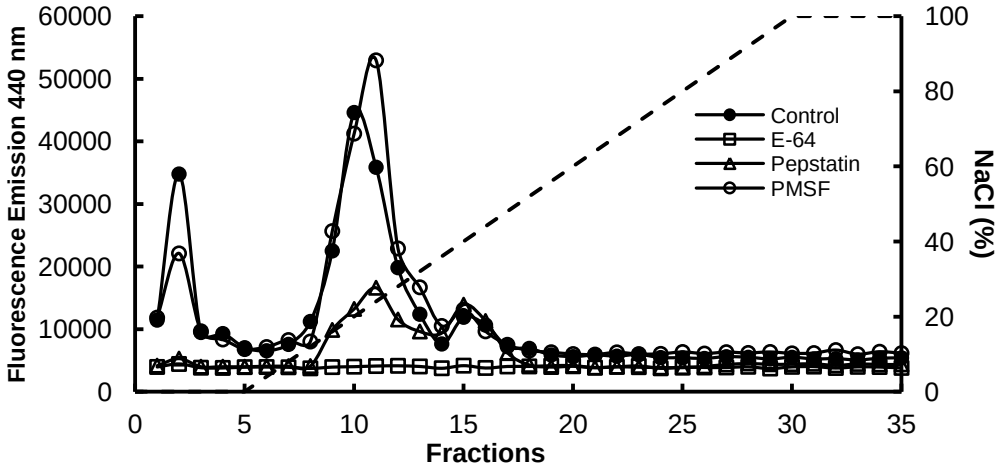
MMG	Non-Activated		Activated	
Fed	**Absolute (mU/MMG)**	**Specific (mU/mg)**	**Absolute (mU/MMG)**	**Specific (mU/mg)**
Z-FR-MCA	213 ± 22	3 ± 0.5	3,695 ± 896	57 ± 8.6
Z-RR-MCA	-	-	258 ± 72	3.5 ± 0.9
Fasting				
Z-FR-MCA	*	*	2,383 ± 353	47 ± 3
Z-RR-MCA	-	-	60 ± 0.5	1.1 ± 0.1

Notes: -activity was not observed.
*the measurements were not used in this table due to its discrepancy.
Non-activated and activated samples hydrolyze ratios were measured with 10 µM of the above substrates diluted in citrate-phosphate buffer pH 5. Activated samples were diluted 10 times in citrate phosphate buffer pH 2.6 and incubated for 2 hours at 30 ºC. After a 5 times dilution in deionised water the activity was measured. All buffers contain 3 mM cysteine and 3 mM EDTA. The values are mean and SEM of at least 3 different biological samples.

Hydrolysis ratios in activated crude MMG extracts from fed or fasting animals were higher over Z-FR-MCA than to Z-RR-MCA, (Fed animals: 3,695 ± 896 mU/MMG and 258 ± 72 mU/MMG, fasting animals: 2,383 ± 353 and 30 ± 0.5 mU/MMG, respectively). Without activation the fed animals' samples activity over Z-FR-MCA was 213 ± 22 mU/MMG and it was not possible to verify Z-RR-MCA

hydrolysis. To confirm that the Z-FR-MCA cleavage was due to the action of cysteine peptidases the samples were submitted to a cation-exchange chromatography and assayed in the presence of different peptidase inhibitors (Figure 4.1). As expected the cysteine peptidase inhibitor (E-64) was the only one to cause 100% inhibition. The serine peptidase inhibitor PMSF did not affect the activity and since EDTA is used in the medium assay metallopeptidase activity over Z-FR-MCA also can be discarded. Curiously pepstatin, an aspartic peptidase inhibitor, caused 47% inhibition. The fact that 1) Z-FR-MCA is not suitable for aspartic peptidase cleavage (78) and that 2) this same feature was observed and confirmed for the cysteine peptidases of *Tityus serrulatus* (Figure 2.8) lead us to believe that this also is a non-specific inhibition.

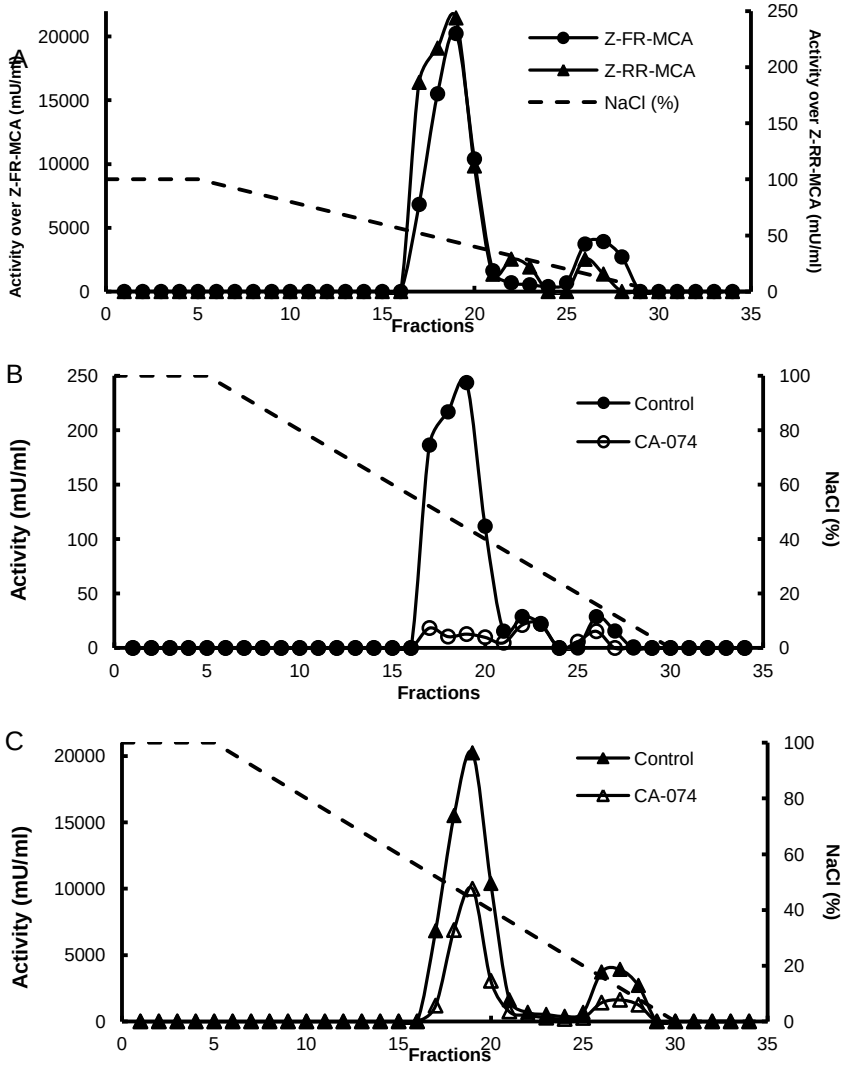
Figure 4.1 - Cation-exchange chromatography of *Nephilengys cruentata* MMG samples on a Hytrap S column assayed in the absence and presence of peptidase inhibitors



The column was equilibrated in 100 mM sodium acetate buffer pH 5.0. The elution was performed using a gradient of 1-0 M NaCl in the same buffer. The activity was measured using 10 µM Z-FR-MCA in citrate-phosphate buffer pH 5.0 containing 3 mM cysteine and EDTA. Control (•); E-64 (□), Pepstatin (△); PMSF (○).

In order to differentiate cathepsin B from cathepsin L activity a hydrophobic interaction chromatography (HIC) was done after an ammonium sulfate fractionation (Figure 4.2). Activity is about 100 times higher with the substrate Z-FR-MCA in contrast to Z-RR-MCA (Figure 4.2A). The cathepsin B inhibitor CA-074 was tested in the active fractions using both substrates in different assays. In figure 4.2B it is possible to see that Z-RR-MCA hydrolysis is 88% inhibited in the main activity peak and only 52% in the other peaks. When Z-FR-MCA is used both peaks are 59% inhibited (Figure 4.2C).

Figure 4.2 - Inhibitory assays with CA-074 using the substrates Z-FR-MCA and Z-RR-MCA after a hydrophobic interaction chromatography

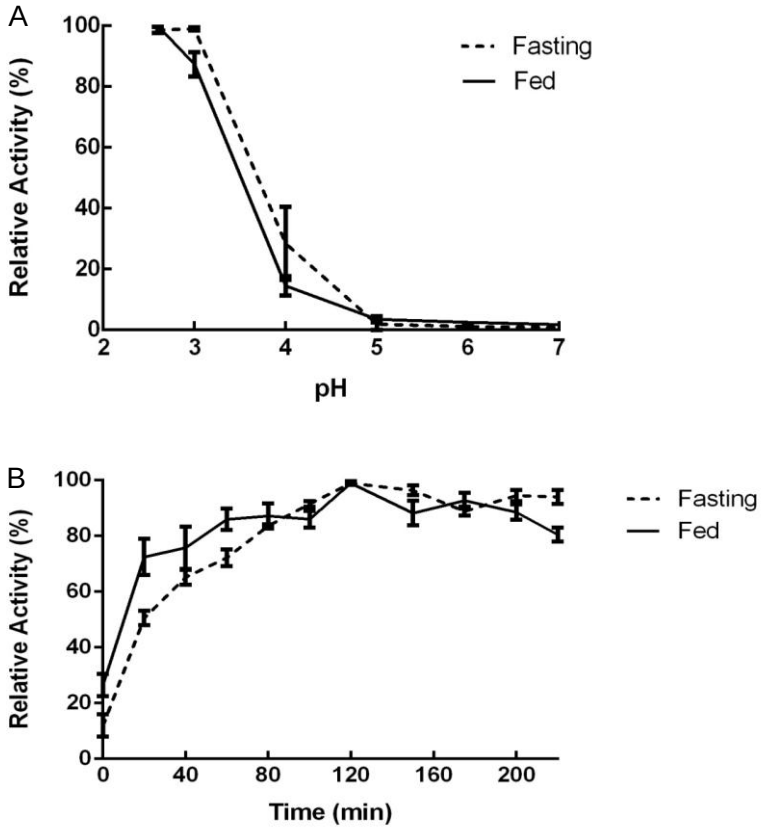


A) Hydrophobic chromatographic fractionation of *Nephilengys cruentata* was assayed with 10 µM Z-FR-MCA (▲) and 10 µM Z-RR-MCA (•) in 0.1 M citrate phosphate buffer pH 5.0. (B) Activity assay with 10 µM Z-RR-MCA in the presence (○) and absence (•) of 10 µM CA-074. (C) Activity assay with 10 µM Z-FR-MCA in the presence (△) and absence (▲) of 10 µM CA-074. All assay buffers contain 3 mM cysteine and 3 mM EDTA.

A further investigation was performed submitting the fractions 17-20 from the first peak (P1) and 26-28 from the last one (P2) to LC-MS/MS experiments. In P1 it was possible to identify cathepsins L1, L2 and B1 with percentage coverage of 18%, 18% and 14%, respectively. In P2, cathepsins L1 and L2 were detected with coverage of 11 and 29%, respectively. The activity peaks P1 and P2 were separated applied onto a cation-exchange chromatography. However, differently from the results obtained with digestive cathepsins from *Tityus serrulatus* (chapter 2), it was not possible to successfully purify the cathepsins from *Nephilengys cruentata* (data not shown).

Once it was confirmed that Z-FR-MCA hydrolysis in the assay conditions was due to the action of cysteine peptidases and the activity was higher over this substrate (Table 4.1) some properties of the enzymes in crude samples were studied using Z-FR-MCA. It is known, in the literature, that cysteine peptidases are synthesized as inactive zymogens that can be activated when incubated in acidic pHs (86, 127, 128). MMG samples from fasting or fed animals were first activated in a pH range from 2.6-7 as described in the methodology. Higher activities were obtained after incubation in pHs 2.6 and 3.0 (Figure 4.3A) showing that these pHs are more suitable to a time-course activation experiment resulting in an increase of activity of 17.3 times (Table 4.2). After choosing the pH 2.6 for activation, samples were submitted to time course experiments which evidenced that maximum activity is reached after 2 hours of incubation in pH 2.6 in samples from fasting and fed animals (Figure 4.3B).

Figure 4.3 - Acidic activation of cysteine cathepsins from *Nephilengys cruentata* MMG in fasting and fed conditions

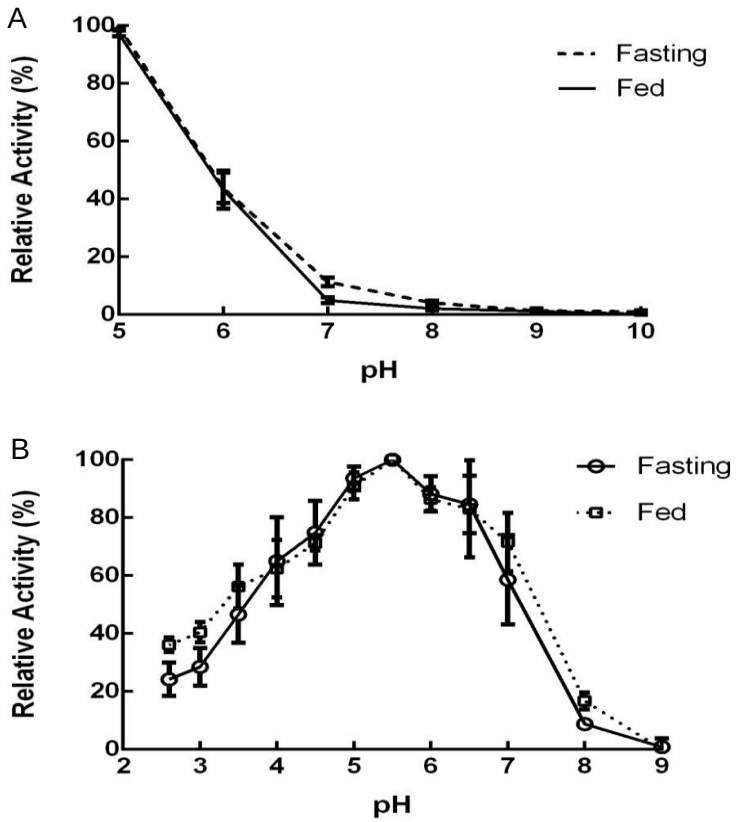


(A) Sample incubation in different pHs for 60 minutes at 30 ºC. (B) The effect of time on the acidic activation of cysteine cathepsins incubated in 0.1 M citrate phosphate buffer pH 2.6. Continous and dashed lines respectively represent fed and fasting spiders. The activities percentage is relative to the highest observed hydrolysis ratio. In both experiments the activity was measured using 10 µM Z-FR-MCA diluted in citrate-phosphate buffer pH 5.0. All buffers used contain 3.0 mM cysteine and 3.0 mM EDTA.

Activated samples were also submitted to pH stability experiments in a pH range from 5.0-10.0. Activated samples kept approximately 100% activity when incubated for 3 hours at 30 ºC in pH 5.0. However remaining activity was only 43% in pH 6. Remaining activities in incubation in pH above 7 were equivalent or lower 11% (Figure 4.4A).

The effect of the pH on crude samples from the MMG of animals in both conditions showed pretty similar profiles (Figure 4.4B). Maximum activity was obtained in pH 5.5 and in pH 5 the activity is higher than 90%. Despite high activities can be observed in pH 7 (between 60 and 70%), the enzymes are stable only up to 30 minutes assay, completely lacking their activities after that (data not shown). In pH 8 the residual activity between 9-16% that can be observed is no longer stable after 15 minutes assay (not shown).

Figure 4.4 - The effect of pH in the activity of *Nephilengys cruentata* cysteine cathepsins from the MMG



(A) The effect of pH in activated samples from fasting and fed animals after 3 hours incubation at 30 ºC in pHs 5.0-10.0. The activity was then measured using 10 µM Z-FR-MCA diluted in 0.1 M citrate phosphate buffer pH 5.0. (B) The effect of pH in the activity of activated samples from fasting (○) and fed animals (□).The activities percentage is relative to the highest observed hydrolysis ratio. The assays were performed in a pH range from 2.6-9.0. Buffers used (all contained 3 mM cysteine and 3 mM EDTA): pHs 2.6-7.0, 0.1 M citrate phosphate; pHs 8.0-9.0, 0.1 M TRIS-HCl; pH 10.0, 0.1 M Gly-NaOH.

Table 4.3 exhibits the activity measured in the digestive juice using Z-FR-MCA. In the digestive juice samples from 9 hours fed animals A1, A2 and A4 an increase in the activity was observed after acidic incubation, in which sample A2 did not have activity without activation. In contrast to that sample A3 had a high activity of 86 mU/ without activation and when incubated in acidic pHs the activity reduce to 6.1 mU/ml. Samples B1 and B2 from starved animals were only able to hydrolyze Z-FR-MCA after activation and sample B5 had a 2.6 increase in activity after this procedure. Sample B3 was not affected by acidic incubation and sample B4 kept only 3% of the activity after activation (Table 4.3).

Table 4.3 - Digestive juice activities of fed and fasting animals over Z-FR-MCA

	Fed				Fasting				
Activity (mU/ml)	**A1**	**A2**	**A3**	**A4**	**B1**	**B2**	**B3**	**B4**	**B5**
Non-activated	1.8	0	86	1.7	0	0	3.2	308	2
Activated*	4.4	3.8	6.1	9	1.6	2.8	3.1	9	5.2
Total change **	2.4	-	0.07	5.3	-	-	0.96	0.03	2.6

Notes: Activity was measured with 10 µM Z-FR-MCA in 0.1 M pH 5.5 citrate phosphate buffer containing 3 mM cysteine and 3 mM EDTA.
*Samples were activated for one hour at 30 ºC after a 10 times dilution in 0.1 M pH 5.5 citrate phosphate buffer containing 3 mM cysteine and 3 mM EDTA.
**The total change is the activity of activated sample/non-activated sample activity

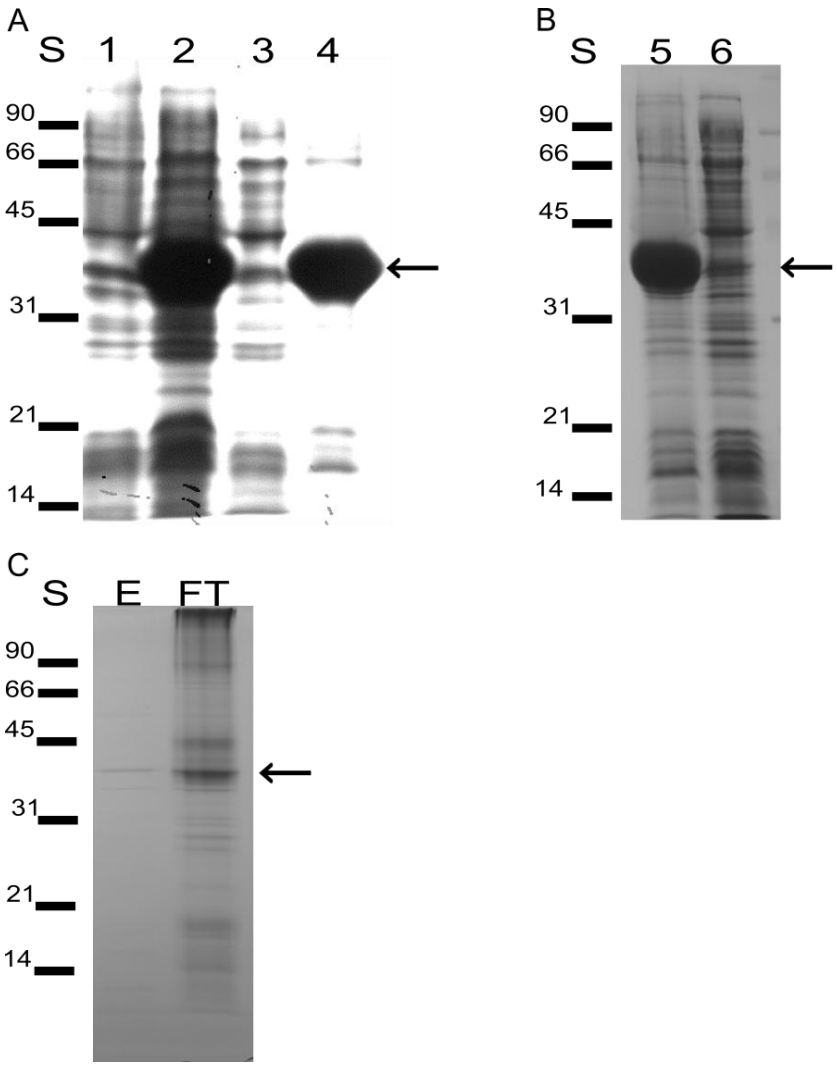
4.3.3 Heterologous expression of cathepsins L1 and L2 from Nephilengys cruentata MMG

The expression of the recombinant cathepsin L1 and cathepsin L2 was well succeeded (Figure 4.5A and 4.5B). All expressions were performed at 37, 30, 25 and 20 ºC (data not shown). For cathepsin L1 most part of the protein was aggregated in inclusion bodies, however in the expression made at 20 ºC activity could be recovered from the soluble fraction. Despite cathepsin L2 could be obtained as a recombinant enzyme even at 20 ºC none activity could be observed in the soluble fraction as well as in the control containing the vector without the insert. The purification of cathepsin L1 to apparent homogeneity was accomplished using an affinity column since the recombinant protein contains a His_6 tag in the N-terminal (Figure 4.5C).

Purified samples from recombinant cathepsin L1 as described in the methodology were used for activation, pH stability and pH effect tests. As for the

MMG samples activation the recombinant enzyme was first incubated in a pH range from 2.6-7. Without activation the samples did not present activity. Incubation in pHs 2.6 and 3 were the most efficient in obtaining activation, however the incubation in pHs 4, 5, 6 and 7 activated the samples 18, 25, 11 and 5% from the maximum activation, respectively (Figure 4.6A). For the time-course activation the pH 3 was chosen, with the samples being incubated in a total of 90 minutes at 30 ºC, with maximal activation after 60 minutes incubation (Figure 4.6B). Activation of 60 minutes at 30 ºC in pH 3 was established as the standard activation procedure.

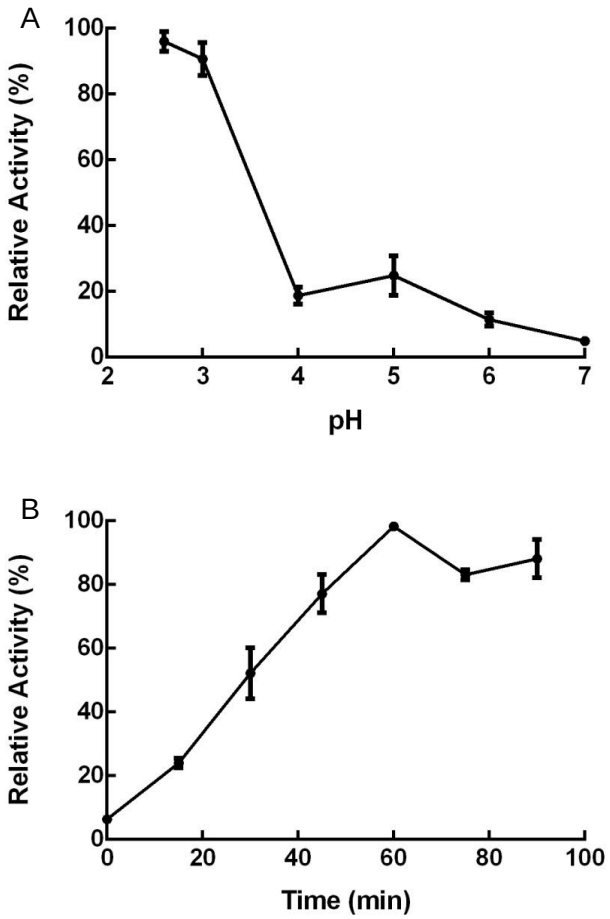
Figure 4.5 - Heterologous expression of cathepsins L1 and 2



(A) Protein profile of expression at 20 ºC using BL21 Star(DE3)pLysS cell transformed with pAE-catLN1. (B) Protein profile of expression at 20 ºC using BL21 Star(DE3)pLysS cell transformed with pAE-catLN2. (C) Purification of recombinant catLN1 using affinity chromatography. S, standard (kDa); lane 1, non-induced cells; lane 2, 1 mM IPTG induced cells; lanes 3 and 6, supernatant from IPTG induced cells after lysis; lanes 4 and 5, pellet from IPTG induced cells after lysis; E, eluted fraction after purification; FT, flow-through after purification procedure. The arrow indicates the recombinant proteins location.

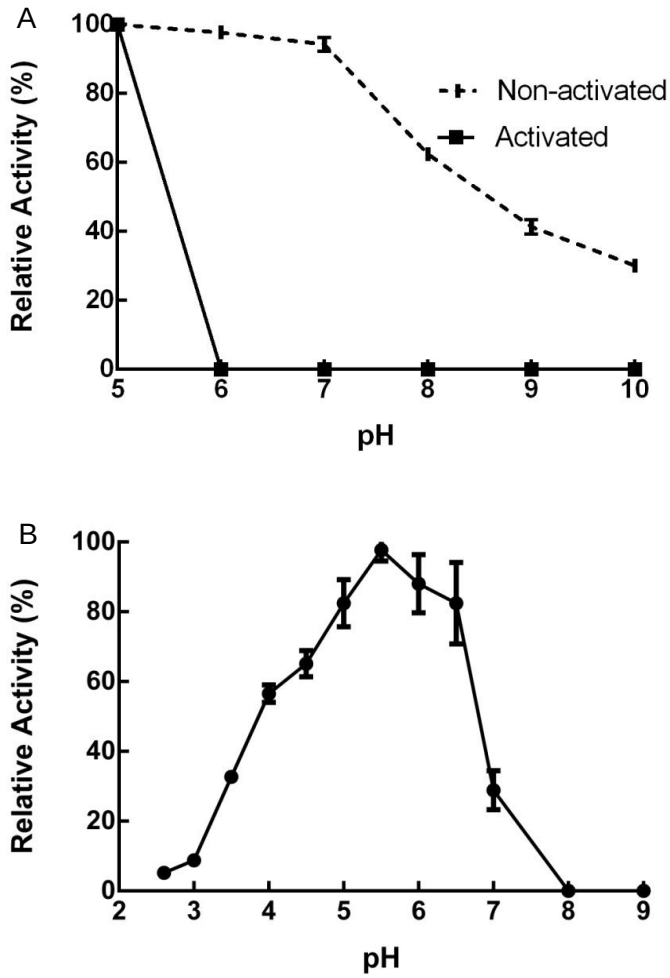
The pH stability was tested with activated and non-activated samples in a pH range from 5.0-10.0. The activated enzyme is stable for 3 hours at 30 ºC only in pH 5, completely losing its activity in pHs 6.0-10.0 (Figure 4.7A). In the zymogen pH stability experiments the incubation step was followed by activation at pH 3. In contrast to the activated enzyme, zymogen remaining activity is near 100% in pHs 5.0, 6.0 and 7.0. Remaining activity in pHs 8.0, 9.0 and 10.0 are, respectively, 62, 42 and 30% (Figure 4.7A). The effect of pH onthe activated recombinant cathepsin L1 was tested in a pH range from 2.6-9.0. The maximum hydrolysis ratio was obtained in pH 5.5 (Figure 4.7B). Kinetic parameters such as K_m (13.5 ± 1.2 µM), V_{max} (672 U/ml ± 20) and K_{cat} (1.2 s^{-1}) were also determined for the recombinant cathepsin L1 using Z-FR-MCA. No hydrolysis over Z-RR-MCA was observed.

Figure 4.6 - Acidic activation of recombinant catLN1



(A) Sample incubation in different pHs for 10 minutes at 30 ºC. (B) The effect of time on the acidic activation of catLN1 incubated in 0.1 M citrate phosphate buffer pH 3.0. In both experiments the activity was measured after the incubation using 10 µM Z-FR-MCA diluted in citrate phosphate buffer pH 5.0. The activities percentage is relative to the highest observed hydrolysis ratio. All buffers used contained 3.0 mM cysteine and 3.0 mM EDTA.

Figure 4.7 - The effect of pH in the activity of the recombinant catLN1



(A) The effect of pH in activated (continuous line) and non-activated* (dashed line) recombinant catLN1 samples after 3 hours incubation at 30 ºC in pHs 5-10. The activity was then measured using 10 µM Z-FR-MCA diluted in 0.1 M citrate phosphate buffer pH 5.0. * In a first step the zymogen was incubated without activation and subsequently was activated for the stability measurements. (B) The effect of pH in the activity of recombinant catLN1 activated samples. The assays were performed in a pH range from 2.6-9.0. Buffers used (all contained 3.0 mM cysteine and 3.0 mM EDTA): pHs 2.6-7.0, 0.1 M citrate phosphate; pHs 8.0-9.0, 0.1 M TRIS-HCl; pH 10.0, 0.1 M Gly-NaOH.

4.4. Discussion

4.4.1 The properties of the native and recombinant cysteine cathepsins from Nephilengys cruentata MMG and its relation with the physiology of digestion

Using a combination of molecular biology, enzymology and mass spectrometry approaches this work reports, for the first time, that cysteine cathepsins are present in the midgut and midgut glands (MMG) and digestive juice of a spider (Tables 4.2 and 4.3). The use of specific inhibitors and mass spectrometry analysis in fractions after hydrophobic chromatography showed that the cleavage of Z-FR-MCA

and Z-RR-MCA was due to the action of cysteine cathepsins (Figures 4.1 and 4.2). Thus, due to the higher activities (Table 4.2) the substrate Z-FR-MCA was chosen to further studies of the properties of these enzymes. The characterization of cysteine peptidases from *Nephilengys cruentata* digestive system will allow correlation of the presence of this protein and its role in the physiology of digestion in this spider. Arachnids, in general, perform an elegant combination of extra-oral with intracellular digestion (5). In spiders, this phenomenon can be observed (21). Nevertheless, the focus of the sporadic studies in proteolytic enzymes in spiders was only related to the alkaline hydrolases present in the digestive juice (48, 49, 51, 52). The only study until now, in which the entire MMG was used as enzyme source, aimed the measure of collagenolytic activity (50). A cysteine cathepsin could be responsible for this activity since it is a nonspecific assay and collagenolytic activity already was observed for cathepsin L-like cysteine peptidases from phylogenetically distant animals (129, 130). However the assays were performed in pH 7.2 for 6 hours at 37 ºC and we observed that after 3 hours incubation at 30 ºC in pH 7 the cysteine cathepsins from *Nephilengys cruentata* are no longer stable (Figure 4.4A).

The recombinant and native enzymes presented optimum pH 5.5 and are not stable in pH 7 or above, with 100% stability only in pH 5 (Figures 4.4 and 4.7). A difference could be observed regarded the stability in pH 6, since the native cathepsins still presented 45% of the activity after the incubation period and the recombinant one totally lost its hydrolysis capacity in this pH (Figures 4.4A and 4.7A, respectively). This difference observed is probably due to the different degrees of purification of native and recombinant samples, in which the native enzymes have more substrates which will protect the enzyme from denaturation. It is curious that the stability in pH 6 is low (about 45%) but 87% of the maximal activity is observed in this pH (Figures 4.4A and 4.4B). The low stability in pH 6 is probably not an artifact due to autolysis since the same thing should have occurred in pH 5 which also have high hydrolysis ratios. Probably the enzymes are really not stable in this pH for 3 hours at 30 ºC and the high activities in this pH could be observed because the activity assay usually last only one hour in contrast to the 3 hours incubation. The same thing (high activity and low stability) was observed in pH 7 as presented in the results. Differently from the activated recombinant enzyme, the recombinant zymogen presents a stability of approximately 100% up to pH 7 and still can keep 30% of the activity even in pH 10 (Figure 4.7A). These results are in accordance with

the literature where it already has been reported that the zymogen of cathepsin L is more stable in neutral and alkaline pHs in comparison to the mature form. Remaining activities of the mature human recombinant cathepsin L after one hour incubation at 37 ºC in pHs 6.0, 6.5 and 7.0 are 65%, 18% and 3% respectively (89).

The K_m of 13.5 µM observed for the recombinant enzyme is in the same range of other cathepsins L in the literature using Z-FR-MCA, such as human cathepsin L, 1 µM (95); cruzain ,10 µM (96); *Rhipicephalus microplus* cathepsin L, 18.8 µM (57); *Tenebrio molitor* cathepsins 1 (50 µM), 2 (9.6 µM) and 3 (16.4 µM) (90, 131); and *Fasciola hepatica* recombinant cathepsins L1 (18 µM) and L2 (39 µM) (132).

The optimum pH of the recombinant cathepsins L from *Rhipicephalus microplus* (57) and *T. molitor* pCAL1a (90) are respectively 5.5 and 5.0. These cathepsins L are lysosomal enzymes and, as cathepsin L1 and the other native cysteine cathepsins from *Nephilengys cruentata,* are not stable and/or do not have activity in neutral and alkaline pHs. Cathepsin L1 is very likely a lysosomal enzyme due to its very acidic characteristics similar to other lysosomal cathepsins L and also it was not found in the digestive juice. Cathepsins L4 and 8 are also probably acting inside the lysosomes because they weren´t identified at the digestive juice. Cathepsin L2 and cathepsin B1 were in both, the digestive juice and MMG (Table 4.1) is in accordance with recent discoveries showing a large variety of roles to these peptidases outside the cells (87). Furthermore, there are reports about secreted cathepsins L acting as digestive enzymes extracellularly (133, 134). In the present work, it was not possible to completely understand the processing and function of these enzymes in the digestive juice, mainly due to sample variation even after the same period of feeding and fasting. Although these results are still unclear and need more experimental data three different patterns were observed. When the enzymes are already activated in the digestive juice they are not stable under acidic conditions. In contrast to that in some cases acidic activation is needed to observe activity. In an intermediary condition a pre-existing activity is increased after activation (Table 4.3). A further investigation needs to be done about the role of the cysteine cathepsins in the digestive juice. Yet in the MMG it seems clearer that these enzymes have an important function in protein digestion under acidic conditions inside lysosome-like vesicles. A functional heterologous expression of cathepsin L2

and cathepsin B1 will help to investigate these unclear aspects of the enzymes found in the MMG secretion.

4.4.2 Zymogen activation and the general mechanism of digestion in spiders

Innumerous works have been reported that cysteine cathepsins zymogen can undergo acidic activation to the mature form (90, 128, 135, 136). In this work, we observed that the cysteine cathepsins present in the MMG of the spider *Nephilengys cruentata* can be activated after incubation in acidic pHs (Figure 4.3A). It was observed that the incubation in very acidic pHs 2.6 and 3 generated the highest activities and that an incubation for 2 hours in the first pH resulted in a full activation of the cysteine cathepsins (Figure 4.3B).

In the case of fasting spiders some samples could be activated while others were already active. Although the fasting period was controlled this differences could be related to a previous metabolic state of the collected spiders. The total activation observed for fed animals was 17.3 times. According to our results after 9 hours of feeding it seems that the enzymes are not completely activated, since activation still can be observed. The presence of zymogen in the MMG of fed animals was also confirmed by mass spectrometry, since for the cathepsins L1, 2 and 8, fragments of the propeptide were also identified (data not shown). In the tick *Ixodes scapularis* the cathepsins L and B are more active during the slow-feeding period between 4 and 6 days after attachment and activity is not observed in the first 2 days (61). However, the authors did not present data of *in vitro* activation and about the presence of cathepsins zymogens. In the spider *Nephilengys cruentata* it seems that the intracellular digestive process has some resemblance with ticks. It has been reported, in a histological study, that the spider *Coelotis terrestris* keeps unchanged intracellular digestive vacuoles after 2 days of feeding despite other ones were already digested (21). This means that not all digestive vacuoles are digested at once, some are kept for a future digestion and how this is regulated in spiders and scorpions is not understood. In all cases, spiders, ticks, and also in scorpions (28), part of the intracellular digestion takes a long time to be fully completed. The observation about the presence of zymogen after 9 hours of feeding together with the histological data above cited it is still not enough to exactly clarify the whole intracellular protein digestion period. However, a reasonable explanation to the

zymogen found in the MMG of fed animals is because the presence of intact and active digestive vacuoles will have as a consequence the enzymes as zymogen and the mature form respectively.

The intracellular digestion of the meal is a common feature to all the arachnids that have been studied so far (5, 61). Previously works showed that spiders use serine (48) and metallopeptidases (52) for the extra-oral digestion. Furthermore, in chapter 5 it will be presented our complete proteomics data analyses of the digestive juice and MMG showing that astacin-like metallopeptidases are the most abundant enzymes in the digestive juice. Despite the present study did not attempt to localize the cysteine cathepsins in the MMG it seems more likely that these enzymes have a main role in the intracellular digestion because of its acidic characteristics (Figures 4.3, 4.4, 4.6 and 4.7). Based on our findings and in the above cited literature the protein digestion in spiders can be depicted as follows: serine and metallopetidases are released by the secretory cells from the MMG to start the extra-oral liquefaction of the prey. After that, the partially digested meal will be absorbed by pinocytosis and the final digestion will take place inside the cells, with cysteine and aspartic (chapter 5) peptidases acting in an acidic environment from the lysosome-like vesicles. The fact that cysteine cathepsins were found at the protein and activity levels in the MMG and digestive juice of fasting animals and by the histological observation that after one day the secretory granules are resynthesized (21) are an indicative that the spider MMG is ready for the next predation event. Thus, in a short time period after prey capture, the enzymes needed for the digestion of a new prey, were already synthesized to start a new digestive cycle. This digestive model will be expanded in the next chapter.

4.5 Conclusions

In this chapter cysteine cathepsins were, for the first time, identified in the MMG homogenate and digestive juice of the spider *Nephilengys cruentata* using different techniques such as enzymology, molecular biology and mass spectrometry. In total, 9 different cathepsins L and 2 cathepsins B were identified at the mRNA level and the presence of 4 sequences of the former and one of the later was also confirmed by mass spectrometry. Their properties were studied and they presented acidic characteristics such as pH stability, pH optimum and acidic activation.

Cathepsins L1 and 2 were successfully expressed as recombinant enzymes but only cathepsin L1 activity could be measured showing acidic characteristics as the native enzymes. Summed to this, cathepsin L1 was not found in the digestive juice which indicates that this enzyme is a lysosomal peptidase. This is in accordance to the known mechanism of spider digestion already studied which combines extra-oral and intracellular digestion and these results biochemically complement previous ones showing alkaline serine and metallopeptidases acting in the digestive juice. The fact the cathepsin L2 and cathepsin B were found also in the digestive juice still needs more investigation but one or both enzymes are not stable in acidic pHs as the other cysteine cathepsins. In conclusion, the MMG of the spider *Nephilengys cruentata* possess cysteine cathepsins active under acidic conditions that are probably related to the intracellular digestion inside lysosome-like vesicles. Future work involves the antibody preparation of cathepsin L1 and 2 and a functional expression of the later. The antibodies will be used to see in which place of the cells these enzymes occur.

CHAPTER 5 – NEW INSIGHTS ABOUT THE MOLECULAR PHYSIOLOGY OF DIGESTION IN THE SPIDER *NEPHILENGYS CRUENTATA* REVEALED BY THE COMBINATION OF HIGH THROUGHPUT TECHNIQUES

5.1 Introduction

Spiders are efficient predators and as most part of arachnids they start to digest their prey extra-orally (EOD) (1) to then finish the process inside the midgut cells (21). Besides that they are capable of capturing and ingesting relative big preys, e.g. bats (46) and birds (45), they are also capable of survive to long fasting periods. Thus, spiders present a digestive system plenty capable to hydrolyze and store as much nutrients as possible from one single meal. Despite this enormous digestive and storage capacity, few studies attempted to describe this process and so far none has used high throughput techniques as next generation sequencing (NGS) and shotgun proteomics.

A histological study of the digestive system of the spider *Coelotis terrestris* during the feeding process brought the first knowledge about it in different physiological conditions (21). They showed that briefly after food uptake in order to initiate the extracellular digestion, the secretory cells discharge their granules into the lumen and that the digestive cells start the intake of predigested food by pinocytosis. The pinocytotic vesicles are assembled giving origin to the big digestive vesicles in which the slow intracellular digestion takes place. So far, none information about the molecules involved nor the possible mechanisms are available.

The enzymological studies revealed important informations about some enzyme classes and pHs of action present in the spider´s digestive juice such as peptidases (48, 49, 51, 52), carbohydrases (113, 137), esterases, phosphatases and nucleases (138). Nevertheless, there are not informations about the enzyme´s sequences and expression as well their subcellular location, with the exception of amino-terminal sequences from astacins (52). Moreover, these studies didn´t observe enzyme presence in fasting and fed spiders.

Thus, this work investigated the midgut and midgut glands (MMG) from the spider *Nephilengys cruentata,* under fasting and fed conditions, using the Illumina® NGS platform followed by the shotgun proteomics. Also, the digestive juice (DJ) was

studied by proteomics to assist the reconstruction of the digestive enzymes location. The use of transcriptomics and proteomics techniques are complementary and has been shown to be a very powerful tool in order to study organisms that still do not have their entire genome sequenced as planaria (102) and anopheline vectors (101). The use of such approaches allowed the proposal of a model for the digestive process in spiders, the subcellular location of digestive enzymes and the molecules involved in the endocytic pathway. We were also capable to identify some proteins expressed and translated in the digestive system which, until now, were exclusively associated to venom glands.

5.2 Materials and methods

5.2.1 Animals and sample obtaining

Adult *Nephilengys cruentata* females were collected in São Paulo city (Brazil), kept under natural photoregime and room temperature conditions with water spraying 4 times per week in their artificial environment. The animals were starved for at least one week and then fed with *Acheta domesticus*. After 1 and 9 hours eating the fed animals were dissected whereas the starved ones were dissected two weeks after start feeding. After anesthetizing the animals in a CO_2 chamber the dissection was performed in a cold isotonic saline solution (300 mM KCl pH 7) and the opisthosomal midgut with its glands (MMG) (figures 1.3A and 1.3B, chapter 1) were removed. In the samples used to RNA extraction, the saline solution was made with autoclaved sterilized water containing 0.1% (v/v) diethyl pirocarbonate (DEPC) and all dissection material was cleaned with 70% ethanol (v/v), placed under UV light for 30 minutes and subsequently heated to 150 ºC for 4 hours. Digestive juice (DJ) samples triplicates were collected by electrical or mechanical stimulus in 2 weeks fastened or 3, 9, 25 and 48 hours fed spiders.

5.2.2 Transcriptomics and proteomics procedures

Essentially, both high throughput techniques were performed as described in chapter 3 (items 3.2.2, 3.2.3 and 3.2.4). The differences will be highlighted in the below description.

The transcriptomics experiment was performed in triplicate for spiders that were under 3 different physiological conditions, fasting, 1 and 9 hours fed. The differential expression was studied using the software DESeq 2 (http://bioconductor.org/packages/2.13/bioc/html/DESeq.html). The gene ontology was obtained using the program BLAST2GO (106) with the non-redundant NCBI database. The enrichment analysis was performed using Fisher´s exact test with multiple testing correcting of false discovery rate using the differentially expressed genes as test group against the entire transcriptome data set in the formerly cited software.

In the proteomics experiment, the total time of the SDS-PAGE run was 5 min for the digestive juice samples and the gels were cut in 9 pieces with a blade.

5.3 Results

5.3.1 Transcriptome and proteome general features

The transcriptomics experiments were performed in triplicates with the opisthosomal midgut and midgut glands (MMG) of animals under three different physiological conditions: fasting, 1 and 9 hours fed. These nine data sets were collectively assembled generating a reference transcriptome. Table 5.1 exhibits the general features of the RNA-seq data of one fasting and one 9 hours fed sample. BLASTX results could be retrieved from 31 and 34% of the contigs in fasting and fed spiders respectively. Figure 5.1 depicts the relation of the contig length to a positive BLASTX result, with the smaller ones (less than 1680 bp) presenting less scores.

The biological process gene ontology (GO) terms associated to the reference transcriptome from the MMG of the spider *Nephilengys cruentata* are shown in the multilevel pie chart (figure 5.2A). A total of 33 different biological processes were identified with the GO term catabolic process being the most abundant present in 368 sequences. Most sequences presented BLAST best hits values with sequences from the tick *Ixodes scapularis* (data not shown).

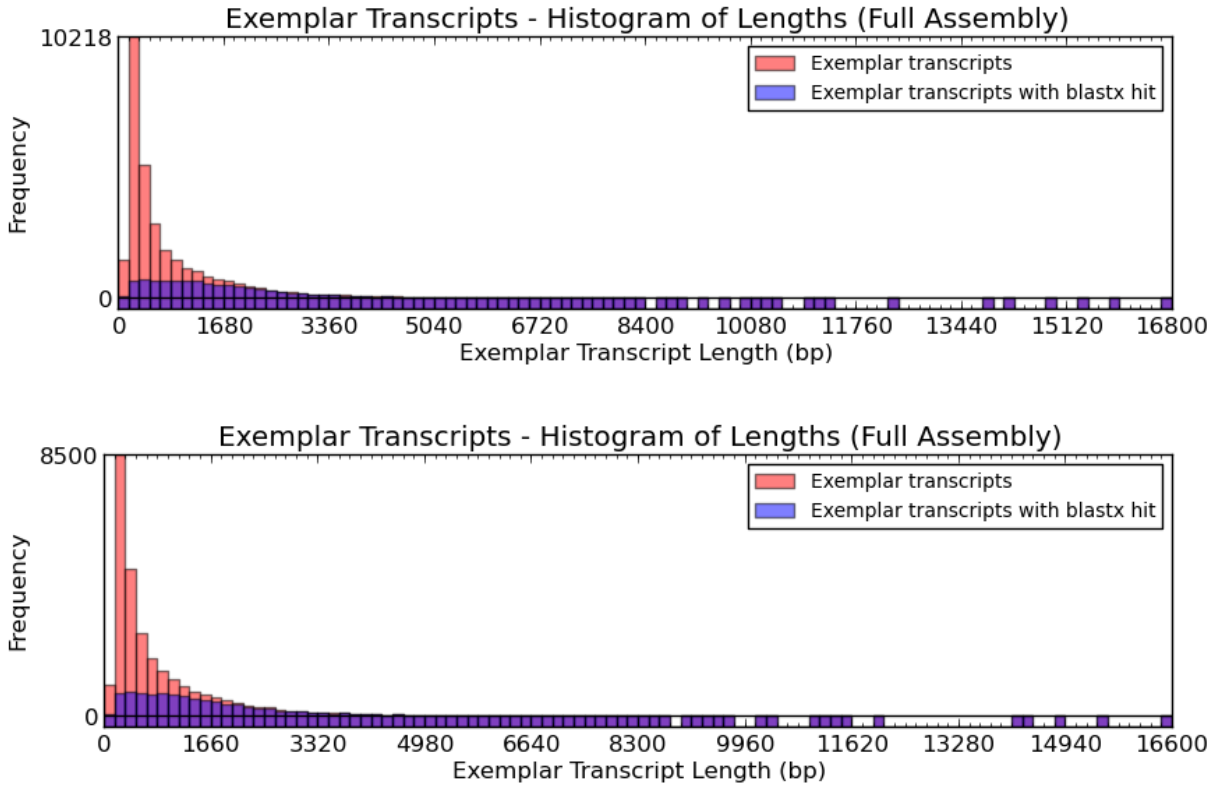
The use of RNA-seq information allowed the construction of a database making possible the identification of 1,150 and 556 protein sequences in the MMG of fasting and fed animals respectively. In the digestive juice (DJ) 310 proteins, in total, were identified summing the samples from fasting, 3, 9, 25 and 48 hours fed spiders.

Table 5.1 - Summary of *de novo* assembly results

Condition	Read Pairs Kept*	Number of Contigs	Mean Length (bp)	N50 Length (bp)	BlastX Hits
Fasting	12,022,570	31,318	846.9	1,530	9,724
Fed	11,751,528	27,925	860.3	1,486	9,621

Note: * After removing rRNA sequences and using Illumina® filtering

Figure 5.1 – Histogram of lenghts and BLASTX hits of the transcriptome assembled contigs from the midgut and midgut glands of *Nepnhilengys cruentata*



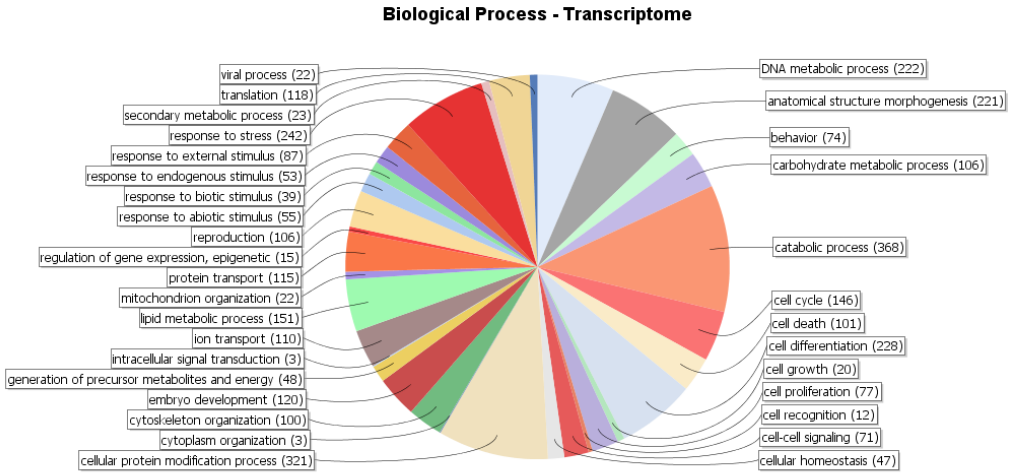
A) Fasting spider. B) Fed spider.

The identification of such large number of proteins in the digestive juice using the database generated from the MMG RNA-seq is the first molecular confirmation that the digestive juice is synthesized in the MMG.

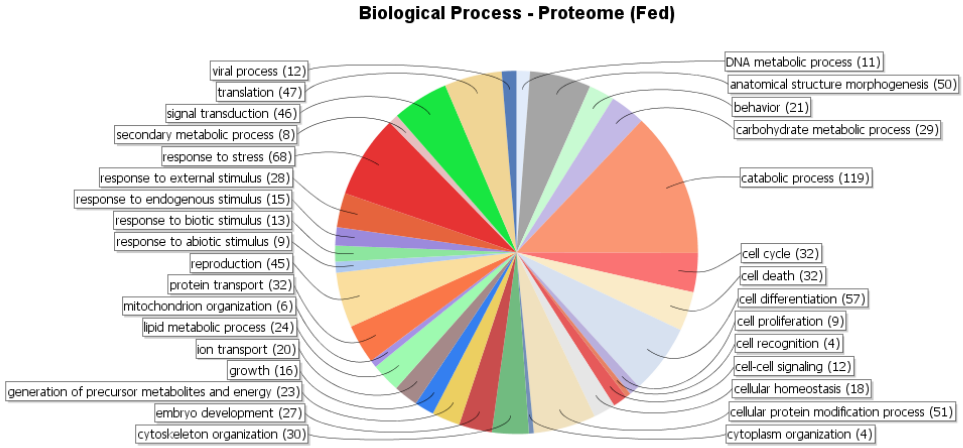
The biological process GO terms for the proteome of fed, fasting and digestive juice samples are respectively shown in figure 5.2B, 5.2C and 5.2D. Catabolic process is the major biological process represented in these three physiological conditions as was already verified at the transcriptomics studies. The multilevel pie charts of the molecular function and cellular component for the digestive juice are represented in figure 5.2E and 5.2F, respectively. Peptidase activity, with 45 sequences, is the most abundant molecular function identified whereas protein complex (38 sequences) is the most abundant cellular component.

Figure 5.2 - Multilevel pie charts of gene ontology terms

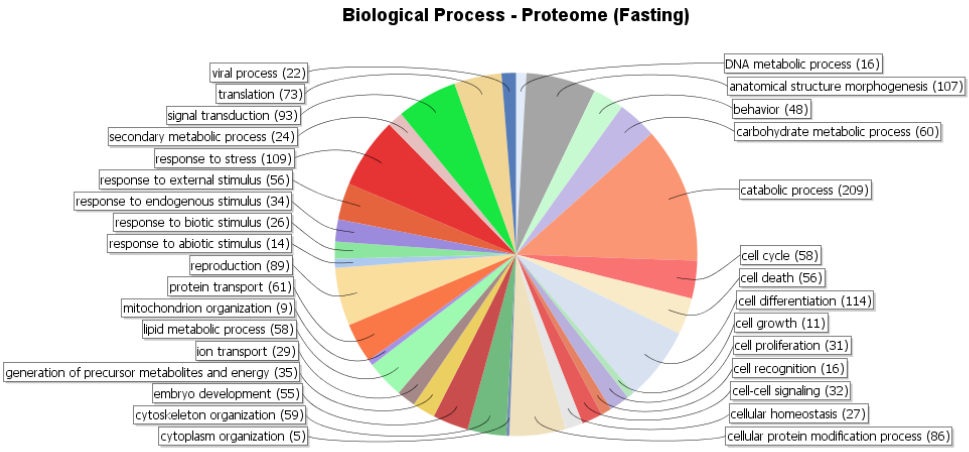
A



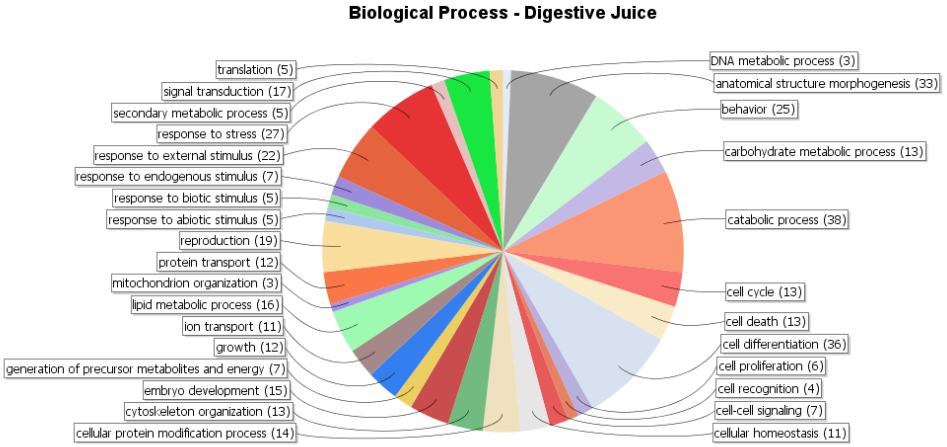
B



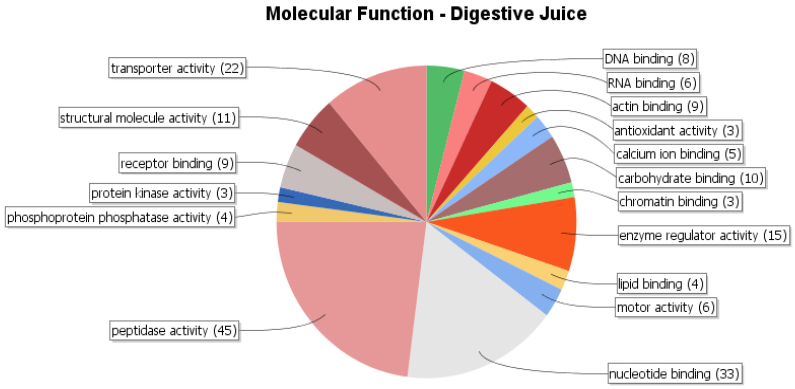
C



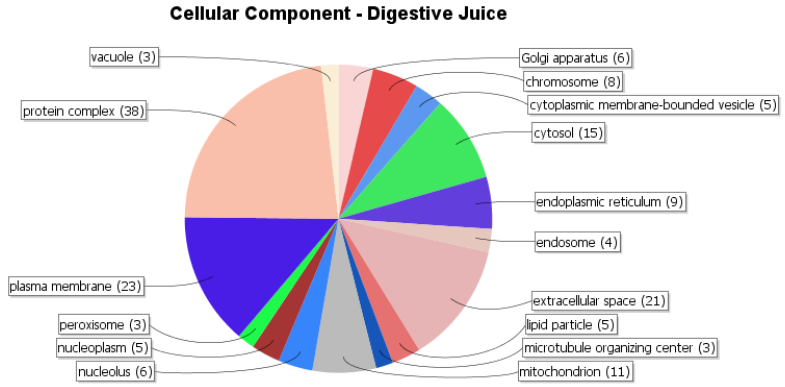
D



E



F



Multilevel pie charts for the biological process of the transcriptome (A) and proteome sequences obtained in the midgut and midgut glands of fed (B), fasting (C) and the digestive juice (D) of *Nephilengys cruentata*. E and F are respectively the charts from molecular function and cellular component from the digestive juice identified proteins.

If about one third of the contigs did not have BLAST score the proteins identified without BLAST hits represented respectively 14, 6.3 and 6.8% of the digestive juice, fasting and fed spiders samples, showing that at the protein level there are less unidentified sequences.

5.3.2 The digestive enzymes identified by proteomics

A total of 81 proteins were identified by mass spectrometry in the MMG of the spider *Nephilengys cruentata* (Table 5.2) with a possible digestive role.

Table 5.2 - List of proteins identified by mass spectrometry according to the sample

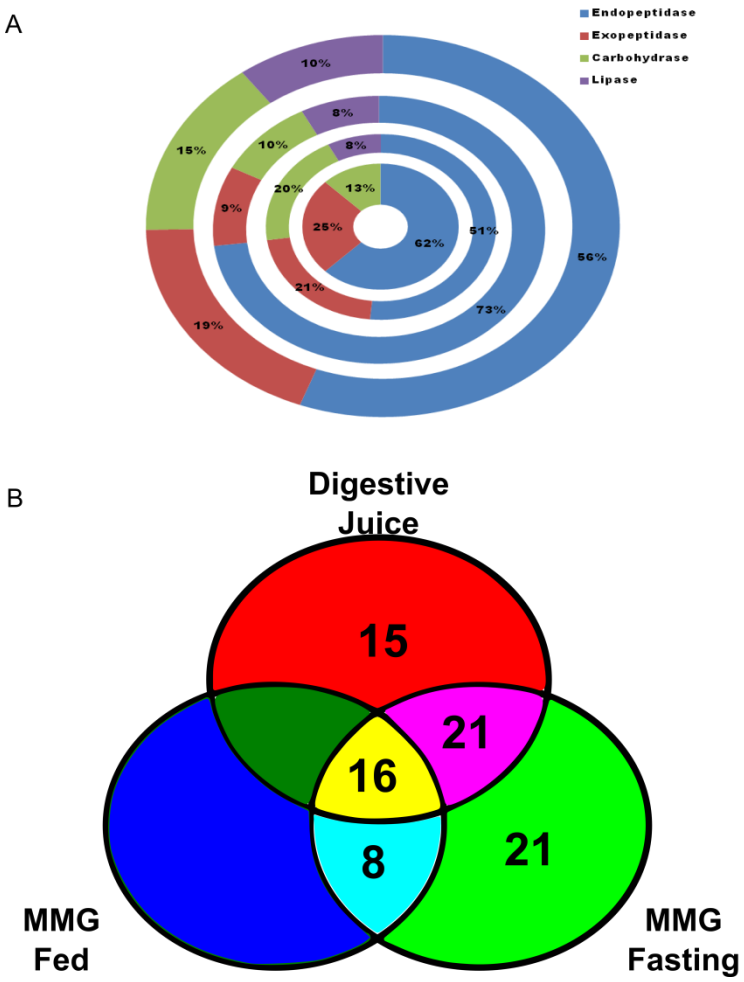
3 conditions (16)	Fasting and Fed (8)	Fasting and Digestive Juice (21)	Fasting (21)	Digestive Juice (15)
Astacin-like metallopeptidase 30	Astacin-like metallopeptidase 26	Astacin-like metallopeptidase 23	lysosomal alpha-mannosidase	Astacin-like metallopeptidase 5
Astacin-like metallopeptidase 1b	Cathepsin L1	Astacin-like metallopeptidase 15	Astacin-like metallopeptidase 36	Astacin-like metallopeptidase 31
Astacin-like metallopeptidase 2	Cathepsin L8	Astacin-like metallopeptidase 18	cathepsin L-like cysteine peptidase-N4	Astacin-like metallopeptidase 6
Astacin-like metallopeptidase 28	Legumain	Astacin-like metallopeptidase 17	Dipeptidyl peptidase 2	Astacin-like metallopeptidase 32
Astacin-like metallopeptidase 11	Dipeptidyl peptidase 1	Astacin-like metallopeptidase 3	Tripeptidyl peptidase 2	Astacin-like metallopeptidase 16b
Astacin-like metallopeptidase 21	Dipeptidyl peptidase 3	Astacin-like metallopeptidase 19a	Carboxypeptidase B 2	Astacin-like metallopeptidase 25
Astacin-like metallopeptidase 43	Alpha-aspartyl dipeptidase	Astacin-like metallopeptidase 9	Probable carboxypeptidase PM20D1	Retinoid-inducible serine carboxypeptidase
Astacin-like metallopeptidase 46	Alpha-L-fucosidase	Astacin-like metallopeptidae 8a	Probable aminopeptidase NPEPL1	Triacylglycerol lipase 1
Cathepsin L 2		Astacin-like metallopeptidase 10	Glutamyl aminopeptidase	Triacylglycerol lipase 2
Cathepsin B1a		Astacin-like metallopeptidase 7a	Glutamyl aminopeptidase	Triacylglycerol lipase 3
Cathepsin D-like aspartic peptidase 1		Astacin-like metallopeptidase 22	Methionine aminopeptidase 2	CUB and LDL domains-containing trypsin-like serine peptidase 2a
Serine carboxypeptidase CPVL		Astacin-like metallopeptidase 14	Maltase-glucoamylase, intestinal	CUB domain-containing trypsin-like serine peptidase 3
carboxypeptidase B 1		CUB domain-containing trypsin-like serine peptidase 4	Lysosomal alpha-glucosidase	CUB and LDL domains-containing trypsin-like serine peptidase 3
Lysosomal protective protein		Alpha-amylase	Neutral alpha-glucosidase AB	CUB and LDL domains-containing trypsin-like serine peptidase 4
lysosomal alpha-mannosidase		Chitotriosidase	Lysosomal alpha-glucosidase	Deoxyribonuclease
Beta-Hexosaminidase subunit beta		Beta-galactosidase	Mannosyl-oligosaccharide alpha-1,2-mannosidase isoform A	
		Triacylglycerol lipase 4	Alpha-galactosidase A	
		CUB and LDL domains-containing trypsin-like serine peptidase 1a	Triacylglycerol lipase 5	
		CUB and LDL domains-containing trypsin-like serine peptidase 2b	Putative phospholipase B-like 2	
		CUB domain-containing trypsin-like serine peptidase 2	Group XV phospholipase A2	
		CUB domain-containing trypsin-like serine peptidase 1	Phospholipase A2	

The endopeptidases are the most representative, containing 56% of the sequences. Exopeptidases, carbohydrases and lipases are respectively 19, 15 and 10% of the identified enzymes. The isolated analysis of digestive juice and MMG from fasting and fed spiders indicated that endopeptidases still are the most abundant enzymes with respectively 73, 51 and 62% of the digestive enzymes in the DJ, fasting and fed spiders (Figure 5.3A). Only one deoxyribonuclease was identified in the DJ and it is not being represented in these charts. These results clearly demonstrate that peptidases (endo + exo) are the most abundant enzymes in number of distinct isoforms.

A Venn diagram (Figure 5.3B) and table 5.2 show the number of shared identified proteins and their identification in each sample. Sixteen sequences were found in the DJ, fasting and fed animals and are mainly related to astacin-like metallopeptidases. However, cathepsins L2, D1 and B1a, in spite of other proteins, are also present in the three conditions (Table 5.2). It is important to highlight that cathepsin D 1 was identified in the digestive juice only in fasting animals. There are 8 sequences that were observed in fed and fasting animals, 21 in the digestive juice and fasting conditions, 21 only in fasting spiders and 15 only in the digestive juice (Table 5.2 and figure 5.3B).

The number of different enzyme types from table 5.2 can be better visualized in figure 5.4A in which the proteins are displayed with the number of different copies. Astacins, with an impressive number of 28 different enzymes, presented the highest number of isoforms and only 2 were not found in the DJ. Serine peptidases are the second in the list presenting 9 sequences. All of them contain the trypsin domain but some of them also have CUB and LDL domains and others have only the CUB domain. Besides that, it is possible that some these proteins annotated as trypsins could have chymotrypsin-like specificities. In the absence of biochemical evidence or a more detailed analysis of the subsites, we will keep their annotation identification as trypsins. Six different cysteine peptidases were found and they are represented by cathepsins B and L and legumain. Five different triacylglycerol lipase sequences were identified. The most diversified carbohydrase were alpha-glucosidase and alpha-mannosidase with 3 different isoforms each one (Figure 5.4A).

Figure 5.3 - General quantitative values relationships between the proteome data of different samples



A) Donut chart of each hydrolyze group percentages identified in the proteome. From outside to inside: proteome sum (all samples), digestive juice, MMG of fasting and fed spiders, respectively. B) Venn diagram showing the relation between the numbers of proteins identified in each sample (digestive juice and MMG of fasting and fed spiders).

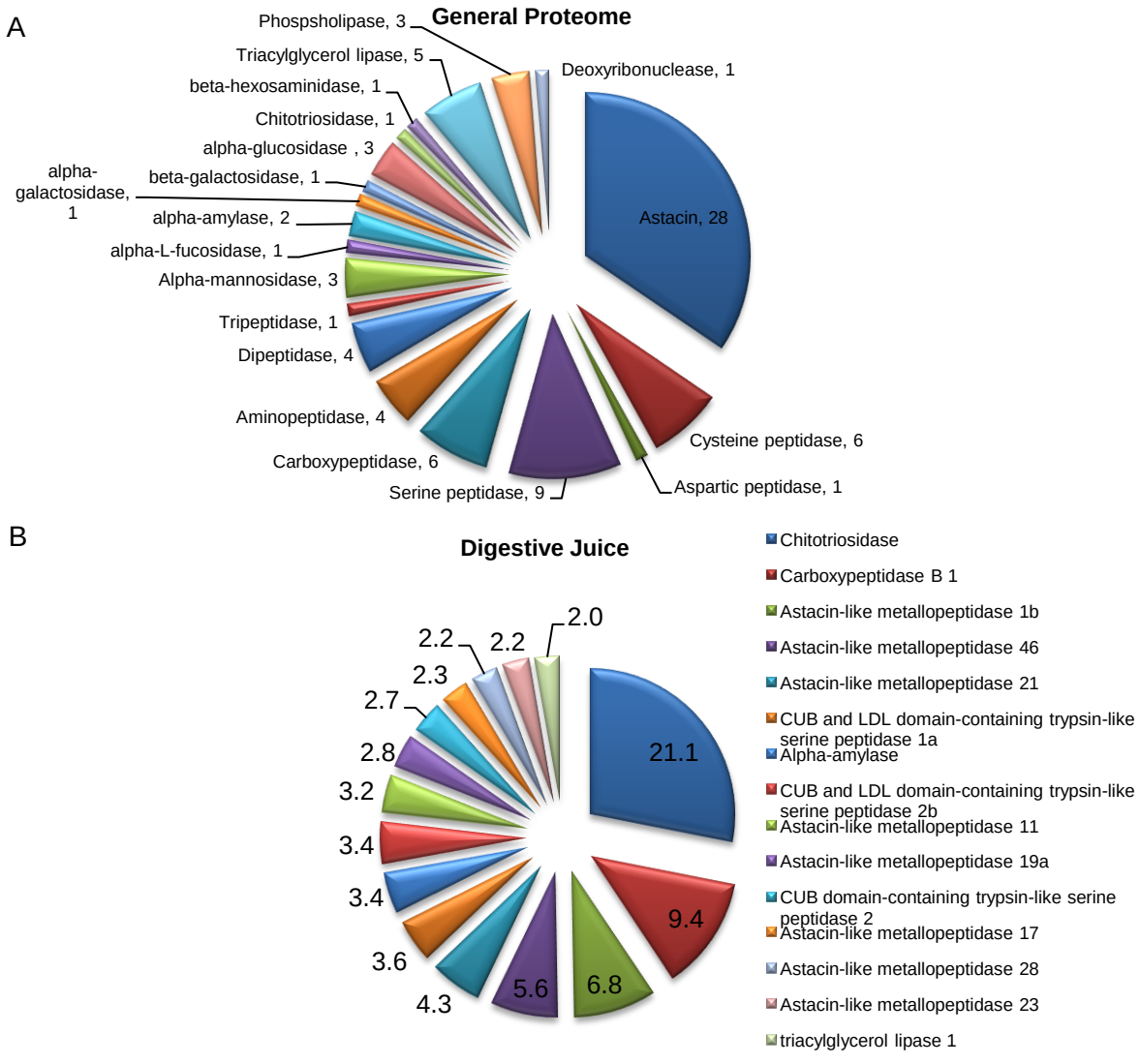
5.3.3 Label-free quantitative analysis

Label-free quantitative proteomics based on the normalized spectra counting (NSC) has been shown to be a simple and efficient method for quantitative proteomics (139). Although the DJ samples were individually analyzed after different periods of feeding for most part of the proteins there was not significant differences between these samples so the data set will be normally treated together. When important differences were observed this will be specified in the text. In the digestive juice, fasting and fed spiders the sequences from table 5.2 represented respectively 38.2, 6.3 and 3.4% of the total NSC. The percentages of the label-free quantitation that it will be subsequently presented are relative only to the possible digestive

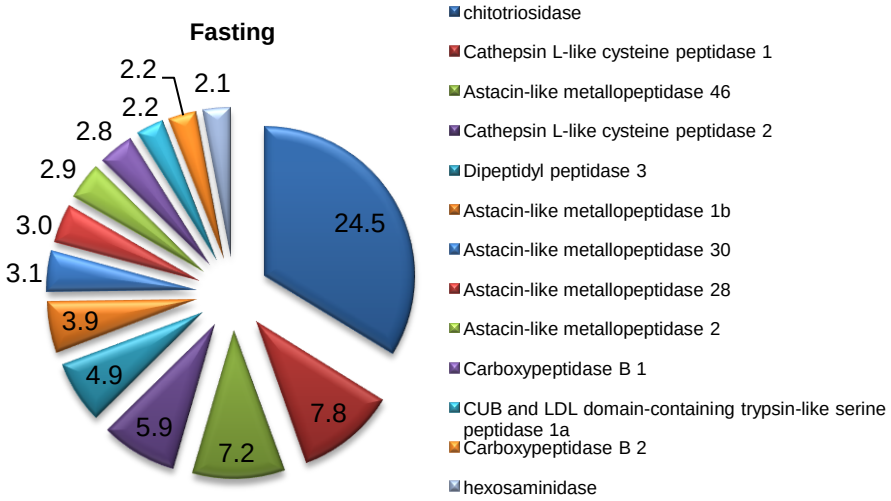
enzymes NSC from table 5.2 for a comparative reason. Chitotriosidase is quantitatively the most abundant protein in the digestive juice and in the MMG of fasting animals with respectively 21.1 and 24.5% of the NSC. Carboxypeptidase B is the second most abundant enzyme in DJ with 9.4% whereas in fasting spiders 7.8% of the digestive enzymes are the cathepsin L 1(figure 5.4B and 5.4C). The most representative digestive enzymes in MMG from fed animals are astacin 30 (25.4%) and cathepsin L1 (15.4%) (Figure 5.4D). Curiously, chitotriosidase was not found in fed animals.

Peptidases, carbohydrases and lipases are respectively 71.4, 25.4 and 3.2% of the digestive enzymes in the DJ whereas in fasting samples they represent 67.1, 31.8 and 1.1%. In the MMG of fed animals mainly peptidases were identified (98.2%) and none lipase sequence.

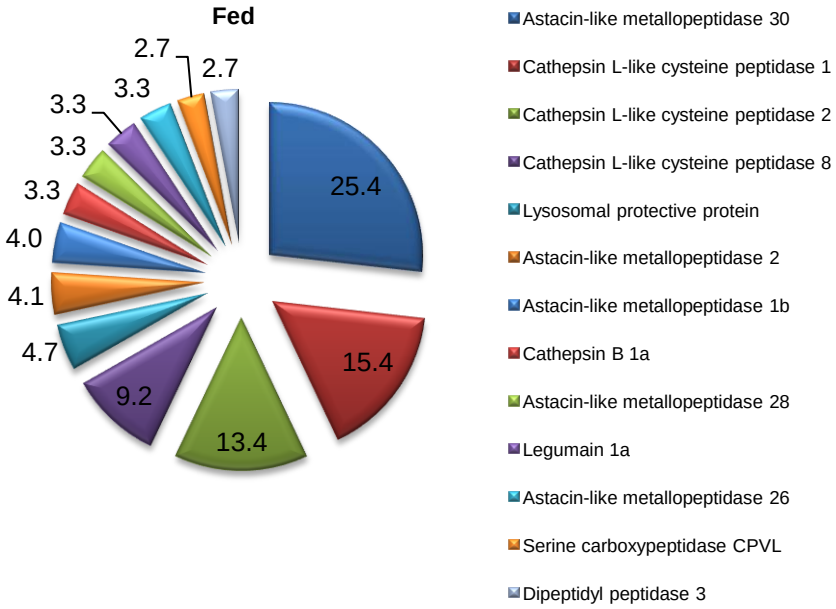
Figure 5.4 - Qualitative and quantitative proteome data



C



D



A) Number of each different enzyme type identified in all the samples. B) Quantitative pie charts of the enzymes identified in the digestive juice and the MMG of fasting (C) and fed (D) animals. The quantitative values are percentages from the normalized spectra (NSC) counting for each individual sample taking in account only the digestive enzymes from table 5.2. The triplicate data of each condition was loaded together and the NSC calculated for each individual pool. The pie charts from items B, C and D display only proteins with at least 2 percent of the NSC.

The astacin endopeptidase diversity is reflected in the percentage that this enzyme represents in each sample, respectively 37.7, 29.6 and 41% in the DJ, fasting and fed animals, with 22, 20 and 11 different isoforms respectively. Some of these astacins are present in more than one sample (Table 5.2). The cathepsins L are more abundant in the MMG tissue, summing 38% in fed and 14% in fasting spiders. In contrast to that, this enzyme is only 0.3% of the digestive enzymes from the DJ. Oppositely to cathepsin L, trypsin-like serine endopeptidases seem to be more important in the digestive juice rather than in the MMG tissue. They represent

17% of the digestive enzymes from the DJ with 9 different isoforms. In fasting animals they represent 4.9% and 5 isoforms were found and in fed animals they were not identified.

Despite carbohydrases represent 25.4% of the digestive enzymes in the DJ only three glycosyl hydrolases were identified, the already mantioned chitotriosidase, an alpha-amylase (3.5%) and a beta-hexosaminidase (0.9%). Fed animals MMG displayed only one alpha-L-fucosidase (1.1%), one beta-hexosaminidase (0.4%) and one alpha-mannosidase (0.3%). However in fasting animals 10 other carbohydrases were also identified in addition to chitotriosidase (Table 5.2) and they sum 7.4% of the digestive enzymes. Lipases, in general, were the less represented digestive enzymes with none identified in fed animals. In fasting samples 2 triacylglycerol lipases (TAGL) and 3 phospholipases were found summing 1.1% whereas in the DJ 4 TAGLs representing 3.2% of the digestive enzymes could be detected. The most abundant was TAGL 1 with 2% of the NSC (Figure 5.4B).

One deoxyribonuclease was found in the DJ and represents 0.1% of the digestive enzymes NSC. This protein was not found by mass spectrometry in the MMG tissue.

5.3.4 Differential expression analysis of the transcriptome data

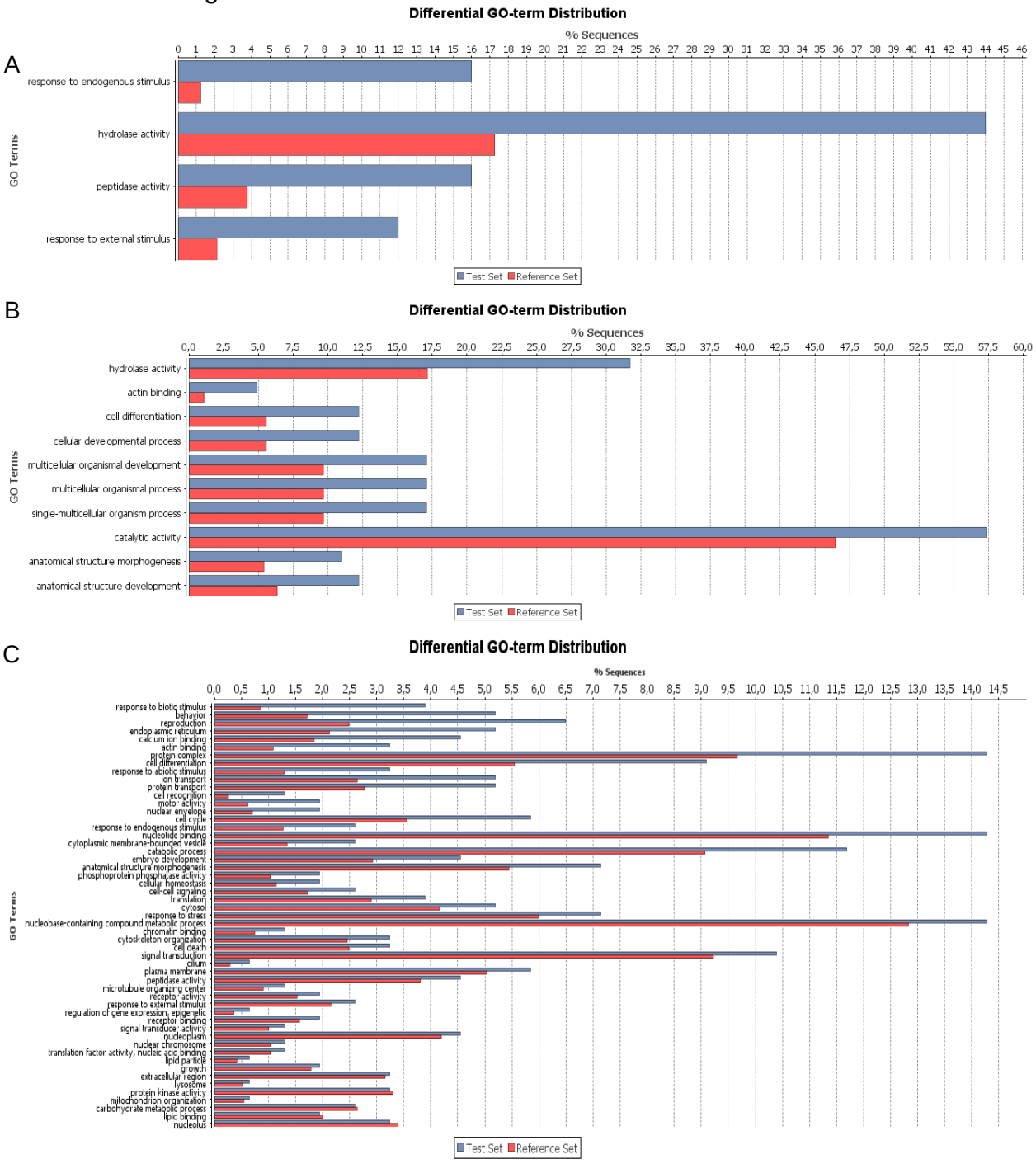
Differential expression study was performed with samples from fasting, 1 and 9 hours fed spiders, which were analyzed in pairs in order to achieve the genes that are up and down regulated. All the numbers related to differentially expressed genes were considered with statistical significance with p-value < 0.01.Comparison between 1 hour after feeding samples versus fasting samples retrieved 48 up and 12 down regulated genes of which 20 in the 1 hour fed sample and 9 in the fasting samples were non-identified contigs (Figure 5.5). The results obtained from the comparison of the 9 hours after feeding sample versus fasting samples analysis are that 169 genes are over expressed while 113 are under expressed, with respectively 69 and 39 unidentified. In the case of the comparison of the 9 hours after feeding sample versus 1 hour after feeding sample, a larger number of up and down regulated genes was obtained, respectively 262 and 219. Both up and down regulated contigs (test set) were submitted to an enrichment analysis based on the GO terms using the

reference transcriptome (reference set). The down regulated expressed genes did not present statistical significance in the enrichment analysis. However, this analysis for the up regulated contigs resulted in statistical significant differential expression (Figure 5.5).

Although in the 1 hour fed samples versus the fasting samples comparison the number of differentially expressed genes was lower than the observer in the other comparisons, the enrichment analysis showed more clearly the differences between the GO terms (Figure 5.5A). All the gene ontology terms of the up regulated genes after 1 hour feeding are clearly related to the feeding stimulus as can be observed in figure 5.5A. The comparison of gene expression between an animal that fed for 9 hours and the ones in fasting conditions evidenced that the up regulated genes in this condition are not only exclusively involved with diet hydrolysis but they are clearly involved with other process like cellular and organism development and cell differentiation. The increase of expression observed in actin binding proteins and also other proteins related to cell structure are probably related to the organization needed for the vesicular trafficking (Figure 5.5B). In the analysis of two animals fed for different periods of time (9 hours versus 1 hour), the number of gene ontology terms obtained related to the up regulated genes was larger than in the other two analyses (Figure 5.5C). Some GO terms that deserve attention are motor activity, protein transport and response to biotic stimulus due to their relation to intracellular digestive process.

Some of these genes differentially expressed related to digestive enzymes and proteins related to the vesicular trafficking mentioned above are listed in table 5.3 with their respective number of $\log_2$ fold change and p-value. The comparison of 1 hour feeding to fasting samples evidenced the up regulation of 4 astacins (named as astacin 9, 33, 42 and 46), 1 CUB domain-containing trypsin, 1 carboxypeptidase B1 and 1 leucine-rich repeat molecule (Table 5.3). None digestive enzyme or related protein is down regulated. In the 9 hours vs fasting samples comparison 7 digestive enzymes are up regulated, for instance astacins, trypsin, carboxypeptidases and chitotriosidase. The up regulation of two Ras proteins and one vesicle-fusing ATPase were also observed in this comparison while the lysosomal-trafficking regulator and Rab 3A are down regulated (Table 5.3). When both fed conditions are confronted, only carboxypeptidase E and astacin 35 are up regulated in the 9 hours fed sample but

Figure 5.5 – Enrichment analysis of the GO terms from the sequences differentially expressed (test set) versus the reference transcriptome (reference set) using the Fisher´s exact test



The graphics were obtained using default parameters in the software Blast2GO (106). A) Up-regulated genes 1 hour fed versus fasting. B) Up-regulated genes 9 hours fed versus fasting. C) Up-regulated genes 9 hours fed versus 1 hour fed.

also proteins involved in the vesicular trafficking and protein processing prior to vesicular targeting (Signal peptidase complex catalytic subunit SEC11A) were up regulated. Down regulated identified contigs were lysosomal-trafficking regulator, cathepsin L 4 and a leucine-rich repeat-containing molecule (Table 5.3).

Table 5.3 - Genes differentially expressed in the three physiological conditions: fasting, 1 and 9 hours fed animals

9 hours vs 1hour		
Up regulated	**log_2 Fold Change**	**p-value**
Vesicle-fusing ATPase 1	0.78	0.00321
Carboxypeptidase E	0.85	0.00204
Signal peptidase complex catalytic subunit SEC11A	0.86	0.00015
Clathrin heavy chain 2	0.92	0.00224
Ras-specific guanine nucleotide-releasing factor RalGPS1	1.21	0.00014
Astacin-like metallopeptidase 35	1.31	0.00122
Down regulated	**log_2 Fold Change**	**p-value**
Lysosomal-trafficking regulator	-1.28	0.00016
Cathepsin L-like cysteine peptidase 4	-1.25	0.00018
Leucine-rich repeat-containing protein 58	-0.90	0.00288
9 hours vs Fasting		
Up regulated	**log_2 Fold Change**	**p-value**
CUB domain-containing trypsin-like serine peptidase 4	1.08	0.00105
Chitotriosidase	1.18	0.00003
Astacin-like metallopeptidase 22	1.19	0.00039
Astacin-like metallopeptidase 46	1.20	0.00040
Carboxypeptidase B 1	1.22	0.00001
Carboxypeptidase E	1.32	0.00000002
Ras-specific guanine nucleotide-releasing factor RalGPS1	1.40	0.00015
Vesicle-fusing ATPase 1	1.40	0.00010
Astacin-like metallopeptidase 19a	1.56	0.00043
Astacin-like metallopeptidase 9	1.60	0.00009
Ras-related protein ced-10	1.62	0.0000002
Down regulated	**log_2 Fold Change**	**p-value**
Lysosomal-trafficking regulator	-1.34	0.00058
Rab-3A-interacting protein	-1.24	0.00004
1 hour vs Fasting		
Up regulated	**log_2 Fold Change**	**p-value**
CUB domain-containing trypsin-like serine peptidase 2	0.96	0.00002
Astacin-like metallopeptidase 42	1.04	0.00000
Astacin-like metallopeptidase 46	1.05	0.00002
Astacin-like metallopeptidase 33	1.30	0.00002
Carboxypeptidase B 1	1.34	0.00000
Leucine-rich repeat-containing protein 15	1.38	0.00002
Biotinidase	1.68	0.00001
Astacin-like metallopeptidase 9	1.76	0.0000002

5.3.5 Non-digestive proteins identified in the DJ and MMG of the spider Nephilengys cruentata

The focus of the non-digestive proteins analysis will be directed to the molecules identified in the digestive juice since their secretion, regardless the mechanism, is an empiric fact. This work provided for the first time strong evidences that some proteins described as toxins in spider venoms can be transcribed and translated in the digestive system of a spider and are secreted to compose the digestive juice. Ten toxins so far associated only to spiders venom were found in the MMG transcripts. Six of them were confirmed as peptides in the digestive juice. Among these toxins there are a venom peptide isomerase, a cysteine-rich secretory protein, 6 different ctenitoxins and 2 aranetoxins. Two ctenitoxins and aranetoxins were found only in the MMG (Table 5.4). Peptidase inhibitors were identified as well. Two cystatins and two serpins were detected by mass spectrometry and only one of each is also in the DJ. One Kunitz-like inhibitor is present in the MMG and 2 are in both MMG and DJ (Table 5.4).

The venom peptide isomerase, cysteine-rich secretory protein and U24-ctenitoxin-Pn1a are quite abundant in the DJ with respectively 6.3, 4.7 and 10.8‰ of the NSC. This latest cited ctenitoxin is also abundant in the MMG with 5 and 2.5‰ in fasting and fed animals. L-cystatin, serpin B3, Kunitz-like 1, and alpha-1 inhibitor 3 are respectively 3.2, 1.4, 1.3 and 3.3 ‰ of the NSC in the DJ whereas in the MMG of fasting spiders they represent respectively 1.2, 0.9, 1.4 and 0.3‰. In the proteome analysis only the serpin B3 was also identified in the MMG of fed animals (Table 5.4). An elastase inhibitor was also identified in the DJ secretion (0.07‰).

All the complete sequences found in the DJ, except for the venom peptide isomerase, present a signal peptide which is usually associated with proteins adressed to secretion or to lysosome. Some of the proteins that are exclusively identified in the MMG did not have this targeting signal, with the exception of one U9-ctenitoxin-Pr1a (Table 5.4).

Transferrin was quite abundant in the DJ with 10.9‰ of the NSC. Fourteen molecules contain the leucine-rich repeat domain and they sum 119‰. All these molecules were automatically annotated with different names but a manual check in their sequences revealed that they usually only contained this domain and not the ones needed for the automated annotation be correct, thus they were grouped

together only as leucine-rich repeat-containing proteins. A peritrophin with six chitin-binding domains was found in the DJ (0.3‰) and in samples from fasting animals (0.55‰).

Table 5.4 - Non-digestive proteins identified by mass spectrometry

Protein	Digestive Juice (‰)	Fasting (‰)	Fed (‰)	SPCS
Venom peptide isomerase (heavy chain)	6.3	-	-	None
Cysteine-rich secretory protein	4.7	-	-	Inc
U24-ctenitoxin-Pn1a	10.8	5.0	2.5	17-18
U24-ctenitoxin-Pn1a	0.7	-	1.5	17-18
U24-ctenitoxin-Pn1a	0.9	0.5	1.7	22-23
U24-ctenitoxin-Pn1a	0.4	0.7	-	17-18
U9-ctenitoxin-Pr1a	1.47**	0.4	0.2	18-19
U9-ctenitoxin-Pr1a	-	0.1	-	Inc
Protease inhibitor U1-aranetoxin-Av1a	-	2.2	3.0	None
U3-aranetoxin-Ce1a	-	0.4	-	None
L-cystatin	3.2	1.2	4.7	17-18
Cystatin A2	-	2.7	-	None
Uncharacterized serpin-like protein TK1782	-	0.2	-	None
Serpin B3	1.4	0.9	0.5	Inc
Serpin B6	-	0.4	-	None
Kunitz-like 1	1.3	1.4	-	15-16
Kunitz-like 2	0.7	0.5	0.3	19-20
Kunitz-like 3	-	0.1	0.3	None
Alpha-1 inhibitor 3	3.3	0.3		27-28
Peritrophin	0.3	0.55	-	
Transferrin	10.9	*	*	Inc
sum of leucine-rich repeat-containing proteins	119	*	*	*
Rab GDP dissociation inhibitor beta	0.8	8.6	10.4	*
Charged multivesicular body protein 2b	3.8	0.3	0.12	*

Notes: Per thousand values are related to the total normalized spectra counting (NSC) for each sample.
Digestive juice NSC is the sum of all different periods of feeding (fasting, 3, 9, 25 and 48 hours).
Only proteins identified with at least two peptides with a false discovery rate of 0.1%.
SPCS: signal peptide cleavage site
Inc: incomplete N-terminal sequence
- protein not identified in that sample
* protein not searched in that sample or SPCS not analyzed.
**identified in a DJ juice sample with uncontrolled feeding time and prey meal.

Molecules involved in the vesicular trafficking were identified in the DJ and MMG. Rab GDP dissociation inhibitor was abundant in the MMG with respectively 8.6 and 10.4‰ in fasting and fed spiders and also was found in the DJ (0.8‰). In contrast to that, charged multivesicular body protein 2b was more represented in the

DJ instead of in the MMG (Table 5.4). Rab-7a is present in the DJ and MMG of fasting animals while Rab-10 only in the DJ. Moreover Rabs 2A, 5C, 6A, 21 and 35 were identified only in fasting spiders. V-type proton ATPases A and B were detected in the DJ only in fasting animals.

Other proteins found in the DJ are: molecules involved in the immune system like 3 peptidoglycan-recognition protein, techylectin and CD109 antigen; chaperons as 78 kDa glucose-related protein, heat shock proteins 83 and 70B2; clathrin; mitochondrial enzymes (malate dehydrogenase, citrate synthase) and cytoplasmatic enzymes (isocitrate dehydrogenase).

5.4 Discussion

5.4.1 General analysis of the transcriptome and proteome data

The data set obtained in this study is the first massive sequencing at both mRNA and protein levels in the midgut and midgut glands of a spider. About one third of *de novo* assembled contigs presented similarity with proteins from the databases while of the proteins identified by mass spectrometry a mean of one tenth were not similar to any database sequence. These results show not only that this is a successful combination in order to study animals with unknown genomes but also that the study of new organisms is important to increase the amount of sequences (mRNA and protein) in the databases with a subsequent better understanding of the biological processes in future studies. The use of these techniquesusing samples from the MMG of the spider *Nephilengys cruentata* allowed the identification of the secreted proteins in this spider digestive juice which is the first molecular evidence that corroborate the previous histological observation (21) that the secretory cells contain the secreted digestive enzymes. Furthermore, some proteins were identified in both secretion and tissue (Table 5.2). This combination was also important for doing a correct assembly of the sequences from the transcriptome data based on the proteome data as already observed by other authors (102). For instance chitotriosidase, peptide isomerase and carboxypeptidase B1 could have their sequences corrected after a comparison of both data.

Another confirmation that both transcriptomics and proteomics experiments were well succeeded is the multilevel pie charts (Figure 5.2). In the reference transcriptome there are 33 gene ontology terms (Figure 5.2A) while in the proteomes of fed and fasting spiders 32 (Figure 5.2B and 5.2C). In all cases many sequences are associated to the GO terms. Such kind of diversified GO terms is a strong indicative that both experiments were able to do a deep molecular analysis in the MMG tissue of the spider *Nephilengys cruentata*.

5.4.2 Corroborating the historical biochemical data with "Omics"

5.4.2.1 Peptidases

The analysis done in the MMG of the spider *Nephilengys cruentata* allowed performing a comparison with the available literature data which so far was based only in biochemical assays without the identification of any complete protein sequence. Table 5.5 summarizes the relationships between biochemical data on literature and the present study. In a series of four different articles, Mommsen was in between the first investigators to do a fine biochemical characterization of many hydrolase activities in the digestive juice of a spider (48, 113, 137, 138). All peptidase activities identified by this author (48) could be associated with the proteins identified by mass spectrometry, excepted for the lack of an aminopeptidase identification in the present study. The carboxypeptidase A activity is related to the carboxypeptidase B 1, since they both belong to the subfamily M14A and blast searches showed that carboxypeptidase B 1 presents high similarities with both carboxypeptidases A and B (data not shown). This author reported that this enzyme was one of the most active hydrolases in the DJ of *Tegenaria atrica*. Figure 5.4B shows that this is the second most abundant enzyme (9.4%) in the DJ of *Nephilengys cruentata*. The protein that Mommsen named "protease" did not present tryptic, catheptic or peptic activity. By exclusion of the endopeptidases found only astacin(s) are likely to be related to this activity. Furthermore there are some reasons why the "protease" could be a mixture of astacins: 1) the measured mass was above 25 kDa and it was not exactly determined. Twenty one of the 26 astacins from the DJ have a mass between 25 and 33 kDa (data not shown); 2) the broad pH range of activity using azocasein (pHs 5-10) is an indicative that more than one enzyme is present in this sample; 3) in our

Table 5.5 - Enzymes obtained in this study and their relationship with the literature data

Reference	Enzyme	This work
Mommsen (48)	Carboxypeptidase A	Carboxypeptidase B 1
	Protease	Astacin(s)
	Chymotrypsin/trypsin	CUB and CUB/LDL Trypsins
	Aryl aminopeptidase	N.I.
Mommsen (113, 137)	Alpha-amylase	Alpha-amylase
	Chitinase	Chitotriosidase
	beta-N-acetylglucosaminidase	beta-hexosaminidase (?)
	Beta-glucuronidase	Beta-galactosidase
	Alpha- and beta-glucosidase	N.I.
Mommsen (138)	Tributyrinase	TAGs 1, 2, 3 and 4
	Carboxylic esterase	TAGs 1, 2, 3, 4 and Abhydrolase domain-containing protein 11
	Lipase	TAGs 1, 2, 3 and 4
	Desoxyribonuclease	Deoxyribonuclease
Kavanagh and Tillinghast (49)	Proteases A, B, C and D	Astacins
Atkinson and Wright (50)	"collagenase"	Astacins and trypsins
Foradori (51)	"collagenase"	Astacins and trypsins
Foradori (52)	p16 an p18	Astacin 19a

Notes: N.I.: not identified
?: probable correlation

group astacin activity was measured using casein as substrate (Fuzita et al., unpublished). Two chymotrypsins- and 2 trypsins-like activities were identified in the *Tegenaria atrica* DJ while we found 9 proteins with a trypsin domain. Two remarks must be done at this point. Firstly, the trypsin domains identified did not have their S1 subsite analyzed, which means that some of these proteins could present chymotrypsin-like activity. Secondly, the measured mass of the trypsin-like activities in Mommsen´s work was 9 kDa which is a small molecular mass for a trypsin even in the absence of the extra domains we have described in this work (data not shown). Despite Mommsen identified 3 different forms of aryl aminopeptidase he stated that these enzymes are quantitatively less important. In our mass spectrometry approach none aminopeptidase was identified in the DJ. In both works dipeptidase was not found secreted in the DJ.

More recent studies identified collagenase- and zinc metallopeptidase-like activities in the DJ and MMG of spiders. Collagenase is a metallopeptidase from M10A subfamily and such type of protein was identified in the present work only at the mRNA level (data not shown). In fact, the spider digestive enzymes named

collagenase in literature was merely a hydrolase which activity was tested using collagen as substrate without further characterization (50-52). Kavanagh and Tillinghast (49) identified 4 zinc metallopeptidase activities in the DJ of *Argiope aurantia* and 2 of them (A and B) had the capability of silk digestion. The molecular mass estimative for the “proteases A and B” were respectively 17 and 20 kDa. The four enzymes identified by these authors are likely astacins (Table 5.5). Atkinson and Wright (50) observed collagen hydrolysis using midgut extracts of 13 spiders. Since it was shown in chapter 2 that the cysteine cathepsins from the MMG are active in acidic conditions and the same is expected for cathepsin D, probably astacins and trypsins are involved in the collagen cleavage due to the alkaline assay conditions. Foradori and collaborators (51) observed cleavage of collagen, fribrinogen, fibrin, fibronectin and elastin using the DJ of *Argiope aurantia*. Only 65% inhibition was observed using EDTA which indicates that possibly astacins and trypsins are involved in these substrates hydrolysis. Moreover some of the trypsins identified in our study were similar to a variety of serine endopeptidases, including plasminogen (data not shown). We have choosen to use a more generic name based on the domains to avoid data misinterpretation without biochemical data. However based on the evidence provided by Foradori (6) it is reasonable to state that these trypsins with similarities to plasminogen likely are involved in avoiding hemolymph/blood prey coagulation during feeding. In another article (52) the same author observed activity over casein with molecular masses of 16, 18, 28, 35 and 120 kDa. The proteins of 16 and 18 kDa (respectively p16 and p18) had their amino-terminal sequences determined and an internal fragment of p16 was also obtained. These enzymes presented high identity (67% identity and 80% similarity) to the astacin 19a identified in the DJ of *Nephilengys cruentata*.

5.4.2.2 Carbohydrases, lipases and nucleases

Mommsen published three works analyzing the carbohydrases from the spiders *Tegenaria atrica* and *Cuppienius salei* DJ (113, 114, 137). The alpha-amylase activity found in these spider species (113, 114) was also identified by mass spectrometry in the present study and it composes 3.4% of the digestive enzymes in the DJ (Figure 5.4B). The molecular mass estimative of 58 kDa for this enzyme (113) is almost the same as the one of 59 kDa predicted in the present study (data not

shown). Only one chitinase was found in the DJ of *Tegenaria atrica* (113) and *Cuppienius salei* (137) and the same was observed by us (Table 5.2). These works reported that the chitinase is very active what could be explained by the high abundance of this enzyme which is 21.1% of the digestive enzymes (Figure 5.4B) and 8% of all the proteins in the DJ. The estimated mass of 48 kDa (137) is exactly the same of the predicted in our work (data not shown). The beta-glucuronidase activity measured by Mommsen (113) could be related to the beta-galactosidase that it was found in the DJ of *Nephilengys cruentata* since these are similar enzymes (140). Despite a detailed study was performed to differentiate beta-N-acetylglucosaminidase from beta-hexosaminidase activity (137), the author concluded that the former is responsible for *N,N´*-diacetylchitobiose digestion but only the latter was identified (Table 5.2) in the DJ samples from the present study. Inferring an enzymatic activity based solely in biochemical evidences can be trick and we identified 11 peptides of the beta-hexosaminidase in all digestive juice samples (data not shown) using a high resolution device. Hence at least in *Nephilengys cruentata* DJ, the more likely enzyme to cleave *N*,*N´*-diacetylchitobiose is beta-hexosaminidase. The alpha- and beta-glucosidase activities measured in *Tegenaria atrica* DJ could not be correlated to the identified enzymes by mass spectrometry since the only possibility left is an alpha-mannosidase (Table 5.2). Alpha-glucosidases were found only in the MMG of fasting spiders (Table 5.2) and none beta-glucosidase could be detected. This is also in accordance with the biochemical data obtained in our group (Genta et al, manuscript in preparation).

The lipases found in the DJ of the spider *Nephilengys cruentata* were basically TAGLs (Table 5.2). Such type of enzymes are probably related to the tributyrinase, lipase and esterase activities observed by Mommsen (138). In our group tributyrinase and esterase activities were observed in the DJ and MMG of the spider under investigation (141). TAGL 4 is present in both, the MMG and DJ, whereas TAGL 5 is only in the MMG. All 5 TAGLs are relatively similar with each other (data not shown) and probably the trybutyrinases observed by Mommsen can be correlated to the TAGLs identified in this work (Table 5.5). The DNase activity in the DJ of *Tegenaria atrica* is probably related to the deoxyribonuclease identified in the present study.

5.4.2.3 Digestive juice composition

Serine peptidase inhibitors in the spider DJ were previously reported in the literature (52, 142). Tugmon and Tillinghast (142) used the *Argiope aurantia* DJ to inhibit the peptidases from grasshopper gut extracts which is mainly constituted of serine peptidases. Subsequently, these inhibitors were further characterized with 10 different molecules identified ranging from 15-32 kDa (52). Peptidase inhibitors were identified in *Nephilengys cruentata* DJ. There is one L-cystatin, one serpin B3, 2 Kunitz-like and one alpha-1 inhibitor3 (Table 5.4). In our group, serine endopeptidase inhibitors were also identified by biochemical assays (Lopes et al., unpublished) which as the ones from *Argiope aurantia* are resistant to boiling (52). In a pooled fraction of partially purified inhibitory activity from MMG the 2 Kunitz-like molecules above cited plus 7 ctenitoxins were identified. The Kunitz-like inhibitors are more clearly involved in serine peptidase inhibition and isoforms 1 and 2 could be among the inhibitors identified by Foradori et al. (5) since the molecular mass of the isoform one is 36 kDa, similar to those found by Foradori. The isoform 2 is only a sequence fragment.

We also identified in the *Nephilengys cruentata* MMG pool with serine peptidase inhibitory activity 7 ctenitoxins from which 5 are in table 5.4 and 2 are not described in this table. They are 6 U24-ctenitoxin-Pn1a (four from table 5.4) and one U9-ctenitoxin-Pr1a. We will focus our analysis in the ctenitoxins in table 5.4. The biological function of these ctenotoxins haven´t still being elucidated but it is possible that they are cysteine peptidase inhibitors due to the presence of 2 thyroglobulin domains. Only the most abundant U24-ctenitoxin-Pn1a had low e-values (10^{-56}) with the original molecule and the thyroglobulin domains. The other 3 molecules had e-values only ranging from 10^{-6}-10^{-9}. However, all these ctenitoxins that we had identified present 6 conserved cysteine residues which could be involved in disulfide bridge formation. This characteristic is common of serine peptidase inhibitors and these molecules could be the thermal resistant serine peptidase inhibitors observed. Such inhibitory activity was already characterized in *Nephilengys cruentata* by our group and also in other spiders by other authors since their molecular mass (range of 16-17 kDa) is also in accordance with our biochemical data (5). The L-cystatin presence is the first clear evidence of a cysteine peptidase inhibitor in the DJ of a spider. The large amount of carboxypeptidase B 1 in the DJ (Figure 5.4B) may have as an important function the digestion of endopeptidase inhibitors from the prey.

A peptide isomerase was identified in the DJ of *Nephilengys cruentata* (Table 5.4). This is the first evidence of this enzyme in a digestive secretion. This isomerase was first isolated from the venom of the spider *Agelenopsis aperta* and it is involved in the conversion of the L-serine 46 to a D-serine (143) from omega-agatoxin-TK. Such characteristic may be important to avoid degradation by prey´s carboxypeptidases since this serine is located 3 amino acids before the C-terminal. Despite is too early to do a more deep speculation this same kind of role should be expected for this enzyme presence in the DJ. Some peptidase inhibitor or neurotoxin could be the substrate for this isomerase to avoid that prey´s carboxypeptidases digest these molecule(s). The less abundant U24-ctenitoxin-Pn1a (Table 5.4) has a serine residue 3 positions before the C-terminal (not shown). Another molecule usually associated to venom is a cysteine-rich secretory protein (CRISP) (144) with resemblance to allergen 5 (Table 5.4). Despite CRISP, ctenitoxins and the peptide isomerase are usually associated to venom glands (143, 144) we found these proteins at the mRNA level in the MMG tissue and at the protein level in the MMG and DJ. Some skepticism about these molecules being transcribed and translated in the MMG is expected. Nevertheless, some remarks will be appointed evidencing that probably this is not a contamination. The peptide isomerase, CRISP and ctenitoxins had their mRNA sequenced in fasting animals. Even if one assumes a holocrine venom secretion which could contain mRNA that was contaminating MMG samples and be further sequenced in the MMG, the fact that these sequences were found in fasting samples discard this hypothesis due to the abscense of venom secretion in these conditions. A venom contamination in the DJ due to the electrical stimulation method for sample obtaining is reasonable. However all the previous cited proteins were also identified in the DJ collected only with mechanical stimulation in the opisthosoma. Hence it seems very likely that these proteins are really original from the MMG tissue. This conclusion leads to an unavoidable consequence which is the fact that the venom toxins sequences were originally duplicated from the MMG proteins. It makes plenty of sense that evolutionarily the emergence of toxins is related to the digestive organ and their earlier presence in the DJ gave advantages in prey immobilization, blocking prey peptidases due to inhibition and avoiding some toxins digestion by the presence of D-amino acids. The primary venom gland probably had these genes duplicated that thereafter presented a paralog evolution in this gland.

Leucine-rich repeat-containing molecules were abundant in the DJ (Table 5.4). This domain is known due to the formation of an alpha/beta horseshoe fold and is associated with protein-protein interactions (145). Some of the molecules with this domain presented high similarities with SLIT proteins which act as guidance molecules (146). It is possible that the proteins in the digestive juice containing this domain are related to the direction of the proteins (digestive enzymes and meal) to the digestive cells. Peritrophin was identified in fasting animals (0.55‰) and in the DJ (0.31‰). It was observed in spiders that the peritrophic membrane was surrounding the fecal pellets (19) and also in mites (16). The fact that it was found in the DJ remains unclear but finding it in only in fasting and not in fed animals agrees with the literature information about fecal involucre.

Cathepsin D was identified in the DJ only in fasting animals and the same was observed for V-type proton ATPase subunits A and B. Cathepsin D due to its acidic nature and usually not being observed as a digestive secreted protein was not expected in the DJ. Our hypothesis is that this protein as well as V-ATPase presence in the DJ only in fasting animals is due to the extrusion of the digestive vacuoles into the lumen (21). Some proteins from intracellular organelles and even from nucleus can be observed in the DJ based on the GO terms (Figure 5.2F). Even though a cellular regeneration was never observed neither in scorpions nor in spiders (21, 28) it is likely this should happen. The biological process pie chart (Figure 5.2D) indicate some regeneration or differentiation since there are proteins related to cell cycle, cell death, cell differentiation, cell proliferation, cell recognition and cell-cell signaling. This pie chart and also figures 5.2E and 5.2F show that the DJ may have many other functions rather than digestion.

Cathepsins L (0.3%) and B (0.1%) despite low abundant in the DJ may have a secondary role in the extracellular digestion. Differently from cathepsin D it is likely that these enzymes are really secreted once they were identified in four out of five samples (data not shown). These cathepsins may be useful to extracellular digestion if, due to the prey tissue nature, a more accentuated acidification occurs. It was already shown that cathepsin L display collagenolytic activity (129, 130). As showed in chapter 2 the cathepsins L from the MMG have an acidic optimum pH and were not stable in neutral pH 7 (Figure 4.4). If cathepsin L2, which was found in both MMG and DJ (Table 5.2), was stable in neutral to alkaline pHs this should be observed in the MMG experiments since this enzyme is 35% of the cathepsins L from the MMG

(Figure 5.4D). Also the zymogen of cathepsin L1 is stable to neutral pHs (Figure 2.7) and in the digestive juice cathepsin L2 seems to be present as zymogen since the propeptide region was sequenced (data not shown). The neutral inference of the digestive juice is based on the data obtained by Mommsen (48).

5.4.3 The digestive process in the spider Nephilengys cruentata

5.4.3.1 Quantitative and qualitative remarks

Data on table 5.2 and figure 5.3B show an important feature about the physiology of digestion in the spider *Nephilengys cruentata*: the digestive juice share more enzymes with the MMG of fasting animals rather than the fed ones. Moreover, the enzymes identified only in the DJ of fasting animals are almost the same from table 5.2 with few exceptions (basically astacins 2 and 25 are not present while astacin 30 was found only in this condition). These results molecularly confirm the previous observation that after 1 hour feeding the secretory granules are being rebuilt and in one day they are completely reconstructed (21), hence both MMG tissue and DJ are prepared prior to the next predation event.

In the quantitative proteomics analysis the comparison of fasting and fed animals MMG composition as well as the DJ showed some important aspects about the spider physiology. Table 5.6 shows the NSC for the most important proteins involved in the digestive process and also the ratio of the NSC percentages between different samples. As observed in scorpions (chapter 3) there is a shift in some protein abundances in fed animals. By the data obtained cysteine peptidases have an important role in intracellular digestion. Despite cathepsins B1a and L2 are secreted, they are much more abundant in the MMG and the same was observed by the activity assays (chapter 4). Legumain 1b is not secreted and together with cathepsin L1 and 2 they are up regulated in fed animals (Table 5.6). Cathepsin D as above discussed is probably not in the secretory granules and acts intracellularly, its ratio between fed and fasting spiders seems to be constant. Opposite to cysteine peptidases, trypsins may be more important in the extracellular digestion. They are present only in the digestive juice or in the MMG of fasting animals (probably in the secretory vesicles) and could not be identified in the MMG of fed animals (Table 5.2). Chitotriosidase as trypsins was found in the DJ and in fasting animals (Table 5.2).

Table 5.6 - Main digestive enzymes and their normalized spectra counting (NSC) in each sample

	NSC percentage of the digestive enzymes			Ratios of the NSC percentages between different samples**		
	MMG					
Enzyme	**DJ**	**Fed**	**Fasting**	**Fed/Fasting**	**Fed/Dj**	**Dj/Fasting**
Legumain 1b	0	3.3	0.2	16.5	*	*
Cathepsin L-like cysteine peptidase 1	0	15.4	7.8	2.0	*	*
Dipeptidyl peptidase 3	0	2.7	4.9	0.6	*	*
Astacin-like metallopeptidase 26	0	3.3	1.4	2.4	*	*
Astacin-like metallopeptidase 19a	2.8	0	0.3	*	*	9.3
Astacin-like metallopeptidase 9	1.3	0	0.2	*	*	6.5
CUB domain-containing trypsin-like serine peptidase 4	1.9	0	0.3	*	*	6.3
Alpha-amylase	3.4	0	0.7	*	*	4.9
Astacin-like metallopeptidase 22	0.6	0	0.2	*	*	3
Chitotriosidase	21.1	0	24.5	*	*	0.9
Astacin-like metallopeptidase 30	1.2	25.4	3.1	8.2	21.2	0.3
Cathepsin L-like cysteine peptidase 2	0.3	13.4	5.9	2.3	44.7	0.1
Cathepsin B 1a	0.1	3.3	1.8	1.8	33.0	0.1
Astacin-like metallopeptidase 2	0.2	4.1	2.9	1.4	20.5	0.1
Cathepsin D-like aspartic peptidase	0.1	1.0	1.2	0.8	10.0	0.1
Astacin-like metallopeptidase 28	2.2	3.3	3.0	1.1	1.5	0.7
Astacin-like metallopeptidase 1b	6.8	4.0	3.9	1.0	0.6	1.7
Astacin-like metallopeptidase 46	5.6	0.4	7.2	0.1	0.1	0.8
Astacin-like metallopeptidase 11	3.2	0.2	1.1	0.2	0.1	2.9
Carboxypeptidase B 1	9.3	0.4	2.8	0.1	0.04	3.3
Astacin-like metallopeptidase 21	4.3	0.1	1.1	0.1	0.02	3.9

Notes: *Calculation not done due to the protein absence in one sample
**Due to the lack of a p-value to be considered as a significant up regulation the ratio should be ≥ 2 while for down regulation ≤ 0.5

Curiously, the ratio DJ/Fasting MMG is almost 1 and this enzyme was not identified in fed animals (Table 5.6). The same identification pattern was observed for alpha-amylase (Table 5.2).

Astacins 26 and 30 had a fold change of respectively 2.4 and 21.2 in the MMG after feeding (Table 5.6). It is possible that these molecules can present a digestive function also intracellularly due to this increase and because astacin 26 was found only in the MMG (Table 5.2). As above discussed the activity that Mommsen named “protease” (48) is likely a mixture of astacins (Table 5.5). The effect of pH curve in

this work using casein as substrate showed a relative activity of 50% in pH 5.5 which is the pH optimum of the intracellular cathepsin L 1 from *Nephilengys cruentata* (Figure 4.7). Carboxypeptidase B1, astacin 21 and astacin 46 decrease 10 times from fed to fasting animals (Table 5.6), which is probably related to the fact that they are more important extracellularly and astacin 11 decreases 5 times for the same reason. In conclusion, the enzymes involved in intracellular digestion usually showed a fold change equal or bigger than 2 in the ratios fed/fasting and also fed/DJ. In contrast to that the ones involved in extracellular digestion displayed inverted ratios or were not even detected in fed animals (Tables 5.2 and 5.6). The fact that exopeptidases were most abundant (number of different proteins) in fasting animals (Table 5.2) rather than the fed ones or in the DJ samples points out to the fact that these proteins will act after the initial phase of digestion, which was not measured in this work.

In the differential expression study few enzymes and proteins related to the digestive process were found differentially expressed. In the comparison between 9 and 1 hour fed spiders all the digestive enzymes from table 5.3 were not found by mass spectrometry, so they may be involved in a later phase of digestion. When fed animals (9 and 1 hour) expression patterns are compared with the fasting ones basically all the proteins up regulated are related to the extracellular digestion (Tables 5.3 and 5.6), with the exception of astacins 33 and 42 that were not identified in the proteomics experiment. These results are in accordance with the observation that after 1 hour the secretory granules are being resynthesized and are mature only after 1 day (21). In order to obtain more details about it the use of quantitative polymerase chain reaction with some chosen genes and more periods of time and protein quantification in parallel will help to elucidate the gene expression process.

5.4.3.2 General picture

Based on the results obtained and literature data the general mechanism of digestion in the spider *Nephilengys cruentata* is depicted as follows. The digestive juice and MMG already contain the digestive enzymes needed for digestion prior to feeding (Table 5.2). After prey capture, the prepared DJ already can be regurgitated and at the same time the secretory vesicles will be discharged into the lumen (21). Astacins and trypsins are the main endopeptidases involved in prey liquefying

(Figure 5.4B) and cysteine cathepsins may be required in cases of a more acidic extracellular digestion. Astacins probably act as multi associated molecules (the same or different isoforms) which would increase the efficiency of digestion. Such association was already observed for spider astacins (52) and also from other arthropods (147). Carboxypeptidase B1 is quite abundant (Figure 5.4B) and has a possible important role in the digestion of inhibitors from the prey. Chitotriosidase is the most representative protein in the DJ and this is likely related to the fact that in spite of vertebrates also be part of a spider diet (45, 46) arthropods are the most common prey, thus the exoskeleton is the first barrier to be crossed in the EOD. The next step in chitin digestion will be performed by beta-hexosaminidase (this study) intra and extracellularly (Table 5.2) or by beta-N-acetylglucosaminidase (137). The alpha-amylase presence in the DJ is probably related to glycogen digestion (10) and triacylglycerol lipases will start the TAG digestion extracellularly. After internalization of the partially digested food other enzymes will become more important. The initial protein intracellular digestion will be accomplished by cathepsins L 1, 2, 4 and 8, cathepsin D 1, cathepsin B 1a, legumain 1b and maybe astacins 26 and 30. Final protein digestion will be performed by dipeptidyl peptidase 1, 2 and 3, serine carboxypeptidase CPVL, alpha-aspartyl dipeptidase, carboxypeptidase B 1 and 2, glutamyl aminopeptidase and cathepsin B 1a. For the carbohydrase intracellular digestion possible enzymes involved are: alpha-L-fucosidase 1 and 2, maltase-glucoamylase, alpha-glucosidase, alpha and beta-galactosidases, alpha and beta-mannosidases. Lipid intracellular digestion may be done by TAGL and phosholipases (Table 5.2).

5.5 Conclusions

This work was the first one to do a combination of next generation sequencing and shotgun proteomics in order to qualitatively and quantitatively study the molecular physiology of digestion in the spider *Nephilengys cruentata*. In summary astacins, trypsins and carboxypeptidase B are the main enzymes responsible for extracellular protein digestion while cathepsins L, D and B and legumain will act digesting the food inside the cells. A shift in the midgut composition is observed in the MMG of fasting and fed spiders at the protein level. Proteins present in fasting animals are mainly secreted proteins which constitute the secretory granules and

digestive juice whereas proteins identified in fed spiders are more important to the intracellular digestion. Peptidase inhibitors such as kunitz, serpin and cystatin are present in the digestive juice. Ctenitoxins and peptide isomerase, so far found only in the venom glands, were identified in the digestive juice secretion, likely being paralogs ancestors of the known venom sequences. Furthermore, many different signaling molecules were sequenced in this secretion, indicating a more complex function than only digestion. The sporadic biochemical characterization of this tissue by other authors could in general be correlated to the enzymes identified in the present study. The presence of unidentified proteins or with unknown function as the ctenitoxins in the digestive juice opens the door for many future researches about the role of these molecules in spider physiology and evolution. Also, these unknown and known new molecules discovered are a huge source of natural active biomolecules that could be tested in a range of uses from industrial biotechnology to diseases treatment.

CHAPTER 6 – MOLECULAR PHYLOGENY AND CONCLUDING REMARKS

6.1 Introduction

In the previous chapters of thfis thesis it was reported the mechanism of digestion in spiders and scorpions using biochemical, transcriptomical and proteomical approaches. A large variety of digestive enzymes mainly constituted of peptidases has been shown to be used for prey digestion. According to the results obtained for the spider *Nephilengys cruentata* it is clear that astacins have an important extracellular role whereas cysteine peptidases will act mainly intracellularly (Chapters 4 and 5) and similar features were observed in the scorpion *Tityus serrulatus* (Chapters 2 and 3). Among the cysteine peptidases, cathepsin L presented the largest number of isoforms.

Astacins (E.C 3.4.24.21) are metallopeptidases belonging to the M12 family with a zinc-binding motif HEXXHXXGXXH that has a catalytic role (71). It is synthesized as a prepropeptide with the pro-segment located at the active site in an inverse orientation to the substrate and bound to the zinc ion (148). Astacin has been shown to be involved in a variety of biological process (149) including food protein digestion (150, 151). Indeed four astacins were identified in the digestive juice of the spider *Argiope aurantia* (52), in the present thesis using a combination of high throughput techniques it was shown that this number is much larger in the spider *Nephilengys cruentata* with 26 astacins identified by mass spectrometry in the digestive juice (Table 5.2).

Another abundant enzyme that showed to be important for digestion in *Nephilengys cruentata* and *Tityus serrulatus* is cathepsin L (Chapters 2, 3, 4 and 5). Until now none study related this cysteine endopeptidase with digestion in scorpions and spiders. The only arachnids known by the use of this enzyme for digestion are ticks (58, 66). Cathepsin L (E.C 3.4.22.15) belongs to the subfamily C1A clan CA and, as other enzymes from this clan, has a catalytic dyad constituted by a Cys^{25} and a His^{159} (papain numbering) (71). Cathepsin L has a preference for hydrophobic residues in P2 position (152) and is synthesized as a prepropeptide that can be activated under acidic conditions (135).

Before the present study, there wasn´t any complete sequence of astacin or cathepsin L originating from the MMG of scorpions and spiders available in the databases. Since these enzymes have been shown as the most important peptidases for intra and extracellular digestion of proteins, in this chapter it will be presented a molecular phylogeny analysis of these enzymes. Furthermore, the concluding remarks will also be discussed along the chapter.

6.2 Materials and methods

The molecular phylogenetic analyses were obtained using the astacin and cathepsin L sequences obtained through next generation sequencing and *in silico* translated as previously described in chapters 3 and 5. The sequences from the scorpion *Tityus serrulatus* have their accession number listed on table 2.2. Sequences from other organisms were retrieved from the Uniprot database. The exception for that are the sequences from *Neosadocus sp*. which were obtained by our group using the same methodology as for the other arachnids (Fuzita et al., unpublished).

6.2.2 Phylogenetic analysis

The gene phylogenies were performed separately for astacin and cathepsin L. All complete sequences transcriptomically identified were used, to these analysis with the only exceptions for some incomplete sequences that has beens also identified by mass spectrometry and that still have a sufficient size which allowed a tree construction (astacin 5b and cathepsin L1 from *Tityus serrulatus*). Firstly, the sequences were aligned using the software clustalW (124) with default parameters. Secondly the multiple sequence alignment was used for the construction of phylogenetic trees using maximum likelihood or neighbor-joining algorithms in the software MEGA 5 (85). A bootstrap analysis was conducted in the same software using 500 replicates for the former and 10,000 for the latter method (83).

6.3 Results

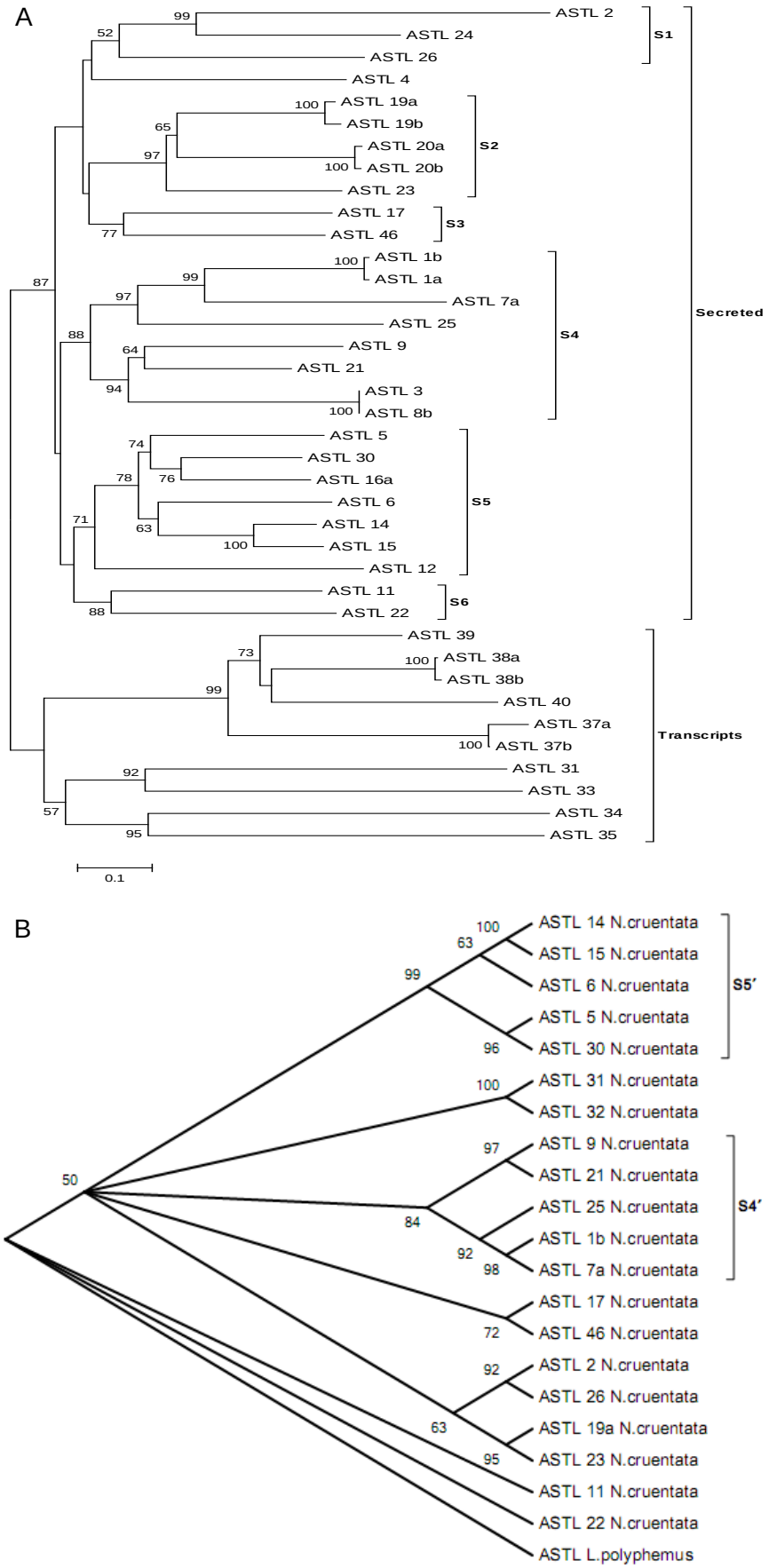
6.3.1 Astacin

In both astacin (ASTL) and cathepsin L (CTSL) phylogenetic analysis the trees were constructed using both maximum likelihood and neighbor-joining algorithms. Although it is known that the maximum likelihood is better in reconstructing the phylogeny (153) the neighbor-joining has also been proved as a good and faster method for this purpose (154). The neighbor-joining was chosen based on the 1) bootstrap values that were higher, 2) for being much less time-consuming, allowing the use of a 10,000 bootstrap value and 3) the topologies obtained were, in general, similar to the ones generated by the maximum-likelihood algorithm (data not shown).

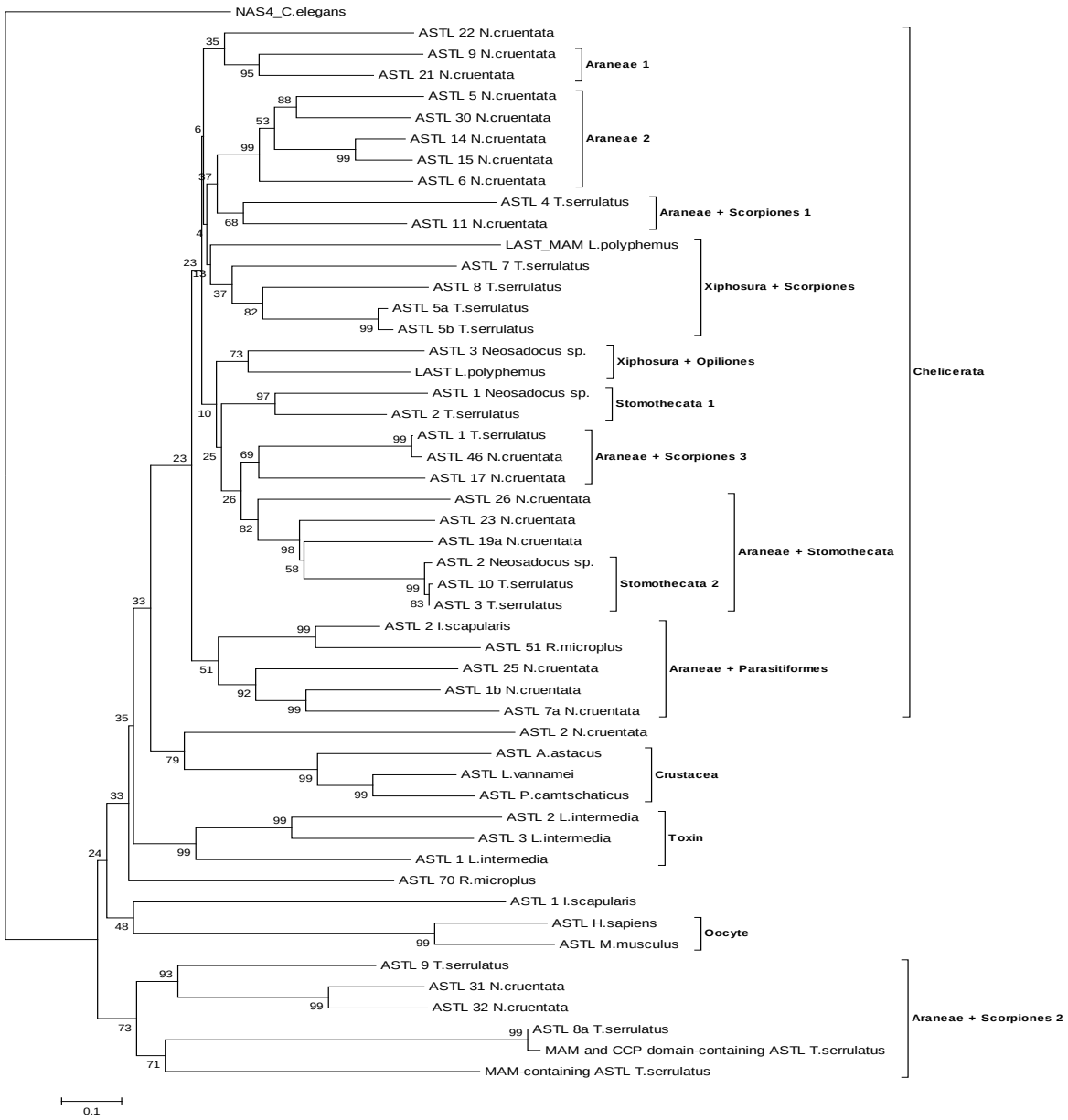
In order to investigate the molecular phylogeny of the astacin gene, firstly only the complete astacin sequences of *Nephilengys cruentata* were used (figure 6.1A) which totaled 46 different sequences. Twenty eight could be confirmed by mass spectrometry and are in the MMG tissue in both, fasting and fed spiders, as well as in the DJ (Table 5.2). Thirty eight sequences were used for the tree construction. The group named transcripts was found only as mRNA and the protein presence could not be confirmed by mass spectrometry, with the exception of astacin 31 (Table 5.2). In the cluster that was considered as the secreted group astacin 26 was found only in the MMG. Astacins 3, 12, 04 and 24 were not identified by mass spectrometry. There are 6 subgroups in the secreted cluster statistically supported named S1-6. Subsequently, only the complete astacins from *Nephilengys cruentata* identified by mass spectrometry were used in order to obtain more functional results, using the *Limulus polyphemus* astacin (LAST) as the tree root (Figure 6.1B). Also the tree was condensed with a 50% bootstrap cut-off for a better understanding of how these enzymes evolved in this single species. Two polytomies are observed in the tree (figure 6.1B). The first one is related to astacins 11 and 22 with LAST, which is an indicative that these 2 enzymes are divergent molecules in relation to the other astacins. The other polytomy split in five branches which are better resolved after that. One of the branches (S5´) is similar to S5 and the S4´ to S4 (Figure 6.1B and 6.1A).

Using only astacin sequences from *Nephilengys cruentata* identified in the proteome and also from other taxa, the rooted tree obtained is shown in figure 6.1C. The digestive astacin NAS4 from *Caernahobidtis elegans* was used as outgroup. A group of astacins named Chelicerata contain sequences from all taxa available in the databases, which includes Parasitiformes and Xiphosura summed with the

Figure 6.1 - Evolutionary relationships of the astacin gene



C



The Evolutionary history was inferred using the Neighbor-Joining method (82). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (10000 replicates) are shown next to the branches (83). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (84) and are in the units of the number of amino acid substitutions per site. Evolutionary analyses were conducted in MEGA5 (85). A) Unrooted tree with *Nephilengys cruentata* astacin complete sequences. B) Rooted tree using only *Nephilengys cruentata* complete astacin sequences identified by mass spectrometry. C) Rooted tree using different taxa sequences. The protein sequences were obtained in Uniprot database and the accession numbers are as follows: *Paralithodes camtschaticus*, AF492483_1, *Ripicephalus microplus* ASTL 70, gi|49558770, ASTL 51 gi|82845051; *Mus musculus*, sp|Q6HA09; *Homo sapiens*, sp|Q6HA08; *Litopenaeus vannamei*, tr|Q20AS7|Q20AS7; *Astacus astacus* sp|P07584, *Limulus polyphemus*, tr|B4F319|B4F319; LAST_MAM *Limulus polyphemus* tr|B4F320|B4F320; *Ixodes scapularis* tr|B7PAE6|B7PAE6 and tr|B7P9W0|B7P9W0; *Loxosceles intermedia* tr|C9D7R3|C9D7R3, tr|C9D7R2|C9D7R2, sp|A0FKN6|VMPA.

sequences from the present work (Araneae, Scorpiones and Opiliones). Crustacea is a sister group of Chelicerata and has the astacin 2 from *Nephilengys cruentata* as a related protein (Figure 6.1C). The astacin toxins from *Loxosceles intermedia* form a sister group with Arachnida + Crustacea and it was named toxin. The oocyte astacins from mammals are clustered in a separated group. The cluster Araneae + Scorpiones 2 contains MAM domain-containing astacins from *Tityus serrulatus* plus other astacins from the same animal and from *Nephilengys cruentata* without such domain. Curiously, the MAM domain-containing astacin from *Limulus polyphemus* (LAST_MAM) which has the same domain in the same position as the one from *Tityus serrulatus* it is not in this group and is closely related to other astacins (5, 7 and 8) of the same scorpion (Figure 6.1B). Such pattern was also observed when maximum likelihood was used (data not shown). Inside the large group Chelicerata there are some clusters well supported by the bootstrap composed exclusively of Araneae sequences and also this taxa form groups with Scorpiones, Stomothecata (Scorpiones + Opiliones) and Parasitiformes. Xiphosura groups with Scorpiones and Opiliones were observed but neither with Araneae nor Parasitiformes (Figure 6.1C).

6.3.2 Cathepsin L

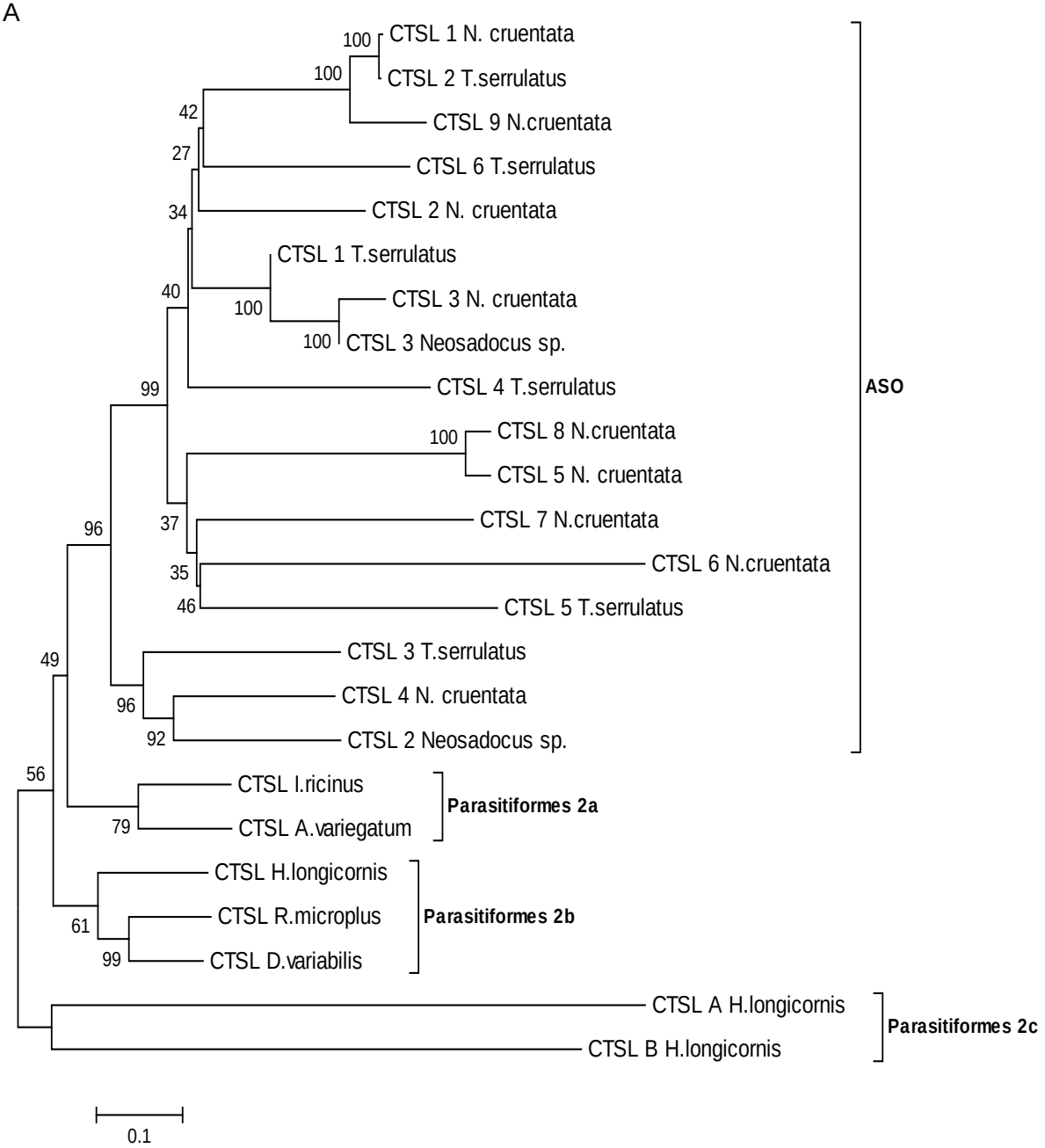
Firstly the tree was constructed using only the arachnid sequences. Figure 6.2A displays the topology obtained in which 2 main groups (1 and 2) can be observed. A clear division was observed between the Acari (here represented only by Parasitiformes) and the other group formed by ASO (Araneae, Scorpiones and Opiliones). ASO is highly supported with a bootstrap value of 96%. Some internal clusters also show high statistical support with sequences from 2 or 3 species. Curiously, the Parasitiformes did not form a monophyletic group, displaying a paraphyletic pattern, which is composed by groups 2a, b and c (Figure 6.2A). If the astacins formed some clusters only with Stomothecata (Figure 6.1C) the same was not observed for cathepsin L (Figure 6.2A and 6.2B). The CTSL 2 and 3 from *Neosadocus sp*. are respectively closest to *Nephilengys cruentata* CTSL 4 and 3 (Figure 6.2A). Nevertheless, all other subgroups in ASO are formed by scorpion and spider sequences only. The most quantitatively important CTSL are in this order 1 and 2 in both spider and scorpion (Chapters 4 and 5.). CTSL 1 (*Nephilengys cruentata*) and CTSL 2 (*Tityus serrulatus*) are very similar while the spider CTSL 2

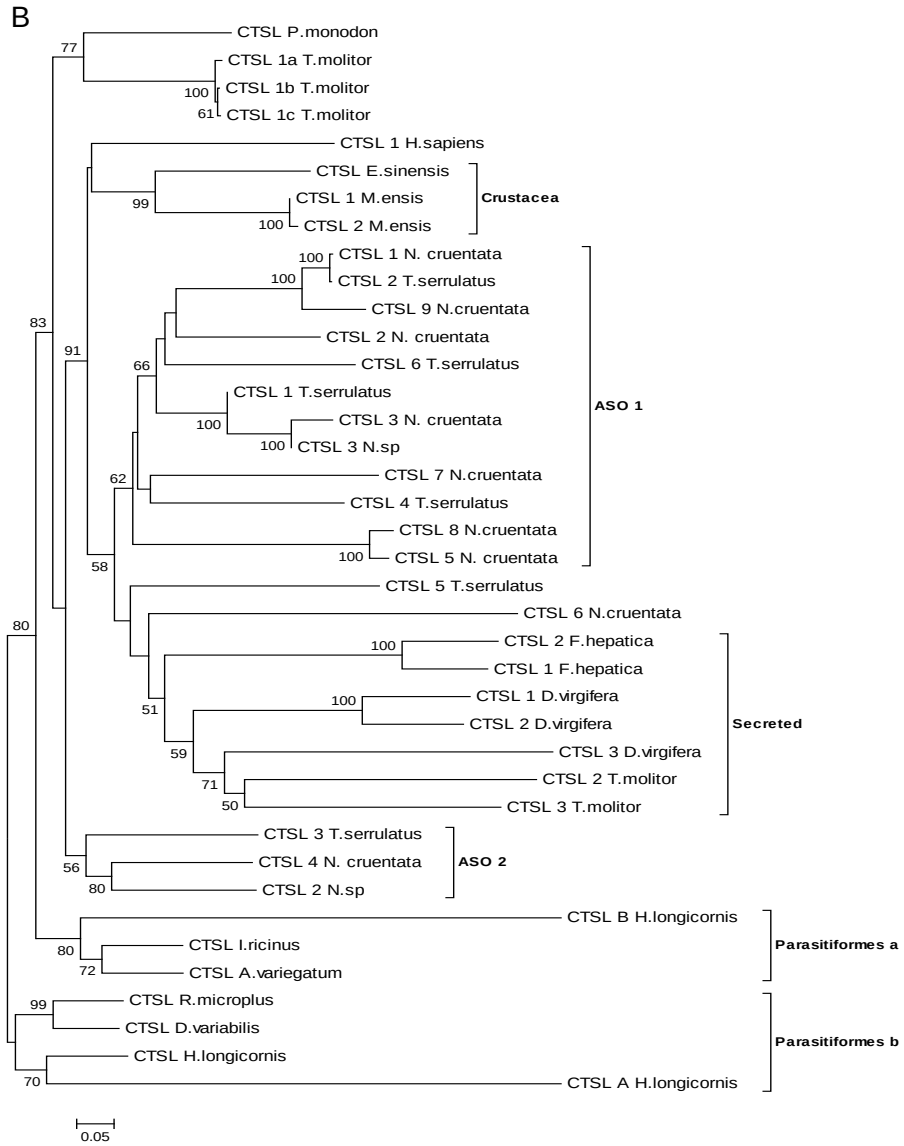
does not have high similarity with any other CTSL. *Tityus serrulatus* CTSL 1 showed high similarities with CTSL 3 from both spider and harvestmen (Figure 6.2A).

The Parasitiformes sequences did not group with other arachnid sequences even in the presence of other arthropods and human sequences forming what was named Parasitiformes a and b (Figure 6.2B). Two ASO groups can be observed with the number 1 being the larger one. Inside this group the CTSL 1 from *Tityus serrulatus* and CTSL 3 from both *Nephilengys cruentata* and *Neosadocus sp*. are highly similar. ASO 2 also has sequences from these three Arachnida groups.

Figure 6.2 - Evolutionary relationships of the cathepsin L gene

A





The evolutionary history was inferred using the Neighbor-Joining method (82). The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (10000 replicates) are shown next to the branches (83). Branches corresponding to partitions reproduced in less than 50% bootstrap replicates are collapsed. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method (84) and are in the units of the number of amino acid substitutions per site. Evolutionary analyses were conducted in MEGA5 (85). A) Unrooted tree with only arachnids sequences. B) Unrooted tree using different taxa sequences. The protein sequences were obtained in Uniprot database and the accession numbers are as follows: *Rhipicephalus microplus*, tr|J9QJ79|J9QJ79; *Haemaphysalis longicornis*, gi|254674508|dbj|BAH86062.1|, tr|O96087|O96087, tr|O96086|O96086; *Ixodes ricinus*, tr|A4GTA6|A4GTA6; *Dermacentor variabilis*, tr|A7LJ78|A7LJ78; *Amblyomma variegatum*, tr|F0JA27|F0JA27; *Metapenaeus ensis*, tr|Q7Z0G8|Q7Z0G8, tr|Q7Z0G9|Q7Z0G9; *Penaeus monodon*, tr|A9U938|A9U938; *Fasciola hepatica*, tr|Q7JNQ8|Q7JNQ8, tr|Q7JNQ9|Q7JNQ9; *Diabrotica virgifera*, tr|Q70EW9|Q70EW9, tr|Q70EW8|Q70EW8, tr|Q70EX2|Q70EX2; *Homo sapiens*, sp|P07711|.

A cluster formed by the crustacean sequences with high bootstrap value is also present in figure 6.2B. A group named “secreted” is formed by enzymes that have been reported as secreted proteins such as the ones from *Fasciola hepatica* (155) and *Tenebrio molitor* (90, 131). Only the enzymes from *Diabrotica virgifera*

(156) had not their subcellular location studied but due to the phylogenetic proximity with *Tenebrio molitor* and also clustering together with the enzymes from *Fasciola hepatica* it seems likely that these proteins form a group of secreted cathepsin L . The secreted CTSL 2 from *Nephilengys cruentata* it is not inside this group. The lysosomal CTSL 1a, b and c from Tenebrio molitor cluster together with the *Penaeus monodon* and separated from all other groups.

6.4 Discussion

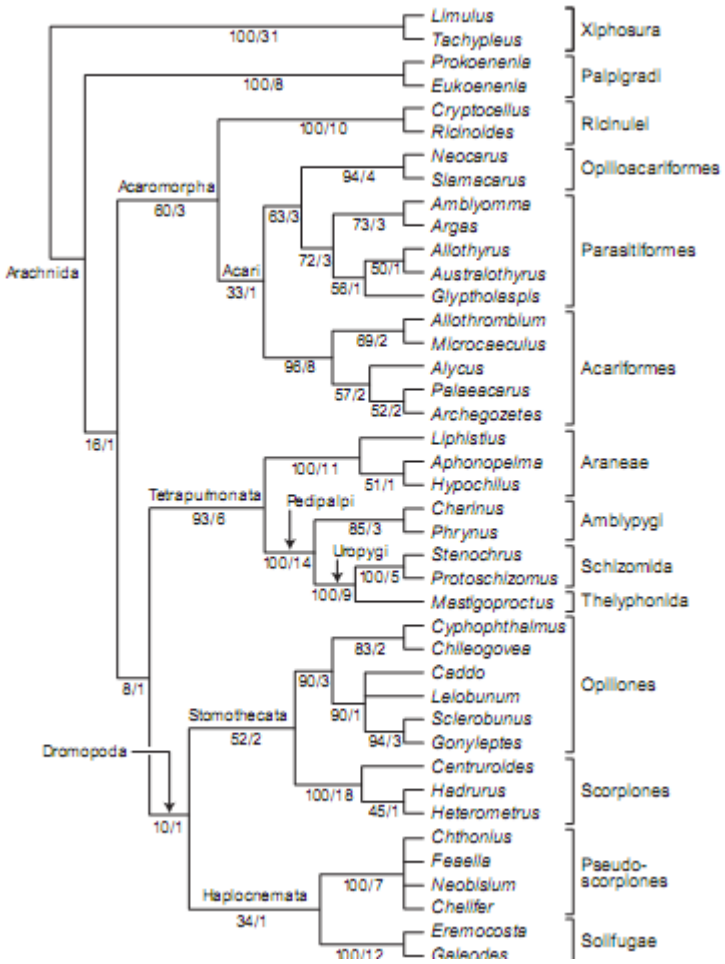
Before the individual discussion about the molecular phylogeny obtained using the astacin and cathepsin L sequences it must be clear that for both cases this is not an attempt of reconstructing the evolutionary history of Chelicerata, Arachnida or even Arthropoda. The phylogenies obtained by other authors will be used as the bases for the discussion that it will be presented and the cladograms chosen (Figure 6.3) for such comparisons are the ones from Shultz (157). The aim of doing a single-gene molecular phylogeny from a digestive enzyme is to understand the evolution of that protein in different organisms (for astacin also only in *Nephilengys cruentata*) and how this can be related to the digestive system functionality of each group according to their feeding habits. Because the feeding and digestive mechanisms are deeply linked with the evolutionary history of a group, the comparison with the phylogenies done by other authors is extremelly important in order to find how such mechanisms evolved. Furthermore, all other evidences obtained in the previous chapters of this thesis will be used to perform a contextual and functional analysis of the digestive process in Arachnida.

6.4.1 Astacin

In this work, to our knowledge, the largest dataset of different astacin sequences of a single organism and tissue-specific was obtained for the spider *Nephilengys cruentata*. Forty six astacin genes were identified expressed with 8 of them presenting isoforms leading to a total of 54 sequences. Twenty six astacins were found secreted in the DJ and until now none study identified such a variety of this enzyme used for meal digestion. *Caernohabditis elegans* have forty astacin

genes, but only five which belongs to the subgroup 1 are associated to digestion(158). In this subgroup, as for all astacins from *Nephilengys cruentata* and *Tityus serrulatus* (except for the MAM astacins), the enzymes are composed of a prepropeptide plus the astacin domain (158). NAS4 was found in the *Caernohabditis elegans* midgut after being secreted by the pharynx and was chosen to the phylogenetic tree construction being used as the root of the tree from figure 6.1C.

Figure 6.3 - Chelicerata morphological phylogeny from Shultz (157)



Although the molecular phylogeny from figure 6.1 still need more information like other arachnid orders and more sequences to each species, it is possible to make some inferences. In the analysis involving only the astacins from *Nephilengys cruentata* it can be observed politomies (Figure 6.1B). This is because the tree parameters were more stringent not accepting monophily supported by bootstrap values smaller than 50%. This is an indicative that at some point a fast gene duplication event happened in Aranae. Subsequently these genes evolved slowly and their ancestrality can be traced. In figure 6.3C the crustacean sequences clearly

separated from the arachnid ones as well as the toxin sequences from *Loxosceles intermedia* and the oocyte astacins from mammals. These results probably indicate that digestive and venom astacins evolved with different functions in an early stage. Astacin 2 from *Nephilengys cruentata* shared high identities with crustacean astacins and forms an external group statistically supported of the Crustacea group from figure 6.1C. This protein could be related to an ancestral astacin from Arthropoda. The S5´ and Araneae 2 groups from, respectively, figures 6.1B and 6.1C are exactly the same and they are related to the subgroup S5 from figure 6.1A. This group does not share similarities with molecules from other animals, thus it is likely exclusive from spiders. The two molecules from Araneae 1 (Figure 6.1C) are also exclusive to this order.

The LAST_MAM clusters with the *Tityus serrulatus* sequences 5a and b, 7 and 8 while LAST with ASTL 3 from *Neosadocus sp* (Figure 6.1C). Although the authors stated that LAST is more likely a neuronal enzyme rather than digestive, this protein was also found expressed in the proventriculus which is a strong indicative of a digestive role as well. It makes sense that the sequences from the basal clade Xiphosura group with Stomothecata, even separately, since these two clades are basal in Arachnida phylogeny (Figure 6.3). Stomothecata 1 and 2 share close-related sequences with the group 2 also sharing similarities with Araneae (Figure 6.1C). Araneae is the most derived Arachnida group in this study (Figure 6.3), what means that this is the only order that shares a common ancestor with all other taxa. This phylogenetic pattern can also be observed for the astacin sequences in figure 6.1C where Araneae clusters with Scorpiones (3 times), Stomothecata and Parasitiformes. The most frequent clustering with Scorpiones is probably due to the fact that: 1) there are more sequences of scorpions in this analysis rather than sequences from other arachnid species (except Araneae); 2) the *Neosadocus sp* (Opiliones) transcriptome is still under analysis so they may have more astacins and also the NGS was done for only one fed animal which leads to a minor identification and also more probability of incomplete sequences. Also *Neosadocus sp*. is an omnivorous species and harvestmen can ingest solid particles not liquefying the food prior to ingestion (17); 3) *Ixodes scapularis*, from the Parasitiformes clade, has its genome sequenced but in the Uniprot database there are only 2 complete astacins sequences. This low number of astacins in Parasitiformes is probably a secondary loss since the feeding habits of these animals do not have an extracellular food liquefying need as other

arachnids do have (61). Thus it seems plausible that animals which rely on extra-oral digestion share more similar sequences and moreover is more parsimonious assuming a loss in Parasitiformes in contrary to the digestive astacin presence as an homoplasy in all taxa. Even if in *Limulus polyphemus* the use of LAST as a digestive enzyme is secondary in relation to a neuronal function, this protein was already expressed in the digestive system of this basal animal (159), indicating that the Chelicerata ancestor also had this sequence. Furthermore, it is likely that *Limulus polyphemus* expresses other(s) astacins(s) since it is not clear how the cDNA library screening was performed and also this is not a technique of deep investigation like NGS.

Biochemically astacins in the digestive system were found in scorpions (Chapter 2), spiders (49, 52), harvestmen (Fuzita et al., unpublished) and in *Astacus astacus* (150). The astacin from *Astacus astacus* cleave substrates in a pH range from 6-8 while the activity of the *Argiope aurantia* was measured in pH 8.8 (52). The optimum pH of the astacins from *Tityus serrulatus* and *Nephilengys cruentata* is respectively 9.0 and 8.3 using casein-FITC as substrate.

Hence, the use of astacin-like metallopeptidases for digestion seems to be an ancestral condition in Chelicerata. The extra-oral digestion performed by Arachnida probably was the trigger for gene duplication events that led to the diversification of many different molecules observed in the digestive juice of the spider *Nephilengys cruentata*. Moreover, the predators under investigation presented a large number of astacin sequences that clustered together more frequently than with the other groups, indicating that the use of these metallo-enzymes may be more important for arachnid predators which depend on prey extra-oral digestion. We were able to identify only 2 astacins at the protein level from the basal arachnid *Tityus serrulatus.* However, it could be detected in the MMG secretion of *Nephilengys cruentata* 26 different molecules, evidencing that, during Arachnida evolution in land, the use of a large variety of peptidases (probably with different specificities) for external prey digestion was an important evolutionary event.

6.4.2 Cathepsin L

Until now cathepsin L activity in Chelicerata was measured only in the Parasitiformes group (56, 58, 66). This thesis provided the first biochemical,

proteomical and transcriptomical evidences in chapters 2, 3, 4 and 5 that cathepsin L is an important enzyme for the digestive process in scorpions and spiders, likely in the intracellular phase. Such aspect was expected since it is known that both spiders and scorpions do intracellular digestion after extra-orally liquefying the prey (21, 28). Cathepsin L is the most abundant endopeptidase in the MMG of fed spiders (37.9%) and scorpions (25.6%) (Chapters 4 and 5). The Xiphosura group also presents cathepsin L (CK086983.1) expressed in the “hepatopancreas” but only a small fragment was identified in *Carcinoscorpius rotundicauda* which was not used for the phylogenetic reconstruction.

The relationship between the arachnid CTSL from figure 6.2A shows that the Parasitiformes sequences are divergent from the ones in ASO group. Ticks are specialized in blood digestion and, although they have also kept the intracellular mechanism of food degradation from arachnids using the same enzyme types, this led to completely different selective pressures over this group. Even the omnivorous harvestmen *Neosadocus sp.,* which does not need to liquefy the prey prior to ingestion, kept cathepsins L similar with the predators ones (Figure 6.2A). The *Nephilengys cruentata* CTSL 2, which is believed to be secreted (Chapter 5), did not exhibit high identity values with the proteins in the secreted group. Thus, more data supporting this hypothesis will have to be obtained. The same is true for CTSL 2 from *Tityus serrulatus* which was predicted as an extracellular enzyme.

The tree obtained using other Arthropoda taxa and human cathepsin L 1 (Figure 6.2) showed that phylogenetic groups as Crustacea and Parasitiformes form separated clusters (although Parasitiformes splits in two and the *Penaeus monodon* sequence did not group with Crustacea and it seems to be more close to the lysosomal *Tenebrio molitor* CTSL 1). The group named secreted has a functional similarity but it also complains a subgroup of Insecta sequences. Human CTSL 1, which is lysosomal, clustered with the Crustacea sequences. In conclusion, this tree, despite still poorly resolved, showed phylogenetic relationships and 2 groups that could be functionality related (secreted and ASO). The former is related to secreted CTSL and the later to food intracellular digestion in digestive vesicles by non-hematophagous arachnids. The “truly” lysosomal enzymes, which means that they act in lysosome and not in big digestive vacuoles, did not clustered together and this can be due to the phylogenetic distance from insects to humans. Furthermore, the cathepsin L presence, as verified for astacin, is probably an ancestral condition in

Chelicerata since this sequence was also found in *Carcinoscorpius rotundicauda* “hepatopancreas”.

6.5 Conclusions

6.5.1 Chapter 6 conclusions

The astacin and cathepsin L phylogenetic analyses showed some important aspects of the feeding habits and evolutionary history of these Chelicerata sequences. Firstly, in both cases the molecular phylogeny obtained showed similar clustering to the ones based in morphological data. Secondly, in general, the Parasitiformes group showed sequences divergent to all other arachnids, except one grouping observed with the astacins from *Nephilengys cruentata*. Such divergent pattern is probably related to the selective pressures over ticks which specialized in blood digestion. Astacin is a very diversified and important group of enzymes used for extra-oral digestion in spiders and maybe in other arachnids. In ticks, astacins are much less abundant and the only remaining ones are closest related to the spider sequences, thus due to the ticks digestive mechanism which uses more an intracellular apparatus, a secondary loss of astacins was observed. In the cathepsin L case, despite of ticks have kept the intracellular digestion and the same classes of enzymes, the blood digestion is an evolutionary pressure which led to the selection of more divergent sequences in comparison to the other Arachnida taxa.

6.5.2 General concluding remarks

Along this work the molecular physiology of digestion of two predators arachnids was characterized. In chapter 2, a biochemical characterization of digestive enzyme was done using *Tityus serrulatus* MMG. Different endopeptidase substrates and inhibitors were used and by the analysis of the proteomics data these activities could be associated to the enzymes identified. Cysteine peptidase activity was abundant in this tissue and the presence of zymogen activated under acidic conditions could be demonstrated. Two cysteine peptidases were purified to apparent homogeneity with a molecular mass of 33kDa. The activity isolated pool that originated this purified samples contained cathepsin L1 and 2 and cathepsin F.

The purified enzymes showed to be competitive inhibited by the aspartic peptidase inhibitor pepstatin. Trypsin-like activity over Z-FR-MCA partially inhibited by benzamidine and totally calcium dependent as well as astacin-like activity over casein were observed. Both enzyme types were identified by mass spectrometry. Cathepsins D, B and legumain were identified by mass spectrometry.

Chapter 3 provided the entire analysis of next generation sequencing followed by proteomics analysis in *Tityus serrulatus* MMG. Transcriptomics analysis allowed the identification of 235 enzymes with a possible digestive role from which 43 were detected by mass spectrometry. In general, it was possible to identify cathepsins B, D, F and L, trypsin, legumain, chitinase, alpha-glucosidase, alpha-fucosidase, alpha-amylase, alpha-manosidase, lipase, phospholipase, dipeptidase, and carboxypeptidase. Most abundant enzymes were cathepsin L, chitinase, alpha-mannosidase, alpha-glucosidase, cathepsin D and dipeptidases. It seems that fasting animals contain the enzymes needed for the next predation event in fed scorpions a shift was observed and enzymes involved in intracellular digestion became more abundant. Since in this scorpion the digestive juice could not be collected, the enzymes subcellular location was inferred using prediction software. The extracellular digestion will be performed by endopeptidases such as astacins, CUB and LDL domains-containing trypsin and cathepsins L2 and F and carbohydrases as chitinase. The intracellular phase will be done by cathepsin L1, cathepsin D, legumain, cathepsin L 2, cathepsin F, alpha-glucosidase and alpha-mannosidase.

In chapter 4 the cysteine cathepsins from *Nephilengys cruentata* MMG were biochemically characterized. These enzymes are stable and can be activated in acidic pHs and losses the activity in the neutral-alkaline range. Two cathepsins L1 and 2 were expressed and obtained in the recombinant form but only cathepsin L1 presented activity. The same acidic characteristics presented by the native enzymes were observed but it was possible to show that the zymogen is stable in alkaline pHs. Cathepsin L 1 was not identified in the digestive juice thus it is probably a lysosomal enzyme.

In chapter 5 the possibility of using the combined high throughput techniques for analyzing the digestive juice and MMG provided important insights about the mechanism of digestion in spiders. It was shown that both MMG and digestive juice of the spider already contain the enzymes needed for the next predation event. Astacins were extremely abundant with 26 different molecules secreted followed by

trypsins which had 8 copies identified, quantitatively they respectively represented 37.7% and 17% of the digestive enzymes. Carboxypeptidase B 1 is probably involved in prey inhibitor digestion and chitotriosidase was the most abundant individual enzyme. In the MMG, cathepsins L showed to be important for digestion and they are up regulated at the protein level in fed animals. Other molecules were also identified in the digestive juice such as ctenitoxins and peptide isomerase. It seems probable that the first toxin-like molecules were developed in the MMG and posteriorly the same genes were requested in the primitive venom gland.

Chapter 6 presented a phylogenetic analysis using the astacins and cathepsins L identified in this work together with other taxa sequences from the databases. In some cases there was a phylogenetic clustering similar to the one obtained by morphological data. However the differences between the Parasitiformes sequences in contrast to the Chelicerata ones was clearly evidenced. Based on that it could be inferred two main hypothesis: The first is that the astacins are important in predator´s extra-oral digestion thus ticks probably lost some astacin sequences. The second hypothesis is that the intracellular cathepsin L, which seems to be important for all arachnids, is more divergent in Parasitiformes due to the selective pressures over this group for blood digestion.

6.5.3 Future perspectives

In this thesis a deep investigation was done in order to comprehend the poor studied mechanism of digestion in predator arachnids. In comparison to other arthropods such as crustaceans, insects and ticks, the digestive process in predator arachnids is still underexplored. This study provided many insightful aspects of the digestive mechanisms in spiders and scorpions and a large advance was obtained in the field, mainly related to the protein sequences expressed in the midgut and midgut glands and secreted in the digestive juice, their composition under different physiological conditions and pH of action of some endopeptidases. A single study about this process certainly is not enough to solve the puzzle of how predator arachnids efficiently combine extra and intracellular digestion in order to digest their prey, even using high throughput techniques combined. Nevertheless, if this puzzle is still far from being solved it is possible metaphorically to say that before the present work the box with the pieces was still closed. Now that we have opened the box and

put a big number of pieces on the table we will need further investigation to a better understanding of the molecular physiology of digestion in these two arachnids. Future studies should concern in at least 6 main points: 1) Studies about the RNA expression of the digestive enzymes using quantitative PCR in the MMG associated with biochemical assays and quantitative mass spectrometry in different periods of feeding, including some days after the extra-oral digestion completion. The expression pattern of selected genes in different tissues can also provide important information; 2) recombinant expression of selected digestive endopeptidases in order to understand their substrate specificities, mechanism of activation and biochemical characterization; 3) use the recombinant enzymes for antibody production with a subsequent immunohistochemical experiment for locating these enzymes during the digestive process. Also cytological studies should complain an observation of the acidification process in the digestive vesicles; 4) study the target of the toxins present in the MMG of spiders and scorpions that could lead to a future biomedical application; 5) investigate the composition of each individual midgut gland of the scorpion *Tityus serrulatus* regarding their composition and observe if the enzymes are equally distributed using quantitative mass spectrometry with stable isotopes; 6) study the subcellular composition of the MMG using differential ultracentrifugation and using quantitative mass spectrometry with stable isotopes in order to identify the enzymes present in the fractions. The items 5 and 6 are already were executed and are under analysis by our group.

REFERENCES*

1. Cohen AC. Extraoral digestion in predaceous terrestrial Arthropoda. Annual Review of Entomology. 1995;40:85-103.

2. Brusca RC, Brusca GJ. Invertebrates. 2nd ed. Sunderland: Sinauer Associates, Inc.; 2002. 895 p.

3. Cavalier-Smith T. Predation and eukaryote cell origins: A coevolutionary perspective. International Journal of Biochemistry & Cell Biology. 2009;41(2):307-22.

4. Pough FH, Janis CM, Heiser JB. A vida dos vertebrados. 4th ed. São Paulo: Atheneu Editora; 2008. 684 p.

5. Ludwig M, Alberti G. Peculiarities of arachnid midgut glands. Acta Zoologica Fennica. 1990:255-9.

6. Hu K-J, Leung P-C. Food digestion by cathepsin L and digestion-related rapid cell differentiation in shrimp hepatopancreas. Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology. 2007;146(1).

7. Resch-Sedlmeier G, Sedlmeier D. Release of digestive enzymes from the crustacean hepatopancreas: effect of vertebrate gastrointestinal hormones. Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology. 1999;123(2):187-92.

8. Chajec L, Rost-Roszkowska MM, Vilimova J, Sosinka A. Ultrastructure and regeneration of midgut epithelial cells in Lithobius forficatus (Chilopoda, Lithobiidae). Invertebrate Biology. 2012;131(2):119-32.

9. Fontanetti CS, Camargo-Mathias MI, Caetano FH. Apocrine secretion in the midgut of Plusioporus setiger (Brolemann, 1901) (Diplopoda, Spirostreptidae). Naturalia (Rio Claro). 2001;26:35-42.

10. Terra WR, Ferreira C. Biochemistry and molecular biology of digestion. In: Gilbert LI, editor. Insect Molecular Biology and Biochemistry. London: Academic Press; 2012. p. 355-418.

11. Grimaldi D, Engel MS. Evolution of the insects. 1st ed. Hong Kong: Cambridge University Press; 2005. 755 p.

12. Selden PA, Dunlop JA. Fossil taxa and relationships of Chelicerates. Arthropod Fossils and Phylogeny. 1998:303-31.

13. Coddington JA, Giribet G, Harvey MS, Prendini L, Walter DE. Arachnida. Assembling the Tree of Life. 2004:296-318.

14. Dunlop JA, Webster M. Fossil evidence, terrestrialization and arachnid phylogeny. Journal of Arachnology. 1999;27(1).

* According to the International Committee of Medical Journal Editors. Uniform requirements for manuscripts submitted to Biomedical Journal. Sample references [2011 July 15]. Available from: http://www.icmje.org

15. Akov S, Samish M, Galun R. Protease activity in female Ornithodoros tholozani ticks. Acta tropica. 1976;33(1):37-52.

16. Sobotnik J, Alberti G, Weyda F, Hubert J. Ultrastructure of the digestive tract in Acarus siro (Acari : Acaridida). Journal of Morphology. 2008;269(1):54-71.

17. Pinto-da-Rocha R, Machado G, Giribet G. Harvestmen: The Biology of Opiliones. Cambridge and London: Harvard University Press; 2007. 597 p.

18. Becker A, Peters W. The ultrastructure of the midgut and the formation of peritrophic membranes in a harvestman, phalangium-opilio (chelicerata, phalangida). Zoomorphology. 1985;105(5):326-32.

19. Vanderbo.O. Peritrophic membranes in Arachnida (Arthropoda). Nature. 1966;210(5037):751.

20. Ludwig M, Alberti G. Ultrastructure and function of the midgut of camel-spiders (Arachnida, Solifugae). Zoologischer Anzeiger. 1992;228(1-2):1-11.

21. Ludwig M, Alberti G. Digestion in spiders - histology and fine-structure of the midgut gland of Coelotes-terrestris (Agelenidae). Journal of Submicroscopic Cytology and Pathology. 1988;20(4):709-18.

22. Polis GA. The Biology of Scorpions. 1st ed. California: Stanford University Press; 1990. 233 p.

23. Brasil. Manual de Controle de Escorpiões. Brasília2009. p. 74.

24. Srivastava DS. Maxilary processes and mechanism of feeding in scorpions. Saugar University Journal. 1955;4:85-91.

25. Auber M. Sur les glandes gnathocoxales des scorpions. Zoologique de France Bulletin. 1960(85):67-77.

26. Sarin E. Uber die Fermente der Verdauungsorgano der Skorpione. Biochemische Zeitschrift. 1922;129:pp 359-66.

27. Farley RD. Scorpiones. In: Harrison FW, Foelix RF, editors. Microscopic Anatomy of Invertebrates. 8A. New York: Wiley; 1999. p. 117-222.

28. Goyffon M, Martoja R. Cytophysiological aspects of digestion and storage in the liver of a scorpion, Androctonus-australis (Arachnida). Cell and Tissue Research. 1983;228(3).

29. Weel PBv. "Hepatopancreas" ? Comaparative Biochemistry and Physiology. 1973;47A:1-9.

30. Pavlovsky EN, Zarin EJ. On the structure and ferments of the digestive organs of scorpions. Quarterly Journal of Microscopical Science. 1926;70(278):221-U23.

31. Alberti G, Storch V. the ultrastructure of the midgut glands of Arachnida (Acorpiones, Araneae, Acari) under different feeding conditions. Zoologischer Anzeiger. 1983;211(3-4):145-60.

32. Bardi JK, George CJ. Digestive glands of the scorpion - a physiological investigation. Journal of the university of Bombay. 1943:91-109.

33. Said EE. On the digestive enzymes of some terrestrial Arthropoda (Butus quinquestriatus H. E. and Scolopendra morsitans L). Proceedings of the Egiptian Academy of Sciences. 1958;13:55-75.

34. Louati H, Zouari N, Fendri A, Gargouri Y. Digestive amylase of a primitive animal, the scorpion: Purification and biochemical characterization. Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences. 2010;878(11-12):853-60.

35. Vijayalakshmi NR, Kurup PA. Alpha-amylase of the hepatopancreas of the scorpion, Heterometrus scaber. Indian Journal of Experimental Biology. 1969;7:266-7.

36. Vijayalakshmi NR, Kurup PA. Metabolism of glycosaminoglycans in Heterometrus-scaber. Indian Journal of Experimental Biology. 1976;14(1):10-3.

37. Zouari N, Miled N, Cherif S, Mejdoub H, Gargouri Y. Purification and characterization of a novel lipase from the digestive glands of a primitive animal: The scorpion. Biochimica Et Biophysica Acta-General Subjects. 2005;1726(1).

38. Zouari N, Bernadac A, Miled N, Rebai T, De Caro A, Rouis S, et al. Immunocytochemical localization of scorpion digestive lipase. Biochimica Et Biophysica Acta-General Subjects. 2006;1760(9):1386-92.

39. Zouari N, Sayari A, Miled N, Verger R, Gargouri Y. Scorpion digestive lipase: Kinetic study using monomolecular film technique. Colloids and Surfaces B-Biointerfaces. 2006;49(1).

40. Zouari N, Miled N, Rouis S, Gargouri Y. Scorpion digestive lipase: A member of a new invertebrate's lipase group presenting novel, characteristics. Biochimie. 2007;89(3):403-9.

41. Shultz JW. EVolutionary morphology and phylogeny of Arachnida. Cladistics-the International Journal of the Willi Hennig Society. 1990;6(1):1-38.

42. Selden PA. Fossil mesothele spiders. Nature. 1996;379(6565):498-9.

43. Kuntner M. A monograph of Nephilengys, the pantropical 'hermit spiders' (Araneae, Nephilidae, Nephilinae). Systematic Entomology. 2007;32(1):95-135.

44. Schuck-Paim C, Alonso WJ. Deciding where to settle: conspecific attraction and web site selection in the orb-web spider Nephilengys cruentata. Animal Behaviour. 2001;62:1007-12.

45. Peloso PL, Souza VPd. Pedration on Todirostrum cinerum (Tyrannidae) by the orb-web spider Nephilengys cruentata (Aranae, Nephilidae). Revista Brasileira de Ornitologia. 2007;15(3):461-3.

46. Nyffeler M, Knornschild M. Bat predation by spiders. PloS one. 2013;8(3):e58120.

47. Foelix R. The Biology of Spiders. 3rd ed. USA: Oxford University Press; 2010. 432 p.

48. Mommsen TP. Digestive enzymes of a spider (Tegenaria-atrica koch) .1. general remarks, digestion of proteins. Comparative Biochemistry and Physiology a-Physiology. 1978;60(4):365-70.

49. Kavanagh EJ, Tillinghast EK. The alkaline proteases of Argiope .2. fractionation of protease activity and isolation of a silk fibroin digesting protease. Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology. 1983;74(2):365-72.

50. Atkinson RK, Wright LG. The involvement of collagenase in the necrosis induced by the bites of some spiders. Comparative Biochemistry and Physiology C-Pharmacology Toxicology & Endocrinology. 1992;102(1):125-8.

51. Foradori MJ, Keil LM, Wells RE, Diem M, Tillinghast EK. An examination of the potential role of spider digestive proteases as a causative factor in spider bite necrosis. Comparative Biochemistry and Physiology C-Toxicology & Pharmacology. 2001;130(2):209-18.

52. Foradori MJ, Tillinghast EK, Smith JS, Townley MA, Mooney RE. Astacin family metallopeptidases and serine peptidase inhibitors in spider digestive fluid. Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology. 2006;143(3):257-68.

53. Jongejan F, Uilenberg G. The global importance of ticks. Parasitology. 2004;129:S3-S14.

54. Sojka D, Franta Z, Horn M, Hajdusek O, Caffrey CR, Mares M, et al. Profiling of proteolytic enzymes in the gut of the tick Ixodes ricinus reveals an evolutionarily conserved network of aspartic and cysteine peptidases. Parasites & Vectors. 2008;1.

55. Mendiola J, Alonso M, Marquetti MC, Finlay C. Boophilus microplus: Multiple proteolytic activities in the midgut. Experimental Parasitology. 1996;82(1):27-33.

56. Mulenga A, Sugimoto C, Onuma M. Characterization of proteolytic enzymes expressed in the midgut of Haemaphysalis longicornis. Japanese Journal of Veterinary Research. 1999;46(4):179-84.

57. Renard G, Garcia JF, Cardoso FC, Richter MF, Sakanari JA, Ozaki LS, et al. Cloning and functional expression of a Boophilus microplus cathepsin L-like enzyme. Insect Biochemistry and Molecular Biology. 2000;30(11).

58. Renard G, Lara FA, de Cardoso FC, Miguens FC, Dansa-Petretski M, Termignoni C, et al. Expression and immunolocalization of a Boophilus microplus cathepsin L-like enzyme. Insect Molecular Biology. 2002;11(4):325-8.

59. Boldbaatar D, Sikasunge CS, Battsetseg B, Xuan X, Fujisaki K. Molecular cloning and functional characterization of an aspartic protease from the hard tick Haemaphysalis longicornis. Insect Biochemistry and Molecular Biology. 2006;36(1):25-36.

60. Sojka D, Hajdusek O, Dvorak J, Sajid M, Franta Z, Schneider EL, et al. IrAE - An asparaginyl endopeptidase (legumain) in the gut of the hard tick Ixodes ricinus. International Journal for Parasitology. 2007;37(7):713-24.

61. Franta Z, Frantova H, Konvickova J, Horn M, Sojka D, Mares M, et al. Dynamics of digestive proteolytic system during blood feeding of the hard tick Ixodes ricinus. Parasites & Vectors. 2010;3:11.

62. Rodriguez de la Vega RC, Possani LD. On the evolution of invertebrate defensins. Trends in Genetics. 2005;21(6):330-2.

63. Chippaux JP, Goyffon M. Epidemiology of scorpionism: A global appraisal. Acta Tropica. 2008;107(2):71-9.

64. Guerrero-Vargas JA, Mourao CBF, Quintero-Hernandez V, Possani LD, Schwartz EF. Identification and Phylogenetic Analysis of Tityus pachyurus and Tityus obscurus Novel Putative Na+-Channel Scorpion Toxins. Plos One. 2012;7(2):13.

65. Louati H, Zouari N, Miled N, Gargouri Y. A new chymotrypsin-like serine protease involved in dietary protein digestion in a primitive animal, Scorpio maurus: purification and biochemical characterization. Lipids in Health and Disease. 2011;10.

66. Franta Z, Sojka D, Frantova H, Dvorak J, Horn M, Srba J, et al. IrCL1-The haemoglobinolytic cathepsin L of the hard tick, Ixodes ricinus. International Journal for Parasitology. 2011;41(12):1253-62.

67. Sojka D, Franta Z, Frantova H, Bartosova P, Horn M, Vachova J, et al. Characterization of Gut-associated Cathepsin D Hemoglobinase from Tick Ixodes ricinus (IrCD1). Journal of Biological Chemistry. 2012;287(25):21152-63.

68. Smith DJ, Maggio ET, Kenyon GL. Simple alkanethiol groups for temporary blocking of sulfhydryl groups of enzymes. Biochemistry. 1975;14(4):766-71.

69. Smith PK, Krohn RI, Hermanson GT, Mallia AK, Gartner FH, Provenzano MD, et al. MEASUREMENT OF PROTEIN USING BICINCHONINIC ACID. Analytical Biochemistry. 1985;150(1):76-85.

70. Beynon R, Bond JS. Proteolytic enzymes. 2nd edition ed. United States: Oxford University Press Inc., New York; 2001. 340 p.

71. Rawlings ND, Salvesen G. Handbook of proteolytic enzymes. 3rd ed. United States: Elsevier Science *Publishing* Co Inc; 2013. 4104 p.

72. Laemmli UK. Cleavage of structural proteins during assembly of head of bacteriophage-t4. Nature. 1970;227(5259):680.

73. Lemos FJA, Terra WR. Properties and intracellular-distribution of a cathepsin-d-like proteinase active at the acid region of musca-domestica midgut. Insect Biochemistry. 1991;21(5):457-65.

74. Sogawa K, Takahashi K. USE OF FLUORESCAMINE-LABELED CASEIN AS A SUBSTRATE FOR ASSAY OF PROTEINASES. Journal of Biochemistry. 1978;83(6):1783-7.

75. Melo RL, Alves LC, Del Nery E, Juliano L, Juliano MA. Synthesis and hydrolysis by cysteine and serine proteases of short internally quenched fluorogenic peptides. Analytical Biochemistry. 2001;293(1):71-7.

76. Cotrin SS, Puzer L, Judice WAD, Juliano L, Carmona AK, Juliano MA. Positional-scanning combinatorial libraries of fluorescence resonance energy transfer peptides to define substrate specificity of carboxydipeptidases: assays with human cathepsin B. Analytical Biochemistry. 2004;335(2):244-52.

77. Chen JM, Dando PM, Rawlings ND, Brown MA, Young NE, Stevens RA, et al. Cloning, isolation, and characterization of mammalian legumain, an asparaginyl endopeptidase. Journal of Biological Chemistry. 1997;272(12):8090-8.

78. Pimenta DC, Oliveira A, Juliano MA, Juliano L. Substrate specificity of human cathepsin D using internally quenched fluorescent peptides derived from reactive site loop of kallistatin. Biochimica Et Biophysica Acta-Protein Structure and Molecular Enzymology. 2001;1544(1-2):113-22.

79. Twining SS. Fluorescein isothiocyanate-labeled casein assay for proteolytic-enzymes. Analytical Biochemistry. 1984;143(1):30-4.

80. Lopes AR, Terra WR. Purification, properties and substrate specificity of a digestive trypsin from Periplaneta americana (Dictyoptera) adults. Insect Biochemistry and Molecular Biology. 2003;33(4):407-15.

81. Lopes AR, Saro PM, Terra WR. Insect chymotrypsins: chloromethyl ketone inactivation and substrate specificity relative to possible coevolutional adaptation of insects and plants. Archives of Insect Biochemistry and Physiology. 2009;70(3):188-203.

82. Saitou N, Nei M. The neighbor-joining method - a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution. 1987;4(4):406-25.

83. Felsenstein J. Confidence-limits on phylogenies - an approach using the bootstrap. Evolution. 1985;39(4):783-91.

84. Zuckerka.E, Pauling L. Molecules as documents of evolutionary history. Journal of Theoretical Biology. 1965;8(2):357.

85. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution. 2011;28(10):2731-9.

86. Turk B, Dolenc I, Lenarcic B, Krizaj I, Turk V, Bieth JG, et al. Acidic pH as a physiological regulator of human cathepsin L activity. European Journal of Biochemistry. 1999;259(3):926-32.

87. Turk V, Stoka V, Vasiljeva O, Renko M, Sun T, Turk B, et al. Cysteine cathepsins: From structure, function and regulation to new frontiers. Biochimica Et Biophysica Acta-Proteins and Proteomics. 2012;1824(1):68-88.

88. Ishidoh K, Kominami E. Processing and activation of lysosomal proteinases. Biological Chemistry. 2002;383(12):1827-31.

89. Nomura T, Fujishima A, Fujisawa Y. Characterization and crystallization of recombinant human cathepsin L. Biochemical and Biophysical Research Communications. 1996;228(3):792-6.

90. Cristofoletti PT, Ribeiro AF, Terra WR. The cathepsin L-like proteinases from the midgut of Tenebrio molitor larvae: Sequence, properties, immunocytochemical localization and function. Insect Biochemistry and Molecular Biology. 2005;35(8):883-901.

91. Miyaji T, Murayama S, Kouzuma Y, Kimura N, Kanost MR, Kramer KJ, et al. Molecular cloning of a multidomain cysteine protease and protease inhibitor precursor gene from the tobacco hornworm (Manduca sexta) and functional expression of the cathepsin F-like cysteine protease domain. Insect Biochemistry and Molecular Biology. 2010;40(12):835-46.

92. Kollien AH, Waniek PJ, Nisbet AJ, Billingsley PF, Schaub GA. Activity and sequence characterization of two cysteine proteases in the digestive tract of the reduviid bug Triatoma infestans. Insect Molecular Biology. 2004;13(6):569-79.

93. Cristofoletti PT, Ribeiro AF, Deraison C, Rahbe Y, Terra WR. Midgut adaptation and digestive enzyme distribution in a phloem feeding insect, the pea aphid Acyrthosiphon pisum. Journal of Insect Physiology. 2003;49(1):11-24.

94. Wang B, Shi GP, Yao PM, Li ZQ, Chapman HA, Bromme D. Human cathepsin F - Molecular cloning, functional expression, tissue localization, and enzymatic characterization. Journal of Biological Chemistry. 1998;273(48):32000-8.

95. Smith SM, Gottesman MM. Activity and deletion analysis of recombinant human cathepsin-l expressed in escherichia-coli. Journal of Biological Chemistry. 1989;264(34):20487-95.

96. Eakin AE, Mills AA, Harth G, McKerrow JH, Craik CS. The sequence, organization, and expression of the major cysteine protease (cruzain) from trypanosoma-cruzi. Journal of Biological Chemistry. 1992;267(11):7411-20.

97. Schmidt PG, Bernatowicz MS, Rich DH. Pepstatin binding to pepsin - enzyme conformation changes monitored by nuclear magnetic-resonance. Biochemistry. 1982;21(26):6710-6.

98. Ladrat C, Verrez-Bagnis V, Noel J, Fleurence J. Milli-calpain from sea bass (Dicentrarchus labrax) white muscle: Purification, characterization of its activity and activation in vitro. Marine Biotechnology. 2002;4(1):51-62.

99. Alim MA, Tsuji N, Miyoshi T, Islam MK, Huang X, Motobu M, et al. Characterization of asparaginyl endopeptidase, legumain induced by blood feeding in the ixodid tick Haemaphysalis longicornis. Insect Biochemistry and Molecular Biology. 2007;37(9):911-22.

100. Quintero-Hernandez V, Ortiz E, Rendon-Anaya M, Schwartz EF, Becerril B, Corzo G, et al. Scorpion and spider venom peptides: Gene cloning and peptide expression. Toxicon. 2011;58(8):644-63.

101. Mohien CU, Colquhoun DR, Mathias DK, Gibbons JG, Armistead JS, Rodriguez MC, et al. A Bioinformatics Approach for Integrated Transcriptomic and Proteomic Comparative Analyses of Model and Non-sequenced Anopheline Vectors of Human Malaria Parasites. Molecular & Cellular Proteomics. 2013;12(1):120-31.

102. Adamidi C, Wang Y, Gruen D, Mastrobuoni G, You X, Tolle D, et al. De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Research. 2011;21(7):1193-200.

103. Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 2008;18(5):821-9.

104. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28(8):1086-92.

105. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403-10.

106. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, et al. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Research. 2008;36(10):3420-35.

107. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Research. 2007;35:W585-W7.

108. Gouzy J, Carrere S, Schiex T. FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics. 2009;25(5):670-1.

109. Liu XW, Inbar Y, Dorrestein PC, Wynne C, Edwards N, Souda P, et al. Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins. Molecular & Cellular Proteomics. 2010;9(12):2772-82.

110. Muth T, Vaudel M, Barsnes H, Martens L, Sickmann A. XTandem Parser: An open-source library to parse and analyse X!Tandem MS/MS search results. Proteomics. 2010;10(7):1522-4.

111. Blanc G, Font B, Eichenberger D, Moreau C, Ricard-Blum S, Hulmes DJS, et al. Insights into how CUB domains can exert specific functions while sharing a common fold - Conserved and specific features of the CUB1 domain contribute to the molecular basis of procollagen C-proteinase enhancer-1 activity. Journal of Biological Chemistry. 2007;282(23):16924-33.

112. Padilha MHP, Pimentel AC, Ribeiro AF, Terra WR. Sequence and function of lysosomal and digestive cathepsin D-like proteinases of Musca domestica midgut. Insect Biochemistry and Molecular Biology. 2009;39(11):782-91.

113. Mommsen TP. Digestive enzymes of a spider (Tegenaria-atrica koch) .2. carbohydrases. Comparative Biochemistry and Physiology a-Physiology. 1978;60(4):371-5.

114. Mommsen TP. Comparison of digestive alpha-amylases from 2 species of spiders (Tegenaria-atrica and Cupiennius-salei). Journal of Comparative Physiology. 1978;127(4):355-61.

115. Jensen AG, Chemali M, Chapel A, Kieffer-Jaquinod S, Jadot M, Garin J, et al. Biochemical characterization and lysosomal localization of the mannose-6-phosphate protein p76 (hypothetical protein LOC196463). Biochemical Journal. 2007;402:449-58.

116. Tellam RL, Wijffels G, Willadsen P. Peritrophic matrix proteins. Insect Biochemistry and Molecular Biology. 1999;29(2):87-101.

117. Terra WR. The origin and functions of the insect peritrophic membrane and peritrophic gel. Archives of Insect Biochemistry and Physiology. 2001;47(2):47-61.

118. Forlino A, Lupi A, Vaghi P, Cornaglia AI, Calligaro A, Campari E, et al. Mutation analysis of five new patients affected by prolidase deficiency: the lack of enzyme activity causes necrosis-like cell death in cultured fibroblasts. Human Genetics. 2002;111(4-5):314-22.

119. Beckmann G, Bork P. AN Adhesive domain detected in functionally diverse receptors. Trends in Biochemical Sciences. 1993;18(2):40-1.

120. Nykjaer A, Willnow TE. The low-density lipoprotein receptor gene family: a cellular Swiss army knife? Trends in Cell Biology. 2002;12(6):273-80.

121. Stephens A, Rojo L, Araujo-Bernal S, Garcia-Carreno F, Muhlia-Almazan A. Cathepsin B from the white shrimp Litopenaeus vannamei: cDNA sequence analysis, tissues-specific expression and biological activity. Comparative Biochemistry and Physiology B-Biochemistry & Molecular Biology. 2012;161(1):32-40.

122. Sambrook J, Russell DW. Molecular cloning: a laboratory manual. 3rd edition ed. New York, United States: Cold Spring Harbor Laboratory Press; 2001.

123. Zhang Y, Frohman MA. Using rapid amplification of cDNA ends (RACE) to obtain full-length cDNAs. Nucleic Acid Protocols Handbook. 2000:267-88.

124. Thompson JD, Higgins DG, Gibson TJ. Clustal-w - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research. 1994;22(22):4673-80.

125. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods. 2011;8(10):785-6.

126. Ramos CRR, Abreu PAE, Nascimento A, Ho PL. A high-copy T7 Escherichia coli expression vector for the production of recombinant proteins with a minimal N-terminal his-tagged fusion peptide. Brazilian Journal of Medical and Biological Research. 2004;37(8):1103-9.

127. Collins PR, Stack CM, O'Neill SM, Doyle S, Ryan T, Brennan GP, et al. Cathepsin L1, the major protease involved in liver fluke (Fasciola hepatica) virulence - Propeptide cleavage sites and autoactivation of the zymogen secreted from gastrodermal cells. Journal of Biological Chemistry. 2004;279(17):17038-46.

128. Rozman J, Stojan J, Kuhelj R, Turk V, Turk B. Autocatalytic processing of recombinant human procathepsin B is a bimolecular process. Febs Letters. 1999;459(3):358-62.

129. Robinson MW, Corvo I, Jones PM, George AM, Padula MP, To J, et al. Collagenolytic Activities of the Major Secreted Cathepsin L Peptidases Involved in the Virulence of the Helminth Pathogen, Fasciola hepatica. Plos Neglected Tropical Diseases. 2011;5(4).

130. Kirschke H, Kembhavi AA, Bohley P, Barrett AJ. Action of rat-liver cathepsin-l on collagen and other substrates. Biochemical Journal. 1982;201(2):367-72.

131. Beton D, Guzzo CR, Ribeiro AF, Farah CS, Terra WR. The 3D structure and function of digestive cathepsin L-like proteinases of Tenebrio molitor larval midgut. Insect Biochemistry and Molecular Biology. 2012;42(9):655-64.

132. Stack CM, Caffrey CR, Donnelly SM, Seshaadri A, Lowther J, Tort JF, et al. Structural and functional relationships in the virulence-associated cathepsin L proteases of the parasitic liver fluke, Fasciola hepatica. Journal of Biological Chemistry. 2008;283(15):9896-908.

133. Lowther J, Robinson MW, Donnelly SM, Xu WB, Stack CM, Matthews JM, et al. The Importance of pH in Regulating the Function of the Fasciola hepatica Cathepsin L1 Cysteine Protease. Plos Neglected Tropical Diseases. 2009;3(1).

134. Beton D. Estrutura e função das cisteína proteinases intestinais do besouro Tenebrio molitor. http://www.teses.usp.br/teses/disponiveis/46/46131/tde-28042010-141010/: University of Sao Paulo; 2009.

135. Jerala R, Zerovnik E, Kidric J, Turk V. pH-induced conformational transitions of the propeptide of human cathepsin L - A role for a molten globule state in zymogen activation. Journal of Biological Chemistry. 1998;273(19):11498-504.

136. Kramer G, Paul A, Kreusch A, Schuler S, Wiederanders B, Schilling K. Optimized folding and activation of recombinant procathepsin L and S produced in Escherichia coli. Protein Expression and Purification. 2007;54(1):147-56.

137. Mommsen TP. Chitinase and beta-n-acetylglucosaminidase from the digestive fluid of the spider, Cupiennius-salei. Biochimica Et Biophysica Acta. 1980;612(2):361-72.

138. Mommsen TP. Digestive enzymes of a spider (tegenaria-atrica koch) .3. esterases, phosphatases, nucleases. Comparative Biochemistry and Physiology a-Physiology. 1978;60(4):377-82.

139. Zhu W, Smith JW, Huang C-M. Mass Spectrometry-Based Label-Free Quantitative Proteomics. Journal of Biomedicine and Biotechnology. 2010.

140. Henrissat B. A classification of glycosyl hydrolases based on amino-acid-sequence similarities. Biochemical Journal. 1991;280:309-16.

141. Filietaz CFT, Lopes AR. Caracterização de lipases em Arthropoda. Revista de Pesquisa e Inovação Farmacêutica. 2010;2(1):23-36.

142. Tugmon CR, Tillinghast EK. proteases and protease inhibitors of the spider Argiope aurantia (Araneae, Araneidae). Naturwissenschaften. 1995;82(4):195-7.

143. Shikata Y, Watanabe T, Teramoto T, Inoue A, Kawakami Y, Nishizawa Y, et al. Isolation and characterization of a peptide isomerase from funnel-web spider venom. Journal of Biological Chemistry. 1995;270(28):16719-23.

144. Sunagar K, Johnson WE, O'Brien SJ, Vasconcelos V, Antunes A. Evolution of CRISPs Associated with Toxicoferan-Reptilian Venom and Mammalian Reproduction. Molecular Biology and Evolution. 2012;29(7):1807-22.

145. Kobe B, Deisenhofer J. The leucine-rich repeat - a versatile binding motif. Trends in Biochemical Sciences. 1994;19(10):415-21.

146. Nasarre P, Potiron V, Drabkin H, Roche J. Guidance molecules in lung cancer. Cell Adhesion & Migration. 2010;4(1):130-45.

147. Titani K, Torff HJ, Hormel S, Kumar S, Walsh KA, Rodl J, et al. Amino-acid-sequence of a unique protease from the crayfish astacus-fluviatilis. Biochemistry. 1987;26(1):222-6.

148. Guevara T, Yiallouros I, Kappelhoff R, Bissdorf S, Stoecker W, Xavier Gomis-Rueth F. Proenzyme Structure and Activation of Astacin Metallopeptidase. Journal of Biological Chemistry. 2010;285(18):13958-65.

149. Gomis-Ruth FX, Trillo-Muyo S, Stocker W. Functional and structural insights into astacin metallopeptidases. Biological Chemistry. 2012;393(10):1027-41.

150. Zwilling R, Pfleider.G, Sonnebor.Hh. Zur evolution der endopeptidasen - eine protease vom molekulargewicht 11000 und eine trypsinahnliche fraktion aus astacus fluviatilis. Hoppe-Seylers Zeitschrift Fur Physiologische Chemie. 1967;348(10):1251.

151. Park JO, Pan J, Mohrlen F, Schupp MO, Johnsen R, Baillie DL, et al. Characterization of the astacin family of metalloproteases in C. elegans. Bmc Developmental Biology. 2010;10.

152. Bromme D, Bonneau P, Lachance P, Storer AC. Engineering the s2 subsite specificity of human cathepsin-s to a cathepsin-l-like and cathepsin-b-like specificity. Journal of Cellular Biochemistry. 1994:152.

153. kuhner mk, felsenstein j. a simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates (vol 11, pg 459, 1994). Molecular Biology and Evolution. 1995;12(3):525.

154. Tateno Y, Takezaki N, Nei M. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Molecular Biology and Evolution. 1994;11(2):261-77.

155. Robinson MW, Menon R, Donnelly SM, Dalton JP, Ranganathan S. An Integrated Transcriptomics and Proteomics Analysis of the Secretome of the Helminth Pathogen Fasciola hepatica Proteins associated with invasion and infection of the mammalian host. Molecular & Cellular Proteomics. 2009;8(8):1891-907.

156. Bown DP, Wilkinson HS, Jongsma MA, Gatehouse JA. Characterisation of cysteine proteinases responsible for digestive proteolysis in guts of larval western corn rootworm (Diabrotica virgifera) by expression in the yeast Pichia pastoris. Insect Biochemistry and Molecular Biology. 2004;34(4):305-20.

157. Shultz JW. A phylogenetic analysis of the arachnid orders based on morphological characters. Zoological Journal of the Linnean Society. 2007;150(2):221-65.

158. Mohrlen F, Hutter H, Zwilling R. The astacin protein family in Caenorhabditis elegans. European Journal of Biochemistry. 2003;270(24):4909-20.

159. Becker-Pauly C, Bruns BC, Damm O, Schuette A, Hammouti K, Burmester T, et al. News from an Ancient World: Two Novel Astacin Metalloproteases from the Horseshoe Crab. Journal of Molecular Biology. 2009;385(1):236-48.