10-2014

Venomic Characterization of the Terebridae and Novel Terebrid Neuropeptides

Mary Elizabeth Wright
Graduate Center, City University of New York

How does access to this work benefit you? Let us know!
Follow this and additional works at: https://academicworks.cuny.edu/gc_etds
Part of the Biochemistry Commons

Recommended Citation
Wright, Mary Elizabeth, "Venomic Characterization of the Terebridae and Novel Terebrid Neuropeptides" (2014). CUNY Academic Works.
https://academicworks.cuny.edu/gc_etds/510

This Dissertation is brought to you by CUNY Academic Works. It has been accepted for inclusion in All Dissertations, Theses, and Capstone Projects by an authorized administrator of CUNY Academic Works. For more information, please contact deposit@gc.cuny.edu.
Venomic Characterization of the Terebridae and Novel Terebrid Neuropeptides

Mary Elizabeth Wright
This manuscript has been read and accepted by the Graduate Faculty in Biochemistry
In satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

Date ______________ Dr. Mande Holford

_____________________________________
Chair of Examining Committee

Date ______________ Dr. Richard Magliozzo

_____________________________________
Executive Officer

Supervisory Committee: Dr. David Foster
Dr. Diego Loayza
Dr. Ines Ibanez-Tallon
Dr. Mark Siddall

THE CITY UNIVERSITY OF NEW YORK
Abstract

VENOMIC CHARACTERIZATION OF THE TEREBRIDAE AND NOVEL TEREBRID NEUROPEPTIDES

By

Mary Elizabeth Wright

Adviser: Dr. Mande Holford

Unravelling the complex mixture of neuropeptides produced by the terebrid venom duct holds the promise of discovering tomorrow’s therapeutics. Cone snails have already demonstrated the potential found in the venom of these unusual marine organisms, through the commercial approval of drugs for pain and other indications. Terebrids, as the sister family to the cone snails, have been much less investigated, but have a species richness that makes them very attractive in the search for novel neuropeptides. The venomics research described in this work encompasses the first comprehensive characterization of the terebrid venom duct transcriptomes of two species, *Cinguloterebra anilis* and *Terebra subulata*. De novo assembly and analysis were performed using next generation sequencing technology and state of the art bioinformatics tools to identify the cocktail of peptides, teretoxins, produced by the venom duct. These disulfide rich peptides often show a remarkable specificity for subtypes of ion channels and ligand-gated receptors, giving them therapeutic potential, but they are only available in vanishingly small amounts from the natural source. As a result, identification of teretoxins using next generation sequencing is a prelude to synthesizing them in sufficient quantities to test for bioactivity. Here recombinant expression and solid phase peptide synthesis have both been utilized for the synthesis of four different teretoxins, with a special focus on recombinant synthesis to design a reproducible strategy for synthesizing disulfide rich peptides greater than 40 amino acids in length. Preliminary characterization of bioactivity was performed by injecting synthesized toxin into the polychaete *N. virens*. A teretoxin identified from *Terebra guttata*, Tg77, has produced promising results in this assay, with repeated trials showing the effects of uncoordinated movement and rigid paralysis. Further testing on Tg77 and other teretoxins will be carried forward, with an evident need for high throughput assays to provide an efficient means for the testing of novel compounds with a variety of potential molecular targets.
Acknowledgements

Five short years ago I began the PhD program in Biochemistry, and it with a real sense of wonder, maybe even disbelief, that here I am on the verge of graduating. Naturally none of this has occurred in a vacuum. Once I found my lab “home” I have been blessed to find a great mix of people. My adviser, Dr. Holford, has brought enormous enthusiasm to the serious business of conducting research. She is willing to tackle almost any subject matter, and has designed her research in such way that her students have the freedom to explore many exciting topics where venomics meets marine biology. Her support, advice and encouragement throughout have been exceptional. The people working in the lab have become not only colleagues but friends. Without the wisdom and knowledge of Prachi Anand on so many different aspects of scientific research I never would have learned as much as I have, or as painlessly. Our merry band of undergraduates have also been terrific in their contributions to the lab, and I am particularly thankful to two talented individuals who have worked on my projects, Girish Ramrattan and John Moon.

My family has been there for me throughout this whole adventure, and my thoughts go out to my Mom, who may not be around to see me graduate, but has who has been the single biggest influence in my life as a role model. She was an independent and resourceful woman who built her own horse jumps from lumber, a homemaker who sewed her own drapes and slipcovers, a public school teacher for all of her career, and a loving mom who singlehandedly raised four children. She was extremely proud of her children’s accomplishments, and I wish she could be here with me today.

I thank my friends who have provided solace when experiments were not working, enthusiasm when things were going well, and put up with my long-winded explanations of my research projects, either feigning interest or even being genuinely intrigued by what I had to say. My committee members have been helpful throughout. I would never be in the program in the first place if it were not for the support of Diego Loayza, who believed in my potential. David Foster and I have shared a love of opera as well science, Inez Ibanez-Tallon never fails to be an inspiring researcher, and Mark Siddall at the American Museum of Natural History has been indispensable in providing research advice and support. In fact, the museum was my home away from home for a good part of the PhD process, which has been a real privilege.
There are many other colleagues and friends too numerous to mention, but to name just a few: Alex Grigoryan and Jasy Ho, fellow graduate students; Christine Tada, en route to becoming a physician assistant extraordinaire; and Dr. Emmanuel Chang at York College, an endless font of scientific acumen and good ideas.

Learning to be a scientist does not happen overnight. It is a process of learning and hard thinking that has no endpoint and is a humbling experience in light of so many great scientific minds and the huge body of both known and unknown matters that power the advance of science. For those who have shared their knowledge and advice (also reagents!), helping to shed light on my corner of the research world, I offer my deepest gratitude.
Table of Contents

Chapter 1 Introduction ................................................................................................................... 1
  1.1 Peptide drug development from venomous organisms .............................................................. 2
  1.2 Conoidean species and venom diversity ................................................................................. 3
  1.3 A brief history of conopeptide research ................................................................................. 5
  1.4 Structural and genetic features of conoidean peptides ........................................................... 10
  1.5 Structural and genetic features of teretoxins ......................................................................... 12
  1.6 Traditional vs. emerging techniques for venom characterization .......................................... 12
    1.6.2 Next generation sequencing applied to venom characterization ......................................... 14
    1.6.3 Technical aspects of next generation sequencing ................................................................. 15
  1.6.4 Bioinformatics tools for de novo assembly and analysis ..................................................... 17
  1.7 Synthesis strategies for neuropeptides ..................................................................................... 19
    1.7.1 Solid phase peptide synthesis (SPPS) of venom peptides .................................................. 20
  1.7.2 Recombinant expression of neuropeptides ......................................................................... 22
  1.8 Characterization of venom peptide bioactivity ......................................................................... 26
  1.9 Objectives of Thesis Project .................................................................................................... 28

Chapter 2 De novo characterization of Terebrid venom duct transcriptomes ............................ 30
  1.9 Background: a revolution of model systems ......................................................................... 31
  2.2 Results and Discussion ............................................................................................................ 34
    2.2.1 Hi-Seq Illumina sequencing and overview of C. anilis transcriptome assemblies ................. 34
    2.2.2 Identification and validation of C. anilis teretoxin transcripts ............................................. 35
    2.2.3 Preliminary analysis of Terebra subulata venom duct transcriptome ..................................... 45
    2.2.4 Comparative analyses of C. anilis and T. subulata venom duct transcriptomes ..................... 46
    2.2.5 Identification of first Teretoxin gene superfamily ................................................................. 47
    2.2.6 Identification of Teretoxin Transcripts with “Pepticomb” .................................................... 49
    2.2.7 Annotation using Gene Ontology and KEGG analyses ....................................................... 52
    2.2.8 Venom Duct Proteins with Peptide Folding and Processing Functions ................................ 54
  2.3 Conclusion ............................................................................................................................... 57
  2.4 Materials and Methods .......................................................................................................... 59
    Sample Collection ...................................................................................................................... 59
    RNA extraction and cDNA preparation ...................................................................................... 60
    Library preparation and Sequencing .......................................................................................... 61
    Read processing and de novo assembly ...................................................................................... 61
    BLAST annotation .................................................................................................................... 61

Chapter 3 Synthesis of novel teretoxins for functional assays .................................................. 63
  3.1 Background: characterizing teretoxin bioactivity .................................................................... 64
    3.1.1 Recombinant expression of teretoxins Tg77, Can6.5, and Can10.1 ....................................... 65
  3.2 Results and Discussion ............................................................................................................ 68
    3.2.1 Design and construction of expression system ..................................................................... 68
    3.2.2 Expression and purification of recombinant fusion proteins ............................................. 70
    3.2.3 Chemical synthesis of Can.14.1 teretoxin ............................................................................ 77
  3.3 Functional characterization of teretoxins in polychaete worm assay ...................................... 80
3.5 Materials and methods .................................................................................................................. 86
  3.5.1 Recombinant plasmid construction .......................................................................................... 86
  3.5.2 Induction and Expression ........................................................................................................ 87
  3.5.3 Protein extraction and His Tag Affinity Purification .................................................................. 88
  3.5.4 Enterokinase cleavage ............................................................................................................. 88
  3.5.5 HPLC purification and Mass Spectrometry ............................................................................ 89
  3.5.6 Solid phase peptide synthesis ................................................................................................ 89
  3.5.7 Polychaete Worm Assay ......................................................................................................... 90

Conclusion and Future Directions ..................................................................................................... 91

Supplementary Information ................................................................................................................. 95

References ........................................................................................................................................... 96
List of Figures

Chapter 1

1. Representatives of the Conoidean superfamily 2
2. Conoidean venom apparatus 4
3. Structures of representative O superfamily conotoxins 6
4. Precursor organization of conoidean peptide toxins 8
5. Gene superfamilies of disulfide rich conopeptides 9
6. Integrated pipeline for peptide toxin discovery 13
7. Illumina sequencing preparation and workflow 16
8. De Bruijn graph methodology 18
9. General scheme for Fmoc solid phase peptide synthesis 22
10. Scheme for recombinant expression of disulfide rich venom peptides 23
11. Ligation independent cloning 25

Chapter 2

1. Terebrid shell and venom apparatus schematic 32
2. Venom apparatus characterization in the Terebridae 33
3. Bioinformatic pipeline for characterizing venom duct transcriptome 34
4. Venn diagram showing assembly overlap of putative toxins transcripts 37
5. Representative BLAST alignments of C. anilis teretoxin transcripts 39
6. Precursor structure of C. anilis conophysin/conopressin alignment 42
7. Alignment of a C. anilis terepressor to Conus conopressins 42
8. Circos representation of C. anilis putative toxin BLAST results 44
9. Alignment of T. subulata toxins to previously identified teretoxins 46
10. Distribution of T. Subulata and C. anilis toxins among Cys frameworks 47
11. Terebrid gene superfamily based on conserved signal sequence 48
12. Comparison of C. anilis venom duct transcriptome to toxin-related genes 52
13. BLAST2GO annotation of C. anilis venom duct transcriptome assembly 53
14. Comparative iPPath annotation of C. anilis and C. geographus metabolic pathways 54
15. Representative terebrid species collected in Inhaca 61
Chapter 3

1. Full precursor structure of terotoxins chosen for recombinant expression
2. Fusion protein organization for soluble expression
3. Sample colony PCR showing correct gene insertion
4. SDS-PAGE analysis of expression and purification of Tg77
5. Elution profiles FPLC His affinity purification of peptides Can.6.5 and Can.10.1
6. SDS-PAGE analysis of Tg77 expression and purification
7. Tris-Tricine SDS-PAGE optimization analysis of Tg77 EK cleavage
8. Semi-preparative RP-HPLC and LC-MS of Tg77 teretoxin
9. Analytical RP-HPLC and LC-MS of Can.6.5 teretoxin
10. Can.14.1 alignment to Conus austini AsXIVA peptide
11. SPPS Attachment of First Amino Acid (Serine) to Wang resin
12. Teretoxin Can.14.1 cleavage from Wang-Cl resin
13. RP-uHPLC elution time profiles for oxidative folding of linear Can.10.1
15. Generalized scheme of polychaete anatomy
16. Polychaete N. virens swimming motion
17. Preliminary results of polychaete moving average speed using WormLab
18. Video still image from polychaete worm assay
List of Tables

Chapter 1

1. Comparison of structural and genetic features of conotoxins and teretoxins 11

Chapter 2

1. Summary of *C. anilis* putative teretoxins 36
2. Pepticomb candidate *C. anilis* toxins 50
3. Annotation of posttranslational enzymes found in terebrid venom duct transcriptome 56

Chapter 3

1. Teretoxins synthesized during thesis work 65
2. Comparison of recombinant expression strategies for neurotoxins 68
3. Testing of Tg77 in *N. virens* for phenotypic response 84
Chapter 1

Introduction
1.1. **Peptide drug development from venomous organisms**

As the pharmaceutical pipeline diminishes alternatives to small molecule drugs have garnered increased attention, and peptide therapeutics have seen a remarkable expansion, resulting in six new approvals for commercial use in 2012.¹ Natural products development of bioactive neuropeptides from venomous marine organisms is a major resource for therapeutic applications. The bioactive venom peptides of marine snails are fast acting, efficient, and highly specific – all essential virtues for a successful drug candidate. Venomous marine snails of the superfamily Conoidea are classified into three major groups that include cone snails (Conidae), terebrids (Terebridae) and turrids (Turridae, which have recently undergone a reclassification that describes 13 monophyletic families(Fig. 1).²³ While conoidean snails express peptide neurotoxins in their venom for both predation and defense, it has been discovered that these peptides are highly effective neuronal tools for manipulating cellular physiologies.⁴–⁶

![Figure 1. Representatives of the Conoidean superfamily.](image)

These predatory marine snails are classified into the families Conidae (cone snails), Terebridae, and “Turridae” which has recently undergone reclassification into 13 monophyletic families. A. *Conus marmoreus* (Conidae) B. *Strioterebra plumbea* (Terebridae) C. *Raphitoma sp.* (Raphitomidae) (Images courtesy of Museum national d’histoire naturelle).

The diversity of snail neuropeptide molecular targets include sodium (Na⁺), potassium (K⁺), and calcium (Ca²⁺) channels, noradrenaline transporters, and nicotinic acetylcholine (nACh) receptors, ensuring a significant impact on a broad range of biomedical problems related to neuronal function.⁷ An example of the potential of snail neuropeptides is Ziconotide (MVIIA), discovered in 1980’s from the venomous marine snail Conus magus. MVIIA, a 25 amino acid peptide, is commercially marketed as
Prialt® and used to alleviate chronic pain in HIV and cancer patients by inhibiting N-type calcium channels. Prior to MVIIA most pharmaceutical companies developed analgesic treatments by focusing on morphine like compounds that targeted opioid receptors. Morphine and other opioid receptor acting compounds have major side effects such as addiction and decreased potency over time. As MVIIA does not target opioid receptors, it does not have the same side effects as opiates. MVIIA was innovative in that it was the first drug developed from a marine snail, and it certified a new strategy for identifying analgesic compounds, namely, inhibitors of N-type calcium channels.

Similar to MVIIA several other peptides from venomous marine organisms are currently being investigated for therapeutic use. Specifically, tumor cell proliferation may be mediated by cone snail peptide k-PVIIA, which binds to hERG, a potassium (K⁺) ion channel protein that increases in concentration on the cell surface of cancer cells. Blocking of HERG inhibits tumor cell proliferation. Bioactive peptides from venomous snails are a proven source for identifying new therapeutics, however less than 2% of identified peptide toxins from venomous snails have been characterized to the functional level and great numbers of venomous species remain undescribed.

This dissertation enhances the knowledge of venomous peptides produced from predatory snails of the Terebridae family, an understudied group of the Conoidean superfamily, by investigating the diversity of terebrid venom peptides, teretoxins, and attempts to characterize the function of selected peptides.

1.2. Conoidean species and venom diversity

The globally distributed Conoidea are one of the most diverse group of venomous organisms in the marine realm. Cone snails are estimated to have somewhere between 500 and 800 species, while terebrids have approximately 300-400 species, and turrids are a paraphyletic group with as many as 10,000 species. Conoidean snails vary dramatically in size. In general, cone snails are by far the largest, and hence the easiest to collect, while the “micro” turrids are often quite small (< 10 mm) and reside in deep water. The Conoidea have been perfecting the art of the hunt for the past 50 million years. Conoideans use a sophisticated venom apparatus to inject a cocktail of neuropeptides that rapidly immobilize their prey of sea worms, mollusks or fish. The conoidean venom apparatus includes a
muscular venom bulb, a slender, looping venom duct and a cleverly designed radular tooth that is employed as both a hypodermic needle to inject venom into the prey and a tether to grasp the animal (Figure 2).\textsuperscript{5,6}

![Figure 2. Conoidean venom apparatus. A. Components of conoidean venom apparatus include a muscular venom bulb, the venom gland for toxin production, a radular sac containing harpoon like radular teeth, and a proboscis structure from which venom laden radular tooth is injected into prey (Norton and Olivera, 2006) B. Electron micrograph of sample conoidean radular tooth](image)

Cone snails and their venoms have been of great interest to the research community for at least three decades, given the potential of their neuropeptides as leads for drug development in mammalian systems, and yet they only represent a fraction of the species richness found in the larger Conoidean superfamily. It would be fair to say that cone snails are currently the rock stars of the venomous marine snail world, but there are plenty of other contenders waiting for their turn in the spotlight. As with cone snails, terebrids and turrids have enjoyed substantial evolutionary and pharmacological success as marine predators, and thrive in marine habitats around the world.\textsuperscript{12,14} Given their species richness, Conoidean snails as a group may represent a reservoir of 500,000-1,000,000 toxins for the discovery of neuropeptides with novel functions.\textsuperscript{12}
Predatory snails, such as conoideans, may express up to 200 neuropeptides in their venom duct. The neuropeptides expressed are primarily short disulfide-rich peptides that incapacitate the prey by functioning primarily as agonists on array of nervous system molecular targets in a concerted manner.\textsuperscript{15} Cone snails have often been compared to combinatorial chemists for their ability to produce libraries of tightly scaffolded neuropeptides arising from disulfide bonds, along with highly variable amino acid composition. The combination of distinct structural folds and the spatial distribution of variable residues enables finely calibrated subtype selectivity for a wide array of ion channels and ligand gated receptors.

The potency and diversity of \textit{Conus} neuropeptides, referred to as conotoxins or conopeptides, has led to research focused on three major fronts: 1. Understanding the evolutionary forces that have shaped such extensive diversity; 2. Utilizing neuropeptides with high specificity for molecular targets to probe nervous system structure and function; and 3. Investigating a multitude of peptidic compounds as leads for drug discovery.\textsuperscript{16–21} The vast array of conopeptides made available by nature, driven both by the number of species and the complex assortment of neuropeptides found in the venom of individual species, provide a natural product platform for the discovery of therapeutics to treat nervous system disorders, such as pain, epilepsy and Parkinson's disease, ensuring that they will continue to be extensively investigated.

Where do terebrids fit in among a host of venom producing organisms that are the subject of ongoing research and investigation? First and foremost, terebrids are of interest as they are a sister group to cone snails, however their species and venom diversity are relatively understudied. Similar to cone snails, terebrids express a diverse array of hypervariable disulfide-rich peptide toxins, teretoxins, that come in an assortment of molecular scaffolds.\textsuperscript{12,22,23} Preliminary research indicates that teretoxins, while similar to conotoxins, have significant differences, suggesting that investigation of terebrid venom has the potential to expand the horizons of marine natural product drug development.

1.3. \textit{A brief history of conopeptide research}

While the potentially rich pharmacological resources of terebrid peptide toxins have just begun to be explored, it is instructive to examine the history and trajectory of cone snail research as a paradigm for teretoxin research. The groundbreaking nature of conotoxin research provides many lessons for the
discovery of natural products that modulate nervous system function, and much of what has been accomplished in this field provides a road map for pursuing the investigation of teretoxins, while keeping in mind that the pace of innovation for both genetic and proteomic research techniques is substantially altering the ways in which marine snail venom can be assessed, characterized and analyzed.

Research groups led by Baldomero Olivera, Paul Alewood, and Richard Lewis have largely shaped the existing knowledge of conopeptide research. Olivera and Terlau were the first to hypothesize that cone snails use an assortment of toxins acting in concert on different nervous system targets in a manner referred to as ‘cabals.’ Specifically, the ‘lightning strike cabal’ enables rapid immobilization of prey through the injection of toxins that simultaneously slow Na\(^+\) channel inactivation and inhibit K\(^+\) channels, leading to massive depolarization across nerve cell membranes and consequent hyperexcitation that quickly results in a paralytic state. In contrast, the ‘nirvana cabal’, better suited to the hunting strategy and anatomy of *Conus geographus*, is composed of peptide toxins that act in concert to deaden the neuronal circuitry and work more slowly to produce a soporific response in the prey.

One notable case of conopeptide discovery from the Olivera laboratory involves work from several talented undergraduates exploring potential conopeptide bioactivity via intracranial injection into mice. Their preliminary injections led to further investigation of the molecular functions of “sleeper” and “shaker” behavioral phenotypes elicited from specific peptide fractions. In the case of the “shaker” peptides, culled from the venom of *Conus magus* and *Conus geographus*, collaborative research demonstrated that N-type calcium channel currents were being potently and selectively inhibited at the

Figure 3. Structures of representative O superfamily conotoxins targeting calcium and potassium channels. Examples of well characterized conotoxins with VI/VII framework. \(\omega\)-MVIIA and \(\omega\)-GVIA were discovered in the lab of Baldomero Olivera (Image courtesy of B. Olivera).
neuromuscular junction\textsuperscript{29–31}. Ultimately two of the “shaker” peptides became known as GVIA and MVIIA (Fig. 3), the former widely used to study the inhibition of synaptic transmission and disruption of N-type Ca\textsuperscript{2+} currents, while the latter was developed into the previously mentioned Ziconotide (Prialt®) for the treatment of severe and chronic pain.\textsuperscript{8} The discovery of conotoxins GVIA and MVIIA, as well several other conotoxins of therapeutic interest, laid the groundwork for a major expansion of interest in cone snail research. Numerous other labs, including ours, have taken up the challenge of collecting venom and/or its genetic material, identifying the components found therein, sequencing and synthesizing peptides of interest and testing these for biological function.

Conopeptides that modulate Ca\textsuperscript{2+} channel function have been of great interest both for their clinical importance and ability to characterize ion channel function. There have been more than 2000 research papers that have used GVIA as a tool for probing molecular features of the nervous system, and the search for antinociceptive agents has led to the characterization of conopeptides with a variety of other targets including Na\textsuperscript{+} channels, K\textsuperscript{+} channels, nAChRs, neurotensin receptors, N-methyl-D-aspartate (NMDA) receptors, norepinephrine transporters and G-protein coupled receptors.\textsuperscript{32} Their broad repertoire of nervous system targets has resulted in clinical trials of several conopeptides as treatments for neurophathic pain, anti-convulsants, anti-cancer, and cardioprotective agents.\textsuperscript{6,33,34}

In spite of the impressive strides that have been made in identifying conotoxins with therapeutic potential, it is estimated that ~ 99% of conotoxins have yet to be characterized.\textsuperscript{35} If the majority of conoidean species can express up to 200 conopeptides in their venom and there are over 10,000 conoidean species, the ~5400 protein and ~2400 nucleotide sequences derived from proteomic and genetic analyses currently listed on Conoserver repository website, is a painfully small fraction of conopeptides characterized.\textsuperscript{36} Advances in proteomics (mass spectrometry), genetic analysis (next generation sequencing), and high throughput bioactivity screening of identified peptide toxins promise to accelerate the discovery process, but it is still remarkable to contemplate the vast landscape yet to be explored.
1.4. **Structural and genetic features of conoidean peptides**

Prior to investigating the potential therapeutic effects of conoidean peptides, it is essential to first elucidate the structural and genetic features of these peptides to gain insight into how they function in the context of complex nervous systems, and to provide an overall organizational scheme for classification into gene superfamilies and pharmacological groupings. Conoidean peptide toxins are single gene products defined by highly conserved pre region at the N-terminal, which serves as a signal sequence, and a more variable intervening pro region, followed by a sequence with a conserved cysteine framework and hyper-variable amino acid content that is the mature peptide toxin (Fig. 4). Mature peptide toxins, which are cleaved by proteolysis from the precursor structure to become bioactive moieties, typically range between 12 and 60 amino acids in length, the shorter versions usually being conopeptides, and the larger versions found in terebrid and turrid peptide toxins.

![Figure 4. Precursor organization of conoidean peptide toxins.](image)

The conserved precursor structure, which is the signal sequence in association with mature peptide cysteine frameworks, forms the basis of conoidean gene superfamilies. These superfamilies, designated by a capital letter, are associated with pharmacological families, designated by a greek letter and indicate the potential molecular channel or receptor target (Fig. 5).

Over time, the number of gene superfamilies has grown significantly, and as a result there has been increasing overlap of both the cysteine frameworks and pharmacological families distributed among
them. As an illustration of these increasingly complex relationships, the M gene superfamily is now associated with five different cysteine frameworks (II, III, VI/VII, and XVI) and four pharmacological families, two of which are known, Na+ and K+ channels. Many of the recently identified gene superfamilies have very few representatives, and no known molecular target. In contrast, conoidean peptides of the well characterized A and O superfamily conotoxins have many representatives and have been widely researched for their inhibitory activity on nAChRs and Ca\(^{2+}\) among other targets.

Figure 5. Gene superfamilies of disulfide-rich conopeptides. A lettered scheme shows classification of conopeptides gene superfamilies based on conserved signal sequence and cysteine patterns. Gene superfamilies are associated with the targeting of specific types of ion channels and receptors (Akondi, et al, 2014).

Conoidean disulfide connectivity varies depending on the number of cysteines (Cys) and their distribution in the mature peptide toxin\(^{37}\). One common and well known bonding pattern is the cysteine knot referred to as the ICK (Inhibitor Cysteine Knot) motif. ICK motif which arises from a 6 Cys framework in which three disulfides form a loop that is threaded by the third disulfide bond (Fig. 3). The ICK motif is common among conoidean and arachnid neuropeptides. One of the most familiar ICK motif peptides are
Huwentoxins (HWTX) from Chinese bird spider *Haplorelma huwena*.\textsuperscript{38,39} Recently, ICK motif peptides were found in the venom of the Australian scorpion, *Liocheles waigiensis*.\textsuperscript{40} The discovery of the ICK structure in scorpion supports the hypothesis of functionally divergent and convergent evolution of the ICK scaffold across evolutionary time and taxa. ICK motif peptides are largely antagonists and are active on a large variety of molecular targets.\textsuperscript{41}

An unusual feature of conotoxins is the extent to which they exhibit post-translational modifications, which confer chemical and structural changes to individual amino acids.\textsuperscript{42,43} A few examples of these modifications include hydroxylation of proline, gamma carboxylation of glutamate and bromination of tryptophan. While most if not all of the PTMs found in cone snails are found in mammalian and other systems, the density of these changes to the relatively short conotoxins sets them apart. From what has been observed thus far, at least 15 different modification enzymes are involved in conferring these alterations. The role of PTMs in bioactivity is not always clear. In some cases they are essential for bioactivity although this is not always the case. It is proposed that PTMs play a role in enhancing the diversity of conotoxin biophysical properties so that they may have greater selectivity for molecular targets in the nervous system.\textsuperscript{43}

### 1.5. Structural and genetic features of teretoxins

In comparison to conotoxins, less is known about teretoxin structural features and genetic organization. However, preliminary research has begun to identify the similarities and significant differences in terms of overall structural features (Table 1). To date the majority of the work characterizing teretoxins has been on the genetic level, where it has been established that a similar precursor structure of conserved signal sequence, intervening pro region, and highly variable mature toxin with conserved cysteine framework is present, although the propeptide region is less likely to be observed\textsuperscript{44,45}. A propeptide region is not always present in conotoxins, but this seems to be a more common feature in teretoxins. As with conotoxins, the mature toxin is released from the precursor structure via proteolytic processing and prior to assuming bioactivity.
Terotoxins, in general, are longer than their conotoxin counterparts. Terotoxins are often 40 amino acids or longer in length, whereas conotoxins average closer to 30 amino acids in length. Many of the conserved cysteine frameworks discovered thus far in terotoxins exhibit the same pattern found in conotoxins, which means that the number of intervening amino acids between cysteine residues is greater than conotoxins. The length of terotoxins is comparable to the neuropeptides of other venomous organisms, such as spiders and scorpions, where cysteine rich peptides targeting ion channels and ligand-gated receptors are often >40 amino acids. The increased length of terotoxins also results from a tendency towards conserved frameworks that have eight Cys, which is not as common in cone snails, and a paucity of the four Cys frameworks that are quite plentiful in cone snails.

Too few terotoxins have been characterized thus far to accomplish the work of classifying their toxins into gene superfamilies, or to evaluate whether similar relationships exist between conserved signal sequences and cysteine frameworks. It is also not known how precursor structures of conserved signal sequence in association with Cys framework may relate to pharmacological function. Thus far, the terotoxin signal sequences identified show little homology with conotoxin signal sequences, perhaps signifying the substantial evolutionary distance between these sister families.

Currently it is an open question as to the degree of post-translational modifications carried by terotoxins, although preliminary evidence has suggested either a lack or a very low level of the types found in conotoxins. This is a surprising finding given that conotoxins are some of the most highly

<table>
<thead>
<tr>
<th>Structural Features</th>
<th>Cone Snails</th>
<th>Terebrids</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Pro Organization</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Mature peptide length</td>
<td>Average 10 – 30</td>
<td>Often &gt;40</td>
</tr>
<tr>
<td>Conserved cysteine pattern</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Highly variable amino acids</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Post translational modification</td>
<td>✔</td>
<td>?</td>
</tr>
</tbody>
</table>

Table 1. Comparison of structural and genetic features of conotoxins and terotoxins
post-translationally modified peptides known, with a wide array of modifications. It has also been
demonstrated that post-translational modifications exist in some species of turrids. As the proteomic
research performed thus far to identify teretoxins arises from only a few different species and a limited
number of toxins, evaluation of the presence and degree of post-translational modification remains
speculative. To address the issue of PTMs in terebrids it is important to increase the database of
annotated transcriptome proteins, one of the outcomes of this work, and then confirm these findings on
the proteomic level.

Overall there is still much work to be done to characterize the composition, structure and function
of teretoxins. Understanding features such as gene superfamily organization, disulfide connectivity, and
the presence or absence of post translational modifications will be essential in carrying this work forward.
Combining proteomic and genetic analysis with the advanced techniques now available, in particular the
latest advances in mass spectrometry and next generation sequencing, promises to quickly advance our
understanding of the particular characteristics of the terebrid species-abundant neuropeptides.

1.6. Traditional vs. emerging techniques for venomic characterization

In the early days of research on the bioactive components of cone snail venom, dating back to
the 1970’s, the traditional approach was “bioactivity guided fractionation.” This process begins with the
extraction of crude venom from the snail either by dissecting the whole duct or milking the pure venom,
performing an extraction process in acetonitrile and subjecting this product to reverse phase HPLC for
venom fractionation. Isolation and selection of major peaks are then used for bioactivity guided assays,
in particular intracranial injection into mice for evaluation of unusual behavioral phenotypes. In this
approach, once a bioactive component is identified the peptide of interest may then be reduced, alkylated
and sequenced via Edman degradation, or using newer methods, mass fingerprinted and de novo
sequenced with mass spectrometry. Bioactivity guided fractionation has served to identify a number of
conotoxins of clinical and biological interest, despite the drawbacks of being time consuming, labor
intensive and expensive, it also requires many animals to get the requisite amount of venom. For smaller
specimens, where the amount of sample venom duct material is minimal, the bioactivity guided
fractionation approach becomes impracticable. Recent technological advances have enabled the
expansion of venom characterization to previously unsuitable specimens, broadening the reach of novel peptide toxin discovery and drug development. This emerging area is referred to as ‘Venomics’ and combines patterns in phylogeny and evolution with innovations in mass spectrometry and genetic sequencing.\textsuperscript{24,49–55}

### 1.6.1. Proteomic elution of venom components using mass spectrometry

Venomics, in the case of sample limited material, and more generally as a comprehensive approach to performing whole venom analysis to identify as many components as possible, the field of peptide toxin discovery employs an integrated platform that combines emerging and powerful techniques to analyze, in concert, both the venom duct proteome and transcriptome (Fig. 6).\textsuperscript{56–58} Driving the proteomic efforts is increased sensitivity, accuracy and resolution of various types of mass spectrometry and the use of matrix assisted laser desorption (MALDI-TOF) and electrospray ionization (ESI) technologies coupled with liquid chromatography.\textsuperscript{55} Developments in the area of tandem MS (MS/MS) assisted in surmounting obstacles to de novo sequencing, which typically was performed with collision activated dissociation (CAD). The incorporation of newer techniques such as electron transfer dissociation (ETD), along with chemical

---

**Figure 6. Integrated pipeline for peptide toxin discovery.** Transcriptomic and proteomic analyses are increasingly combined to identify novel peptide toxins with transcripts matching across platforms for database validation.
derivatization of cysteine residues to increase peptide charge state, have enabled more accurate
reconstruction of longer peptide sequences than achievable through CAD alone.\textsuperscript{54,59}

In spite of these advances, de novo sequencing with mass spectrometry to obtain peptide primary
structure is a difficult exercise, and integrating information from cDNA libraries and/or next generation de
novo transcriptome assembly can facilitate the process significantly. For example, genetic sequencing is
capable of identifying the full precursor structure of a toxin transcript, rarely achievable on the proteomic
level, whereas the identification of posttranslational modifications still relies on proteomic analyses of the
peptide itself.

1.6.2. \textit{Next generation sequencing applied to venom characterization}

Advances in next generation sequencing technology over the last decade can only be described as
revolutionary.\textsuperscript{60–62} Until recently, much of the genetic research to characterize the venom duct
transcriptomes of marine snails has relied on the construction of cDNA libraries.\textsuperscript{63,64} This process begins
with many of the same elements used in next generation sequencing, that is, the extraction of total RNA
from the venom duct, followed by reverse transcribing the mRNA content into cDNA. In some cases,
specific primers may be designed to target specific peptide toxin transcripts of interest, but more often
primers targeting the polyA tails of all mRNA transcripts have been used to obtain as many peptide toxin
transcripts as possible. Cloning into a bacterial host followed by plasmid extraction and Sanger
sequencing is then used to obtain as many transcripts as possible. The random nature of the process
inevitably means that not all peptide toxin transcripts will be captured, with a bias towards the most highly
expressed. Construction of cDNA libraries has served the venom discovery process well as numerous
peptide toxin transcripts from a diverse array of species have been identified in this manner. The
drawbacks to cDNA library construction include the lack of completeness, high cost, and the time and
labor intensive nature of the process.

The advent of next generation sequencing owes a great debt to the genius of the late, great Frederick
Sanger, who invented the technology in which dideoxynucleotides (ddNTPs) are incorporated into DNA
synthesis as chain terminating inhibitors of DNA polymerase, making sequence reconstruction a viable
endeavor.\textsuperscript{65} Next generation sequencing has advanced the Sanger technique, with several significant
alterations, which include the use of reversible terminating dNTPs in the sequencing by synthesis (SBS) reaction, and the ability to perform these reactions on a massively parallel scale. Each type of next generation sequencing technology varies in its particulars, such as the exact nature of the sequencing reaction, error rates and the length of reads generated, but all hold in common the ability to generate a significant amount of sequencing data from astonishingly small amounts of starting material.

Elucidation of peptide toxin coding transcripts in venomous organisms with NGS have been largely focused on RNA-seq technologies to perform whole transcriptome sequencing. For non-model organisms such as ours, obtaining as complete a snapshot as possible of the venom duct transcriptome has been favored over attempts to sequence full genomes given their large size and high repeat content, requiring huge amounts of data and a difficult assembly process. RNA-seq is a more lithe approach that gets the job done, in spite of the inherent challenges of de novo reconstruction in the absence of a reference genome. RNA-Seq provides the raw material for an assembly that can be analyzed for the presence of transcripts of interest either on its own, or ideally in combination with proteomics analysis. Advances in sequencing technology and assembly algorithms continue to improve the results obtainable for transcriptome reconstruction, especially where high read coverage allows for the capture of rare and alternatively spliced transcripts, enabling a more complete portrait of the transcriptome.

1.6.3. Technical aspects of next generation sequencing

A significant portion of the research that will be described here focuses on characterizing the terebrid venom duct transcriptome and has made use of the latest advances in high-throughput sequencing technologies and de novo assembly. As such it is instructive to lay out the groundwork of how this process works in its broad strokes, and what some of the challenges are in achieving both good sequencing data and a high quality assembly.

Given the efficiency of their sequencing technology, the low read error rates, and the generation of an enormous amount of data at a reasonable cost and quick turnaround time, Illumina technology has come to dominate the NGS field. While the details of the Illumina technology are proprietary, the overall scheme proceeds as follows: after generation of cDNA from RNA starting material, the cDNA is randomly fragmented with a shearing process, and adapters are ligated to the fragment ends (Fig.7). These
adapters are complementary to oligonucleotides that have been hybridized to the flow cell, providing a dense lawn of primers for adapter attachment at both ends. Bridge amplification for the generation of double stranded DNA, followed by denaturation to generate single stranded DNA, sets the stage for the sequencing by synthesis reaction to occur.

Figure 7. Illumina sequencing preparation and workflow. Construction of cDNA library from tissue-extracted RNA, and subsequent steps to prepare DNA (shearing and adapter ligation), prior to application to Illumina flow cell, cluster generation and sequencing by synthesis (SBS).

Sequencing by synthesis includes all four fluorescently labeled, reversibly blocked dNTPs, primers, and a DNA polymerase for the incorporation of one base per sequencing cycle. Excitation and fluorescent emission allow the recording of each base incorporated into the growing chain, with the cycles repeated until a segment of defined length has been sequenced. With Illumina HiSeq technology, one of their most utilized applications, the sequencing length has reached a maximum of 150 bp. Illumina
sequencing is also designed to allow for paired end sequencing. In other words, clusters are regenerated after the first sequencing round so that both ends of the template molecule may be sequenced.

It is remarkable to note that solid phase amplification on the Illumina flow cell surface allows for 1000 copies per single template molecule, with a proximity of one micron or less between clusters, leading to the generation of 10 million clusters per centimeter squared. In 2007 a single sequencing run could generate a maximum of 1 GB of sequencing data, and by 2011 this number increased to 1TB of data, a thousand fold increase. In 2014, Illumina released the HiSeq X 10, designed primarily for the purpose of sequencing an individual human genome, which can generate up to 1.8 TB of data in the space of three days and also achieved a much anticipated low price of $1000 per genome.

1.6.4. Bioinformatics tools for de novo assembly and analysis

Having a robust amount of high quality sequence data is an excellent first step on the road to studying a transcriptome or genome of interest, but in reality it is only the very beginning of a journey that requires many strategic decisions about assembly and analysis, given the biological questions under consideration and the many choices of bioinformatics tools that can help achieve specific goals. In step with the advances of next generation sequencing technologies, the computer programs designed to assemble the data have also had to evolve. Today the researcher is confronted with numerous software options, and it is often necessary to try multiple approaches for comparative and evaluation purposes.

The outcome of a de novo transcriptome assembly is a set of contigs, or “contiguous assembled pieces of DNA” that are representative of the transcripts coding for peptides and proteins of interest. Contigs are dependent on the assembly process and require further analysis to mine for the open reading frames that provide a basis for gene prediction. Assembly programs are designed to take raw reads and use specific algorithms to reconstruct as faithfully as possible the global genetic information from which the sequencing data was derived, but inevitably there are errors in the assembly process that lead to fragments and chimeric sequences that in addition to the well assembled contigs tend to inflate the assembly. Part of the challenge of de novo assembly in the absence of a reference genome is to identify and validate contigs as being truly representative of genes of interest.
The assembly process most often applies graph based reconstruction approaches that fall into two camps, overlap layout consensus (OLC) or de Bruijn methodologies. The OLC approach computes pairwise overlaps between reads and builds these into a graph that takes into account the global relationships among the reads. OLC is a traditional approach that is tailored towards longer reads where pairwise overlaps can be assessed with a reasonable degree of confidence, and was initially developed for handling Sanger sequencing in the early stages of piecing together the first human genome. While it is still used in software programs such as Mira for the assembly of 454 and other long reads, it does not scale well with an increasing number of short reads and becomes too computationally intensive for practical application.

With the arrival of sequencing technologies such as Illumina HiSeq, capable of generating hundreds of millions of reads that are typically 100 bp in length, developers of assembly programs turned towards the de Bruijn graph based approach to tackle the computational problem of assembly. In this approach, reads are subdivided into kmers of a specified length, which can vary depending on the number specified by k (Fig. 8). Each kmer successively overlaps the entire length of the read by itself.

![Figure 8. De Bruijn graph methodology. Simplified overview of De Bruijn graph approach to assembly with kmer of length 7. Each read is subdivided into kmer of with an overlap of length k-1. Independent of the reads themselves, assembly requires that each kmer ‘node’ is an exact overlap of k-1 with a directed edge connecting successive kmers (http://www.homolog.us/blogs/blog/2011/07/28/de-bruijn-graphs-i/)](image-url)
minus 1 (k-1), so that the kmer itself “slides” forward by one base across the read with each additional kmer. Once all the reads have been subdivided, the kmers themselves are used for graph construction, independent of the reads, by searching for exact overlaps of length k-1. Each unique kmer represents a node and an exact overlap represents an edge. As these connections are built up, genetic sequence can be reconstructed, one base at a time. However, as there are many potential paths that could be followed in finding these exact matches, the graph can become enormously complex, and the ultimate task is to traverse and compress the graph as correctly as possible while resolving alternatives that arise from true SNPs vs. sequencing errors, as well as numerous other stumbling blocks.  

Software programs such as Trinity and Velvet Oases are based in the de Bruijn graph approach, and are some of the most widely used programs for de novo assembly. As read quality continues to improve and new versions of these assembly programs are released, it is possible to acquire more accurate assemblies even in the absence of a reference transcriptome or genome. The current tendency, even for Illumina, which has specialized in producing shorter reads at high volume, is to generate ever longer reads. It will be interesting to see how the underlying algorithms of assembly programs will respond to these developments, and if ultimately there will be a return to an OLC approach, or if other computational alternatives will be explored. Historically, longer reads have shown significantly higher error rates, but this is destined to change as the technology keeps improving. Staying abreast of new developments both in terms of sequencing technologies and the software tools available to analyze data, not only for assembly, but for all the other facets of downstream analysis, is crucial to obtaining the best outcomes possible for the faithful reconstruction of either transcriptomes or genomes. Once a transcript sequence has been validated it is then possible to synthesize the peptide and apply it to biological assays to investigate its function.

1.7. Synthesis strategies for neuropeptides

As outlined above, identification of potential peptide toxins can be determined either from direct venom faction of the primary sequence (proteomics) or by genetic sequencing of the venom specific tissue (transcriptomics). As with all roads leading to Rome, identification of the toxin sequence, no matter how this is achieved, leads to chemical or recombinant synthesis of the peptide for application to
functional assays. Unlike milking snakes, conoidean snails produce significantly less venom, and effective functional characterization is greatly enhanced by synthetic peptide synthesis to obtain sufficient quantities for bioactivity assays. Choices for cysteine rich peptide synthesis fall into two camps, either the chemical method of solid phase peptide synthesis (SPPS)\textsuperscript{86}, or by taking advantage of nature’s machinery for recombinant expression of the peptide in a bacterial or other type of living host.\textsuperscript{87}

There are pros and cons to SPPS and recombinant expression of peptides, and the choice is generally dictated by features such as the presence or absence of post-translational modifications, the peptide length, and biochemical properties. For conotoxins, which tend to average close to 30 amino acids in length, as well as carrying numerous post translational modifications, SPPS is directly suited to their synthesis.\textsuperscript{88} However, SPPS becomes increasingly difficult and expensive with increasing length of peptide, and longer peptides may require chemical ligation reactions that are often problematic. In the case of longer peptides that lack post translational modifications, recombinant expression in heterologous systems offers an attractive alternative, with the most common approach being expression in an \textit{E. coli} host.

\subsection*{1.7.1. Solid phase peptide synthesis (SPPS) of venom peptides}

SPPS is a chemical synthesis technique pioneered by Robert Merrifield in the 1960s and for which he received the Nobel Prize in 1984.\textsuperscript{89} The major breakthrough for improving the synthesis process was conducting the reaction on a solid support, or resin, rather than in the liquid phase. With the use of a solid phase, the first amino acid is attached at its C-terminus to a resin via a linker, and subsequent elongation of the peptide is immobilized on the support until cleavage. The basic cycle of synthesis involves repeated rounds of amino acid couplings and deprotection, washing away excess reagents between steps. Every amino acid that is incorporated into the growing polypeptides carries a N-terminal protecting group that must be removed prior to coupling, while the addition of an activating group facilitates the formation of the peptide bond (Fig. 9).

The two major approaches to SPPS are acid labile and base labile, with the latter being favored in many life science laboratories for the use of less hazardous reagents and overall milder conditions.\textsuperscript{90} Base labile synthesis is referred to as Fmoc (9-flourenylmethylxocarbonyl) SPPS. Fmoc is a protecting group
that shields the alpha amino group from unwanted side reactions until it is deprotected under basic conditions in piperidine and Dimethylformamide (DMF). Fmoc also has the advantage of allowing for the use of acid labile protecting groups for amino acid side chains, as the Fmoc cleavage takes place under very mild basic conditions that are generally not reactive towards these groups. Upon completion of peptide synthesis, cleavage of the peptide from the resin is performed with trifluoroacetic acid (TFA). In contrast, Boc SPPS requires special equipment for cleavage from the resin using anhydrous hydrogen fluoride (HF). Upon completion of SPPS, the linear peptide requires the subsequent step of oxidative folding in which the formation of disulfide bonds is promoted by air oxidation and enhanced by a mixture of disulfide exchange shuffling reagents such as oxidized/reduced glutathione or other low molecular weight thiols.91

Many improvements have been implemented since the early days of SPPS, including optimization of resins, linkers, protecting groups and activating groups, as well as automated options, yielding a wide array of options to obtain the desired peptide sequence.92 Automated synthesizers and the variety of technologies behind them have helped to revolutionize the field in terms of both speed and labor.92 SPPS will undoubtedly continue to be the predominant approach to the synthesis of marine snail peptide toxins given its effectiveness for the synthesis of short peptides in high yield. Even so, there is still a useful role for recombinant expression in synthesizing toxins of interest, especially for longer peptides or those that prove difficult to synthesize under synthetic chemical conditions.
Figure 9. General scheme for Fmoc Solid Phase Peptide Synthesis. Synthesis process is based in rounds of deprotection, activation and coupling until final product is cleaved from resin. (http://www.sigmaaldrich.com/content/dam/sigma-aldrich/life-science/solid-phase-syn-schem.gif)
1.7.2. Recombinant expression of neuropeptides

Recombinant DNA technology has enabled scientists to produce a large number of diverse proteins, in various types of prokaryotic or eukaryotic hosts, which were previously unavailable, relatively expensive, or difficult to obtain in quantity.⁹³ *Escherichia coli* (*E. coli*) bacteria is a primary host for the cloning and expression of proteins of interest, however other eukaryotic systems, such as yeast or insect cells, have also been used effectively for the same purpose. Harnessing nature to do nature's work has a certain appeal, and it is remarkable the extent in which the genetic machinery of microorganisms can be manipulated with the introduction of foreign genetic material put under the control of an inducible promoter. The overarching scheme for recombinant expression of disulfide rich venom peptides requires a gene construct coding for the peptide of interest, a vector of choice, and an expression host (Fig. 10).⁹⁴,⁹⁵

![Diagram of recombinant expression process](image)

**Figure 10. Scheme for recombinant expression of disulfide rich venom peptides.** Major steps for expression of a disulfide rich peptide in an *E. coli* host with an emphasis on achieving solubility and folding.
Performing expression in *E. coli* has the advantage of several vectors and host strains that have been developed for maximizing expression while remaining economically attractive. Currently there are also a wealth of options for the molecular cloning step of inserting a gene construct into a plasmid vector that are more efficient and reliable than a traditional restriction enzyme approach.94,96

For many years the standard technique for inserting a gene into a plasmid has begun with PCR amplification of the gene to introduce restriction enzyme sites that are subsequently cut with the enzyme to generate “sticky ends” for ligation into a plasmid. The multiple cloning site (MCS) of a given plasmid vector offers many types of restriction enzyme sites given the need to identify unique cleavage sites at the ends of the gene insert. Both insert and vector are treated with the same enzymes to generate matching overhangs that can be stitched together in a directional orientation upon the addition of DNA ligase. While the restriction enzyme method is certainly functional, the success rate can be relatively low given the variable efficiency of restriction endonucleases, and may require multiple attempts to ensure that the gene is inserted correctly in frame into the vector of choice.94

In response to the difficulties encountered with restriction enzyme cloning, a variety of alternative techniques have been developed over the last several decades to improve the efficiency and speed of the important first step of recombining insert with vector. One such method is TA cloning, based in the addition of adenines to PCR amplified inserts via Taq polymerase activity, which may then be ligated into a plasmid vector carrying thymine extensions.96,97 While this method is simpler and faster than restriction enzyme cloning, and can be made especially rapid with the addition of topoisomerase as the ligase (TOPO-TA), the major drawback is that this type of cloning is not directional.
The development of ligation independent cloning (LIC) is an efficient process that has freed the molecular cloning process from the use of restriction enzymes and provides for directional cloning. LIC takes advantage of the simultaneous exo and endonuclease capabilities of T4 DNA polymerase to generate complementary 10 to 15 base pair overhangs on both insert and vector that enable hybridization by a simple annealing process (Fig. 11). The LIC approach requires that the flanking sequence at insert and plasmid terminal ends contain only three of the four possible nucleotide bases, so that the 3-5' exonuclease activity of the polymerase in the presence of a single dNTP corresponding to the missing base balances polymerization against base removal until a single stranded overhang of a specific length is generated. This process depends only upon the sequence of the flanking regions, and is independent

Figure 11. Ligation-Independent Cloning (LIC) strategy for recombinant expression of venom peptides. An efficient process for directional cloning based on the generation of long single “sticky ends” and subsequent annealing into vector (image adapted from Novagen Xa/LIC Cloning Kit Protocol).
of the coding sequences of the target gene and the functional sequences of the vector. Control over the length of the flanking regions allows for directional cloning, and the extended length of the overhangs enables ligase free hybridization of insert and vector for circularization of the plasmid.

Finding the most streamlined and effective approach for molecular cloning sets the stage for transformation into the bacterial host and ultimately, expression of the peptide or protein with attention to maximizing yield. The overexpression of cysteine-rich peptides can lead to problems with either proteolytic degradation or aggregation, and the undesirable formation of inclusion bodies. Choosing an environment favorable to disulfide bond formation is an important consideration in experimental design in order to circumvent the typical reducing environment found in many *E. coli* host strains.\(^ {100,101}\) To facilitate folding and avoid peptide aggregation, cysteine-rich peptides are often expressed in conjunction with carrier proteins that remain soluble and express at high levels.\(^ {102,103}\) Examples of these carrier proteins include maltose binding protein, glutathione S-transferase, and thioredoxin, with the latter having the advantage of facilitating disulfide bond formation.

Peptide expression as part of a larger fusion construct requires that a cleavage site be engineered into the gene construct for release of the peptide from the carrier protein post-expression. There are a number of options for the insertion of a cleavage site for proteolytic processing, including enterokinase, factor Xa, TEV protease and thrombin. While any of these may be used successfully for cleavage, a problem that may arise is non-specific cleavage, or limited access to the cleavage site due to protein structural features, and consequent low yield of final product.\(^ {88}\) The use of proteolytic enzymes can also add significantly to cost, and has led to the development of alternative methods such as the use of segments of proteins that are able to excise themselves by a self-catalytic mechanism (inteins). For recombinant expression work, inteins have been modified to undergo temperature and pH dependent cleavage of only the peptide bond between the C-terminus of the intein and the start of the downstream target protein, allowing for its release.\(^ {104,105}\)

The aspects of recombinant expression described here only begin to scratch the surface of options for obtaining a disulfide rich peptide of interest in sufficient yield while optimizing conditions for folding and solubility. A number of different conopeptides, as well as peptide toxins from scorpions, spiders sea
anemone have been successfully expressed in *E. coli* using a variety of approaches. In cases where soluble protein cannot be obtained, the use of strong denaturing reagents and subsequent refolding have been necessary. Recombinant expression will undoubtedly continue to play a role in the production of longer neuropeptides, not only in *E. coli* but also in eukaryotic systems such as the yeast strain *P. pastoris*, which are inexpensive to grow in culture, while providing an intracellular folding environment similar to that of mammalian cells. With some effort, it is possible to incorporate unnatural amino acids in recombinant systems, using a unique codon-tRNA pair and corresponding aminoacyl synthetase to recombinantly expressed modified or unnatural amino acids at specified sites in venom peptides.

### 1.8. Characterization of venom peptide bioactivity

Preliminary characterization of conoidean peptide toxins of interest is typically conducted via a phenotypic assay, such as intracranial injection in mice as pioneered by Olivera and colleagues, or teretoxin injection into prey organisms such as *Caenorhabditis elegans* (*C. elegans*). Injection of bioactive conoidean peptides into a mouse brain or prey organism elicits a phenotypic response that provides a global indication of function. Worm assays in both leech and *C. elegans* have been used to verify peptide toxin effects on behavioral activity, and changes in movement analyzed with the use of computer tracking algorithms can provide a quantitative assessment of motion changes such as paralysis or uncoordinated behavior.

To identify the specific target of a peptide toxin in the nervous system, a gamut of approaches can be applied, including classical methods such as two electrode voltage clamp or patch clamp electrophysiology performed in a variety of cell types, including mammalian DRG neurons and HEK cell lines. One widely used approach for electrophysiology studies is the heterologous expression of ion channels that is performed with either transfection of genes into tissue culture cells or with the injection of genetic material (cDNA or mRNA) into the larger egg progenitor cells (oocytes) of the African frog, *Xenopus laevis*. Today the oocyte model is still widely in use since its development in 1982, due to the reliable expression of genes coding for ion channels, their large size, and the absence of endogenous ion channels that would complicate electrophysiological recordings for a target channel. However, this
approach is time and labor intensive in cases where there is little information of the type of ion channel or ligand-gated receptor targeted by a novel uncharacterized peptide toxin, such as the ones discovered in this dissertation.

In an interesting twist on the oocyte approach, Dr. Ibañez-Tallon and colleagues developed a tethered-toxin method that takes advantage of some novel qualities of Lynx-1, a membrane-bound prototoxin identified as a modulator of the nAChRs in mammalian neurons. In short, recombinant expression is used to build a tethered-toxin construct where the Lynx-1 prototoxin sequence is replaced with a sequence coding for the neurotoxin of interest. Lynx-1 is bound to the membrane with a glycophasphatidylinistiol (GPI) anchor or transmembrane domain, ensuring the peptide is close to its molecular target. This modular system in principle allows for various peptide toxins to be inserted into the Lynx-1 scaffold and screened for activity in electrophysiology assays.

Beyond a classical electrophysiology approach to elucidate novel peptide toxin function, other techniques to assess changes in membrane potential include radioligand competitive binding assays, the use of ion sensitive fluorescent dyes and ion flux assays. Alternative approaches have been useful to profile compounds of interest, including conotoxins, but provide only indirect measures of ion channel activity. The need for high-throughput ion channel screening utilizing electrophysiology has led to the development of a variety of automated platforms for screening compounds in mammalian cell lines and oocytes. Recently described high throughput approaches may help to overcome the labor-intensive and slow nature of patch clamp electrophysiology, so that novel venom peptides may be screened more rapidly and efficiently for drug discovery and development.

1.9. Objectives of Thesis Project

The primary focus of this research has been to apply a venomics approach via state of the art genetic techniques to characterize novel neuropeptides found in venom duct transcriptome of terebrid marine snails, and to use this information to synthesize peptides of interest, using either chemical or recombinant strategies, in sufficient quantities for bioactivity assays. Characterization of the venom duct transcriptome made use of next generation sequencing in combination with a range of bioinformatics tools for de novo assembly and analysis to identify as many terebrid toxin transcripts as possible from two
terebrid species, *C. anilis* and *T. subulata*. This work highlights the ability of deep sequencing to surpass previous genetic efforts of cDNA library construction and Sanger sequencing at peptide toxin characterization. De novo assembly in the absence of a reference genome requires the deployment of several bioinformatics tools to achieve the best results, and a significant effort has been made to utilize multiple approaches for cross validation of results and achieve the highest degree of accuracy possible in terebrid venom duct transcriptome reconstruction. The teretoxin transcripts identified herein have been evaluated in the context of the large body of research regarding the genetic organization, structure and function of conotoxins, the closest evolutionary relative. In characterizing venom peptides from terebrid transcriptomes, an analysis of non-toxin genes was conducted, using homology based searches to categorize transcripts by gene ontology function and obtain a global portrait of the highly specialized venom duct tissue and the evolution of convergent and divergent venom compounds.

A further goal of this research has been the synthesis and folding of select terebrid toxins to obtain sufficient quantities for characterization of bioactivity. The two major methods used were solid phase peptide synthesis (SPPS) and recombinant expression. While solid phase peptide synthesis has been the method of choice for synthesizing conotoxins for the last several decades, due to the longer residue length of teretoxins, which can be intractable using SPPS, the focus here has been on the optimization of recombinant expression for the expression of disulfide rich peptides that are not amenable to chemical synthesis. The ultimate goal of teretoxin synthesis has been to provide a sufficient quantity for bioactivity assays, and perform preliminary tests of bioactivity, such as a phenotypic worm assay.

The research outlined in the remaining pages of this dissertation relay the first comprehensive genetic analyses of Terebridae venom peptides.
Chapter 2

De novo characterization of Terebrid venom duct transcriptomes
2.1. Background: a revolution of model systems

In the postgenomic era the concept of a model system is rapidly evolving. Customary model organisms such as D. melanogaster, C. elegans, and M. musculus, while still gold standards, no longer corner the market, and it is possible with ever advancing sequencing techniques to literally scrape the ocean floor for organisms that produce novel genes and gene products. Previously unsuitable organisms, due to size, availability, or gene complexity, are being investigated at an increasing pace because of their unique qualities and the possibility of identifying novel ligands for understanding the molecular basis of biological processes. With decreasing costs and increasing technology, molecular and functional genomics studies enable venomous organisms to become model organisms in the drug discovery arena. In the search for new drug discovery model systems, marine conoidean organisms are leading the charge. The enormous diversity of conoidean venom peptide toxins greatly outnumber that of snakes, a prior pharmaceutical industry favorite due to ease of collection and venom milking capabilities. This thesis project applies next generation sequencing techniques to study the venom duct transcriptome of Terebridae marine snails, to characterize terebrid unique features in comparison to cone snails, particularly as it pertains to elucidating novel genes and gene products that can be used to manipulate cellular communication for the development of new therapeutics.

As high throughput sequencing technology and computational methods continue to improve, the accurate reconstruction of complex venom duct transcriptomes is accelerating at an equal rate, even for venom duct tissue only available in vanishingly small amounts. Previous cone snail venom duct transcriptome surveys have produced results that are pertinent to drug discovery and development, such as identifying new venom gene superfamilies and peptide toxin ligands. Interestingly, the majority of recently published studies on venom duct transcriptomes of cone snails and snakes have used Roche 454 sequencing. The bias towards 454 technology in venomics research has likely resulted from the long average 454 read length, which has the potential to cover an entire peptide transcript prior to assembly. 454 technology, however, is hindered by low read coverage and a high error rate especially with respect to indels. Illumina technology was chosen to generate high throughput terebrid venom duct transcriptome sequencing data for this thesis. Compared to 454 Illumina has better read accuracy and far
superior depth of coverage. Longer read lengths, depth of coverage, and deployment of ever expanding bioinformatics tools allow unparalleled insight into the diversity of animal venom peptide transcripts. It should be noted that de novo characterization of a transcriptome, in the absence of a reference genome, continues to be a challenge, but specific measures may be taken to improve the confidence level of transcripts by strict assessment of read quality, the right choice of assembly tools, and comparison of results among different bioinformatics programs to validate results.\textsuperscript{66,76}

The first de novo assembly and analysis of terebrid venom duct transcriptomes using RNA-seq are presented in this work. Two species of terebrid were selected for characterization, \textit{Cinguloterebra anilis} (\textit{C. anilis}) and \textit{Terebra subulata} (\textit{T. subulata}). Both are vermivorous (worm hunting) species that similar to cone snails, uses a sophisticated venom apparatus to inject venom into prey via a hollow, harpoon shaped radular tooth (Fig. 1). Not all terebrid snails have a venom apparatus to subdue their

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{\textbf{Terebridae shell and venom apparatus schematic.} \textbf{A.} Photo of \textit{Cinguloterebra anilis} and \textbf{B.} \textit{Terebra subulata}. \textbf{C.} Illustration of conoidean venom apparatus with salivary gland (sg) venom gland (vg), muscular bulb (mb) and radular sac (rs) \textbf{D.} SEM of radular tooth from Terebridae clade C (Venom apparatus illustration and SEM image courtesy of Yuri Kantor)}
\end{figure}
prey. In 1970, Miller characterized three different hunting physiologies for the Terebridae. Recent studies have facilitated the identification of terebrid lineages that produce venom for prey capture by correlating the molecular phylogeny of the Terebridae to the evolution of its venom apparatus (Fig. 2). Using this biodiversity derived discovery approach C. anilis and T. subulata were selected for venom characterization. Both C. anilis and T. subulata are part of the Terebra genus (Clade C), in which all species of terebrid have a venom apparatus (Fig 2).

Figure 2. Venom apparatus characterization in the Terebridae. The presence or absence of venom duct (gland) mapped to the molecular phylogeny of the Terebridae. (Castelin et al., 2012)

The primary goal of terebrid transcriptome research is to characterize disulfide rich peptides expressed in the venom duct in order to identify ligands that may be subsequently used to manipulate signaling in the nervous system. This research has also been designed to provide an assessment of other proteins expressed in the terebrid venom duct and to organize identified proteins by function with gene ontology (GO) and enzyme (KEGG) classifications in order to enhance the knowledge set of venom genes for model system comparative analyses. Finally, this work establishes a robust bioinformatics pipeline that runs the entire gamut of sequencing and analysis that can be applied to ongoing research on other terebrid specimens (Fig. 3).
Figure 3. Bioinformatic pipeline for obtaining and characterizing terebrid venom duct transcriptome data. The outlined work flow is based on tissue preparation and Illumina sequencing (Phase I). Upon obtaining sequences, quality checks are performed (Phase II), and valid sequences are then assembled and evaluated (Phase III). The final step in the workflow is analysis of the assemblies (Phase IV).

2.2. Results and Discussion

2.2.1. Hi-Seq Illumina sequencing and overview of C. anilis transcriptome assemblies

Reconstructing a transcriptome with next generation sequencing technology requires several major steps that include extraction of RNA, construction of a cDNA library, high-throughput sequencing, and de novo assembly. Here total RNA extracted from eight pooled C. anilis venom ducts was used to generate a polyA enriched cDNA library for sequencing on the Illumina HiSeq platform. A single lane from the Illumina flow cell generated 288,959,674 paired end reads of 100 bp length, which after quality control as described in materials and methods reduced the total number to 280,143,112 reads.

Given the large number of reads and high depth of coverage, digital normalization was applied to the trimmed reads prior to assembly with Trinity and Velvet Oases respectively, in order to reduce the size of the data, while retaining the information found in the full dataset by eliminating highly redundant read data. This digital normalization technique also eases the computational burden of assembling
hundreds of millions of reads, reducing the time and memory requirements of de novo assembly, which is advantageous for performing multiple assemblies. In the case of Trinity, assemblies were constructed for both the full set of raw reads and digitally normalized reads. Given the large memory requirements of Velvet Oases, de novo assemblies with this program were run only on digitally normalized reads, over a range of kmers from 25 to 55. Statistics for all assemblies are reported in Supplementary Table 1.

A total of four assemblies, including Trinity assemblies and Velvet Oases assemblies run at kmers 25 and 51, respectively, were used for the identification of novel neuropeptides. The four assemblies are: 1. TrinDN (Trinity assembly from digitally normalized reads), 2. TrinAllReads (Trinity assembly run on full set of reads), 3. VO25 (Velvet Oases assembly run at kmer=25), and 4. VO51 (Velvet Oases assembly run at kmer=51). Four assemblies were used to maximize the capture of putative teretoxins and for comparative analyses across assemblies to support the de novo assembly process. Assemblies that are run over a range of kmers, as is the case with Velvet Oases, may subsequently be meta-assembled using a tool such as CAP3, CD-HIT-EST, or program specific tools such as Oases merge. However, the risk of meta-assembly lies in the artificial collapse of isoforms into one transcript and the potential for misassemblies, exacerbating the problem of chimeric transcripts already found in individual assemblies. Toxin transcripts may be particularly at risk for loss of discrete transcripts or chimeric assembly under the conditions of merging, as they are rich in isoforms that can vary from minor amino acid substitutions to increasing degrees of variability. The Trinity assembly of digitally normalized data was chosen as the primary tool for downstream analysis of C. anilis sequences, as the software is frequently updated and is specifically designed to capture transcript isoforms.

2.2.2. Identification and validation of teretoxin transcripts

Terebrids, like their cone snail counterparts, express a broad repertoire of highly variable, disulfide rich peptides. Precursor teretoxin transcripts contain a mature peptide toxin with a highly conserved cysteine (Cys) framework in combination with hyper-variable amino acids found in between cysteines. The abundance of peptide toxins in conoidean venom duct is believed to stem in part from high rates of gene duplication, while the hypervariable mature peptide region likely results from strong diversifying selection at specific loci, leading to significant allelic variation. Part of the challenge of
parsing a de novo assembly is identifying with confidence as many peptide toxin transcripts as possible. To begin this work, the four assemblies referenced above were blasted against a dataset of conotoxins downloaded from www.conoserver.org and an in house database of teretoxins representing a variety of terebrid species. Curation of these results yielded 84 putative teretoxins, which are organized by Cys frameworks using the same Roman numeral nomenclature as for conotoxins. A similar approach was applied to identify *T. subulata* toxin transcripts, using only one assembly, Trinity, on all reads. (Table 1).

**Table 1: Summary of *C. anilis* putative teretoxins**

<table>
<thead>
<tr>
<th>Framework</th>
<th>Cysteine Pattern</th>
<th><em>C. anilis</em> Transcripts</th>
<th><em>T. subulata</em> Transcripts</th>
<th>Connectivity (cone snail)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>CC-C-C</td>
<td>1</td>
<td>1</td>
<td>I-III, II-IV</td>
</tr>
<tr>
<td>VI/VII</td>
<td>C-C-CC-C-C</td>
<td>16</td>
<td>16</td>
<td>I-IV, II-V, III-VI</td>
</tr>
<tr>
<td>VIII</td>
<td>C-C-C-C-C-C-C-C-C-C-C</td>
<td>16</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>IX</td>
<td>C-C-C-C-C-C-C</td>
<td>16</td>
<td>12</td>
<td>I-IV, II-V, III-VI</td>
</tr>
<tr>
<td>XI</td>
<td>C-C-CC-CC-C-C-C</td>
<td>2</td>
<td>2</td>
<td>I-IV, II-VI, III-VII, V-VIII</td>
</tr>
<tr>
<td>XIII</td>
<td>C-C-C-CC-C-C-C-C-C-C</td>
<td>5</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>XIV</td>
<td>C-C-C-C-C</td>
<td>4</td>
<td>1</td>
<td>I-III, II-IV</td>
</tr>
<tr>
<td>XV</td>
<td>C-C-CC-C-C-C-C-C</td>
<td>5</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>XVI</td>
<td>C-C-C-C-C</td>
<td>-</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>XVII</td>
<td>C-C-CC-C-CC-C-C-C</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>XXII</td>
<td>C-C-C-C-C-C-C-C-C-C-C</td>
<td>16</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>XXV</td>
<td>C-C-C-C-C-C-C-C-C-C-C-C</td>
<td>-</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Novel – 10 cys</td>
<td>C-C-C-C-C-C-C-C-C-C-C-C</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Novel – 12 cys</td>
<td>C-C-C-C-C-C-C-C-C-C-C-C-C</td>
<td>1</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td><strong>Total putative toxin transcripts:</strong></td>
<td><strong>84</strong></td>
<td><strong>65</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For *C. anilis*, the 84 putative teretoxins that were identified comprise nine Cys frameworks that are known in conotoxins, and two novel frameworks of 10 and 12 Cys respectively. The diversity of Cys frameworks found suggests that teretoxins have a wide array of pharmacological targets. The 84
transcripts were analyzed for the presence of canonical peptide features as identified in conotoxins, namely the presence of a N terminal signal sequence, an intervening propeptide region ending with several basic residues identifying a cleavage site, and a C-terminal Cys rich mature peptide. Of the 84 neuropeptide transcripts identified from BLAST results against the combined cono/teretoxin database, 64 contain the full signal-pro-mature toxin canonical structure, and 20 have signal-mature sequence without a pro region, which also occasionally occurs in cone snails.

To establish a high confidence level for the 84 newly identified teretoxin transcripts, where possible, the transcripts were cross-validated with all four de novo assemblies. A Venn comparison was used to identify multiple sequence alignments of the same putative peptide toxin identified separately in each assembly (Fig.4). In the best case scenario, transcripts have unequivocal support from all four assemblies, although it is possible for a transcript to be valid even if only identified from one assembly.

Since different assembly programs run with different parameters may be more effective at picking up specific transcripts where others fail to do so. For example, Trinity assembly algorithms are designed to tease out isoforms, while Velvet Oases assemblies constructed from both high and low kmer values

Figure 4. Venn diagram showing overlap of putative toxin transcripts for Velvet Oases and Trinity assemblies. A. Comparative analysis of the four assemblies conduct on C. anilis transcriptome sequences. VO25 = Velvet Oases assembly with kmer25. TrinDN = Trinity digital normalization assembly. TrinAllReads = Trinity assembly without digital normalization. VO51 = Velvet Oases assembly with kmer51. B. Representative alignments of transcripts from all four assemblies that were used to generate Venn diagram. Levels of support range from all four assemblies to single assembly.
provide different parameters for coverage of both rare and more highly expressed transcripts.\textsuperscript{81,85} Venn analysis identified 27 transcripts with complete overlap in all four assemblies, indicating there is a high probability these are valid sequences.

The disulfide rich proteins of venomous peptides present a complex array of cysteine (Cys) patterns in terms of both arrangement and number. Cys patterns are so important in conotoxins that they have been organized extensively by conserved frameworks, in association with conserved signal sequence, in order to better understand how gene families of peptides relate to pharmacological activity. Importantly, two identified \textit{C. anilis} putative teretoxins have a novel Cys framework with respect to established cone snail frameworks. \textit{Can\_10Cys}, with the cysteine framework C-CC-C-C-C-C-C-C-C, and \textit{Can\_12cys}, with the cysteine framework C-CC-CC-C-C-C-C-C-C-C. \textit{Can\_10Cys} and \textit{Can\_12Cys} precursor structures have identical signal sequence and pro regions, while the mature toxin varies in terms of both cysteine framework and intra-Cys amino acid residues. Conotoxin classification has shown increasing evidence that gene superfamilies often associate with more than one Cys framework.\textsuperscript{37} With only two transcripts identified in this case, further analysis from other terebrid transcriptomes is required to determine if these complex associations between signal sequence and frameworks also holds true for teretoxins.

The most prevalent neuropeptide transcripts identified in the four assemblies of \textit{C.anilis} sequences fall into four different Cys frameworks (VI/VII, VIII, IX, and XXII), of varying number and pattern. The VI/VII framework of the O superfamily (C-C-CC-C-C) in cone snails, is the most heavily represented in \textit{C. anilis} with 19 unique sequences with full precursor transcripts that include signal sequence, pro region, and mature toxin. O superfamily conopeptides, which are active on Na\textsuperscript{+}, K\textsuperscript{+}, and Ca\textsuperscript{2+} channels, are also a predominant gene family in cone snail venom.\textsuperscript{70} O superfamily peptides produce immobilization and neuromuscular block in the prey by inhibiting ion channel flux.\textsuperscript{137} The other heavily represented Cys frameworks include: 15 precursor teretoxin transcripts of the 6 Cys framework IX peptides (C-C-C-C-C-C), 16 precursor teretoxin transcripts of the eight Cys framework XXII (C-C-C-C-C-C-C-C-C-C), and 16 precursor teretoxin transcripts of the ten Cys framework VIII (C-C-C-C-C-C-C-C-C-C-C). Of the four predominant Cys frameworks, VI/VII and VIII conotoxins are known to target serotonin receptors,
nAChRs, Na^+, K^+, and Ca^{2+} channels, while the molecular targets of frameworks IX and XXII are unidentified.

Frameworks with five or fewer *C. anilis* teretoxin transcripts, include the eight Cys frameworks XI (C-C-CC-CC-C-C-C-C-C), XIII (C-C-CC-C-C-C-C-C-C) and XV (C-C-CC-C-C-C-C-C-C), the four Cys framework XIV (C-C-C-C-C), and a single transcript for the four Cys framework I (CC-C-C). The low representation of the latter is somewhat surprising as Framework I peptides are ubiquitous in cone snails, and are known to frequently target nAChRs. Representative teretoxin transcripts for each Cys framework found in *C. anilis* assembly are illustrated in Figure 5, with alignment to its closest BLAST hit.

---

**Figure 5.** Representative BLAST alignments of *C. anilis* teretoxin transcripts. Roman numerals on the right indicate the corresponding Cys framework assigned to each transcript. Alignments show homology of *C. anilis* toxins to the top blast hit, either to existing teretoxin or conotoxin, and conservation of cysteine patterns in mature toxins.
An interesting feature of the of *C. anilis* framework VI/VII peptides, is the repeated appearance of a conserved amino acid motif between the first and second Cys residues, best defined as PXY (Pro-X-Tyr), where X can be any intervening residue. An example of this is seen in Can.6.14, which has strong homology to a teretoxin (Tg77) previously identified from a *Terebra guttata* cDNA library (Holford, unpublished data). Similar to the ICK (Inhibitor Cysteine Knot) motif in conopeptides, the PXY motif may determine the structural and functional selectivity of teretoxins to their molecular target. Proline residues are known to affect the secondary structure of proteins. Unlike other amino acids, Proline is a secondary amine whose cyclic side chain provides rigidity to peptide chains that is usually represented by a ‘bend’ in structure. Tyrosine residues have a reactive phenol (–OH) that can act as an acceptor of phosphate groups. Phosphorylation of phenol groups via receptor tyrosine kinases is a key feature of signal transduction processes. Proline and Tyrosine together in the PXY motif impart significant conformational and functional properties that may be important to teretoxin peptide function. Framework VI/VII conotoxins often target Na$^+$, K$^+$, and Ca$^{2+}$ channels, but it remains to be seen if a similar pattern will emerge for teretoxins with a PXY motif.

As illustrated by Can.9.13, the 6 Cys framework IX teretoxins have structural features similar to conotoxins, namely a full precursor organization (pre-pro-mature) and tightly knit mature peptide resulting from a limited number of intra-Cys residues. Alternatively, framework VIII and XXII teretoxins display a precursor structure where signal sequence is either followed by a very short pro region or directly by the mature peptide. In addition, the mature peptide has more intra-cysteine residues than commonly seen in cone snails, resulting in a longer overall sequence. This alternative precursor structure is seen in Can.8.14 and Can.22.12. While the lack of a pro region is occasionally observed in conotoxins, the loss of this region in some teretoxin precursor transcripts raises interesting questions about the purpose of the pro region, which is currently not well understood, but is implicated in the protein disulfide assisted mature peptide folding process. The eight Cys framework XXII was only recently observed in conotoxins, but is quite prevalent in *C. anilis* transcripts. As noted above, the function of peptides with this framework is unknown.
Relatively few conotoxin sequences are reported for framework XI, (I superfamily) XIII (G superfamily) and XV (V superfamily) peptides, all of which contain 8 Cys in different conserved patterns. The I superfamily, composed of three subfamilies, is the best characterized in terms of pharmacology, with some conotoxins functioning as \( \text{Na}^+ \) channel agonists and \( \text{K}^+ \) channel modulators. Framework XIII conotoxins are extremely rare and have only recently been assigned to a novel G superfamily.\(^{142}\) These frameworks are all present in \( C. \text{anilis} \) teretoxin transcripts.\(^{37}\) The representative framework XI sequence shown, Can.11.1, has its closest BLAST hit to a teretoxin identified from \( T. \text{guttata} \), and as such differs significantly in sequence from its conotoxin counterparts.

Several framework XIV transcripts are present in the \( C. \text{anilis} \) venom duct, with one in particular, Can.14.1, displaying a high degree of homology in the mature peptide region to asXIVa and VilXIVa conotoxins. This is the only example from this assembly of a teretoxin displaying such a high degree of homology to a conotoxin. Framework XIV conotoxins are a complex assortment of peptides classified into multiple superfamilies that include A,I2,J,L,M,O1, and O2.\(^{143}\) Several framework XIV conotoxins have been shown to elicit nAChR and \( \text{K}^+ \) channel inhibition.\(^{144,145}\) Given its similarity to asXIVa and VilXIVa conotoxins, Can.14.1 was synthesized and an attempt was made to characterize its bioactivity (see Chapter 3).

Finally, a single \( C. \text{anilis} \) transcript with 16 Cys residues shows striking similarity to the conopressin/conophysin family of peptides identified from various cone snail species (Fig. 6). Conopressins (C-C) are short, nine residue peptides with two cysteines, which were originally discovered in the venoms of \( C. \text{geographus} \) and \( C. \text{striatus} \), and have been characterized as vasopressin/oxytocin homologs based on their strong sequence similarity to these hormonal neurotransmitters.\(^{146}\) Despite the strong similarity to vasopressin, conopressin-T, isolated from the venom of \( C. \text{tulipa} \), has been shown to act as a V1-vasopressin receptor antagonist and a partial oxytocin receptor agonist, generating considerable interest in having conopressins serve as templates for drug design.\(^{147}\)
The putative *C. anilis* conopressin-like transcript, identified here as "terepressin," shares strong sequence homology with those identified from cone snails (Fig. 7). With such short constructs, it is possible for a single amino acid change to alter the affinity of the cono/terepressin for the receptor. This is the first such transcript to be identified from a terebrid snail, and it is a distinct possibility that other variations are present in other terebrid species that can be identified via transcriptome characterization.

Where conopressins are among the shortest conopeptides identified, conophysins, which belong to the neurophysin peptide family, are one of the longest peptides ever identified in cone snail venom, with 14 cysteine residues (C-C-C-CC-C-C-C-C-CC-C-CC-C-C) and 7 disulfide bridges. Conophysins were first characterized as a major component in the venom of *Conus radiatus*, but no physiological role is

---

**Figure 6.** Precursor structure of *C. anilis* transcript aligned to conophysin and conopressin sequences from cone snails. Conopressin and conophysin are expressed in one construct as demonstrated by alignment of *C. anilis* contig to recently identified, similar construct from *C. geographus*.

The putative *C. anilis* conopressin-like transcript, identified here as "terepressin," shares strong sequence homology with those identified from cone snails (Fig. 7). With such short constructs, it is possible for a single amino acid change to alter the affinity of the cono/terepressin for the receptor. This is the first such transcript to be identified from a terebrid snail, and it is a distinct possibility that other variations are present in other terebrid species that can be identified via transcriptome characterization.

Where conopressins are among the shortest conopeptides identified, conophysins, which belong to the neurophysin peptide family, are one of the longest peptides ever identified in cone snail venom, with 14 cysteine residues (C-C-C-CC-C-C-C-C-CC-C-CC-C-C) and 7 disulfide bridges. Conophysins were first characterized as a major component in the venom of *Conus radiatus*, but no physiological role is

---

**Figure 7.** Alignment of a *C. anilis* terepressin to Conus conopressins. Alignment of *C. anilis* conopressin-type (terepressin) sequence in alignment with all known conopressins.
currently understood for these peptides. This *C. anilis* full precursor transcript provides evidence for cono/terepressin and cono/terephysin being expressed together, similar to vasopressin/neurophysin homologs from other organisms. While the terepressin portion of the transcript shows close homology to conopressin-G (Fig. 7), the signal and terephysin sequences are much more variable compared to the signal and conophysin sequences of the full precursor structure recently identified in *C. geographus* (Fig. 6). Vasopressin/oxytocin hormonal neurotransmitters are emerging as viable targets for mental disorders, such as autism, social anxiety disorder, and schizophrenia. It is conceivable that co-expression of vasopressin/neurophysin hormonal neurotransmitters may facilitate how they are regulated in biological processes.

It should be noted that even when a potential teretoxin with a well-established conotoxin framework is identified in this study, the overall amino acid composition of the mature peptides is in most cases radically different from conotoxins in terms of identity, number and distribution of intra-cysteine residues (Fig. 8). As such, it remains to be seen whether teretoxins with the same conserved Cys frameworks found in cone snails will act upon similar ion channels and receptors. Given their significant differences from conotoxins, teretoxins may have novel molecular targets, thus identifying new mechanisms in drug discovery and development for therapeutic agents that modulate the nervous system.
Figure 8. Circos representation of *C. anilis* putative toxin blast results to known teretoxins and conotoxins. As illustrated by the increased number of BLAST hits of *C. anilis* putative toxins (green lines) compared to known teretoxins and specific conotoxins (blue lines), they may have different functional selectivity and specificity for neuronal molecular targets.
2.2.3. Preliminary analysis of Terebra subulata venom duct transcriptome

Having established a working pipeline for characterization of the *C. anilis* venom duct transcriptome, a similar approach was applied to the venom duct transcriptome characterization of *Terebra subulata* (*T. subulata*). Reports in the literature have presented research on *T. subulata* primarily on the phylogenetic level, although a few of its venom components as well as some preliminary bioactivity have also been characterized.\(^{12,47,134,150}\)

As with *C. anilis*, the sequencing effort began with the pooling of four venom ducts and extraction of total RNA. Paired end reads of 100 bp in length were generated on the Illumina HiSeq platform, producing 176,799,164 raw reads with 92.42% having a quality score of Q30 or better. Further read quality assessment was performed as described above and the full set of reads was assembled with the Trinity program. Assembly generated 193,988 contigs > 200 bp in length, with an N50 of 924bp and an average contig size of 633 bp. In addition to Illumina sequencing, the read data was augmented by performing 454 sequencing to take advantage of the longer read lengths. This sequencing effort generated 117,750 reads of average length 426.6 bp. 454 reads were assembled separately with the program Mira, which generated 9,464 contigs.

The Trinity assembly of 454 contigs and 454 raw reads were blasted against an in house database of conoidean peptide toxins, which includes all teretoxins sequenced to date, including the *C. anilis* teretoxins, conotoxins, and any turrid or terebrid toxins reported in NCBI. A first pass of combined BLAST results of all data yielded 65 putative *T. subulata* teretoxins in their full precursor structure, distributed across 11 cysteine frameworks. The initial *T. subulata* BLAST results accurately captured all *T. subulata* teretoxins either sequenced in house via cDNA work or those reported in the literature, confirming the quality of the *T. subulata* assembly and the integrity of the putative teretoxins identified thus far (Fig. 9).
Figure 9. Alignment of *T. subulata* toxins to previously identified teretoxins. The *T. subulata* toxins identified from Trinity assembly in alignment to *T. subulata* toxins identified by cDNA cloning work (performed by M. Holford) or identified in the literature.
2.2.4. Comparative analyses of C. anilis and T. subulata venom duct transcriptomes

Comparative analysis of the types and number of Cys frameworks found in C. anilis and T. subulata transcriptomes highlight strong similarities between the two venoms (Fig. 10). The most abundant Cys frameworks are VI/VII (C-C-CC-C-C), VIII (C-C-C-C-C-C-C-C-C-C), IX (C-C-C-C-C-C) and XXII (C-C-C-C-C-C-C-C-C-C), which confirm the tendency of teretoxins towards an increased number of Cys residues and generally longer peptides than those found in cone snails. The absence of framework I peptides (CC-C-C) in T. subulata, as also seen in the C. anilis venom profile, is in striking contrast to the abundance of this category in conotoxins, where peptides of this framework are known to most frequently target nAChRs.

Figure 10. Distribution of T. subulata and C. anilis toxins among different Cys frameworks. Comparison of how T. subulata and C. anilis Cys frameworks are distributed show a strong degree of similarity between the two venom duct transcriptomes.

It may be argued that terebrids have evolved a different Cys framework to target ligand gated nAChRS given that these are a common target for venomous organisms and an effective target for quickly disrupting nervous system transmission. It remains to be seen what type of terebrid toxin this might be,
and whether it arises from a shorter cysteine framework as in cone snails, or from longer, more cysteine rich peptides.

2.2.5. Identification of the first Teretoxin gene superfamily

Identification of venom components from two different transcriptomes in combination with previous cDNA work done on other terebrid species have set the stage for the analysis of gene superfamilies that define the association between conserved signal sequences and specific frameworks (Fig. 5, chapter 1). At least 60% sequence identity in the signal peptide is the expectation for designating a superfamily. Investigation of one conserved signal sequence observed in different terebrid species has provided the basis for characterizing the first gene superfamily identified in terebrids (Fig. 11).

**Figure 11. Terebrid gene superfamily based on conserved signal sequence.** The conserved signal sequence highlighted (pink bar) was used to identify the first gene superfamily for terebrids. The conserved sequence is seen here in association with the six Cys framework C-C-CC-C-C, typical of the O superfamily in conotoxins.

The conserved signal sequence, MATSGRLLCLCLVLGLVF (with minor variations) was identified in three different terebrid species, *T. subulata*, *T. guttata*, and *C. anilis*, in association with the six Cys framework C-C-CC-C-C. In cone snails, this framework is found extensively in the O superfamily, although recently a new H gene superfamily has been defined with this framework present, based on venomics work done in *Conus marmoreus* and *Conus victoria*.

Analysis of terebrid toxin signal peptides thus far has shown low homology with the signal peptides found in cone snails, but here there is an exception insofar as the
The teretoxin signal peptide identified has strong sequence identity (>80%) with that of the H superfamily, as demonstrated by the alignment with H_vc7.2 from *C. victoria* (Fig. 11). It is also interesting to note that while the mature peptide sequences diverge at many positions, there is 100% conservation of glycine at position 68, phenylalanine at position 74, aspartate at position 80 and leucine/isoleucine at position 85. To the best of our knowledge, no bioactivity has yet been investigated for the H superfamily, which was only recently been characterized.

Identification of the first gene superfamily in terebrid toxins is an important breakthrough that frames the context in which teretoxins will be characterized.

2.2.6. Identification of Teretoxin Transcripts with “Pepticomb”

As terebrids and their toxins are in the preliminary stages of investigation, and currently have very low representation in public databases, it is important to have alternatives to BLAST as a means to identify putative toxin transcripts. To this end an in house software program, termed Pepticomb, was developed and used to analyze the *C. anilis* Trinity DN assembly. Pepticomb is largely based in source code kindly provided by the developers of the conoprec tool found on the conoserver website ([www.conoserver.org](http://www.conoserver.org)). The major objective was to use this code so that a conoprec type of analysis could be run as a standalone program on the large dataset produced by *de novo* assembly of the terebrid venom duct transcriptome. In stepwise fashion, Pepticomb examines a set of contigs for the presence of a signal sequence, looks for a pro-region that terminates in a basic residue cleavage site, and uses regular expressions to mine for permutations of Cys frameworks in the mature peptide with a limited number of intra-cysteine residues set by the user. Running this program on the Trinity DN assembly successfully identified all the putative teretoxins we identified via BLAST, as well as additional candidate toxins not identified in BLAST results. These additional candidates include peptides of frameworks I, IX, VI/VII, VIII, XIII, XIV and XXII, which are Cys frameworks already present in the putative teretoxins identified via BLAST. Pepticomb results also included candidate teretoxins of framework XII (eight Cys), framework XVI (four Cys), framework XVII (eight Cys), framework XXIV (four Cys) and two unknown frameworks with 10 and 12 Cys, respectively. These latter candidate teretoxins with Cys frameworks not identified via BLAST sequence similarity search are shown in Table 2.
It is difficult without strong BLAST support for these additional peptides, especially given the inherent potential for misassembly or chimeric transcripts in de novo assembly, to present these as high confidence transcripts. As such these results are not included in the final count of 84 putative \textit{C. anilis} teretoxins peptides presented above, pending further verification by proteomic or other means. Nonetheless, Pepticomb is an important tool for exploring the frontiers not captured by BLAST, and is also extremely useful for organizing and validating putative toxin transcripts that are generated from BLAST results. It is also ripe for further development with the incorporation of other alternatives to BLAST techniques, such as the use of the profile Hidden Markov Model (PHMM) to aid in the identification and classification of terebrid gene superfamilies, as has recently been done for conotoxins.\textsuperscript{72,151}

\textbf{Table 2: Pepticomb candidate toxins.} These peptides were identified using pepticomb software to parse for signal sequence, pro region and cysteine framework. The algorithm is designed to identify toxins that may not be identified by BLAST-based homology searches.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Pro region</th>
<th>Mature peptide</th>
<th>Framework</th>
</tr>
</thead>
<tbody>
<tr>
<td>comp43602_c1_seq15_1</td>
<td>MKTFIIVLATLLL FPGLILG</td>
<td>HPISGERRTPYKQER</td>
<td>LRAVPGEEGGD C HTAYR C SDKA C KLORY C NEYY CC QESKKPALGLSTPEHT C EL C KSH</td>
</tr>
<tr>
<td>comp44060_c2_seq6_6</td>
<td>MSRMGMALLVV LLLLPAAPSQ EG</td>
<td>EPDRRAEPRMTATKR</td>
<td>SVNHLNIRGISDVLPE NC S C DEYPK C C IYY</td>
</tr>
<tr>
<td>comp22024_c0_seq1_1</td>
<td>MMRPRTLLLTL LVSSMTMDSDG</td>
<td>SMNEKRAAQG PSKR</td>
<td>C NND C DGRSDFKP CC DNSGK C LQFE CC LHGSTD C</td>
</tr>
<tr>
<td>comp30533_c0_seq2_1</td>
<td>MRLLLTVLTML TIMMMMTREVQ P</td>
<td>APAPLDSELHNGQSAVDRSLP RQKR</td>
<td>AVSWGMKILL C LTPLRN CC GGRWG C Q</td>
</tr>
<tr>
<td>comp42177_c0_seq1_5</td>
<td>MRRSLFVLLAVL AVLAVLLDSNIA</td>
<td>TDETLQR</td>
<td>ONFEPRDA C GSO C RR C SR CC QSTEKHGE C RSP C KR C ON C TGDADRG C</td>
</tr>
<tr>
<td>comp43790_c0_seq16_2</td>
<td>MKPGFLTFLSV LTASILIFDAET</td>
<td>ANHARLRLLP GRR</td>
<td>DD C DTA C NT C NSER C NAPPDP C ATA C AE C TQ C QWGSQGSPIEKV C PT C NT C DR C PGRTRST</td>
</tr>
</tbody>
</table>
2.2.7. Annotation of teretoxin-related transcripts to other venomous organisms

In the interest of exploring *C. anilis* venom duct transcriptome for genes outside of the scope of the short disulfide rich peptides categorized by Cys frameworks, a BLASTx search at an e-value cut off of 1 e-10 was performed using the *C. anilis* Trinity DN assembly against the Uniprot animal toxin database, a highly curated database that includes venomous compounds from a variety of toxin producing organisms\(^{152}\). This search yielded BLAST hits to peptide toxins from a number of different toxin families, primarily found in snake and spider venom (Fig. 12). This search yielded BLAST hits to peptide toxins from a number of different toxin families, primarily found in snake and spider venom.

The greatest number of transcripts identified relate to the latrotoxin family of peptide toxins, which are high molecular weight compounds implicated in the presynaptic release of neurotransmitters such as acetylcholine by inducing neurotransmitter exocytosis.\(^ {69,153}\) The latrotoxin family forms a major component of *Latrodectus* spider venom and comprises several types including alpha, beta, gamma, delta and epsilon-latroinsectotoxins. A recent analysis of the venom transcriptome of the black widow spider *Latrodectus hesperus* revealed the latrotoxin family to be highly diverse with a majority of the identified venom specific transcripts (39 in total) representing the latrotoxin family.\(^ {154}\)

A number of *C. anilis* teretoxin transcripts with homology to acetylcholinesterase were also identified from the comparative toxin BLAST. Acetylcholinesterase is a component of snake venom that rapidly degrades acetylcholine via hydrolysis. Acetylcholinesterase is present in the Bungarus genus of snakes in large quantity, although it is unclear what role it might play in toxicity to prey.\(^ {155,156}\) Other transcripts from the *C. anilis* transcriptome show homology to the toxin families of metalloprotease, phospholipase, prothrombin, ryncolin and CRiSP, components common to the venom of other organisms.\(^ {157,158}\) While it would require further analysis to understand the role that these transcripts play in terebrid venom duct, and whether they have a significant effect on prey capture, analysis of the de novo terebrid assembly led to the identification of potential toxic components other than the classical short peptides that are better understood for their specific role in inhibiting nervous system transmission. The latrotoxin type transcripts identified in our *C. anilis* assembly are potentially of the greatest interest, since they are the most prevalent and could have an important effect on pre-synaptic acetylcholine neurotransmission.
Figure 12. Comparative analysis of *C. anilis* venom duct transcriptome to other venom toxin-related genes. A. Highlights similar venom peptides toxins identified. B. Venomous organisms with toxin-related genes.

2.2.8. Terebrid venom duct transcriptome annotation using Gene Ontology and KEGG analyses

In the interest of gaining an overview of non-venom related proteins present in the terebrid venom duct transcriptome, gene ontology (GO) analyses were performed on *C. anilis* transcriptome by combining BLAST results from the Trinity DN assembly against NCBI NR, Uniprot SwissProt, and Uniprot Trembl databases. BLAST2GO analysis led to the annotation of 13,055 contigs, which were then visualized with
WEGO according to second level GO classifications of cellular component, molecular function and biological process (Fig. 13). Compared to previous analyses of this type performed on the venom duct transcriptomes of cone snails and other venomous organisms (e.g. snake and spider), we have annotated significantly more transcripts (13,055), undoubtedly due to the substantial depth of coverage provided by our Illumina sequencing data. The large number of annotated contigs with GO terms indicate a robust de novo assembly of terebrid venom duct transcriptome.

![Figure 13. BLAST2GO annotation of *C. anilis* venom duct transcriptome assembly. BLAST2GO level 2 analysis annotates contigs according to cellular component, molecular function and biological process. The high number of genes identified confirms the robust de novo assembly of the *C. anilis* venom duct transcriptome.](image)

A high level of assembly coverage based on illumine data is also illustrated by the mapping of metabolic pathways using KEGG annotation and visualization of the associated pathways with the software tool iPATH (Fig. 14). In the case of the *C. anilis* transcriptome a preponderance of metabolic pathways are outlined, which compares favorably with a similar analysis of a *C. geographus* cone snail venom duct based in 454 data, where far fewer pathways were identified.
Figure 14. Comparative iPath annotation of *C. anilis* and *C. geographus* metabolic pathways. Ipath is a web based tool that provides visualization of *C. anilis* metabolic pathways using a global pathways map constructed using approximately 120 KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. A. Visualization of metabolic pathways from *C. anilis* illumina assembly. B. Previous research effort presenting similar visualization of metabolic pathways based on 454 assembly of the *Conus geographus*, with annotation of fewer pathways (Hu et al. 2011).
Previous analyses of GO annotation have suggested that enrichment of proteins found within certain GO categories are related to specific functional aspects of the venom duct, in particular the potential high metabolic cost of toxin production. Such studies have suggested that the high representation of proteins implicated in processes such as binding, catalytic activity, transcription and translation, and protein secretion are specific to the venom duct. However, a comparison of second level GO annotation of the venom duct transcriptome to the tissue specific transcriptomes from other organisms, including non-venomous marine animals, shows a very similar distribution across GO categories. Irrespective of organism or tissue, binding and catalytic activity are always highly represented, and speculation that this is specific to the venom duct appears to be misplaced. It is important to note that distribution and number of hits among GO categories does not provide information about gene expression levels, and as such is limited in providing insight into which GO identified proteins may be present in greater abundance for venom production.

2.2.9. Venom Duct Proteins with Peptide Folding and Processing Function

Assignment of GO terms provides the means to investigate for the presence of proteins that are known to be important to folding and processing of peptide toxins, including peptide disulfide isomerases, cis-trans peptidyl prolyl isomerases, heat shock proteins, and proteins that play a role in post translational modification, which is of great importance in cone snails. Based on limited proteomic data, preliminary research suggested that teretoxins lack the heavy post translational modifications found in cone snails. However, preliminary investigation of the C.anilis transcriptome for the presence of post-translational enzymes similar to those found in the cone snail venom duct yielded a number of candidate proteins that may serve a similar function for teretoxins (Table 3).

Vitamin K dependent γ-glutamyl carboxylase is a post-translational enzyme found in the cone snail venom duct that catalyzes the addition of a carboxyl group to specific glutamate residues. One transcript from our Trinity assembly displays a near perfect match to Conus textile vitamin K-dependent γ-glutamyl carboxylase, with a BLAST hit of e-value of 0.0 and a bit score of 828. While carboxylation of glutamate residues is an established post-translational modification in conotoxins, it should be noted that this carboxylase is highly conserved in a wide variety of organisms, including human and drosophila. In
mammalian systems it plays an important regulatory role in the blood clotting cascade through the carboxylation of blood clotting proteins such as prothrombin.\textsuperscript{165}

**Table 3. Annotation of posttranslational enzymes and chaperones found in terebrid venom duct transcriptome.** All proteins listed here were identified from the *C. anilis* transcripts annotated by gene ontology and analyzed for the strength of the BLAST hit.

<table>
<thead>
<tr>
<th>Post-translational enzyme</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ-glutamyl carboxylase</td>
<td>Gamma-glutamyl carboxylase adds CO\textsubscript{2} to protein-bound glutamic acid to form gamma-carboxyglutamic acid</td>
</tr>
<tr>
<td>Peptidyl-α-amidating monooxygenase (PAM)</td>
<td>PAM modifies peptides containing a C-terminal glycine residue through cleavage of the glycine followed by amidation of the preceding residue</td>
</tr>
<tr>
<td>Prolyl 4-hydroxylase (PH4)</td>
<td>PH4 catalyzes hydroxylation of proline residue to yield (2S,4R)-4-hydroxyproline (Hyp)</td>
</tr>
<tr>
<td>Glutaminyl-peptide cyclotransferase</td>
<td>Glutaminyl-peptide cyclotransferase converts protein bound glutamine residues to cyclic pyroglutamyl residues</td>
</tr>
<tr>
<td>Tyrosyl sulfotransferase</td>
<td>Tyrosyl sulfotransferase catalyzes sulfation of tyrosine residues</td>
</tr>
<tr>
<td>Protein disulfide isomerase (PDI)</td>
<td>PDI catalyzes the formation and breakage of disulfide bonds between cysteine residues within proteins as they fold</td>
</tr>
<tr>
<td>Peptidyl-prolyl cis-trans isomerase (PPIase)</td>
<td>PPIase interconverts the cis and trans isomers of peptide bonds of the amino acid proline.</td>
</tr>
<tr>
<td>Heat Shock Proteins (HSP)</td>
<td>HSPs are class of functionally related proteins involved in the folding and unfolding of other proteins</td>
</tr>
</tbody>
</table>

Another important enzyme of interest in conotoxin post-translational modification is peptidyl-glycine α-amidating monooxygenase (PAM), which modifies peptides containing a C-terminal glycine residue through cleavage of the glycine followed by amidation of the preceding residue.\textsuperscript{166} Annotation of *C. anilis* assembly yielded a transcript with strong homology to peptidyl-glycine α-amidating monooxygenase A, found in the mollusk *Crassotrea gigas*, with a BLAST hit of 1 e-79, as well as a BLAST hit to the same protein found in *Conus bullatus*, with an e-value of 4 e-56.

Hydroxylation of conotoxin mature peptide prolines happens in as many as one out of two residues, presumably through the activity of prolyl 4-hydroxylase (PH4), although this specific enzyme
has yet to be characterized for any cone snail species. Preliminary studies have indicated that hydroxyprolinination may be important for both conopeptide bioactivity and oxidative folding. Proline hydroxylation is also widely present in humans as the result of PH4 activity, with diverse protein substrates, although the best known role is the stabilization of the collagen triple helix. Several PH4 transcripts are present in the proposed terebrid transcriptome assembly, with the most predominant being PH4-1, one of the important catalytic subunits of the larger protein. The PH4-1 transcript identified from our assembly has a BLAST hit to *Crassotrea gigas* PH4-1 with an e-value of 1 e-138.

Other identified *C. anilis* protein transcripts with potential post-translational function include, glutaminyl-peptide cyclotransferase and tyrosyl sulfotransferase, which are responsible for N-terminal cyclization of glutamine to pyroglutamate and sulfation of tyrosine in conotoxins, respectively. The terebrid glutaminyl-peptide cyclotransferase identified had a BLAST hit at e-value 4 e-86 to that of *Crassotrea gigas*, while the terebrid tyrosyl sulfotransferase transcript identified displayed excellent homology that of the sea snail *Littorina sitkana*, with an e-value of 0.0 and bit score of 635.

In addition to these post-translational enzymes, assembly of terebrid venom duct transcriptome also yielded a number of transcripts identified as having protein disulfide isomerase (PDI) activity. In particular, one *C. anilis* transcript shows a very strong match to a *Conus eberneus* PDI, with an e-value of 0.0 and a bit score of 708. Also identified was a transcript with strong homology (e-value 3 e-92) to a peptidylprolyl cis-trans isomerase of *Conus novaehollandiae*, which interestingly was used to facilitate the oxidative folding of selected conotoxins *in vitro*. A wide variety of heat shock proteins (HSP) are also present in the terebrid assembly, including HSPs 22, 40, 60, 70 and 90 and beta-11, with the lowest e-values found for BLAST hits to other mollusk proteins. The identification of PDI and other folding enzymes was expected given the disulfide rich nature of teretoxins.

While the identification of protein transcripts potentially implicated in post-translation modification of teretoxins is of great interest, it remains to be seen whether these modifications will be identified on the proteomic level. It is possible the posttranslational enzymes identified are involved in common housekeeping processes and not used to modify teretoxins in the venom duct. The recognition sequences for posttranslational modification enzymes within the conotoxin precursor structure have yet to
be defined, and it is not yet elucidated how, for example, some prolines are selected for hydroxylation while others remain unchanged. It has been proposed that glutamate carboxylation in conotoxins depends on a specific sequence found in the pro region of the peptide, but this is one of the few cases where the process is characterized.

2.3. Conclusion

This study applied next generation sequencing and state of the art bioinformatics tools to present the first comprehensive look at terebrid venom duct transcriptomes, with a particular focus on identification of disulfide rich peptides that can be used as lead compounds for drug development and as probes for investigating the structure and function of the nervous system. One hundred and forty nine putative teretoxins were identified from *C. anilis* and *T. subulata* combined (Table 1). Apart from the significant number of potential new teretoxins for investigating cellular processes, the highlights of comparative analyses of the assembled terebrid venom duct transcriptome presented include: 1. Identification of two novel Cys frameworks, Can_10Cys and Can_12Cys, with conserved pre signal sequences, but varying Cys frameworks and intra-Cys amino acid residues, a break from the conventional gene superfamily and Cys framework relationship found in conopeptides. 2. Identification of PXY motif, which, similar to the ICK motif in conotoxins, may have structural and function implications on teretoxin bioactivity, 3. Identification of the first teretoxin gene superfamily, which has sequence homology to the H superfamily of conotoxins, 4. The first characterization of enzymes that could be linked to posttranslational modification of teretoxins, pending validation by proteomic analyses of terebrid venom duct components, and 5. A paucity of Cys framework XI (I superfamily) transcripts, which directly contrasts cone snail transcriptomes in which I superfamily conotoxins are largely abundant and are among the best characterized as Na⁺ channel agonists and K⁺ channel modulators. Analyses of venom duct transcriptome implies that while terebrid toxins share organizational features with conotoxins, they differ substantially in terms of distribution of Cys frameworks, amino acid composition, and average length. The structural differences observed between teretoxins and conotoxins highly suggest teretoxins are novel ligands with molecular targets that may be distinct from conotoxins, making terebrid marine snails a good resource for identifying novel ligands that manipulate cell signaling in the nervous system.
De novo assembly of the transcriptomes of non-model organisms presents an ongoing challenge as a result of potential misassembly or the presence of chimeric transcripts. To ensure a high quality of the putative teretoxins identified, two different assembly programs, Trinity and Velvet Oases, were run at different parameters, and cross-correlated to validate identified teretoxin transcripts. Digital normalization techniques were applied to ease the computational load of assembling in excess of 200 million reads, while maintaining assembly coverage and quality. Cross validation, especially in cases where all four assemblies provide an exact match, provide a high confidence level in the integrity of these transcripts. At the same time, the capture of unique transcripts from different assemblies may represent rarer peptide toxins that are identified as a result of different kmer lengths or assembly algorithms. Subsequent validation on the proteomic level would be one potential avenue to establish further confidence in these sequences.

The expansive number of Illumina reads that formed the basis of terebrid venom duct transcriptome assembly allowed for a high depth of coverage that is reflected not only in the number of teretoxins captured, but also in the analysis of other components of the transcriptome, as reflected in identification of toxin-related genes, of transcripts that can be assigned gene ontology terms to form a global portrait of expression, and of transcripts implicated in metabolic pathways via the assignment of KEGG terms. The presence of different transcripts coding for enzymes involved in post-translational modifications, the same modifications found extensively in conotoxins, raises interesting questions regarding previous research suggesting that these modifications are low or absent in terebrid toxins. All posttranslational terebrid transcriptome transcripts can only be validated with certainty on the proteomic level, which will be part of the future directions of this research.

Finally, a working bioinformatics pipeline was established for subsequent transcriptome efforts that can be easily replicated for other species of terebrids, which will lay the foundation for further understanding of how teretoxins compare to those of cone snails and any forthcoming research of other conoidean toxins. This research also paves the way for the identification of gene superfamilies with the
identification of numerous precursor structures and the identification of conserved signal sequences that form the basis of the evolutionary relationships among teretoxins. The large numbers of peptides identified in this work as well as the numerous conotoxins being identified from transcriptome work point towards the need for the expansion of high throughput methods for both synthesis and bioactivity characterization to ensure that newly identified compounds with therapeutic potential can be screened rapidly and efficiently. Progress on this front will surely lead to the identification of compounds with novel function, given the vast universe of neuropeptides generated by the naturally occurring combinatorial chemistry practiced by species rich venomous organisms.

2.4. Materials and Methods

2.4.1. Sample Collection

The C. anilis specimens used in this study were collected on a 2011 expedition to Inhaca, an island off of the coast of Mozambique, where its populous presence suggests successful adaptation to this ecological niche. The expedition was a collaboration of the Holford group of The City University of New York and the American Museum of Natural History and the Bouchet group of the Muséum national d'Histoire naturelle (MNHN) of Paris. Materials from this expedition was used to obtain venom ducts and salivary glands for next generation sequencing projects and to enhance existing phylogenetic reconstructions of the Terebridae. A significant diversity of terebrid species were acquired in Inhaca using a combination of diving, dredging, vacuum accumulation, and intertidal collection techniques. As terebrids are nocturnal hunters, night dives yielded the most specimens. One particularly successful night dive yielded in excess of 100 specimens of C. anilis. Overall, 37 species of terebrids were collected during the 15 days expedition to Inhaca. At least eight of the collected Inhaca species are new to phylogenetic tree, including T. nitida, G. eddunhami, T. Sandrinae (Fig. 16). Tissue samples for barcoding and shell vouchers were fixed in 80% ethanol and sent to the Paris museum for DNA extraction and sequencing, while venom ducts and salivary glands were dissected from specimens and stored in RNALater and brought back to The American Museum of Natural History for transcriptome research.
2.4.2. RNA extraction and cDNA preparation

Total RNA was extracted from eight pooled *C. anilis* venom ducts with Qiagen RNeasy kit, with DNase digestion on column, according to manufacturer’s instructions. 10ng of *C. anilis* total RNA was used as template for Clontech’s SMARTer™ Ultra Low RNA Kit for Illumina Sequencing to perform 1st strand cDNA synthesis and 12 cycles of PCR amplification according to manufacturer’s instructions. *C. anilis* cDNA was assessed for quality and concentration with Agilent Bionalyzer, using their high sensitivity DNA chip.

2.4.3. Library preparation and Sequencing

The cDNA library was sequenced using Illumina HiSeq™ 2000 technology at the New York University Center for Genomics and Systems Biology. Library construction was performed with the Kapa Biosystems Kit for sequencing on two paths of an 8-lane Illumina flow cell. cDNA was fragmented using a Covaris S2 Sonicator (range 300-600 bp) and libraries were amplified 8 cycles. Library quality was analyzed on the Agilent Bioanalyzer using the DNA High Sensitivity Chip.

Figure 15. Representative terebrid species collected in Inhaca.
2.4.4. Read processing and de novo assembly

Illumina HiSeq generated 288,959,674 paired end reads of 100 bp length with 89.43% bases having a quality score ≥ Q30. Raw read quality assessment was performed with FastQC to determine the need for base trimming and adapter removal (http://www.bioinformatics.babraham.ac.uk/projects/.fastqc/). Seqtk was used to trim bases from the reads using a default Mott algorithm based on phred score (https://github.com/lh3/seqtk) while trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) was used to remove adapter contamination. Ultimately the quality control process yielded a set of paired end reads 66.5 GB in size for a total of 280,143,112 reads.

The de novo assembly programs Trinity (release date 2014-04-13) and Velvet Oases ((v. 1.2.10 and v. 0.2.08) were run using default parameters to assemble quality trimmed and normalized reads. Trinity assemblies were run on both Trinity in silico normalized reads, which reduced the number of reads to 9,204,374, and the full read set prior to normalization. For the Velvet Oases assemblies, two-pass digital normalization reduced the dataset to 23,607,286 paired end reads and 102,935 single end or “orphaned” reads. The Velvet Oases assemblies were constructed over a range of odd-numbered kmer values from 25 to 55.

2.4.5. BLAST annotation

Annotation of the C. anilis venom duct transcriptome was performed by using BLASTx to search for assembly contig hits against different databases, including (1) a combined database of conotoxins downloaded from Conoserver (www.conoserver.org) and our in house database of teretoxin sequences, (2) UniprotKB/Animal Toxin Annotation Program (release 2014_03), (3) NCBI non-redundant (nr) protein database, and (4) UniProtKB/Swiss-Prot (release 2014_03) and UniProtKB/TrEMBL (release 2014_03).

Assignment of gene ontology terms was performed in two parts: phase one of annotation followed a bioinformatics pipeline made freely available online by the Palumbi lab (http://sfg.stanford.edu/guide.html)169. After blasting against all three databases, a python script was run on significant BLAST hits in combination with a download of Uniprot flatfiles to extract gene names, general descriptions, and Gene Ontology (GO) categories, which are then combined in a master
annotation metatable. Running this pipeline on our Trinity DN assembly generated 18,390 annotated contigs with an evalue of 1e-6 or lower.

The second phase of GO analysis incorporated the use of BLAST2GO (B2G) tools on the SFG results for further mapping, annotation and in particular, the use of GoSlim, which employs a reduced set of GO terms to provide a broad overview of the gene ontology content without as much specific detail.
Chapter 3

Synthesis of novel teretoxins for functional assays
3.1. **Background: characterizing teretoxin bioactivity**

To date the functional activity of a purified teretoxin has yet to be described. A recent publication characterized crude venom extracts from several terebrid species on rat neuronal nAChRs (α3β2, α3β4, α4β2, α4β4, α7) and muscle subtypes (α1β1γδ), and K\(^+\) (Kv1.2 and Kv1.3) and Na\(^+\) channels (Nav1.2, 1.3, 1.4, 1.6) expressed in *Xenopus* oocytes.\(^{170}\) The crude extracts displayed inhibitory activity on nAChRs receptors, but had no activity on Na\(^+\) or K\(^+\) subtypes analyzed. While this result of terebrid crude venom analyses is promising, it is a very preliminary attempt at characterization and did not successfully identify what fraction in the crude venom was responsible for the inhibitory response in nAChR subtypes. To successfully identify teretoxins that are bioactive it is imperative to be able to synthesize large amounts of peptide toxins for characterization in functional assays. In this dissertation four teretoxins were synthesized for biological characterization (Table 1).

<p>| Table 1. Teretoxins synthesized during thesis work. |</p>
<table>
<thead>
<tr>
<th>Species</th>
<th>Peptide name</th>
<th>Peptide sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>T. guttata</em></td>
<td>Tg77</td>
<td>SLTELKAGGCPLYCSSQIFCCHGRKCRNVDGRLKCVTEASMLGK</td>
</tr>
<tr>
<td><em>C. anilis</em></td>
<td>Can6.5</td>
<td>ATKHPEEQYSDCPDNQCCIPSGKTVAFGVSHGVKICVTCL</td>
</tr>
<tr>
<td><em>C. anilis</em></td>
<td>Can10.1</td>
<td>QLCKCCAKLCHEEDRTIPCSGGNSQFECFKRTQKIRKDCNTQNEAKALCVGYFTGEC</td>
</tr>
<tr>
<td><em>C. anilis</em></td>
<td>Can14.1</td>
<td>GGLGRCHNCMNSRGGLNFIQCKTMCS</td>
</tr>
</tbody>
</table>

The four teretoxin peptides selected were identified from transcriptomes of *Terebra guttata* (Tg77) and *Cinguloterebra anils* (Can6.5, Can10.1 & Can14.1). The teretoxins chosen, Tg77, Can6.5, Can10.1, and Can14.1, were selected they as have features that are both similar and different from known conodiean peptide toxins. Tg77 is a 44 residue peptide identified from cDNA cloning library (Holford, unpublished data). Can6.5, a 40 residue peptide with six cysteines, Can10.1, 60 residues in length with 10 cysteines, and Can14.1, 27 residues with four cysteines, were identified in the next generation sequencing and de novo assembly work described in Chapter 2. Tg77 and Can6.5 both have a cysteine scaffold, C-C-CC-C-C, corresponding to framework VI/VII of the O gene superfamily. O superfamily conotoxins target Na\(^+\), K\(^+\), and Ca\(^{2+}\) channels, suggesting that Tg77 and Can6.5 may act on similar
molecular targets. It should be noted however, that apart from their cysteine scaffold Tg77 and Can6.5 have no strong sequence homology to known conotoxins and it is therefore hard to predict if they will have similar functional activity. Can10.1 has a novel framework not previously identified in conopeptides, with cysteine motif C-CC-C-C-C-C-C-C-C. At 60 amino acids, Can10.1 is significantly larger than most known conopeptides and is similar in size to scorpion or spider peptide toxins. The unique scaffold of Can10.1 suggests it may have functional activity different from previously characterized conotoxins.

Can.14.1 has framework XIV, C-C-C-C, with strong homology to mature conotoxins AsXIVA from Conus Austini, and vil14a from Conus villepini. Conotoxin AsXIVA was shown to elicit scratching and grooming activity in mice, and both peptides potentially target K+ channels, although the specific target has yet to be fully characterized for either peptide.\(^{171}\) The synthesis and bioactive characterization of teretoxins, Tg77, Can6.5, Can10.1, and Can14.1 would significantly enhance existing knowledge of teretoxins and provide the first analyses of specific teretoxin bioactivity.

Two synthesis strategies were employed to produce teretoxins Tg77, Can6.5, Can10.1, and Can14.1 in sufficient quantities for functional characterization: recombinant synthesis and solid phase peptide synthesis (SPPS). Due to their large size and cysteine scaffold complexity, Tg77, Can6.5, and Can10.1 were synthesized using recombinant expression. Can14.1, which is significantly shorter at 27 residues, was synthesized using SPPS. Given the structural characteristics of teretoxins, the main focus of the research presented here is on designing and executing a reproducible strategy for the recombinant expression of neuropeptides of at least 40 residues in length and having six or more cysteines. The synthesis and functional characterization of Tg77, Can6.5, Can10.1, and Can14.1 is described below.

3.1.1. Recombinant expression of teretoxins Tg77, Can6.5, and Can10.1

Teretoxins Tg77, Can6.5, and Can10.1 were all synthesized using recombinant expression, after having been identified on the genetic level from the full precursor structure that includes signal sequence, pro-region and mature peptide (Fig. 1). Due to the difficulties of collecting sufficient quantities of each species, and the limiting amounts of venom found in many specimens, recombinant expression has been widely used to produce peptide toxins in amounts that are not generally obtainable from natural sources.
For Cys rich peptides from a venom source, the synthesis method chosen must be cost and time effective,

**Figure 1. Full precursor structure of teretoxins chosen for recombinant expression.** Tg77 was previously identified from a *Terebra guttata* cDNA library and two *Cinguloterbra anilis* peptides, Can.6.5 and Can.10.1, were identified from next generation sequencing and de novo transcriptome assembly (see Chapter 2).

...avoid the common problem of peptide aggregation in the form of inclusion bodies, and facilitate the correct formation of disulfide bridges. A diversity of vectors, hosts, and cloning strategies available for recombinant expression in *E. coli* have been successfully manipulated to optimize venom peptide synthesis strategies of several conotoxins, and other peptide toxins from venomous organisms.

Table 2 summarizes several recombinant expression strategies used for the synthesis of conotoxins as well as peptide toxins from scorpion, spider and sea anemone. Some of the important considerations include the choice of a fusion tag, where the fusion protein is expressed (i.e. periplasmic space or cytosol), and type of cleavage site for release of the mature peptide. Expression as
a fusion protein has the advantage of increasing the amount of protein expressed and averting proteolytic degradation of the short peptide construct, but in some cases the final enzymatic cleavage step to release the mature peptide may prove problematic in terms of specificity of the reaction, as well as obstruction of the cleavage site by protein aggregation, resulting in low yield of final product. Interestingly, in several different research initiatives that include the use of thioredoxin tag, S-tag and GST-tag, leaving the fusion
protein intact has not impeded bioactivity of the toxin. Ideally, however, steps are taken to successfully troubleshoot cleavage problems and produce the desired peptide without the presence of any additional residues. Subsequent to vector construction there are numerous choices to be made in the overall research design, even though the overarching steps follow the same progression, proceeding through the sequential steps of transformation, induction and expression, protein purification, protease cleavage, and verification of the final product with HPLC and mass spectrometry.

As a first step, designing an effective recombinant expression protocol benefits from a cloning method that has a high rate of efficiency for correctly inserting the gene construct into the plasmid. Ligation independent cloning (LIC) has proven to be an adroit method for recombinant plasmid construction with a reported 70% efficiency rate.\textsuperscript{104,176} In contrast to conventional restriction enzyme cloning, LIC balances the exonuclease and polymerization activities of T4 DNA polymerase to create specific single-stranded 5’ tails on the gene insert complementary to the single-stranded overhangs in the target vector. Fragment and vector are combined in the absence of ligases, with circularization of the plasmid vector only occurring after annealing of the gene insert through its cohesive ends. LIC was applied to the recombinant synthesis of Tg77, Can6.5, and Can10.1.

3.2. Results and Discussion

3.2.1. Design and construction of expression system

Tg77, Can6.5, and Can 10.1 were initially expressed in \textit{E. coli} strains Nova Blue and Origami B (DE3) as pET-32 Xa/LIC plasmid fusion proteins with a thioredoxin tag to aid in solubility and folding, a His tag for purification purposes, and an enterokinase site for peptide cleavage (Fig. 2).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fusion_protein.png}
\caption{Fusion protein organization for soluble protein expression. All teretoxin neuropeptides were expressed as fusion protein construct under the control of a pET-32a T7 promoter, with a thioredoxin tag to aid in solubility and folding, a 6x histidine tag for purification and and enterokinase site for cleavage of the final product.}
\end{figure}
The pET-32 Xa/LIC plasmid, and *E. coli* strains Nova Blue and Origami B (DE3) (Novagen) are important components of an expression system tailored toward expression of cysteine rich peptides. The pET-32 Xa/LIC vector was chosen due to the presence of an N-terminal thioredoxin fusion adduct which aids in solubility and folding. The vector is pre-linearized with single overhangs in place, which allows for the efficient insertion of the gene, and carries an N-terminal 6X Histidine (His) tag for affinity purification. For cloning, the NovaBlue host provides high transformation efficiency, good plasmid yield and can also be used to produce single stranded plasmid DNA for mutagenesis. The expression host Origami B contains a chromosomal copy of the T7 gene for induction of expression, is deficient in ompT proteases to prevent protein degradation, and has a trxB/gor double mutation that aids in disulfide bond formation.177,178

The primary amino acid sequence of each teretoxin was back translated into codons optimized for expression in *E. coli*, and a synthetic oligonucleotide was prepared for insertion into the plasmid using LIC. In the case of Tg77, primers were designed for PCR amplification of the gene insert to add the specific sequence required for the generation of the long single stranded ends that are the hallmark of the LIC process. For Can6.5, and Can 10.1, the PCR reaction was dispensed with by ordering the synthetic oligonucleotides as a double stranded “gene blocks” with the additional LIC sequence incorporated at the 5’ and 3’ end of both strands. In frame insertion of the synthetic oligonucleotide into the pET 32 plasmid was verified by colony PCR and Sanger sequencing for all three teretoxin plasmids (Fig. 3).

![Figure 3](image.png)

**Figure 3. Sample colony PCR showing correct gene insertion.** In frame insertion of Can6.5 peptide into pET 32+ LIC plasmid with verification of His Tag, EK site, and peptide sequence followed by stop codon.
Purified plasmid was then transformed into the expression host, and colonies were screened for positive transformants to be used for inoculation of culture and subsequent expression.

3.2.2. Expression and purification of recombinant fusion proteins

Tg77, Can6.5, and Can10.1 were transformed expression host E. coli bacteria and grown in terrific broth culture. IPTG was added during log phase growth at a concentration of 0.4 mM for induction of fusion protein expression. Induction and growth parameters were varied to find the optimal conditions for overexpression of soluble protein, including IPTG concentration, time and temperature of growth, and lysis methods. Expression of Tg77 fusion protein was determined to be the highest with terrific broth culture medium, with induction of expression at close to 1.0 OD and 0.4mm IPTG, and overnight growth at 25 degrees Celsius. After overnight induction, lysis by sonication and centrifugation of bacterial culture, Tg77 was His tag purified with Nickel-NTA affinity chromatography using a batch purification method. Expression of Tg77 in the Origami host gave an average yield of 20.9 mg per Liter of growth medium. Expression and purification of Tg77 were analyzed with 12% SDS-PAGE, confirming overexpression of soluble fusion protein at molecular weight of 23.8 kD (Fig. 4).

Figure 4. SDS-PAGE analysis of expression and purification of Tg77. The lanes of the gel are represented as follows: M = protein molecular mass marker. Lane 1 = Tg77 crude lysate. Lane 2 = Tg77 crude lysate supernatant post binding to nickel-NTA resin. Lane 3 = wash buffer 1 supernatant. Lane 4 = wash buffer 2 supernatant. Lane 5 = combined imidazole eluted fractions of Tg77 fusion protein.

In the interest of generating product more efficiently and in larger amounts, the recombinant expression protocol was modified for teretoxins Can6.5, and Can10.1 by using BL21(DE3)
as the expression host, and performing His-tag purification using Fast Protein Liquid Chromatography (FPLC). BL21(DE3) is used routinely for recombinant expression of disulfide rich conotoxins (Table 2). The strain BL21(DE3) is known for relatively high expression levels, and the presence of a thioredoxin tag in the fusion construct may be sufficient for the objective of disulfide bond formation. A recent study comparing recombinant expression of disulfide rich peptides using a variety of different fusion tag proteins in combination with three *E. coli* strains, including OrigamiB and BL21(DE3), demonstrated that BL21 more often allowed for the expression of soluble protein. With respect to His-tag affinity purification, the FPLC method can be readily scaled to larger amounts of protein, and in industrial settings is used to generate kilogram amounts of purified protein.

FPLC chromatograms for His-tag affinity purification of terotoxins Can.6.5 and Can10.1 show different elution profiles. The elution profile of Can.6.5 depicts two distinct, sharp peaks at 170 mM and 250 mM of imidazole, respectively, while Can.10.1 displays a single broad peak at 170 mM (Fig. 5). Can6.5 and Can10.1 vary in size, 40 vs. 60 residues, and complexity, which is illustrated by the different elution profiles. The two peaks of Can6.5 and the single peak of Can10.1 elute in a comparable range of 150 to 280 mM imidazole, and 150 to 260 mM imidazole, respectively. It is unclear why Can.6.5 elutes in two distinct peaks, whereas Can.10.1 elutes in a single broad peak. These chromatographic results may be explained by variable elution times of monomers versus multimers of the fusion protein, and the implications for the final step of protease cleavage. Recently, in the FPLC His-tag purification of spider toxin Huwentoxin-IV, Sermadiras et al. observed that monomers eluted more readily (i.e. earlier) than multimers with intermolecular disulfide bridges that are present as aggregates.
Figure 5. Elution profiles FPLC His affinity purification of peptides Can.6.5 and Can.10.1. A. Can.6.5 elution shows two distinct peaks of variable size over an imidazole gradient of 150 to 250 mM imidazole. B. Can.10.1 elutes in a single broad peak over a similar gradient, 150 to 260 mM imidazole.
Sermadiras, et al. report that optimization of the imidazole elution gradient increased the elution of monomers only. Without further analysis varying imidazole concentration to manipulate elution profiles of Can.6.5 and Can.10.1 it is difficult to know with certainty if a similar pattern holds, however, Can10.1 has 10 cysteines in the peptide, which would make the formation of multimers more likely due to misfolding or intercysteine bonding.

To fully characterize the presence of correctly folded monomers as opposed to multimeric aggregates, a first step would be to analyze the His-tag purified fusion protein under native PAGE conditions, and/or with size exclusion chromatography. These approaches, while not applied to Can6.5 and Can10.1, would enable the physical separation of monomers from multimers should both be present, and allow for the optimization of imidazole based elution of monomeric fusion protein.

SDS-PAGE analysis of Can6.5 and Can10.1 fusion protein expression and purification is illustrated in Figure 6. Yields in the bacterial host BL21(DE3) are approximately double the amount in BL21(DE3) compared to expression of Tg77 in the Origami bacterial host. Induction and expression under the same time, temperature and IPTG concentration used for Origami expression led to a yield of 23 mg of His-tag purified Can6.1 fusion protein, and 20 mg of Can10.1 fusion protein per 500 ml of growth medium.

![Figure 6. SDS-PAGE analysis of expression and purification of terotoxins Can.6.5 and Can.10.1. The gel lanes are labeled as follows: A. M = protein molecular mass marker. Lane 1 = Can.6.5 crude lysate. Lane 2 = Can.10.1 crude lysate. B. M = protein molecular mass marker. Lane 1 = His affinity purified Can.6.5. Lane 2 = His affinity purified Can.10.1. The expected protein fusion mass for Can6.5 and Can 10.1 is 22.2 and 24.7 KDa respectively.](image-url)
Enterokinase cleavage to release Tg77 mature peptide from fusion protein construct was optimized by varying the amount of protein substrate relative to enzyme over a range of specified values. A ratio of 1 unit rEK enzyme to 50 ug of protein generated the best cleavage results (Fig. 7). However, enterokinase cleavage of the Tg77 mature peptide from the fusion construct produced a low yield of the final product even with optimization of rEK usage. Enterokinase cleavage may produce poor results due to lack of specificity and problems with cleavage site accessibility. Notably, yield of the mature peptide after EK cleavage appeared to have an inverse relationship with increasing levels of expression of fusion protein after optimization of expression, suggesting that access to the cleavage site was being inhibited by problems of protein aggregation. In order to address this obstacle, varying concentrations of urea, ranging from 1M to 4M, were added to the cleavage reaction to facilitate enterokinase access to the cleavage site. The partially denaturing effects of urea led to a substantial improvement in yield of the cleavage reaction, resulting in consistent yields 2.5 mg/ml of Tg77 per L of growth media.

**Figure 7. Tris-Tricine SDS-PAGE optimization analysis of Tg77 EK cleavage.** Varying concentrations of rEK were applied to Tg77 fusion protein in 1:10, 1:20, and 1:50 dilutions. The gel lanes are labeled as follows: M = protein molecular mass marker. Lanes 1-3 = EK cleavage with rEK to Tg 77 fusion protein ration with no protease inhibitor. Lanes 4-6 = EK cleavage with rEK to Tg 77 fusion protein ration with protease inhibitor.
Fusion cleaved Tg77 teretoxin was purified by semi-preparative Reverse Phase High Performance Liquid Chromatography (RP-HPLC) and the mass was confirmed using Liquid Chromatography Mass Spectrometry (LC-MS) (Fig. 8).

Figure 8. Semi-preparative RP-HPLC and LC-MS of Tg77 teretoxin. A. Semi-preparative HPLC purification of folded Tg77 with peak elution at 22.45 minutes. B. LC-MS characterization folded Tg77 (expected mass = 4756.2 Da, observed mass = 4756.2 Da). The +4, +5, +6, and +7 ion charge states are shown.
LC-MS analysis confirmed the mass of folded Tg77 peptide, and the HPLC and LC-MS analysis of fusion cleaved teretoxin Can6.5 also yielded the 4360.9 Da expected mass of a folded Can.6.5 peptide (Fig. 9).

Figure 9. Analytical RP-HPLC and LC-MS of Can.6.5. HPLC purification of folded Can.6.5 on an analytical scale with peak elution at 14.65 minutes, followed by confirmation with Q-TOF mass spectrometry of the expected molecular weight of 4360.91423 Da, with an observed mass of 4360.918.

The overall yield of fusion cleaved Can6.5 is low, and the addition of urea as a partial denaturant did not have the same remedial effect as seen in Tg77. As noted previously, increasing expression levels of the Can6.5 fusion protein may have an inverse relationship with the efficiency of EK cleavage and it may be beneficial to investigate alternative enzymatic cleavage methods such as Factor Xa, which is incorporated in the pET-32 Xa/LIC plasmid construct, to improve enzymatic cleavage specificity and Can6.5 yield.
Factor Xa was originally not applied as enzymatic cleavage with this protease would leave additional nonnative residues to the mature Can6.5 peptide sequence, which may affect functional characterization of the peptide.

HPLC and LCMS characterization of fusion cleaved Can10.1 did not yield the expected product (data not shown.) As illustrated in Fig 6, while it was possible to successfully express fusion Can10.1 protein, subsequent EK cleavage did not produce the mature Can10.1 peptide. Complications of Can10.1 EK cleavage may be due to the formation of Can10.1 multimers as suggested by FPLC purification (Fig. 5).

3.2.3. Chemical synthesis of Can.14.1 teretoxin

*Cinguloterebra anilis* teretoxin Can.14.1 was identified from de novo transcriptome assembly and is homologous to *Conus austini* conotoxin AsXIVA (Fig 10). Chemical synthesis of Can.14.1 teretoxin was accomplished using base labile Fmoc solid phase peptide synthesis (SPPS).

![Figure 10. Can.14.1 teretoxin alignment to Conus austini AsXIVA peptide.](image)

Shown is the full precursor structure of Can.14.1 showing signal peptide and precursor structure aligned to AsXIVA mature peptide. Blue boxes indicate sites of amino acid difference between the mature peptides AsXIVA and Can.14.1.

Can14.1 synthesis was initiated by attaching the first Can.14.1 carboxy-terminal amino acid serine (Ser) to a Wang-chloride polystyrene resin, which resulted in a load volume of 0.4 mm/g (Fig. 11).
After the initial loading of Ser, chain elongation was carried out on a CEM Liberty Microwave Peptide Synthesizer for the production of linear Can14.1 peptide. The linear Can14.1 teretoxin was then cleaved from the Wang resin using a Reagent K cocktail composed of scavengers to prevent unwanted reaction of the sidechain protecting groups being removed (Fig. 12).

After the initial loading of Ser, chain elongation was carried out on a CEM Liberty Microwave Peptide Synthesizer for the production of linear Can14.1 peptide. The linear Can14.1 teretoxin was then cleaved from the Wang resin using a Reagent K cocktail composed of scavengers to prevent unwanted reaction of the sidechain protecting groups being removed (Fig. 12).

Figure 11. SPPS attachment of Can14.1 First Amino Acid (Serine) to Wang resin. Fmoc protected Serine amino acid (Fmoc-Ser(tBU)-OH) was reacted with Wang-chloride resin at room temperature in preparation for Can14.1 peptide chain elongation on CEM automated microwave synthesizer. The remaining residues of Can14.1 are added in stepwise manner as described in Chapter 1 until the mature Can14.1 peptide is synthesized.

After the initial loading of Ser, chain elongation was carried out on a CEM Liberty Microwave Peptide Synthesizer for the production of linear Can14.1 peptide. The linear Can14.1 teretoxin was then cleaved from the Wang resin using a Reagent K cocktail composed of scavengers to prevent unwanted reaction of the sidechain protecting groups being removed (Fig. 12).

Figure 12. Teretoxin Can.14.1 cleavage from Wang-cl resin. The assembled Can.14.1 peptide chain is cleaved from the resin by treatment of Reagent K, which contains trifluoroacetic acid (TFA), ethanediethiol (EDT), Triisopropylsilane (TIS), and water. The resin was incubated in Reagent K for 4 hours at RT and the peptide was recovered by ether precipitation at 4°C.
Despite the numerous Cys residues present, folding of conoidean peptides is readily achieved through thiol assisted air oxidation. Linear Can14.1 was oxidatively folded in solution with reduced and oxidized forms of glutathione (GSH/GSS), which is used to initiate disulfide bond formation of the four cysteine residues in the peptide. Can14.1 folding reaction was monitored via Ultra High Performance Liquid Chromatography (UHPLC) over 0 to 24 hours at room temperature. Folding of the peptide proceeded rapidly, with the folded species present within one half hour. The transition from linear Can14.1 to folded Can14.1 is illustrated in Figure 13. As peptides fold they become more hydrophilic as hydrophobic residues are buried in the folded conformation, resulting in a shift to the left in UHPLC traces. Folded Can14.1 elutes as a single peak at 1.7 minutes in more polar gradient, compared to linear Can14.1 which elutes at 2.4 minutes, in a less polar gradient (Fig. 13).

Figure 13. RP-uHPLC elution time profiles for oxidative folding of linear Can.10.1. Overlay of oxidative folding reaction of Can.14.1 at 0 minutes folding (2.4 min elution), at 15 minutes folding (1.8 min elution) and at 30 minutes folding (1.7 min elution). Based on these results subsequent folding reactions were carried out at 37 deg C and for a duration of 30 minutes.
Folded Can14.1 was purified by semi-preparative RP-HPLC and the mass was confirmed by LC-MS (Fig. 14).

**3.3. Functional characterization of teretoxins in polychaete worm assay**

Teretoxins Tg77, Can 6.5, Can 10.1, and Can14.1 are novel venom peptides, and as such their molecular targets are unknown. As described in Chapter 1, conoidean peptides are active on a plethora of ion channels and receptors in the nervous system. Identifying the specific molecular target of a novel teretoxin is similar to finding a needle in a haystack as there are a myriad of potential molecular targets. As a preliminary screen for teretoxin bioactivity, a global phenotypic assay using *Nereis virens*, *N. virens*, polychaete worms was applied. Polychaete worms are the natural prey of terebrid marine snails and...
assays of this type have been previously performed to analyze conoidean toxin effect in *N. virens* and the nematode *C. elegans*.\(^{47,180}\)

*N. virens* is a typical free moving or “errant polychaete” (L. *errare*, to wander) that lives in shallow marine waters and is widely dispersed in the North Atlantic region.\(^{181}\) They average 20 to 40 cm in length, with highly segmented bodies and parapodia that function in both locomotion and gas exchange. The nervous system consists of a brain, which is found in the head region and is connected to the ventral nerve cord and the sub-pharyngeal ganglia, while the circulatory system is composed of two main blood vessels, with smaller vessels to supply the parapodia and the gut (Fig. 15).\(^{182}\) Injections of teretoxin and control peptides are targeted to the double and ventral nerve cord.

**Figure 15. Generalized scheme of polychaete anatomy.** Body cross section shows location of double and ventral nerve cords. Injection of teretoxin and control peptides is targeted to this area (Image adapted from http://lanwebs.lander.edu/faculty/rsfox/invertebrates/nereis.html).

*N. virens* can engage in either crawling or swimming. In swimming they throw their bodies into sinusoidal S-shaped curves generated by the contraction and relaxation of the longitudinal muscles in the segmented body divisions. In contrast to most marine organisms, the swimming forward motion results from “reversed” posterior to anterior waves that nonetheless propel the worm forward when combined with the turbulence created by the parapodia (Fig 16).\(^{183}\) Observation of *N. virens* swimming pattern upon injection of teretoxin and control peptides provides a phenotypic response that can be related to teretoxin bioactivity.
Purified Tg77 was assayed five times in *N. virens* to determine its bioactivity in terms of phenotypic disruption of polychaete worm movement (Table 3). Worms were injected 12 to 20 µl of 21µM Tg77, alongside positive and negative controls that included injection with 1µM omega agatoxin, injection with normal saline solution, respectively, and observation of a worm that was not injected. Test worms are placed in 4 identical saltwater baths at room temperature. Directly after injection, the worm response is videotaped over an approximate 3-4 hour period. After one hour, a cold stimulus is applied by placing worms in a 4 deg C water bath to determine recovery time.

Injection of Tg77 results in an initial partial or complete paralysis of *N. virens*, followed by uncoordinated swimming movement, twisting, and rigid extension of parapodia. This effect is observable within 15 to 30 seconds post injection. After a short period of uncoordinated movement, the worm assumes a state of complete immobility. Each assay was video recorded to monitor the worms over a 3-4 hour period and was reviewed both by visual observation and digitally tracked with the computer software WormLab. A video still of the worm assay is provided in Figure 18.

*Figure 16. Polychaete* *N. virens* *swimming motion.* During swimming contraction of longitudinal muscles in inner body segments and relaxation of out body segments throw their bodies into sinusoidal S-shaped curves. The parapodia paddle backward exerting thrust, during their power strokes. The main thrust during swimming, however, comes from the force exerted by the undulating body on the water (Image adapted from http://cronodon.com/BioTech/Polychaete.html).
A preliminary analysis using WormLab to analyze the moving average speed of worms injected with Tg77, as well as positive and negative controls, is presented in Figure 17. The positive and negative numbers along the Y-axis represent the worm movement in opposite directions. Although N. virens has a generally sinusoidal movement, it is not a particularly good swimmer leading to irregular movement and alternation with crawling. As a result, WormLab was not always tracking the motion of the worm effectively, as its tracking algorithm is based on a geometric analysis that looks for more regular sinusoidal swimming motion found in C. elegans. When N. virens motion deviates from a predictable...

<table>
<thead>
<tr>
<th>Table 3. Testing of Tg77 in N. virens for phenotypic response</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Polychaete body mass (g)</strong></td>
</tr>
<tr>
<td>Assay 1</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Assay 2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Assay 3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Assay 4</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Assay 5</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
wave pattern, WormLab cannot track the worm and has to be manually manipulated to recapture the tracking. Thus, while this graph represents a first step towards achieving a true quantitative assessment of changes to worm motion upon injection of teretoxin, it is not yet able to capture with fidelity the overall phenotypic response from the peptide. Efforts are ongoing to rectify the tracking algorithm so that a comprehensive picture of worm movement over a period of time can be obtained. Visual assessment of the teretoxin effect is currently the best measure of behavioral change in looking for signs of uncoordinated movement or paralysis, which is clearly observable in the case of Tg77 as seen in the video footage within a few minutes of injection.

Figure 17. Preliminary results of polychaete moving average speed using WormLab. Video software footage over a 40 minute period was used to analyze moving average speed of *N. virens* after injection with Tg77 and controls. Video is taped at 30 frames/sec and moving average speed is analyzed in microns/sec with positive and negative values reflecting opposite bending angles. WormLab tracking coverage may be dropped temporarily if the worm movement deviates too much from regular sinusoidal movement, resulting of varying coverage of individual worms.
Based on repeated phenotypic response of paralysis observed in *N. virens* upon injection of Tg77, Tg77 was submitted to collaborator Dr. Ibanez-Talon for further characterization to determine the specific molecular target using electrophysiology. Thus far, it has not been possible to identify a specific target for Tg77. Additional functional assays are being pursued to continue this effort. Similar to Tg77, Can6.5 and Can14.1 will be provided to collaborators for identification of their molecular target, based on the results of initial phenotypic assays.

### 3.4. Conclusion

One of the main goals of this dissertation project has been to design and execute a reproducible strategy for the synthesis of cysteine rich teretoxins in good yield for downstream functional assays. To this end, four novel teretoxins, Tg77, Can6.5, Can10.1, and Can14.1 were synthesized using recombinant expression and chemical synthesis methods.

The recombinant synthesis approach employed LIC as a high efficiency technique for reliable insertion of the gene of interest into the plasmid, avoiding the common problems inherent to a restriction enzyme approach. To optimize the yield of teretoxins being expressed, several modifications in bacterial
expression strains and enzymatic cleavage conditions were conducted. Specifically, while expression of Tg77 in OrigamiB plasmid and RP-HPLC was effective, expression of Can10.1 and Can 14.1 teretoxins BL21(DE3) followed by FPLC His-tag affinity purification doubled the expression of Can 6.5 and Can 10.1 fusion proteins compared to Tg77. Yield of enterokinase cleavage was greatly enhanced for Tg77 by the addition of 2-3M urea to the cleavage reaction, but did not have the same positive effect for cleavage of Can.6.5 and Can.10.1 mature peptides. Cleavage has been partially successful for Can.6.5 and an accurate mass for the folded confirmation has been obtained. Enzymatic cleavage and synthesis of Can.10.1 was inconclusive, which is not entirely surprising considering the high number of cysteines and the increased potential for aggregation and misfolding.

There are a number of potential avenues to pursue in order to increase the yield of the cleavage reaction. A first step would be to evaluate more closely problems of aggregation and misfolding even in species that are isolated from the soluble expression fraction, as indicated by the presence of monomeric versus multimeric species. Very interesting questions are raised as to when the formation of inclusion bodies results from aggregation during the expression process itself, and when multimeric or misfolded species may still be present as soluble forms but as a result prove resistant to successful protease cleavage. Given the weak conditions for protein folding in *E. coli*, even in the presence of engineered strains, it has been proposed that folding may actually occur ex-vivo in lysis and purification steps where the presence of a favorable pH (typically close to 8.0) and high concentrations of molecular oxygen provide favorable folding conditions. If this is in fact the case, these steps could be optimized for correct folding by the addition of oxido shuffling reagents to the buffers, or possibly ab initio denaturation and refolding of the peptide in the fusion form prior to cleavage. On a different front, if lack of specificity is leading to low cleavage, then altering proteases, such as using FactorXa or Tev may prove more effective. Although it would be worth exploring both alternatives, it appears more likely that aggregation is the biggest hurdle, given the greater difficulty of cleavage in the ten Cys Can.10.1

In this body of research the chemical method of SPPS has also been explored as a means to synthesize teretoxin peptides and was successfully applied to the synthesis of Can.14.1, a 27 residue peptide with four cysteines. SPPS has proven to be very effective the production of this short peptide,
with at least 8 milligrams of folded peptide prepared for bioactivity assays. However, it still remains the case that SPPS becomes more difficult and cost prohibitive with increased peptide length, which makes recombinant expression an attractive alternative in many cases.

Functional characterization of synthesized teretoxins was performed using a phenotypic polychaete *N. virens* worm assay. Characterization of Tg77 was performed chiefly by visual observation of the videotaped footage. Compared to saline injected worms, Tg77 injection caused partial or complete paralysis. Injection of Tg77 generated uncoordinated movement and paralysis in repeated trials, similar to the effect of the positive control omega-agatoxin IVA, which acts as a P-type calcium channel antagonist. These preliminary findings are promising, but subsequent attempts to identify a specific molecular target for Tg77 were inconclusive.

In summary, with the successful synthesis of Tg77, Can6.5, Can10.1 and Can14.1, significant strides have been made towards the recombinant and chemical synthesis of teretoxin during this research period. Characterization of the molecular targets of these teretoxins is an ongoing effort in the Holford laboratory.
3.5. Materials and methods

3.5.1. Recombinant plasmid construction

Synthetic oligonucleotide gene inserts for all three peptides were obtained by back translation of respective peptide amino acid sequence and the addition of an N terminal enterokinase cleavage site, according to standard E. coli codon usage. PCR primers were designed to add the requisite LIC oligonucleotide sequence to the 5’ and 3’ ends of the gene specific insert for the subsequent generation of long sticky ends and directional cloning:

5’-ATATTATATTATTAGATGACGACGACAAGAGCCTGACCGAACTGAAAGCGGGCGGCGGCTCGCCGTATTGCAGCAGCCAGATTTTTTGCTGATGGCCGCAAATGCCGAACGATGGC

CGCCTGAAATGCGTGACCGAAGCGAGCATGCTGGGCAAA

TAAATATTATTA-3’

5’-GGTATTGAGGGTGC-3’ Forward LIC extension

GATGACGACGACAAG Enterokinase site

5’-AGAGGAGAGTTAGAGCCA-3’ Reverse LIC extension

The commercially available pET-32 Xa/LIC Vector Kit (EMD Millipore) supplied the plasmid vector as ready-to-use linear double-stranded DNA with single stranded overhangs for annealing to target DNA. PCR amplified oligonucleotide was purified with SpinPrep Gel DNA Kit (Novagen), treated with T4 DNA Polymerase in the presence of dGTP to generate vector-compatible overhangs, and annealed to the pre-linearized vector according to the manufacturer’s protocol. The plasmid construct was initially transformed into the E. Coli NovaBlue competent cells and screened for successful transformation with ampicillin/carbenicillin and kanamycin resistance markers. Verification of target gene insertion in the LIC vector in the correct reading frame was verified by colony PCR, according to manufacturer’s protocol.

The two C. anilis peptides, Can.6.5 and Can.10.1 were handled in the same manner as the above, with one exception. Rather than PCR amplification of the synthetic oligonucleotide, double stranded DNA
gene blocks (IDT) were designed with LIC extensions pre-incorporated and ready for immediate treatment to generate single stranded ends for annealing to vector. Treatment to generate single stranded overhangs followed the same protocol of treatment with T4 DNA polymerase in the presence of dGTP, and correct insertion of target gene was also verified with colony PCR. Oligonucleotide synthesis of double stranded gene blocks were based on the following oligonucleotide sequences:

**Can6.5**

5'GGTATTGAGGGTGCAGATGACGACGACAAAGGCAGCCAAACATCCGGAATGCGAACAGT
 ATAGCGATTGCCCGGATAACCCAGTCGATCCTCCGAGCGCAGGAAACCCTGCGCTTTGCGCT
 AGCCATGCGTGAATAATTGTGCTGACCTGCTGTGAATCATTGCTCTAATCTCTCTCTCT
3'

**Can10.1**

5'GGTATTGAGGGTGCAGATGACGACGACAAAGGCAGCTGTGCAAATGCTGCGCGAAACTGTG
 CCATGAAGATCGCGATCGCACCATTCCGTCAGCGCCGCGGCAACAGCCAGCTGTGCTCTCTCTCT
 GCAAACGCACAAACAGAAAATTCGCAAAGATTGCAACACCCAGAACGAGCGCAAAGCGC
 TGTGCGTGGGCTATTGACCGCAATGCTTAATGGCTCTAATCTCTCTCTCTCTCTCT
3'

### 3.5.2. Induction and Expression

The Trx-His-neuropeptide construct was plasmid purified with QIAprep Spin Miniprep Kit (Qiagen) and transformed into the *E. coli* Origami strain. A single colony from a fresh plate was used to inoculate a primary culture, which was grown overnight at 37°C and 250 RPM to saturation. The following morning, a larger culture was inoculated using the overnight pre-culture and incubated at 37°C and 250 RPM until optical density (OD) measured at 600nm reached a value of 0.8 -1.0. IPTG was then added to a final concentration of 0.4mM to induce expression of the Trx-His-peptide fusion protein. The culture was grown overnight at 25°C and 250 RPM for overexpression of soluble protein. The bacteria were harvested by centrifugation (8,000g, 10 minutes, 4°C) and the pellet was stored at -20°C until his tag purification.
3.5.3. Protein extraction and His Tag Affinity Purification

The bacterial pellet was resuspended in sodium phosphate buffer (100mM Na-PO4, pH 8.0, 300 mM NaCl, 10% glycerol) and lysed by sonication with Fisher Scientific Model 120 Sonic Dismembrator at three rounds of 70% power for 30s and three rounds at 90% power for 30s for soluble protein extraction. Cleared lysate was generated by centrifugation at 13,000g for 45 min and filtered (0.2 micron).

Two methods were pursued for 6x His tag purification. Tg77 was treated with a batch purification method using Nickel-NTA resin (Qiagen) pre-equilibrated with lysis buffer (100 mM Na-PO4, 300mM NaCl, 10% glycerol, 1 mg/ml lysozyme, pH 8.0). Binding of cleared fusion protein lysate to nickel-NTA resin was followed by treatment with two wash buffers [wash buffer 1 (50 mM Na-PO4 pH 7.7, 300 mM NaCl, 10mM imidazole, 10% glycerol) and wash buffer 2 (50mM NaPO4 pH 7.7, 2M NaCl, 10mm imidazole, 10% glycerol)]. His tagged protein was eluted sequentially with two elution buffers [elution buffer 1 (50mM Na-PO4 pH 7.7, 300mM NaCl, 20mM imidazole) and elution buffer 2 (50 mM Na-PO4 pH 7.7, 300 mM NaCl, 500 mM imidazole, 10% glycerol)].

C. anilis fusion proteins were purified with FPLC using a 5ml HisTrap HP column (GE Healthcare) charged with NiSO4 on an ÄKTA Purifier system (GE Healthcare). Two buffers (Buffer A: 50mM Tris/HCl pH 8.5, 10 mM imidazole, 500mM NaCl and Buffer B: 50mM Tris/HCl pH 8.5, 1M Imidazole, 500M NaCl) were used to generate the imidazole gradient. The flow rate was maintained at 5ml/min. The column was equilibrated for 5 column volumes (CV) with buffer A. The lysate was then loaded (typically 150ml), followed by at 15 column volume wash step. Elution was performed using an imidazole gradient from 0 to 50% Buffer B (500 mM imidazole) followed by an increase to 100% Buffer (1 M imidazole to clear the column. Elution fractions of 2ml each were collected and analyzed by SDS PAGE for the presence of His purified fusion protein.

3.5.4. Enterokinase cleavage

Enterokinase protease (EMD Millipore and Syd Labs) was used to cleave peptide from the fusion protein at the recognition site DDDK according to manufacturer’s protocol. Optimization of reaction conditions was conducted at enzyme-substrate ratios of 1:10, 1:20 AND 1:50, and over a time range of 5 hours to overnight. After optimization, the reaction was run at enzyme to substrate ratio of 1:50 and overnight in
duration to ensure full cleavage. Cleavage yield was enhanced by adding urea to the enterokinase reaction, from 1-4 M concentrations, in order to generate partially denaturing reaction conditions.

3.5.5. **HPLC purification and Mass Spectrometry**

Cleaved Tg77 was purified by RP-HPLC using an X-Bridge C18 semi-preparative column (10 x150 mm, 5 um particle size, Waters Corporation, Milford, MA, USA) pre-equilibrated with 95% buffer A (0.1% TFA)). Elution was carried out at 5 mL/min over a linear gradient of buffer B (80% acetonitrile 0.1% TFA) from 5% to 75% in 30 min. HPLC buffer composition remained the same all through the work.

ESI-mass spectra were recorded on an Agilent Technologies 6520 Accurate-Mass Q-TOF LC/MS. Samples were delivered to the mass spectrometer through chromatographic separation on the Agilent HPLC 1290 and monoisotopic average masses of peptides were calculated from sequence information using the UCSF ProteinProspector MS-Product tool ([http://prospector.ucsf.edu/prospector/mshome.htm](http://prospector.ucsf.edu/prospector/mshome.htm)). Observed mass was calculated from m/z charged states using MassHunter Bioconfirm Qual B.05 software and compared to expected mass allowing for the reduction in mass of the folded peptide (1 Da per oxidized cysteine in formation of disulfide bridges).

3.5.6. **Solid Phase Peptide Synthesis**

*Method for attachment of the first residue to resin:* Wang chloride resin (1.0 g, 1.1mmol/g) was swelled in 10 ml DMF at room temperature for 30 minutes. Fmoc-Ser(tBU)-OH (3.3 mmol, 3 equivalents), Potassium Iodide [KI (0.33 mmol, 0.3 equivalents)] and N,N-Diisopropylethylamine [DIEA (3.3 mmol, 3 equivalents)] were added to the resin and the suspension was shaken for 16 to 24 hours. The loaded resin was successively filtered and washed with Dichloromethane (DCM), Dimethylformamide (DMF) and methanol, and dried under vacuum. The final substitution rate was determined photometrically from the amount of Fmoc chromophore released upon treatment of the resin with piperidine/DMF, by taking the UV absorbance at 301 nm and using a standard formula to calculate loading. Unreacted resin was then endcapped with 11 equivalents of DIEA and 10 equivalents of acetic anhydride to prevent unwanted side reactions during peptide elongation.

*Peptide elongation and cleavage of final product:* TA1 peptide was synthesized by microwave assisted
Fmoc SPPS on a CEM Liberty synthesizer using standard side chain protection. Standard Fmoc cysteine side chain protecting group Trityl (Fmoc-Cys (Trt)-OH) was used for all cysteine residues. Cleavage was performed by treatment of peptidyl resin with Reagent K 92.5% Trifluoroacetic acid (TFA), 2.5% Triisopropylsilane (TIS), 2.5% 1,2 Ethanedithiol (EDT) and 2.5% water] for four hours.

Oxidative Folding of Tv1: A one-step thiol-assisted oxidation was used to prepare folded Ta1 peptide. The linear peptide (20 mM) was incubated in 0.1 M Tris–Hydrochloride (Tris_H), 0.1 M Sodium Chloride (NaCl), 100 mM Ethylenediaminetetraacetic acid (EDTA), 1 mM Glutathione (GSH), 1 mM Glutathione Disulfide (GSSG), pH 7.5 at room temperature and 37 deg C. The folding reaction was terminated by acidification with 8% formic acid at 15 min, 30 min, 1, 2, 3, 4, and 24 h and the folding yield monitored using UHPLC. UHPLC injections were run over a 0 to 75% gradient of buffer B (80% acetonitrile in water, 0.1% TFA) in buffer A (water, 0.1% TFA) gradient at a flow rate of 0.4 ml/ min for 3.5 minutes. A preparative scale folding reaction was then conducted at an optimized time of 30 min and 37 deg C, and the folded peptide was purified using an X-Bridge semipreparative column over gradient B from 30 to 55% acetonitrile.

3.5.7. Polychaete Worm Assay

N. virens were obtained either from the commercial vendor Marine Biolabs or from local bait shop and then maintained in a saltwater and sand environment at 4° C. Lyophilized neuropeptides were dissolved into normal saline solution at the appropriate concentration and volume to prepare for injection. Injections were performed with A Becton Dickinson ½ cc U-100 insulin syringe and targeted to the anterior ventral nerve between the 5th and 7th body segments, at a depth of ~0.5 mm. Prior to injection the worms are placed in 4 identical salt water baths at room temperature and allowed to acclimate. Once injected with either peptide or saline solution, the worm is place immediately back in the salt water bath and videotaped over an approximate 4-6 hour period. After one hour, a cold stimulus is applied by placing worms in a 4 deg C water bath for 60 seconds, after which they are returned to RT water baths to assess recovery time.
Conclusion and future directions

Venom from marine organisms is a gift from nature to biomedical research. The diversity of compounds, the precise selectivity for molecular targets, and rapid efficiency when dispersed cannot be replicated without considerable effort. The small fraction of conoidean neuropeptides that have been characterized to date suggests investigation of conoidean venom is still in its early stages, yet the power of peptides found in snail venom to deliver therapeutic agents for the treatment of pain, a hugely underserved market for drug development, has already been established. Beyond neurological disorders, conoidean peptides also hold promise as anti-microbials, anti-cancer agents, and cardioprotective compounds, with other possibilities restricted only by the rate at which they can be identified, synthesized, and functionally characterized.

This dissertation project significantly characterized the venom composition of an understudied member of the conoidean superfamily, the Terebridae, and provided a reproducible research strategy for further discovery and characterization of teretoxins in the Holford lab. A venomics approach which combines a phylogenetic framework with next generation sequencing and bioinformatics analysis was applied to unravel the complexities of the terebrid venom duct transcriptome. The results of this work include: comprehensive analysis of the venom duct transcriptome of two terebrid species, the identification of the first teretoxin gene superfamily, two novel conoidean peptide cysteine frameworks, a reproducible recombinant synthesis strategy for obtaining large quantities of teretoxins for functional assays. Four novel teretoxins, Tg77, Can6.1, Can10.1, and Can14.1, were successfully synthesized and preliminary attempts were made to characterize their bioactivity. It is important to note, that characterization of teretoxin bioactivity and identification of their molecular target is not a trivial pursuit given the multiplicity of ion channels and receptors on which they are active. The astonishing diversity of methods for manipulating the nervous system displayed by predatory marine snails makes the search for a specific target a complicated endeavor. This research has shed light on the path towards identifying
therapeutic peptides for manipulating cellular communication in the nervous system. Summarized below are the specific details that generated the results achieved in this dissertation.

As demonstrated by this research, the first step in the discovery process is identification of the multitude of peptide toxins expressed by an individual species of venomous marine snail, in this case, the terebrid. Classic Sanger sequencing techniques have been almost completely replaced by next generation sequencing, which may be used to elucidate entire transcriptomes and genomes with greater speed and at a fraction of the cost compared to older technologies. The research performed here has employed the latest next generation sequencing technology and bioinformatics tools to reconstruct the first terebrid transcriptome, in order to begin the work of characterizing as many teretoxins as possible from an individual species and lay the groundwork for synthesis and testing. The analysis pipeline developed for the de novo assembly and analysis of the C. anilis venom duct transcriptome was then performed for the preliminary characterization of the T. subulata. Taken together this work has led to the identification of ~150 novel teretoxins, laying the groundwork for subsequent peptide synthesis and functional characterization. In addition, the analysis of the full teretoxin precursor structures identified through transcriptome analysis will enable the identification of gene superfamilies, which is ultimately important to understanding pharmacological activity. This research has led to the characterization of the first teretoxin gene superfamily, supported by an analysis of teretoxins from three different species that present the same signal sequence in association with a canonical six Cys framework.

Bioinformatics analyses were also conducted to obtain a portrait of the venom duct transcriptome in its broad outlines, through annotation of the transcriptome of possible venom-related genes, housekeeping genes as categorized by gene ontology function, and enzyme classification for a brief look at enzymatic pathways. Of particular interest was the annotation of proteins with a post-translational modifying function. Heavy post-translational modification is a salient feature of conotoxins, but is less well understood in teretoxins. Future work on the proteomic level, using advanced techniques in mass spectrometry, will be important to validate not only the teretoxin transcripts identified here, but also the degree of post-translational modifications on the teretoxins themselves.
The world of next generation sequencing and bioinformatics has evolved at a dramatic rate over the last few decades, and will undoubtedly continue to provide the research community with new tools and technologies. De novo assembly in the absence of a reference genome remains can still be improved, making it vital to stay abreast of the latest sequencing technologies and downstream analysis tools to obtain the most accurate and comprehensive reconstruction of genetic information. In particular, sequencing technology is a rapidly changing field, with competing technologies such as Illumina, IonTorrent, and Oxford Nanopore, as well as others, vying for dominance to produce ever longer and more accurate reads. At the same time, without bioinformatics software tools that can provide reliable assemblies and downstream analysis, having vast amounts of “big data” is not remotely useful. Bioinformatics programs tend to co-evolve with the sequencing technologies available, and in the future are likely to undergo as substantial alterations as the sequencing technology itself.

Proceeding transcriptome assembly and analysis, identification of teretoxins leads to the important step of synthesis, either by recombinant expression or SPPS. As a naturally occurring resource, teretoxins are only available in minute quantities, making synthesis a necessary task to obtain the milligram or greater quantities needed for functional assays. In this body of research, four teretoxins, Tg77, Can6.1, Can10.1, and Can14.1, have been synthesized, three by recombinant expression and one by SPPS. Devising a reproducible strategy for recombinant expression in *E. coli* has been a particular focus, since the length of teretoxins makes them more suitable for this approach than chemical synthesis. Here, Tg77 teretoxin from *T. guttata* was expressed successfully in good yield, with the application of mass spectrometry confirming the presence of the folded species. Using largely the same approach, two peptides identified from the *C. anilis* transcriptome work described in Chapter 2 were recombinantly expressed with modifications of immobilized metal affinity purification and choice of E. coli host employed to improve yield. Finally, a shorter, 27 residue peptide also identified the *C. anilis* transcriptome work was synthesized using SPPS and oxidatively folded to produce milligram quantities of peptide.

The recombinant expression of disulfide rich proteins in an *E. coli* host has varying degrees of success depending on the number of cysteines present and the amino acid composition. Efforts will be ongoing to tackle problems of protein aggregation that interfere with the final step of proteolytic cleavage.
that reduces yield of the final product. There are a wealth of options for optimizing the production of disulfide rich proteins in a bacterial host, with the emerging availability of newly engineered strains that enhance disulfide formation, as well as alternative fusion tags and proteolytic cleavage sites providing new avenues for exploration. High throughput techniques for experimenting with different constructs are becoming more readily available and facilitate the optimization of conditions in micro format prior to scale up. It is also worth noting the potential of expressing disulfide rich proteins in a eukaryotic host, which has been performed successfully in *P. pastoris*, as this approach has many of the advantages of *E. coli* in terms of low cost and well developed protocols, while providing a more favorable folding environment.\textsuperscript{110}

Synthesis of peptides in quantity has set the stage for functional assays, which were performed primarily through the use of a worm assay, using the polychaete *N. virens*. Injection of Tg77 into polychaete worms in micromolar quantities produced uncoordinated swimming and paralysis in repeated trials, similar to the effect observed upon injection of the positive control $\omega$-agatoxin IVA. Subsequent electrophysiological testing in the lab of Dr. Ines Ibanez-Tallon did not produce any specific information about molecular target, and other tests will be ongoing to see if the Tg77 mechanism of action can be pinpointed. Electrophysiology is a time and labor intensive endeavor, with no single assay providing a catch all for potential neuropeptide target, especially where phenotypic evidence does not suggest a particular line of attack.

As such, both synthesis, and to a greater degree, testing in functional assays, is the bottleneck impeding a more rapid characterization of the functional aspects of neuropeptides produced by venomous organisms. Identifying sheer numbers of potential toxins of interest is scaling up rapidly with the advent of transcriptomics and genomics, which may be especially powerful when integrated with state of the art proteomics.\textsuperscript{49} Certainly there are many advances in peptide synthesis and electrophysiology techniques, but realizing these on a genuinely high throughput level is still in need of advancement. Putting together the whole package promises to further capitalize on the therapeutic promise of conoidean neuropeptides.

In summary, this dissertation assisted in unraveling the complexities of the terebrid venom duct transcriptome. As such, it successfully identified and characterized specific venom components for further development in the Holford lab.
## Supplementary Information

### Supplementary Table 1. Trinity and Velvet Oases assembly statistics for *C. anilis*

<table>
<thead>
<tr>
<th></th>
<th>Number of reads</th>
<th>Assembly size (bp)</th>
<th># contigs &gt; 200 bp</th>
<th>Mean contig size (bp)</th>
<th>Median contig size (bp)</th>
<th>N50 (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trinity digital normalization</td>
<td>9,204,374</td>
<td>71,641,414</td>
<td>105,768</td>
<td>441</td>
<td>677</td>
<td>964</td>
</tr>
<tr>
<td>Trinity all reads</td>
<td>280,143,112</td>
<td>172,425,581</td>
<td>290,811</td>
<td>420</td>
<td>592</td>
<td>757</td>
</tr>
<tr>
<td>Velvet Oases Kmer 27</td>
<td>23,607,286</td>
<td>122,663,519</td>
<td>187,143</td>
<td>452</td>
<td>655</td>
<td>863</td>
</tr>
<tr>
<td>Velvet Oases Kmer 33</td>
<td>23,607,286</td>
<td>132,650,654</td>
<td>181,356</td>
<td>509</td>
<td>731</td>
<td>994</td>
</tr>
<tr>
<td>Velvet Oases kmer 39</td>
<td>23,607,286</td>
<td>126,084,601</td>
<td>147,925</td>
<td>550</td>
<td>852</td>
<td>1,271</td>
</tr>
<tr>
<td>Velvet Oases kmer 45</td>
<td>23,607,286</td>
<td>118,783,256</td>
<td>139,914</td>
<td>551</td>
<td>848</td>
<td>1,271</td>
</tr>
<tr>
<td>Velvet Oases kmer 51</td>
<td>23,607,286</td>
<td>104,256,362</td>
<td>127,804</td>
<td>532</td>
<td>815</td>
<td>1,213</td>
</tr>
</tbody>
</table>
REFERENCES


