Novel chymotrypsins from *Loligo opalescens* and *Sepioteuthis lessoniana*: Isolation, Purification and Molecular Characterization

by

Nana Akyaa Ackaah-Gyasi

Department of Food Science and Agricultural Chemistry
McGill University, Montreal
December, 2015

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Suggested Short Title:

Novel chymotrypsins from *L. opalescens a*nd *S. lessoniana* viscera

ABSTRACT

Chymotrypsins are widely distributed among living species and have found widespread use in different industrial applications. However, until the last two decades, most studies on chymotrypsin have been restricted to mammalian species with few reported works on marine invertebrates. The high catalytic activity of some aquatic enzymes at low temperatures, coupled with high pH and the relatively low thermal stability makes them robust in certain industrial applications where cold temperatures are preferred. In this study, chymotrypsin was purified to homogeneity and characterized from the viscera of two squid species (*Loligo opalescens*, cold water adapted and *Sepioteuthis lessoniana,* warm water adapted). Cold adapted chymotrypsin can transform substrates at low temperature thereby reducing loss of organoleptic properties and nutritional value of heat-sensitive substrates and products mostly found in industrial processing of food such as fish. They also provide economic benefit through energy savings during processing. This study also looks at the possibility of producing the enzymes using recombinant DNA technology, since it is currently not feasible or practical to extract these proteases from crude sources such as viscera for commercial application. The enzymes were purified about 300-fold, with a recovery of 44% and specific activities between 273.25 and 280 U/mg protein using ultrafiltration and affinity chromatography. Both enzymes migrated as single polypeptide chains with molecular masses of 22.0 ± 2.7 kDa and 18 ± 1.7 kDa, respectively, by SDS-polyacrylamide gel electrophoresis. The enzymes showed temperature and pH optima between 25°C - 35°C and 7.5 – 8.5, respectively. The kinetic constants K_m and k_{cat} for hydrolysis of benzoyl tyrosine ethyl ester (BTEE) were determined based on Hanes plots. K_m was 1.43 mM and k_{cat} was 103.43 sec^{-1} with a catalytic efficiency (k_{cat}/K_m) of 72.33 sec^{-1} mM^{-1} for *Sepioteuthis*; while for *Loligo*, K_m was 0.4 mM and k_{cat} was 349.21 sec^{-1} with a catalytic efficiency (k_{cat}/K_m) of 873.01 sec^{-1} mM^{-1}. *De novo* assembly from massively parallel sequencing data were generated from total RNA from *L. opalescens* (due to its biochemical characteristics) on an Illumina Genome Analyzer platform. The transcriptome was assembled on normalized reads using the Trinity assembler. Each component longest transcript was aligned against the uniprot_sprot_2013_11 protein database using the Blastx program from the NCBI BLAST family. Overall, 61661 transcripts were obtained with a total transcript length of 46232292 bp. Maximum transcript length obtained was 16932 bp, while the minimum length was 201 bp. A partial cDNA encoding *L. opalescens* chymotrypsin was identified from the *de novo* assembled

transcripts using BLAST algorithms in NCBI. A full- length cDNA was obtained using nested PCR. The deduced sequence consists of 292 amino acid residues, being longer than its vertebrate analogs. A search of the non-redundant protein database showed highest identity to a hypothetical protein from *Octopus bimaculoides* (69%), chymotrypsinogen A-like protein from *Lingular anatina* (45%) and a serine protease from *Aplysia californica* (45%). The catalytic triad involving histidine, aspartate and serine was conserved in *L. opalescens* chymotrypsin. The primary structure also contained higher content of methionine, arginine, histidine and glutamic acid relative to chymotrypsin from mammalian homologues. A substitution was observed in the *Loligo* chymotrypsin sequence corresponding to position 124 in bovine chymotrypsin sequence from proline to alanine. A maximum likelihood phylogenetic tree comparing chymotrypsin from *L. opalescens* with other vertebrate and invertebrate chymotrypsin sequences suggests the existence of two main groups representing chymotrypsin from vertebrates and invertebrates and shows that *L. opalescens* chymotrypsin shares a common ancestor with most insect chymotrypsins. A homology model of *L. opalescens* chymotrypsin was built using the crystal structure of bovine chymotrypsin (PDB ID: 1t8o) as a scaffold. The model was assessed for stereochemical quality and side chain environment. Sequence alignment shows that the target and the template (1t8o) share 36% sequence identity. The model had the characteristic 2 β-barrel domains typical of chymotrypsin. The Ramachandran plot using 258 residues gave a total stereochemistry of 88.5% with 8.3% in the additional allowed region. The percentage of residues found in the disallowed region was 0.9%. Based on crystallographic data, *Loligo* chymotrypsin and bovine chymotrypsin showed almost superimposable tertiary structures.

Résumé

On retrouve les enzymes de type chymotrypsine dans nombre d'organismes vivants; ces enzymes servent également à différentes applications industrielles. Jusqu`à la fin du siècle dernier, la majorité des études sur cet enzyme concernaient surtout celles provenant de mammifères, avec de rares travaux visant les invertébrés marins. En général, les enzymes d'origine aquatiques sont reconnus pour leur haute activité catalytique à basses températures, et en environnements plutôt alcalins. Leur thermostabilité leur confère une robustesse qui s`avère utile dans certaines conditions industrielles. Une chymotrypsine adaptée au froid peut ainsi transformer des substrats à faible température dès lors préservant les propriétés nutritionnelles et organoleptiques de substrats thermosensibles tels que les produits marins transformés en industrie agro-alimentaire. Par le fait même de telles conditions catalytiques permettent d'économiser l'énergie, un intrant parfois coûteux. Dans le cadre de la présente étude, nous avons purifié une chymotrypsine jusqu`à homogénéité, et effectué sa caractérisation à partir des viscères de deux espèces de calmar (*Loligo opalescens*, organisme d`eau froide, et *Sepioteuthis lessoniana,* adapté aux eaux chaudes). Nos travaux ont également évalué l'opportunité de produire ces protéases commercialement par génie génétique, ceci comme alternative rentable à une méthode peu rentable comme l'extraction conventionnelle à partir de biomasses marines. Les deux enzymes ont été purifiés par un facteur de 300 X, avec rendement de 44% et des activités spécifiques entre 273.25 et 280 U/mg de protéine via ultrafiltration et chromatographie par affinité. Lors de l'électrophorèse sur gel de SDS-polyacrylamide, les deux chymotrypsines migrent sous forme de chaines polypeptidiques simples, de poids moléculaires respectifs 22.0 ± 2.7 kDa et 18 ± 1.7 kDa. Nous avons déterminé leurs optima de température et pH comme étant respectivement 25°C - 35°C et 7.5 – 8.5. Quant aux constantes cinétiques K_m et k_{cat} durant l'hydrolyse du benzoyl tyrosine ethyl ester (BTEE), les diagrammes de Hanes ont permis de calculer un K_m de 1.43 mM et k_{cat} de 103.43 sec^{-1} pour l'enzyme de *Sepioteuthis* avec efficacité catalytique (k_{cat}/K_m) de 72.33 sec^{-1} mM^{-1}. Dans le cas de la protéine provenant de *Loligo*, nous avons déterminé un K_m de 0.4 mM et k_{cat} de 349.21 sec^{-1}, avec (k_{cat}/K_m) s`élevant à 873.01 $sec^{-1}$ $mM^{1}$. L'assemblage *de novo* à partir de données de séquençage parallèle massif a été généré avec l`ARN complet de *L. opalescens* à l'aide de la plateforme Illumina Genome Analyzer. Le système Trinity a permis d'assembler le transcriptome en utilisant les lectures normalisées. Pour chaque composante, la plus longue transcription a été alignée vs la base de données uniprot_sprot_2013_11 à l'aide du

logiciel Blastx provenant de la famille NCBI BLAST. Un total de 61661 transcriptions ont ainsi été obtenues, avec longueur complète de 46232292 paires de bases (pb), longueur maximale 16932 pb et minimum de 201 pb. Un ADNc partiel codant pour la chymotrypsine de *L. opalescens* a été identifié parmi les transcriptions de l'assemblage *de novo* via algorithmes BLAST du NCBI. D'autre part, l'utilisation du PCR par amorces incluses a permis d'obtenir un ADNc complet; la séquence ainsi déduite comprend 292 résidus d'acides aminés, donc d'une longueur supérieure aux analogues provenant de vertébrés. Un examen par bases de données sur protéines non-répétitives a indiqué un degré de similarité élevé avec une protéine hypothétique de *Octopus bimaculoides* (69%), avec une protéine de type chymotrypsinogène-A de *Lingular anatina* (45%) et également avec une sérine protéase de *Aplysia californica* (45%). La chymotrypsine de *L. opalescens* conserve la triade catalytique qui comprend les résidus histidine, aspartate et serine. En contraste avec les chymotrypsines homologues provenant de mammifères, la structure primaire présentait des proportions plus élevées de méthionine, arginine, histidine et acide glutamique. De plus, une substitution chez la protéine de *Loligo* a été notée à la position correspondant au résidu 124 de la chymotrypsine bovine, alanine remplaçant proline. La phylogénétique à haute probabilité comparant l'enzyme de *L. opalescens* avec les séquences d'autres protéines analogues de vertébrés et invertébrés suggère à la fois (i) l'existence de deux principaux groupes (chymotrypsines de vertébrés et d'invertébrés) et (ii) un ancêtre commun pour la protéine de *L. opalescens* et la plupart de celles analogues provenant d'insectes. Nous avons élaboré un modèle homologique pour la chymotrypsine de *L. opalescens* en utilisant la structure cristalline de l'enzyme bovin (PDB ID:1t8o) comme échafaudage; ce modèle a aussi fait l`objet d`évaluation pour la qualité stéréochimique et l'environnement des chaines latérales. L`alignement des séquences indique un partage d'identité de 36 % entre la cible et la matrice (1t8o). Le modèle présentait par ailleurs les deux domaines (tours β) particuliers aux chymotrypsines. Un diagramme de Ramachandran utilisant 258 résidus a démontré une stéréochimie totale atteignant 88.5% incluant 8.3% en zone favorable additionnelle. Le pourcentage de résidus situés en zone défavorable était de 0.9%. Sur la base des données cristallographiques, la chymotrypsine de *Loligo* et celle d'origine bovine présentent des structures tertiaires quasi superposables.

CONTRIBUTION OF AUTHORS

This thesis consists of eight chapters and is presented in manuscript format.

Chapter 1 is a general introduction and gives a brief perspective on distribution of chymotrypsin in nature. It also provides the outline of the research, rationale and objectives of the present study.

Chapter II provides a detailed review of the literature on enzymes including chymotrypsin, purification techniques, next generation sequencing techniques, cloning, expression and industrial applications of chymotrypsin.

Chapters III, IV, V, VI and VII, constitute the main body of the thesis and have or will be submitted for publications. Connecting statements provide logical bridges between different aspects of the research.

In Chapter III, chymotrypsins from two squid species (*Loligo opalescens* and *Sepiotethis lessoniana*) are purified and biochemically characterized. In Chapter IV, a partial cDNA encoding *S. lessoniana* chymotrypsin is cloned using degenerate primers designed from tryptic fragments derived from LC-MS/MS data. Chapter V consists of the de novo sequencing and assembly of the *L. opalescens* transcriptome using RNA seq. In chapter VI, molecular biology techniques and bioinformatics tools are used to isolate a novel chymotrypsin cDNA from the *L. opalescens* transcriptome assembly. Chapter VII describes the 3-dimensional structure of the isolated gene using a homology modelling approach. General conclusions are drawn in Chapter VIII. This dissertation format is in accordance with the guidelines for thesis preparation provided by the Faculty of Graduate and Postdoctoral Studies.

The author was responsible for all the experiments conducted, the data analysis, and the preparation of manuscripts for submission and publication. Professor Benjamin K. Simpson, the supervisor of the student, guided all biochemistry aspects of the work, provided laboratory space and supplies for the experiments, advice on direction, and edited the manuscripts, while Professor Timothy Geary; supervised both the molecular biology and bioinformatics aspects of the work and provided laboratory space, supplies for the experiments, advice on direction, content of the research, and edited the manuscripts. Dr. Traian Sulea provided technical and editorial assistance for Chapter VII

LIST OF PUBLICATIONS AND CONFERENCE PRESENTATIONS

Nana Akyaa Ackaah-Gyasi, Timothy Geary and Benjamin Simpson (2015). **Novel chymotrypsin-like enzymes from squid (*Sepioteuthis lessoniana* and *Loligo opalescens*) viscera; Purification, biochemical characterization and peptide identification using LC-MS/MS.** Submitted to Journal of Food Chemistry

Nana Akyaa Ackaah-Gyasi, Timothy Geary and Benjamin Simpson (2015). **Peptide sequencing and protein identification of a chymotrypsin-like enzyme from squid (*Loligo opalescens*).** Presented at Biotechnology Industry Organization conference, July 22, 2015, Montreal, Canada

Nana Akyaa Ackaah-Gyasi, Priyanki Patel, Yi Zhang and Benjamin K. Simpson. 2014. **Enzymes Current and Future Applications**. In: Improving and tailoring enzymes for food quality and functionality. (Ed. R. Yada) pp.103-122.

Nana Akyaa Ackaah-Gyasi, Priyanki Patel, Julie Ducharme, Hui Yin Fan and Benjamin K. Simpson. 2014. **Enzymes and Inhibitors in Food and Health**. In: Functional Polymers in Food Science: From Technology to Biology. (Francesca Iemma, U.G. Spizzirri and G. Cirillo, Eds). Scrivener Publishing LLC in cooperation with John Wiley and Sons Ltd. pp. 289-328.

Nana Akyaa Ackaah-Gyasi, Yi Zhang and Benjamin K. Simpson. 2014. **Enzymes Inhibitors: Food and Non-Food Impacts**. In: Advances in Food Biotechnology. (Ravishankar Rai, Ed). John Wiley and Blackwell Publishers. In press.

Nana Akyaa Ackaah-Gyasi, Timothy Geary and Benjamin Simpson (2014). **Purification and biochemical characterization of chymotrypsin from squid viscera. P**resented at the 17th World Congress of Food Science and Technology, August 17 – 20, 2014. Montreal, Canada.

ACKNOWLEDGEMENT

I would like to thank my supervisor, Professor Benjamin K. Simpson, for his support, direction, mentorship and keen interest in this work. I am very grateful. To Professor Timothy Geary, I say a big thank you. Could not have done this without you. Anytime I came to your office, I left with a big smile feeling more hopeful and encouraged.

I am grateful to all my colleagues, Saranya, Chijioke, Chloe, Yi and Tina, Ebenezer, Gabby, Ella and Nike for their time, friendship and encouragement throughout my studies.

I would also like to thank all the teaching and non-teaching staff of the Department of Food Science and Agricultural Chemistry and Institute of parasitology, McGill University, especially Leslie, Trish and Vicki.

I am profoundly grateful to my husband, Mike Quashie and daughter, Naana Adadzewa,for their support and encouragement throughout my studies. To my parents and siblings, I say thank you.

Finally, I am very grateful to the Almighty God for the strength and direction He gave me throughout my studies

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS AND ACRONYMS

BAC……..	Bacteria artificial chromosome
BLAST……..	Basic Local Alignment Search Tool
BTEE……..	Benzoyl-L-tyrosine-ethyl-ester
cDNA…….	Complementary Deoxyribonucleic acid
DTT……...	Dithiothreitol
EDTA…….	Ethylenediaminetetraacetic acid
EST…….	Express sequence tags
LB……..	Luria-Bertani
LC……...	Liquid chromatography
MS……..	Mass spectrometry
NCBI……..	National Center for Biotechnology Information
PCR……	Polymerase chain reaction
PDB……	Protein Data Bank
PMSF…….	phenylmethylsulfonyl fluoride
SBTI…….....	Soybean Trypsin Inhibitor
TPCK…….	Tosyl phenylalanyl chloromethyl ketone
YAC……...	Yeast artificial chromosome

CHAPTER 1

1. INTRODUCTION

Enzymes such as chymotrypsin have found widespread use in different industrial applications (e.g., pharmaceuticals, food processing, detergents and leather industries) due to their catalytic properties and capacity to produce uniform products consistently (Bucholz et al., 2005; El Enshasy et al., 2008). Chymotrypsin has been purified and characterized from different sources, including mammals such as bovine (Balti et al., 2012), fish such as crucian carp (Yang et al., 2009) and Atlantic cod (Asgeirson and Bjarnason, 1991), mollusks such as cuttle fish (Balti et al., 2012), invertebrates such as cut worm (Zhang et al., 2010) and microorganisms such as *Metarhizium anisopliae* (Screen and Leger, 2000). However, enzymes have limitations in their ability to adapt to extreme environmental changes (Simpson, 2000). This has led to a heightened interest in finding alternate sources of chymotrypsin which are superior to known enzymes, or an improvement on current forms of the enzymes that would suit specific applications in industry (Psochiou et al., 2007).

Research has shown that the source of an enzyme is an important variable for application, as those obtained from fish and other aquatic organisms have been reported to possess several distinct properties suitable for distinct industrial applications compared to those derived from mammalian sources (Haard, 1992). The distinct properties of enzymes of aquatic origin are attributed to the psychrophilic nature and adaptation of the parent organism. For example, fish or aquatic source enzymes are better able to adapt to extreme environmental conditions of pressure, temperature, salinity and alkalinity compared to homologous enzymes from species acclimated to warmer environments (Simpson, 2000). This makes it attractive to study enzymes of aquatic origin such as squids for homologues of chymotrypsin.

Squids are important marine invertebrates and members of the marine ecosystems serving as food for many fishes (e.g., tuna) and marine mammals (e.g., seals). They in turn prey on other fishes and crustaceans, and preference for food is dependent on season (Guerra, 2006). Squids comprise one of the most economically important aquatic families in the world and together with cuttlefish and octopus, account for ≥ 3% of world's total capture of all aquatic organisms (FAO 2013). The species has a high tolerance for environmental changes and undergoes seasonal migrations between shallow coastal waters in summer for spawning and deeper waters in winter (Gauvrit et al., 1997). To date, no published studies have investigated major proteases in squid.

Also, few studies have been done on chymotrypsin from marine invertebrates (Balti et al., 2012). This research will explore chymotrypsin in squid up to the molecular level and how differences in amino acid and 3 – D structure accounts for some observed biochemical characteristics.

Apart from source, enzyme purification and proper characterization can be a major challenge (Mowery and Seidman, 2005). Methods used for purification involve precipitation, filtration and chromatography (Nielsen, 2010). These methods are based on the properties of the enzyme of interest (Nielsen, 2010; Chaplin and Bucke, 1990). In most cases, the target protein for purification may constitute as little as 0.001% of the total protein in the cell, making it difficult to get a good yield (Mowery and Seidman, 2005). At most, a target protein may be 10% of the starting crude extract, unless over-expression is induced by genetic engineering methods (Walsh, 2002). Affinity chromatography is one of the most powerful tools used in the purification of enzymes with respect to biological function or individual chemical structure, and is a relatively simple and bio-specific technique (Nielsen, 2010).

Many researchers have purified and characterized chymotrypsin, a major digestive protease, from aquatic animals, and have shown its potential in various industries to be enormous. To date, however, no such work has been carried out for squid, an important marine mollusk with a highly adaptive life style. It is not currently feasible or practical to extract these proteases from crude sources such as squid viscera for commercial application. This study therefore aimed to isolate and characterize chymotrypsin form squid viscera and then use recombinant DNA technology to clone the enzyme in a microorganism. Recombinant DNA technology has been successfully employed for the large scale production of many industrial enzymes (Walsh, 2002; Chaplin and Bucke, 1990). The research also aims at elucidating the amino acid sequence of the isolated enzymes with subsequent modeling to produce a 3D structure of the enzyme. The prediction of a bioactive conformation of the enzyme will give deeper insights into the molecular basis of the biological function of squid cold-adapted enzymes. Apart from bridging the knowledge gap on chymotrypsin from marine invertebrates such as squid, the enzyme could also serve as an additional or alternative source of catalyst for industrial processing such as meat tenderization, fermentation, protein hydrolysate production, fish scale removal and bone protein removal, Based on these points, it is hypothesized that squid viscera can yield a form of chymotrypsin that is endowed with cold-tolerant properties that are well-suited for particular industrial applications based on its adaption to peculiar habitats and

changing feeding pattern. Furthermore, the use of combined classical techniques and cloning can provide an easy and cheap source of this chymotrypsin. To test these hypotheses, the objectives of the study have been set to:

1. Isolate and purify chymotrypsin from 2 squid species (*Loligo opalescens* and *Sepioteuthis lessoniana*) viscera
2. Characterize and study the biochemical properties of the purified enzymes in terms of temperature, pH, inhibitor interactions and kinetic properties
3. Determine amino acid sequences of the enzymes by de novo sequencing, assembly and polymerase chain reaction
4. To predict a 3-dimensional structure of the enzyme by homology modelling

CHAPTER 2

2. LITERATURE REVIEW

This chapter highlights the general characteristics of enzymes, the rationale for their study, common techniques and difficulties encountered during the purification and characterization of native enzymes from whole cell extracts. It also presents general techniques employed in mass production of these enzymes with more emphasis on techniques used in identifying and cloning novel genes for commercial purposes, as well as their current and potential applications and the future of enzymes.

2.1 Nature of enzymes and their mode of action

Enzymes are biological catalysts that are produced by living cells (Chaplin and Bucke, 1990). In living organisms, enzymes accelerate biochemical reactions under the physiological conditions of pH, pressure, temperature, ionic strength and other conditions that exist in living cells (Buchholz et al., 2005). One of the most important properties of enzymes is specificity since they lead to consistent product formation. In general, enzymes can be grouped into four main categories based on their specificities (Simpson, 2012), as follows:

- Absolute specificity – This group of enzymes catalyzes only one reaction involving a specific substrate. Examples include carbonic anhydrase, which acts only on carbonic acid, and arginine amidinase, which acts only on arginine
- Group specificity – This group acts only on closely-related molecules that have specific functional groups, such as amino, phosphate or methyl groups. Examples include chymotrypsin, which cleaves peptide bonds in which the c-terminal is an aromatic amino acid, and aminopeptidases, which cleave peripheral amino acids at the n-terminal of a peptide chain.
- Linkage specificity - Enzymes in this group act on a particular type of chemical bond regardless of the rest of the molecular structure. Examples include amylase and lipase, with the former acting on α 1-4 glycosidic linkages in starch, dextrins and glycogen, and the latter hydrolyzing ester bonds in triglycerides.
- Stereo-chemical specificity – This group of enzymes act on a particular steric or optical isomer. An example is α-glycosidase, which acts only on α-glycosidic bonds in starch, dextrins or glycogen.

2.2 Methods of studying enzymes

Enzyme reactions may be studied using *in vivo* or *in vitro* techniques. An example of an *in vivo* enzyme technique is the use of collagen or gelatin labeled with fluorescent probes to visualize the activity of matrix-degrading enzymes, such as cathepsin B (Sameni et al., 2000) and matrix metallo-proteinases (Mook et al., 2003). However, in the laboratory, most enzyme activities are studied *in vitro*. An example of an *in vitro* enzyme assay is the chymotrypsin assay using benzoyl-L-tyrosine-ethyl-ester (BTEE), a synthetic substrate (Yang et al., 2009; Asgeirson and Bjarnason, 1991). The use of any of these techniques is dependent upon the availability of suitably characterized substrates and methods of determining the changes in their properties caused by enzymes (Speers and Cravatt, 2004).

2.2.1 *In vivo* techniques

In vivo methods involve studying changes in a known substrate due to the presence of an enzyme. This technique has been widely used in investigating various enzymes such as cytochrome P4503A (CYP3A) (Thummel and Wilkinson, 1998) and members of the kinase family (Higashi et al., 1997). The method has the advantage of mimicking the natural situation more closely (Speers and Cravatt, 2004). However, there is always a difficulty in identifying reactions of specific enzymes, and also separating and identifying intermediate products of reaction since they are found in mixture with final products (Wang, 2001).

2.2.2 *In vitro* techniques

In vitro techniques usually involve extraction of the enzyme of interest from a culture fluid or cells of an organism, followed by purification, laboratory observations and analysis (Wang, 2001). Though this method is more specific, it is artificial in nature and does not always mimic real life catalysis in situ (Wang, 2001).

2.3 Enzyme purification Techniques

Purification of enzymes is a multistep process which ensures that the enzyme of interest is separated from other components that may cause undesirable changes in systems to which the enzyme is to be applied (Walsh, 2002). The primary purification step usually involves disruption of cells and the separation of cell debris from liquid containing the enzyme using centrifugation

or membrane filtration (Bonner, 2007). Secondary purification methods are based on physico-chemical principles such as solubility, charge density, charge distribution, hydration, size, shape, density, specific reactive group, and hydrophobicity (Simpson, 2012).

2.3.1 Separation by different solubility characteristics

Precipitation is one of the most commonly used methods in the purification of enzymes (Mowery and Seidman, 2005). Enzymes are charged molecules and disruption of their charge can lead to the aggregation of enzymes in solution in the form of a precipitate (Nielsen, 2010). Precipitation may also be achieved via hydrophobicity, hydration and solubilization of the enzyme. The most common precipitants used are neutral salts, acids, bases, organic solvents and soluble non-ionic polymers (Walsh, 2012).

2.3.1.1 Precipitation with neutral salts

Due to the unique solubility behavior of enzymes in neutral salts, precipitation can be achieved in a process called salting out. Salts are able to change the structure of solvents which alters the electrostatic interaction between charged groups on the protein surface (Simpson, 2012). Salts also compete with enzymes for water of hydration, thereby lowering its solvation (Walsh, 2002). At a point which depends on the enzyme in question, enzymes are forced out of solution. The most common salt used is ammonium sulfate because of its high solubility and low cost (Nielsen, 2010). Sodium chloride, sodium sulfate and potassium chloride can also be used in this regard (Bonner, 2007).

2.3.1.2 Precipitation with acids and bases

The addition of acid or base to an enzyme mixture directly affects its overall charge (Bonner, 2007; Mowery and Seidman, 2005). The isoelectronic point (pI), also known as the isoionic point, is the pH at which the total net charge on a protein is zero. At pH values above the pI, proteins are negatively charged. At pH values below the pI, the surface of a protein is predominantly positively charged and repulsion between proteins occurs, causing them to be in solution (Nielsen, 2010). At the pI, where negative and positive charges cancel, repulsive electrostatic forces are reduced and attraction forces predominate, causing aggregation and

precipitation of the protein out of solution. Acid and bases, however, can inactivate enzymes, thus limiting their use as precipitants (Nielsen, 2010).

2.3.1.3 Precipitation with organic solvents and water

Organic solvents such as acetone, methanol, ethanol and 2, 4-methylpentanediol (MPD) are used in this respect. The use of organic solvents as precipitants for enzymes is related to the fact that protein solubility at a fixed pH and ionic strength is a function of the dielectric constant of a given solution (Simpson, 2012). Water-miscible organic solvents usually lower the dielectric properties of an aqueous solution, leading to a decrease in the ionization of charged amino acids and protein aggregation and precipitation (Nielsen, 2010). Solvent fractionation is usually done at or below 0°C to prevent denaturation of proteins.

2.3.1.4 Precipitation with water soluble non-ionic polymers

The most commonly used polymers as precipitants are polyethylene glycol (PEG), alginate, pectinate, carboxymethyl cellulose, polyethylenimine, polyacrylic and polymeta acrylic acids (Walsh, 2002). Even though all these polymers are able to cause proteins to precipitate out of solution, PEG is most commonly used. The mechanism of precipitation involves competition between the polymer and proteins for water, thereby changing the dielectric properties of the solution (Nielsen, 2010).

2.3.2 Separation by charge differences

Each protein has numerous amino acid side chains, some of which are charged. Therefore, each protein carries a net charge. Separation of proteins based on charge difference is achieved through methods such as sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) and electro-focusing, which are based on this principle.

2.3.2.1 SDS-PAGE

SDS-PAGE is one of the best-known methods used to separate proteins (Mowery and Seidman, 2005). The method takes advantage of the movement of charged molecules including proteins when placed in an electric field (Nielsen, 2010; Deutscher, 1990). Proteins can migrate through a

polyacrylamide gel, with velocity dependent on the pore size, which varies with the concentration of polyacrylamide used, and the voltage applied to the gel (Deutscher, 1990). Separation of proteins by SDS-PAGE occurs by a sieving mechanism with smaller proteins moving fastest towards the anode (Walsh, 2002). Proteins to be separated by electrophoresis are first incubated with SDS and mercaptoethanol to denature the proteins and break disulfide linkages. This step generally linearizes proteins and confers the same negative charge density to all protein by binding directly to the proteins (Bonner, 2007). SDS-PAGE can also be used to determine the relative molecular mass as well as the purity of a protein (Walsh, 2002). Purified proteins in native forms can also be separated by native-PAGE without denaturation (Mowery and Seidman, 2005).

2.3.2.2 Isoelectric focusing (IEF)

IEF, also known as electro-focusing, is a technique used for the analysis of proteins and for micropurification of proteins (Walsh, 2002). It is an electrophoretic technique in which mixtures of low molecular mass organic acids and bases are used to achieve protein separation. The acids and bases, also known as ampholytes, distribute in a gel under the influence of an applied electric field, creating a pH gradient (Walsh, 2002; Deutscher, 1990). Protein samples to be applied are usually denatured by incubation with urea, causing the amino acid side chain to be exposed (Walsh, 2002). The exposed side chains contribute to the pI (Neilson, 2010). In a pH gradient, the proteins migrate towards the anode or the cathode which corresponds to the pH value of their isoelectric point. Combination of SDS-PAGE with isoelectric focusing is a powerful tool for protein purification. It is referred to as two-dimensional electrophoresis and can resolve between 1000-2000 protein bands (Walsh, 2002).

2.3.3 Separation by chromatographic behavior

Chromatographic techniques are based on the separation of compounds by adsorption to or desorption from the surface of a solid support by an eluting solvent (Nielsen, 2010). The solid support media used are usually inert matrices packed in a glass or steel tube. Most commonly used column materials include finely divided solids such as silica, alumina, calcium phosphate and hydroxyapatite (Walsh, 2002). Many different types of adsorption chromatography are used with each method operating on a specific principle.

2.3.3.1 Affinity Chromatography

Affinity chromatography is a purification technique that purifies biomolecules with respect to biological function or individual chemical structure, hence its classification as a bio-specific technique (Nielsen, 2010; Simpson, 2012). This technique is one of the most powerful tools used for purification of proteins (Walsh, 2002). The substance to be purified is specifically and reversibly adsorbed to a ligand (binding substance). The ligand is usually covalently attached to an inert support matrix (Walsh, 2002). Most ligands are first attached to spacer arms which are then bonded to the matrix. Most common ligands are enzyme inhibitors, enzyme substrates, coenzymes, antibodies or dyes. Samples are applied under favorable conditions for their specific binding to the ligand. Target molecules are reversibly bound to the ligand on specific sites while unbound substances are washed away under specific buffer conditions. Recovery of proteins of interest can be achieved by changing experimental conditions such as pH, ionic strength, temperature and polarity to favor desorption. An average of a 100-fold purification can be achieved by this technique, and a 1000-fold increase in purification has been reported (Nielsen, 2010).

2.3.3.2 Ion-exchange chromatography (IEC)

IEC involves the reversible adsorption between charged molecules and ions in solution and a charged solid support matrix (Bonner, 2007). It is the most commonly used technique for protein separation and results in an average eight-fold purification (Nielsen, 2010). The adsorption of proteins to the solid support is driven by the ionic interaction between the oppositely charged ionic groups in the sample molecule and in the functional ligand on the support. Usually, the protein to be purified is first adsorbed to an ion exchanger under specific buffer conditions. Contaminating proteins pass through the exchanger without interacting with the ions in the support matrix. Elution of the protein of interest from the column is achieved by changing the ionic strength or pH of the eluting buffer. This causes the displacement of the proteins from charges by a new counter-ion with a greater affinity for fixed charges than the protein of interest, causing it to be eluted from the column (Nielsen, 2010).

2.3.3.3 High Performance Liquid Chromatography (HPLC analysis)

This method of separation is characterized by very high peak resolution and fast fraction speed (Walsh, 2002). This technique is rapid and highly sensitive due to the use of sensitive protein detectors. Usually, two or more different HPLC column types are used to assess protein purity. Most common are reverse phase columns and ion-exchange columns. In reverse phase HPLC, a nonpolar stationary phase such as modified silica is used with an elution gradient of decreasing polarity (Rhodes et al., 2009). The presence of the organic solvent in the mobile phase decreases the affinity of the protein for the stationary phase and the protein elutes. Detection of the protein is usually based on UV absorbance at 215–220 nm, which will identify peptides as opposed to measurements based on aromatic amino acids at 280 nm (Rhodes et al., 2009). HPLC can also be used to check protein purity (Walsh, 2002).

2.3.4 Separation by Size

Enzymes vary in size depending on the number of amino acids present and may be categorized as low or high molecular weight enzymes (Walsh, 2002). The ability of different enzymes to pass through different systems due to their size forms the basis of their separation by size (Bonner, 2007)

2.3.4.1 Centrifugation

Centrifugation is a widely applied research technique that uses centrifugal force (g-force) to isolate suspended particles from a solution in a batch or a continuous-flow basis (Nielsen, 2010). When a suspension is rotated at a certain speed centrifugal force causes the particles to move radially away from the axis of rotation. The force on the particles relative to gravity is called Relative Centrifugal Force (RCF). Separation by centrifugation may be primarily based on the size of the particles in a process known as differential centrifugation (Walsh, 2002). On the other hand separation can be achieved through difference in density gradient (Bonner, 2007). Differential centrifugation is commonly used in simple pelleting and obtaining partially-pure preparation of subcellular organelles and macromolecules including proteins. Density gradients on the other hand can be to purify subcellular organelles and macromolecules.

2.3.4.2 Ultrafiltration

This is membrane filtration in which hydrostatic pressure forces a solution against a semipermeable membrane (Bonner, 2007; Chaplin and Bucke, 1990). The primary removal mechanism is by size, achieved through the use of a semipermeable membrane with a defined range of pore sizes leading to purification, separation, and concentration (Simpson, 2012). The overall shape of proteins affects their ultrafiltration characteristics and most ultrafilters cut-off points apply to globular proteins (Walsh, 2002). It is also important to note that certain post-translational modifications of proteins, such as glycosylation, may also affect ultrafiltration behavior of proteins (Walsh, 2002). A major drawback of using ultrafiltration is the build-up of a concentrated layer of high molecular weight materials directly over the membrane surface, preventing other molecules from passing through the membrane (Nielsen, 2010). Ultrafiltration is relatively simple, rapid and has little adverse effect on bioactivity of proteins.

2.3.4.3. Dialysis

Dialysis is one of the oldest techniques employed in the removal of low molecular weight molecules, usually solutes, through exchange of buffers against a semi-permeable membrane (Chaplin and Bucke, 1990). The technique allows the free passage of molecules below a certain molecular weight cut-off while larger molecular weight molecules are unable to penetrate the pores of the membrane. The driving force of this method is the difference in concentration of solutes on the two sides of the membrane (Walsh, 2002). The diffusion of solutes becomes equal in both directions when the concentration reaches equilibrium (Nielsen, 2010). Increasing the ratio of the membrane area to the volume of the solution enhances the speed and the rate. Dialysis is affected by temperature and viscosity of the solution.

2.3.4.4 Gel Filtration

This, column technique, can be used to separate proteins based on their size and shape. The sample containing the proteins to be separated is percolated through the column packed with porous beads, as it travels down the column, large proteins are the first to be eluted because they interact less with the column due to inability to enter the gel beads (Chaplin and Bucke, 1990). In this technique, protein molecules are eluted from the column in order of decreasing molecular

size (Walsh, 2002). The most important factors that affect gel filtration are the diameter of the pores that allow access to the internal volume and the hydrodynamic diameter of the protein molecule (Nielson, 2010).

2.3.5 Separation based on hydrophobic properties of protein

The hydrophobicity of a protein is related to its amino acid composition (Walsh, 2002). Hydrophobic portions of most proteins are folded such that these portions are found buried in the internal structure of the molecule, shielding these groups from surrounding water (Neilson, 2010). There are, however, a few hydrophobic amino acids present on the surface of most proteins which contribute to an important property of native proteins with regard to hydrophobic interaction chromatography (Bonner, 2007).

2.3.5.1 Hydrophobic interaction chromatography (HIC)

HIC fractionates proteins by exploiting their different degrees of hydrophobicity due to interactions between hydrophobic patches on a protein's surface and hydrophobic groups covalently attached to a support matrix. Protein separation by HIC is usually done in aqueous salt solutions which increases the hydrophobicity of protein molecules. Samples are loaded unto the matrix in a high-salt buffer and elution is by descending salt gradient (Bonner, 2007).

2.4 Evidence of enzyme purity

Purity of an enzyme may be referred to as the absence of contaminants in the final product (Walsh, 2002). Depending on the end usage, enzymes may be purified to different extents (Chaplin and Bucke, 1990). The type of purification strategy employed is related to maximum recovery of the target protein; minimal loss of biological activity; and maximum removal of contaminating proteins (Mowery and Seidman, 2005). Purity of an enzyme can be assessed using activity testing, electrophoresis techniques and mass spectroscopy.

2.4.1 Activity testing

Biological activity studies aim to elucidate the biological function and activity of purified proteins and how various factors may contribute (Bonner, 2007). The protein type as well as its intended use may affect the type of activity testing carried out. For enzymes, determination of

specific activity is a pertinent parameter to study. During enzyme purification, specific assays permit the determination of yield, which gives an idea about the level of purity. Enzymes are usually assayed by their ability to catalyze a specific reaction which can be monitored either by product formation or substrate disappearance (Mowery and Seidman, 2005). In this regard, colorimetric assays are employed based on a component of the reaction that absorbs light at a given wavelength (Chaplin and Bucke, 1990). Usually, an additional assay is needed to determine the amount of all proteins, called the protein/enzyme concentration assay, which enables computation of the specific activity of the enzyme (Mowery and Seidman, 2005). Each purification step is characterized by a level of purity and yield. As the purification process proceeds, the specific activity should increase because contaminants are being removed, thereby decreasing the total weight of all proteins in the sample. This indicates greater purity due to fewer contaminating proteins (Chaplin and Bucke, 1990).

Many other methods are used to investigate enzyme purity, such as mass spectroscopy. Other methods outlined earlier, such as SDS-PAGE, isoelectric focusing, and HPLC can also be employed to check the purity of a protein. Methods based on the chromatographic behavior of a protein are also used to check purity.

2.4.2 Mass spectrometry (MS)

MS is one of the most recent methods used to study the purity of proteins. Apart from its high sensitivity and simplicity, it also provides a direct measure of the mass distribution in a sample (Rhodes and Laue, 2009). Its routine application to proteins is due to the development of suitable ionization techniques that allow generation of gas-phase ionized proteins (Walsh, 2002). Apart from detection of impurities in a protein sample, it also characterizes proteins and any impurities by their mass. It is therefore often possible to identify the origin of the impurity (Rhodes and Laue, 2009). MS can provide an accurate estimation of mass for proteins up to 500 kDa in size with accuracy in the region of 0.01% (Walsh, 2002). In addition to mass analysis, MS can be used to generate protein sequence due to the development of ionization techniques such as electrospray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI) that can transform biomolecules into ions. The process typically involves digestion of protein with enzymes such as trypsin to produce peptide fragments. The products of digestion are then concentrated under vacuum and injected into an HPLC system connected online to an ESI or

MALDI mass spectrometer. Sometimes, two or more sequential separation of ions can be done by coupling two or more mass analyzers referred to as tandem mass spectrometry (MS/MS). Peptide sequences can then be derived from spacing between adjacent product ions of the same series to produce spectra. The spectra produced can be searched against theoretical product ion spectra through a number of database algorithms such as Mascot, Sequest and XTandem to produce peptides. These peptides can then be used to search and match protein sequences.

2.4.3 Enzyme concentration assays

Various methods are used to detect and quantify enzyme/protein concentration in biological samples. The different methods come with some advantages and disadvantages as well as sensitivities

2.4.3.1 Absorbance at 280 nm

This is the most common method used. Proteins in solution have a maximum absorbance of ultraviolet light or optical density at 280 nm. The basis of this method is the fact that aromatic amino acids (tyrosine, tryptophan, and to a lesser extent phenylalanine) absorb UV rays strongly at 280 nm (Walsh, 2002). This method is fast, simple and non-destructive; however, it has low sensitivity (between 0.02-3.0 mg/ml of protein) (Nielsen, 2010).

2.4.3.2 Acid digestion (ninhydrin method)

It involves the hydrolysis of protein into constituent amino acids by incubation with sulphuric acid (H_2SO_4) at 100°C and then quantification of amino acids by subsequent reaction with ninhydrin, which forms a derivative with maximal absorbance at 570 nm (Bonner, 2007). The sensitivity of this method is between 20-50 μg protein per ml (Walsh, 2002). Due to the use of hazardous chemical at elevated temperatures, the method is not often used. It is also time consuming.

2.4.3.3 Bradford Method

Detection by this method is based on the reversible pH-dependent binding of Coomassie brilliant blue G-250 dye to protein, mostly to basic and aromatic amino acids which form a complex with

maximal absorbance at 595 nm (Walsh, 2002). Concentrations are determined relative to a standard curve based on bovine serum albumin (BSA). The method is moderately sensitive (150-750 μg/ml), fast and simple. However, some buffers, alkaline pH and detergents can interfere with the assay.

2.4.3.4 Hartree - Lowry Method

In this assay, proteins are reacted with a combination of copper and phosphomolybdic/ phosphotungstic acid. The resultant solution has an absorbance maximum at 750 nm from which total proteins can be estimated. The assay is laborious and subject to interference by detergents and chelating agents, however, it is sensitive (30-150 μg/ml) (Walsh, 2002).

2.4.3.5 Bicinchoninic acid (BCA) method

The BCA method involves the reaction of a copper-containing reagent with protein which causes the copper in the reagent to be reduced due to its interaction with the peptide bonds in the protein. The amount of copper reduced is proportional to the amount of protein present in the solution. When protein is reacted with bicinchoninic acid, it yields a derivative which absorbs maximally at 562 nm. The amount of protein present in a solution can be quantified using a standard curve prepared with protein solutions of known concentration. This method is not affected by the presence of detergents in the reaction mixture (Walsh, 2002). It is one of the most sensitive assays (20-100 μg/ml) and very convenient, but the reagents are very expensive.

2.5 Chymotrypsin

Chymotrypsin is one of the many enzymes whose three-dimensional structure has been well characterized. It is a serine protease that catalyzes the hydrolysis of an amide group or ester group present in a substrate which contains a leaving alcohol group (Blow, 1971). The active site of chymotrypsin is part of a hydrophobic pocket called the “tosyl hole” in which specific and non-specific substrates are bound and hydrolyzed by a general acid-base catalyzed reaction (Parker and Wang, 1968). The enzyme cleaves peptide bonds involving amino acids with bulky side chains and non-polar amino acids such as phenylalanine, tyrosine, leucine and tryptophan (Simpson, 2000).

2.5.1 Structure/forms of Chymotrypsin

They consist of two differently charged molecues which are both single-polypeptide chains with molecular weight (MW) of 26.6 ± 1.5 kDa (Asgeirsson and Bjarnason, 1991). Klomklao (2008) reported a molecular weight of between 25 and 28 kDa for these proteins. Chymotrypsin is synthesized in the pancreas as the precursor, chymotrypsinogen. It is known to have about 245 residues arranged in two six-stranded beta-barrels and 5 pairs of disulfide linkages (Blow, 1971). It exists in three inactive forms (chymotrypsinogens A, B and C) in the pancreas (Raae, 1995). These three forms have been found in mammals but only two forms of chymotrypsin (A and B) have been found in fish (Yang et al., 2009).

2.5.2 Chymotrypsin in Fish

Most species of fish lack a well-defined pancreatic organ (Fong et al., 1998). For example, in cod and rainbow trout, pancreatic tissue is associated with the mesentery of the pyloric caeca (Kristjansson and Nielsen, 1992; Asgeirsson and Bjarnason, 1991). In other species, the pancreatic tissue is fused with the liver (Fong et al., 1998). Overnell (l973) was the first author to report fish chymotryptic activity in crude extracts from the pyloric caeca of the Atlantic cod. Kristjansson and Nielsen (1992) isolated chymotrypsin from the pyloric caeca of rainbow trout, obtaining two peaks of activity, I and II, which migrated as single bands on SDS-PAGE. Fish chymotrypsins usually have cationic (chymotrypsin A) and anionic (chymotrypsin B) forms. Work by Yang et al., (2009) found two isoforms of chymotrypsin in the hepatopancreas of crucian carp. This observation was in agreement with studies on chymotrypsin from rainbow trout (Kristjansson and Nielsen, 1992), Atlantic cod (Asgeirsson and Bjarnason, 1991) and grass carp (Fong et al., 1998). Initial N-terminal sequencing by Edman degradation in the presence of β-mercaptoethanol revealed that native Atlantic cod chymotrypsin consist of two polypeptide chains (Raae et al., 1995), of which chain A was made up of 13 amino acid residues and chain B, 230 amino acid residues, in accordance with previous studies from other fish.

2.5.3 Effect of pH of fish chymotrypsin

Chymotrypsins from fish are mostly stable in the alkaline pH range. Yang et al., (2009), observed maximum activity for chymotrypsins A and B from crucian carp at pH 7.5 and 8.0**,** respectively. Both chymotrypsin A and B from crucian carp showed high stability over a broad

pH range of 6.0-11.0 (Yang et al., 2009). Similar pH stability was obtained for two trout chymotrypsins, with optimal pH values around 9.0 (Kristjansson and Nielsen, 1992). Chymotrypsin isolated from Atlantic cod also had an optimum pH of 7.8 for the hydrolysis of BTEE (Asgeirsson and Bjarnasson, 1991). Anchovy chymotrypsin showed an optimum pH of 8.0; while pH optima of chymotrypsins from less related fish were around 9.0 (Heu et al., 1995; Kristjansson and Nielsen, 1992). Monterey sardine chymotrypsin activity showed high relative activity in the pH range from 8 to 10 (Castillo-Yanez et al., 2006) which was dependent on factors such as substrate, substrate concentration, ionic strength and temperature. Loss in activity of chymotrypsin at low pH has been reported. For example, there was a considerable loss in activities of both chymotrypsins A and B in the crucian carp when pH was reduced to 5.0 (Yang et al., 2009). Atlantic cod chymotrypsin also displayed marked instability at pH values < 5.0 (Asgeirsson and Bjarnasson, 1991). Similar observations were made for Monterey sardine chymotrypsin (Castillo-Yanez et al., 2006), rainbow trout chymotrypsin (Kristjansson and Nielsen, 1992) and common carp chymotrypsin (Cohen et al., 1981). The observed reduced stability of the enzyme corresponds with a change of their net charge that occurs when the enzymes are at pH below their isoelectric points, which affects their tertiary structures (Castillo-Yanez et al., 2006)**.**

2.5.4 Effect of temperature on fish chymotrypsin

Temperature stability of chymotrypsin has been studied extensively. Yang et al. (2009) found the maximal activity of chymotrypsin A and B from crucian carp to be 40ºC and 50ºC, respectively, even though they exhibited some activity at lower temperatures. Chymotrypsin A from crucian carp maintained > 60% of its activity after 30 min incubation at 20°C, while chymotrypsin B maintained only 27% of activity (Yang et al., 2009). Similar observations were made by Heu et al. (1995), who found maximal activity of anchovy chymotrypsin at 45ºC. These results were comparable to trout chymotrypsin which increased in activity up to about 55ºC (Kristjansson and Nielsen, 1992) and Monterey sardine which showed an optimum temperature for activity at about 50ºC (Castillo –Yanez et al., 2006). Differences in optimal temperatures of chymotrypsin from different fish species might be related to their preferred environments (Yang et al., 2009).

The thermal stability of chymotrypsin has also been well elucidated. Studies by Asgeirsson and Bjarnasson (1991) showed that Atlantic cod chymotrypsin displayed a significantly lower

tolerance towards heat and that the loss of half-maximal activity occurred at 52ºC. At temperatures above the optimum, the activity of the two chymotrypsin isozymes were affected, suggesting different stabilities of the two enzymes (Asgeirsson and Bjarnasson, 1991). In a similar study, Kristjansson and Nielsen (1992) observed that chymotrypsin A from rainbow trout was more thermally stable that it's variant chymotrypsin B at 40°C and above. For crucian carp chymotrypsins, similar activity profiles were observed for both enzymes at 37°C, with relative activities decreasing sharply above 45°C. This result was similar to chymotrypsins from cod (Raae and Walther, 1989) and dogfish (Racicot and Hultin, 1987). Unlike results from Asgeirsson and Bjarnasson (1991), who found a significant difference in the stability of the two chymotrypsin isozymes from rainbow trout at temperatures above 40°C, Yang et al. (2009) only observed slight and insignificant changes in both chymotrypsins above 45°C. Differential thermal stabilities of chymotrypsin isozymes are attributed to fewer disulphide linkages and decreased hydrophobic interactions in the interior of the less thermostable isozyme (Simpson and Haard, 1984).

2.5.5 Effect of metal ions and protease inhibitors

The enhancing effect of calcium on the activity of chymotrypsin has been studied by several researchers. Chymotrypsin binds strongly to one calcium ion which makes it more stable against denaturation (Delaage et al., 1968). Lakowski (1955) observed that calcium exerted a favorable effect on stability of chymotrypsins, with chymotrypsin A more stable in the presence of calcium. Yang et al. (2009) found that both chymotrypsins A and B from crucian carp were activated slightly by calcium and magnesium at concentrations of 5 mM. On the other hand, the enzymes were strongly inhibited by iron and copper and partially by manganese and cadmium.

Chymotrypsin from crucian carp is strongly inhibited by serine protease inhibitors such as phenyl methane sulfonyl fluoride (PMSF), pefabloc SC and chymostatin (Yang et al., 2009). Similar observations were made when cod chymotrypsin was treated with N-tosyl-L-phenylalanylchloromethane (TosPheCH_2Cl), PMSF and chymostatin (Asgeirsson and Bjarnasson, 1991). Monterey sardine chymotrypsin activity was almost completely inhibited by PMSF and soybean trypsin inhibitor (SBTI) (Castillo–Yanez et al., 2006). Rainbow trout chymotrypsin was also inhibited by standard serine protease inhibitors (Kristjansson and Nielsen, 1992). Aprotinin caused about 86% loss in activity of rainbow trout chymotrypsin I but

apparently increased the activity of chymotrypsin II (Kristjansson and Nielsen, 1992). However, cod chymotrypsin was, considerably less sensitive to, inhibition by aprotinin (Asgeirsson and Bjarnasson, 1991). Yang et al. (2009) reported inhibition of carp chymotrypsin by the chymotrypsin specific inhibitor N-tosyl-L-phenylalanylchloroketone (TPCK), with a minimal effect by the metallo-protease inactivator EDTA and the trypsin specific inhibitor benzamidine. Similar inhibition patterns were observed with Monterey sardine (Castillo –Yanez et al., 2006), Atlantic cod (Asgeirsson and Bjarnason, 1991), anchovy (Heu et al., 1995) and rainbow trout chymotrypsins (Kristjansson and Nielsen, 1992).

2.5.6 Substrate specificity and kinetic parameters of chymotrypsin

In theory, the affinity of an enzyme for its substrate has evolved to attain optimal functioning within the concentration range of the substrate found *in vivo* (Copeland, 2000). Chymotrypsin-catalyzed hydrolysis is a three-step process in which an enzyme-substrate complex and an acyl-enzyme intermediate are formed (Schellenberger et al., 1991). Specificity of chymotrypsin is largely determined by the binding and acylation step (Hedstrom, 2002). When evaluating the kinetics of serine protease reactions, including chymotrypsin, the Michaelis-Menten parameters k_{cat}, K_m, and k_{cat}/K_m are composites of the rate constants (Hedstrom, 2002). According to Schellenberger et al. (1991), the hydrolysis of low molecular weight substrates by chymotrypsin usually obeys Michaelis-Menten kinetics, but the kinetic constants k_{cat} and K_m are not necessarily related to individual rate constants of the process. Of all the kinetic parameters, k_{cat}/K_m most accurately reflects the substrate specificity of chymotrypsin activity (Brot and Bender, 1969). Chymotrypsin hydrolyzes peptides with values of k_{cat}/K_m of approximately 10^7 M^{-1} s^{-1}, and k_{cat} approximately 100 s^{-1} (Hedstrom, 2002). Yang et al. (2009) studied the kinetic constants K_m and k_{cat} of chymotrypsin from crucian carp based on Lineweaver-Burk plots and found K_m of chymotrypsin A and B to be 1.4 μM and 0.5 μM, respectively. Many methods can be used estimate these kinetic parameters, such as the Michaelis-Menten plot, Lineweaver-Burk plot, Hanes-Woolf plot, Eadie-Hoftsee plot, although the most commonly used are Lineweaver-Burk and Hanes-Woolf plots.

2.5.6.1 Michaelis-Menten plot

The Michaelis-Menten plot is based on the assumptions that the amount of enzyme does not change during the reaction and hence the total enzyme concentration $[E]_{total} = [E]_{free} + [ES]$, and that the affinity of an enzyme for its substrate is independent of the concentration of enzyme and substrate. It uses these assumptions and the steady state equation to derive an equation for velocity as a function of substrate concentration [S]. This important equation is known as the Michaelis-Menten equation.

$$v = \frac{Vmax[S]}{Km+[S]} \quad \text{(Michaelis-Menten, 1913)}$$

The Michaelis-Menten equation is applicable to most enzymes (except allosteric enzymes) and can be used to derive the Michaelis constant (K_m) and maximal velocity or rate (V_{max}) from an enzyme kinetic assay using varying concentrations of substrate (Simpson, 2012). The K_m gives an estimate of an enzyme's affinity for its substrate and also indicates the substrate concentration at which half of the enzyme's active site is filled with substrate, hence the lower the K_m the stronger the affinity and vice versa. Using the Michaelis-Menten approach to determine K_m and V_{max} values accurately is challenging, especially for multi-substrate enzymes and enzyme-inhibitor interactions, leading to inaccurate values due to the non-linear nature (hyperbolic curve) of the graph. To overcome these inaccuracies, various modifications of the Michaelis-Menten equation have been developed to find K_m and V_{max}, such as Lineweaver-Burk and Hanes-Woolf plots

2.5.6.2 Lineweaver-Burk plots

Hans Lineweaver and Dean Burk derived a linear form of the Michaelis-Menten equation which is a double reciprocal plot (inverting both sides of the Michaelis-Menten equation) and called the Lineweaver-Burk plot in 1934. This plot is useful in estimating values of K_m and V_{max} more accurately, especially for multi–substrate enzymes and for studying the interaction between enzymes and inhibitors (Simpson, 2012). The equation reveals that a plot of $^1/v$ versus $^1/_{[S]}$ has a slope of K_m/V_{max} and a y-intercept of $^1/V_{max}$. Further examination shows that the x-intercept = $-^1/K_m$. The Lineweaver-Burk plot is very useful for illustrative purposes**;** however**,** it has been

shown to yield uncharacteristic values of K_m and V_{max} at low substrate concentrations. This problem can be alleviated by preparing an evenly spaced series of substrate concentrations along the $^1/_{[S]}$ axis (Copeland, 2000).

2.5.6.3 Hanes-Woolf plots

Hanes-Woolf plots are obtained from Lineweaver-Burk plots by multiplying both sides of the Lineweaver-Burk equation by [S]. The resulting equation gives rise to a slope equal to $^1/V_{max}$ and y intercept equal to $-K_m$ when $^{[S]}/v$ is plotted against [S]. The plot has the advantage of yielding kinetic values without distorted data (Hanes, 1932). This plot has been shown to be more accurate than the Lineweaver-Burk plot (Simpson, 2012)

2.5.6.4 Eadie Hoftsee, plots

The Eadie Hoftsee plot is another linearized plot obtained from the Michaelis-Menten equation through rearrangement by multiplying both sides of this equation by K_m + [S]**,** and dividing both sides of the equation by [S]. A plot of v against $^v/_{[S]}$ yields a slope which is equivalent to $-K_m$ and a y intercept equivalent to v. It has the advantages of decompressing high substrate concentrations and gives the opportunity to observe a range of values for v from zero to V_{max} (Cornish-Bowden, 2013). This plot is, however, prone to error (Simpson, 2012)

2.6 Cloning and expression of proteins

Genetic manipulation using recombinant DNA techniques has played a central role in the large-scale production of proteins from different sources as well as in protein engineering. Most common methods employed in large scale protein production involve isolation of total RNA from specific tissues, synthesis of complementary DNA (cDNA), amplification of the target gene using the polymerase chain reaction using specific primers, insertion of the amplified gene into a plasmid/vector and expression of this gene in bacteria, yeast or other expression systems. The expression of a recombinant protein which does not naturally occur in the host cell is termed heterologous protein production (Walsh, 2002) as opposed to homologous protein production, in which the host cell produces a protein found naturally in its cell.

2.6.1 Isolation of cDNA

cDNA is double-stranded DNA that is derived from mRNA obtained from prokaryotes or eukaryotes (Prescot et al., 2004). An mRNA population isolated from a specific developmental stage should contain mRNAs specific for any protein expressed during that stage**,** making it possible for the isolation and further studies of the gene of interest (Bainbridge, 2000). A DNA copy of isolated mRNA can be generated by the use of reverse transcriptase, an enzyme that can convert mRNA into complementary DNA (cDNA) (Bainbridge, 2000). The second DNA strand is generated by DNA polymerase and the double-stranded product can be introduced into an appropriate vector for cloning and eventual expression of the encoded protein**.** Amplification of a specific cDNA of interest is usually achieved by the polymerase chain reaction (PCR) (Prescot et al., 2004).

2.6.2 Cloning vectors

Cloning vectors are very important tools for genetic engineering, allowing modification of an organism through the introduction of new or modified proteins which can lead to improvements of some inherent properties through the introduction of desirable characteristics or removal of unwanted traits. A cloning vector is a DNA molecule that carries foreign DNA into a host cell, replicates inside that cell and produces many copies of itself and the foreign DNA (Prescot et al., 2004). A typical cloning vector must have a sequence that permits propagation of itself in a host cell, cloning sites to insert foreign DNA which can be cut by restriction enzymes, and a way of selection, usually accomplished by selectable markers for drug resistance or correction of nutritional auxotrophies. Different types of cloning vectors are used for different types of cloning experiments. The choice of a vector is usually based on size and type of DNA to be cloned. Common vectors include plasmids, bacteriophages, cosmids, bacterial artificial chromosomes (BAC) and yeast artificial chromosomes (YAC).

2.6.2.1 Plasmids

A plasmid is an extra-chromosomal circular DNA molecule that autonomously replicates inside bacterial cells. Plasmid vectors are used to clone DNA ranging in size up to several thousand base pairs (100 bp -10 kb) (Bainbridge, 2000). They are introduced into the host cell by first ligating the desired DNA fragment into the plasmid at the cloning site (Figure 2.1) followed by

transformation of competent cells. Plasmids have the advantage of been easily isolated and manipulated due to their small size. They are also stable and replicate independently of the host cell (Prescot et al., 2004). Since several copies of plasmids are present in one cell, replication is facilitated. One most important factor which makes plasmids advantageous is the fact that they typically encode genes for proteins that endow resistance to an antibiotic (Prescot et al., 2004). On the other hand, plasmids are limited in their use because they cannot accept large fragments of DNA (Casali and Preston, 2003).

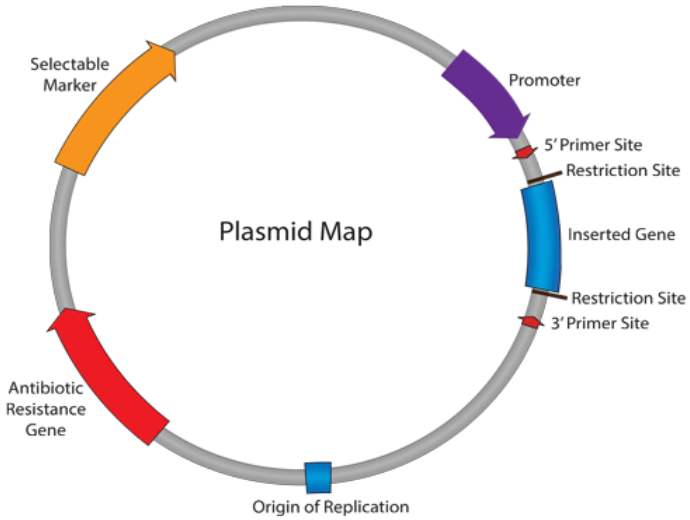

Figure 2.1: Structure of plasmid (source: www.addgene.org)

2.6.2.2 Bacteriophage (phage)

A phage is a virus particle consisting of a head (also known as a capsid) which contains the phage's double-stranded circular DNA (Figure 2.2). The most common bacteriophage vector is the phage lambda. Its structure consists of a head, tail and tail fibers (Dale, 2004). The lambda viral genome consist of a 48.5 kb linear DNA with a 12 base single stranded DNA (ssDNA) "sticky end" at both ends; these ends are complementary in sequence and can hybridize to each other (Casali and Preston, 2003)

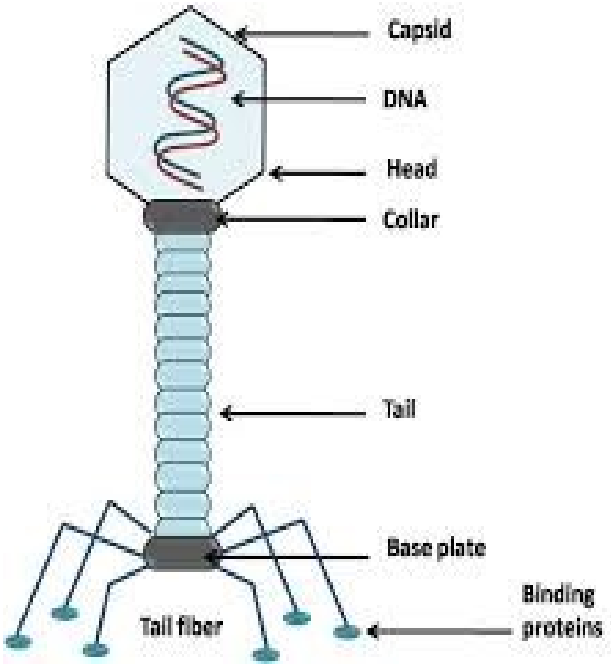

Figure 2.2: structure of bacteriophage (Source: nptel.ac.in/courses)

These complementary ends are usually referred to as the 'cos site' meaning cohesive ends (Figure 2.3). COS is a recognition sequence that allows the DNA to be packaged into infectious phage particles (Prescot et al., 2004). Lambda tail fibres normally adsorb to a cell surface receptor and the tail contracts, causing DNA to be injected into the host cell. The DNA circularizes at the cos site, and lambda begins its life cycle in the host. The cloning limit for the phage lambda is 8-20 kbp (Bainbridge, 2000).

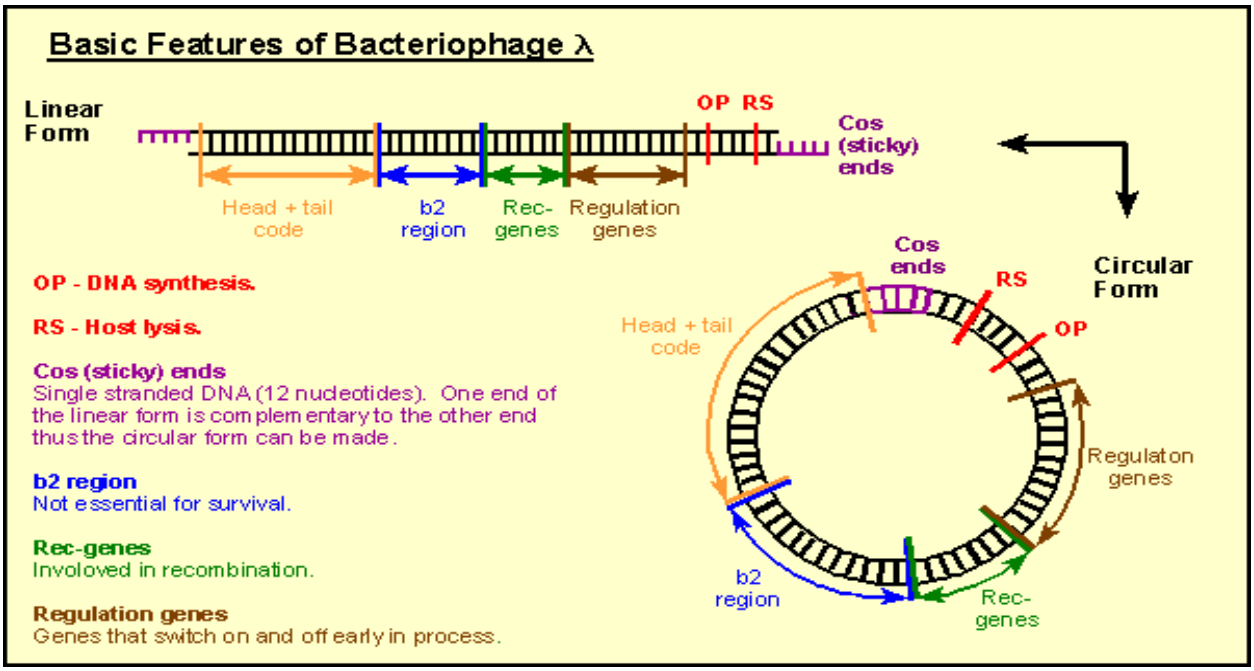

Figure 2.3: Bacteriophage lambda (Image source: http://www.bio.davidson.edu)

2.6.2.3 Cosmids

Cosmids are extrachromosomal circular DNA molecules that combine features of plasmids and phages (Casali and Preston, 2003). They are usually described as plasmids with one or two lambda cos sites (Prescot et al., 2004). Presence of the cos site permits *in vitro* packaging of cosmid DNA into lambda particles. Unlike plasmids or phages, cosmids can be used to clone DNA between 35-45 kbp (Prescot et al., 2004).

2.6.2.4 Bacterial Artificial Chromosomes (BAC)

Bacterial Artificial Chromosomes are engineered chromosomes based on a functional fertility plasmid. They are similar to bacterial plasmid vectors except for the presence of genes coding for proteins that replicate in very large natural plasmid called F-factor (Casali and Preston, 2003). They are usually cloned as a plasmid in a bacterial host, and its natural stability generally permits cloning of large pieces of insert DNA. BACs can be used to clone fragments as long as 300 kbp (Casali and Preston, 2003).

2.6.2.5 Yeast Artificial Chromosomes (YAC)

This is an artificial chromosome (genetically engineered) that contains telomeres, an origin of replication, a yeast centromere, and a selectable marker for identification in yeast cells. It can be used to clone DNA fragments as large as 1000 kbp (Casali and Preston, 2003). YACs represent good tools for the study of eukaryotic genomes and for mobilization of large genetic elements among bacteria and eukaryotes.

2.6.3 Gene expression systems

An expression system refers to factors (expression vector, DNA or insert, and host) that work together to yield a particular gene product. Most homologous proteins are expressed in bacteria, yeast, insect cells or mammalian cells (Walsh, 2002). The type of expression system used is dependent on the vector and the nature of the expected product, for example, if the protein needs to go through post-translational modifications or not.

2.6.3.1 Bacteria expression systems

By far the most common bacterial species used in this regard is *E. coli* (Walsh, 2002). This is because most prokaryotic genetic studies used *E. coli* as a model system, leading to a wealth of information on the genetic characteristics of *E. coli*. Also, suitable plasmids as well as potent inducible promoters are readily available to express foreign genes in *E. coli* (Walsh, 2002). However, most heterologous genes expressed in *E. coli* are produced intracellularly and accumulate in the cell cytoplasm, sometimes making downstream processing of the protein more complicated (Casali and Preston, 2003). Another limitation of using *E. coli* is the formation of insoluble aggregates known as inclusion bodies, leading to proteins which are not in their native form but are partially folded. Furthermore, *E. coli* cannot carry out post translational modification of proteins such as glycosylation, amidation and acetylation (Prescot et al., 2004). Other bacterial systems used to produce heterologous proteins are lactic acid bacteria and *Corynebacterium*. Some species in the Bacillus genus are also used (Prescot et al., 2004). These groups of organisms have the advantage of secreting proteins into extracellular fluids, making downstream processing simple; however, high quantities of exogenous proteins can be produced that can destroy the homologous protein (Walsh, 2002).

2.6.3.2 Yeast expression systems

Yeasts are very important platforms for the production of heterologous proteins, especially in the industry. Yeast genetics have been widely studied and characterized making it easier to work with (Walsh, 2002), and yeast fermentation technology is well established. Most common yeasts used in this regard include *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Kluyveromyces lactis*, *Pichia pastoris*, *Hansenula polymorpha*, and *Yarrowia lipolitica*. Yeasts have the ability to carry out post translational modifications of heterologous proteins, making it superior to most prokaryotic expression systems in this regard (Tatsumi et al., 1989). On the other hand, yeast expresses heterologous proteins at relatively low levels. Though post translational changes can occur in yeast, the end product sometimes differs significantly from end-products achieved by other expression systems such as insect or animal cells (Walsh 2002).

2.6.4 Cloning and expression of fish/invertebrate enzymes in microorganisms

For at least the the last decade, researchers have been producing enzymes from marine organisms on large scale using recombinant DNA technology (Walsh, 2002). Different cloning and expression systems have been used, ranging from bacteria to yeasts, for cloning cDNAs encoding various fish enzymes.

2.6.4.1 Chymotrypsin/chymotrypsin-like enzymes

Despite much progress in recent years, there is still a lot to learn about this group of proteases. Molecular biology tools have helped in the generation of new information about its primary and tertiary structures and mode of action**.** Chymotrypsin and chymotrypsin-like enzymes from fish and some invertebrate organisms, are of interest due to their special properties as model enzymes and potential applications in industry. Psochiou et al. (2007) cloned and expressed chymotrypsinogen from *S. aurata* (sea bream) using a lambda vector and an *E. coli* expression system. In a similar way, the cut worm (*S. litura*) chymotrypsin-like protease was cloned using a plasmid and an *E. coli* expression system (Zhang et al., 2010). A chymotrypsin-like enzyme was also cloned from *Metarhizium anisopliae* using the lambda ZAP vector and expressed in *P. pastoris* by Screen and Leger (2000). A similar study on *M. anisopliae* chymotrypsin was conducted by Volante et al. (2011), who instead used a plasmid and expressed the gene in *E.*

coli. Jiang et al. (1997) cloned and expressed a chymotrypsin-like enzyme from *Aedes aegypti* (mosquito) using a plasmid vector and an *E. coli* expression system.

2.6.4.2 Trypsin and trypsin-like enzymes

Trypsin is one of the most widely characterized serine proteases and much effort has been put into their production using recombinant techniques. Male et al. (1995) used cloning techniques to describe the nucleotide sequences and the structural models of five trypsin variants from Atlantic salmon. In a similar way, Klein et al. (1996) isolated and studied cDNA of the 5 trypsin isoforms from *Penaeus vannamei*. Yang et al. (2010) cloned trypsin from *Helicoverpa. armigera* using the *E. coli* expression system.

2.7 Finding genes in novel genomes

A genome is the complete set of DNA molecules in an organism, including all the genes needed to build and maintain that organism. Genomic sequencing has become very popular and much cheaper recently due to improvements in instrumentation coupled with the development of high performance computing and bioinformatics. For the past 20 years, Sanger sequencing and fluorescence based electrophoresis technologies have been extensively used in somatic and germline genetic studies (Reis-Filho, 2009). However, with the evolution of this technology occurring at an unprecedented pace, there has been a shift from traditional methods to microarrays and massively parallel sequencing (also known as next generation sequencing), which allows for the buildup of qualitative and quantitative information about any type of nucleic acid in a given sample at an incredible throughput (Korf, 2004). Next-generation sequencing refers to non-Sanger-based high-throughput DNA sequencing technologies in which millions or billions of DNA strands can be sequenced in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that are often used in Sanger sequencing of genomes**.** Sequencing throughput provided by this technology is attained at relatively low cost compared with traditional sequencing methods. Next-generation sequencing generates much shorter reads (~21 to ~400 bp), but millions of them compared to long reads generated from PCR-amplified samples (Stratton et al., 2009; Morozova and Marra, 2008). The limitation of short read lengths is that it makes *de novo* sequence assembly more difficult and less complete, particularly for novel genomes or massively repetitive and rearranged DNA

segments (Tucker et al., 2009). In addition to the ability to sequence DNA, parallel sequencing can be applied to sequencing RNA (Wang et al., 2009) and can also be applied to paired-end RNA sequencing, and small and noncoding RNA sequencing (Reis-Filho, 2009).

Table 2.1: Summary of parallel sequencing methods

Method	Amplification approach	Read length	Sequencing chemistry
Illumina	Bridge PCR	75+ bp	Reverse terminator
ABI 3730x1	PCR	900-1100 bp	Sanger method
454 FLX Roche	emulsion PCR	400 bp	pyrosequencing
SOLiD	emulsion PCR	50 bp	ligation-based
Helicos Heliscope	None (single-molecule sequencing)	30 - 35	polymerase-based
Pacific Biosciences*	None	>1000	Single molecule real time
ZS Genetics *	None		ZSG atomic labelling and electron spray

Source (Reis-Filho, 2009; Genet, 2009) *Technologies still under development

Once sequencing is achieved, the output needs to be assembled and annotated to determine structure of protein-coding genes (Korf, 2004). Various approaches have been used to accurately predict genes. In the past, the process mostly involved time- consuming experimentation on living cells. With recent advances in computational biology and software, this limitation has largely been overcome.

2.7.1 Genome assembly

A genome assembly may be defined as the process of stitching together an organism's chromosomes from fragmented reads of DNA. Genome assembly is usually achieved computationally, using programs that compile data consisting of short sequenced fragments known as single reads. These reads are joined through overlapping regions into a continuous

sequence known as a 'contig'. Factors such as repetitive sequences, polymorphisms, missing data and mistakes have been shown to limit the length of the contigs that assemblers can build. Contigs can be oriented and ordered and linked to each other to form scaffolds (Baker, 2012). In the absence of a high-quality reference genome, new genome assemblies are usually evaluated on the basis of the number of scaffolds and contigs required to represent the genome, the proportion of reads that can be assembled, the absolute length of contigs and scaffolds, and the length of contigs and scaffolds relative to the size of the genome (Yandell and Ence, 2012). To verify the completeness and contiguity of an assembly, several summary statistics are used; by far the most important is N50; the longer the N50 value, the better the assembly. The average gap size of a scaffold and the average number of gaps per scaffold can also be used to verify the completeness of an assembly (Yandell and Ence, 2012).

2.7.2 Gene annotation

Gene annotation usually describes two distinct processes; ‘structural’ genome annotation, the process of identifying protein-coding genes (Korf, 2004) and ‘functional’ genome annotation , the process of attaching meta-data such as gene ontology terms to structural annotations (Yandell and Ence, 2012). Many structural gene findings rely on current understanding of cellular biochemical processes such as transcription, translation, protein-protein interactions and various cell regulation processes. Computational approaches and databases are used in this regard and the accuracy of a gene finder depends on many factors, most significantly proper training, which is often laborious (Korf, 2004). The annotation process is intrinsically complicated and involves many different tools generally referred to as annotation pipelines (Yandell and Ence, 2012). These pipelines differ in their details, but share a core set of features. (Treangen and Salzberg, 2012)

2.7.2.1 The computational phase of gene annotation

The first step in computation is usually repeat identification and masking. Repeat identification helps in the detection of ‘low-complexity’ transposable (mobile) elements, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs) (Kapitonov and Jurka, 2008). Identifying repeats is complicated by the fact that repeats are often poorly conserved; thus, accurate repeat detection usually requires users to create a repeat library for

their genome of interest (Yandell and Ence, 2012). 'Masking' on the other hand simply means transforming every nucleotide identified as a repeat to any base 'N' or, in some cases, to a lower case a, t, g or c (Majoros, 2007). Left unmasked, repeats can seed millions of spurious Basic Local Alignment Search Tool (BLAST) alignments producing false evidence for gene annotations (Korf, 2004). The next step after masking usually involves alignment of identified transcripts and proteins from the organism whose genome is being annotated to provide evidence of proper annotation. In principle, UniProtKB/SwissProt databases provide an excellent core resource for protein sequences (UniProt Consortium, 2011) supplemented by the National Center for Biotechnology Information (NCBI) taxonomy browser (Sayers et al., 2009), while TBLASTX can be used to align ESTs and RNA-seq data from phylogenetically distant organisms (Camacho et al., 2009).

The resulting alignments are usually filtered to identify and to remove marginal alignments on the basis of metrics such as percent similarity or percent identity (Yandell and Ence, 2012). Splice-site-aware alignment algorithms, such as Splign and Spidey, can be used to realign matching and highly similar ESTs, mRNAs and proteins to genomic sequences (Kapustin et al., 2008). RNA-seq data has been shown to have the greatest potential to improve the accuracy of gene annotations and its use currently lies at the cutting edge of genome annotation, and the available toolset is evolving quickly (Garber et al., 2011). Currently, RNA-seq reads are usually handled in two ways. The first is a *de novo* assembly which is independent of the genome using tools such as Assembly by Short Sequences (ABySS), a parallel, paired-end sequence assembler (Simpson, 2009) and Short Oligonucleotide Analysis Package (SOAPdenovo) (Li et al., 2010). Alternatively, the RNA-seq data can be directly aligned to a genome using tools such as TopHat, GSNAP or Scripture (Yandell and Ence, 2012).

2.7.3 Ab initio gene prediction

Ab initio gene prediction is an intrinsic method that uses mathematical models rather than external evidence to identify genes (Yandell and Ence, 2012). This approach uses organism-specific genomic traits, such as codon frequencies and distributions of intron–exon lengths, to distinguish genes from intergenic regions and to determine intron–exon structures, and hence, requires extensive data training (Korf, 2004). Given enough training data, the gene-level sensitivity of ab initio tools can approach 100% (Coghlan et al., 2008). To achieve highly

accurate gene predictions with the ab initio method, large numbers of preexisting, high-quality gene models or near base perfect genome assemblies are usually required (Reese and Guigo, 2006). In the absence of preexisting gene models, however, alignments of ESTs, RNA-seq and protein sequences to a genome can be used to train gene predictors (Yandell and Ence, 2012).

2.7.3 Evidence-driven gene prediction

This method of gene prediction uses external evidence to improve the accuracy of predictions and has great potential to improve the quality of gene prediction in newly sequenced genomes (Howe et al., 2002). The process typically involves firstly aligning ESTs and proteins to the genome; followed by RNA-seq data alignment (if available). Splice sites must then be identified, and the assembled evidence must be post-processed before a synopsis of these data can be passed to the gene finder (Yandell and Ence, 2012). This process involves a lot of time and a lot of work in order to produce reasonable data.

2.7.4 The annotation phase

In the past, annotation was done manually by reviewing evidence for each gene to decide on their intron–exon structures (Souvorov et al., 2010). Though laborious and time consuming, the process resulted in high-quality annotation (Yandell et al., 2005). Because of time and budgetary reasons, genome projects are increasingly being forced to rely on automated annotations (Yandell and Ence, 2012). Various strategies are available for creating automated annotations. The simplest form of automated annotation is to run a battery of different gene finders on the genome and then to use a 'chooser algorithm' (also known as a 'combiner') to select the single prediction whose intron–exon structure best represents the consensus of models from among overlapping predictions that define each putative gene locus (Haas et al., 2008; Li et al., 2008). Another popular approach is to feed the alignment evidence to the gene predictors at run time to improve the accuracy of the prediction process (Yandell and Ence, 2012). A chooser algorithm can also be added in order to attain greater accuracies especially with RNA-seq and EST data. A full run by an annotation pipeline can take weeks, but because these pipelines align evidence to the genome, their outputs provide starting points for annotation, curation and downstream analyses, such as differential expression analyses using RNA-seq data (Yandell and Ence, 2012).

Once annotation is complete, output data which include the transcript and protein sequences of every annotation can be visualized in FASTA format (Pearson and Lipman, 1988).

2.8 Amino acid sequence determination of proteins

The amino acid sequence of a protein provides a wealth of information about its function and evolutionary history. This can be determined directly through chemical sequencing or by physical fragmentation and analysis. The basis of chemical sequencing is Edman degradation, which involves sequential labeling followed by subsequent removal and identification of amino acid residues beginning at N terminal of the polypeptide (Walsh, 2002). Multipeptide proteins as well as any disulfide linkage must be broken by incubation with a suitable reducing agent prior to Edman sequencing. Larger polypeptides must also be fragmented either chemically or enzymatically and the fragments sequenced individually. Trypsin and chymotrypsin are most often used for this, while chemicals such as cyanogen bromide can also be used for fragmentation. The sequencing begins with incubation of the polypeptide with phenylisothiocyanate (PITC) which reacts with the uncharged terminal amino group of the peptide to form a phenylthiocarbamoyl derivative. Under mildly acidic conditions, a cyclic derivative of the terminal amino acid is liberated, which leaves an intact peptide shortened by one amino acid. The cyclic compound is a phenylthiohydantoin (PTH)-amino acid, which can be identified by chromatographic procedures or mass spectrometry (Walsh, 2002). The cycle is repeated until the amino acid sequence is determined. Samples must be free of contaminants such as Tris, glycine, SDS and acrylamide, since they have the potential to interfere with the sequencing machine and also clutter the chromatogram with large peaks.

Other methods discussed previously like mass spectrometry can also be used to elucidate the primary structure of proteins

2.9 Higher structure determination of protein/enzymes

Previous methods for 3-dimensional analysis of proteins employed X-ray diffraction. Currently methods, including nuclear magnetic resonance (NMR), electron microscopy and bioinformatics tools, can also be employed to study protein structure. Electron microscopy (EM) detects structural information at very low resolution, making it almost impossible to detect atomic

details of the protein structure though this is changing through cryo-EM. X-ray diffraction and NMR give very high resolution with detailed atomic structural information (Walsh, 2002).

2.9.1 X-ray diffraction

X-ray diffraction involves bombarding a crystalline protein sample with a beam of X-rays. Most of the rays pass straight through the sample while some are diffracted. The resultant diffraction pattern is recorded on a detector. This diffraction pattern is then analyzed by mathematical expression known as Fourier transformation, from which the 3D structure can be determined. The initial pre-requisite is the generation of protein crystals. Protein crystals suitable for X-ray diffraction must have a minimum diameter of 50 μm although larger diameters may be required for older equipment (Walsh, 2002). This method has been used to resolve the crystal structures of many proteins including bovine chymotrypsin (Birktoft and Blow, 1972), trypsin (Walter et al., 1982), and elastase (Thayer et al., 1992). Proteins are extremely large; exhibit irregular surfaces, contain solvent-filled channel or pores and may co-exist in solution in multiple conformations or states, creating the major drawbacks in this method.

2.9.2 Nuclear magnetic resonance (NMR)

NMR involves the application of a very strong magnetic field to a protein sample. It is based on the principle that a number of atomic nuclei, including 1H, 13C, 15N and 31P, display a magnetic moment due to their spinning-like movement about an axis. Since these spinning nuclei are positively charged, they act as tiny magnets and therefore can interact with an applied magnetic field. This interaction can be detected and measured. The exact frequency emitted by any nucleus is influenced by its molecular environment and can be used to provide the 3-dimensional structure of a protein. Reckel et al. (2011) resolved the 3D structure of proteorhodopsin in solution using NMR. Unlike X-ray crystallography, NMR may be used to determine the structure of proteins in solution (Walsh, 2002).

2.9.3 Cryo Electron microscopy (cryo-EM)

Cryo-EM involves imaging radiation-sensitive specimens in a transmission electron microscope under cryogenic conditions. In electron microscopy, the electrons, emitted by a source housed under a high vacuum, is accelerated down the microscope column at accelerating voltages of typically 80–300 kV. After passing through the specimen, scattered electrons are focused by the

electromagnetic lenses of the microscope to produce a highly resolved image. This method is becoming popular since microscopy is becoming a mainstream technology for studying cells, viruses and protein assemblies at molecular resolution. The possibility of imaging biological structures with electrons was demonstrated in 1975 with the determination of the structure of bacteriorhodopsin at ~ 7 Å resolution (Henderson and Unwin, 1975) and most recently membrane proteins such as aquaporin Gonen et al., 2005)

2.9.4 Bioinformatics tools

The use of bioinformatics tools for the prediction of 3-dimensional structure of a protein is based on methods such as threading. Threading consists of several steps centered on the idea that if an experimentally determined protein with known structure can be found that is similar to the protein of interest; this known structure's atomic coordinates can be used as a scaffold on which the new protein can be built. Many protein structures have been solved using bioinformatics tools, including cannabinoid receptor-2 (Diaz et al., 2009), human adenosine A2A receptor (Michielan et al., 2008) and alpha-1-adrenoreceptors (Li et al., 2008)

2.10 Industrial Application of chymotrypsin

Chymotrypsin has found widespread application in the production of valuable products for human material needs, including food, animal feed, pharmaceuticals, fine and bulk chemicals, fibers, hygiene, and environmental technology, as well as in a wide range of analytical purposes (Buchholz et al., 2005).

2.10.1 Food Industry

Chymotrypsin has been used to improve the nutritional value of proteins and to lower the protein denaturation temperature and cleavage specificity in certain foods (Haard, 1992). It is used in combination with other proteases for the tenderization of meat, fermentation, protein hydrolysate production and bone protein removal (Haard, 1998). Exogenous proteases from squid hepatopancreas, including chymotrypsin, for example**,** is used to ferment fish as it can accelerate the process and also yield products with superior sensory properties (Raksakulthai et al., 1986). In the dairy industry, chymotrypsin is used with trypsin to hydrolyze casein. (Kilara and Panyam, 2003; Korhonen and Pihlanto, 2006)

2.10.2 Medicine and pharmaceutical industry

Chymotrypsin can be administered by mouth, inhaled or as a shot for the relief of different symptoms (Natural Database, 2013). Generally, the primary uses of chymotrypsin are as a digestive aid and as an anti-inflammatory agent (Swamy and Patil, 2008). Chymotrypsin, along with other pancreatic enzymes, is most often used in the treatment of pancreatic insufficiency (Natural Database 2013). As an anti-inflammatory agent, chymotrypsin and other protease enzymes prevent tissue damage during inflammation and the formation of fibrin clots (Swamy and Patil, 2008). Chymotrypsin has also been reported to aid in the liquefaction of mucus secretions (Zhang, 2007). Alpha chymotrypsin is widely used to separate cataractous lens from the zonular attachment sites during cataract surgery (Rhee et al., 1999; Hill et al., 1960). Chymotrypsin decreases the manipulating force required during the surgery to make it easy to remove the cataract which guarantees higher success rates (O'Malley et al., 1961).

2.10.3 Detergent

Chymotrypsin is added to laundry detergent or dish detergents to enhance decontamination ability. The enzyme remains stable in both ionic and non-ionic surfactants and maintains around 80% of its catalytic ability after one hour of incubation with chemical detergents (Espósito et al., 2009). Due to the specificity of chymotrypsin for proteins and peptides, it can be used to break down proteinaceous contaminants (blood and foods) on cloth. Another advantage of using chymotrypsin is its ability to work under mild conditions such as low water temperature or natural pH environments which may lower the damage to cloth and body (Kamini et al., 1999).

CONNECTING STATEMENT 1

Chapter III investigates two novel chymotrypsin-like enzymes from the viscera of *Sepioteuthis lessoniana* and *Loligo opalescens*. The study addressed purification of the enzymes using affinity chromatography and also compares pH, temperature and kinetic parameters between these two enzymes. Finally, peptide fragments produced by tryptic digestion of the enzymes are elucidated using LC-MS/MS

The results of this study have been submitted to the Journal of Food Chemistry as:

Nana Akyaa Ackaah-Gyasi, Timothy Geary and Benjamin Simpson (2015). **Novel chymotrypsin-like enzymes from squid (*Sepioteuthis lessoniana* and *Loligo opalescens*) viscera; Purification, biochemical characterization and peptide identification using LC-MS/MS.**

Part of the results were also presented at the 17th World Congress of Food Science and Technology, August 17 – 20, 2014. Montreal, Canada.

Nana Akyaa Ackaah-Gyasi, Timothy Geary and Benjamin Simpson (2014). **Purification and biochemical characterization of chymotrypsin from squid viscera.**

CHAPTER III

Novel chymotrypsin-like enzymes from squid (*Sepioteuthis lessoniana and Loligo opalescens*) viscera: Purification, biochemical characterization and peptide identification using LC-MS/MS

3.1 Abstract:

Chymotrypsin-like enzymes were isolated and characterized from two species of squid viscera. The crude enzymes were extracted with 10 mM tris-HCl, 5 mM $CaCl_2$ and 20 mM NaCl (pH 8.0, 4°C, 30 min) and purified to homogeneity using ultrafiltration and affinity chromatography on phenyl-butyl-amine coupled to cyanogen bromide activated Sepharose 4B. The purified enzymes migrated as a single polypeptide chains with molecular mass of 22.0 ± 2.7 kDa and 18 ± 1.7 kDa by SDS-polyacrylamide gel electrophoresis. The enzymes showed temperature and pH optima between 25°C - 35°C and 7.5 – 8.5, respectively, using 1 mM benzoyl-L-tyrosine-ethyl-ester (BTEE) as substrate. *Loligo* chymotrypsin was stable at 10°C, retaining about 90% of its activity whiles *Sepioteuthis* retained ≈ 70% of its activity . The hydrolytic activities of the enzymes were inhibited totally by TPCK. The enzymes, however, were less sensitive to inhibition by chymostatin and EDTA. Ca^{2+} had greater activating effects (≈300%**)** on the enzymes than Mg^{2+} and Zn^{2+}, which both elicited ≈200% activation. Exposure to 50% isopropanol (v/v) completely inactivated the enzymes, compared with 67% inhibition in 50% ethanol (v/v). *Loligo opalescens* chymotrypsin was more active toward the ester substrate N-benzoyl-tyrosine ethyl ester. In-gel tryptic cleavage and peptide fragment analysis using LC MS/MS produced unique peptides that matched a serine protease from another cephalopod in the NCBI database. Observed differences in kinetic properties and temperature properties shows that the *Loligo opalescens* chymotrypsin-like enzyme is more adapted to cold temperatures and hence has the potential for industrial applications in which lower temperatures are desirable.

Keywords: chymotrypsin, enzyme purification, LC MS/MS, peptides, squid viscera

3.2 Introduction

Enzymes including chymotrypsin have found widespread use in different industrial applications (e.g., pharmaceuticals, food processing, detergents and leather industries) due to their catalytic properties and capacity to consistently produce uniform products (Buchholz et al., 2005; El Enshasy et al., 2008). Chymotrypsins have been purified and characterized from many different sources including mammals such as bovine (Balti et al., 2012), fish such as crucian carp (Yang et al., 2009) and Atlantic cod (Asgeirson and Bjarnason, 1991), mollusks such as cuttlefish (Balti et al., 2012), invertebrates such as cut worm (Zhang et al., 2010) and microorganisms such as *Metarhizium anisopliae* (Screen and Leger, 2000). However, enzymes have limited ability to adapt to extreme environmental changes (Simpson, 2000). This has led to a heightened interest in finding alternative sources of enzymes from organisms living in different environments that may have unique features better suited for specific industrial applications (Haard, 1992).

Research has shown that enzymes obtained from fish and other aquatic organisms exhibit several properties suitable for distinct industrial applications compared to those derived from mammalian sources (Haard, 1998). Some of these properties are attributed to the psychrophilic nature conferred by adaptation of the source organisms to cold temperature regimes. For example, fish or other aquatic source enzymes often function better at extreme environmental conditions of pressure, temperature, salinity and alkalinity compared to homologous enzymes from species acclimated to warmer environments (Simpson, 2000). This makes it attractive to study enzymes of aquatic origin, such as from squids, to identify homologues of currently used enzymes that have more favorable properties.

Squids are important marine invertebrates, and are members of marine ecosystems that serve as food for many fishes such as tuna and marine mammals such as seals. They in turn prey on other fishes and crustaceans and their preference for food is dependent on season (Guerra, 2006). Squids comprise one of the most economically important aquatic groups and together with cuttlefish and octopus account for ≥3% of the global capture of all aquatic species (FAO, 2013). Squids have high tolerance for environmental changes (Gauvrit et al., 1997). Various species have been shown to adapt to different environmental conditions especially with regard to temperature. Examples are the northern calamari, *Sepioteuthis lessoniana*, a neritic warm water-dwelling squid found in tropical regions of the Indian Ocean and the western Pacific Ocean and California market squid, *Loligo opalescens*, an Eastern Pacific and Atlantic Ocean species that

can be found as deep as 500 m. Very few studies have reported on chymotrypsin-like enzymes from marine invertebrates (Balti et al., 2012). The few comparative studies on chymotrypsin-like enzymes usually compare cold adapted species to warm adapted mammalian homologues. Information on enzymes from organisms in the same evolutionary family with different environmental adaptation properties, such as temperature, is therefore of interest. In this study, we explored biochemical properties of chymotrypsin obtained from two squid specie, and identified peptide components using LC MS/MS to identify homologous enzymes in the NCBI database.

3.3 Materials and methods

3.3.1 Materials

Squid samples were purchased frozen from OCN Imports, a local fish market in Montreal, QC, Canada. Samples were stored on ice and transported to the laboratory for enzyme isolation and characterization. Viscera were excised immediately upon arrival and frozen at -20°C. The frozen viscera were cut into pieces (1.5 cm x 1.5 cm) and refrozen in liquid nitrogen. The liquid nitrogen frozen viscera samples were blended into fine powder using a Warring blender and the powder was kept at -20ºC for further characterization. The purification experiments were replicated three times

3.3.2 Extraction of crude enzyme

All procedures were performed at 4ºC. The crude enzyme was extracted by mixing the powdered viscera sample with cold extraction buffer (10 mM Tris-HCl, pH 8.0 containing 5 mM $CaCI_2$, 20 mM NaCl) in a ratio of 1:1 (w/v) followed by centrifugation at 5000 g for 30 min. The supernatant was termed the crude enzyme extract.

3.3.3 Determination of enzyme activity

Chymotrypsin activity was assayed according to the modified method of Hummel (1959). The standard reaction mixture had a total volume of 3 ml, comprised of 2.4 ml extraction buffer and 0.5 ml 1 mM BTEE dissolved in 50% w/w methanol and distilled water. The reaction was initiated by adding 0.1 ml of enzyme extract. The production of benzoyl-tyrosine was measured

by monitoring increase in absorbance at 256 nm with a spectrophotometer (Beckman Coulter UV-Vis DU 800). Measurements were taken at 15 sec intervals over a period of 5 min. One unit (U) of chymotrypsin activity was defined as the amount of enzyme that releases 1 µmol benzoyl-tyrosine per min.

3.3.4 Purification of chymotrypsin

All centrifugation steps were conducted at 4°C. The crude enzyme extract was first ultrafiltered by centrifugation through a 30 kDa molecular weight cut-off (MWCO) filter membrane (Millipore Ontario, Canada) at 5000 g for 30 min. The filtrate was then centrifuged through a 10 kDa MWCO membrane (Millipore Ontario, Canada) at 5000 g for 30 min. The retentate was applied to an affinity column (1.5 x 10 cm) containing phenyl-butyl-amine coupled to cyanogen-bromide activated Sepharose 4B (Sigma St. Louis, MO, USA. The column was equilibrated with extraction buffer. After sample application, the column was thoroughly washed with the extraction buffer to remove unbound proteins. This was followed by elution with 1 mM HCl at a flow rate of 1.0 mL/min. Fractions of 3.0 mL were collected and those containing chymotrypsin activity were pooled. The purified enzyme was stored at -20 ºC and used for characterization studies.

3.3.5 Total protein determination

Protein concentration was determined by measuring the absorbance at 280 nm of the sample solution and also by the BCA method using bovine serum albumin (BSA) as standard. (Walter 1996)

3.3.6 Purity and molecular weight determination

SDS-PAGE was carried out by the method of Laemmli (1970), using a 4-15% mini gradient gel (Biorad, Mississauga, ON, Canada). The protein was dissolved in sample buffer comprised of 3.55 ml deionized water, 1.25 ml 0.5 M Tris-HCl pH 6.8, 2.5 ml glycerol, 2 ml 10% SDS (w/v) and 0.5 ml 5% bromophenol blue (w/v); for reducing conditions, samples were denatured with β-mercaptoethanol and heated for 5 min at 100°C. Samples were subjected to electrophoresis at 100 V for 45 min. Gels were stained with Coomassie Brilliant Blue R-250 for 60min. Destaining was done by successive washes in destaining solution comprised of 700 ml distilled water, 200 ml methanol and 100 ml glacial acetic acid.

3.3.7 pH stability and activity profile

The effect of pH on purified chymotrypsin activity was determined using 1 mM BTEE as substrate from pH range 4.0 - 11.0 at 25°C in the following buffer systems: sodium acetate buffer (0.1 M; pH 4.0 - 5.0), sodium phosphate buffer (0.1 M; pH 5.5 - 7.0), Tris-HCl buffer (0.1 M; pH 7.5 - 8.5), and carbonate-bicarbonate buffer (0.1 M; pH 9.5 - 11.0) (Yang et al., 2009). The effect of pH on chymotrypsin stability was evaluated by measuring the residual enzymatic activity at 25ºC after incubating the enzyme for 30 min at pH 4.0 – 11.0 at 25°C (Balti et al., 2012).

3.3.8 Temperature stability and activity profile

To establish a temperature activity profile, enzyme activity was assayed as described above at temperatures from 10 to 70°C over a 5 min interval (Balti et al., 2012). To measure the effects of temperature on enzyme stability, the enzyme was incubated at temperatures from 10 to 70ºC for 30 min and immediately cooled to 25°C under running tap. Residual activity was determined at pH 8.0 and 25°C over a 5 min interval using BTEE as substrate. The un-heated enzyme was used as the control (Balti et al., 2012).

3.3.9 Effect of inhibitors, solvents and metal ions

The effect of the enzyme inhibitors, chymostatin, N-α-p-tosyl-L-phenylalanine chloromethyl ketone (TPCK) and ethylenediaminetetraacetic acid (EDTA) (all from Sigma, St. Louis, MO, USA) on chymotrypsin activity was studied as described (Balti et al. (2012). The purified enzyme was incubated with inhibitors at various concentrations (Table 3.5) in 10 mM Tris–HCl buffer (pH 8.0) for 30 min at 25ºC. The remaining activity of the enzyme was measured as described above. The activity of the enzyme assayed in the absence of inhibitors served as control. The same procedure was followed to study the effect of isopropanol, ethanol and metal ions (Zn^{2+}, Ca^{2+} and Mg^{2+}, as the chloride salts) on enzyme activity

3.3.10 Kinetic studies

The kinetic parameters K_m, V_{max} and k_{cat} of the purified enzyme were determined as described by Yang et al., (2009). Purified squid chymotrypsin was prepared at a concentration of 30 µg/ml in extraction buffer and reacted with different concentrations of BTEE (0.5 mM – 3 mM) at 25ºC

for 5 min. The reaction was initiated by adding 0.1 ml of the enzyme solution to 2.4 ml buffer, 0.5 ml substrate and activity determined as described above. Kinetic parameters including maximum velocity (V_{max}), the apparent Michaelis–Menten constant (K_m) and the turn over number (k_{cat}) were evaluated based on Hanes plots (Ritchie and Prvan 1996)

3.3.11 Proteomics analysis

Sequencing of chymotrypsin isolated from squid viscera was conducted at the CHU Laval Quebec Proteomics platform in Canada and included in-gel digestion, mass spectrometry and database searching.

3.3.11.1 In-gel digestion and mass spectrometry

Single protein bands migrating with apparent molecular masses of 22 kDa for *S. lessoniana* and 18 kDa for *L. opalescens* were obtained after SDS-PAGE of the active fractions obtained by affinity chromatography. These bands were, excised from the gel, reduced with 10 mM DTT and alkylated with 55 mM iodoacetamide. The bands were individually digested by incubation with 105 mM modified porcine trypsin (Sequencing grade, Promega, Madison, WI) at 58˚C for 1 h in 25 mM NH_4HCO_3. The products of digestion were extracted with 1% formic acid/2% acetonitrile followed by 1% formic acid/50% acetonitrile. The extracts were concentrated under vacuum and re-suspended in 8 µl 0.1% formic acid and injected into a capillary HPLC system connected online to a nanospray ionization mass spectrometer (LTQ, ThermoFisher).

3.3.11.2 Database searching

MS/MS samples were analyzed using Mascot (Matrix Science, London, UK; version 2.4.1) and X-Tandem (The GPM, thegpm.org; version CYCLONE (2010.12.01.1)). Mascot was set up to search the UR14_2_Cephalopoda_6605_20140415 database (3068 entries) assuming the digestion enzyme trypsin. X-Tandem was set up to search a subset of the UR14_2_Cephalopoda_6605_20140415 database (only samples "Ingel19019r_MGFPeaklist (F054576) and "Ingel19020r_MGFPeaklist (F054575). Mascot and X-Tandem were searched with a fragment ion mass tolerance of 0.100 Da and a parent ion tolerance of 0.100 Da.

Carbamidomethyl of cysteine was specified in Mascot and X-Tandem as a fixed modification. Dehydration of the N-terminus, glu->pyro-Glu of the N-terminus, ammonia-loss of the N-terminus, Gln->pyro-Glu of the N-terminus, deamidation of asparagine and glutamine and oxidation of methionine were specified in X-Tandem as variable modifications. Glu->pyro-Glu of the N-terminus, Gln->pyro-Glu of the N-terminus, deamidation of asparagine and glutamine and oxidation of methionine were also specified in Mascot as variable modifications.

3.3.11.3 Criteria for protein identification

Scaffold (version Scaffold_4.3.2, Proteome Software Inc., Portland, OR) was used to validate MS/MS based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0% probability by the Peptide Prophet algorithm (Keller**,** et al., 2002) with Scaffold delta-mass correction. Protein identifications were accepted if they could be established at greater than 99.9% probability and contained at least 2 identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm (Nesvizhskii**,** 2003). Proteins that contained similar peptides and could not be differentiated based on MS/MS analysis alone were grouped to satisfy the principles of parsimony.

3.4 Results and discussion

3.4.1 Purification of Squid chymotrypsin

Purification of chymotrypsin from the two squid viscera was achieved through ultrafiltration and affinity chromatography on a PBA-Sepharose column and yielded identical results. Both enzymes eluted as single symmetric peaks with constant specific activity indicating homogeneity of the preparation (Figure 3.1). The enzymes ware purified about 300-fold, with minimum recovery of approximately 44% and specific activities between 273.25 and 280 U/mg protein (Table 3.1). The enzymes were fully active upon extraction and purification from the viscera. These findings are similar to those reported by Yang et al., (2009) for crucian carp and by Kristjansson and Nielsen (1992) for rainbow trout, but different from the findings of Fong et al. (1998) for grass carp chymotrypsin, which required activation of the zymogen by the addition of exogenous trypsin.

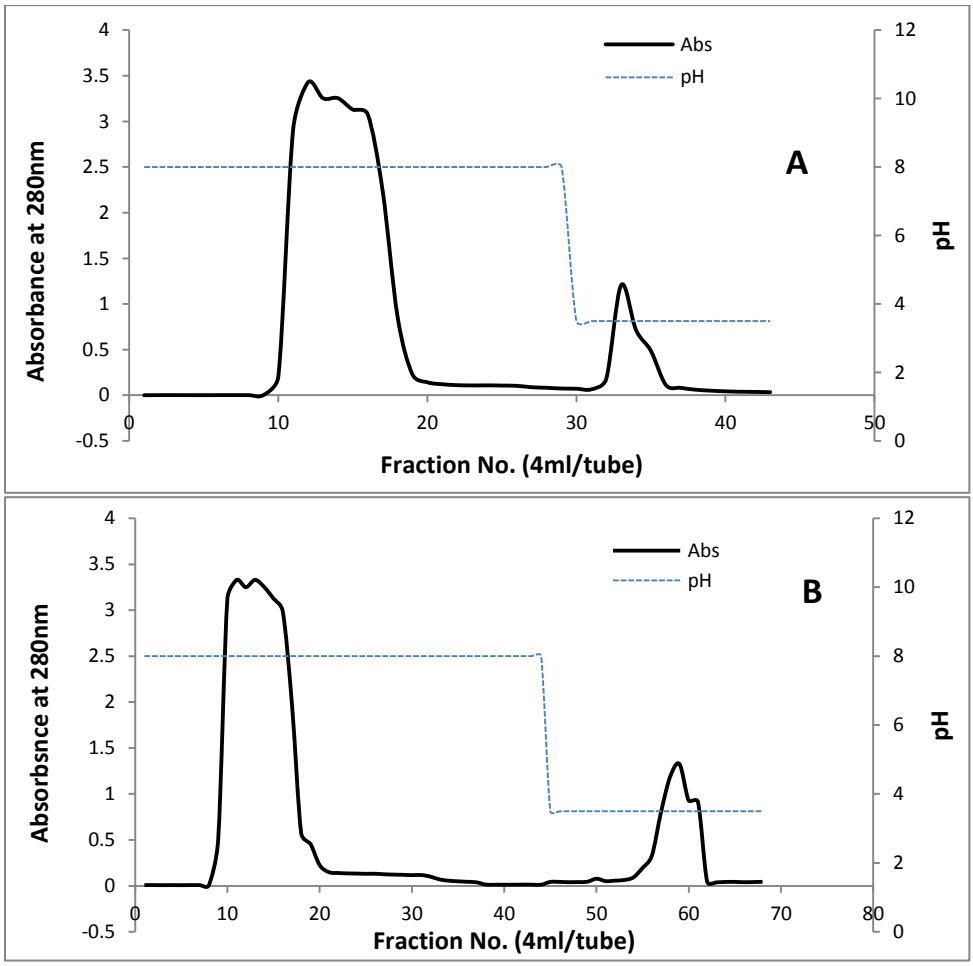

Figure 3.1: Absorption spectra indicating protein content in fractions collected during affinity chromatography. A: *Loligo opalescens* purification, B: *Sepioteuthis lessoniana* purification

Table 3.1: Summary of steps involved in the purification of chymotrypsin-like enzymes from squid viscera

Fraction	Protein (mg)	Protein concentration (mg/ml)	Total activity (Units)	Specific activity (U/mg)	Purity (fold)	Yield (%)
Sepioteuthis lessoniana						
Crude extract	713.15	35.66	597.62	0.84	1.00	100
Ultrafiltered fraction	121.32	12.13	354.8	2.92	3.48	59.37
Affinity fraction	0.98	0.16	267.79	273.25	326.07	44.81
Loligo opalescens						
Crude extract	700.2	32.97	612.50	0.87	1.00	100
Ultrafiltered fraction	122.2	14.2	372.31	3.05	3.51	60.79
Affinity fraction	1.08	0.32	281.43	260.58	299.52	45.95

3.4.2 Purity and molecular weight determination

SDS-PAGE of the purified enzymes gave single bands with estimated molecular weights of 22.0 ± 2.70 kDa for *S. lessoniana* and 18.0 ± 1.7 for *L. opalescens* (Figure. 3.2). Chymotrypsins have been reported by many authors to have molecular weights between 20 – 30 kDa depending on the source, including cod (27.0 kDa, Raae and Walther, 1989), anchovy (26.1 kDa, Heu et al., 1995), Japanese sea bass (27.0 and 27.5 kDa, Jiang et al., 2010), and cuttlefish (28.0 kDa, Balti et al., 2012). However, Houseman et al. (1987) reported a 13.8 kDa chymotrypsin from the barn fly (*Stomoxys calcitrans*). The large variation in size of chymotrypsins from different sources can be attributed to the differences in amino acid composition of the proteins and not to post-translational modifications. These differences in size usually do not affect the substrate binding of chymotrypsins since most enzymes have conserved amino acid domains required for binding and product formation (Petsko and Ringe, 2004). For chymotrypsin, irrespective of the source, the catalytic triad of serine, histidine and aspartate is conserved (Polgar, 2005). Differences in molecular mass may have resulted from proteolytic cleavage of different chains of the pro-enzyme by trypsin and chymotrypsin itself, leading to fragmentation with the shorter fragment running out of the gel during electrophoresis (Bender and Killheifer, 1973)

Under reducing conditions, the purified proteins revealed single bands on polyacrylamide gels (Figure 3.2), suggesting that both chymotrypsins are single polypeptide chains. This observation is similar to those made for chymotrypsins from other aquatic invertebrates, such as cuttlefish (Balti et al., 2012), white shrimp (Hernandez-Cortes et al., 1997) and abalone (Groppe and Morse, 1993). Aquatic vertebrates, in contrast, have been reported to have different isoforms of chymotrypsin, similar to mammals. For instance, Jiang et al. (2010) reported 2 isoforms of chymotrypsin in Japanese seabass, while Yang et al. (2009) also reported 2 isoforms in crucian carp.

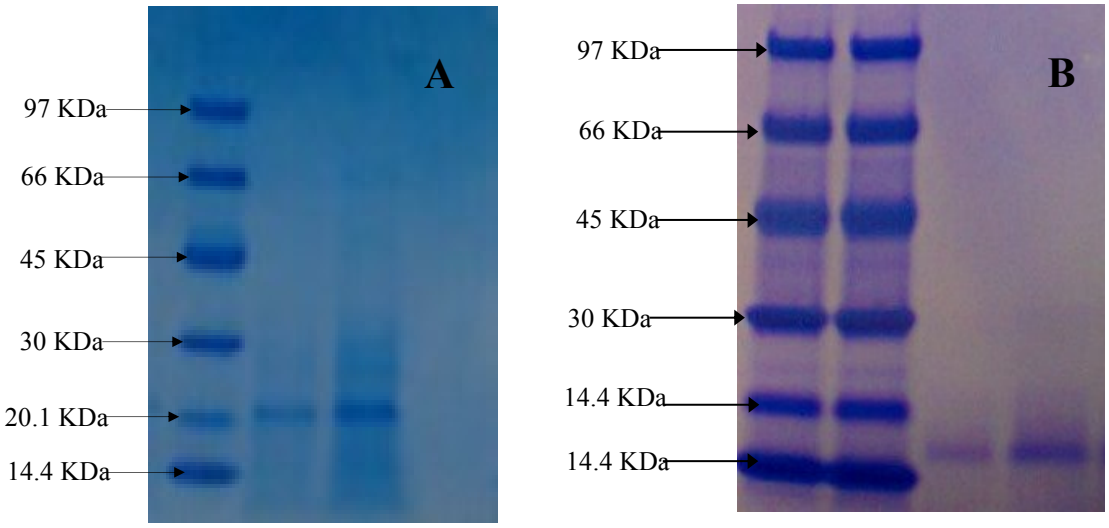

Figure 3.2: SDS-PAGE: A: *S. lessoniana*, lane 1: low molecular weight markers, lanes 2 and 3: affinity chromatography fraction; B: *L. opalescens*, lanes 1 and 2: low molecular weight markes, lanes 3 and 4: affinity fraction. Proteins were visualized by staining with Coomassie brilliant blue

3.4.3 Effect of pH on enzyme activity and stability

Chymotrypsins from the two squid species had a bell-shaped pH-profile for the hydrolysis of BTEE with a pH optimum at pH 7.5 for *S. lessoniana* and 8.5 for *L. opalescens* (Figure 3.3A). pH stability of both enzymes are in the alkaline pH range with *L. opalescens* more stable at pH > 8.5 (Figure 3.3B). Chymotrypsins from aquatic sources have been reported to show high activity and stability at pH > 7 (Asgeirsson and Bjarnason, 1991; Kristjansson and Nielsen, 1992; Simpson, 2000), being more stable between pH 7.5 - 8.5 (Yang et al., 2009; Balti et al., 2012).

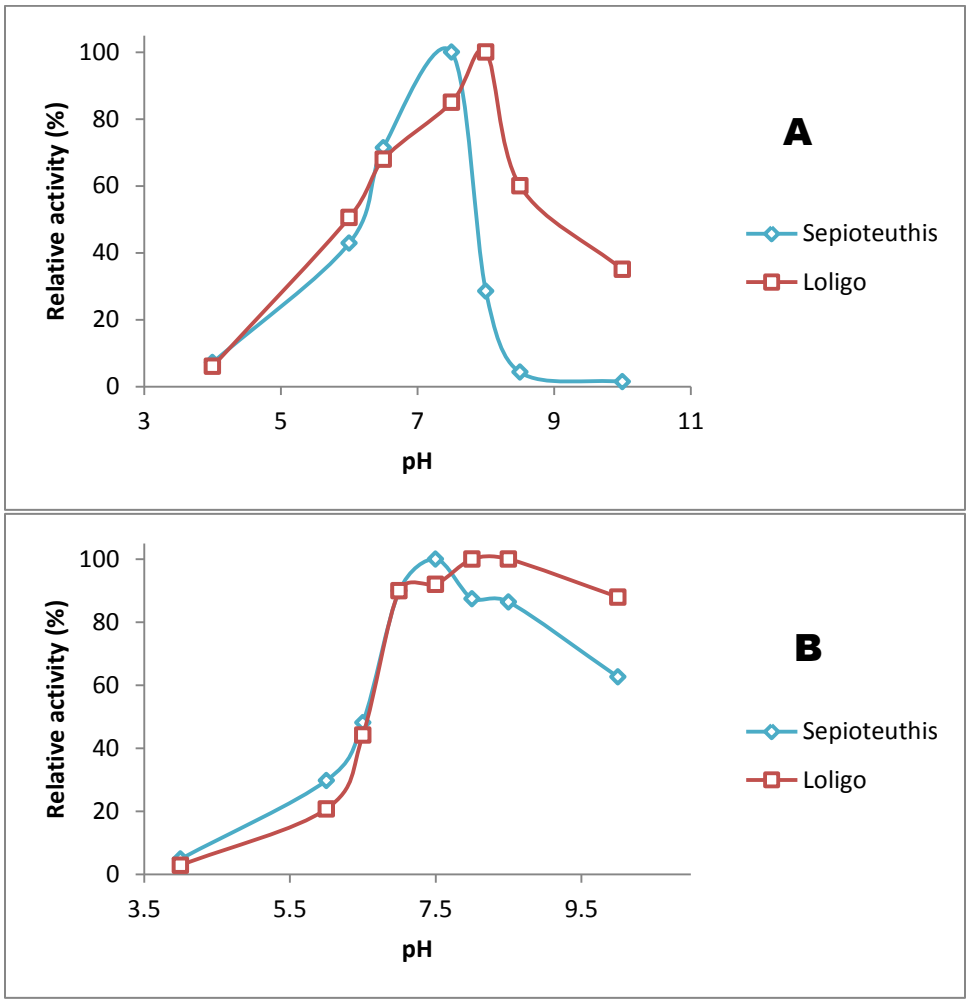

Figure 3.3: Effect of pH on *S. lessoniana* and *L. opalescens* chymotrypsins. A: pH optimum. B: pH stability

3.4.4 Effect of temperature on enzyme activity and stability

L. opalescens chymotrypsin displayed a significantly lower tolerance for heat, with a temperature optimum of 25°C, than *S. lessoniana* (35°C) (Figure 3.4A). The loss of half-maximal activity of the *Loligo* enzyme occurred at 52°C compared to 58°C for *Sepioteuthis* (Figure 3.4B). Temperature plays a very critical role in enzyme activity with optimum temperatures often related to habitat temperatures (Simpson, 2000). The different habitats of the two squid species may account for the observed temperature profiles.

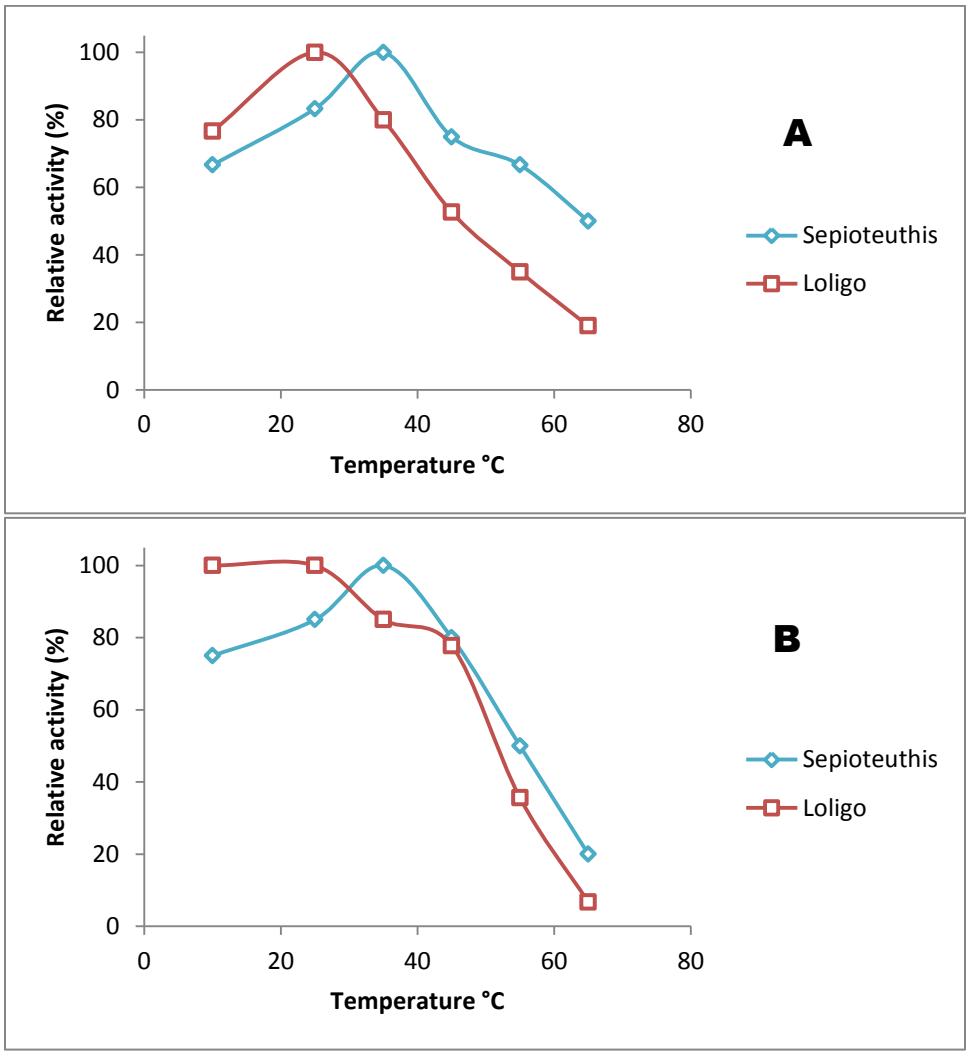

Figure 3.4: Effect of temperature on *S. lessoniana* and *L. opalescens* chymotrypsins. A. Temperature optimum. B. Temperature stability

3.4.5 Effect of protease inhibitors/metal ions/solvents

Protease inhibitors were used to ascertain if the purified enzymes were serine proteases and to confirm that they were chymotrypsin. The activities of both enzymes were strongly inhibited by the chymotrypsin-specific inhibitors TPCK and chymostatin, with the enzymes losing all activity with TPCK and approximately 85% of activity with chymostatin (Table 3.2). These results support the hypothesis that the purified enzymes are serine proteases and are chymotrypsin-like. EDTA also inhibited the enzymes, leading to about 95% loss of activity. Similar effects were

reported for crucian carp (Yang et al., 2009), cuttlefish (Balti et al., 2012) and Monterey sardine (Castillo–Yanez et al., 2006) chymotrypsins. EDTA is a metal chelator and may remove Ca^{2+} ion required for stability and activity. Metal ions had a positive activity effect on both chymotrypsins with Ca^{2+} exhibiting higher activity than Mg^{2+} and Zn^{2+} (Table 3.2). Even though both Mg^{2+} and Zn^{2+} enhanced enzyme activity, only Ca^{2+} satisfies structural requirement for stabilization in chymotrypsin. The enhancing effect of Ca^{2+} and Mg^{2+} on the activity of chymotrypsin has been reported by several authors (Yang et al., 2009; Jiang et al., 2010). Chymotrypsin binds strongly to one Ca^{2+} in its active structure which stabilizes it against denaturation (Delaage et al., 1968). Besides temperature and pH, a good industrial processing protease is expected to be stable in the presence of common commercial solvents. Results from the solvent studies showed that isopropanol and ethanol had no effect on squid chymotrypsin activity at concentrations up to 25% v/v; however, inhibitory effects became more significant above 25% v/v concentration (Table 3.2).

Table 3.2: Effect of proteinase inhibitor metal ions and solvents on squid chymotrypsin activity

Chemical	Concentration	Relative activity (%) *S. lessoniana*	Relative activity (%) *L. opalescens*
Control	-	100	100
TPCK	3.1µM	0	0
Chymostatin	3.1µM	14.48	16.50
EDTA	3.1µM	4.93	6.05
$CaCl_2$	5mM	296.64	281.00
$MgCl_2$	5mM	195.21	150.30
$ZnCl_2$	5mM	183.54	180.20
Isopropanol	2.5%	99.96	99.96
	20%	99.94	99.94
	50%	0	0
Ethanol	2.5%	99.90	99.50
	20%	66.67	64.62
	50%	32.92	40.02

3.4.6 Kinetics studies

The kinetic constants K_m and k_{cat} for hydrolysis of BTEE were determined based on Hanes plots (Table 3.3). K_m was 1.43 mM and k_{cat} was 103.43 sec^{-1} with a catalytic efficiency (k_{cat}/K_m) of 72.33 $sec^{-1} mM^{-1}$ for *S. lessoniana*. For *L. opalescens*, K_m was 0.4 mM and k_{cat} was 349.21 sec^{-1} with a catalytic efficiency (k_{cat}/K_m) of 873.01 $sec^{-1} mM^{-1}$. The values for *Sepioteuthis* chymotrypsin were generally similar to those reported for bovine chymotrypsin, while those of the *Loligo* enzyme were comparable to cod chymotrypsins (Table 3.3).

Table 3.3: Kinetic properties of squid chymotrypsins compared to chymotrypsin from other sources for the hydrolysis of BTEE at 25°C

Enzyme source	K_m (mM)	k_{cat} (sec^{-1})	k_{cat}/K_m ($sec^{-1} mM^{-1}$)	Reference
S. lessoniana	1.43	103.43	72.33	Current study
L. opalescens	0.41	349.21	873.01	Current study
Cod A	0.14	207	1479	Asgeirsson and Bjarnson (1991)
Cod B	0.20	214	1019	Asgeirsson and Bjarnson (1991)
Bovine	1.07	185	172	Atassi and Manshouri (1993)

3.4.7 Proteomics analysis

Mass spectroscopic analysis of tryptic peptides from purified enzymes identified three peptides for *S. lessoniana* (Figures 3.5A–C and Table 3.4). The mass spectrum of peak A (Figure 3.5A) shows 10 main peaks representing the charged ions of a peptide with a calculated molecular mass of 1071.62 (Table 3.4). The fragmentation spectrum of A contained a major ion at m/z 632.36 which was identified as a "y type" fragment ion. For spectrum B, a molecular mass of 1227.72 was calculated from 11 ion species and showed two prominent ions at m/z 303.21 and 675.38 both identified as "y type" ions. Spectrum C yielded a calculated mass of 1109.55 from

10 ion species with a modification (carbamidomethyl) at its amino terminal and showed 3 prominent ions at m/z 248.16, 565.33 and 735.44.

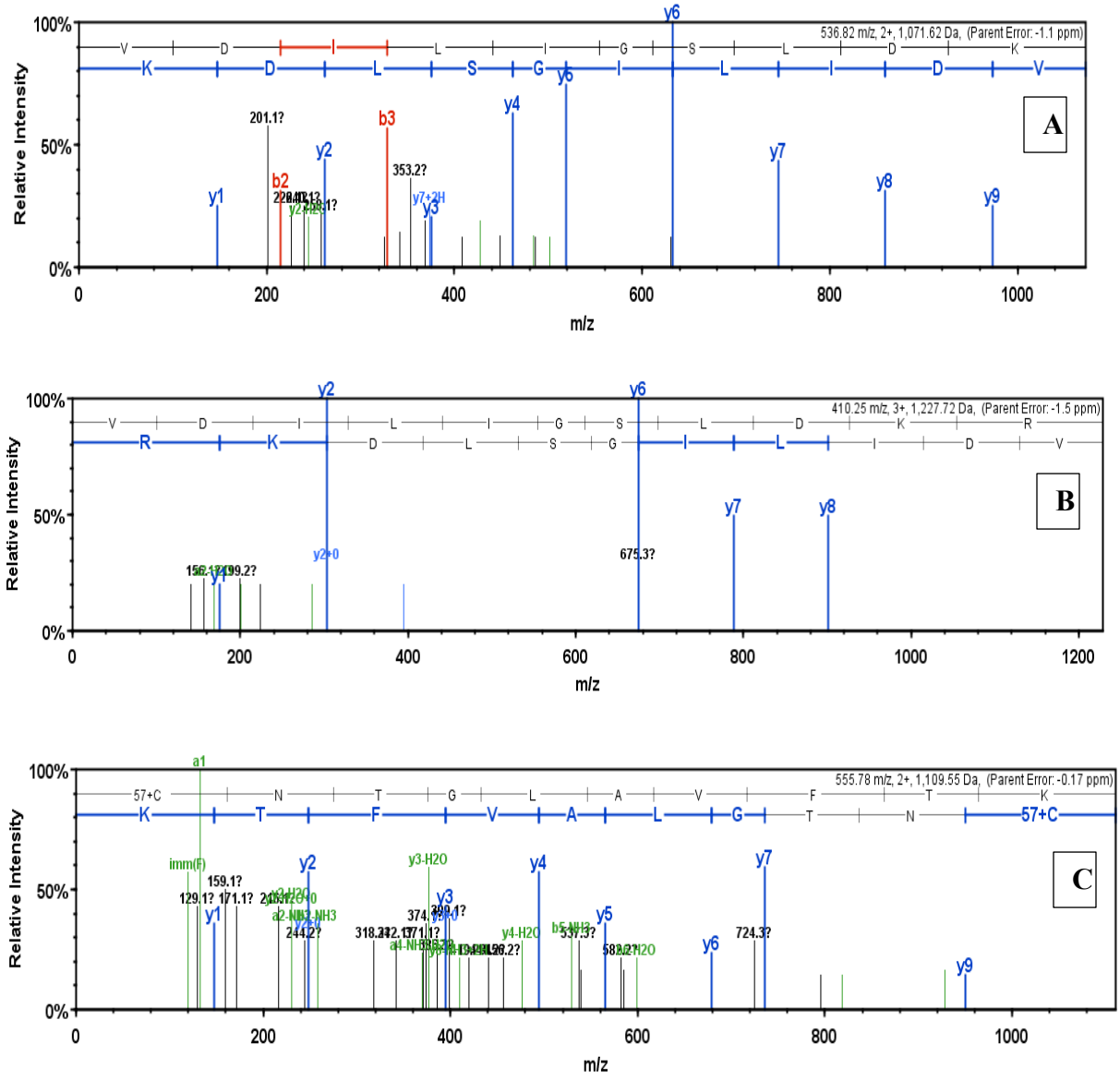

Figure 3.5: Electrospray mass spectra of region a-c corresponding to different peptide fragments from tryptic digest of *S. lessoniana* chymotrypsin-like enzyme

Table 3.4: Fragmentation table showing B and Y ions of peptides from the different mass spectra from *S. lessoniana*

Spectrum **A**			Spectrum **B**			Spectrum **C**		
Amino acid	B Ions	Y Ions	Amino acid	B Ions	Y Ions	Amino acid	B Ions	Y Ions
V	100.08	1,072.63	V	100.08	1,228.73	C+57	161.04	1,110.56
D	215.1	973.56	D	215.1	1,129.66	N	275.08	950.53
I	328.19	858.53	I	328.19	1,014.63	T	376.13	836.49
L	441.27	745.45	L	441.27	901.55	G	433.15	735.44
I	554.35	632.36	I	554.35	788.46	L	546.23	678.42
G	611.38	519.28	G	611.38	675.38	A	617.27	565.33
S	698.41	462.26	S	698.41	618.36	V	716.34	494.3
L	811.49	375.22	L	811.49	531.33	F	863.41	395.23
D	926.52	262.14	D	926.52	418.24	T	964.46	248.16
K	1,072.63	147.11	K	1,054.61	303.21	K	1,110.56	147.11
			R	1,228.73	175.12			

Table 3.5: Peptide fragments of chymotrypsin-like enzyme identified by chemical sequencing from *S. lessoniana*

Mass spectrum	Sequence	Modifications	Observed	Actual Mass	Charge	Delta Da
A	(K)VDILIGSLDK(R)	None	536.8156	1,071.62	2	-0.00122
B	(K)VDILIGSLDKR(T)	None	410.2463	1,227.72	3	-0.0019
C	(K)CNTGLAVFTK(V)	Carbamido methyl (+57)	555.7842	1,109.55	2	-0.00019

MLPFIYLISA ILVSTLSVQG SEVHHIVGGT KAKRCEFPHI VFIYTAKNGH
YFGCGGSLID NKHVLTAAHC MAGDVTKVDI LIGSLDKRTM PIWAPVARFI
KNSKYAKLRS TVVNDIAVLT LAKPVRFTSC IKPIRMATPD EAFVGDCVIA
GWGRTGFNLP TSQILLRANV PIMDHATCAN RLPIILKQHL CVGSGKILDT
TTCKGDSGGP LMCKSAVDGS QVLAGIVSYG WKCNTGLAVF TKVSYYLDWV
NSVRKLIP

Figure 3.6: Amino acid sequence of best hit showing exclusive unique peptides (highlighted) from spectra with 8% sequence coverage of an uncharacterized serine protease belonging to the spear squid (*Loligo bleekeri*)

Using the UR14_2_cephalopoda NCBI database to find the identity of the three peptide fragments, a match was obtained between all three peptides and portions of an uncharacterized serine protease from the spear squid (*Loligo bleekeri*) with accession number D2KX88, assigned to the trypsin superfamily of enzymes. The results showed approximately 8% of the coverage of the hit sequence from the NCBI database (Figure 3.6). The low coverage may be attributed to missing sequences resulting from the inability to match the short tryptic peptides obtained from the in-gel digest and the relatively few deposited corresponding proteins in the NCBI databases (Darville et al., 2012). Nonetheless, this level of identity is fully consistent with a high level of homology between the two squid enzymes. Coupled with the enzyme assay data presented here, both enzymes should be considered as chymotrypsins. In addition, the tryptic digest of *L. opalescens* chymotrypsin identified one peptide that matched the same uncharacterized serine protease in the spear squid (*Loligo bleekeri*) (Figure 3.7). This could have been due to an accidental hit since it was determined to be statistically insignificant.

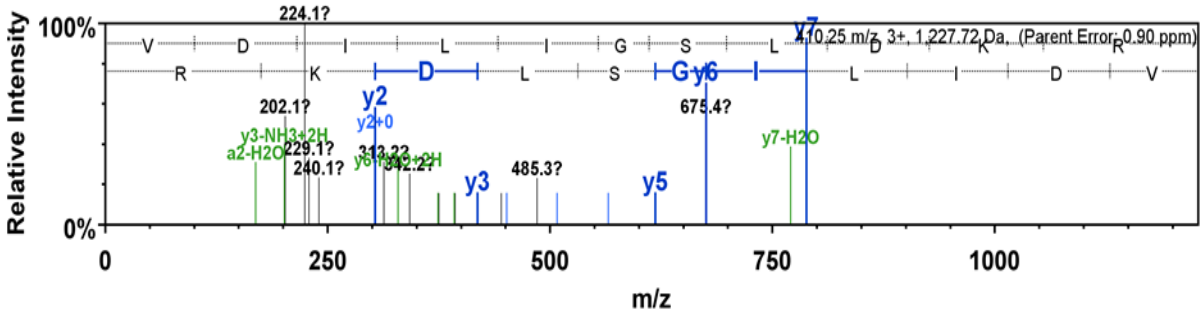

Figure 3.7: Electrospray mass spectra corresponding of peptide fragment from tryptic digest of *L. opalescens* chymotrypsin-like enzyme

3.5 Conclusions

In the present study, two new chymotrypsin-like enzymes from the viscera of *L. opalescens* and *S. lessoniana* were purified and characterized based on molecular mass, inhibitor sensitivity, pH optima, temperature sensitivity and substrate specificity. The purified enzymes were homogeneous on SDS–PAGE, and their molecular masses were estimated to be approximately 18 and 22 kDa, respectively. The enzymes displayed varying tolerance to of temperature and pH. Overall, *L. opalescens* chymotrypsin demonstrated a better adaptation to cold temperature and a higher catalytic activity compared to its homologue in *S. lessoniana*. Three peptides identified from a tryptic digest of *Sepioteuthis* chymotrypsin were identical to peptides found in a serine protease from another cephalopod. This report enhances knowledge on the distribution of chymotrypsin-like enzymes among the class Cephalopoda, family Loliginidae. Further work is needed to isolate cDNAs encoding the two chymotrypsin-like enzymes and to determine the properties of these proteases as possible biotechnological tools in the food processing and pharmaceutical industries.

CONNECTING STATEMENT 2

In the previous chapter, three peptides from a tryptic digest of a *Sepioteuthis lessoniana* chymotrypsin-like enzyme were found to be identical to peptides found in a serine protease of another cephalopod. In this chapter, a detailed study of a partial cDNA encoding the chymotrypsin-like enzyme from *S. lessoniana* is presented. The partial cDNA was isolated using degenerate primers manually designed from peptide fragments obtained from LC-MS/MS in the previous study. The use of degenerate primers to amplify genes based on mass spectrometry data is a useful alternative strategy when information on gene sequence is not available.

CHAPTER IV

cDNA cloning of a partial chymotrypsin-like gene from squid (*Sepioteuthis lessoniana*) using degenerate primers

4.1 Abstract

Reverse transcription (RT)-PCR with non-degenerate and degenerate primers was used to amplify partial mitochondrial cytochrome oxidase 1(MtCO1) and partial chymotrypsin-like genes, respectively, from the warm-water squid *Sepioteuthis lessoniana*. A partial length MtCO1 gene showed 98% homology to a *Sepioteuthis lessoniana* CO1 gene in the NCBI database, confirming the species of squid used in this study. Translated amino acid sequence from deduced from a partial chymotrypsin-like cDNA showed 48% amino acid identity to a serine protease from *Heterololigo bleekeri*, 58% to an S1 peptidase from *Sepia latimus* and 45% to a hypothetical protein from *Octopus bimaculoides*. The translated protein also showed one domain hit belonging to the trypsin-like super family of enzymes.

Keywords: *Sepioteuthis lessoniana,* chymotrypsin, LC-MS/MS, degenerate primers

4.2 Introduction

Most applications of the polymerase chain reaction (PCR) are based on the use of primers that precisely match a known target sequence (Henikoff and Henikoff, 1993). In some instances, precise primers cannot be used to exactly match a target sequence due to unavailability of genomic information for the gene of interest. In such cases, PCR primer design is usually based on reverse translation of multiple aligned sequences across conserved regions of proteins. These primers, referred to as degenerate primers, are designed from consensus sequences of evolutionary conserved portions of protein families referred to as BLOCKs (Henikoff,. and Henikoff, 1991) or ancient conserved regions (ACRs) (Green et al., 1993) and can be used in PCR to generate candidate cognate cDNAs (Rose et al., 1998; D 'Esposito et al., 1994). These blocks or ACRs are thought to have originated early in evolution and have remained much the same throughout phylogeny (Henikoff and Henikoff, 1993). They sometimes correspond to a part of a gene which has been duplicated in its entirety and diverged to form a more or less dispersed family of genes and pseudogenes (D' Esposito et al., 1994). The very high sequence homology in these gene families is probably due to a continual process of turnover by unequal crossing over, gene conversion, and transposition (Bateman et al., 1999).

Various approaches have been used experimentally to isolate distantly-related sequences by PCR. The synthesis of a pool of degenerate primers containing most or all of the possible nucleotide sequences implicit in a multiple alignment is the most common (Henikoff et al., 1998). Such ACR/BLOCK-derived oligonucleotides can be considered as degenerate sequence-tagged sites (STSs) (Olsen et al., 1989), and as such, can facilitate the study of multigenic families as well as provide markers for genes whose encoded proteins contain specific motifs (D' Esposito et al., 1994). A limitation of this approach is that the concentration of primer drops with increasing degeneracy due to increase in divergence among genes. As a result, the number of primer molecules in a PCR reaction that can be amplified during a cycle drops, as these primers are used up early in the reaction (Henikoff and Henikoff, 1993).

In the current study, we confirm the identity of the squid species used by amplifying a partial mitochondrial cytochrome oxidase gene with specific primers and report a partial chymotrypsin cDNA using degenerate primers designed from peptide fragments .These peptides were obtained from LC-MS/MS data from a tryptic digest of a chymotrypsin-like enzyme from *Sepioteuthis lessoniana*. This approach was used because of a limited number of mollusk chymotrypsin gene

sequences available in the NCBI database, coupled with high divergence amongst the few such genes that are available, preventing the use of block sequences for degenerate primer design. Few studies have reported work on chymotrypsins from marine invertebrates (Balti et al., 2012), with just a handful describing cloning and characterization of chymotrypsin cDNA and genomic DNA sequences. This study sought to amplify and clone cDNA encoding chymotrypsin from this squid for the first time using degenerate primers.

4.3 Materials and Methods

Peptide fragments used in this study were obtained from results of previous chapter III.

4.3.1 Total RNA and Genomic DNA extraction

Total RNA was extracted from the squid sample using the Trizol reagent (Invitrogen, USA) according to the manufacturer's protocol, and potential genomic contamination was removed by treatment with DNase I (Ambion, Life technologies) according to the manufacturer's protocol.. RNA quality was determined by agarose gel electrophoresis and quantification was done with an ND-1000 NanoDrop UV spectrophotometer (NanoDrop Technologies, Wilmington, USA). Treated RNA was purified using an RNeasy MinElute clean up kit (Qiagen) according to the manufacturer's instruction. Genomic DNA was extracted and purified using a PureLink Genomic DNA mini kit (Invitrogen, USA) according to the manufacturer's instruction.
To synthesize first stranded cDNA, aliquots of total RNA (2 µg) were reverse transcribed using a reversed first strand cDNA synthesis kit (Invitrogen, USA) and stored at -20 °C.

4.3.2 PCR primers

Both partial mitochondrial cytochrome oxidase 1 and partial chymotrypsin gene were amplified in this study. Partial CO1 sequence was obtained from the NCBI database and primers designed based on this sequence (Table 4. 1). Degenerate oligonucleotide primers were manually designed from putative chymotrypsin peptides obtained from the LC-MS/MS results. This approach was used because a search of different databases showed that the few genomic sequences available for squid chymotrypsins were highly divergent (not shown).

4.3.3 PCR amplification of partial CO1 gene

Amplification of a partial CO1 gene was done as a positive control and also to confirm the species of *Sepioteuthis* used in the study. Amplifications were performed on 5 µl of the resulting cDNA and genomic DNA. In amplifying the partial CO1 gene, primer set SEPCO1F2 and SEPCO1R1 were used. The amplification was carried out in a 50 µl reaction mixture containing 5 µL 10× buffer, 1 mM $MgCl_2$, 1 µl 10 mM dNTPs, 1 µl each primer at 10 µmole concentration, 0.5 µl Taq polymerase, 5 µl DNA template and 36.5 µl RNase-treated water. Reaction mixtures were preheated at 95°C for 5 min, followed by 30 cycles of amplification (95°C for 40 s, 50°C for 30 s, and 72°C for 1.5 min), final extension at 72°C for 1 min and holding at 4°C. PCR products (20 µl each) of both CO1 and chymotrypsin gene were resolved on a 1% agarose gel using a 1 kb ladder as a marker. Bands of the expected size were cut from the gel and purified using the EZ-10 spin columns and gel extraction kits according to the manufacturer's instructions.

4.3.4 Degenerate primers

The degenerate oligonucleotide primers SEPIOF1, SEPIOF2, SEPIOF3, SEPIOF4, SEPIORR1 and SEPIORR2 (Table 4.1) were manually designed from peptides obtained from LC/MS/MS results. The sequences and combinations of the sense (SEPIOF1, SEPIOF2, SEPIOF3 and SEPIOF4) and antisense (SEPIORR1and SEPIORR2) oligonucleotide primers are presented in Table 4.2.

4.3.5 PCR amplification of partial chymotrypsin gene

Degenerate primers in different combinations were used, resulting in 8 different primer combinations (Table 4.2). The amplification was carried out in a 50 µl reaction mixture containing 5 µL 10× buffer, 1 mM $MgCl_2$, 1 µl 10 mM dNTP, 5 µl each primer at 10 µmole concentration, 0.5µl Taq polymerase, 1.5 µl DNA template and 32 µl RNase treated water. The reaction mixtures were preheated at 95°C for 5 min, followed by 40 cycles of amplification (95°C for 1 min), different annealing temperatures (40, 45, 50, 55 and 60°C) for 1 min, and 72°C for 2 min), final extension at 72°C for 5 min and holding at 4°C.

PCR products (20 µl) of both CO1 and chymotrypsin were resolved by electrophoresis through a 1% agarose gel using a 1 kb ladder as a marker. Bands of the expected size were cut from the gel and purified on EZ-10 spin columns and using gel extraction kits according to the manufacturer's instructions.

4.3.6 Cloning and sequencing of PCR products

Purified PCR fragments were ligated into the pGEMT cloning vector (Promega, Madison, WI) following the manufacturers instruction. The ligation mixture was used to transform competent *E. coli* (DH5α). Competent cells were incubated in lysogeny broth (LB) supplemented with ampicillin. Positive recombinant clones for CO1 and chymotrypsin were sequenced at McGill University and Génome Québec Innovation Centre, Montreal Canada using T7 and SP6 specific primers (Invitrogen Corp., San Diego, CA)

Table 4.1: Primers used in this study

Primer name	Primer Sequence	Purpose
SEPCO1F2	5′ ACA AAT CAT AAA GAT ATT GG 3′	Mt CO1 amplification
SEPCO1R1	5' GTA AAT ATA TGT TGG GCT CA 3'	Mt CO1 amplification
SEPIOF1	5' GAY ATI CTN ATH GGI AGY CTN GA 3'	Chymotrypsin gene amplification
SEPIOF2	5' GAY ATI TTR ATH GGI TCN TTR GA 3'	Chymotrypsin gene amplification
SEPIOF3	5' GAY ATI CTN ATH GGI AGY TTR GA 3'	Chymotrypsin gene amplification
SEPIOF4	5' GAY ATI TTR ATH GGI TCN CTN GA 3'	Chymotrypsin gene amplification
SEPIORR1	5' GTR AAI ACI GCN AAN CCN GTA T 3'	Chymotrypsin gene amplification
SEPIORR2	5' TRA AIA CIG CAN GNC CNG TGT T 3'	Chymotrypsin gene amplification
T7	5' TAA TAC GAC TCA CTA TAG GG 3'	Sanger sequencing
SP6	5' TAT TTA GGT GAC ACT ATA G 3'	Sanger sequencing

Table 4.2: Degenerate primer combination used in study

Tube	Primer combination
1	SEPIOF1 + SEPIORR1
2	SEPIOF1 + SEPIORR2
3	SEPIOF2 + SEPIORR1
4	SEPIOF2 + SEPIORR2
5	SEPIOF3 + SEPIORR1
6	SEPIOF3 + SEPIORR2
7	SEPIOF4 + SEPIORR1
8	SEPIOF4 + SEPIORR2

4.4 Results and Discussion

4.4.1 Mass Spectroscopy results

MS analysis of a squid chymotrypsin tryptic digest identified three peptides: KVDILIGSLDKR, KVDILIGSLDKRT and KCNTGLAVFTKV (Table 4.3)

Table 4.3: Peptide fragments of chymotrypsin-like enzyme from *Sepioteuthis* identified by MS/MS

Mass spectrum	Sequence	Modifications	Observed	Actual Mass	Charge	Delta Da
A	(K)VDILIGSLDK(R)	None	536.8156	1,071.62	2	-0.00122
B	(K)VDILIGSLDKR(T)	None	410.2463	1,227.72	3	-0.0019
C	(K)CNTGLAVFTK(V)	Carbamido methyl (+57)	555.7842	1,109.55	2	-0.00019

4.4.2 PCR amplification and Sequencing of partial CO1 gene

Ten independent clones of the mitochondrial CO1 gene were sequenced on both strands. All ten yielded essentially identical amplicons, differing by 1 to 4 nucleotides over the 600 bp sequence (Figure 4.1). Alignment of the sequence using the BLAST algorithm at the NCBI database showed the query sequence to share 98% sequence identity to partial CO1 gene from *S. lessoniana* (accession number KF052480.1). (Default Blast parameters were: E cut-off = 10, mask low complexity = yes). This result confirmed the species of *Sepioteuthis* to be *lessoniana*. The CO1 gene has been shown to be highly conserved and hence is routinely used as a molecular marker for distinguishing closely related species. The *Sepioteuthis* genus has overlapping morphological characteristics which can lead to uncertainties in identification and taxonomy, hence the necessity to identify them on the molecular level.

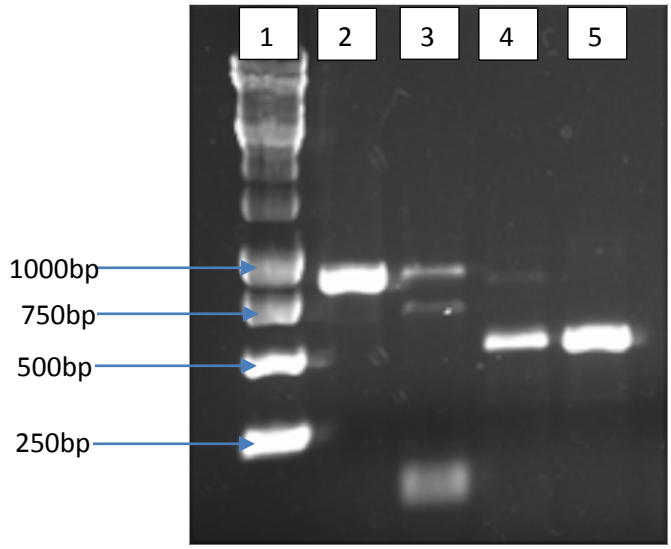

Figure 4.1: 1% agarose gel with 1 kb base pair ladder and amplicons from RT-PCR of chymotrypsin-like gene from *Sepioteuthis lessoniana*, well 1: 1 kb ladder, well 2: positive control, wells 4and5: amplicons from CO1 amplification

CTTTGGTTTGAGCAGGATTAGTTGGTACCTCACTAAGGTTAATAATTCGAACCGAATTAGGTAAACCCGGCTCATTACTAAATGAT
GACCAATTATATAATGTTGTAGTTACTGCACACGGTTTTATTATAATTTTCTTTATAGTTATACCTATTATAATTGGAGGCTTCGG
TAACTGACTTGTCCCTCTCATACTAGGAGCACCTGATATAGCATTCCCACGAATAAATAATATAAGATTCTGATTGCTACCTCCAT
CACTAACACTCCTTTTAGCATCCTCAGCAGTTGAAAGAGGAGCCGGTACAGGATGAACCGTCTATCCGCCCCTCTCAAGTAACCTA
TCTCATGCTGGACCTTCAGTTGATCTTGCTATCTTCTCACTACATTTAGCTGGTATCTCTTCTATCCTAGGAGCAATTAACTTTAT
ACAACCATTATTAATATACGATGAGAAGGTTTACTTATAGAGCGCTTACCTTTATTGCCGATCTGTCTTTATTACTGCTATCTTAC
TCCTTCTATCAGTACCTGTTTTTAGCGGGGGCCATTCAATATTCTTACGACCGAAACTTAATACACTTTCTTCACCCA

Figure 4.2: Nucleotide sequence of partial Mt CO1 gene from *Sepioteuthis lessoniana*

4.4.3 Cloning of partial chymotrypsin gene

To obtain cDNA, the enzyme was purified, cleaved using trypsin and peptides identified using LC/MS-MS. On the basis of the identified peptides, degenerate primers were designed (Table 4.1) for both the forward and reverse strand. The primers were combined (Table 4.2) and used in RT-PCR. Of the eight primer combinations tried, four produced visible and distinct bands. Single bands ranging between 500 and 1000 bp were amplified. Each obvious band was cloned. Six independent clones for each combination were sequenced on both forward and reverse strands, four of which shared similar identity with only 4-6 nucleotide differences among clones. The predicted amino acid sequence was used to search the non-redundant protein databases and showed 48% amino acid identity to a serine protease from *Heterololigo bleekeri*, 58% to an S1 peptidase from *Sepia latimus* and 45% to a hypothetical protein from *Octopus bimaculoides*. The translated protein also showed a domain hit belonging to the trypsin-like super family of enzymes (Figure 4.3).

```
  1 gataccaaagatattctgattggcagcctggataaaaaacagtggtatcagcgccattgc 60
    D  T  K  D  I  L  I  G  S  L  D  K  K  Q  W  Y  Q  R  H  C

 61 ctgaaatatgcgaaactgcgcagcaaaaccgtggtgaacgatattgcggtgctgtatacc 120
    L  K  Y  A  K  L  R  S  K  T  V  V  N  D  I  A  V  L  Y  T

121 attccgattcatccggatgaagcgtttgtgattctgattattcgcaacccggcgattatg 180
    I  P  I  H  P  D  E  A  F  V  I  L  I  I  R  N  P  A  I  M

181 gatcatgcggtgctgatgctgctgcatagctttctgcgctttgtgaccggcaaaattctg 240
    D  H  A  V  L  M  L  L  H  S  F  L  R  F  V  T  G  K  I  L

241 gataccggcgcgctgatgtgcaaaagcgcggtggatggcctggcgggcattgtgagctat 300
    D  T  G  A  L  M  C  K  S  A  V  D  G  L  A  G  I  V  S  Y

301 ggctggaaatgcttttgcaacaccggcctggcggtgtttaccaaagtgagc 351
    G  W  K  C  F  C  N  T  G  L  A  V  F  T  K  V  S
```

Figure 4.3: Nucleotide and deduced amino acid of a chymotrypsin-like cDNA clone from *Sepioteuthis lessoniana*.

Sequence alignment of the predicted amino acid sequence from the translated gene with other chymotrypsins (bovine, Atlantic cod, salmon and Bleeker's squid) showed that most of the N-terminal amino acid sequence was missing. The active site aspartate was present on all aligned sequences, while histidine and serine were not present in the partially amplified gene (Figure 4.4).

```
Bovine                      -------------------CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQ--DKTGF
sp|P80646|CTRB_GADMO        -------------------CGSPAIQPQVTGYARIVNGEEAVPHSWPWQVSLQ--QSNGF
tr|B5XB02|B5XB02_SALSA      -MAFLWFVSCLAFVSAAYGCGIPAIKPEVSGYARIVNGEEAVPHSWPWQVSLQ--QTSGF
tr|D2KX88|D2KX88_HETBL      MLPFIYLISAIL---------VSTLSVQGSEVHHIVGGTKAKRCEFPHIVFIYTAKNGHY
Sepioteuthis                ------------------------------------------------------------

Bovine                      HFCGGSLINENWVVTAAHCGVTTSDV--VVAGEFDQGSS--SEKIQKLKIAKVFKNSKYN
sp|P80646|CTRB_GADMO        HFCGGSLINENWVVTAAHCNVRTYHR--VIVGEHDKASD---ENIQILKPSMVFTHPKWD
tr|B5XB02|B5XB02_SALSA      HFCGGSLINENWVVTAAHCNVATYHR--VIIGEHKKGSGNNAEDIQILKPAKVFTHPKWN
tr|D2KX88|D2KX88_HETBL      FGCGGSLIDNKHVLTAAHCMAGDVTKVDILIGSLDKRTMPI--WAPVARFIKNSKYAKLR
Sepioteuthis                ------------------------DTKDILIGSLDKKQW-----YQR----HCLKYAKLR
                                                     :: *. .:                 .  *

Bovine                      SLTINNDITLLKLSTAASFSQTVSAVCLPSASDDFAAGTTCVTTGWGLTRYTNANTPDRL
sp|P80646|CTRB_GADMO        SRTINNDISLIKLASPAVLGTNVSPVCLGESSDVFAPGMKCVTSGWGLTRYNAPGTPNKL
tr|B5XB02|B5XB02_SALSA      PSTINNDISLIKLSTPAVLNTNVSPVCLAETADVFAPGMTCVTTGWGLLRYNALNTPNEL
tr|D2KX88|D2KX88_HETBL      ST-VVNDIAVLTLAKPVRFTSCIKPIRMATPDEAFV--GDCVIAGWGRTGFNLPTS-QIL
Sepioteuthis                SKTVVNDIAVLYT-IP------------IHPDEAFV----------------------IL
                               : ***:::                  : *.                      *

Bovine                      QQASLPLLSNTNCKKYWGTK-IKDAMICAGA---SGVSSCMGDSGGPLVCKK--NGAWTL
sp|P80646|CTRB_GADMO        QQAALPLMSNEECSQTWGNNMISDVMICAGA---AGATSCMGDSGGPLVCQK--DNVWTL
tr|B5XB02|B5XB02_SALSA      QQAALPLLSNEQCKTHWGSS-ISDVMICAGG---AGATSCMGDSGGPLVCEK--DNVWTL
tr|D2KX88|D2KX88_HETBL      LRANVPIMDHATCANRLPI--ILKQHLCVGSGKILDTTTCKGDSGGPLMCKSAVDGSQVL
Sepioteuthis                IIRNPAIMDHAV-----LM--LLHSFLRFVTGKIL--------DTGALMCKSAVDG---L
                                  ::..           : .  :               . * *:*:.  :    *

Bovine                      VGIVSWGSSTCSTSTPGVYARVTALVNWVQQTLAAN-
sp|P80646|CTRB_GADMO        VGIVSWGSSRCSVTTPAVYARVTELRGWVDQILAAN-
tr|B5XB02|B5XB02_SALSA      VGIVSWGSSRCSTTTPAVYARVTELRSWVDQTLAAN-
tr|D2KX88|D2KX88_HETBL      AGIVSYGWKC--NTGLAVFTKVSYYLDWVNSVRKLIP
Sepioteuthis                AGIVSYGWKCFCNTGLAVFTKVS--------------
                            .****:* .    :  .*:::*:
```

Figure 4.4: Multiple sequence alignment of a predicted *Sepioteuthis* partial chymotrypsin-like protein with homologues from bovine, Atlantic cod, salmon and Bleeker's squid

4.5 Conclusion

The use of degenerate primers designed based on MS data to amplify genes has been shown in this study as an alternative strategy when information on genomic sequence is not available. It is, however, important to note that in the case of very high primer degeneracy, complications may arise especially in trying to find suitable annealing temperatures and primer lengths. Similarly, as the degeneracy increases due to more divergent genes, the concentration of any single primer drops. As a result, the number of primer molecules in a PCR reaction that can prime synthesis during the amplification cycles drops. Isolation of an unknown sequence using peptides generated from mass spectrometry can be an alternative method for isolating genes and investigating biological function.

CONNECTING STATEMENT 3

The previous study on temperature optima and kinetic properties of *Loligo opalescens* chymotrypsin (Chapter 3) illustrated how well this enzyme is adapted to function in a cold habitat compared to chymotrypsins from different sources. The results from those studies showed that the catalytic activity of *Loligo* chymotrypsin was > 10 times higher than that of a chymotrypsin from a closely related squid (*Sepioteuthis*) of the same family (Loliginidae). This led to an interest in trying to understand the molecular basis of these observations; however, limited information on genomic sequences of *Loligo* is available. We therefore decided to sequence and assemble the transcriptome of *Loligo opalescens* using de novo methods to provide a facile route to the cloning of a full-length cDNA encoding chymotrypsin from this organism.

The results of this study have been prepared with the aim of submitting it for publication

CHAPTER V

De novo transcript assembly and analysis of *Loligo opalescens* from RNA-Seq

5.1 Abstract

De novo transcriptome sequencing and assembly has the potential to open up new frontiers by providing insight into biodiversity and functional capabilities. In the current study, de novo assembly from massively parallel sequencing data were generated from total *Loligo opalescens* RNA using Illumina Genome Analyzer. The transcriptome was assembled on normalized reads using the Trinity assembler. Each component longest transcript was aligned against the uniprot_sprot_2013_11 protein database using the blastx program from the NCBI BLAST family. For each longest transcript, the best BLAST hit was then used to annotate its associated component in Differential Expression. Overall, 61661 transcripts were obtained with total transcript length of 46232292 bp. Maximum transcript length obtained was 16932 bp while the minimum length was 201 bp. A further analysis of open reading frames from the data has the potential to identify important genes for function al investigation.

Keywords: De novo sequencing, De novo assembly, RNA-Seq, Loligo opalescens

5.2 Introduction

Squids comprise one of the most economically important aquatic groups and together with cuttlefish and octopus account for ≥3% of the global capture of all aquatic species (FAO, 2013). Interest in the squids is stimulated mainly by their high tolerance for environmental changes and seasonal migrations between shallow coastal waters in summer for spawning and deeper waters in winter (Gauvrit et al., 1997). A convenient example is the California market squid, *Loligo opalescens*, an Eastern Pacific ocean species that can be found as deep as 500 m.

Fully sequenced genomes are superior to genome fragments since they provide very accurate reference for interpreting transcriptomes. With recent advancement in next generation sequencing technologies, whole genomes can be sequenced in very short time frame. Various approaches including RNA-Seq can be used to characterize genomes. RNA-Seq is a powerful tool for studying the effect of the transcriptome on phenotypes and generally applied to transcript identification, splice variant analysis and differential expression. It is one of the most recently developed approach to transcriptome profiling that uses deep-sequencing technologies and provides more precise measurement of levels of transcripts and their isoforms than other methods (Wang et al 2009). This approach can catalogue all kinds of transcript, including mRNAs, non-coding RNAs and small RNAs; to determine the transcriptional structure of genes, in terms of their start sites, splicing patterns and other post-transcriptional modifications if present; and to quantify the changing expression levels.

No studies have been reported on gene characterization in *Loligo opalescenc* and indeed, few such reports are available from marine invertebrates in general. We therefore De novo sequence and assemble the genome of this squid using massively paralleled sequencing approach and Trinity software. This data has the ability to complement in vitro studies and give a better understanding of observations made at the gene level and open up new perspectives.

5.3 Materials and Methods

5.3.1 Total RNA extraction

Total RNA was extracted from the squid (*Loligo opalescens*) muscle using the Trizol reagent (Invitrogen, USA) according to the manufacturer's protocol. Potential genomic contamination was removed by treatment with DNase I (Ambion, Life technologies) according to the manufacturer's protocol. . Treated RNA was purified using an RNeasy MinElute clean up kit (Qiagen) according to the manufacturer's instruction. RNA quality was determined by agarose gel electrophoresis and quantification was done with an ND-1000 NanoDrop UV spectrophotometer (NanoDrop Technologies, Wilmington, USA).

5.3.2 Genome sequencing and data assembly

Purified RNA was directly sequenced using Illumina standard technology (Illumina sequence analyzer system at the McGill University and Génome Québec Innovation Centre, Montreal Canada). This sequencing pipeline was developed following the protocol as described by Haas et al. (2013) and Grabherr (2011). Paired reads per library were generated using the Illumina Hiseq 2000/2500 sequencer. Base calls were made using the Illumina CASAVA pipeline. Base quality was encoded in phred 33.

5.3.3. Analysis of sequence data

5.3.3.1 Trimming

Read trimming and clipping were performed using the Trimmomatic software (Usadel Lab, 2014). Reads were trimmed from the 3' end with a minimal phred score of 30 with Illumina sequencing adapters removed. Read minimal length was set to 32 bp.

5.3.3.2 Normalization

Normalization of data was performed to reduce memory requirement and decrease assembly runtime by reducing the number of reads, using the Trinity normalization utility (Grabherr et al., 2011) inspired by the Diginorm algorithm (Brown et al., 2012): first, a catalog of k-mers from all reads was created; then, each RNA-Seq fragment (single read or pair of reads) was

probabilistically selected based on its k-mer coverage value and the targeted maximum coverage value.

5.3.4 De Novo Assembly

The transcriptome was assembled on normalized reads using the Trinity assembler Haas et al., 2013; Grabherr et al., 2011) based on three algorithms: Inchworm, Chrysalis and Butterfly

5.3.5 BLAST Annotation

Each component longest transcript was aligned against the uniprot_sprot_2013_11 protein database using the blastx program from the NCBI BLAST family. For each longest transcript, the best BLAST hit was used to annotate its associated component in Differential Expression.

5.3.6 Differential Expression

The analysis design is presented in Table 5.1. Each line represents a sample, each column an experimental design, each value (0, 1 or 2) the sample group membership:

Table 5.1: Sample names and experimental designs for differential expression

Sample	1vs2	1vs3	2vs3
loligo1	1	1	0
loligo2	2	0	1
loligo3	0	2	2

- **0**: the sample is not a member of any group.
- **1**: the sample is a member of the control group.
- **2**: the sample is a member of the test case group.

Gene abundance estimation for each sample was performed using RSEM (RNA-Seq by Expectation-Maximization) (Li and Dewey, 2011). Differential gene expression analysis was performed using DESeq (Anders and Huber, 2010) and edgeR (Robinson et al., 2010) R Bioconductor packages.

5.4 Results

5.4.1 Trimming

Exactly 200,908,285 Paired Reads (raw pair reads) were obtained from all three samples. This value was reduced to 187,037,223 after trimming resulting in an output percentage of 93% (Table 5.2).

Table 5.2: Trimming metrics

Sample	Raw Paired Reads	Surviving Paired Reads	%
loligo3	68,193,187	63,105,824	93
loligo1	65,629,553	61,423,621	94
loligo2	67,085,545	62,507,778	93
TOTAL	200,908,285	187,037,223	93

- **Raw Paired Reads**: number of paired reads obtained from the sequencer
- **Surviving Paired Reads**: number of paired reads remaining after the trimming step
- **%**: percentage of surviving paired reads / raw paired reads

5.4.2 Normalizing

Normalizing the data using the Trinity normalization utility (Grabherr et al., 2011) inspired by the Diginorm algorithm (Brown et al., 2012) resulted in 5,073,076 surviving paired reads (Table 5.3).

Table 5.3: Normalization metrics

Surviving Paired Reads	%
5,073,076	3

- **Surviving Paired Reads**: number of paired reads remaining after the normalization step
- **%**: percentage of surviving paired reads after normalization / surviving paired reads after trimming

5.4.3 De novo assembly

Trinity was used to create fasta files with a list of contigs representing the transcriptome isoforms. The transcripts were then grouped in components mostly representing genes. After removal of highly similar contigs, the final assembly had 61661 contigs (46232292 bp) with an average size of 749 bp (Table 5.4). The N50 value was calculated to be 1298 bp (Figure 5.1).

Table 5.4: Assembly metrics

Description	**Value**
Number of Transcripts	61661
Number of Components	50234
Total Transcripts Length (bp)	46232292
Min. Transcript Length (bp)	201
Median Transcript Length (bp)	393
Mean Transcript Length (bp)	749.78
Max. Transcript Length (bp)	16932
N50 (bp)	1298

- N50 is a statistic calculated such that 50% of the total length of the assembly is contained in contigs equal or larger than this value.

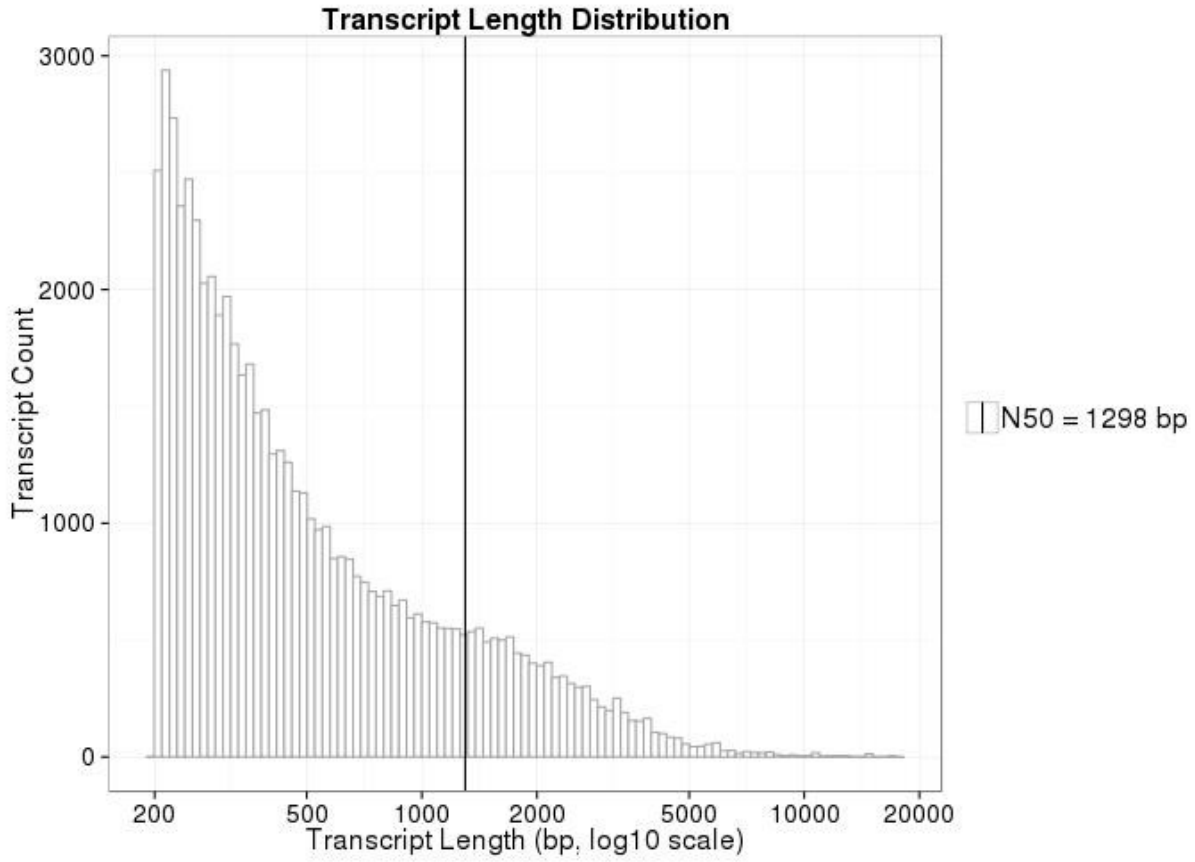

Figure 5.1: Sequence coverage for all de novo assembled transcripts (contigs).

5.4.4 Differential Gene Analysis Description

After normalization, the lists of differentially expressed genes (DEGs) using the statistic tests from the edgeR and DESeq packages were determined. In the case of each data set, gene expression levels between minimum of two types of biological samples were compared and ranked according to adjusted values. Genes that had adjusted values < 0.05 were selected as differentially expressed. Tables 5.5, 5.6 and 5.7 shows that 6 genes were differentially expressed in three samples used with p values and adjusted p values approaching zero.

Table 5.5: Differential Gene Expression (1vs2)

Gene	Symbol	log_ FC	log_ CPM	deseq.p -value	deseq.adj. pvalue	edger. p.value	edger.adj. p.value
c40452_g1	.	4.9	13	3.1e-31	4.7e-27	5e-19	5.2e-15
c40524_g2	sp|Q8WZ42|TITIN_HUMAN	5.2	5.7	3.6e-31	4.7e-27	5e-20	8.1e-16
c40254_g4	sp|Q8WZ42|TITIN_HUMAN	5.6	4.1	8.9e-30	7.8e-26	6.6e-21	2.8e-16
c40452_g2	.	4.7	13	1.7e-29	1.1e-25	5.3e-18	3.7e-14
c40276_g1	sp|Q8TGM6|TAR1_YEAST	4.4	13	3.7e-27	2e-23	1.3e-16	6e-13
c39865_g1	sp|Q3E811|RRT15_YEAST	4.3	14	4.2e-26	1.9e-22	5.4e-16	2.2e-12

Table 5.6: Differential Gene Expression (1vs3)

Gene	Symbol	log_ FC	log_ CPM	deseq.p -value	deseq.adj. pvalue	edger. p.value	edger.adj. p.value
c40452_g1	.	5	13	3e-57	7.9e-53	3.3e-21	3.4e-17
c40452_g2	.	4.9	13	3.9e-55	5.2e-51	1.8e-20	1.3e-16
c40524_g2	sp|Q8WZ42|TITIN_HUMAN	5.2	5.7	2.2e-52	1.9e-48	9.2e-22	1.7e-17
c40254_g4	sp|Q8WZ42|TITIN_HUMAN	5.7	4.2	2.7e-46	1.8e-42	6.9e-23	2.9e-18
c40276_g1	sp|Q8TGM6|TAR1_YEAST	4.3	13	1.9e-45	1e-41	5e-17	2.1e-13
c39865_g1	sp|Q3E811|RRT15_YEAST	4.2	13	1.8e-44	7.9e-41	1.1e-16	4.1e-13

Table 5.7: Differential Gene Expression (2vs3)

Gene	Symbol	log_ FC	log_ CPM	deseq.p -value	deseq.adj. pvalue	edger. p.value	edger.adj. p.value
c38943_g1	sp|Q6YHK3|CD109_HUMAN	4	4.8	2.7e-25	7.7e-21	2.7e-16	3e-12
c40482_g6	.	4.4	3.4	8.8e-24	1.2e-19	1.1e-17	1.6e-13
c28550_g1	.	3.8	4.2	3e-22	2.8e-18	6.1e-15	5.4e-11
c38820_g1	sp|Q9NPA2|MMP25_HUMAN	3.6	4	3.6e-20	2.5e-16	8.2e-14	5e-10
c16504_g1	.	6	1.8	5.2e-20	2.9e-16	8.2e-20	3.6e-15
c36468_g1	.	5	2.2	8.7e-20	4.1e-16	4.4e-18	9.7e-14

- **log_FC**: log2 Fold Change of gene level expression
- **log_CPM**: log2 Counts Per Million of gene level expression
- **deseq.p-value**: DESeq nominal p-value
- **deseq.adj.pvalue**: DESeq False Discovery Rate (FDR) adjusted p-value
- **edger.p-value**: edgeR nominal p-value
- **edger.adj.pvalue**: edgeR False Discovery Rate (FDR) adjusted p-value

5.5 Discussion

Next generation sequencing (NGS) platforms, including RNA-Seq technologies, have led to vast improvements in the ability to conduct transcriptome-based research in various organisms. De novo sequencing technologies generate many short sequence reads of the transcriptome which can be assembled into longer contigs that approximate full-length transcripts. In the current study, the transcriptome of the cold water squid *L. opalescens* was assembled on normalized reads using the Trinity assembler. Data normalization is one of the most crucial steps of data processing in massively paralleled sequencing data. Studies on both simulated and real datasets suggest that Trinity assembler has high sensitivity and high specificity relative to other assemblers for transcriptome assembly (Ren et al., 2012) and consists of three sequential steps; Inchworm, Chrysalis and Butterfly.

Inchworm assembles the RNA-Seq data into unique sequences of transcripts, often generating full-length transcripts for a dominant isoform, but then reports just the unique portions of

alternatively spliced transcripts (Grabherr et al., 2011). Chrysalis clusters the Inchworm contigs into clusters and constructs complete de Bruijn graphs for each cluster. Each cluster represents the full transcriptional complexity for a given gene (or sets of genes that share sequences in common). Chrysalis then partitions the full read set among these disjoint graphs. De Bruijn graph-based assemblers are able to handle the huge size of an NGS output in a relatively short time and space efficacy. Butterfly then processes the individual graphs in parallel, tracing the paths of reads and pairs of reads within the graph, ultimately reporting full-length transcripts for alternatively spliced isoforms, and teasing apart transcripts that corresponds to paralogous genes (Haas et al., 2013). Current protocols employed in RNA-seq are based on an mRNA fragmentation approach prior to sequencing. This fragmentation process leads to much higher sequence coverage of the entire transcript (Oshlack and Wakefield, 2009). The N50 value, an important statistical parameter, is defined as the length of the longest transcript such that all transcript of at least that length compose at least 50% of the bases of the assembly (Miller et al., 2010). The motivation for this measure is that better assemblies will result from a larger number of identified overlaps between the input reads and thus will have more reads assembled into longer transcripts (Miller et al., 2010).

A primary task in RNA-Seq data analysis is the detection of differentially expressed genes. To achieve this, RNA-Seq read counts can be obtained for each non-overlapping gene. Read counts are, to good approximation, linearly related to the abundance of the target transcript (Mortazavi et al., 2010): if reads were independently sampled from a population with given, fixed fractions of genes, read counts would follow a multinomial distribution, approximated by the Poisson distribution. Thus, statistical tests can decide whether, for a given gene, an observed difference in read counts is significantly greater than natural random variation expectation (Oshlack and Wakefield 2009). In this situation, since all the RNA populations were obtained from a single organism and were not differentially exposed to stressors, the % of DEG reflects the reliability of the analysis; that we found few DEG suggests high accuracy of the transcriptome assembly.

5.6 Conclusion

The current study reports for the first time RNA-Seq data on *Loligo opalescens*. High throughput sequencing is becoming more popular and less expensive and likely to become the platform of choice for most transcriptome analysis and provide new insight into the complexity of biological systems. These data can complement experimental data and improve our understanding of *Loligo* biology. Further analyses of this dataset can lead to discovering novel protein coding genes and offers the potential to identify genes that are important in regulating key aspect of the *Loligo opalescens* life cycle.

CONNECTING STATEMENT 4

The previous Chapter described the assembly of the transcriptome of *Loligo opalescens*. The results of this study provided new insight into the complexity of *Loligo opalescens* transcriptome. Building on a successful *de novo* sequencing and assembly of this transcriptome, the tools of bioinformatics were used in Chapter VI to find partial sequences of a chymotrypsin-like enzyme and used molecular biology techniques to experimentally amplify the full-length cDNA that encodes chymotrypsin.

The results of the current study will be submitted for publication.

CHAPTER VI

Combining molecular biology techniques and bioinformatics tools to isolate a novel chymotrypsin gene in cold adapted squid (*Loligo opalescens*)

6.1 Abstract

A cDNA encoding California squid (*Loligo opalescens*) chymotrypsin was identified using de novo sequencing, Trinity software and BLAST algorithms in NCBI. Its deduced amino acid sequence consists of 292 amino acid residues, being longer than vertebrate analogs. A search of the non-redundant protein database showed highest identity to a hypothetical protein from *Octopus bimaculoides* (69%), a chymotrypsinogen A-like protein from *Lingular anatina* (45%) and a serine protease from *Aplysia californica* (45%). Generation of a maximum likelihood phylogenetic tree of *Loligo opalescens* with other vertebrate and invertebrate chymotrypsin sequences suggest two main groups representing chymotrypsin from vertebrates and invertebrates and suggests that *Loligo opalescens* shares a common ancestor with most insect chymotrypsins.

Keywords: Massively parallel sequencing, bioinformatics, Loligo opalescens, chymotrypsin

6.2 Introduction

Molecular biology has provided powerful techniques for high-throughput nucleic acid analysis. For the past 30 years, Sanger sequencing and fluorescence-based electrophoresis technologies have been extensively used in somatic and germline genetic studies (Reis-Filho, 2009). However, with the evolution of technology at unprecedented pace, there has been a shift from these traditional methods to more robust massively parallel sequencing techniques (also known as next generation sequencing). This has made it easier for the buildup of qualitative and quantitative information about any type of nucleic acid in a given sample at an incredible throughput and also at relatively low cost (Korf, 2004).

Next-generation sequencing refers to non-Sanger-based high-throughput DNA sequencing technologies in which millions or billions of DNA strands can be sequenced in parallel, yielding substantially more throughput and minimizing the need for the fragment-cloning methods that are often used in Sanger sequencing of genomes. This method generate much shorter reads (~21 to ~400 bp), but millions of them compared to long reads generated from PCR-amplified samples (Stratton et al., 2009; Morozova and Marra, 2008). Various approaches have been used to accurately predict genes. In the past, the process mostly involved time consuming experimentation on living cells. With recent advancements in computational biology and f software, the requirement for painstaking experimentation on living cells to attain gene identification has largely been overcome.

Chymotrypsin belongs to one of the largest gene families, the serine proteases (SP), with the catalytic triad consisting of His, Asp and Ser (Zhou et al., 2012). In the current study, we combine RNA-Seq technology to assemble a *Loligo opalescens* transcriptome de novo and use software and databases, including Trinity and NCBI blast family algorithms, to predict a full-length chymotrypsin cDNA and validate this prediction using experimental data. These data are expected to help interpret experimental observations on the catalytic activity and optimum activity conditions of this enzyme at the structural level.

6.3 Materials and Methods

Squid samples were purchased frozen from OCN Imports, a local fish market in Montreal, QC, Canada. Samples were frozen and stored at -80°C pending further analyses.

6.3.1 Total RNA extraction

Total RNA was extracted from the squid (*Loligo opalescens*) muscle using the Trizol reagent (Invitrogen, USA) according to the manufacturer's protocol. Potential genomic contamination was removed by DNase I treatment. Treated RNA was purified using an RNeasy MinElute clean up kit (Qiagen) according to the manufacturer's instruction. RNA quality was determined by agarose gel electrophoresis and quantification was done with an ND-1000 NanoDrop UV spectrophotometer (NanoDrop Technologies, Wilmington, USA).

6.3.2 Verification of *Loligo* specie used in study

Amplification of partial CO1 gene in this study was done as a positive control and to confirm the specie of *Loligo* used in the study. Amplifications were performed on 5µl of the resulting cDNA and genomic DNA. In amplifying the partial CO1 gene forward primer 5′-ACA AAT CAT AAA GAT ATT GG-3′and reverse primer 5'-GTA AAT ATA TGT TGG GCT CA-3' were used. The amplification was carried out in a 50µl reaction mixture containing 5 µL of 10× buffer, 1 mM $MgCl_2$, 1 µl of 10mM deoxyribonucleoside triphosphate (dNTP), 1 µl of each primer at 10µmole concentration, 0.5µl of Taq polymerase, 5µl of DNA template and 36.5µl RNase treated water. The reaction mixtures were preheated at 95°C for 5 min, followed by 30 cycles of amplification (95°C for 40 s, 50°C for 30 s, and 72°C for 1.5 min), final extension at 72°C for 1 min and holding at 4°C. PCR products (20µl each) were resolved on a 1% agarose gel using a 1kb ladder as a marker. Bands of the expected size were cut from the gel and purified using the EZ-10 spin columns and gel extraction kits according to manufacturer's instructions. Purified samples were then sequenced using the Sanger method.

6.3.3 De novo sequencing and assembly of *Loligo* transcriptome

Transcripts used in this study were results obtained in de novo sequencing and assembly reported in chapter V

6.3.4 Gene annotation and prediction

All contigs > 32 kb were searched for putative protein genes by aligning against the uniprot_sprot_2013_11 protein database using the blastx program from the NCBI BLAST family. For each contig, the best BLAST hit in the NCBI non redundant (nr) database was used to annotate proteins

6.3.5 Chymotrypsin gene prediction

To optimize gene predictions for members of the S1 peptidase family, the BLASTp algorithm was re-run on all data sets using amino acid sequences of the vertebrate family S1 serine proteases as queries. The best hits for predicted S1 protein coding genes were compared to annotated proteins deposited in the NCBI nr protein database to identify genes most similar to chymotrypsin.

6.3.6 Amplification and cloning of the full chymotrypsin gene

To test whether these genes were transcribed in *Loligo*, three gene-specific primers were designed (5'-CATGGGGAGATGGACACACC-3'; 5'-CCACACGTGTAAAGACGCCA-3' and 5'-TCAGGATGGTCCTTGGCTTGT-3') to bind different parts of the gene on the reverse strand. A nested PCR approach was used to obtain full-length gene using the APAgene GOLD genome walking kit (RT) (BIO SandT Inc, Montreal, QC, Canada) in accordance to the manufacturer's instruction. Respective amplicons were cloned into pGEMT Easy Vector System (Promega, Madison, WI) for propagation in DH4α *E. coli* competent cells. Purified plasmids were sequenced by Sanger methods using T7 forward and SP6 reverse primers (Invitrogen Corp., San Diego, CA) at the McGill University and Génome Québec Innovation Centre, Montreal Canada).Once the identity of the gene was confirmed, we used reverse transcription-PCR to amplify the full gene using two locus-specific primers (forward primer 5'-ATGGCACTTATAGGAGGTCACTC-3' and reverse primer 5' ACTATACAATCTCATTGACT-3'

6.3.7 Sequence alignments and reconstruction of phylogenetic trees

Clustal Omega (Sievers et al., 2011) was used to align the protein sequences of vertebrate and invertebrate members of the S1 serine protease family. Maximum likelihood phylogenetic trees were constructed using the one-click online tool at Phylogeny.fr (Guindon et al., 2010; Dereeper et al., 2008; Castresana, 2000).

6.4 Results and Discussion

3.1 Amplification of Mt CO1 gene

To test the quality of the transcriptome assembly, ten independent clones of the mitochondrial CO1 gene were sequenced on both strands. All ten yielded identical amplicons over the 1750 bp of sequence (Figure 6.1). Alignment of the obtained sequence using BLAST algorithm at the NCBI database showed the query sequence share 100% sequence identity to partial CO1 gene from *Loligo opalescens*. This result confirmed the specie of *Loligo* in this study to be opalescens.

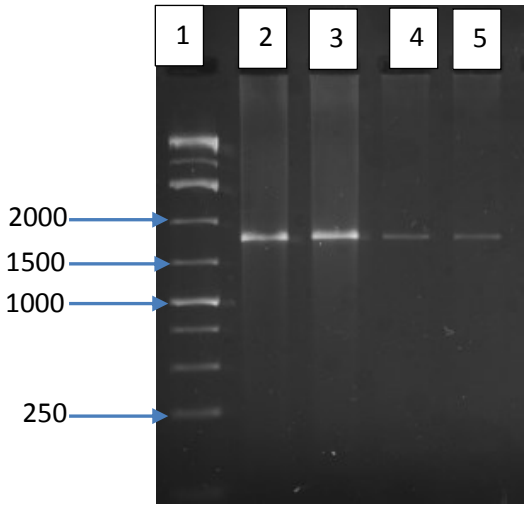

Figure 6.1: Amplicons (lane 2-5) from RT-PCR of mtCO1 gene resolve by electrophoresis through a 1% agarose gel. Lane 1: 1 kb ladder

6.4.3 Gene annotation and domain prediction

Alignment of transcripts against the uniprot_sprot_2013_11 protein database using the blastx program from the NCBI BLAST family revealed two genes (c25528_gl_il and c25528_g2_il) with significant homology to the S1 serine protease family. Both sequences are predicted to

encode domains belonging to the trypsin-like superfamily of enzymes (Figures 6.2 and 6.3). The first sequence was shown to be an N-terminal sequence (Figure 6.2B). The second sequence encoded a putative substrate binding site at positions 342-344, 417-419, and 423-425.

>c25528_gl_il
AAATCCCAGAATGCATCAGACTTTGTGCACTTCCTGTTTATTCCTCTTCCTGTCATTGTGTACTGTACTA
ACAATGGGGACAGTAACTGAAACAGTCCACAATGACACAAATTTTCACCCACCTGTGGCAATGGCACTTA
TAGGAGGTCACTCAATACCCCGAGGACGCTGGCCATGGCTGGTCTCTTTACAGTCAAAGGTTGTCATCAA
TCGAATATTTGGCATCTTTCCGGTCTACAGATACTATTGGTGTGGAGGATCACTCATCTCTGATCAGTGG
GTTATGTCTGCAGCTCATTGCTTTTTTGGTGAAAACTCAAAAGCAATTCCTGCAACAAGTTGGACAGCTC
GAATGGCAACAAATTCTTTAAAACCAAACATTAGAGAAAGG

Figure 6.2A: Nucleotide sequence of the putative peptidase transcript c25528_gl_il.

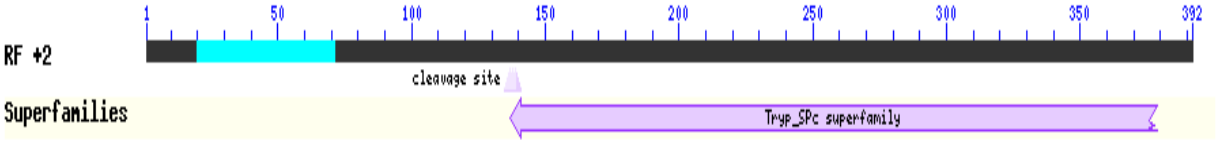

Figure 6.2B: Putative conserved domains in sequence >c25528_gl_il above using the blastx suite in NCBI database

>c25528_g2_i1
GCCAATGGAATGTCAAAGTTGAAAGAATAATTGTTTATCCTGACTTCCAAATAAATGATTTTCATGACGA
TATTGCACTGTTAAAGTTGACACACCCAGTTCCAACTGAAATTATATCAACTCTTCAGCCAGTTCCTTTG
GCTCAAAAGGAGGATATCCTTTTCCCTGATGCAGGAACTCTTTGTATGATGAAAGGATGGGGTTGCACTG
AAGGTGGTGGCAGTGTAACGAAATATGCACAAGAACTGAGCCTTCCTATAATGTCCAATGATGAATGTTC
TCACTATTTTGGAGCTGTTAGTGATGAGAAAATCTGTGCTGGATTTACAAATCATGCAAAAGGAATATGC
AGAGGTGACAGTGGTGGTCCCCTGGTGTGTCCATCTCCCCATGGTTGGGTGCAGGTTGGAGTGGCATCAT
TTACAAGCAAGGACCATCCTGAAAATTATCCTGGCGTCTTTACACGTGTGGCCAAATACAGAGACTGGGT
GGACCAGTCAATGAGATTGTATAGTTGATCATCTGTTCTTTAGTGAATGGGCCATCATCAGGATCCATAC
GTCTTACTATTGTTCTCAAATAATTCCACATTGTTTAC

Figure 6.3A: Nucleotide sequence of putative S1 peptidase transcript c25528_g2_il. Codons for the amino acids in the catalytic triad are highlighted in green

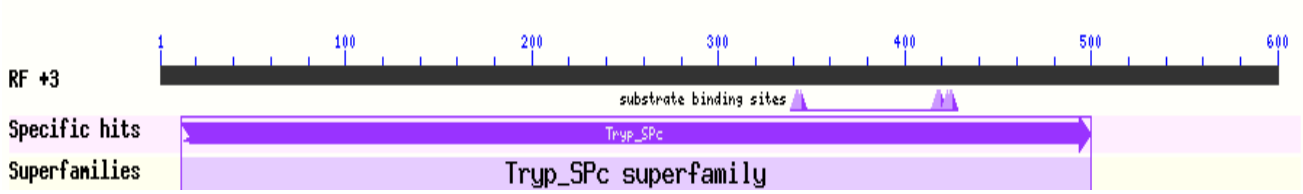

Figure 6.3B: Putative conserved domains in sequence >c25528_g2_il above using the blastx suite in NCBI database

6.4.4 Amplification and cloning of a full-length cDNA for a chymotrypsin-like protein

Nested PCR amplification showed that the two sequences were both part of a single mRNA with an internal fragment missing. The fragment was identified after Sanger sequencing of nested and RT-PCR products as 5'-gttctacacgtctttggaaaaatcttccgaaagatgaaatggcgccaatggaat-3'. The intact sequence contained an open reading frame (ORF) and a stop codon and consisted of 876 nucleotides (292 amino acids) (Figure 6.4). A search of the non-redundant protein database showed highest identity to a hypothetical protein from *Octopus bimaculoides* (69%), a chymotrypsinogen A-like protein from *Lingular anatina* (45%) and a serine protease from *Aplysia californica* (45%). The amino acid sequence contains the three conserved residues of the catalytic triad present in serine proteases (His, Asp and Ser) corresponding to His 75, Asp 144 and Ser 241 in the *Loligo* sequence.

```
    M  G  T  V  T  E  T  V  H  N  D  T  N  F  H  P  P  V  A  M
  1 ATGGGGACAGTAACTGAAACAGTCCACAATGACACAAATTTTCACCCACCTGTGGCAATG 60

    A  L  I  G  G  H  S  I  P  R  G  R  W  P  W  L  V  S  L  Q
 61 GCACTTATAGGAGGTCACTCAATACCCCGAGGACGCTGGCCATGGCTGGTCTCTTTACAG 120

    S  K  V  V  I  N  R  I  F  G  I  F  P  V  Y  R  Y  Y  W  C
121 TCAAAGGTTGTCATCAATCGAATATTTGGCATCTTTCCGGTCTACAGATACTATTGGTGT 180

    G  G  S  L  I  S  D  Q  W  V  M  S  A  A  H  C  F  F  G  E
181 GGAGGATCACTCATCTCTGATCAGTGGGTTATGTCTGCAGCTCATTGCTTTTTTGGTGAA 240

    N  S  K  A  I  P  A  T  S  W  T  A  R  M  A  T  N  S  L  K
241 AACTCAAAAGCAATTCCTGCAACAAGTTGGACAGCTCGAATGGCAACAAATTCTTTAAAA 300

    P  N  I  R  E  R  V  L  H  V  F  G  K  I  F  R  K  M  K  W
301 CCAAACATTAGAGAAAGGGTTCTACACGTCTTTGGAAAAATCTTCCGAAAGATGAAATGG 360

    R  Q  W  N  V  K  V  E  R  I  I  V  Y  P  D  F  Q  I  N  D
361 CGCCAATGGAATGTCAAAGTTGAAAGAATAATTGTTTATCCTGACTTCCAAATAAATGAT 420

    F  H  D  D  I  A  L  L  K  L  T  H  P  V  P  T  E  I  I  S
421 TTTCATGACGATATTGCACTGTTAAAGTTGACACACCCAGTTCCAACTGAAATTATATCA 480

    T  L  Q  P  V  P  L  A  Q  K  E  D  I  L  F  P  D  A  G  T
481 ACTCTTCAGCCAGTTCCTTTGGCTCAAAAGGAGGATATCCTTTTCCCTGATGCAGGAACT 540

    L  C  M  M  K  G  W  G  C  T  E  G  G  G  S  V  T  K  Y  A
541 CTTTGTATGATGAAAGGATGGGGTTGCACTGAAGGTGGTGGCAGTGTAACGAAATATGCA 600

    Q  E  L  S  L  P  I  M  S  N  D  E  C  S  H  Y  F  G  A  V
601 CAAGAACTGAGCCTTCCTATAATGTCCAATGATGAATGTTCTCACTATTTTGGAGCTGTT 660

    S  D  E  K  I  C  A  G  F  T  N  H  A  K  G  I  C  R  G  D
661 AGTGATGAGAAAATCTGTGCTGGATTTACAAATCATGCAAAAGGAATATGCAGAGGTGAC 720

    S  G  G  P  L  V  C  P  S  P  H  G  W  V  Q  V  G  V  A  S
721 AGTGGTGGTCCCCTGGTGTGTCCATCTCCCCATGGTTGGGTGCAGGTTGGAGTGGCATCA 780

    F  T  S  K  D  H  P  E  N  Y  P  G  V  F  T  R  V  A  K  Y
781 TTTACAAGCAAGGACCATCCTGAAAATTATCCTGGCGTCTTTACACGTGTGGCCAAATAC 840

    R  D  W  V  D  Q  S  M  R  L  Y  S  *
841 AGAGACTGGGTGGACCAGTCAATGAGATTGTATAGTTGA 879
```

Figure 6.4: Nucleotide and predicted amino acid sequence of *Loligo opalscens* chymotrypsin

6.4.5 Sequence alignments

Alignment of the *Loligo* chymotrypsin sequence shows that the first 20 or so amino acids of all sequences are quite variable and are likely to be signal peptides. Regions that are highly conserved are highlighted in blue, while regions in grey and white are poorly conserved. The binding pocket residues in *Loligo* chymotrypsin are Gly 235, Ser 260 and Thr 262. Chymotrypsins have been shown to possess variable binding pocket residues, but usually consist of Ser, Gly, and Ala/Gly (Jiang et al., 1997). Gly 216, which lines the entry of the binding pocket, is conserved.

Figure 6.5: Multiple sequence alignment of Loligo opalesencs chymotrypsin with chymotrypsins form bovine*, Heterololigo bleekeri* (D2KX88), Atlantic salmon (B5XB02), Atlantic cod (P80646), cotton ball worm (O18445), and *Spodoptera* (E2D741)

6.4.6 Phylogenetic analysis

Maximum likelihood analysis of *L. opalescens* chymotrypsin with homologs from vertebrates and invertebrate chymotrypsin sequences (Figure 6.6) suggests that there are two main groups representing chymotrypsin, from vertebrates and invertebrates. The divergence event that separated the lineage of vertebrates and invertebrates gave rise to *Tenebrio molitor* and all the other species. *Loligo opalescens* also shares a common ancestor with most insect chymotrypsins.

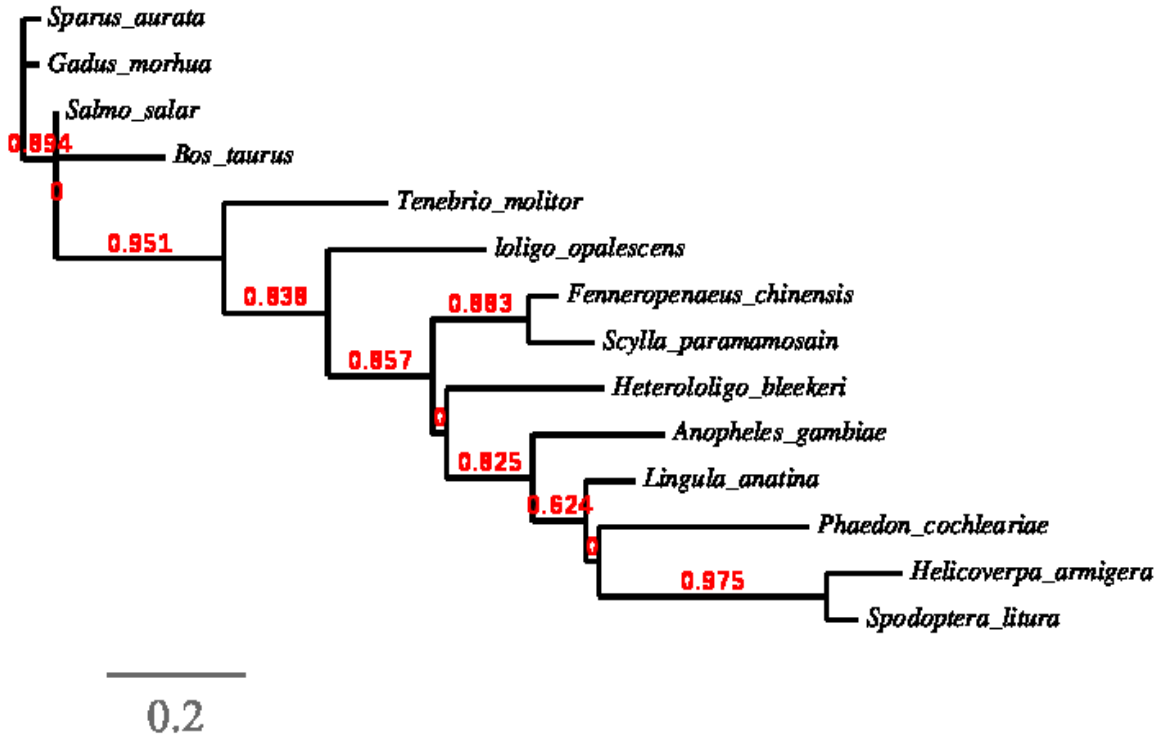

Figure 6.6: Maximum likelihood phylogenetic tree of *Loligo* with other chymotrypsin enzymes. Branch length is proportional to the number of substitutions per site

The branch support values shown are the expectation values and give an idea that a branch is statistically realistic or reliable. The closer the support value is to 1, the more reliable that branch is. The branch that separates vertebrates from invertebrates has a branch support value of 0.951, making it both statistically significant and realistic.

6.5 Conclusion

Eukaryotic transcriptomes are believed to contain a wealth of information at the gene level with regard to transcription, translation and other related processes that occur in the cell. With recent advancements in both molecular biology techniques and computational biology, it is expected that more genomes will be available in the near future. The current study provides comprehensive data on de novo gene sequencing and the use of computational biology to find a novel chymotrypsin from a cold adapted squid, *L. opalescens* and shows how this enzyme is related to other chymotrypsins through phylogeny. Also now we can produce recombinant enzyme for characterization and possible application to food and etc.

CONNECTING STATEMENT 5

Three-dimensional structures of proteins give a wealth of information on functionality. In this chapter, the 3 dimensional structure of the *Loligo opalescens* chymotrypsin is modelled based on the x-ray resolved structure of bovine chymotrypsin. The quality of the model is analyzed using various statistical methods. The amino acid components of the template and query sequences are compared for similarities and differences.

CHAPTER VII

Three-dimensional structure of a novel chymotrypsin from squid (*Loligo opalescens*) as predicted by homology modelling

7.1 Abstract

Chymotrypsins have been thoroughly studied, but mostly in mammals. In contrast, very little is known about chymotrypsins from aquatic sources, especially from aquatic mollusks, including squids. In this study, a homology model of *Loligo opalescens* chymotrypsin was built using the crystal structure of bovine chymotrypsin (PDB ID: 1t8o) as a scaffold. The model was assessed for stereochemical quality and side chain environment. Sequence alignment shows that the target and the template (1t8o) share 36% sequence identity. The model predicts the characteristic 2 β barrel domains typical of chymotrypsin. The catalytic triad involving Histidine, Aspartate and Serine, as well hydrophobic and electrostatic interactions typical of serine proteases, were conserved. The Ramachandran plot using 258 residues gave a total stereochemistry fit of 88.5% with 8.3% in the additional allowed region. The percentage of residues found in the disallowed region was 0.9%. The 3D structure of *Loligo* chymotrypsin based on crystallographic data for bovine chymotrypsin showed them to be almost superimposable proteins.

Keywords: Homology modeling, chymotrypsin, bioinformatics, squid

7.2 Introduction

Chymotrypsins are members of S1 serine peptidase family (Copeland, 2000) produced by pancreatic acinar cells as inactive chymotrypsinogen in mammals containing a signal sequence of 16–18 residues and a 15-residue pro-peptide (Smillie et al., 1968). Activation of chymotrypsinogen occurs when trypsin cleaves off the pro-peptide, which remains attached to the mature enzyme by a disulfide bond between Cys1 and Cys122 (Hedstrom, 2002). This disulfide linkage plays a role in keeping the zymogen stable against nonspecific activation (Venekei et al., 1996). The active enzyme catalyzes hydrolysis of ester / amine bonds formed by L isomers of aromatic residues (tryptophan, phenylalanine or tyrosine). The main player in the catalytic mechanism is the catalytic triad located in the active site of the enzyme, consisting of histidine, aspartic acid and serine at positions 57, 102 and 195 in bovine chymotrypsin. These three residues form a charge relay system that serves to make the active site serine nucleophilic during proteolysis

Chymotrypsins from different sources display different optimal conditions for activity mostly due to differences in amino acid composition (Hedstrom, 2002). For instance, chymotrypsin from Atlantic cod has higher activity at low temperatures and increased catalytic efficiency (k_{cat}/K_m) compared to analogs from mesophilic organisms (Spilliaert and Gudmundsdóttir 2000). In addition, the fish enzymes are less thermos- and acid stable (Simpson, 2000).

Experimental studies have shown a lower optimal temperature for a squid chymotrypsin compared to mammalian homologs. It is important to understand the basis of this difference with respect to both amino acid composition and 3D structure and elucidate how the differences may confer such properties.

7.3 Materials and Methods

7.3.1 Primary sequence of chymotrypsin gene

Primary sequence used in this study was obtained from results of previous study in chapter VI.

7.3.2 PCR amplification and cloning of products

Full-length sequence was obtained using the APAgene GOLD genome walking kit (RT) (BIO SandT Inc, Montreal, QC, Canada) in accordance to manufacturer's instruction. Respective amplicons were cloned into pGEMT Easy Vector System (Promega) for propagation in DH4α E. coli cells.

7.3.3 Template protein

To find an adequate template for homology modeling of *Loligo* chymotrypsin, alignments of its amino acid sequence against PDB (spell out) were performed using the BLAST algorithm in HHpred (Soding et al., 2005). Sequence alignment showed that the target and the template (1t8o) share 36% sequence identity. Because protein structures are more conserved than DNA sequences, detectable levels of sequence similarity usually imply significant structural similarity. Based on the significant e-value (6.2^{e-50}) and alignment among the investigated templates, bovine chymotrypsin complexed with P1 BPTI variants (PDB ID: 1t8o) was selected as the template to build the model.

7.3.4 Model building

The homology model of *Loligo* chymotrypsin was constructed using MODELER (version 9.15) (Sali et al., 1995). This program allows automatic generation of one or more hypothetical models based upon one or more templates with known crystal structures.

7.3.5 Model validation

To determine if the models generated were reasonable, three methods were used; VERIFY 3D, SOLVX, and ANOLEA (Melo and Feytmans, 1998; Luthy et al., 1992; Holm and Sander, 1992) and PDBsum, including full PROCHECK analysis, to check stereochemical quality by Ramachandran plots (Laskowski et al., 1993).

7.4 Results and Discussion

7.4.1 Full length chymotrypsin gene

The full-length cDNA sequence contained an open reading frame (ORF) of 816 base pairs (bp) coding for 272 amino acids (Figure 7.1). The protein shows highest local identity with other chymotrypsins with respect to active site residues and the substrate binding site important for catalysis.

```
    M  G  T  V  T  E  T  V  H  N  D  T  N  F  H  P  P  V  A  M
  1 ATGGGGACAGTAACTGAAACAGTCCACAATGACACAAATTTTCACCCACCTGTGGCAATG 60

    A  L  I  G  G  H  S  I  P  R  G  R  W  P  W  L  V  S  L  Q
 61 GCACTTATAGGAGGTCACTCAATACCCCGAGGACGCTGGCCATGGCTGGTCTCTTTACAG 120

    S  K  V  V  I  N  R  I  F  G  I  F  P  V  Y  R  Y  Y  W  C
121 TCAAAGGTTGTCATCAATCGAATATTTGGCATCTTTCCGGTCTACAGATACTATTGGTGT 180

    G  G  S  L  I  S  D  Q  W  V  M  S  A  A  H  C  F  F  G  E
181 GGAGGATCACTCATCTCTGATCAGTGGGTTATGTCTGCAGCTCATTGCTTTTTTGGTGAA 240

    N  S  K  A  I  P  A  T  S  W  T  A  R  M  A  T  N  S  L  K
241 AACTCAAAAGCAATTCCTGCAACAAGTTGGACAGCTCGAATGGCAACAAATTCTTTAAAA 300

    P  N  I  R  E  R  V  L  H  V  F  G  K  I  F  R  K  M  K  W
301 CCAAACATTAGAGAAAGGGTTCTACACGTCTTTGGAAAAATCTTCCGAAAGATGAAATGG 360

    R  Q  W  N  V  K  V  E  R  I  I  V  Y  P  D  F  Q  I  N  D
361 CGCCAATGGAATGTCAAAGTTGAAAGAATAATTGTTTATCCTGACTTCCAAATAAATGAT 420

    F  H  D  D  I  A  L  L  K  L  T  H  P  V  P  T  E  I  I  S
421 TTTCATGACGATATTGCACTGTTAAAGTTGACACACCCAGTTCCAACTGAAATTATATCA 480

    T  L  Q  P  V  P  L  A  Q  K  E  D  I  L  F  P  D  A  G  T
481 ACTCTTCAGCCAGTTCCTTTGGCTCAAAAGGAGGATATCCTTTTCCCTGATGCAGGAACT 540

    L  C  M  M  K  G  W  G  C  T  E  G  G  G  S  V  T  K  Y  A
541 CTTTGTATGATGAAAGGATGGGGTTGCACTGAAGGTGGTGGCAGTGTAACGAAATATGCA 600

    Q  E  L  S  L  P  I  M  S  N  D  E  C  S  H  Y  F  G  A  V
601 CAAGAACTGAGCCTTCCTATAATGTCCAATGATGAATGTTCTCACTATTTTGGAGCTGTT 660

    S  D  E  K  I  C  A  G  F  T  N  H  A  K  G  I  C  R  G  D
661 AGTGATGAGAAAATCTGTGCTGGATTTACAAATCATGCAAAAGGAATATGCAGAGGTGAC 720

    S  G  G  P  L  V  C  P  S  P  H  G  W  V  Q  V  G  V  A  S
721 AGTGGTGGTCCCCTGGTGTGTCCATCTCCCCATGGTTGGGTGCAGGTTGGAGTGGCATCA 780

    F  T  S  K  D  H  P  E  N  Y  P  G  V  F  T  R  V  A  K  Y
781 TTTACAAGCAAGGACCATCCTGAAAATTATCCTGGCGTCTTTACACGTGTGGCCAAATAC 840

    R  D  W  V  D  Q  S  M  R  L  Y  S  *
841 AGAGACTGGGTGGACCAGTCAATGAGATTGTATAGTTGA 879
```

Figure 7.1: Predicted amino acid sequence of Loligo chymotrypsin

A general comparison of chymotrypsin from *L. opalescens* to bovine chymotrypsin in terms of amino-acid composition (Table 7.1) shows higher content in the squid enzyme of methionine, arginine, histidine and glutamic acid, with lower content of threonine and serine, similar to features of most fish serine proteases (Male et al, 1995; Gudmundsdottir et al., 1996; Leth-Larson et al., 1996). In general, cold-adapted serine proteases contain a proportionally higher

number of methionine residues compared with mammalian homologues (Smalas et al., 1994). These results follow the same trend. Methionine offers greater side-chain conformational flexibility than most other residues (Gudmundsdottir et al., 1996). The total number of charged residues (Table 7.2) is higher in squid than bovine chymotrypsin, but the molar ratio of polar hydrogen bond-forming residues is lower. This may play a role in flexibility and account for the less thermostable structure (Somero and Hochachka, 1984) commonly observed in enzymes adapted to lower temperatures (Leth-Larson et al., 1996).

Table 7.1: Distribution of amino acids in *L. opalescens* and bovine chymotrypsin

	Total number of residues				
Residue	*L. opalescens*	Bovine	Residue	*L. opalescens*	Bovine
Alanine	16	22	Methionine	8	2
Cysteine	8	10	Asparagine	9	14
Aspartic acid	13	9	Proline	17	9
Glutamic acid	10	5	Glutamine	9	10
Phenylalanine	13	6	Arginine	14	4
Glycine	23	23	Serine	21	28
Histidine	9	2	Threonine	12	23
Isoleucine	19	10	Valine	21	23
Lysine	15	14	Tryptophan	10	8
Leucine	17	19	Tyrosine	9	4

Table 7.2: Distribution of properties in *L. opalescens* and bovine chymotrypsin primary sequence

property	*L. opalescens*	bovine
aliphatic	73	74
Aromatic	41	20
Non-polar	161	136
Polar	112	109
Charged	61	34
Basic	38	20
Acidic	23	14

An arg-pro substitution was observed in *Loligo* chymotrypsin at position 124 compared to the bovine enzyme (Figure 7.2). Proline at position 124 is highly conserved in mesophilic chymotrypsins (Gudmundsdottir et al., 1994) and its substitution by alanine in *Loligo* may have an effect on adaptation to cold temperatures in the squid enzyme. A similar substitution of has also been observed in a cod chymotrypsin A variant and chymotrypsins from salmon and spear squid, all organisms of aquatic habitat (Figure 2). Eight out of ten cysteines that form five disulfide bridges in bovine chymotrypsin are found in the *Loligo* homologue, suggesting some differences in flexibility of tertiary structure. Spilliaert and Gudmundsdóttir (2000) also observed eight cysteine residues that form four disulfide bridges in Atlantic cod chymotrypsin B. Three positions with uncharged residues (233, 235, 239 in bovine) are substituted with charged residues in the *Loligo* enzyme. A similar observation was made for cod chymotrypsins (Leth-Larson et al., 1996)

```
tr|D2KX88|D2KX88_HETBL      MLPFIYLISAILVST----------LSVQGSEVHHIVGGTKAKRCEFPHIVFIYTAK----
loligoopalescens            ----------MGTVTETVHND-TNFHPP--VAMALIGGHSIPRGRWPWLVSLQSKVVINR
Bovine                      -------------------CGVPAIQPVLSGLSRIVNGEEAVPGSWPWQVSLQDKT----
sp|P80646|CTRB_GADMO        -------------------CGSPAIQPQVTGYARIVNGEEAVPHSWPWQVSLQQSN----
tr|B5XB02|B5XB02_SALSA      -MAFLWFVSCLAFVSAAYGCGIPAIKPEVSGYARIVNGEEAVPHSWPWQVSLQQTS----
                                                :         :: * .     :*  * :

tr|D2KX88|D2KX88_HETBL      ------NGHYFGCGGSLIDNKHVLTAAHCMAGDVTKVDILIGSLDKR-------------
loligoopalescens            IFGIFPVYRYYWCGGSLISDQWVMSAAHCFFGENSKAIP--ATSWTARMATNSLKPNIRE
Bovine                      --------GFHFCGGSLINENWVVTAAHCGVTTSDVVVA--GEFDQGS-S-------SEK
sp|P80646|CTRB_GADMO        --------GFHFCGGSLINENWVVTAAHCNVRTYHRVIV--GEHDKAS-D--------EN
tr|B5XB02|B5XB02_SALSA      --------GFHFCGGSLINENWVVTAAHCNVATYHRVII--GEHKKGS-GN-----NAED
                                    :. ******.:: *::****       .    .

tr|D2KX88|D2KX88_HETBL      -TMPIW-------------APVARFIKNSKYAKLRSTVVNDIAVLTLAKPVR--FTSCIK
loligoopalescens            RVLHVFGKIFRKMKWRQWNVKVERIIVYPDFQ--INDFHDDIALLKLTHPVPTEIISTLQ
Bovine                      IQK----------------LKIAKVFKNSKYN--SLTINNDITLLKLSTAAS--FSQTVS
sp|P80646|CTRB_GADMO        IQI----------------LKPSMVFTHPKWD--SRTINNDISLIKLASPAV--LGTNVS
tr|B5XB02|B5XB02_SALSA      IQI----------------LKPAKVFTHPKWN--PSTINNDISLIKLSTPAV--LNTNVS
                                              .:   .:      . :**:::.*:  .   :   :.

tr|D2KX88|D2KX88_HETBL      PIRMATPDEAFV----GDCVIAGWGRTGFNLP-TSQILLRANVPIMDHATCANRLPI--I
loligoopalescens            PVPLAQKEDILFPDAGTLCMMKGWGCTEGGGS-VTKYAQELSLPIMSNDECSHYFGA--V
Bovine                      AVCLPSASDDF--AAGTTCVTTGWGLTRYTNANTPDRLQQASLPLLSNTNCKKYWGTK-I
sp|P80646|CTRB_GADMO        PVCLGESSDVF--APGMKCVTSGWGLTRYNAPGTPNKLQQAALPLMSNEECSQTWGNNMI
tr|B5XB02|B5XB02_SALSA      PVCLAETADVF--APGMTCVTTGWGLLRYNALNTPNELQQAALPLLSNEQCKTHWGSS-I
                             : :    : :       *:  ***        . .   .  :*::..  *        :

tr|D2KX88|D2KX88_HETBL      LKQHLCVGSGKILDTTTCKGDSGGPLMCKSAVDGSQVLAGIVSYGWKC--NTGLAVFTKV
loligoopalescens            SDEKICAGFTN-HAKGICRGDSGGPLVCPSP--HGWVQVGVASFTSKDHPENYPGVFTRV
Bovine                      KDAMICAGASG---VSSCMGDSGGPLVCKKN--GAWTLVGIVSWGSSTCSTSTPGVYARV
sp|P80646|CTRB_GADMO        SDVMICAGAAG---ATSCMGDSGGPLVCQKD--NVWTLVGIVSWGSSRCSVTTPAVYARV
tr|B5XB02|B5XB02_SALSA      SDVMICAGGAG---ATSCMGDSGGPLVCEKD--NVWTLVGIVSWGSSRCSTTTPAVYARV
                             .  :*.*         * *******:* .      . .*:.*:  .    .  .*:::*

tr|D2KX88|D2KX88_HETBL      SYYLDWVNSVRKLIP
loligoopalescens            AKYRDWVDQSMRLYS
Bovine                      TALVNWVQQTLAAN-
sp|P80646|CTRB_GADMO        TELRGWVDQILAAN-
tr|B5XB02|B5XB02_SALSA      TELRSWVDQTLAAN-
                            :    **:.
```

Figure 7.2: Multiple sequence alignment of chymotrypsin from different sources.

7.4.2 Model building

Sequence identity ≥ 25% between two proteins is indicative of similar 3-dimensional structures (Yang and Honig, 2000). Sequence alignment shows that the target and the template (1t8o) share 36% sequence identity (Figure 7.3). Because protein structures are more conserved than DNA sequences, detectable levels of sequence similarity usually imply significant structural similarity. Based on the significant e-value (6.2^{e-50}) and alignment among the investigated templates, bovine chymotrypsin complexed with P1 BPTI variants (PDB ID: 1t8o) was selected as a template to build the model. The model had the characteristic 2 β barrel domains typical of chymotrypsin (Figure 7.4).

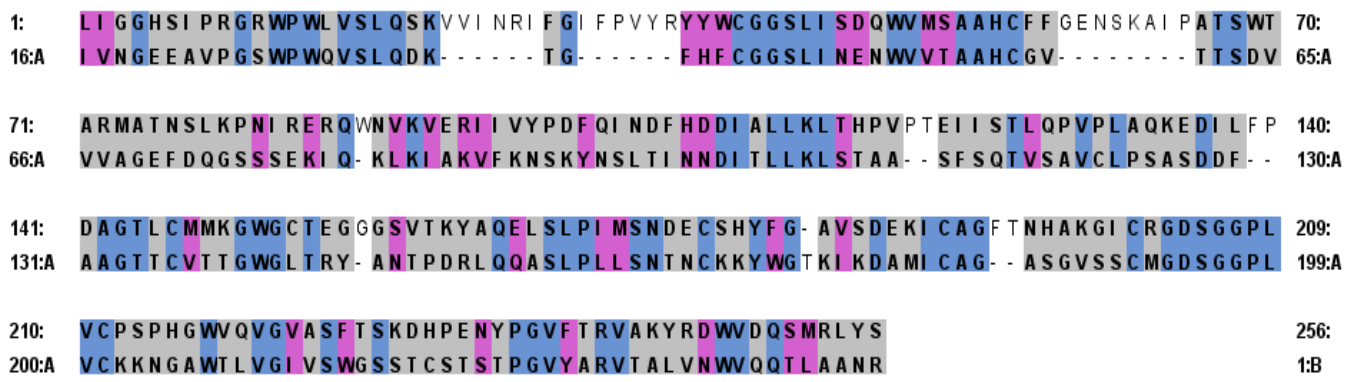

Figure 7.3: Sequence alignment between target and the template (1t8o) protein

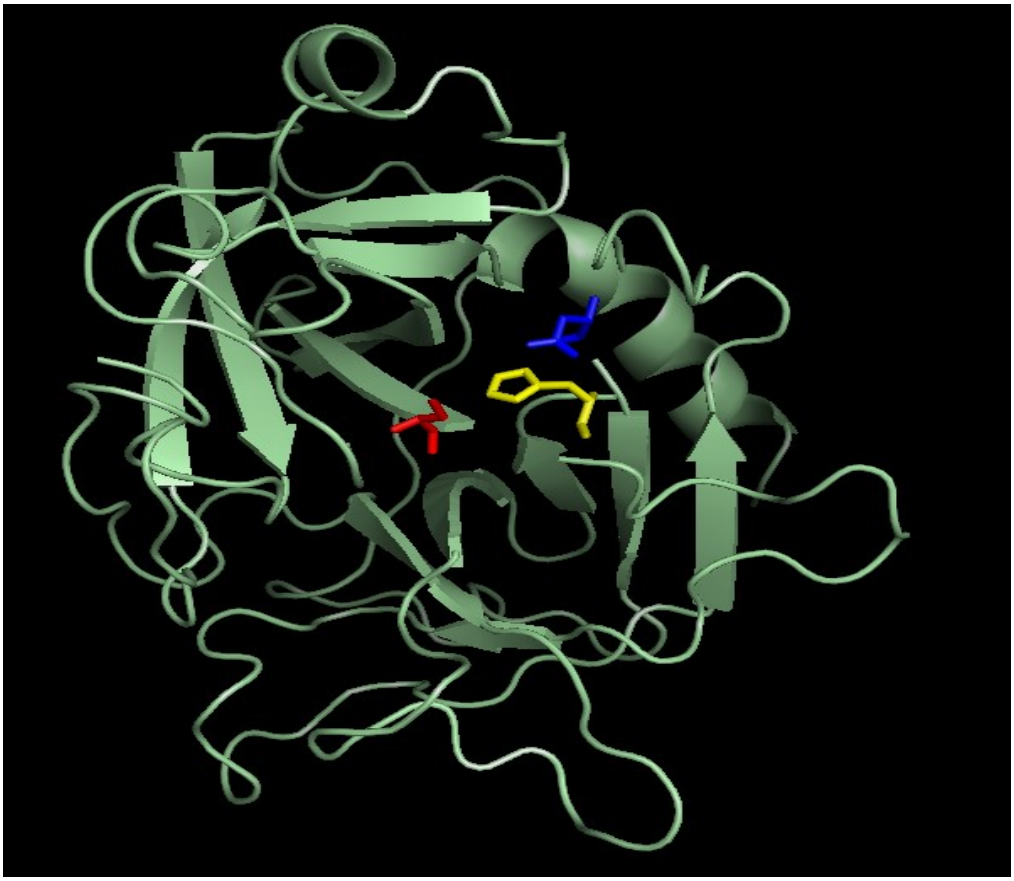

Figure 7.4: Cartoon representation of a 3-dimensional model of *L. opalescens* showing active site residues Histidine (yellow), Aspartate (blue), Serine (red)

7.4.3 Model quality

The Verify 3D method assesses protein structures using 3-dimensional profiles by analyzing the compatibility of an atomic model with its own primary amino acid sequence (Guex and Peitsch, 1997). The scores range from -1 (bad score) to +1 (good score). The scores for individual amino acids in the current study were all found in the very good region (Figure 7.5). Similarly for the SOLVX plot, most of the residues in the model were found in the good region (white region of the plot) (Figure 7.6). ANOLEA performs energy calculation on protein chains and evaluates the environment of each heavy atom in the molecule. Most of the residues in the model built were found in the favored region of the ANOLEA plot (Figure 7.7)

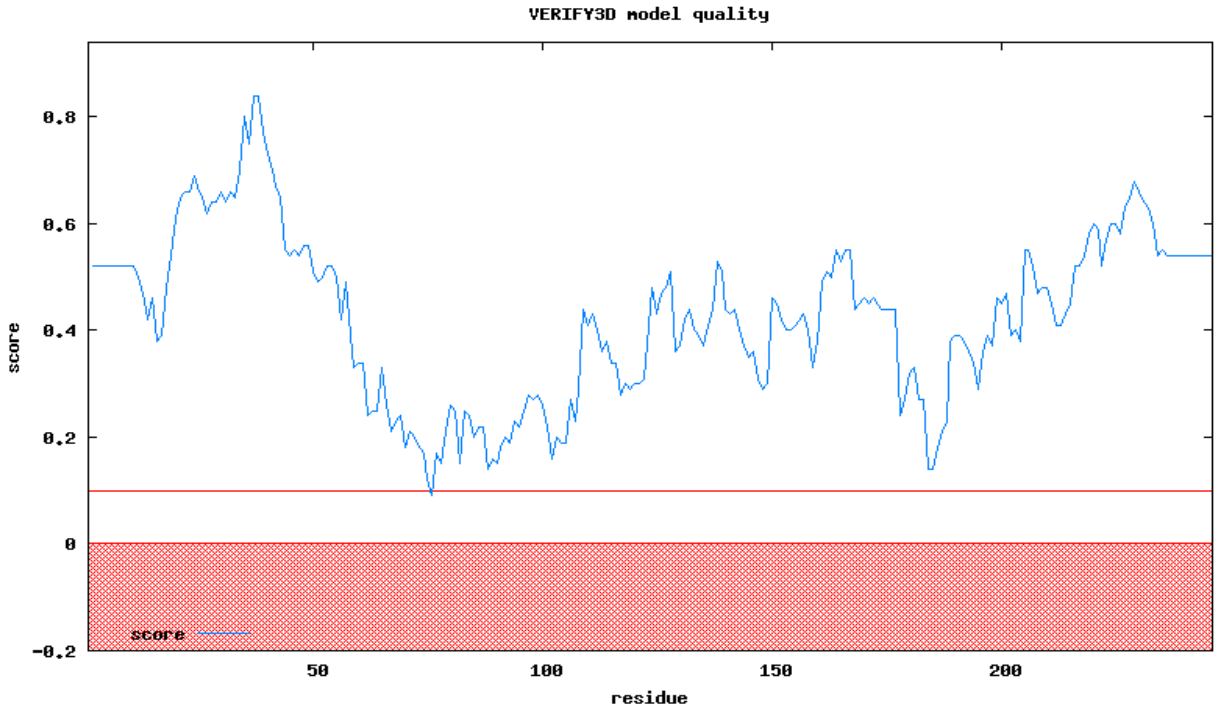

Figure 7.5: Verify 3D plot of model *Loligo* chymotrypsin built using bovine chymotrypsin as template.

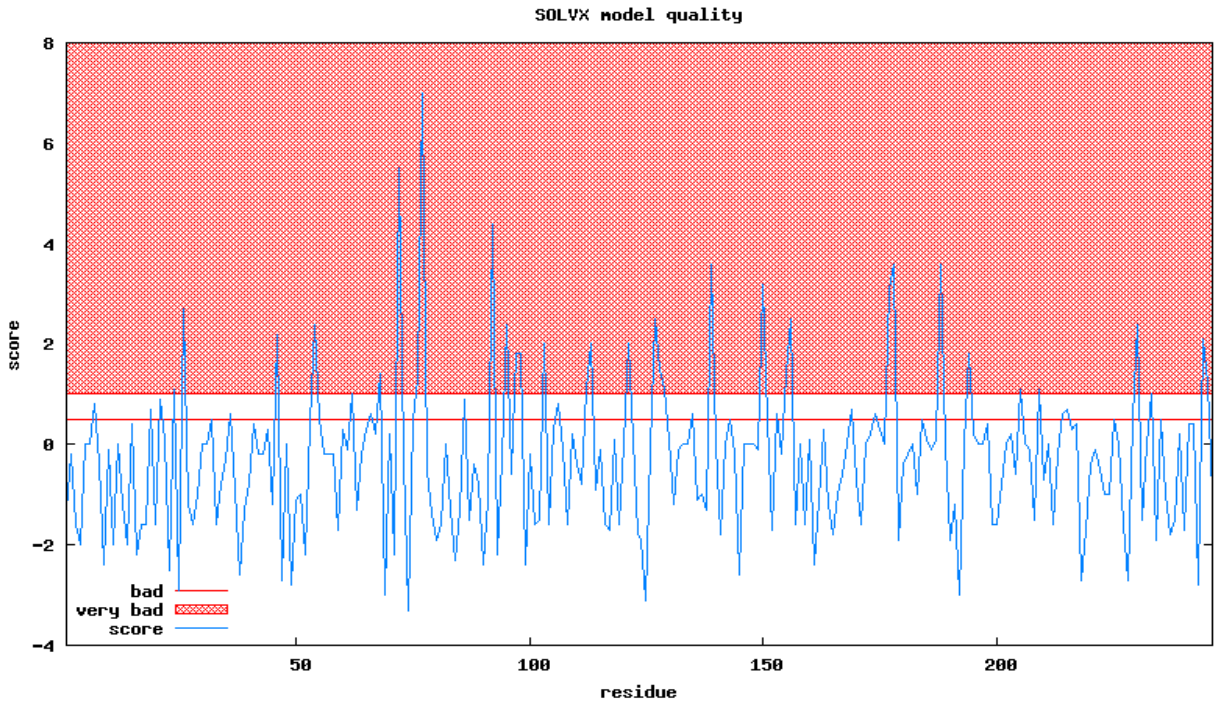

Figure 7.6: SOLVX plot of model Loligo chymotrypsin built using bovine chymotrypsin as template.

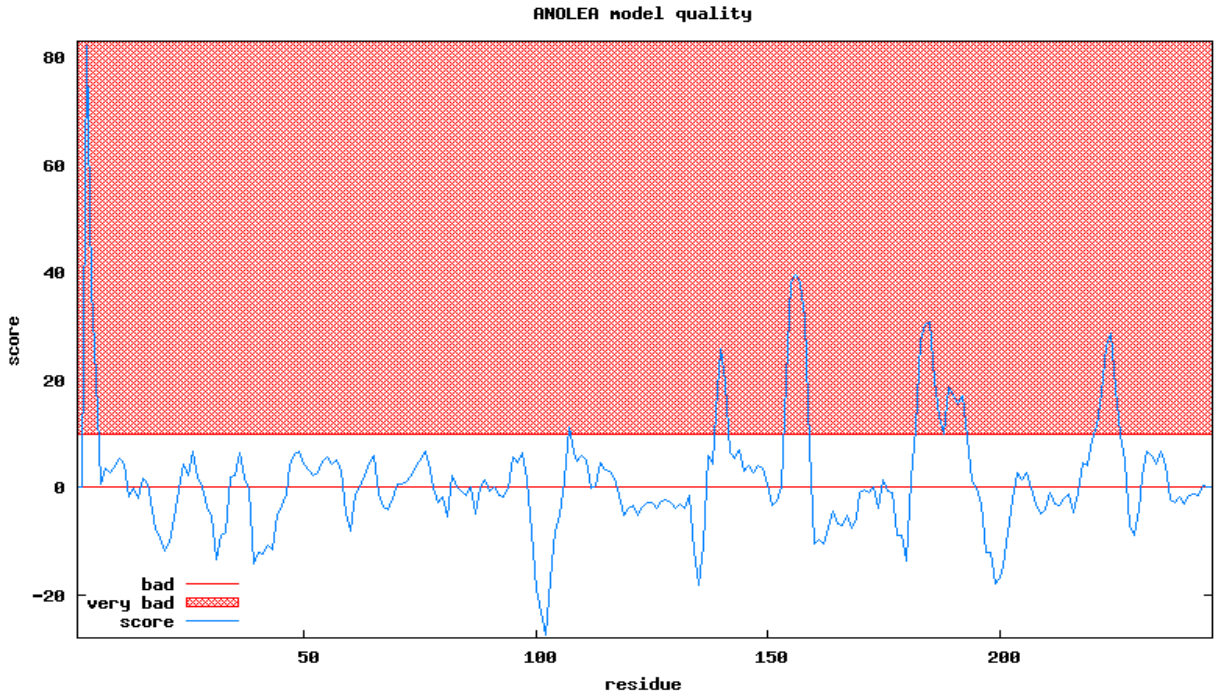

Figure 7.7: ANOLEA plot of model *Loligo* chymotrypsin built using bovine chymotrypsin as template

The Ramachandran plot using 258 residues gave a total stereochemistry of 88.5% with 8.3% in the additional allowed region (Table 7.3). The percentage of residues found in the disallowed region was 0.9% (2 residues) (Figure 7.8). The crystal structure on which the model was built was resolved at 1.7Å with 97.9% residues in the favored region. There were 3 outliers for the crystal structure. The 3D structure of *Loligo* chymotrypsin based on crystallographic data for bovine chymotrypsin showed almost superimposable proteins with an RMSD of 3.25. The 3D structure of the superimposed proteins (Figure 7.9) showed that *Loligo* chymotrypsin contains more loops compared to the bovine homologue. The presence of increased loops in cold adapted enzymes has been described as one of the main reasons accounting for their flexibility and higher catalytic activity at low temperatures.

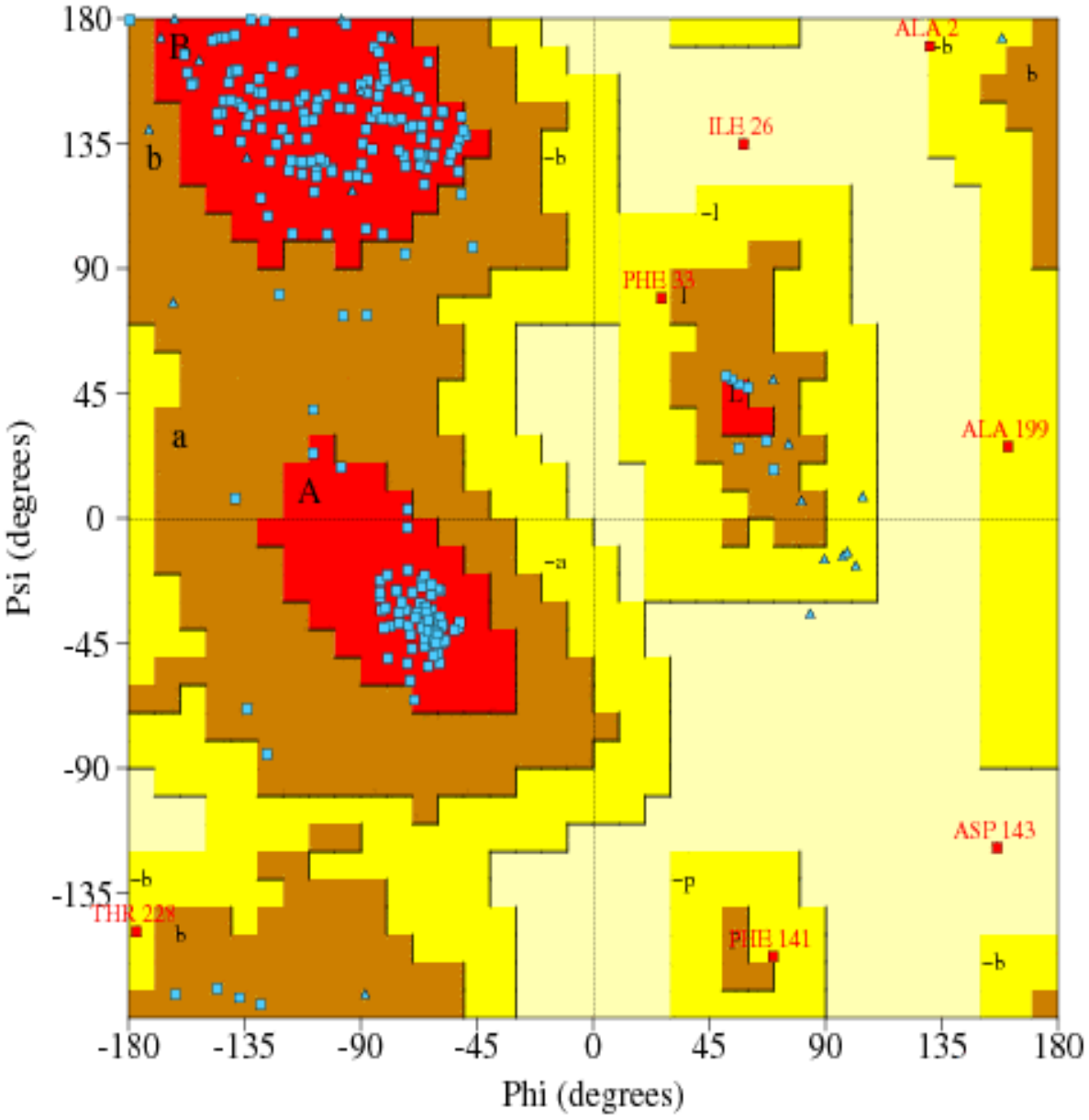

Figure 7.8: Ramachandran plot of *Loligo* chymotrypsin homology model

Table 7.3: Plot statistics of Ramachandran plot

Residues in most favoured regions [A,B,L]	192	88.5%
Residues in additional allowed regions [a,b,l,p]	18	8.3%
Residues in generously allowed regions [~a,~b,~l,~p]	5	2.3%
Residues in disallowed regions	2	0.9%
	----	------
Number of non-glycine and non-proline residues	217	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	22	
Number of proline residues	17	

Total number of residues	258	

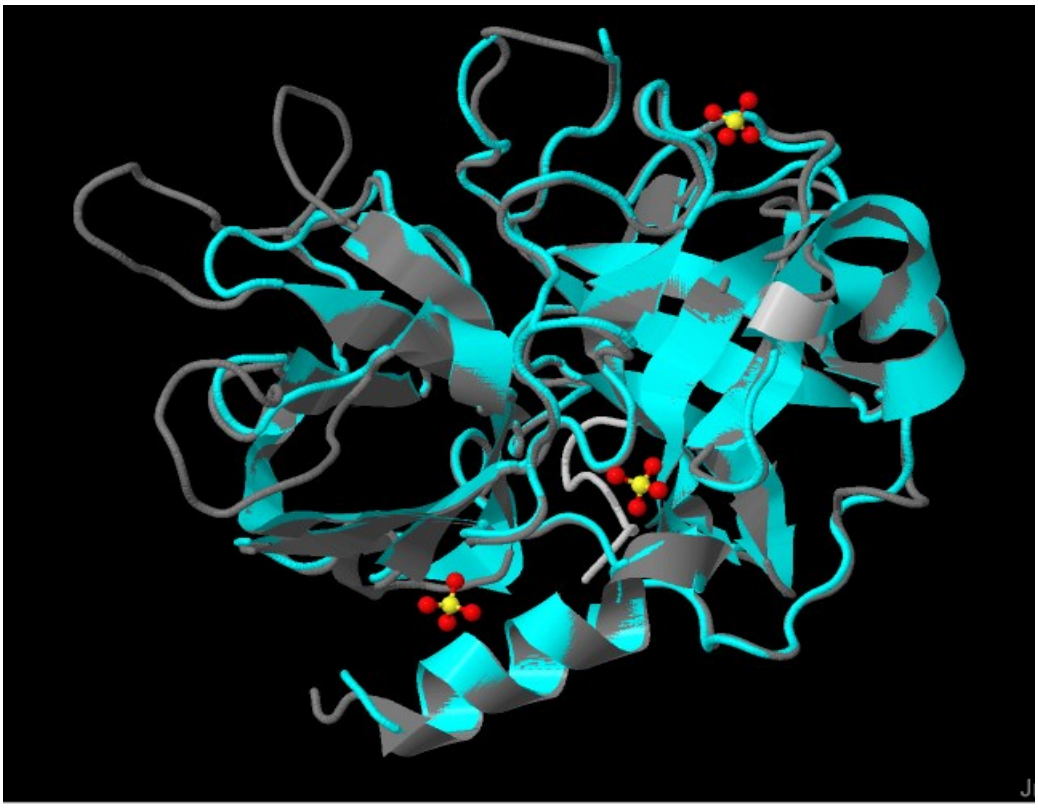

Figure 7.9: Structure of *Loligo* homology model superimposed on crystal model from bovine. Crystal structure in cyan, *Loligo* model in grey

7.5 Conclusions

The results of this study allow interesting comparisons to be made between a chymotrypsin enzyme from squid and a bovine homologue. Although *Loligo* chymotrypsin has several sequence characteristics not observed in mammalian homologues, only very tentative conclusions can be drawn as to which alterations affect catalytic and stability characteristics. Experiments using site-directed mutagenesis are required to provide experimental evidence for or against the suggestions made here

CHAPTER VIII: General conclusions, contribution to knowledge and recommendations for future work

8.1 General Conclusions

In the present study, two new chymotrypsin-like enzymes from the viscera of *Loligo opalescens* and *Sepioteuthis lessoniana* were purified and characterized based on molecular mass, inhibitor sensitivity, pH profile, temperature sensitivity and substrate specificity. The purified enzymes were homogeneous on SDS–PAGE, and their molecular masses were estimated to be 18 and 22 kDa. The enzymes displayed varying tolerance to temperature and pH. Overall, *L. opalescens* chymotrypsin demonstrated better adaptation to cold temperature and higher catalytic activity than its homologue from S*epioteuthis*. Three peptides identified from tryptic digestion of purified *Sepioteuthis* chymotrypsin were identical to peptides found in a serine protease from another cephalopod. This report enhances knowledge on the distribution of chymotrypsin-like enzymes among the class Cephalopoda, family Loliginidae.

The use of degenerate primers derived from MS –based peptide sequences to amplify genes has been demonstrated in this current study as an alternative strategy when information on gene sequence is not available. Isolation of an unknown sequence using peptides generated from MS can be an alternative method for isolating genes and investigating biological function.

The current study reports for the first time RNA-Seq data on *L. opalescens*. High- throughput sequencing is becoming more popular and less expensive and is likely to become the platform of choice for transcriptome analysis to provide new insights into the complexity of biological systems. These data can complement experimental data and improve our understanding of *Loligo* biology. Further analyses of this dataset can lead to the discovery of novel protein coding genes and offers the potential to identify genes that are important in regulating key aspects of the *L. opalescens* life cycle

Eukaryotic transcriptomes contain a wealth of information at the gene level with regards to transcription, translation and other related processes that occur in the cell. With recent advances in both molecular biology techniques and computational biology, it is certain that more genomes will be available in the near future. The current study provides comprehensive data on de novo gene sequencing and the use of computational biology to find a novel chymotrypsin from a cold

adapted squid, *L. opalescens* and shows how this enzyme is related to other chymotrypsins through phylogeny.

The results of this study have allowed interesting comparisons to be made between a chymotrypsin from squids and a mammalian homologue. Although *Loligo* chymotrypsin has several sequence characteristics which are not observed in mammalian homologues only very tentative conclusions can be drawn as to which alterations are in the key positions which affect both the catalytic and stability characteristics. Experiments using site-directed mutagenesis are required to provide experimental evidence for or against the suggestions made here

8.2 Contribution to Knowledge

- We report for the first time two chymotrypsin-like enzymes from *Loligo opalescens* and *Sepioteuthis lessoniana* with a comparative study on temperature and pH properties as well as their kinetic properties.

- We report for the first time a de novo assembly and transcript annotation of *Loligo opalescens* genome and a full gene and cDNA sequence that encodes *Loligo opalescens* chymotrypsin. The annotated transcripts produced has been submitted to Genbank non-redundant nucleotide database (accession number KU558921) and is available to all researchers for future studies

- The full protein sequence and the three-dimensional model of the enzyme has given a deeper insight to the molecular basis of the biological function of this and other cold-adapted enzymes from marine invertebrates.

- The possibility of using the enzyme as an alternative or additional processing catalyst for industry has been established in this study.

8.3 Recommendations for future work

Large scale production of *Loligo opalescens* chymotrypsin using recombinant DNA techniques

Effect of site directed mutagenesis on the catalytic efficiency, pH, temperature and pressure stability of the enzyme to confirm the differences in amino acid sequences observed between *loligo* chymotrypsin and homologues from other sources.

REFERENCES:

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. 11(10) 106

Addgene, (2015). www.addgene.org date accessed: September 12, 2015

Appel, W. (1986). Chymotrypsin: molecular and catalytic properties. Clinic. Biochem. 19 (6), 317-322.

Asgeirsson, B., and Bjarnason, J. B. (1991). Structural and kinetic properties of chymotrypsin from Atlantic cod (*Gadus morhua*). Comparison with bovine chymotrypsin. Comparative Bioch and Physiol, 99B, 327–335.

Atassi, M. Z. and Manshouri, T. (1993) Design of peptide enzymes (pepzymes): Surface-simulation synthetic peptides that mimic thande chymotrypsin and trypsin active sites exhibit the activity and specificity of the respective enzyme. Proc. Nat. Acad. Sci. 90, 8282-8286.

Bainbridge, B.W. (2000) Microbiological techniques for molecular biology: bacteria and phages. In: *Essential Molecular Biology: A Practical Approach,* Vol. 1 (ed. T.A. Brown), 2nd edn, Oxford University Press, Oxford. 21-54.

Baker, M. (2012). De novo genome assembly: what every biologist should know. Nat. meth .9 (4): 333 - 337

Balti, R., Bougherra, F., Bougatef, A., Hayet B. K., Nedjar-Arroume, N., Dhulster, P., Guillochon, D., and Nasri, M. (2012). Chymotrypsin from the hepatopancreas of cuttlefish (*Sepia officinalis*) with high activity in the hydrolysis of long chain peptide substrates: Purification and biochemical characterization. Food Chem. 130, 475–484

Bateman A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D. and Sonnhammer, E.L.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res., 27, 260–262.

Bender, M.L. and Killheffer, J.V. (1973). Chymotrypsins. Crit. Rev. Biochem. 1 (2), 149-199.

Birktoft, J.J. and Blow, D.M. (1972) Structure of crystalline chymotrypsin. V. The atomic structure of tosyl- -chymotrypsin at 2 A resolution. J.Mol.Biol. 68: 187-240

Blow, D. M. (1971). The structure of chymotrypsin. In P. D. Boyer (Ed.). The enzymes (Vol. 3), New York: Academic Press,185–212)

Bonner, P.L.R. (2007). Protein purification. Taylor and Francis Group, New York, NY.

Brot, F.E. and Bender, M.L. (1969). Use of the specificity constant of alpha-chymotrypsin J. Am. Chem. Soc. 91, 7187-7191.

Brown, C. T, Howe, A., Zhang, Q., Pyrkosz, A. B., Brom, T. H. (2012) A reference-free algorithm for computational normalization of shotgun sequencing data. 480

Buchholz, K., V. Kasche, U. T. Bornscheuer Biocatalysts and Enzyme Technology. (2005). WILEY-VCH Verlag GmbH and Co. KGaA, Weinheim

Camacho, C. et al. (2009). BLAST+: architecture and applications. BMC Bioinformatics 10, 421

Casali, N. and Preston, A. (2003) *E. coli* Plasmid Vectors : Methods and Applications: In *E. coli* Plasmid Vectors : Methods and Applications ,Vol 235, Springer City and pages

Castillo-Yanez, F. J., Pacheco-Aguilar, R., Garcia-Garreno, F. L., Toro, M. A. N., and Lopez, M. F. (2006). Purification and biochemical characterization of chymotrypsin from the viscera of Monterey sardine (*Sardinops sagax caeruleus*). Food Chem 99, 252–259.

Castresana J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol.(4):540-52.

Chaplin M. F. and Bucke C. (1990) Enzyme Technology, Cambridge University Press, Cambridge, UK

Coghlan, A. et al. (2008). nGASP—the nematode genome annotation assessment project. BMC Bioinformatics 9, 549

Cohen, T., Gertler, A. and Birk, Y. (1981). Pancreatic proteolytic enzymes from carp (*Cyprinus carpio*)-I. Purification and physical properties of trypsin, chymotrypsin, elastase and carboxipeptidase B. Comparative Bioch. and Physio, 69B, 639–646.

Copeland, R.A. (2000). Enzymes: A Practical Introduction to Structure, Mechanism, and Data Analysis. Wiley-VCH, Inc. N. Y., USA

Cornish-Bowden, A. (2013). The origins of enzyme kinetics. FEBS Lett., 587 2725–2730

Dale, I. W. (2004) Molecular Genetics of Bacteria, 4th edn. John Wiley, Chichester.

Darville, L. N., Merchant, M. E., Maccha, V., Siddavarapu, V. R., Hasan, A., and Murray, K. K. (2012). Isolation and determination of the primary structure of a lectin protein from the

serum of the American alligator (*Alligator mississippiensis*). Comparative Biochem. and Physiol, 161(2), 161-169.

Delaage, M., Abita, J. P., Lazdunski, M. (1968). Physico-chemical properties of bovine chymotrypsinogen B. A comparative study with trypsinogen and chymotrypsinogen. European J. of Biochem, 5, 285-291

Dereeper, A., Guignonm V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M., Gascuel, O. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 1; 36

D'Esposito, M., Pilia,G. and Schlessinger,D. (1994). BLOCK-based PCR markers to find gene family members in human and comparative genome analysis. Hum. Mol. Genet., 3, 735–740

Deustcher M. P. (1990). Guide to Protein Purification, Academic Press Ltd, San Diego

Diaz P, Phatak SS, Xu J, Astruc-Diaz F, Cavasotto CN, Naguib M. (2009). 6-Methoxy-N-alkyl isatin acylhydrazone derivatives as a novel series of potent selective cannabinoid receptor 2 inverse agonists: design, synthesis, and binding mode prediction. J Med Chem. 52:433–44.

El Enshasy, H., Abuoul-Enein, A., Helmy, S. and El Azaly, Y. 2008. Optimization of the industrial production of alkaline protease by *Bacillus licheniformis* in different production scales. Australian J. of Basic and Appl Sciences, 2(3): 583-593.

Espósito, T.S., Amaral, I.P.G., Oliveira, G.B., Carvalho, J.L.B. and Bezerra, R.S. (2009). Fish processing waste as a source of alkaline proteases for laundry detergent. Food Chem. 112 (1), 125-130.

FAO Yearbook (2013). Fishery statistics-capture production, Rome. Food and Agricultural organization of the United Nations

Fong, W.P., Chan, E.Y., Lau, K. K. (1998). Isolation of two chymotrypsins from grass carp. Biochem. Mol. Bio. Int. 45, 409-418.

Gauvrit, E., Le Goff, R. and Daguzan, J. (1997). Reproductive cycle of the cuttlefish. *Sepia officinalis* (L.) in the northern part of the Bay of Biscay. J. of Molluscan Stud. 63: 19–28.

Gonen, T., Cheng, Y., Sliz, P., Hiroaki, Y., Fujiyoshi, Y., Harrison, S. C., Walz, T. (2005). Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. Nature.438:633–8.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A. (2011). Full-length transcriptome assembly from RNA-seq data without a reference genome, Nature Biotechnology, 29(7):644-652

Green, P., Lipman. D., Hillier, L., Waterston, R., States, D. and Claverie, J. M. (1993). Ancient conserved regions in new gene sequences and the protein databases. Science, 259, 1711-1716.

Groppe, J., and Morse, D. (1993). Molluscan chymotrypsin-like protease; structure, localisation and substrate specificity. Archives of Biochemistry and Biophysics, 305, 159–169.

Guerra, A. (2006). Ecology of Sepia officinalis, J.of Life and Environ.56: 97–107.

Gudmundsdottir, A., Oskarssona, S., Eakinb, A. E., Craikb, C.S. and Bjarnason, J.B. (1994). Atlantic cod cDNA encoding a psychrophilic chymotrypsinogen. Biochimica et Biophysica Acta 1219, 1:211-214

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 59(3):307-21.

Haard, N.F. (1998) Specialty enzymes from marine organisms. Food Technol., 52 (7), 64-67.

Haard, N.F. (1992). A review of proteolytic enzymes from marine organisms and their application in the food industry. J. Aqua. Food Prod. Technol., 1 (1), 17-36.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., Macmanes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R. D., Friedman, N., Regev, A.(2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nature Protocols, 8(8):1494-1512

Haas, B. J. et al. (2008). Automated eukaryotic gene structure annotation using Evidence Modeler and the Program to Assemble Spliced Alignments. Genome Biol. 9, R7

Hanes, C. S. (1932). The effect of starch concentration upon the velocity of hydrolysis by the amylase of germinated barley. Biochem J. 26(5): 1406–1421.

Hedstrom, L. (2002). Serine Protease Mechanism and Specificity, Chem. Rev. 102, 4501–4523.

Henderson, R. and Unwin, P. N. (1975). Three-dimensional model of purple membrane obtained by electron microscopy. Nature.257:28–32.

Henikoff, J. G., Pietrokovski, S., McCallum C. M. and Henikoff, S. (1998) Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences Nucleic Acids Res. 26(7):1628-35

Henikoff, S. and Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. Nucleic Acids Res., 19, 6565-6572.

Henikoff, S. and Henikoff, J. G. (1993). Performance evaluation of amino acid substitution matrices. Proteins 17: 49–61

Hernandez-Cortes, P., Whitaker, J. R., and Garcia-Carreno, F. L. (1997). Purification and characterisation of chymotrypsin from *Penaeus vannamei* (Crustacea: decapoda). Journal of Food Biochem, 21, 497–514.

Heu, M. S., Kim, H. R., and Pyeun, J. H. (1995). Comparison of trypsin and chymotrypsin from the viscera of anchovy (*Engraulis japonica*). Comparative Bioch and Physiol, 112B, 557–568.

Higashi, H. et al. (1997) Imaging of cAMP-dependent protein kinase activity in living neural cells using a novel fluorescent substrate. FEBS Lett. 414, 55–60

Hill, H.F. and Me, W. (1960). Technique of cataract extraction with alpha-chymotrypsin. Arch. Ophthalmol. 64 (4), 601-605.

Holm, L., and Sander, C. (1992) Evaluation of protein models by atomic solvation preference, J Mol Biol 225, 93-105.

Houseman, J. G., Campbell, F. C., and Morrison, P. E. (1987). A preliminary characterization of digestive proteases in the posterior midgut of the stable fly *Stomoxys calcitrans* (L.) (Diptera: Muscidae). Insect Biochem, 17, 213- 218.

Howe, K. L., Chothia, T. and Durbin, R. (2002). GAZE: a generic framework for the integration of gene-prediction data by dynamic programming. Genome Res. 12, 1418–1427

Hummel, B. C. W. (1959). A modified spectrophotometric determination of chymotrypsin, trypsin and thrombin. Canadian J Bioch Physiol, 37, 1393–1399.

Jiang, Q. J., Hall, M., Noriega, F.G., Wells, M. (1997). cDNA cloning and pattern of expression of an adult, female-specific chymotrypsin from *Aedes aegypti* midgut. Insect Biochem. Mol. Biol. 27:283–289

Jiang, Y. K., Sun, L. C., Cai, Q. F., Liu, G. M., Yoshida, A., Osatomi, K., et al. (2010). Biochemical characterisation of chymotrypsins from the hepatopancreas of Japanese sea bass (*Lateolabrax japonicus*). Journal of Agricultural and Food Chemistry, 58, 8069–8076.

Kamini, N. R., Hemachander, C., Geraldine Sandana Mala, C. and Puvanakrishnan (1999), R. Microbial enzyme technology as an alternative to conventional chemicals in leather industry. http://www.iisc.ernet.in/currsci/jul10/articles16.htm Retrieved 28th August 2013

Kapitonov, V. V. and Jurka, J. (2003). A novel class of SINE elements derived from 5S rRNA. Mol. Biol. Evol. 20, 694–702

Kapustin, Y., Souvorov, A., Tatusova, T. and Lipman, D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. Biol. Direct 3, 20

Keller, A., Nesvizhskii, A., Kolker, E. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Analytical Chem, 74(20):5383-92

Kilaraa, A., and Panyam D. (2003). Peptides from Milk Proteins and Their Properties. Critical Reviews in Food Sci and Nutr, 43, (6) 607-633

Klein, B. LeMoullac, G., SIlos, D. and Van Wormhaudjt, A. 1996. Molecular cloning and sequencing oftrypsin cDNAs from *P. vannamei*: use in assessing gene expression during the moult cycle. Int. J. Biochem. BioL 28:551-563.

Klomklao, S. (2008). Digestive proteinases from marine organisms and their applications J. Sci. Technol. 30 (1), 37-46

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics, 5, 59

Korhonen, H. and Pihlanto, A. (2006). Bioactive peptides: Production and functionality. Intl Dairy Journal, 16, (9), 945–960

Kristjansson, M. M., and Nielsen, H. H. (1992). Purification and characterization of two chymotrypsin-like proteases from the pyloric ceca of rainbow trout (*Oncorhynchus mykiss*). Comparative Biochem and Physiol, 101B, 247–253.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature, 277, 680–685.

Laskowski, M. (1955). Chymotrypsinogens and chymotrypsins. Methods Enzymol. II, 8–26.

Laskowski, R. A., MacArthur, M. W., Moss, D. S. and Thornton J M (1993). PROCHECK - a program to check the stereochemical quality of protein structures. J. App. Cryst., 26, 283-291.

Leth-Larsen, R.; Àsgeirsson, B.; Thorolfsson, N.M.; HØjrup, P. (1996). Structure of chymotrypsin variant B from Atlantic cod, *Gadus morhua*. Biochim. Biophys. Acta. 1297 (1), 49-56.

Li, B. and Dewey, C. N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics, 12(323)

Li M, Fang H, Du L, Xia L, Wang B. (2008). Computational studies of the binding site of alpha1Aadrenoceptor antagonists. J Mol Model14:957–66.

Luthy, R., Bowie, J. U., and Eisenberg, D. (1992) Assessment of protein models with three-dimensional profiles, Nature356, 83-85.

Majoros, W. H. (2007). Methods for Computational Gene Prediction 2 (Cambridge Univ. Press) U.K.

Male, R, Lorens, J.B., Smalas, A.O. and Torrissen K.R. (1995). Molecular cloning and characterization of anionic and cationic variants of trypsin from Atlantic salmon. Eur. 1. Biochem. 232:677-685.

Melo, F., and Feytmans, E. (1998) Assessing protein structures with a non-local atomic interaction energy, J Mol Biol 277, 1141-1152.

Michaelis, L. and Menten, M.L. (1913). Kinetik der Invertinwirkung, Biochem. Zeitung, 49 (1913), 333–369

Michielan L, Bacilieri M, Schiesaro A, Bolcato C, Pastorin G, Spalluto G, et al. (2008). Linear and nonlinear 3D-QSAR approaches in tandem with ligand-based homology modeling as a computational strategy to depict the pyrazolo-triazolo-pyrimidine antagonists binding site of the human adenosine A2A receptor. J Chem Inf Model. 48:350–63.

Miller, J. R., Koren, S. and Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics, 95:315-327

Mook, O.R. et al. (2003) In situ localization of gelatinolytic activity in the extracellular matrix of metastases of colon cancer in rat liver using quenched fluorogenic DQ-gelatin. J. Histochem. Cytochem. 51, 821–829

Morozova, O. and Marra, M. A. (2008). Applications of next-generation sequencing technologies in functional genomics. Genomics. 92(5):255-64

Mortazavi, A., Williams, B.A., McCue, K., Schaeffers, L and Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq, Nature Methods 5:621-628

Mowery, J. and Seidman, S. (2005). Protein Purification Manual, The Biotechnology Project.

Naturaldatabase, (2013) http://naturaldatabase.therapeuticresearch.com/home.aspx?cs=&s=ND&AspxAutoDetectCookieSupport=1. Date accessed, December 15, 2013

Nesvizhskii, A., Keller, A. and Kolker, E. (2003). A statistical model for identifying proteins by tandem mass spectrometry. Analytical Chem, 75(17):4646-58)

Nielsen S. S. (2010). Food Analysis, 4th Edition, Springer science business media, New York

Olson,M.V., Hood.L., Cantor,C. and Botstein.D. (1989). A common language for physical mapping of human genome. Science, 245, 1434-1435

O'Malley, M.M. and Straatsma, B.R. (1961). Experimentally induced adverse effect of alpha-chymotrypsin. Arch. ophthalmol., 66: 539-544.

Oshlack, A. and Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. Biol Direct 4: 14.

Overnell, J. (1973). Digestive enzymes of the pyloric caeca and of their associated mesentery in the cod (*Gadus morhua*). Comparative Biochem and Physiol, 46B, 519-531.

Parker, L. and Wang, J. H. (1968). On the mechanism of action at the acylation step of the α-chymotrypsin-catalyzed hydrolysis of anilides. J. Biol. Chem. 243, 3729-3734.

Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc. Natl Acad. Sci. USA 85, 2444–2448

Petsko, G. A. and Ringe, D. (2004). Protein structure and function. Sinauer associates Inc publishers. Sunderland, USA. 141

Prescott, L.M., Harley, I. P. and Klein, D.A. (2004). Microbiology, 5th cdn. McGraw Hill Education, Maidenhead. [A good introduction to microbiology, including plasmids and phages.] From: Methods in Molecular Biology, Vol. 235: E. coli Plasmid Vectors Edited by: N. Casali and A. Preston © Humana Press Inc., Totowa, NJ

Polgar, L. (2005). The catalytic triad of serine peptidases. Cell, Molecular and Life Science. 62: 2161–2172

Psochiou E., Mamuris Z., Panagiotaki P., Kouretas D., Moutou K.A. (2007). The response of digestive proteases to abrupt salinity decrease in the euryhaline sparid Sparus aurata L. Comparative Biochem and Physiol, Part B, 147, 156-163.

Raae, A. J., Flengsrud, R., and Sletten, K. (1995). Chymotrypsin isoenzymes in Atlantic cod; differences in Comparative Biochem and Physiol, 112B, 393–398.

Raae, A. J. and Walther, B. T. (1989). Purification and characterization of chymotrypsin, trypsin and elastase like proteinases from cod (*Gadus morhua* L.). Comparative Biochem and Physiol, 93B, 317–324.

Racicot, W. F., and Hultin, H. O. (1987). A comparison of dogfish and bovine chymotrypsins. Archives of Biochem and Biophy, 256, 131–143.

Raksakulthai, N., Lee1, Y. Z., Haard N. F. (1986). Effect of Enzyme Supplements on the Production of Fish Sauce from Male Capelin (Mallotus villosus). Canadian Institute of Food Science and Technology Journal, 19 (1), 28–33

Reese, M. G. and Guigo, (2006). R. EGASP: Introduction. Genome Biol. 7 (Suppl. 1), 1–3

Reckel, S., Gottstein, D., Stehle, J., Lohr, F., Verhoefen, M.K., Takeda, M., Silvers, R., Kainosho, M., Glaubitz, C., Wachtveitl, J., Bernhard, F., Schwalbe, H., Guntert, P., Dotsch, V. (2011) Solution NMR structure of proteorhodopsin. Angew.Chem.Int.Ed.Engl. 50: 11942-11946

Reis-Filho, J. S. (2009). Next-generation sequencing. Breast Cancer Res. 11(3):12

Ren, X., Liu, T., Dong, J., Sun, L., Yang, J., Zhu, Y. and Jin, Q. (2012). Evaluating de Bruijn Graph Assemblers on 454 Transcriptomic Data. PLoS One. 2012; 7(12)

Rhee, D.J.; Deramo, V.A.; Connolly, B.P.; Blecher, M.H.(1999). Intraocular pressure trends after supranormal pressurization to aid closure of sutureless cataract wounds. J. Catar. Refr. Surg. 25 (4), 546-549.

Rhodes, D. G., Bossio, R. E., Laue, T. M. (2009). Determination of size, molecular weight, and presence of subunits. Methods Enzymol. 463:691-723.

Ritchie, R. J. and Prvan, T. (1996). Current Statistical Methods for Estimating the Km and Vmax of Michaelis-Menten Kinetics. Biochemical Education, 24 (4) 196-206

Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics* 26 (1):139-140

Rose T. M., Schultz, E. R., Henikoff, J. G., Pietrokovski, S., McCallum, C. M. and Henikoff, S. (1998). Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences Nucleic Acids Res., 26, 1628–1635

Sali, A., Potterton, L., Yuan, F., van Vlijmen, H., and Karplus, M. (1995) Evaluation of comparative protein modeling by MODELLER, Proteins 23, 318-326.

Sameni, M. et al. (2000) Imaging proteolysis by living human breast cancer cells. Neoplasia 2, 496–504

Sayers, E. W. et al. (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 37, D5–D15

Schellenberger, V., Braune, K., Hoffman, H. J. and Jakubke, H.-D. (1991). The specificity of chymotrypsin. A statistical analysis of hydrolysis data. Eur. J. Biochem. 199, 623–636.

Screen, S.E. and Leger R.J. (2000). Cloning, expression, and substrate specificity of a fungal chymotrypsin. Evidence for lateral gene transfer from an actinomycete bacterium. J. Biol. and Chem, 275(9):6689-94

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol.7:539

Simpson. B. K. (2012). Food biochemisty and food processing. Wiley Blackwell, Iowa, USA.

Simpson, B. K. 2000. Digestive proteinases from marine animaIs. In: Seafood Enzymes. Chapter 8, pp.191-213 (N.F. Haard and B.K. Simpson, eds.) Marcel Dekker, N.Y., U.S.A

Simpson, B. K. and Haard, N. F. (1994). Proteases from aquatic organisms and their uses in the seafood industry. In Fish Processing: Biotechnological Applications. A. Martin (ed.). Elsevier, New York, U. S. A. pp. 132-154.

Simpson, J. T. et al. (2009). ABySS: a parallel assembler for short read sequence data. Genome Res. 19, 1117–1123

Smalls, A.O., Heimstad, E.S., Hordvik, A., Willassen, N.P. and Male, R. (1994) Proteins Struct. Funct. Genet. 20, 149-166.

Smillie, L.B., Furka, A., Nagabhushan, N., Stevenson, K.J., Parkes, C.O. (1968). Structure of chymotrypsinogen B compared with chymotrypsinogen A and trypsinogen. Nature, 218: 343-346

Söding J., Biegert, A. and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction Nucleic Acids Res. 33:W244–W248

Somero, G.N. and Hochachka, P.W. (1984) Biochemical Adaptation, Princeton University Press, Princeton, NJ.

Souvorov, A. et al. (2010). Gnomon — the NCBI eukaryotic gene prediction tool. National Center for Biotechnology Information [online], http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml

Speers, A. E. and Cravatt, B. F. (2004). Profiling Enzyme Activities In Vivo Using Click Chemistry Methods. Chemistry and Biology, Vol. 11, 535–546

Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. Nature 458, 719-724

Swamy, A. H. M. V.; Patil, P. A. (2008) Effect of some clinically used proteolytic enzymes on inflammation in rats. Ind. J. Pharm. Sci. 70 (1):114-117.

Tatsumi, H., Ogawa, Y., Murakami, S., Ishida, Y. et,al. (1989). A full length cDNA clone for the alkaline protease from Aspergillus oryzae: Structural analysis and expression in Saccharomyces cerevisiae. Molecular and General Genetics, 219, (1) 33-38

Thayer, M.M., Flaherty, K. M., McKay, D. B. Three-dimensional structure of the elastase of Pseudomonas aeruginosa at 1.5-A resolution. (1991) J.Biol.Chem. 266: 2864-2871

Thummel, K. E. and Wilkinson, G. R. (1998). In vitro and In vivo drug interactions involving human CYP3A. Annu. Rev. Pharmacol. Toxicol. 38:389–430

Treangen, T. J. and Salzberg, S. L. (2012).Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Rev. Genet. 13, 36–46

Tucker T, Marra M, Friedman J. M. (2009). Massively parallel sequencing: the next big thing in genetic medicine. Am J Hum Genet, 85:142-154

UniProt Consortium. (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. 39, D214–D219

Usadel Lab (2014), Trimmomatic: A flexible read trimming tool for Illumina NGS data

Venekei, I., Szilágyi, L., Gráf, L. and Rutter, W. J. (1996). Attempts to convert chymotrypsin to trypsin. FEBS Letters, 383, 133

Walsh G. (2002). Protein Biochemistry and Biotechnology. John Wiley and Sons Ltd, England.

Walter, J. .M. (1996). The protein protocols handbook. Humana press Inc. Totowa, NJ.pg 11-14

Walter, J., Steigemann, W., Singh, T. P., Bartunik, H., Bode, W., Huber, R. (1982) Acta On the Disordered Activation Domain in Trypsinogen. Chemical Labelling and Low-Temperature Crystallography Crystallogr.,Sect.B 38: 1462-1472

Wang, J. S. (2001). In vivo and in vitro studies on drug metabolism and interactions involving mibefradil, isradipine, lidocaine, selegiline and metronidazole. Academic Dissertation presented, with the permission of the Medical Faculty of the University of Helsinki, for public examination Helsinki, Haartamaninkatu

Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics

Yandell, M. and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics 13: 329 – 342

Yang, A. and Honig, B. (2000). An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. Journal of Mol. Biol, 301 (3), 665–678

Yang, F., Su, W. J., Lu, B. J., Wu, T., Sun, L. C., Hara, K., et al. (2009). Purification and characterisation of chymotrypsins from the hepatopancreas of crucian carp (*Carassius auratus*). Food Chem, 116, 860–866.

Yang, Z., Xia, X., Wang, X. and He, G. (2010) cDNA Cloning, Heterogeneous Expression and Biochemical Characterization of a Novel Trypsin-Like Protease from Nilaparvata lugens. Z. Naturforsch. 65 c, 109 – 118

Zhang, C., Zhou, D., Zheng, S., Liu, L., Tao, S.,Yang, L., Hu, S., Feng, S. (2010). A chymotrypsin-like serine protease cDNA involved in food protein digestion in the common cutworm, *Spodoptera litura*: Cloning, characterization, developmental and induced expression patterns, and localization. Journal of Insect Physiol. 56 (7) 788-799

Zhang, Y.C. (2007). Progress of clinical application of chymotrypsin. Asia-Pacific. Trad. Med. 3 (7).

Zhou, L.M., Wu, S. G., Liu, D. C., Xu, B., Zhang, X. F. and Zhao, B. S. (2012). Characterization and expression analysis of a trypsin-like serine protease from planarian Dugesia japonica. Mol Biol Rep 39:7041-7047