IMPROVED METHODS FOR NEXT GENERATION SEQUENCING-BASED CONOTOXIN DISCOVERY

by

Qing Li

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

May 2017

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Qing Li**

has been approved by the following supervisory committee members:

Mark Yandell , Chair **March 2nd, 2017**
Date Approved

Karen Eilbeck , Member **March 2nd, 2017**
Date Approved

Jay Gertz , Member **March 2nd, 2017**
Date Approved

Lajos Horvath , Member
Date Approved

Robert Weiss , Member **March 2nd, 2017**
Date Approved

and by **Lynn B. Jorde** , Chair/Dean of

the Department/College/School of **Human Genetics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Cone snails (genus Conus) have attracted scientific interest for the great neuropharmacological potential of their venoms to treat chronic pain, which consist of a complex mixture of peptides known as conotoxins. For discovery purposes, we have carried out a survey of the venom-ducts of 22 Conus species using next generation high throughput RNAseq (NGS). *In silico* analyses of these data are complicated because paralogous conotoxin precursors display both highly conserved, as well as hyper varied regions. As a result, NGS-based discovery involves an inherent trade off between fidelity of transcript assembly and sensitivity towards novel discovery. On the one hand, overly lenient assembly parameters create a few, long, but misassembled chimeric transcripts, which lessen the true discovery potential of NGS. On the other hand, overly stringent assembly parameters can mistake sequencing artifacts as novel discoveries. Moreover, many new conotoxins likely remain undiscovered. This fact can complicate homology-based discovery efforts using tools such as BLAST because reference databases may lack homologous peptides, leading to false negative results.

With these problems in mind, I developed a comprehensive pipeline for discovery of conotoxins and their modification enzymes from high throughput RNAseq data. My pipeline includes (1) simulation software for benchmarking purposes, (2) a ‘partial extension pipeline’ that employs a novel kmerization tool called Taxonomer to rapidly cluster and taxonomically classify reads prior to assembly, and (3) a discovery engine that can identify novel conotoxins even when they lack significant homologs.

Collectively, my pipeline maximizes the discovery potential of Conus RNAseq data, identifying on average ~ 30% more full length toxins per sample than any other than approach in use today.

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to express my sincere thanks to my advisor, Dr. Mark Yandell, for his excellent guidance and patience. Mark is the best mentor I can imagine – he has given me freedom and encouraged me to learn and to try out ideas. I'm also grateful to my committee member, Dr. Karen Eilbeck, for her great suggestions and timely help whenever I turned to her. I also thank my other committee members: Dr. Robert Weiss, Dr. Jay Gertz and Dr. Lajos Horvath, for their supervision and advice during my PhD studies. I also would like to give thanks to my former committee member, Gillian Stanfield, for her kindness supervising my preliminary exam.

I would like to give special thanks to my collaborators, especially the ones for the Conus project. These include Helena Safavi, Pradip Bandyopadhyay, Sam Robinson, Aiping Lu and Baldomero Olivera. I will always remember the time we spent developing the Conus pipeline as a team.

I also appreciate the help I got from other Yandell lab members: Carson Holt, Barry Moore, Hao Hu, Mike Campbell, Daniel Ence, Zev Kronenberg, Steven Flygare, EJ Osborne, Aurelie Kapusta, Edgar Hernandez, Scott Watkins and Edwin Lin. I will remember the days we spent in the lab together – and all of the scientific/silly/funny/crazy discussions we had.

I would also like to thank my Lord, my parents, and my brothers and sisters at Church for their unconditional support and love.

CHAPTER 1

INTRODUCTION

Chronic neuropathic pain

15 of every 100 people in the United States report some form of chronic neuropathic pain (MITKA 2003; ALFORD *et al.* 2008). Chronic neuropathic pain is intractable pain caused by damage of peripheral or central nervous systems, which can be directly caused by lesions or indirectly through disease, like diabetes and shingles. The damaged nerve fibers send incorrect signals to other pain centers and numerous classes of ion channels and receptors participate in the propagation and processing of pain signals (JULIUS AND BASBAUM 2001; BOULPAEP *et al.* 2009).

For many patients, chronic neuropathic pain severely interferes with the quality of their daily lives, since poor management of the pain can cause significant social, psychological and financial consequences (SCHOLZ AND WOOLF 2002; BREIVIK *et al.* 2006; ALFORD *et al.* 2008). Moreover, the presence of untreated pain can limit the patients participating in rehabilitation programs, which will result in reduced functional recovery of them (PERRET AND LUO 2009).

Traditional chemically synthesized analgesic drugs are only marginally effective at best to this disease, with all kinds of side effects (STAATS *et al.* 2004). Because the molecular targets of traditional analgesic drugs serve crucial roles in both normal physiology function and the pathology pain pathway, severe side effects are not unexpected. New drugs targeting with greater specificity for the pain pathway are highly desired in the treatment of chronic neuropathic pain (PERRET AND LUO 2009).

Conus venoms are a unique resource for drug discovery

Pharmacologists and physicians have long been interested in venomous animals (RASH AND HODGSON 2002; LEWIS AND GARCIA 2003; BOGIN 2005). The marine Cone snails are one of the largest and most venomous animal clades. Cone snails belong to the large genus *Conu*s of predatory marine molluscs (~730 extant species according to WoRMS – accessed on 09.28.2016). Cone snails hunt a diverse range of animals, including fish, worms and other snails, and produce complex venom of peptide toxins, which are known as conotoxins or conopeptides (OLIVERA *et al.* 1990; TERLAU AND OLIVERA 2004).

Conotoxins have attracted the attention of pharmacologists and physicians for their unique ability to block critical components of the nervous or muscular systems, especially ion channels (SHEN *et al.* 2000; MCINTOSH AND JONES 2001; LIVETT *et al.* 2004). By utilizing the conotoxins that target receptors and ion channels in prey's nervous system with remarkable potency and specificity, there is promise for drugs with greater efficacy and reduced off target side effects (OLIVERA *et al.* 1985). Because homologs of many of these molecular targets are also found in humans, conotoxins have become invaluable tools in neuroscience, for drug design, and as pharmacotherapeutics. One conotoxin is already an approved drug for pain treatment, seven are in preclinical and clinical trials, two are used as clinical diagnostics, one was used to co-crystallize the AMPA receptor, and many others serve as probes to study specific receptors and ion channels of our nervous system (MOTOMURA AND IWANAGA 2001; POPE AND DEER 2013; CHEN *et al.* 2014; BAKER *et al.* 2015).

Previous work has shown that each of the ~730 species of cone snail produces a distinct set of 70-400 conotoxins with numerous and unusual posttranslational modifications (HU *et al.* 2012; BARGHI *et al.* 2015a; PHUONG *et al.* 2016). Thus, Conus venoms can be regarded as a unique source of mostly uncharacterized, highly specific compounds of tremendous pharmacological interest. Conus venoms have been increasingly studied over the past two decades, and conotoxins have demonstrated medical, translational and research values (SHEN *et al.* 2000; MCINTOSH AND JONES 2001; LIVETT *et al.* 2004; PERRET AND LUO 2009; POPE AND DEER 2013; BAKER *et al.* 2015).

Discovery effort

Traditionally, venoms were separated by assay-guided fractionation to isolate individual peptides. However, this approach is time intensive, and requires a large amount of venom, which is not always available (PRASHANTH *et al.* 2012). In the past few years, with the reduced cost of 454 Pyrosequencing and Illumina, which only require a moderate amount of starting material, sequencing the venom gland transcriptome has become an affordable and relatively quick way to fingerprint the venom profile of animals (HU *et al.* 2011; HU *et al.* 2012).

Compared to Illumina platform, 454 Pyrosequencing platform has the advantage of producing longer reads, which can cover the entire conotoxin precursor cDNA (~300bp) in a single read, thus circumventing the issue of assembly. However, 454 Pyrosequencing platform is limited by its relatively low sequencing quality and throughput, making it less suitable than the Illumina platform to comprehensively profile highly diverse Conus venoms (HUSE *et al.* 2007; ARCHER *et al.* 2012). However, its relatively short read lengths complicate analyses (BARGHI *et al.* 2015a; MACRANDER *et al.* 2015; SCHIRMER *et al.* 2015).

The conotoxin sequences themselves are a big part of the challenge. Conotoxins are translated from mRNA as peptide precursors, which can be readily divided into three distinct regions: (1) an N-terminal signal sequence for targeting to the endoplasmic reticulum; (2) an intermediate propeptide region that has been suggested to play a role in secretion, posttranslational modification and folding; and (3) a single copy of the mature toxin region at the C terminus (BANDYOPADHYAY *et al.* 1998; CONTICELLO *et al.* 2003; BUCZEK *et al.* 2004). Conotoxins that belong to the same toxin superfamily share a highly conserved signal, sequence but the mature toxin region is hyper variable, with the exception of a conserved cysteine framework (BUCZEK *et al.* 2005). The conotoxin sequence pattern can be seen in Figure 1.1.

Subtle variations in the sequences of mature toxins often confer important changes in target specificity and potency; these subtle variations also complicate venom discovery and characterization efforts (ELLISON *et al.* 2003; SAFAVI-HEMAMI *et al.* 2011; HU *et al.* 2012). The challenge is how to confidently assemble paralogous conotoxins using reads, as a single amino acid substitution may define to pharmacologically distinct species. In such cases, even today's best practice assembly approaches often fail (Figure 1.2). Similar problems have been reported by groups attempting to assemble anemone venoms and human CDR3 genes (MACRANDER *et al.* 2015; LI *et al.* 2016). Therefore, conotoxin researchers are not the only ones facing these difficulties. Thus, a solution to this problem not only benefits *Conus* researchers, but other groups as well.

Summary

Conotoxins have attracted more and more scientific interest because of their medical, translational and research value (SHEN *et al.* 2000; MCINTOSH AND JONES 2001; LIVETT *et al.* 2004; PERRET AND LUO 2009; POPE AND DEER 2013; BAKER *et al.* 2015). Now next generation RNAseq (NGS) enables conotoxin discovery at speeds and scale never before possible. However, NGS-based approaches for conotoxin discovery are far from perfect (HU *et al.* 2011; HU *et al.* 2012; BARGHI *et al.* 2015b; SCHIRMER *et al.* 2015). Thus, for my dissertation, I developed a comprehensive pipeline for conotoxin discovery using Illumina NGS data. The comprehensive pipeline includes a simulation pipeline for benchmarking and discovery of optimal parameter settings (Chapter 2), a partial extension pipeline to extend the truncated conotoxin transcripts to full length (Chapter 2) and a discovery pipeline to discover the potential novel conotoxins that have no known homologs identifiable with tools such as BLAST (Chapter 4).

In Chapter 3, I explain how I have repurposed my partial extension pipeline for identification of novel conotoxin modification enzymes (conotoxin specific protein disulfide isomerase). My dissertation demonstrates the power of this discovery pipeline and documents its discoveries.

References

Alford, D. P., J. Liebschutz, I. A. Chen, C. Nicolaidis, M. Panda *et al.*, 2008 Update in pain medicine. J Gen Intern Med 23**:** 841-845.

Archer, J., G. Baillie, S. J. Watson, P. Kellam, A. Rambaut *et al.*, 2012 Analysis of high-depth sequence data for studying viral diversity: a comparison of next generation sequencing platforms using Segminator II. Bmc Bioinformatics 13:47.

Baker, B. J., B. M. Olivera, H. Safavi-Hemami, M. P. Horvath and R. W. Teichert, 2015 Conopeptides, Marine Natural Products from Venoms: Biomedical Applications and Future Research Applications, pp. 463-496 in *Marine Biomedicine: From Beach to Bedside*. CRC Press.

Bandyopadhyay, P. K., C. J. Colledge, C. S. Walker, L.-M. Zhou, D. R. Hillyard *et al.*, 1998 Conantokin-G precursor and its role in γ-Carboxylation by a vitamin K-dependent carboxylase from a conus snail. J. Biol. Chem 273**:** 5447-5450.

Barghi, N., G. P. Concepcion, B. M. Olivera and A. O. Lluisma, 2015a Comparison of the venom peptides and their expression in closely related Conus species: insights into adaptive post-speciation evolution of Conus exogenomes. Genome Biol. Evol 7**:** 1797-1814.

Barghi, N., G. P. Concepcion, B. M. Olivera and A. O. Lluisma, 2015b High conopeptide diversity in Conus tribblei revealed through analysis of venom duct transcriptome using two high-throughput sequencing platforms. Mar. Biotechnol 17**:** 81-98.

Bogin, O., 2005 Venom peptides and their mimetics as potential drugs. Modulator 19**:** 14-20.

Boulpaep, E. L., W. F. Boron, M. J. Caplan, L. Cantley, P. Igarashi *et al.*, 2009 Medical physiology a cellular and molecular approach. Signal Transduction 48**:** 27.

Breivik, H., B. Collett, V. Ventafridda, R. Cohen and D. Gallacher, 2006 Survey of chronic pain in Europe: prevalence, impact on daily life, and treatment. Eur J Pain 10**:** 287-333.

Buczek, O., G. Bulaj and B. M. Olivera, 2005 Conotoxins and the posttranslational modification of secreted gene products. Cell Molec Life Sci 62**:** 3067-3079.

Buczek, O., B. M. Olivera and G. Bulaj, 2004 Propeptide does not act as an intramolecular chaperone but facilitates protein disulfide isomerase-assisted folding of a conotoxin precursor. Biochemistry 43**:** 1093-1101.

Chen, L., K. L. Durr and E. Gouaux, 2014 X-ray structures of AMPA receptor-cone snail toxin complexes illuminate activation mechanism. Science 345**:** 1021-1026.

Conticello, S. G., N. D. Kowalsman, C. Jacobsen, G. Yudkovsky, K. Sato *et al.*, 2003 The prodomain of a secreted hydrophobic mini-protein facilitates its export from the endoplasmic reticulum by hitchhiking on sorting receptors. J. Biol. Chem. 278**:** 26311-

26314.

Ellison, M., J. M. McIntosh and B. M. Olivera, 2003 alpha-conotoxins ImI and ImII - similar alpha 7 nicotinic receptor antagonists act at different sites. J. Biol. Chem. 278**:** 757-764.

Hu, H., P. K. Bandyopadhyay, B. M. Olivera and M. Yandell, 2011 Characterization of the Conus bullatus genome and its venom-duct transcriptome. Bmc Genomics 12:60.

Hu, H., P. K. Bandyopadhyay, B. M. Olivera and M. Yandell, 2012 Elucidation of the molecular envenomation strategy of the cone snail Conus geographus through transcriptome sequencing of its venom duct. Bmc Genomics 13:284.

Huse, S. M., J. A. Huber, H. G. Morrison, M. L. Sogin and D. Mark Welch, 2007 Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8.
Julius, D., and A. I. Basbaum, 2001 Molecular mechanisms of nociception. Nature 413**:** 203-210.

Lewis, R. J., and M. L. Garcia, 2003 Therapeutic potential of venom peptides. Nature Reviews Drug Discovery 2**:** 790-802.

Li, B., T. W. Li, J. C. Pignon, B. B. Wang, J. Z. Wang *et al.*, 2016 Landscape of tumor-infiltrating T cell repertoire of human cancers. Nature Genet 48**:** 725-+.

Livett, B. G., K. R. Gayler and Z. Khalil, 2004 Drugs from the sea: Conopeptides as potential therapeutics. Curr. Med. Chem. 11**:** 1715-1723.

Macrander, J., M. Broe and M. Daly, 2015 Multi-copy venom genes hidden in de novo transcriptome assemblies, a cautionary tale with the snakelocks sea anemone Anemonia sulcata (Pennant, 1977). Toxicon 108**:** 184-188.

McIntosh, J. M., and R. M. Jones, 2001 Cone venom - from accidental stings to deliberate injection. Toxicon 39**:** 1447-1451.

Mitka, M., 2003 "Virtual textbook" on pain developed effort seeks to remedy gap in medical education. JAMA. 290**:** 2395-2395.

Motomura, M., and H. Iwanaga, 2001 [Lambert-Eaton myasthenic syndrome: diagnosis and treatment]. Clin Calcium 11**:** 1468-1474.

Olivera, B. M., W. R. Gray, R. Zeikus, J. M. Mcintosh, J. Varga *et al.*, 1985 Peptide neurotoxins from fish-hunting Cone snails. Science 230**:** 1338-1343.

Olivera, B. M., J. Rivier, C. Clark, C. A. Ramilo, G. P. Corpuz *et al.*, 1990 Diversity of Conus neuropeptides. Science 249**:** 257-263.

Perret, D., and Z. D. Luo, 2009 Targeting voltage-gated calcium channels for neuropathic pain management. Neurotherapeutics 6**:** 679-692.

Phuong, M. A., G. N. Mahardika and M. E. Alfaro, 2016 Dietary breadth is positively correlated with venom complexity in cone snails. Bmc Genomics 17:401.

Pope, J. E., and T. R. Deer, 2013 Ziconotide: a clinical update and pharmacologic review. Expert. Opin. Pharmacother. 14**:** 957-966.

Prashanth, J. R., R. J. Lewis and S. Dutertre, 2012 Towards an integrated venomics approach for accelerated conopeptide discovery. Toxicon 60**:** 470-477.

Rash, L. D., and W. C. Hodgson, 2002 Pharmacology and biochemistry of spider venoms. Toxicon 40**:** 225-254.

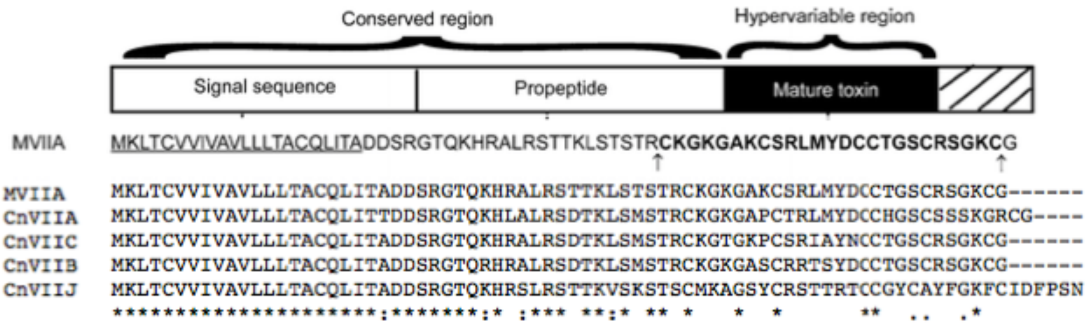
Safavi-Hemami, H., W. A. Siero, Z. H. Kuang, N. A. Williamson, J. A. Karas *et al.*, 2011 Embryonic toxin expression in the Cone snail Conus victoriae PRIMED TO KILL OR DIVERGENT FUNCTION? J. Biol. Chem. 286**:** 22546-22557.

Schirmer, M., U. Z. Ijaz, R. D'Amore, N. Hall, W. T. Sloan *et al.*, 2015 Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res 43:1341.

Scholz, J., and C. J. Woolf, 2002 Can we conquer pain? Nat Neurosci. 5**:** 1062-1067.
Shen, G. S., R. T. Layer and R. T. McCabe, 2000 Conopeptides: From deadly venoms to novel therapeutics. Drug Discovery Today 5**:** 98-106.

Staats, P. S., T. Yearwood, S. G. Charapata, R. W. Presley, M. S. Wallace *et al.*, 2004 Intrathecal ziconotide in the treatment of refractory pain in patients with cancer or AIDS - A randomized controlled trial. JAMA. 291**:** 63-70.

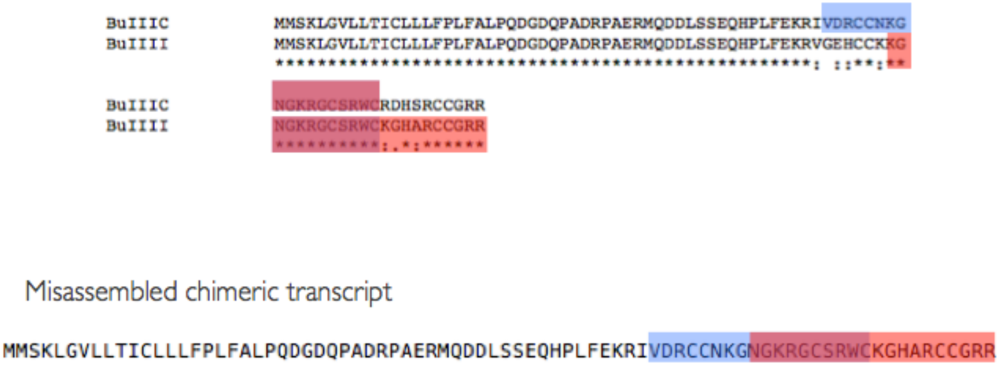
Terlau, H., and B. M. Olivera, 2004 Conus venoms: A rich source of novel ion channel-targeted peptides. Physiol Rev. 84**:** 41-68.



```
MVIIA    MKLTCVVIVAVLLLTACQLITADDSRGTQKHRALRSTTKLSTSTRCKGKGAKCSRLMYDCCTGSCRSGKCG

MVIIA    MKLTCVVIVAVLLLTACQLITADDSRGTQKHRALRSTTKLSTSTRCKGKGAKCSRLMYDCCTGSCRSGKCG------
CnVIIA   MKLTCVVIVAVLLLTACQLITTDDSRGTQKHLALRSDTKLSMSTRCKGKGAPCTRLMYDCCHGSCSSSKGRCG----
CnVIIC   MKLTCVVIVAVLLLTACQLITADDSRGTQKHRALRSDTKLSMSTRCKGTGKPCSRIAYNCCTGSCRSGKCG------
CnVIIB   MKLTCVVIVAVLLLTACQLITADDSRGTQRHRALRSDTKLSMSTRCKGKGASCRRTSYDCCTGSCRSGKCG------
CnVIIJ   MKLTCVVIVAVLLLTACQLITADDSRGTQKHRSLRSTTKVSKSTSCMKAGSYCRSTTRTCCGYCAYFGKFCIDFPSN
         *********************:*******:* :*** **:* ** *   *  *      **   ..  .*
```

Figure 1.1. The sequence motifs of conotoxins. The sequence pattern of conotoxins showed three distinct regions.


```
BuIIIC          MMSKLGVLLTICLLLFPLFALPQDGDQPADRPAERMQDDLSSEQHPLFEKRIVDRCCNKG
BuIIII          MMSKLGVLLTICLLLFPLFALPQDGDQPADRPAERMQDDLSSEQHPLFEKRVGEHCCKKG
                ***************************************************; ::**:**

BuIIIC          NGKRGCSRWCRDHSRCCGRR
BuIIII          NGKRGCSRWCKGHARCCGRR
                **********:.*:******
```

Misassembled chimeric transcript

```
MMSKLGVLLTICLLLFPLFALPQDGDQPADRPAERMQDDLSSEQHPLFEKRIVDRCCNKGNGKRGCSRWCKGHARCCGRR
```

Figure 1.2. Misassembled chimeric conotoxin transcript. Current standard RNAseq assembly produces misassembled chimeric conotoxin transcripts.

CHAPTER 2

SIMULATION AND PARTIAL EXTENSION PIPELINE FOR IMPROVED CONOTOXIN ASSEMBLY

Abstract

Next generation RNA sequencing makes possible comprehensive expression profiling, but faithfully reconstructing complete transcripts from such massive data remains difficult, especially for transcripts with both highly conserved and hyper variable regions, such as paralogous conotoxin precursors. Here, I describe a simulation pipeline designed to discover optimal parameters for conotoxin assembly, and a partial extension pipeline designed to extend partially assembled transcripts to their full lengths. By using these two pipelines, I demonstrate marked improvements in sensitivity and specificity over current standard RNAseq assembly on both simulated and real data.

Introduction

Next generation RNA now provides cost-effective means for investigating the transcriptomes of organisms and tissue types (BIROL *et al.* 2009; TRAPNELL *et al.* 2010). Ideally, these data would allow us to identify all expressed transcripts and (following assembly) to extract full length, even multiple alternatively spliced isoforms by assembly (GUTTMAN *et al.* 2010). Reality proves more difficult.

The RNAseq reads are assembled into transcripts using short read transcript assembly programs. A number of assembly programs are available. Trinity is commonly used to assemble transcriptomes when investigating venom composition (GRABHERR *et al.* 2011; HE *et al.* 2013; HARGREAVES *et al.* 2014), because it tends to perform well compared to other tools (YANG AND SMITH 2013).

Standard Trinity assembly uses a single kmer with size fixed at 25 base pairs (bp). Trinity first builds linear contigs using a greedy extension algorithm, starting from the most abundant kmer. Then Trinity groups overlapping contigs into connected components, and constructs a de Bruijn graph for each component. It then maps reads to the graphs, simplifies and corrects the graphs according to mapped reads, and extracts likely isoforms for each component (GRABHERR *et al.* 2011).

However, no tool is perfect. And no one set of parameters works optimally for every class of transcript of the dataset. This is especially true for *Conus* venom transcripts, which have both highly conserved and hyper variable regions. This makes *de novo* reconstruction and parsing of gene copies extremely difficult. This problem is further compounded when paralogous toxin genes are expressed at low frequencies. These phenomena notoriously produce chimeric truncated conotoxin transcripts (VIJAY *et al.* 2013; LIGHTEN *et al.* 2014).

NGS-based transcriptome mining involves an inherent trade off between the fidelity of transcript assembly and sensitivity for novel discovery. On the one hand, overly lenient assembly parameters create a few, long, but misassembled chimeric transcripts.. (MORAN *et al.* 2008; SABOURAULT *et al.* 2009; KOZLOV AND GRISHIN 2011; NICOSIA *et al.* 2013; YANG AND SMITH 2013). On the other hand, overly stringent assembly parameters can mistake sequencing errors for novel discoveries (MACRANDER *et al.* 2015).

Previous studies have developed ad hoc postassembly filtering approaches to minimize assembly chimeras and maximize total gene coverage (YANG AND SMITH 2013). Still other studies have tried to recover the hidden transcripts by aligning raw reads back to the candidate transcripts after assembly (MACRANDER *et al.* 2015). However, no study to date has ever attempted to discover and optimize assembly parameters for Conotoxin discovery

Here, we present a simulation pipeline to optimize conotoxin transcripts assembly by simulating and testing a number of different assembly parameters. Moreover, conotoxin assembly is further optimized by using the partial extension pipeline, which employs a novel kmerization tool called Taxonomer, which can very rapidly cluster and taxonomically classify reads prior to assembly, enabling targeted and precise micro reassembly for truncated transcripts (FLYGARE *et al.* 2016). This method can be generalized to customize RNAseq assembly for any application. Our method is implemented in Perl as a complete transcriptome assembly package under the name ConusPipe. It is available from https://github.com/Yandell-Lab/ConusPipe.

Materials and methods

Our transcriptome assembly package proceeds by first using its simulation pipeline to simulate and test a number of different assembly parameters to discover best practice assembly parameters for the target transcripts (here, a set of reference conotoxin transcripts are our targets). Next, RNAseq reads from real data are assembled with Trinity using the best practice assembly parameters, and all the conotoxin transcripts are pulled out. The partial extension pipeline is then used to extend the truncated conotoxin transcripts to full length by doing targeted local reassembly for each truncated one (Figure 2.1).

The simulation pipeline first simulates RNAseq reads from targeted transcripts (here, we use the manually curated Conus *bullatus* venom transcriptome as the gold standard reference). Next, the pipeline modifies a series of Trinity assembly parameters, and for each modification, the pipeline launches Trinity to do a separate assembly with simulated RNAseq reads. Then the pipeline uses BLAST to search each new assembly dataset and the gold standard reference transcripts against each other, and identifies a reciprocal best hit (RBH) for each reference transcript. The accuracy (the average of sensitivity and specificity) is then calculated for each new Trinity assembly dataset based upon the numbers and completeness of the RBHs recovered. These parameters are then used to assemble the new Trinity assembly dataset, and so on until a set of parameters is identified that has the highest accuracy. These are defined as the best practice parameters for the target transcripts (Figure 2.2).

The partial extension pipeline first employs Taxonomer to kmerize truncated conotoxin transcripts and RNAseq reads from sequence data and find the shared kmers between truncated transcripts and RNAseq reads (FLYGARE *et al.* 2016). Next, the RNAseq reads that have shared kmers with truncated transcripts are mapped back to the truncated transcripts. Then the pipeline does targeted local reassembly for each truncated transcript with additional reads recovered by kmer matching. In this way, the truncated transcripts are extended to full length in targeted local reassembly (Figure 2.3). After partial extension, all the putative full-length conotoxins are combined into one file and the pipeline removes the redundant conotoxins.

For benchmarking, simNGS (v1.6)(http://www.ebi.ac.uk/goldman-srv/simNGS/) was used to simulate strand-specific paired end RNAseq reads from the manually curated Conus *bullatus* venom transcriptome with a mean fragment length of 220 nt and various read abundances (SAFAVI-HEMAMI *et al.* 2016b). Paired end RNAseq data from two biological replicates of *C. andremenezi*, a single replicate of *C. praecellens* and a single replicate of *C. geographus* were generated by Illumina HiSeq 2000 platform (Table 2.1) (Li *et al.* submitted to Genome Biology and Evolution). Simulated and real RNAseq reads were assembled using standard Trinity settings, best practice Trinity settings and best practice Trinity settings plus partial extension pipeline, respectively, to compare the three approaches.

Results

I compared the performance of our assembly package - best practice Trinity settings plus partial extension pipeline with standard Trinity settings and best practice Trinity settings alone, on a number of different datasets.

Simulated data

For the simulation study, a dataset of strand specific paired end RNAseq reads with various read lengths and abundances were simulated from the manually curated Conus *bullatus* venom transcriptome using simNGS (v1.6)(http://www.ebi.ac.uk/goldman-srv/simNGS/) (SAFAVI-HEMAMI *et al.* 2016b). The read lengths and abundances were simulated according to RNAseq data for Conus. The main measures of performance were sensitivity and specificity. Sensitivity was defined as the fraction of correctly reconstructed reference transcripts (TP) over the number of reference transcripts that were either correctly reconstructed (TP) or had no RBH pair (FN). Specificity was defined as the fraction of correctly reconstructed reference transcripts (TP) over the sum of number of reference transcripts that were correctly reconstructed (TP) and number of assembled transcripts that completely missed the rbh pair (FP).

Our method – the combination of simulation pipeline and partial extension pipeline – exhibited both higher sensitivity and specificity than the other two methods – standard assembly or applying simulation pipeline alone, with all read abundances (Figure 2.4, 2.5). Our method has an average sensitivity of 93.5% and specificity of 72.2% across all read abundances, while the standard assembly method only has average sensitivity of 62.5% and specificity of 57.7%. Interesting increased sequence coverage depth did not necessarily guarantee higher assembly specificity using standard assembly parameters.. My pipeline, however does (Figure 2.5).

Real data

I benchmarked the performance of my pipeline on real data, using paired end RNAseq data from two biological replicates of *C. andremenezi*, a single replicate of *C. praecellens* and a single replicate of *C. geographus*; all four datasets were generated using the Illumina HiSeq 2000 platform. Importantly, these datasets represent different *Conus* species, library construction protocols and sequencing depths. As there is no gold standard transcriptome reference for real data, all assembled conotoxin transcripts were annotated using NCBI-BLASTX against a combined ConoServer and UniProtKB database and then manually checked and confirmed by conotoxin experts (KAAS *et al.* 2012).

My method found more verified full-length conotoxin transcripts both in absolute numbers and relative to the total number of predicted transcripts for each Conus species, thus demonstrating its superior sensitivity and specificity, respectively (Table 2.2).

Discussion

I have devised an approach that improves RNAseq short read assembly for highly similar sequences. Compared to previous studies, which only attempted to minimize assembly chimeras by postassembly filtering or recover hidden transcripts by post-assembly aligning, mine is the first tool for optimizing assembly parameters to minimize chimeras and maximize discovery yields. Mine operates by simulating and testing a number of different assembly parameters and uses Taxonomer to rapidly cluster and taxonomically classify reads prior to assembly, enabling targeted and precise micro reassembly for truncated transcripts.

Discovering the best practice assembly parameters took 72 hours on 30 CPU cores to for a 58,612nt venom transcriptome (6.5 million simulated paired end reads), and 96 hours on 30 CPU cores to run partial extension pipeline to extend the truncated transcripts to full length. For simulated data, my approach outperformed the current standard RNAseq. For real data benchmarking, my approach found more verified full length conotoxin transcripts both in absolute numbers and relative to the total number of predicted transcripts. My approach is a significant advance for t *de novo* conotoxin assembly with short reads, and therefore has great pharmacological discovery potential, it also has the strong potential to improve the quality and quantity of data mining any transcriptome. My pipeline has been employed in several *Conus* projects and has led to novel discoveries and publications, such as hormone-like peptides, venom insulin and species-specific conotoxin diversity (ROBINSON *et al.* 2015; SAFAVI-HEMAMI *et al.* 2015; SAFAVI-HEMAMI *et al.* 2016b). Moreover, I have also used the partial extension pipeline for discovery of conotoxin-specific protein disulfide isomerases (csPDI) (SAFAVI-

HEMAMI *et al.* 2016a). Crucially, the functionally distinct isoforms of this novel enzyme family were invisible using current standard transcriptome assembly methods. I describe these findings in more detail in Chapter 3.

References

Birol, I., S. D. Jackman, C. B. Nielsen, J. Q. Qian, R. Varhol *et al.*, 2009 De novo transcriptome assembly with ABySS. Bioinformatics 25**:** 2872-2877.

Flygare, S., K. Simmon, C. Miller, Y. Qiao, B. Kennedy *et al.*, 2016 Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. Genome Biology 17:111.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29**:** 644-U130.

Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson *et al.*, 2010 Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nature Biotechnology 28**:** 503-U166.

Hargreaves, A. D., M. T. Swain, M. J. Hegarty, D. W. Logan and J. F. Mulley, 2014 Restriction and recruitment—gene duplication and the origin and evolution of snake venom toxins. Genome Biol. Evol. 6**:** 2088-2095.

He, Q. Z., Z. G. Duan, Y. Yu, Z. Liu, Z. H. Liu *et al.*, 2013 The venom gland transcriptome of Latrodectus tredecimguttatus revealed by deep sequencing and cDNA library analysis. Plos One 8: e81357.

Kaas, Q., R. L. Yu, A. H. Jin, S. Dutertre and D. J. Craik, 2012 ConoServer: updated content, knowledge, and discovery tools in the conopeptide database. Nucleic. Acids. Res. 40**:** D325-D330.

Kozlov, S., and E. Grishin, 2011 The mining of toxin-like polypeptides from EST database by single residue distribution analysis. Bmc Genomics 12:88.

Lighten, J., C. Van Oosterhout and P. Bentzen, 2014 Critical review of NGS analyses for de novo genotyping multigene families. Molecular Ecology 23**:** 3957-3972.

Macrander, J., M. Broe and M. Daly, 2015 Multi-copy venom genes hidden in de novo transcriptome assemblies, a cautionary tale with the snakelocks sea anemone Anemonia sulcata (Pennant, 1977). Toxicon 108**:** 184-188.

Moran, Y., H. Weinberger, J. C. Sullivan, A. M. Reitzel, J. R. Finnerty *et al.*, 2008 Concerted evolution of sea anemone neurotoxin genes is revealed through analysis of the Nematostella vectensis genome. Molecular Biology and Evolution 25**:** 737-747.

Nicosia, A., T. Maggio, S. Mazzola and A. Cuttitta, 2013 Evidence of accelerated evolution and ectodermal-specific expression of presumptive BDS toxin cDNAs from Anemonia viridis. Marine Drugs 11**:** 4213-4231.

Robinson, S. D., Q. Li, P. K. Bandyopadhyay, J. Gajewiak, M. Yandell *et al.*, 2015 Hormone-like peptides in the venoms of marine cone snails. General and Comparative Endocrinology 244:11-18.

Sabourault, C., P. Ganot, E. Deleury, D. Allemand and P. Furla, 2009 Comprehensive EST analysis of the symbiotic sea anemone, Anemonia viridis. Bmc Genomics 10:333.
Safavi-Hemami, H., J. Gajewiak, S. Karanth, S. D. Robinson, B. Ueberheide *et al.*, 2015

Specialized insulin is used for chemical warfare by fish-hunting cone snails. Proceedings of the National Academy of Sciences of the United States of America 112**:** 1743-1748.

Safavi-Hemami, H., Q. Li, R. L. Jackson, A. S. Song, W. Boomsma *et al.*, 2016a Rapid expansion of the protein disulfide isomerase gene family facilitates the folding of venom peptides. Proceedings of the National Academy of Sciences of the United States of America 113**:** 3227-3232.

Safavi-Hemami, H., A. P. Lu, Q. Li, A. E. Fedosov, J. Biggs *et al.*, 2016b Venom insulins of Cone snails diversify rapidly and track prey taxa. Molecular Biology and Evolution 33**:** 2924-2934.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan *et al.*, 2010 Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28**:** 511-U174.

Vijay, N., J. W. Poelstra, A. Kunstner and J. B. W. Wolf, 2013 Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Molecular Ecology 22**:** 620-634.

Yang, Y., and S. A. Smith, 2013 Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. Bmc Genomics 14:328.

Table 2.1 RNAseq data sets used for the benchmarking

Conus Species	Illumina HiSeq 2000	
	Number of reads	Read length(nt)
C.andremenezi	63,598,020	101
C.andremenezi 2	54,177,324	91
C. geographus	158,004,874	101
C. pracellens	76,907,910	101

Table 2.2 Benchmark assembly results using real data. Simulation plus partial extension pipeline find 30% more full length conotoxins than current standard assembly methods

Conus Species	Number of full length conotoxins		
	Standard assembly	Apply simulation pipeline: best practice	Apply simulation+ partial extension pipeline
C.andremenezi	95	103	129
C.andremenezi 2	85	88	108
C. geographus	80	84	105
C. pracellens	115	124	155

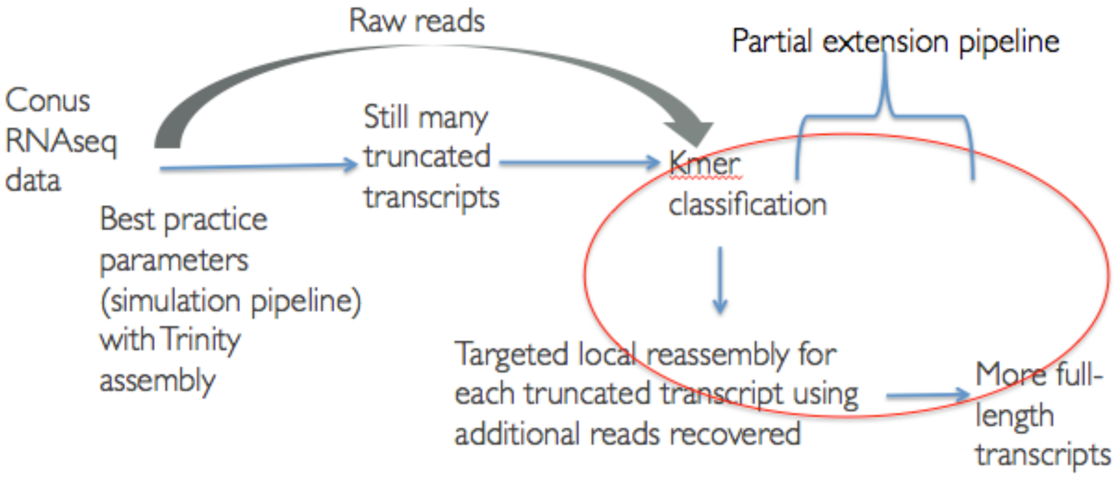

Raw reads
Partial extension pipeline
Conus RNAseq data
Still many truncated transcripts
Kmer classification
Best practice parameters (simulation pipeline) with Trinity assembly
Targeted local reassembly for each truncated transcript using additional reads recovered
More full-length transcripts

Figure 2.1. Simulation and partial extension pipeline. Illustrations of how simulation and partial extension pipeline improve conotoxin assembly.

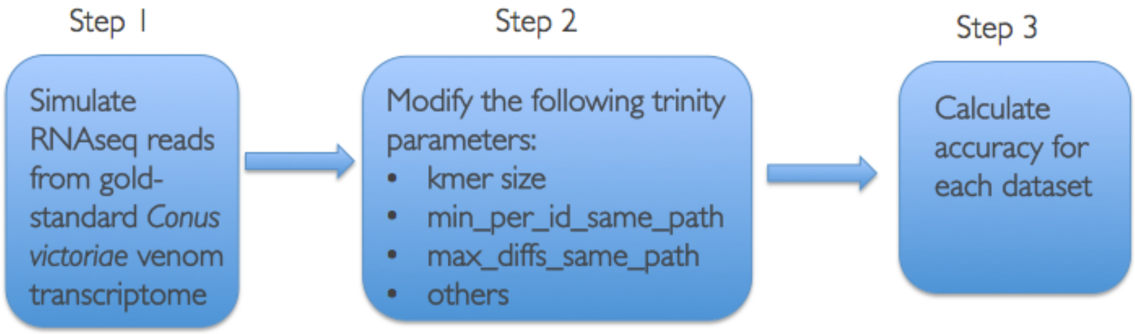

Step 1
Simulate RNAseq reads from gold-standard *Conus victoriae* venom transcriptome
Step 2
Modify the following trinity parameters:
• kmer size
• min_per_id_same_path
• max_diffs_same_path
• others
Step 3
Calculate accuracy for each dataset

Figure 2.2. Simulation pipeline. A simulation pipeline discovers the best practice assembly parameters.

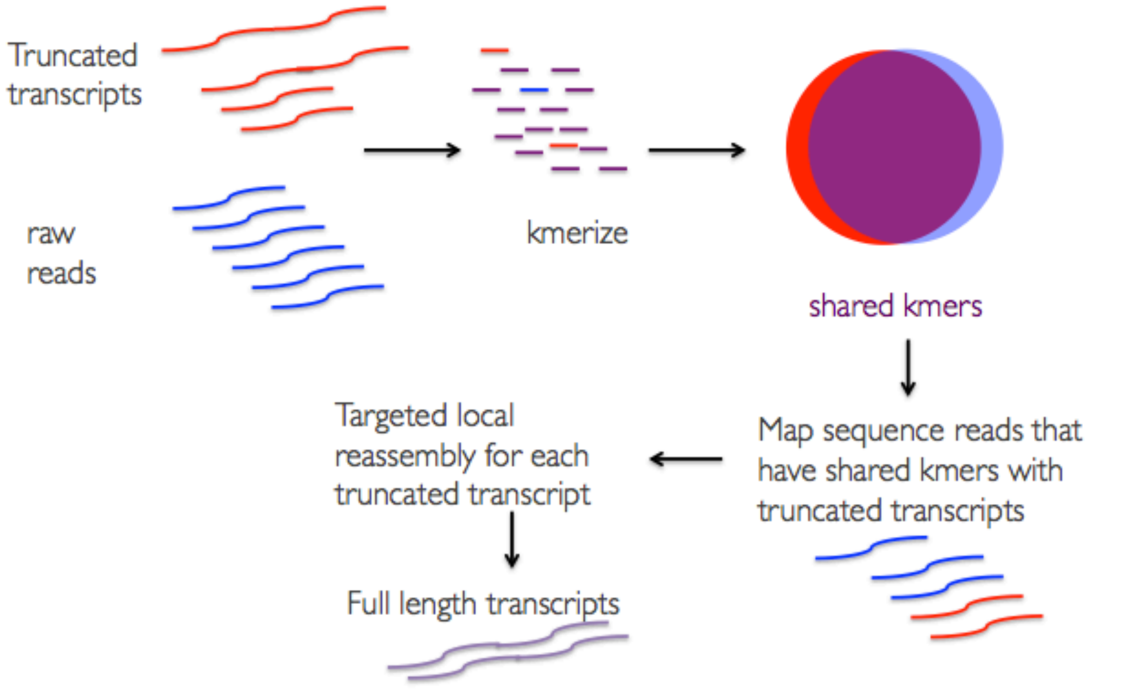


Figure 2.3. Partial extension pipeline. My partial extension pipeline uses kmers to aid classification of additional reads to truncated transcripts for their extension to full length.

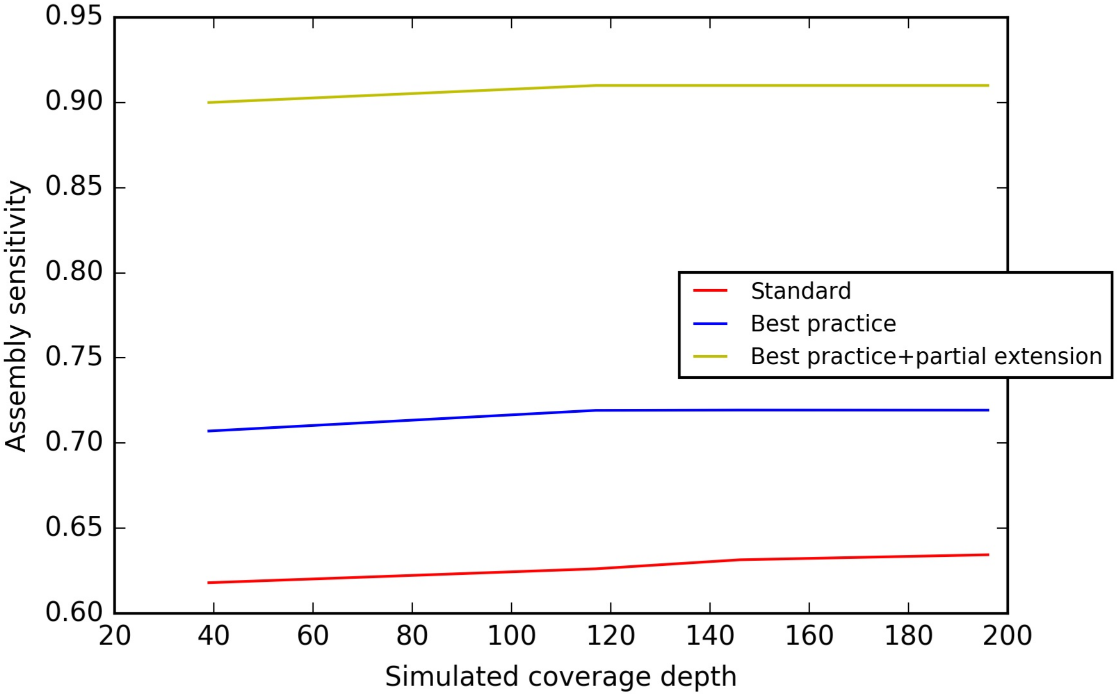


Figure 2.4. Sensitivity plot. The combination of simulation pipeline and partial extension pipeline has the highest assembly sensitivity across different sequence coverage depth.

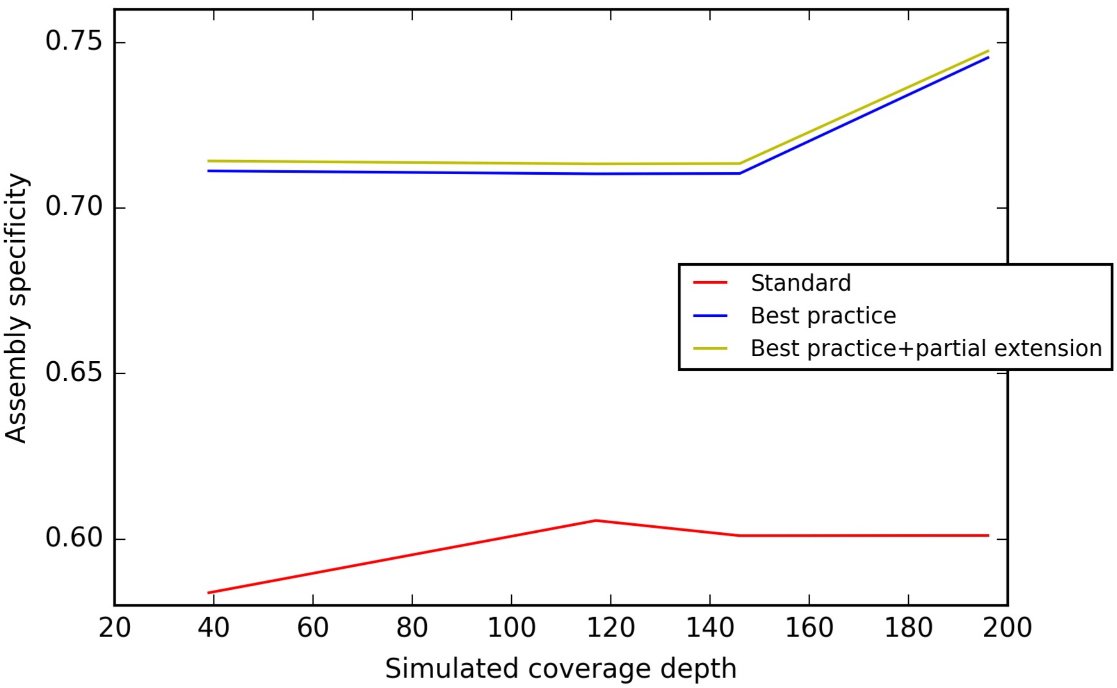


Figure 2.5. Specificity plot. The combination of simulation pipeline and partial extension pipeline has the highest assembly specificity across different sequence coverage depth.

CHAPTER 3

RAPID EXPANSION OF THE PROTEIN DISULFIDE ISOMERASE GENE FAMILY FACILITATES THE FOLDING OF VENOM PEPTIDES

The following chapter is reprinted with permission from the Proceedings of National Academy of Sciences (PNAS). Safavi-Hemami, H., Q. Li, R. L. Jackson, A. S. Song, W. Boomsma *et al.*, 2016a Rapid expansion of the protein disulfide isomerase gene family facilitates the folding of venom peptides. Proceedings of the National Academy of Sciences of the United States of America 113**:** 3227-3232.

Rapid expansion of the protein disulfide isomerase gene family facilitates the folding of venom peptides

Helena Safavi-Hemami[a,b,1], Qing Li[c], Ronneshia L. Jackson[a], Albert S. Song[a], Wouter Boomsma[b], Pradip K. Bandyopadhyay[a], Christian W. Gruber[d,e], Anthony W. Purcell[f,g], Mark Yandell[c,h], Baldomero M. Olivera[a,1], and Lars Ellgaard[b]

[a]Department of Biology, University of Utah, Salt Lake City, UT 84112; [b]Department of Biology, University of Copenhagen, DK-2200 Copenhagen, Denmark; [c]Eccles institute of Human Genetics, University of Utah, Salt Lake City, UT 84112; [d]Center for Physiology and Pharmacology, Medical University of Vienna, A-1090, Vienna, Austria; [e]School of Biomedical Sciences, University of Queensland, St. Lucia, Queensland 4072, Australia; [f]Infection and Immunity Program, Monash Biomedicine Discovery Institute, Monash University, Clayton, Victoria 3800, Australia; [g]Department of Biochemistry and Molecular Biology, Monash University, Clayton, Victoria 3800, Australia; and [h]Utah Science Technology and Research Center for Genetic Discovery, University of Utah, Salt Lake City, UT 84112

Contributed by Baldomero M. Olivera, February 17, 2016 (sent for review November 30, 2015; reviewed by Johannes M. Herrmann and Maido Remm)

Formation of correct disulfide bonds in the endoplasmic reticulum is a crucial step for folding proteins destined for secretion. Protein disulfide isomerases (PDIs) play a central role in this process. We report a previously unidentified, hypervariable family of PDIs that represents the most diverse gene family of oxidoreductases described in a single genus to date. These enzymes are highly expressed specifically in the venom glands of predatory cone snails, animals that synthesize a remarkably diverse set of cysteine-rich peptide toxins (conotoxins). Enzymes in this PDI family, termed conotoxin-specific PDIs, significantly and differentially accelerate the kinetics of disulfide-bond formation of several conotoxins. Our results are consistent with a unique biological scenario associated with protein folding: The diversification of a family of foldases can be correlated with the rapid evolution of an unprecedented diversity of disulfide-rich structural domains expressed by venomous marine snails in the superfamily Conoidea.

protein disulfide isomerase | peptide folding | gene expansion | cone snail venom | conotoxins

Formation of correct disulfide bonds is essential for the structural stability and functional integrity of many secreted proteins and peptides, such as protease inhibitors, hormones, antimicrobial peptides, and toxins from venoms. Recent advances in nucleotide and protein sequencing have revealed that toxins from marine snails in the superfamily Conoidea, such as cone snails (*Conus*), comprise a remarkable diversity of cysteine-rich polypeptides (1, 2). Toxin expression and folding takes place in the endoplasmic reticulum (ER) of venom gland cells, where, at any given time, hundreds of distinct cysteine-rich peptides are properly folded and secreted in preparation for envenomation (3, 4). No other tissue type has been reported to produce such a high density and diversity of cysteine-rich peptides. Although a fraction of these peptides contain domains that are widely distributed in the animal and plant kingdom [e.g., the inhibitor cysteine knot (5), kunitz-type domains (6), and the insulin/relaxin-like fold (7)], the majority represent unique structural domains, expressed only in conoidean venoms. How these structural scaffolds are efficiently folded in the venom gland is not well understood, but it is clear that ER-resident helper proteins must be involved. For example, members of the well-characterized O superfamily of conotoxins contain six cysteine residues that can form three disulfide bonds. Despite the fact that these peptides could potentially adopt 15 different disulfide-bonded scaffolds, only one native fold is commonly found in cone snail venom (8). Conversely, in vitro folding of these toxins commonly results in low folding yields, as well as accumulation of misfolded or aggregated products (8), highlighting the need for a better understanding of the molecular processes guiding in vivo disulfide-bond formation.

Several common ER-resident foldases—including peptidyl prolyl *cis–trans* isomerase (PPI) (9), and protein disulfide isomerase (PDI) (10)—have been shown to assist in the oxidative folding of conotoxins. Whether specialized adaptations in the venom gland oxidative folding machinery have evolved to enable the folding of such a remarkably diverse set of cysteine-rich peptides has not been addressed.

Here, a systematic interrogation of 17 cone snail venom gland transcriptomes led to the identification and subsequent characterization of a large, previously undescribed PDI gene family that likely plays a critical role in the folding of conotoxins. Comparative sequence analysis revealed that this gene family arose by gene duplication and positive selection, complementing the rapid evolution of conotoxin-encoding genes. Thus, the evolution of the conotoxin-specific PDI (csPDI) family can be regarded as a key adaptation for the high-throughput production of cysteine-rich venom peptides.

Results

New PDI Sequence from *Conus geographus* Defines the First Member of a Diverse Gene Family. Analyses of the published venom gland transcriptome of *Conus geographus* (1) identified a sequence resembling other known cone snail PDIs (e.g., ~96% identity to PDI from *Conus marmoreus*), but also revealed the presence of an additional related sequence sharing only ~67% identity to

Significance

The majority of secreted proteins contain disulfide bonds that provide structural stability in the extracellular environment. The formation of correct disulfide bonds is assisted by the enzyme protein disulfide isomerase (PDI). Most secreted structural domains are ancient and widely distributed in all metazoans; in contrast, diverse sets of unique disulfide-rich structural domains have more recently evolved in venomous marine snails (superfamily Conoidea comprising >10,000 species). We have discovered a previously undescribed gene family encoding PDIs of unprecedented diversity. We suggest that these enzymes constitute an important part of the supporting molecular infrastructure required for properly folding the plethora of structural domains expressed in the venoms of snails in different conoidean lineages.

Author contributions: H.S.-H., A.W.P., B.M.O., and L.E. designed research; H.S.-H., R.L.J., A.S.S., and P.K.B. performed research; C.W.G., A.W.P., and M.Y. contributed new reagents/analytic tools; H.S.-H., Q.L., and W.B. analyzed data; and H.S.-H., B.M.O., and L.E. wrote the paper.

Reviewers: J.M.H., University of Kaiserslautern; and M.R., University of Tartu.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. KT874559–KT874574).

[1]To whom correspondence may be addressed. Email: helena.safavi@utah.edu or olivera@biology.utah.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1525790113/-/DCSupplemental.

PDI from other cone snail species (Fig. 1). This previously undescribed sequence represented, to our knowledge, the first member of the csPDI gene family. Like canonical PDI, csPDI consists of four thioredoxin-like domains: an **a** and **a′** domain containing the active site CGHC motif and noncatalytic **b** and **b′** domain. Further transcriptome mining identified two variants of csPDI in the transcriptome dataset. Reverse-transcription PCR, cloning, and Sanger sequencing of *C. geographus* venom gland cDNA confirmed these variants and led to the identification of two additional csPDI sequences, previously undetected in the RNA sequencing (RNA-Seq) dataset.

Thus, a total of five distinct csPDI sequences that share 87–97% identity with each other and 61–65% identity with canonical PDI were retrieved (*SI Appendix*, Table S1). Notably, several of the variable amino acid residues were located between the conserved cysteine residues in the active-site motif (CXXC). These amino acids are known to affect the redox state of PDI and therefore the ability of the enzyme to form, reduce, and isomerize disulfide bonds in client proteins (11). Although canonical PDI in other organisms, including *C. geographus*, contains a glycine followed by histidine (CGHC), *C. geographus* csPDIs have a diverse set of residues in both catalytic domains: CGAC and CDAC in the **a** domain and CGLC and CEFC in the **a′** domain (*SI Appendix*, Fig. S1). How these changes may affect the oxidoreductase activity of these enzymes, especially in respect to conotoxin folding, is discussed further below.

csPDIs Are Hypervariable and Ubiquitously Expressed in the Venom Gland of Cone Snails. To determine whether csPDIs are expressed in the venom glands of other cone snail species, the venom gland transcriptomes of 15 additional cone snails were obtained. The published transcriptome of *Conus victoriae* (2) was also examined, and a close relative, *Conus textile*, was examined by RT-PCR (see *SI Appendix*, Table S2, for all species used in this study). All species examined were found to express csPDIs, demonstrating the importance and wide distribution of this protein family in the genus *Conus*. Sequences shared between 76.2% and 98.2% identify, with no exact matches even between very closely related species (e.g., *Conus praecellens* and *Conus andremenezi*).

Unlike the sequence diversity observed for *C. geographus* csPDIs, only one sequence per species was retrieved from assembled datasets, suggesting that *C. geographus* csPDIs were exceptionally diverse or that the true diversity of csPDI sequences was being missed using standard RNA-Seq assembly protocols. To better investigate the diversity of csPDIs, we applied a recently developed tool for next-generation sequencing read classification called Taxonomer (ref. 12; see *SI Appendix*, *SI Materials and Methods* for more details). Taxonomer specifically identified all RNA-Seq reads derived from the csPDI gene family before data assembly, thus enabling faster and more accurate assemblies of highly similar sequences. Taxonomer identified an average of 2.6 csPDI sequences per species, confirming the expansion of the csPDI family in cone snails. In total, 43 unique full-length and 4 partial csPDI sequences were identified from 18 species. Applying the same methodology, only one canonical PDI was identified per species with the exception of *Conus textile*, which expressed two distinct variants of PDI, a finding that has been reported (13).

Phylogenetic analysis clearly resolved the PDI and csPDI gene families and revealed that these enzymes have evolved by duplication from an ancestral gene (Fig. 2, black arrow). csPDI sequences resolved into two groups that correlate with the "primitive" and "complex" group of cone snails, previously described based on mitochondrial phylogenetics (14). However, within these groups, csPDI sequences from the same species do not group together,

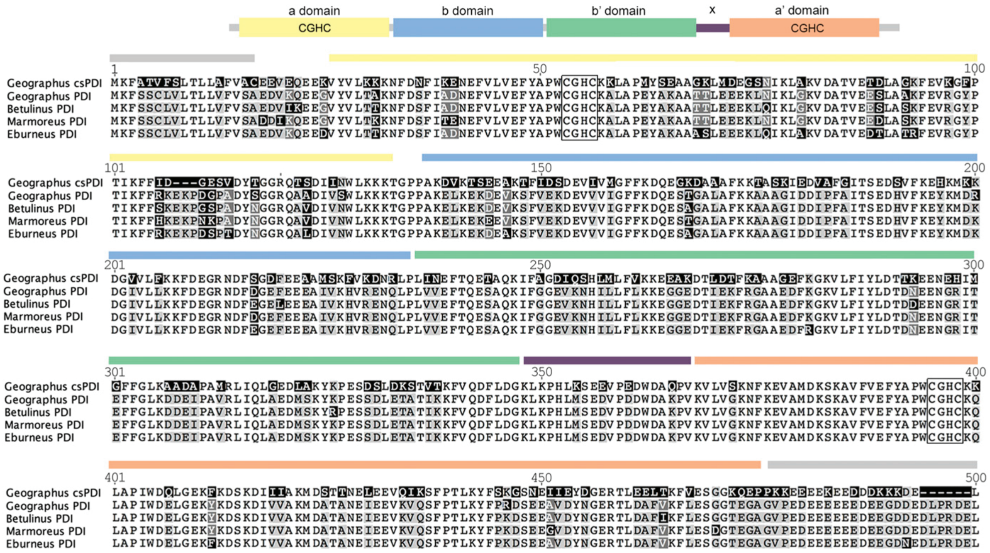


Fig. 1. Identification of a previously unidentified PDI sequence (csPDI) in the venom gland of *C. geographus* (GenBank accession no. KT874567). Multiple sequence alignment with canonical PDI from the same species (GenBank accession no. KT874559) and three additional species [*C. betulinus* (ADZ76593), *C. marmoreus* (ABF48564), and *C. eburneus* (ADZ76591)] identifies regions of divergence between csPDI and PDIs (white, 100% identity; light gray, 99–80% identity; dark gray, 80–60% identity; black: <60% identity). The alignment was performed in Geneious by using the Blosum62 similarity option for coloring (Version 8.1.3). csPDI and PDI sequences share 65–66% identity. Thioredoxin domain organization is depicted above the sequences and was predicted by using known boundaries for human PDI (31). Signal sequences (gray bar above sequence) were predicted by using InterProScan (32). The C terminus containing ER-retention motifs is also depicted with a gray bar. Active site CGHC motifs are boxed.

suggesting that the evolution of csPDIs is more closely tied to molecular function than phylogenetic relationships.

Similar to observations made for *C. geographus*, several csPDI sequences contained unusual variations in the active-site motifs. The most prominent difference was the replacement of a histidine for alanine in the **a** domain, a motif that was found in 13 of 18 species. Phylogenetic analysis suggested that this mutation evolved several times within the csPDI family (Fig. 2). Additional variations included CGIC in the **a** domain and CAHC and CEKC in the **a**′ domain. Two partial sequences retrieved from the *C. victoriae* transcriptome (2) contained CEFC and CRPC variations in the **a**′ domain. Hereafter, the two amino acids located between the active-site cysteines will be provided as superscript letters—e.g., csPDI$_{GA/GH}$, where the first two letters (GA) represent residues found in the **a** domain and the last two (GH) residues in the **a**′ domain of the enzyme. BLAST searches could not retrieve a gene resembling the csPDI family from any other organism in the NCBI nonredundant protein and nucleotide collection, suggesting that the csPDI family specifically evolved in the genus *Conus*. Several other members of the PDI family were identified in transcriptome datasets (e.g., PDIA3 and PDIA5). Comparative phylogenetic analysis of these and *Conus* PDIs and csPDIs illustrated that the csPDIs are more closely related to PDI than to other PDI family members (*SI Appendix*, Fig. S2).

PDI and csPDIs Are Highly Expressed in the *Conus* Venom Gland and Among the Most Abundant Proteins. Analysis of transcriptomic data highlighted that PDI and csPDIs are highly expressed in the venom glands of all cone snail species, ranging from 0.03% to 0.27% of all sequenced reads. Expression ratios for csPDI/PDI ranged from 0.4 to 2.2 (mean ratio: 1.2), demonstrating that the csPDI family has similar expression values to PDI (*SI Appendix*, Fig. S3). Furthermore, csPDIs are preferentially expressed in the venom gland with very low expression levels in other tissues, as determined by quantitative real-time PCR (qPCR) and RNA-Seq analysis on the foot, venom bulb (a venom "pump" located at the inner end of the venom gland), esophagus, nerve ring, salivary gland, and venom gland of *C. geographus* and *C. rolani,* supporting a specialized role of the csPDI family in conotoxin folding (*SI Appendix*, Fig. S4; the esophagus and nerve ring were not available for *C. rolani*).

To investigate relative abundances of these proteins in the venom gland, the published proteome of *C. geographus* was revisited (15). Separation of venom gland proteins by 2D gel electrophoresis resolved two distinct gel areas that were identified as different isoforms of PDI in the original study (15). Reanalysis of mass spectrometric data by using a proteogenomic strategy revealed that these areas correspond to PDI and csPDIs (Fig. 3). Migration patterns are consistent with differences in the predicted isoelectric point (pI) for PDI (pI: 4.6) and members of the csPDI family (pI: 4.8–5.0).

Although gel analysis could not resolve individual csPDI members, matching of tryptic peptides obtained by mass spectrometry to csPDI sequences identified between one and nine unique peptides corresponding to each csPDI sequence (Fig. 3 and *SI Appendix*, Fig. S5). This finding strongly suggests that all *C. geographus* csPDI variants are translated into functional proteins.

The intensity of gel spots identified as PDI and csPDIs highlights that these enzymes are clearly among the most abundant soluble proteins present in the venom gland of *C. geographus* (Fig. 3).

csPDI Family Is Rapidly Evolving with High Sequence Variability at Functionally Important Sites. Several distinct csPDI sequences were identified for each cone snail species. This diversification suggests that the csPDI family is expanding and subject to strong positive selection. Additionally, comparative sequence alignments and phylogenetic analyses suggested that the genetic variability was higher for members of the csPDI gene family than for PDI-encoding genes. Evolutionary pressures can be quantified by the rates of substitutions at silent sites (dS), which are presumed neutral, relative to the rate of substitutions at nonsilent sites (dN), which possibly experience selection. To investigate whether the csPDI family contained sites that experience high positive selection rates, the mixed effects model of

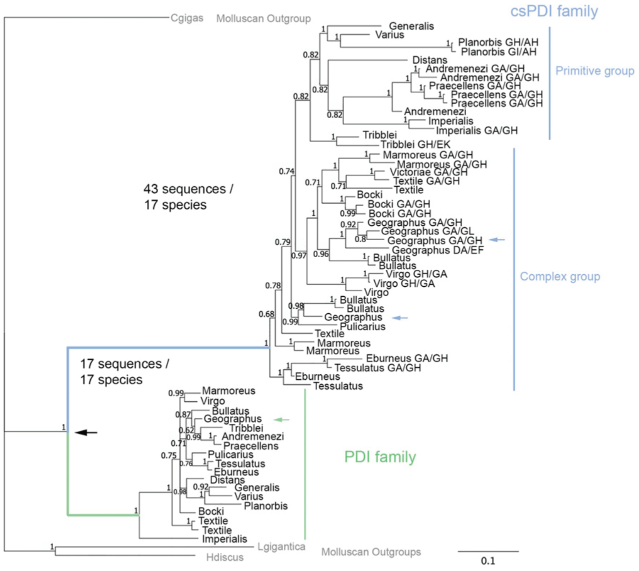
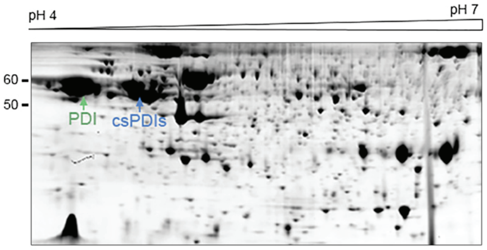


Fig. 2. Phylogenetic analysis of full-length PDI and csPDI protein sequences supports the presence of two gene families originating from an ancestral gene duplication event (black arrow, posterior probability: 1). Diversity and genetic variance for the csPDI family are apparent by more than one distinct csPDI sequence per species and long branch lengths, respectively. Two groups (the primitive and complex groups) resolved within the csPDI branch and correlate with molecular phylogenetics analyses (14). Amino acids of the active site motif are provided for sequences with divergent active sites. Arrows indicate *C. geographus* sequences selected for subsequent functional characterization. Bayesian tree was constructed using MrBayes (Version 3.2.2; ref. 33) with two runs each of four Markov chains sampling every 200 generations. The log-likelihood score stabilized after 1,100,000 generations. The consensus tree was calculated after omitting the first 25% of the samples as burn-in.



	MW	pI	Peptide Matches	Unique Matches
PDI	54034	4.6	26	17
csPDI GH/GH	53869	5.0	30	9
csPDI GA/GH	54023	4.9	23	2
csPDI GA/GH 2	53778	4.9	31	5
csPDI GA/GL	53894	4.8	21	2
csPDI DA/EF	54123	4.9	25	1

Fig. 3. Analysis of the venom gland proteome of *C. geographus* shows high abundances for PDI and csPDIs as determined by 2D gel electrophoresis and subsequent mass spectrometric protein identification. Data deposited in the original study (15) were revisited and examined for mass spectrometric peptide hits that matched PDI and csPDI sequences obtained in the present study. Gel spots identified as PDI and csPDIs are depicted and correlate with predicted molecular weights (MW) and isoelectric points (pI). The number of total and unique peptide matches obtained for PDI and different members of the csPDI family are provided (score > 99 using Protein Pilot; Version 3.0; AB SCIEX). Sequences and position of matched peptides onto the full-length sequences are provided in *SI Appendix*, Fig. S5. Reproduced from ref. 15, copyright the American Society for Biochemistry and Molecular Biology.

evolution (MEME) implemented in datamonkey (16) was used. MEME analysis revealed a total of 12 and 35 sites ($P < 0.1$), with positive selection for PDI and csPDI families, respectively, demonstrating that the csPDI family experiences higher selection rates than PDI. Interestingly, for the PDI family, episodes of positive selection were absent in the **b**′ domain, a domain known to be important for substrate binding (17), whereas the csPDI family displayed 13 positive selection events in this domain (*SI Appendix*, Fig. S6). To graphically illustrate protein sequence variation for PDIs and csPDIs, a sequence variation score was generated based on multiple sequence alignments for the two enzyme families (see *SI Appendix*: a sequence logo representation for the two multiple sequence alignments compared with the sequence of human PDI is provided in *SI Appendix*, Fig. S7). This score was subsequently converted into a red-white color range, where darker color represents more sequence variation, and mapped onto the crystal structure of full-length human PDI (18) (Fig. 4 *A* and *B*). Modeling revealed that PDIs and csPDIs show widely distributed sequence variations in the **a** and **b** domains. Notably, the PDIs show only very moderate variation in the **b**′ and **a**′ domains, compared with csPDIs. This difference is most pronounced for the **b**′ domain, where by far most residues are strictly conserved among PDIs, but vary considerably in the csPDIs (Fig. 4*C*). This finding is consistent with the MEME analysis as discussed above. In human PDI, the **b**′ domain harbors a hydrophobic patch known to bind substrates directly (17). In addition, we noticed that two equivalent positions in the **a** and **a**′ domains, located two residues C-terminal of the second cysteine of the CXXC active-site motif, show high sequence variation that is restricted to the csPDIs (arrows in Fig. 4*B*). Overall, csPDIs show higher sequence variation compared with the PDIs on key positions implicated in substrate binding and found at or in close proximity to the active site of both redox active domains.

C. geographus **csPDIs Have Distinct Effects on the Folding Kinetics of Conotoxin Substrates.** To determine whether csPDIs can assist in the folding of conotoxins, two csPDI family members from *C. geographus* containing the two most widely distributed active site motifs (csPDI$_{GH/GH}$ and csPDI$_{GA/GH}$) were expressed for oxidative folding studies (*SI Appendix*, Fig. S8). PDI was analyzed for comparison. Three O-superfamily conotoxins containing six cysteine residues were selected for oxidative folding studies based on their distinct folding characteristics (Fig. 5). Omega-GVIA, from the venom of *C. geographus*, folds rapidly with very little misfolded byproducts (19). PDI and both csPDIs significantly accelerated the folding of ω-GVIA compared with no-enzyme control reactions (half-time $t_{1/2}$ for accumulation of native product: 42.7 min). Folding was fastest in the presence of csPDI$_{GH/GH}$ (7.1 min) followed by csPDI$_{GA/GH}$ (8.7 min) and PDI (19.4 min) (Fig. 5*A*). The folding of μ-SmIIIA, a peptide with faster folding kinetics than ω-GVIA (20), was significantly accelerated only in the presence of csPDI$_{GH/GH}$ and csPDI$_{GA/GH}$ (Fig. 5*B*). PDI had no significant effect. Most remarkably, when the recombinant enzymes were tested on δ-PVIA, a member of the delta conotoxin family that is characterized by very slow in vitro folding kinetics (21), folding was accelerated by a factor of ∼32 in the presence of csPDI$_{GH/GH}$ ($t_{1/2} = 9.7$ min) compared with no-enzyme controls ($t_{1/2} = 315.8$ min; Fig. 5*C*). PDI was slightly less efficient ($t_{1/2} = 20.6$ min) followed by csPDI$_{GA/GH}$

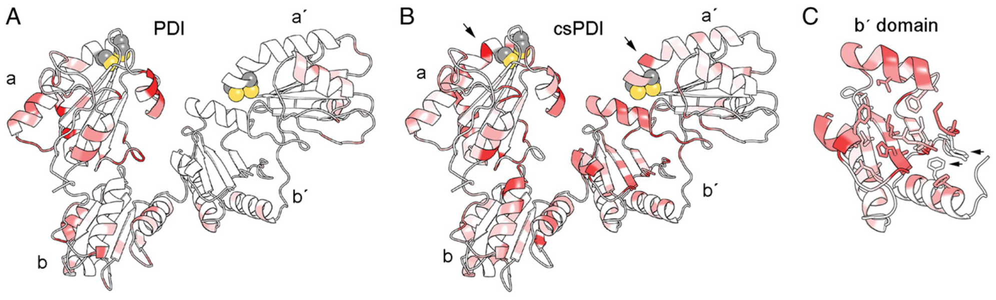


Fig. 4. (*A* and *B*) Sequence variation in cone snail PDIs (*A*) and csPDIs (*B*) mapped onto a representation of the crystal structure of full-length human PDI (Protein Data Bank ID code 4EKZ). Multiple sequence alignments of PDIs and csPDIs were used to assign a variation score for each position in the alignment. This score was then converted to a red–white color range, where darker color indicates higher sequence variation. Heavy atoms of active-site cysteines are depicted as space-filling models (gray, C; yellow, S). The four thioredoxin-like domains are indicated, and arrows point to the +2 position C-terminal of the CXXC motifs in the **a** and **a**′ domains of the csPDIs, which shows sequence variation only in this group of enzymes and not in the PDIs (see text for details). (*C*) Residues of the hydrophobic patch of the **b**′ domain are shown as stick models. Sequence variation is apparent in all but two of these residues (arrows).

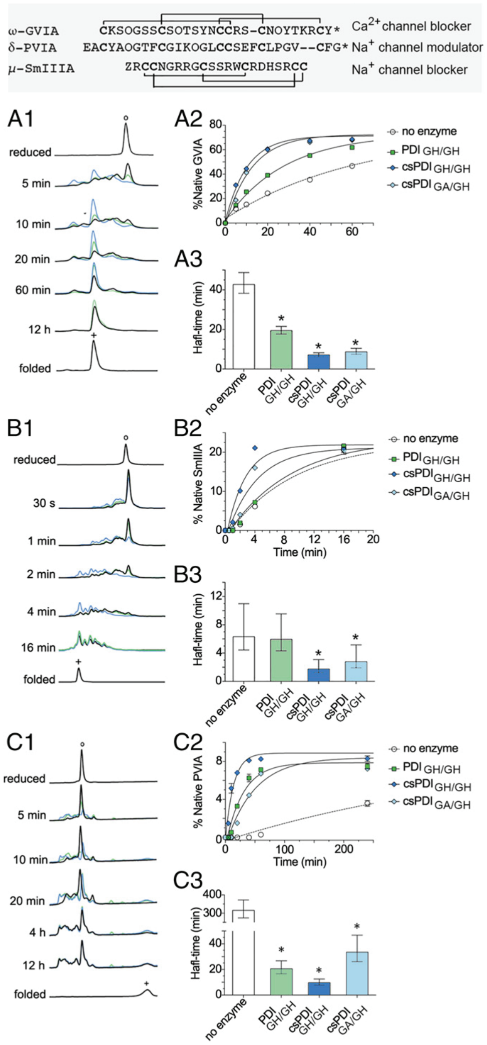


Fig. 5. Oxidative folding of conotoxin substrates in the presence of PDI and two members of the csPDI family from *C. geographus*. Sequences of the three conotoxin substrates tested are shown with their names, molecular targets, and disulfide connectivities. Amino acids: Z, pyroglutamate; O, hydroxyproline; *C-terminal amidation. (*A* and *B*) Folding assays for ω-GVIA (*A*) and μ-SmIIIA (*B*) were carried out at room temperature in the absence and presence of 2 μM enzyme in 100 mM Tris-HCl (pH 7.5), 1 mM EDTA, 0.4 mM reduced glutathione, and 0.2 mM oxidized glutathione. (*C*) Folding of δ-PVIA was performed at 4 °C and in the presence of 1% Tween-20. Reactions were initiated by adding 20 μM reduced toxin, quenched at different time points with formic acid (final 10% vol/vol), and analyzed by reverse-phase chromatography on a C_{18} column. Chromatograms of reactions without enzyme (black), with PDI (green), and with $\text{csPDI}_{GH/GH}$ (blue) are shown in *A1*, *B1*, and *C1*, respectively. The area under the curve was determined for the native, fully folded toxin and plotted against the reaction time (*A2*, *B2*, and *C2*). Half-time for the appearance of folded toxins (95% confidence values) was calculated in Prism (Version 6.0e; GraphPad) and is shown in *A3*, *B3*, and *C3*. Reactions that were significantly different from no-enzyme controls are indicated. *$P < 0.01$ (unpaired *t* test with Welch's correction).

($t_{1/2}$ = 33.5 min). Together, folding studies demonstrate that csPDIs are highly efficient in accelerating conotoxin folding and have distinct effects on the kinetics of disulfide-bond formation.

Discussion

Key evolutionary events can induce a rapid expansion and diversification of gene families to promote fitness and survival. An example is the parasitic liver fluke *Fasciola hepatica.* Cathepsin peptidases are important for the migration of the parasite thorough host tissue. The cathepsins in *F. hepatica* have greatly expanded and diverged to form multigenic families (22). These presumably play an important role at the host–parasite interface. The vast expansion of cathepsins was suggested to contribute to the high evolutionary potential of *F. hepatica* for infecting novel hosts and adapting to changes in the environment (22).

Similar observations have been made for venomous cone snails, in which a rapid expansion of multigene toxin families has facilitated exceptional rates of species diversification (14, 23). The molecular mechanisms behind the accelerated evolution of conotoxin-encoding genes are not fully understood, but high rates of gene duplication and positive selection have been repeatedly proposed (24, 25). Conotoxins are disulfide-rich peptides that highly selectively target a specific receptor or ion channel expressed in the nervous system of their prey, predators, or competitors. A conotoxin gene duplication could lead to advantageous neofunctionalization in one of the copies, which might act directly at the predator–prey interface, for which positive selection could be extremely high (26).

Here, we report on the expansion of a gene family involved in oxidative folding, a crucial step in conotoxin biosynthesis. Phylogenetic analysis revealed that this gene family evolved by gene duplication of an ancestral PDI gene. In humans, PDI is highly abundant and expressed in nearly all tissues types, where it serves in the formation, reduction, and isomerization of disulfide bonds (27). To date, no viable PDI knockout mouse has been reported demonstrating a crucial role of this enzyme in survival (28). The initial duplication of the PDI gene presumably allowed neofunctionalizations of the new PDI gene copy in the *Conus* venom gland, while maintaining the fundamental enzymatic properties of canonical PDI. As suggested by the presence of multiple csPDI variants in almost all cone snail species examined, the initial generation of the csPDI gene was followed by additional duplication events accompanied by high mutation rates that resulted in further gene specializations. Thus, csPDI expansion and diversification complemented the evolution of their conotoxin substrates, implying a rapidly changing need for oxidative folding of newly evolved disulfide-rich structural domains. A specialized function of the csPDIs in conotoxin folding is further supported by the finding that csPDIs are found in relatively very high abundance in the venom gland compared with other tissues (*SI Appendix*, Fig. S4).

Despite thousands of different conotoxin sequences, only a limited number of disulfide scaffolds are found in vivo, a phenomenon that has been referred to as the conotoxin folding puzzle (8). Conotoxins that significantly differ in their amino acid sequence efficiently adopt the same structural fold. However, in vitro, even toxins that contain the same cysteine framework often display an array of different folding properties and commonly adopt nonnative structures (8). To our knowledge, the csPDI family provides the first insight into addressing this biological conundrum. By guiding the folding of conotoxins into their native structural fold, csPDIs may eliminate the effects of extensive sequence variations observed in these peptide substrates.

Variation within the csPDI family is specifically found in regions that play an important role for enzyme activity and substrate binding. The greatest diversity was observed in *C. geographus*: Four of the five csPDI enzymes had mutations in the two amino acids located between the active-site cysteine residues (CXXC) of the **a** and **a′** domain. Comparative alignment of all csPDI sequences further detected a conspicuous sequence variation on the +2 position C-terminal of the CXXC active-site motif in both **a** and **a′** domains. We are not aware of any systematic investigation of the potential functional consequence of mutating residues at this position in redox-active thioredoxin-like domains. Still, the close proximity to the active site could well indicate an influence of the residue at this position in modulating the active-site reduction potential and thereby the redox activity of the given enzyme. If so, the csPDIs could use sequence variation at this position to modulate their redox activity to assist the folding of specific conotoxins. In contrast, the active site motif of canonical PDI was conserved in all species, pointing to a more constrained role of this enzyme.

Analysis of position-specific sequence variations demonstrated that the **b′** domain showed pronounced variation in the csPDIs, but remains highly conserved for PDI (Fig. 4 and *SI Appendix*, Fig. S7). In human PDI, a hydrophobic patch in **b′** is important for domain–domain interactions between **b′** and **a′** and for binding substrates directly (29–31). This hydrophobic patch is clearly conserved in the *Conus* PDIs. Notably, many of the residues of the hydrophobic patch show sequence variation in the csPDIs (Fig. 4*C*). Despite sequence variation, the hydrophobic nature of this patch is kept intact in the csPDIs. Thus, we speculate that this region is also involved in substrate binding in csPDIs, but may have evolved to accommodate a more diverse set of substrate peptides.

Functional characterization was carried out with two csPDI variants from *C. geographus* that had active site motifs found in almost all other cone snail species: csPDI$_{GH/GH}$ and csPDI$_{GA/GH}$. Oxidative folding assays using several conotoxin substrates confirmed that these enzymes are highly efficient in accelerating conotoxin folding and showed distinct effects on the kinetics of disulfide bond formation compared with PDI.

In conclusion, the discovery and subsequent characterization of the csPDI gene family represents an evolutionary showcase for the dynamic interplay between enzymes and their hypervariable substrates and provide important insight into the complex folding machinery evolved in conoidean venoms.

Materials and Methods

Detailed material and methods are provided in *SI Appendix*, *SI Materials and Methods*. Briefly, transcriptomes were sequenced on an Illumina HiSeq instrument, assembled by using Trinity software, and annotated by BLASTx. Additional csPDI sequence variants were discovered by using the recently developed software tool Taxonomer. Sequences were confirmed from several species by RT-PCR. The 2D gel electrophoresis coupled with mass spectrometric analysis confirmed the presence of PDI and csPDIs in the venom gland of *C. geographus*. qPCR and RNA-Seq on different cone snail tissues demonstrated high expression levels of the csPDI family in the venom gland. Recombinant *C. geographus* enzymes were expressed in *Escherichia coli* and purified by metal affinity and size-exclusion chromatography. Oxidative folding studies were carried out by using Fmoc synthesized linear conotoxins. Folding reactions were analyzed by reverse-phase chromatography.

ACKNOWLEDGMENTS. We thank Dr. William Low for MS analysis of synthetic peptides; the sequencing and genomics core facility at the University of Utah for DNA sequencing; My Thi Thao Huynh for help with image presentation; and Dr. Alexander Fedosov for dissecting different cone snail tissues. This work was supported in part by National Institutes of Health Grants GM 48677 (to B.M.O. and P.K.B.) and GM 099939 (to M.Y. and P.K.B.); and European Commission International Outgoing Fellowship Grant CONBIOS 330486 (to H.S.-H. and L.E.). A.W.P. acknowledges fellowship support from the Australian National Health and Medical Research Council. C.W.G. is an Australian Research Council Future Fellow. W.B. is supported by the Villum Foundation.

1. Hu H, Bandyopadhyay PK, Olivera BM, Yandell M (2012) Elucidation of the molecular envenomation strategy of the cone snail *Conus geographus* through transcriptome sequencing of its venom duct. *BMC Genomics* 13(284):284.
2. Robinson SD, et al. (2014) Diversity of conotoxin gene superfamilies in the venomous snail, *Conus victoriae*. *PLoS One* 9(2):e87648.
3. Lu A, Yang L, Xu S, Wang C (2014) Various conotoxin diversifications revealed by a venomic study of *Conus flavidus*. *Mol Cell Proteomics* 13(1):105–118.
4. Tayo LL, Lu B, Cruz LJ, Yates JR, 3rd (2010) Proteomic analysis provides insights on venom processing in *Conus textile*. *J Proteome Res* 9(5):2292–2301.
5. Norton RS, Pallaghy PK (1998) The cystine knot structure of ion channel toxins and related polypeptides. *Toxicon* 36(11):1573–1583.
6. Bayrhuber M, et al. (2005) Conkunitzin-S1 is the first member of a new Kunitz-type neurotoxin family. Structural and functional characterization. *J Biol Chem* 280(25): 23766–23770.
7. Safavi-Hemami H, et al. (2015) Specialized insulin is used for chemical warfare by fish-hunting cone snails. *Proc Natl Acad Sci USA* 112(6):1743–1748.
8. Bulaj G, Olivera BM (2008) Folding of conotoxins: Formation of the native disulfide bridges during chemical synthesis and biosynthesis of *Conus* peptides. *Antioxid Redox Signal* 10(1):141–155.
9. Safavi-Hemami H, Bulaj G, Olivera BM, Williamson NA, Purcell AW (2010) Identification of *Conus* peptidylprolyl cis-trans isomerases (PPIases) and assessment of their role in the oxidative folding of conotoxins. *J Biol Chem* 285(17):12735–12746.
10. Safavi-Hemami H, et al. (2012) Modulation of conotoxin structure and function is achieved through a multienzyme complex in the venom glands of cone snails. *J Biol Chem* 287(41):34288–34303.
11. Ellgaard L, Ruddock LW (2005) The human protein disulphide isomerase family: Substrate interactions and functional properties. *EMBO Rep* 6(1):28–32.
12. Graf EH, et al. (2016) Unbiased detection of respiratory viruses using RNA-seq-based metagenomics: A systematic comparison to a commercial PCR panel. *J Clin Microbiol* 03060-15.
13. Bulaj G, et al. (2003) Efficient oxidative folding of conotoxins and the radiation of venomous cone snails. *Proc Natl Acad Sci USA* 100(Suppl 2):14562–14568.
14. Puillandre N, et al. (2014) Molecular phylogeny and evolution of the cone snails (Gastropoda, Conoidea). *Mol Phylogenet Evol* 78:290–303.
15. Safavi-Hemami H, et al. (2014) Combined proteomic and transcriptomic interrogation of the venom gland of *Conus geographus* uncovers novel components and functional compartmentalization. *Mol Cell Proteomics* 13(4):938–953.
16. Murrell B, et al. (2012) Detecting individual sites subject to episodic diversifying selection. *PLoS Genet* 8(7):e1002764.
17. Pirneskoski A, et al. (2004) Molecular characterization of the principal substrate binding site of the ubiquitous folding catalyst protein disulfide isomerase. *J Biol Chem* 279(11):10374–10381.
18. Wang C, et al. (2013) Structural insights into the redox-regulated dynamic conformations of human protein disulfide isomerase. *Antioxid Redox Signal* 19(1):36–45.
19. Price-Carter M, Gray WR, Goldenberg DP (1996) Folding of omega-conotoxins. 2. Influence of precursor sequences and protein disulfide isomerase. *Biochemistry* 35(48): 15547–15557.
20. Fuller E, et al. (2005) Oxidative folding of conotoxins sharing an identical disulfide bridging framework. *FEBS J* 272(7):1727–1738.
21. DeLa Cruz R, Whitby FG, Buczek O, Bulaj G (2003) Detergent-assisted oxidative folding of delta-conotoxins. *J Pept Res* 61(4):202–212.
22. Cwiklinski K, et al. (2015) The Fasciola hepatica genome: Gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. *Genome Biol* 16:71.
23. Stanley SM (2008) Predation defeats competition on the seafloor. *Paleobiology* 34: 1–21.
24. Chang D, Duda TFJ, Jr (2012) Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Mol Biol Evol* 29(8):2019–2029.
25. Puillandre N, Watkins M, Olivera BM (2010) Evolution of Conus peptide genes: Duplication and positive selection in the A-superfamily. *J Mol Evol* 70(2):190–202.
26. Kordis D, Gubensek F (2000) Adaptive evolution of animal toxin multigene families. *Gene* 261(1):43–52.
27. Okumura M, Kadokura H, Inaba K (2015) Structures and functions of protein disulfide isomerase family members involved in proteostasis in the endoplasmic reticulum. *Free Radic Biol Med* 83:314–322.
28. Hatahet F, Ruddock LW (2009) Protein disulfide isomerase: A critical evaluation of its function in disulfide bond formation. *Antioxid Redox Signal* 11(11):2807–2850.
29. Denisov AY, et al. (2009) Solution structure of the bb′ domains of human protein disulfide isomerase. *FEBS J* 276(5):1440–1449.
30. Nguyen VD, et al. (2008) Alternative conformations of the x region of human protein disulphide-isomerase modulate exposure of the substrate binding b′ domain. *J Mol Biol* 383(5):1144–1155.
31. Wang C, et al. (2012) Human protein-disulfide isomerase is a redox-regulated chaperone activated by oxidation of domain a′. *J Biol Chem* 287(2):1139–1149.
32. Zdobnov EM, Apweiler R (2001) InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17(9):847–848.
33. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17(8):754–755.

Supporting Information

SI Material and Methods

Transcriptome sequencing and RT-PCR. Total RNA was isolated from venom ducts using the TRIzol® Reagent (Life Technologies Corporation) or RNAEasy kit (Qiagen) following the manufacturers' instructions. The transcriptomes of 16 species (see Table S1) were sequenced on an Illumina HiSeq 2000 platform (Sanger/Illumina1.9 reads, 101 bp paired-end) at Cofactor Genomics (St. Louis MO, USA). RNAseq data were de novo assembled using Trinity (1) and annotated using BLASTx. Additionally, the published venom gland transcriptomes of *C. geographus* (2), *C. tribblei* (3)and *C. victoriae* (4) were interrogated for PDI and csPDI sequences. Based on sequences identified in the assembled *C. geographus* dataset oligonucleotides were designed to confirm sequences for *C. geographus*, *C. bullatus* and *Co. bocki* and obtain additional sequences for *C. textile*, a species for which no transcriptome data was available. First strand cDNA was generated using SMARTscribe reverse transcriptase (Clontech) following the manufacturer's protocol. RT-PCR was performed using the Clontech Advantage 2 PCR Kit with the following oligonucleotide sequences (5'-3'): PDI Sense: AATCTCTCCCACGAGTTACATAG; PDI Antisense: AAGAACACAAGTGCAGTCAGAC; csPDI Sense: TACCTTTCGGCTGTTGTCTCT; csPDI Antisense: AACACAAGATGAAACTGAATTATGT. PCR was carried out for 25 cycles at an annealing temperature of 55 °C. To avoid the formation of heteroduplexes, amplicons were diluted 1:5 and subjected to three additional PCR cycles in the presence of fresh buffer, dNTPs, oligonucleotides and polymerase. PCR amplicons were gel-purified (Qiagen gel purification kit), cloned into the pGEM-T *Easy* Vector (Promega) and transformed into *E. coli* (DH10B strain). Plasmids were purified (DNA extraction kit, Viogene-Biotek Corporation) and either sequenced at the University of Utah Microarray and Genomic Analysis Core Facility or the Beckman Coulter Genomics Facility using Sanger DNA sequencing. A total of 20-100 plasmids were sequenced per species. For *C. geographus* and *C. bullatus*, purified PCR amplicons were further subjected to Illumina sequencing. Briefly, libraries were prepared using the TruSeq DNA HT Sample Prep Kit (Illumina) and sequenced on the Illumina MiSeq platform (150 cycle, paired-end sequencing) at the University of Utah Microarray and Genomic Analysis Core Facility. Sequenecs obtained by RT-PCR were submitted to GenBank under the accession number provided in Table S3. All PDI and csPDI sequences obtained in this study are provided in File S1.

RNASeq and quantitative real-time PCR on different cone snail tissues

Several different tissues (foot, venom bulb, oesophagus, nerve ring, salivary gland and venom gland) were dissected from two specimens of *C. geographus*. Total RNA was extracted using the Direct-zol RNA MiniPrep Plus (Zymo Research) following the manufacturer's instruction. An on-column DNAse treatment step was included during RNA purification. Total RNA integrity, quantity and purity were determined on a 2100 Bioanalyzer (Agilent Technologies). 300 ng per tissue of RNA was reverse transcribed using the SuperScript III First-Strand Synthesis System (Invitrogen) with a 1:1 mixture of random hexamers and oligo-dT primers. Quantitative real-time PCR (qPCR) was performed on a CFX96 instrument (Bio-Rad) with an initial enzyme activation step of 30 sec at 95°C followed by 40 cycles of denaturing (5 sec at 95°C) and annealing/extension (10 sec at 54°C). Melt curve analysis was performed after completion of the run to ensure single amplicon formation. Reactions were carried out in SsoFast EvaGreen supermix (Bio-Rad) with a cDNA template concentration of 18 ng. Oligonucleotides were designed based on a unique region in the csPDI transcript that allowed for the amplification of all csPDI family members but not PDI (Sense: TTACGCACCATGGTGTGG, Antisense: GCCACTGATGAAGAATTTG). Thioredoxin-Related Transmembrane Protein 1 (TMX1, Sense: GTCACAGAGAGTCCAGGACTT, Antisense: CTGCAGAGAGTTAGGATGGAT) and mitochondrial NADH dehydrogenase (Sense: TAATGCACCATGTCTGCAAG, Antisense: CCTGCCTGGTACTTGCTGTT) were used for normalization. All tissues with the exception of the nerve ring, that did not provide sufficient RNA yields for qPCR, were analyzed. Reactions were run in duplicate and data was analyzed using the CFX Manager software (Bio-Rad).

In addition to qPCR analysis, different tissues from one specimen of *C. geographus* (foot, venom bulb, oesophagus, nerve ring, salivary gland and venom gland) and *C. rolani* (foot, venom bulb, salivary gland and venom gland) were subjected to next-generation RNASeq. Libraries were prepared using the TruSeq Stranded mRNA Sample Prep with poly(A) selection (Illumina) and sequenced on a HiSeq instrument (Illumina, 125 Cycle Paired-End) at the University of Utah Microarray and Genomic Analysis Core Facility. Reads were mapped onto the open reading frame of PDI and csPDIs using Geneious (version 8.1.3, default medium sensitivity setting) and normalized to the total number of reads obtained for each RNASeq dataset.

Identification of additional csPDI variants by Taxonomer. RT-PCR sequencing and transcriptome data obtained by 454 pyrosequencing revealed several variants of csPDI per species. However, only one csPDI sequence could be retrieved from Illumina read datasets suggesting that the true diversity of csPDI was being missed during the assembly of short reads. In order to identify all csPDIs sequences a bioinformatics pipeline that was recently developed for the ultrafast discovery of pathogens in clinical next-generation sequencing datasets was applied (5). Briefly, a reference database was generated by extracting all PDI and csPDI sequences from assembled RNASeq files. Sequences obtained by RT-PCR were also added. The following 4-level database hierarchy was used: PDI and csPDI were categorized as different members of the PDI superfamily. All PDIs/csPDIs from different species were further categorized as isoforms of the same gene family. RNAseq reads from each species were individually run through this database and classified by k-mer matching (21-mer). Each raw read with a k-mer matching to either PDI or csPDI was assigned to the best-match level in the reference database. Read ties were reported and Taxonomer outputs were parsed to classify reads to the end of the taxonomy tree: the individual PDI/csPDI isoform. By pre-filtering only reads derived from PDI and csPDI sequences from the RNASeq dataset this approach enabled micro-assemblies of complete, high fidelity PDI/csPDI transcripts. Transcripts were de novo assembled using Trinity (1) and annotated using BLASTX. The specificity of this approach was verified by comparing assembled contigs to sequences obtained by RT-PCR (for *C. geographus*) and by re-mapping of all reads to de novo assembled contigs.

Analysis of expression levels. To calculate expression levels for PDI and csPDIs from the different transcriptome datasets raw reads were mapped onto the open reading frame of each sequence using Geneious (version 8.1.3, custom sensitivity setting: no gaps allowed, maximum mismatch per read 5%, maximum ambiguity 1). Expression ratios for PDI over csPDIs were calculated by dividing the average number of reads per PDI per species by the average number of reads per csPDI per species. Total expression values for PDI and csPDIs were expressed as the average number of reads divided by the total number of reads obtained per transcriptome.

Determining protein abundance of PDI and csPDI in the venom gland of *C. geographus*. To investigate relative abundances of PDI and csPDI the previously published proteome of *C.*

geographus was revisited (6). Briefly, as outlined in the previous study (6) proteins were extracted from the venom gland of *C. geographus* and separated by two-dimensional gel electrophoresis (2DGE). Gel spots were excised and subjected to in-gel tryptic digestion followed by MS/MS sequencing on an TripleTOF 5600 mass spectrometer (AB SCIEX). Proteins were identified using Protein Pilot MS/MS data were searched against the in-house cone snail database using Protein Pilot software (version 3.0, AB SCIEX). Peptide matches for proteins identified as PDI in the original study (score >99) were re-analyzed here by mapping all non-redundant peptide sequences onto the *C. geographus* PDI and csPDI sequences using Geneious (version 8.1.3).

Analysis of evolutionary rates. Codon sites experiencing episodes of positive selection were determined using the mixed effects model of evolution (MEME) implemented in datamonkey (7) which is based on calculating rates of non-synonymous to synonymous substitutions (dN/dS). This measure is most meaningful for distantly diverged sequences rather than within the same population (8). Thus, only the most highly expressed PDI and csPDI sequence was taken from each species. Codon-based sequence alignments were generated in Geneious (version 8.1.3) and uploaded into the datamonkey analysis tool for MEME analysis (7).

Position-specific variation scores. To illustrate sequence variations within the csPDI and PDI families their sequences were aligned using the muscle program (9) using standard settings. Subsequently, each position was assigned a variation score based on the entropy of the amino acid distribution at that location: $H = \sum_{i=1..20} -p_i \log_2 p_i$, where p_i is the normalized frequency of observing amino acid i. The entropy H is measured in bits, with 0 as its mininum value (only one type of amino acid occurs at this position), and 4.32 as its maximum value (all amino acid types are equally probable). For the structure color annotation in Fig. 4, the same procedure was used but including the sequence of the PDB entry 4EKZ (full-length human PDI in the reduced form, (10)) in the alignment as a reference sequence. The resulting variation scores were mapped to a red-white color range, using only those columns in the alignment for which the reference sequence had a non-gap entry.

Cloning, expression and purification of *C. geographus* PDI and csPDIs. Four *C. geographus* csPDI members were selected for recombinant expression based on differences in their active site motifs. PDI was expressed for comparison. Enzymes lacking the N-terminal signal sequences were cloned into the pET22b+ expression vector (Novagen). Briefly, transcripts were PCR-amplified from pGEM-T *Easy* plasmids obtained by RT-PCR and cloning as described above and ligated into pET22b+ using the NdeI (5') and XhoI (3') restriction sites (New England Biolabs*)*. The constructs, containing a C-terminal 6x His-tag, were transformed into *E. coli* (Rosetta strain, Novagen) by heat treatment. For expression of recombinant proteins, LB broth containing 100 µg/mL ampicillin was inoculated with overnight cultures and incubated at 37°C with shaking until the A_{600} spectrophotometric reading was ~ 0.6. Expression was induced by adding 0.1 mM Isopropyl-β-D-thiogalactopyranoside followed by incubation for 3 h at 25°C with shaking. Bacteria were harvested and resuspended in native lysis buffer (50 mM NaH_2PO_4, 300 mM NaCl and 10 mM imidazole, pH 8.0) containing 1x SigmaFAST protease inhibitor cocktail (EDTA-free, Sigma). Bacterial cells were lysed by probe-tip sonication. Cellular debris and insoluble protein were pelleted by centrifugation at 20,000 x *g* for 20 min and the supernatants were used for subsequent protein purifications. Recombinant proteins were purified on nickel-nitrilotriacetic acid resin (Thermo Scientific) using the gravity-flow purification method. Protein lysates were loaded onto the resin in native lysis buffer (NB: 300 mM KCl, 50 mM KH_2PO_4, pH 8.0) and nonspecifically bound proteins were removed with NB containing 5 mM imidazole followed by a second wash with NB containing 15 mM imidazole. His-tagged fusion-proteins were eluted with 250 mM imidazole in NB. Further purification and buffer exchange into 10 mM Tris-HCl, 150 mM NaCl, pH 8 was accomplished by size exclusion chromatography (Superdex 75, HiLoad 16/60, GE Healthcare) at 1 mL/min. The purified recombinant proteins were analyzed by SDS-PAGE. Protein concentrations were determined spectrophotometrically using the proteins' molar absorption coefficients. Recombinant proteins were concentrated using AMICON Ultra centrifugal filter devices (cut-off 30 kDa, Millipore) and stored at -80ºC.

Peptide synthesis. Based on their well-characterized folding properties, three conotoxins were selected for oxidative folding studies: ω-GVIA originally isolated from the venom of *C. geographus*, µ-SmIIIA isolated from the venom of *C. stercusmuscarum* and δ-PVIA, isolated from *C. purpurascens*. Peptide resins were obtained from the peptide synthesis core facility at the University of Utah, UT, USA. Peptides were cleaved from the resin by treatment with reagent K (TFA/thioanisole/ethanedithiol/water/phenol (82.5/5/2.5/5/5 by volume)) for 3.5 h, 2.5

h and 5.5 h for ω-GVIA, μ-SmIIIA and δ-PVIA, respectively. The peptides were subsequently filtered, precipitated and washed with cold methyl *tert*-butyl ether. Linear peptides were purified by RP-HPLC on a semi-preparative C_{18} column (Vydac, 5μm particle size, 10 mm x 250 mm, Grace) using a linear gradient from 5 to 50% buffer B (90% ACN/0.1%TFA) over 45 min. Buffer A was 0.1% TFA/water. Absorbance was monitored at 220 nm and 280 nm. Concentrations were determined spectrophotometrically using the peptides' molar absorption coefficient at 280 nm. Correctly folded peptides were obtained from the peptide synthesis facility at the Salk Institute, CA, USA.

Oxidative folding studies. Oxidative folding reactions were carried out in 0.1 M Tris (pH 7.5), 1 mM EDTA, 0.4 mM GSH, 0.2 mM GSSG at room temperature with the exception of δ-PVIA which was folded at 4°C in the presence of 1% Tween-20. Reactions were pre-incubated with or without 2 μM recombinant enzyme for 30 min. Folding was initiated by adding 20 μM reduced synthetic toxin. Aliquots were taken at various time points and reactions were quenched by acidification with formic acid to a final concentration of 10%. Folding reactions were analyzed by RP-HPLC on a C_{18} column (Vydac, 5 μm particle size, 4.6 mm x 250 mm, Grace) using the following gradients: For ω-GVIA the gradient was 10-35% buffer B (90% ACN/0.1% TFA) over 30 min. Buffer A was 0.1% TFA/water. For μ-SmIIIA the gradient was 5-40% buffer B over 20 min and for δ-PVIA the gradient was 25-65% buffer B over 50 min. Native peptides were distinguished from linear forms based on characteristic elution profiles (11-13), by comparing the elution profiles to native standard peptides and by mass spectrometric (MS) analyses of manually collected reversed-phase fractions (MALDI-TOF mass spectrometer, positive reflector mode, Voyager, AB SCIEX). To determine the kinetics of the reaction, the area under the curve was calculated for the fully folded peptide and plotted against the folding time (n=2 for each time point, mean ± STDEV). Half-times for the appearance of folded peptides were calculated in Prism (GraphPad) using nonlinear regression (curve fit) analysis. Values were plotted with their 95% confidence intervals. Statistical analysis was performed using two-tailed Student's t-tests with unequal variance.

Supporting Tables and Figures

Table S1. Identity matrix for translated protein sequences of PDI and csPDIs sequenced from the venom gland of *C.geographus*. Values were calculated using Geneious (version 8.1.3).

% Identity	csPDI$_{GH/GH}$	csPDI $_{GA/GH}$	csPDI $_{GA/GH\ 2}$	csPDI$_{GA/GL}$	csPDI $_{DA/EF}$	PDI$_{GH/GH}$
csPDI$_{GH/GH}$		89	91	88	87	65
csPDI$_{GA/GH}$	89		96	97	93	62
csPDI $_{GA/GH\ 2}$	91	96		96	91	63
csPDI$_{GA/GL}$	88	97	96		91	62
csPDI $_{DA/EF}$	87	93	91	91		61
PDI$_{GH/GH}$	65	62	63	62	61	

Table S2. Species examined in this study

Species Name	Subgenus	Source data	Prey Type
C.geographus	*Gastridium*	Illumina; 454 sequencing published: (2); RT-PCR	Fish
C.bullatus	*Textilia*	Illumina; RT-PCR	Fish
C.bocki	*Asprella*	Illumina, RT-PCR	Fish
C.victoriae	*Cylinder*	454 sequencing, published: (4)	Snail
C.textile	*Cylinder*	RT-PCR	Snail
C.marmoreus	*Conus*	Illumina	Snail
C.tessulatus	*Tesselliconus*	Illumina	Worm
C.eburneus	*Tesselliconus*	Illumina	Worm
C.tribblei	*Splinoconus*	Illumina; 454 sequencing published: (3)	Worm
C.praecellens	*Turriconus*	Illumina	Worm
C.andremenezi	*Turriconus*	Illumina	Worm
C.varius	*Strategoconus*	Illumina	Worm
C.generalis	*Strategoconus*	Illumina	Worm
C.planorbis	*Strategoconus*	Illumina	Worm
C.imperialis	*Stephanoconus*	Illumina	Worm
C.pulicarius	*Puncticulis*	Illumina	Worm
C.distans	*Fraterconus*	Illumina	Worm
C.virgo	*Virgiconus*	Illumina	Worm

Table S3. GenBank Accession numbers for sequences obtained by RT-PCR. All sequences obtained in this study are provided in Supporting File 1.

Species	Name	Accession Number
Conus geographus	PDI $_{\text{GH/GH}}$	KT874559
Conus geographus	csPDI $_{\text{GH/GH}}$	KT874567
Conus geographus	csPDI $_{\text{GA/GH}}$	KT874564
Conus geographus	csPDI $_{\text{GA/GH}}$ variant 2	KT874565
Conus geographus	csPDI $_{\text{GA/GL}}$	KT874566
Conus geographus	csPDI $_{\text{DA/EF}}$	KT874563
Conus bullatus	PDI $_{\text{GH/GH}}$	KT874562
Conus bullatus	csPDI $_{\text{GH/GH}}$	KT874571
Conus bullatus	csPDI $_{\text{GH/GH}}$ variant 2	KT874572
Conus bullatus	csPDI $_{\text{GH/GH}}$ variant 3	KT874573
Conus bullatus	csPDI $_{\text{GH/GH}}$ variant 4	KT874574
Conus textile	PDI $_{\text{GH/GH}}$	KT874560
Conus textile	PDI $_{\text{GH/GH}}$ variant 2	KT874561
Conus textile	csPDI $_{\text{GH/GH}}$	KT874569
Conus textile	csPDI $_{\text{GH/GH}}$ variant 2	KT874570
Conus textile	csPDI $_{\text{GA/GH}}$	KT874568

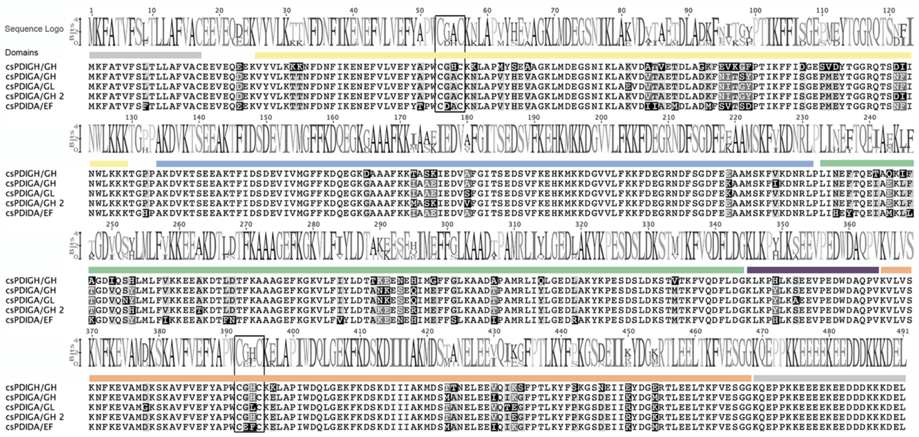


Fig. S1. Diversity of csPDI sequences in *Conus geographus*. Sequence alignment identifies regions of divergence (white: 100% similarity, light gray: 100-80% identity, dark gray: 80-60% identity, black: less than 60% identity). The alignment was performed in Geneious using the Blosum62 similarity option for coloring (version 8.1.3). Sequences share 87-97% identity. Sequence logo (generated in Geneious) and domains are shown above the alignment (domain boundaries were predicted based on human PDI (14)). The C-terminal tail containing ER-retention motitfs are also depicted with a gray bar. Active site CGHC motifs are boxed. Active site CXXC motifs are boxed.

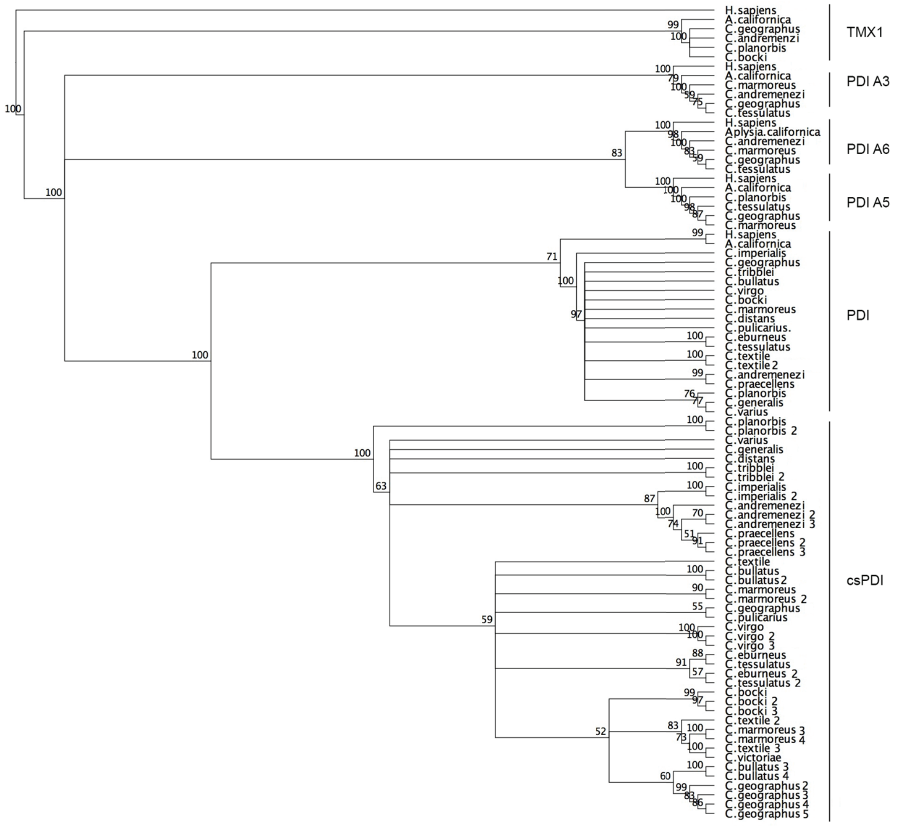


Fig. S2. Phylogenetic analysis of different members of the PDI gene family (TMX1, PDI A3, PDI A5, PDI A6, PDI and csPDI) from cone snails, the mollusc *Aplysia californica* and human resolves the PDI and csPDI family and illustrates close relatedness of the two families when compared to other members of the PDI protein family. Trees were reconstructed using Neighbour-joining analyses (Genetic Distance Model: Jukes-Cantor, number of replicates: 100, Outgroup: Human TMX1) using Geneious (version 8.1.3). Consensus support values (%) are provided.

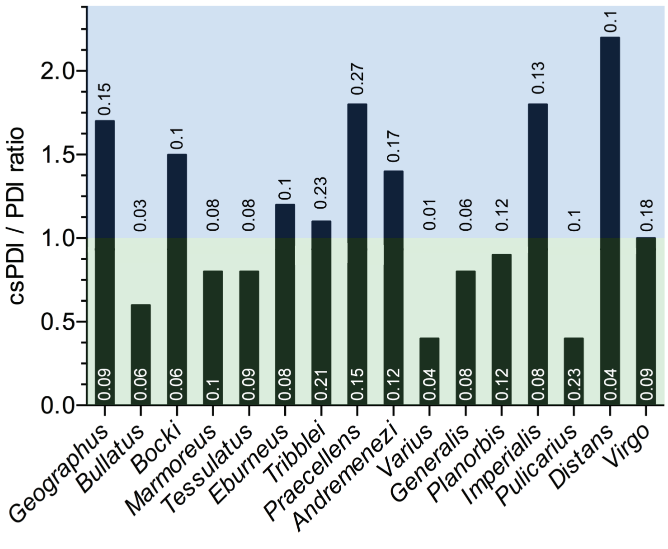


Fig. S3. Relative expression levels of PDI and csPDI in various cone snail venom glands. Bars represent expression ratios of the average read counts (csPDI reads divided by PDI reads for each species). Percentages of PDI and csPDI reads relative to the total number of reads obtained for each species are provided at the bottom and top of bars, respectively. Total read counts for individual transcriptomes ranged from ~21-98 million reads per species. Read counts for PDI and members of the csPDI family were obtained by mapping total reads onto the open reading frame of a particular sequence using Geneious software (version 8.1.3). Data from *Conus victoriae* was excluded because libraries were normalized prior to sequencing.

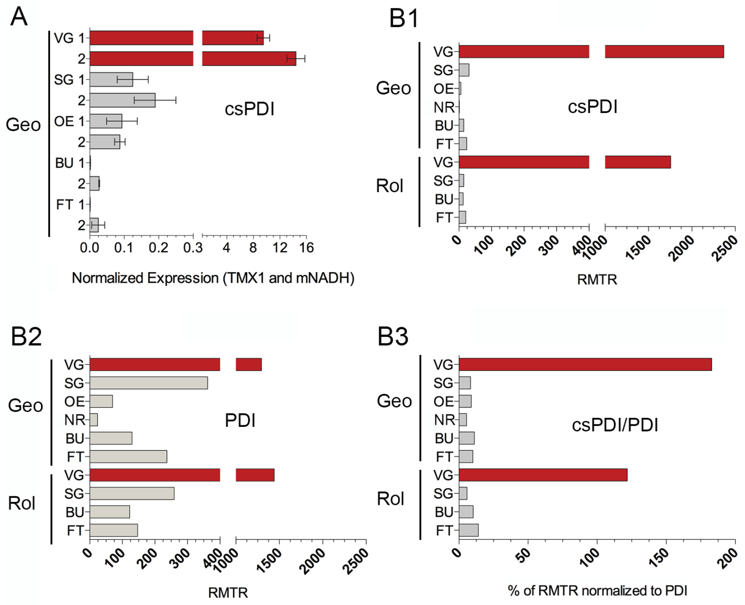
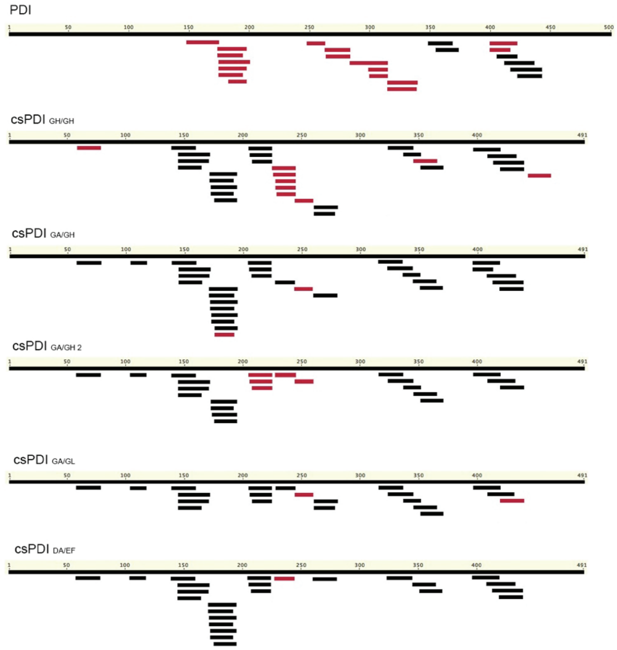


Fig. S4. Expression analysis of csPDIs in different cone snail tissues by (A) quantitative real-time PCR (qPCR) and (B1-B3) RNASeq analysis. qPCR performed on different tissue types from two specimens of *C. geographus* (n=2 specimens) demonstrated high expression levels of csPDI in the venom gland (colored in red) compared to all other tissues tested (gray). VG: venom gland, SG: salivary gland, OE: oesophagus, BU: venom bulb, FT: foot, NR: nerve ring. qPCR was performed on a CFX96 instrument (Bio-Rad) and analyzed in CFX Manager software (Bio-Rad) using TMX1 and mitochondrial NADH dehydrogenase for normalization. Each sample was run in duplicate and plotted values represent normalized mean expression levels ± SEM. csPDI expression in the venom gland of *C. geographus* was between 80 - 930 times higher in the venom gland when compared to other tissues. High expression levels of csPDI in the venom gland were also demonstrated by RNASeq analysis of different tissues from *C. geographus* and *C. rolani* (B1). Plotted values represent reads per million total reads (RMTR) obtained for each RNASeq dataset. Expression levels of PDI were also elevated in the venom gland (B2). However, PDI expression was also high in other tissues when compared to csPDI expression. This is demonstrated in B3; RMTR values of csPDI were normalized against PDI for each tissue tested. For example, RMTR values of csPDI were only 8% of values obtained for csPDIs for the salivary gland but 183 % for the venom gland of *C. geographus*.



PDI	$csPDI_{GH/GH}$	$csPDI_{GA/GH}$	$csPDI_{GA/GH\ 2}$	$csPDI_{GA/GL}$	$csPDI_{DA/EF}$
TEFFGLKDDEIPAVR	LAPMYSEAAGKLMDEGSNIK	LFTGDVQSYLMLFVK	ASKIEDVVFGITSEDSVFKEHK	MDSTANELEEVQTEGFPTLK	DNRLPLIHEYTQEIAMK
VLFIYLDTDNEENGRITEFFGLKDDEIPAVR	DNRLPLINEFTQETAQK	IEDVAFGITSEDSVFK	LFTGDVQSHLMLFVK	LFTGDVQNYLMLFVK	
KAAAGIDDIPFAITSEDHVFKEYK	NRLPLINEFTQETAQK		ASKIEDVVFGITSEDSVFK		
KAAAGIDDIPFAITSEDHVFK	VKDNRLPLINEFTQETAQK		IEDVVFGITSEDSVFKEHK		
ITEFFGLKDDEIPAVR	GSNEIIEYDGERTLEELTK		SKIEDVVFGITSEDSVFKEHK		
DELGEKYKDSKDIVVAK	IFAGDIQSHLMLFVK				
PFAITSEDHVFKEYK	DNRLPLVNEFSQETAQK				
KEGGEDTIEKFRGAAEDFKGK	FVKDNRLPLINEFTQETAQK				
IFGGEVKNHILLFLK	FVQDFLDGKLKPHLK				
AAAGIDDIPFAITSEDHVFKEYKMDK					
AAAGIDDIPFAITSEDHVFKEYK					
LIQLAEDMSKYKPESSDLETATIKK					
SFVEKDEVVVIGFFKDQESTGALAFKK					
LIQLAEDMSKYKPESSDLETATIK					
QLAPIWDELGEKYKDSKDIVVAK					
EGGEDTIEKFRGAAEDFKGK					
AAAGIDDIPFAITSEDHVFK					

Fig. S5. Mass spectrometric (MS) identification of peptides belonging to PDI and csPDIs from the venom gland of *Conus geographus*. As described in the original study (6) gel spots shown in Fig. 3 were excised and subjected to in-gel tryptic digestion followed by MS analysis on a 5600 TripleTOF mass spectrometer (5600 AB SCIEX). Data was analyzed in Protein Pilot software (version 3.0, AB SCIEX) using the *C. geographus* transcriptome database for protein/peptide identification. Here, peptide sequences identified in the original study were mapped against PDI and members of the csPDI family. Peptide matches are shown in black. Unique peptide matches are shown in red. Sequences for unique peptides are provided.

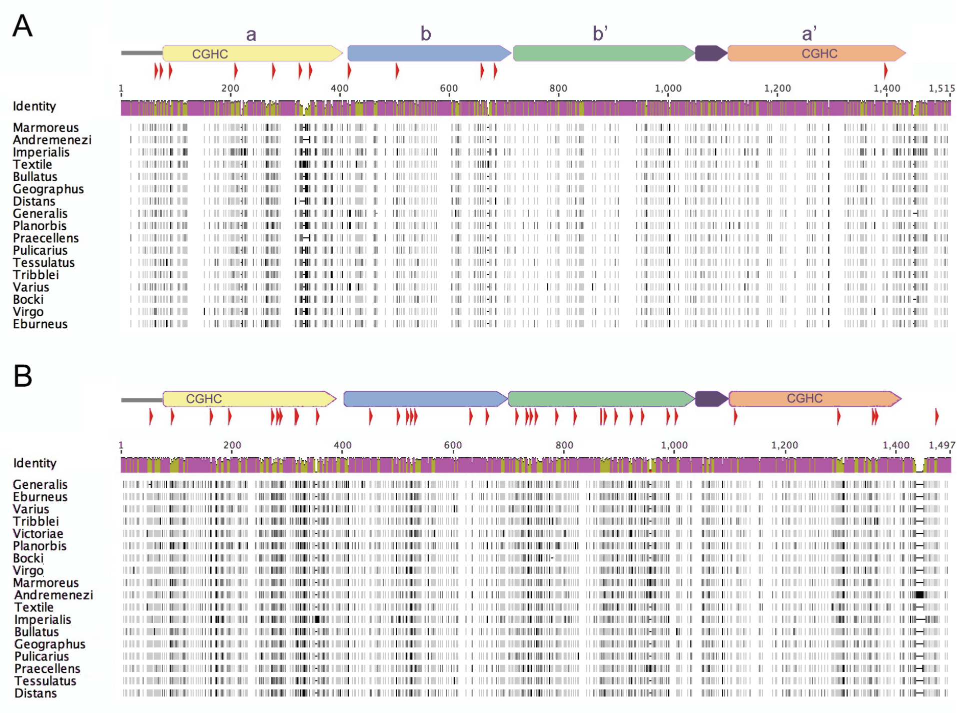


Fig. S6. Analysis of sites undergoing episodic positive selection as performed by MEME implemented in datamonkey (7). Position of sites experiencing positive selection are shown as red arrows above sequence alignments for PDIs (A) and csPDIs (B). A total of 12 and 35 sites (p <0.1) were identified for PDIs and csPDIs, respectively. Multiple sequence alignments were generated in Geneious (version 8.1.3) with the Blosum62 similarity option for coloring (white: 100% identity, light gray: 100-80% identity, dark gray: 80-60% identity, black: less than 60% identity). Schematics of the sequence identity between aligned sequences are provided in purple above the alignment (areas of low or no identify are shown in yellow and red, respectively. Domain architectures are depicted above schematics and follow the same color-coding provided in Fig. 1.

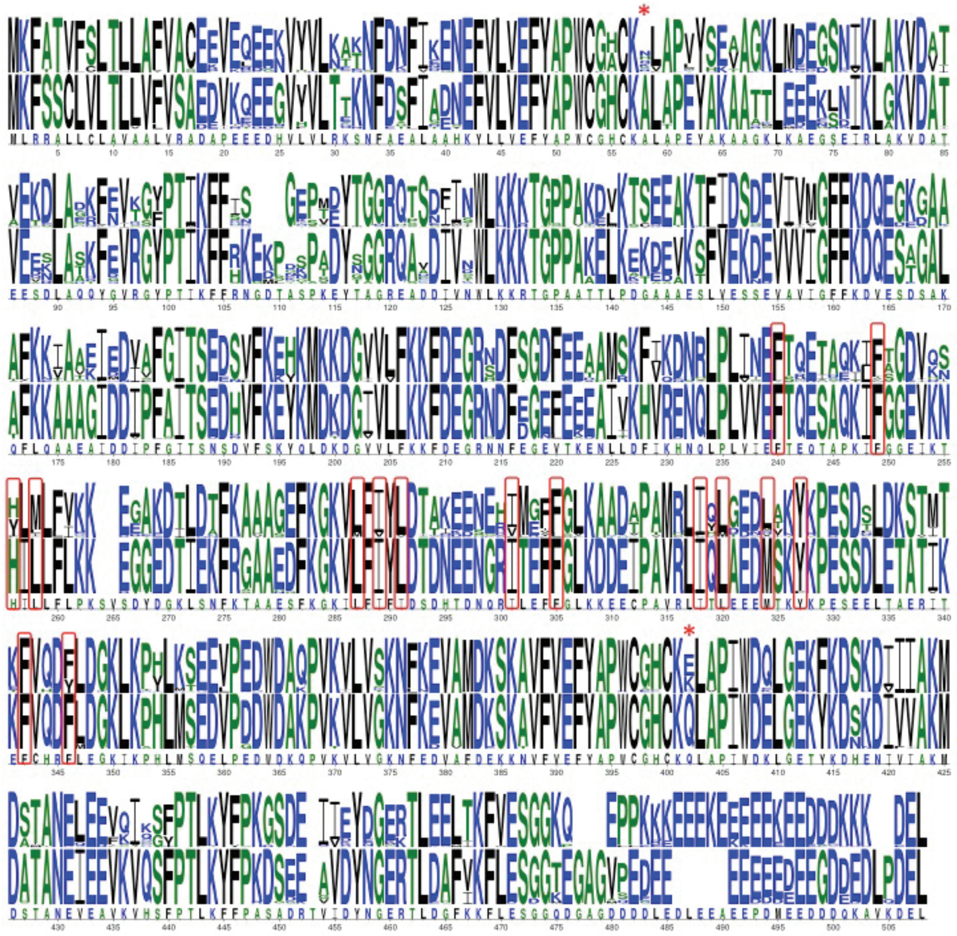


Fig. S7. A logo plot of the csPDI sequence alignment (top row) compared to the corresponding PDI alignment (middle row), and the human PDI sequence (Uniprot entry P07237) as reference (bottom row). The alignment was obtained using the muscle program with standard settings (9). Residues of the hydrophobic, substrate-binding patch mapped in human PDI are boxed, and the "+2 positions" C-terminally of the two CXXC active-site motifs are marked with an asterisk.

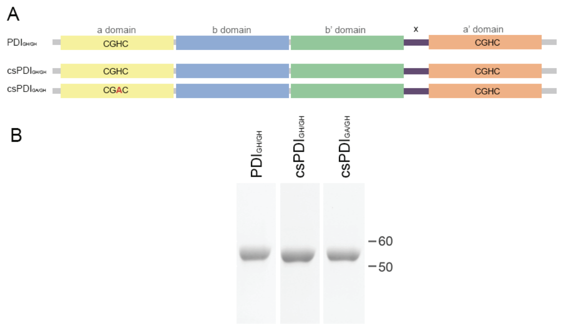


Fig. S8. Recombinant expression of PDI and two members of the csPDI family from *Conus geographus*. A. Two csPDI members were selected for recombinant expression. PDI was expressed for comparison. All proteins had the characteristic domain architecture of the PDI gene family (schematic). B. SDS-PAGE gel analysis of purified PDI and csPDIs from *C. geographus* used for functional assays. His-tagged fusion proteins were purified by metal affinity and size exclusion chromatography to >95% purity.

References

1. Grabherr MG, *et al*. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology* 29(7):644-652.
2. Hu H, Bandyopadhyay PK, Olivera BM, & Yandell M (2012) Elucidation of the molecular envenomation strategy of the cone snail *Conus geographus* through transcriptome sequencing of its venom duct. *BMC genomics* 13(284):1-12.
3. Barghi N, Concepcion GP, Olivera BM, & Lluisma AO (2015) High conopeptide diversity in Conus tribblei revealed through analysis of venom duct transcriptome using two high-throughput sequencing platforms. *Marine biotechnology (New York, N.Y.)* 17(1):81-98.
4. Robinson SD, *et al*. (2014) Diversity of conotoxin gene superfamilies in the venomous snail, Conus victoriae. *PLoS One* 9(2):e87648.
5. Flygare S, *et al*. (2015) Interactive web-based metagenomics analysis portal for universal pathogen detection and host response-based diagnosis and discovery. *Genome biology In Press*.
6. Safavi-Hemami H, *et al*. (2014) Combined proteomic and transcriptomic interrogation of the venom gland of *Conus geographus* uncovers novel components and functional compartmentalization. *Molecular & cellular proteomics : MCP* 13(4):938-953.
7. Murrell B, *et al*. (2012) Detecting individual sites subject to episodic diversifying selection. *PLoS genetics* 8(7):e1002764.
8. Kryazhimskiy S & Plotkin JB (2008) The population genetics of dN/dS. *PLoS genetics* 4(12):e1000304.
9. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32(5):1792-1797.
10. Wang C, *et al*. (2013) Structural insights into the redox-regulated dynamic conformations of human protein disulfide isomerase. *Antioxidants & redox signaling* 19(1):36-45.
11. DeLa Cruz R, Whitby FG, Buczek O, & Bulaj G (2003) Detergent-assisted oxidative folding of delta-conotoxins. *Journal of Peptide Research* 61:202-212.
12. Fuller E, *et al*. (2005) Oxidative folding of conotoxins sharing an identical disulfide bridging framework. *The FEBS journal* 272:1727-1738.
13. Price-Carter M, Gray WR, & Goldenberg DP (1996) Folding of ω-Conotoxins. 1. Efficient Disulfide-Coupled Folding of Mature Sequences in Vitro. *Biochemistry* 35:15537-15546.
14. Wang C, *et al*. (2012) Human protein-disulfide isomerase is a redox-regulated chaperone activated by oxidation of domain a'. *The Journal of biological chemistry* 287(2):1139-1149.

```
>PDI_GH/GH[Conus andremenezi]
ATGAAGTTTTCATCTTGTCTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAAGATGTCAAACAGGAGGAAGGTGTCTACGTTTTGACGACGAAAAATTTTGACTC
CTTCATAGCAGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAACTTTGGA
GGAAGAGAAGTTAAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGAAAGTTTGGCCTCCAAATTTGAAGTTCGTGGATACCCAACAATCAAGTTCTTCCG
TAAAGAGAAGCCTACAGACTACGGTGGTGGTCGCCAAGCTGCCGATATTGTTGACTGGCTGAAGAAGAAGACAGGACCACCTGCCAAGGAACTGAAGGAGAAA
GATGAAGTCAAGTCTTTTGTGGAAAAAAACGAAGTTGTTGTCATTGGATTCTTCAAGGATCAAGAATCCGCAGGTGCTTTGACCTTCAAAAAGGCAGCTGCCGGC
ATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATCGTGCTGCTGAAGAAGTTTGATGAGGGCCGT
AATGACTTTGATGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCCCAGAAGAT
CTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAAAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCTGCTGAGGATTTCAAAGGAA
AGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGCATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCTCATCCAGCT
GGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGACCTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCTTGGATGGAAAACTGAAGCCCCATCT
GATGTCCGAGGATGTGCCTGATGACTGGGACGCCAAGCCCGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCAATGGACAAATCAAAAGCTGTCTTTG
TGGAGTTCTATGCACCTTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGACATTGTTGTTGCCAAGA
TGGATGCCACTGCCAACGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGAGGAGGGTGTGGACTACAATGGCGAG
AGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCACGGAAGGAGCTGGAGTGTCAGAGGATGAGGAAGAGGAAGATGATGAGGAGGAGGGTGATG
AAGAAGATCTGCAAAGAGATGAGCTGTAG
>PDI_GH/GH_[Conus praecellens]
ATGAAGTTTTCATCTTGTCTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAAGATGTCAAACAGGAGGAAGGTGTCTACGTTTTGACGACGAAAAATTTTGACTC
CTTCATAGCAGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAACTTTGGA
GGAAGAGAAGTTAAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGAAAGTTTGGCCTCCAAATTTGAAGTTCGTGGATACCCAACAATCAAGTTCTTCCG
TAAAGAGAAGCCTACAGACTACGGTGGTGGTCGCCAAGCTGTTGATATTGTTGACTGGCTGAAGAAGAAGACAGGACCACCTGCCAAGGAACTGAAGGAGAAA
GATGAAGTCAAGTCTTTTGTGGAAAAAAACGAAGTTGTTGTCATTGGATTCTTCAAGGATCAAGAATCCGCAGGTGCTTTGACCTTCAAAAAGGCAGCTGCTGGC
ATTGATGACGTTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATCGTGCTGCTGAAGAAGTTTGATGAGGGCCGT
AATGACTTTGATGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCCCAGAAGAT
CTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAAAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCTGCTGAGGATTTCAAAGGAA
AGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGCATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCTGCTGTGCGTCTCATCCAGCT
GGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGACCTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCTTGGATGGAAAACTGAAGCCCCATCT
GATGTCCGAGGATGTGCCTGATGACTGGGACGCCAAGCCCGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCAATGGACAAATCAAAAGCTGTCTTTG
TGGAGTTCTATGCACCTTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGACATTTTTGTTGCCAAGA
TGGATGCCACTGCCAACGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTAAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTACAATGGCGAG
AGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCAAGGAAGGAGCTGGAGTGTCAGAGGATGAGGAAGAGGAAGATGATGAGGAGGAGGGTGATG
AAGAAGATCTGCAAAGAGATGAGCTGTAG
>PDI_GH/GH_[Conus bocki]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAAGATGTCAAACAGGAGGAAGGTGTCTACGTTTTGACGACGAAAAATTTTGACTC
CTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATATGCCAAAGCTGCAACAATTTTGGAG
GAAGAGAAGTTAAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGCCAATTTGGCCACCAAATTCGAAGTTCGTGGATACCCAACAATCAAGTTCTTCCAT
AAAGAGAAGTCCAACAGTCCAGCAGACTACGGTGGTGGTCGCCAAGCTGTCGATATTGTTACCTGGCTGAAGAAGAAGACAGGACCACCTGCCAAGGAACTGAA
GGAGAAAGATGAAGTCAAGTCTTTTGTGGAAAAAGACGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCACAGGCGCTTTGGCCTTCAAAAAGGCAGC
TGCTGGCATCGATGACATTCCATTTGCCATCACCTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTGCTACTGAAGAAGTTTGATGA
GGGCCGTAATGACTTTGAGGGGGAGTTTGAGGAGGATGCCATCGTCAAGCACGTCAGGGAAAACCAGCTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCCC
AGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATTTC
AAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGCATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCTC
ATCCAGCTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGATTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCATGGACGGGAAACTGAA
GCCCCATCTGATGTCTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCAATGGACAAATCAAAAG
CTGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGTTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGACATTGTTG
TTGCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTAC
AATGGCGAGAGAACCTTGGATGCTTTCATCAAATTCCTCGAGAGCGGTGGCAAGGAAGGTGCTGGAATGCCAGAGGATGAGGAGGAAGAGGAAGATGAGGAGG
GTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus eburneus]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAAGATGTCAATCAGGAGGAAGGTGTCTACGTGCTGACGACGAAAAATTTTGACTC
CTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCTTTGACACCAGAATACGCCAAAGCTGCAGCAACTTTGGAG
GAAGAGGAGTCAAACATCAAGTTGGCCAAAGTGGATGCTACCGTGGAGCAGAGTTTGGCCTCCAAATTCGATGTTCGTGGATACCCAACAATCAAGTTCTTCCGT
AAAGAGAAGCCTGACAGTCCAGCAGACTACAATGGTGGTCGCCAAGCTGTCGATATTGTTAACTGGCTGAAGAAGAAGACAGGACCACCTGCCAAGGAACTGAA
GAAGAAAGATGATGCCAAGTCTTTTGTGGAAAAAGATGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCGCAGGTGCTTTGGCCTTCAAAAAGGCAGC
TGCCGGCATAGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTACTGCTGAAGAAGTTTGATGA
GGGCCGTAATGACTTTGAGGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCCC
AGAAGATCTTCGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATTTC
AAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGCATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCTC
ATCCAGCTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGATTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGAAACTGAAG
CCCCATCTGATGTCTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCGATGGACAAATCAAAAGC
TGTCTTTGTGGAGTTCTACGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAACAAGGACATTGTTGT
TGCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTACA
ATGGCGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCAAGGAAGGTGCTGGAGTGCCAGAGGATGAGGAAGAGGAAGAGGAAGATGAGGA
GGGTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus tessulatus]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAAGATGTCAAACAGGAGGAAAGTGTCCACGTGCTGACGACGAAAAATTTTGACT
CCTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCTTTGACACCAGAATACGCCAAAGCTGCAGCAACTTTGGA
GGAAGAGGAGTCAAACATCAAGTTGGCCAAAGTGGATGCTACCGTGGAGCAGAGTTTGGCCTCCAAATTCGATGTTCGTGGATACCCAACAATCAAGTTCTTCCG
TAAAGAGAAGCCTGACAGCCCAGCAGACTACAATGGTGGTCGCCAAGCTGTCGATATTGTTAACTGGCTGAAGAAGAAGACAGGACCACCTGCCAAGGAACTGA
AGAAGAAAGATGATGCCAAGTCTTTTGTGGAAAAAGATGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCGCAGGTGCTTTGGCCTTCAAAAAGGCAG
CTGCCGGCATAGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTACTGCTGAAGAAGTTTGATG
AGGGCCGTAATGACTTTGAGGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCC
CAGAAGATCTTCGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATTT
CAAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGCATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCT
CATCCAGCTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGATTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGAAACTGAA
GCCCCATCTGATGTCTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCGATGGACAAATCAAAAG
CTGTCTTTGTGGAGTTCTACGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAACAAGGACATTGTTG
TTGCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTAC
AATGGCGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCACGGAAGGTGCTGGAGTGCCAGAGGAGGAGGAAGAGGAAGAGGAAGATGAGG
AGGGTGATGATGAAGATCTGCCAAGAGACGAACTGTAG
>PDI_GH/GH_[Conus pulicarius]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTGTCTGCCGAAGATGTCAAAGAGGAGGAAGGTGTCTACGTTTTGACGACGAAAAATTTTGACTC
CTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAACTTTGATG
```

```
GAAGAGAACTCAAACATCAAGTTGGGTAAAGTGGATGCTACTGTGGAGGACAGTTTGGCCGCTAAATTCGAAGTTCGTGGATACCCAACAATCAAGTTCTTCCGT
AAAGAGAAGCTTGGCAGTCCAGCAGACTACAATGGTGGTCGCCAAGCTGCCGATATTGTTAATTGGCTGAAGAAGAAGACAGGACCACCTGCCAAGGAACTGAA
GGAGAAAGCTGAAGTCAAGTCTTTTGTGGAAAAAGATGATGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCACAGGTGCCTTGGCCTTCAAAAAAGCAGC
TGCCGGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTACTGCTGAAGAAGTTTGATGA
GGGCCGTAATGATTTTGAGGGGGAGTTCGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAGCTGCCACTGGTCGTAGAGTTCACTCAAGAGTCTGCCC
AGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATTTC
AAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCTC
ATCCAGCTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCAGATTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGAAACTGAAG
CCCCATCTGATGTCTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCGATGGACAAATCAAAAGC
TGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAACAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAACAAGGACATTGTTGT
TGCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTACA
ATGGCGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCACGGAAGGTGCTGGAGTGCCAGAGGATGAGGAAGAGGAAGAGGAAGATGAGGA
GGGTGATGATGAAGATCTGCCAAGAGACGAACTGTAG
>PDI_GH/GH_[Conus marmoreus]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTGGTATCTGCCGATGATATCAAACAGGAGGAAGGTGTCTACGTTTTGACGAAGAAAAATTTTGACTC
CTTCATAACTGAGAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAACTTTGGA
GGAAGAGAAGTTAAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGAGGATTTGGCCTCCAAATTTGAAGTTCGTGGATACCCAACAATCAAGTTCTTCCA
TAAAGAGAAGCCTAACAAACCAGCAGACTACAATGGTGGTCGCCAAGCTGTCGATATTGTTAACTGGCTGAAGAAGAAGACAGGACCACCAGCCAAGGAACTG
AAGGAGAAAGAAGAAGTCAAGTCTTTTGTGGAAAAAGATGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCACAGGTGCTTTGGCCTTCAAAAAGGCA
GCTGCCGGTATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTACTGCTGAAGAAGTTTGAT
GAGGGCCGTAATGACTTCGACGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGC
CCAGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATT
TCAAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAAAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCTGCTGTGCGTC
TCATCCAACTGGCAGAGGACATGTCCAAGTACAAACCTGAGTCCTCGGATTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGAAACTGA
AGCCCCATCTGATGTCTGAAGATGTGCCTGATGACTGGGATGCCAAACCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCGATGGACAAATCAAAA
GCTGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAAGACATTGTT
GTTGCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTTCCCACCCTCAAGTACTTCCCCAAGGACAGTGAGGAGGGTGTGGACTA
CAATGGCGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGATGGCACGGAAGGTGCTGGAGTGCCAGAGGATGAGGAAGAGGAAGAGGAAGATGAG
GAGGGTGATGACGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus virgo]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAAGATGTCATAACGGAGGAAGGTGTCTACGTTTTGACGACGAAAAATTTTGACTC
CTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCATGGTGTGGCCATTGCAAAGCATTGGCACCAGAATATGCCAAAGCTGCAAGAACTTTGGAG
GAAGAGAAGTTACAGATCAAGTTGGGCAAAGTGGATGCTACTGTAGAGGAAAATTTGGCCTCCAAATTCGGAGTTCGTGGATACCCAACAATCAAGTTCTTCCAC
AAAGAGAAGCCTCAGAGTCCAGCAGACTACAATGGTGGTCGCCAAGCTGTTGATATTGTTAACTGGCTGAAGAAGAAGACCGGACCACCTGCCAAGGAACTGAA
GGAGAAAGAGGAAGTCAAGTCTTTTGTGGAAAAAGATGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCGCAGGTGCTTTGGCCTTCAAAAAGGCAGC
TGCCGGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTGCTGCTGAAGAAGTTTGATGA
GGGCCGTAATGACTTTGATGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCCC
AGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATTTC
AAAGGAAAGGTCCTGTTCATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGGCTC
ATCCAGCTGGCAGAAGACATGTCCAAGTACAAGCCTGAGTCCTCAGACTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGAAACTGAAG
CCCCATCTGATGTCTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCCATGGACAAATCAAAAGC
TGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGACATTGTTGT
TGCCAAGATGGATTCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTACA
ATGGTGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCACGGAAGGAGCTGGAGTGCCAGAGGATGAGGAAGAGGAAGAGGAAGATGAGGA
GGGTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus tribblei]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAAGATGTCAAAGAGGAGGACGGTGTCTACGTTTTGACGCAGAAAAATTTTGACTC
CTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAATGTTGGAG
GAAGAGAAGTTAAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGAGAAATTGGCCTCCGATTTTGAAGTTCGTGGATACCCAACAATCAAGTTCTTCCGT
AAAGAGAAGCCTAAGAATCCAACAGACTACAGTGGTGGTCGCCAAGCTGCCGATATTGTCAGTTGGCTGAAGAAGAAGACGGGACCACCTGCCAAGGAACTGA
AGGAGAAAGAGGAAGTCAAGTCTTTTGTGGAAAAAGACGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCACAGGTGCTTTGGCCTTCAAAAAGGCGG
CTGCTGGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTGCTGCTGAAGAAGTTTGATG
AGGGCCGAAATGACTTTGATGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCC
CAGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATTT
CAAAGGAAAGGTCCTGTTCATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCT
CATCCAGCTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGACTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGAAACTGAA
GGCCCACCTGATGTCTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGTGATGGACAAATCAAAAG
CAGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGTTGGGTGAAAAGTACAAGGACAGCAAGGACATTGTTG
TTGCCAAGATGGATGCCACTGCCAACGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTAC
AATGGCGAGAGAACCTTGGATGCTTTCATCAAATTCCTCGAGAGTGGTGGCATGGAAGGAGCTGGAGTGTCAGAGGATGAGGAAGAGGATGAGGAAGAGGAGG
AGGGTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus distans]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAGGATGTCAAACAGGACGAAGGTGTCTACGTTTTGACGACGAAAAATTTTGACTC
CTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCAGCAACAACTTTGGA
GGAAGAGAAGTTAAACATCAAGTTGGGCAAAGTGGATGCTATTGTGGAGGAGAAGTTGGCCACCAGATTCGAAGTTCGTGGATACCCAACAATCAAGTTCTTCA
GTAAAGAGAACAAACCATCAGACTACACTGGTGGTCGCCAAGCTTCCGATATTGTTCAATGGCTGAAGAAGAAGACAGGACCACCTGCCAAGGAACTGAAGGAG
AAAGATGAAGTCAAGTCTTTTGTGGAAAAAGACGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAGGAATCTGCAGGTGCTTTGGCATTCAAAAAGGCAGCTGCC
GGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAGGATGGCGTTGTGCTGCTGAAGAAGTTTGATGAGGGC
CGTAATGACTTTGAGGGGGAGTTTGAGGAGGAGGCCATTGTCAAGCACGTCAGGGAAAACCAGCTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCCCAGAA
GATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGGGGTGCAGCTGAGGGTTTCAAAG
GAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACCGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCTCATCC
AGCTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGACTTGGAAACTGCCACCATCAAGAAATTTGTTCAGGATTTCCTGGATGGAAAACTGAAGCCCC
ATCTGATGTCTGAGGATGTGCCTGATGACTGGGACGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCGATGGACAAATCAAAAGCTGTC
TTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAACAAGGACATTGTTGTTGCC
AAGATGGATGCCACTGCCAACGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACTCTCAAGTATTTCCCCAAGGACAGCGAGGAGGCGGTGGACTATAATGG
CGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCACGGAAGGAGCTGGAGTGCCAGAGGATGAGGAAGAGGAAGAGGAAGATGAGGAGGGT
GATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus generalis]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTTTGGTTTTTGTGTCTGCCGAAGAAGTCAAACGGGAGGAAAATGTCTACGTTTTGACGACGAAAAATTTTGACTC
CTTCATAGCTGAAAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAACTTTGGA
GGAAGAGAAGTTAGACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGAGAATTTGGGCGCCAAATTCGGAGTTCGTGGATACCCAACAATCAAGTTCTTCC
GTAAAGAGGATCCCAGCAATCCATCAGACTACAGTGGTGGTCGCCAAGCTGACGATATTGTTAAATGGCTGAAGAAGAAGACAGGACCACCGGCCGCAGAACTG
AAGCAGACAGAGGAAGTCAAATCTTTTGTGGAAAAAGATGTTGTTGTCATTGGTTTCTTCAAGGATCAGGAATCTGCAGGTGCTTTGGCCTTCAAAAAGGCAGCT
```

```
GCTGGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGTATTGTGCTGCTGAAGAAGTTTGATGAG
GGCCGTAGTGACTTTGAGGGGGAGTTTGAGAAGGAGGCCATCATCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCACA
GAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAAGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGCTGCAGCCGAGGATTTCA
AAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACGGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCTCA
TCCAGTTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGACTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCATGGATGGAAAACTGAAGC
CCCATCTGATGTCTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCGATGGACAAATCAAAAGCT
GTCTTTGTGGAGTTCTATGCCCCCTGGTGTGGACACTGCAAGCAGCTGGCTCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGACATTGTTGTA
GCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACTCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTACAA
TGGCGAGAGAACCTTGGATGCTTTCATCAAATTCCTCGAGAGCGGTGGCACGGAAGGAGCTGGAGTGCCAGAGGAAGAGGAAGAGGAAGATGAGGAGGGTGAT
GATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus varius]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTTTGGTTTTTGTGTCTGCCGATGAAGTCAAACAGGAGGAAGGTGTCTTCGTTTTGACGACGAAAAATTTTGACAC
CTTCATAGCTGAAAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAACTTTGGA
GGACGAGAAGTCAGACATCAAGTTGGGCAAAGTAGATGCTACTGTGGAGGACAAGTTGGCCGCCAAATTCGAAGTTCGTGGATACCCAACAATCAAGTTCTTCC
GTAAAGAGGATCCCACCAATCCAACAGACTACAGTGGTGGTCGCCAAGCTGATGATATTGTTACCTGGCTGAAAAAGAAGACAGGACCACCGGCCGCAGAACTG
AAGGAGACGGATGAAGTCAAGTCTTTTGTGGAAAAAGACGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCTGCAGGTGCTTTGGCCTTCAAAAAGGCA
GCTGCTGGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACACAGATGGCATTGTGCTGCTGAAGAAGTTTGAT
GAGGGCCGTAATGACTTTGAGGGGGAGTTTGAGAAGGAGGCCATCATCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGC
CCAGAAGATCTTTGGAGGTGAGGTGAAGAACCACATACTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATT
TCAAAGGAAAGGTCTTGTTTATCTACTTGGACACTGATAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTC
TCATCCAGTTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGACTTGGAAACTGCCACCGTCAAGAAATTTGTCCAGGATTTCCTGGATGGAAAACTGA
AGCCCCATCTGATGTCTGAGGATGTGCCTGATGACTGGGACGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGATGTGGCAATGGACAAATCAAAA
GCTGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGACATTGTT
GTTGCCAAGATGGATGCCACTGCCAACGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACTCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTA
CAATGGCGAGAGAACCTTGGATGCTTTCATCAAATTCCTCGAGAGCGGTGGCACGGAAGGAGCTGGAGTGCCAGAGGATGAGGAAGAGGATGAGGAAGAGGAG
GAGGGTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus planorbis]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTTTGGTTTTTGTTTCTGCCGAAGAAGTCAAAAAGGAGGAAGGTGTCTACGTTTTGACGGAGAAAAATTTTGAATC
CTTCATTGCTGAAAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGTCATTGCAAGGCATTGGCGCCAGAATATGCTAAAGCTGCAGCAACTTTGGAG
AAAGAGAACTTAGACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGAAGATTTGGCCCGGAAATTCGAAGTCCGTGGATACCCAACAATCAAGTTCTTCCG
TAAAGAGGATCCAAAAACTGCATCAGACTACACTGGTGGTCGCCAATCTGGCGATATTGTTAACTGGCTGAAGAAGAAGACGGGACCACCAGCCGCAGAACTGA
AGGAGACAGATGAAGTCAAGACTTTTGTGGAAAAAGACGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCTGCAGGTGCTTTGGCCTTCAAAAAGGCAG
CTGCTGGCATTGATGACATTCCGTTTGCCATCACTTCAGAGGATAATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTGCTGCTGAAGAAGTTTGATG
AGGGCCGTAATGACTTTGAGGGGGAGTTTGAGAAGGAGGCCATCATCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAGGAGTCTGCC
CAGAAGATCTTCGGAGGCGAGGTGAAGAACCACATTCTGCTTTTCCTGAAGAAGGATGGTGGAGAAGACACGATTGAGAAGTTCAGAAGTGCAGCTAATGATTT
CAAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCTGCTGTGCGTCT
CATCCAGTTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGACTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGAAAACTGAA
GCCCCATCTGATGTCTGAGGATGTACCTGATGACTGGGATGCCAAGCCTGTGACGGTCCTAGTGGGCAAAAATTTCAAGGAAGTGGCGATGGACAAATCAAAAG
CTGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAACAAGGACATTGTTG
TTGCCAAGATGGATGCCACTGCCAACGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACTCTCAAGTACTTCCCCAAGGACAGCGAGGAGGTTGTGGACTAC
AATGGTGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGTACAGAAGGAGCTGGAGTGCCAGAGGATGAGGAAGAGGAAGAGGAAGATGAGG
AGGGTGATGAAGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus imperialis]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTGGTATCTGCCGAAGATGTCAAAGAGGAGGAAGGTGTCCACGTTTTGACGAACAACAATTTTGACTC
CTTCATAAACGAGTATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATATGCCAAAGCTGCACAAAAATTGAA
GGATGAGGGCAACGAGAATATCAAGTTGGCCAAAGTGGATGCTACTGTGGAGGACAAATTGGCCACCAAATTCCAAGTTCGTGGATACCCAACAATCAAGTTCT
TCCATAAAGAGAAGTCTGACAGTCCAGTAGACTACAGCGCTGGTCGCCAGGCTGAGGACATTGTTAACTGGCTGAAGAAGAAAACAGGACCACCTGCCAAGGAA
CTGAAGGACAAAGATGCAGCCAAGACTTTTGTGGAAAAAGACGAAGTTGTTGTCATTGGGTTCTTCAAGGATCAAGAATCCGAAGGTGCTTTGGCCTTCAAAAAG
GCAGCTGCCGGCATTGATGACATTCCATTTTCCATCACTTCAGACGATGATGTTTTCAAGGAGTACAAGATGGACAGAGATGGCGTTGTGCTGCTGAAGAAGTTT
GACGAGGGCCGTAATGACTTTGAGGGGGAGTTTGAGGCAGAGGCCATCACCAAGCACGTCAGGGATAACCAACTACCACTGGTTGTAGAGTTCACTCAAGAGTC
TGCCCAGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGGGG
ATTTCAAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGC
GTCTCATCCAGCTGGCAGAGGACATGTCCAAGTTCAAGCCTGAGTCCTCAGACCTGGAAACTGCCACCATCAAGAAATTTGTCCAAGATTTCCTGGATGGAAAAC
TGAAACCCCATCTGATGTCTGAGGATGTGCCCGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGATGTGGCGATGGACGAATCA
AAAACTGTCTTTGTGGAGTTCTATGCCCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAATGACATT
GTTATTGCCAAGATGGATGCCACTGCCAACGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACGGTGGGAAGGTTGTGGA
CTACAATGGTGAGAGAACCTTGGAGGGTTTCGTCAAATTCCTCGACAGTGATGGCAAGGAAGGAGCTGGAGCGCCAGAGGGAGAGGAGGAGGAAGAGGATGAG
GAGGAGGAGGGTGATGAAGATGATCTGCCAAGGGATGAACTGTAG
>PDI_GH/GH_[Conus_geographus]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCTGCCGAAGATGTCAAACAGGAGGAAGGTGTCTACGTTTTGACGGCGAAAAATTTTGACTC
CTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAACTTTGGAA
GAAGAGAAGTTAAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGAGAGTTTGGCCGCCAAATTCGAAGTTCGTGGATACCCAACAATCAAGTTCTTCCGT
AAAGAGAAGCCTGACGGTCCAGCAGACTACAGTGGTGGTCGCCAAGCTGCTGATATTGTTAGCTGGCTGAAGAAGAAGACAGGACCACCTGCCAAGGAACTGAA
GGAGAAAGATGAAGTCAAGTCTTTTGTGGAAAAAGACGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCACAGGTGCTTTGGCCTTCAAAAAGGCAGC
TGCCGGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAGAGATGGCATTGTACTGCTGAAGAAGTTTGATGA
GGGCCGTAATGACTTTGACGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCCC
AGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATCGAGAAGTTCAGAGGTGCAGCTGAGGATTTC
AAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCTC
ATCCAGTTGGCAGAGGACATGTCCAAGTACAAGCCTGAGTCCTCGGATTTGGAAACGGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGAAACTGAAG
CCTCATCTGATGTCTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCAATGGACAAATCAAAAGC
TGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGACATTGTTGT
TGCCAAGATGGATGCCACTGCCAACGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAGGGACAGTGAGGAGGCTGTGGACTACA
ATGGCGAGAGAACCTTGGATGCTTTCGTCAAATTCCTTGAGAGCGGTGGCACGGAAGGTGCTGGAGTGCCAGAGGATGAGGAAGAGGAAGAGGAAGACGAGGA
GGGTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus_textile]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCAGCCGAAGATGTCAAACAGGAGGAAGGTGTCTACGTTTTGACGGAGAAAAATTTTGACG
GCTTCATATCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATATGCCAAAGCTGCAACAACTTTGGA
GGAAGAGAAGTCGAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGTGAACTTGGCCACCAAATTCGAAGTTCGTGGATACCCAACAATCAAGTTCTTCC
ATAAAGAGATGCCTGCTGGCAGTCCAGCAGACTACAGTGGTGGTCGCCAAGCTCCAGATATTGTTGGCTGGCTGAAGAAGAAGACAGGACCACCAGCCAAGGAA
CTGAAGGCGAAAGATGAAGTCAAGACTTTTGTGGAAAAAGATGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCACAGGTGCTTTGGCCTTCAAAAAG
GCAGCTGCCGGCATTGATGACATTCCACTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTACTGCTGAAGAAGTTT
GATGAGGGCCGTAATGACTTCGAGGGGAATCTGGAGGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGA
GTCTGCCCAGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGAGGAGAAGACACAATTGAAAAGTTCAGAGGTGCAGCTG
```

```
AGGATTTCAAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCAGCTG
TGCGTCTCATCCAGCTGGCAGAGGACATGTCAAAGTACAAGCCCGAGTCCTCGGATTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGA
AACTGAAGCCCCATCTGATGTCTGAGGATGTGCCTGGTGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCGATGGACAAA
TCAAAGGCTGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGAC
ATTGTTGTTGCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGATGAGGCTGTG
GACTACAATGGCGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCACGGAAGGTGCTGGAGTGCAAGAGGATGAGGAAGAGGAAGAGGAAG
ATGAGGAGGGTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_variant_2_[Conus_textile]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTCTGGTTTTTGTATCAGCCGAAGATGTCAAACAGGAGGAAGGTGTCTACGTTTTGACGGAGAAAAATTTTGACG
CCCTCATATCTGATAGTGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATATGCCAAAGCTGCAACAACTTTGGA
GGAAGAGAAGTCGAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGTGAACTTGGCCACCAAATTCGAAGTTCGTGGATACCCAACAATCAAGTTCTTCC
ATAAAGGGATGCCTGCTGGCAGTCCAGCAGACTACAGTGGTGGTCGCCAAGCTCCAGATATTGTTGGTTGGCTGAAGAAGAAGACAGGACCACCAGCCAAGGAA
CTGAAGGCGAAAGATGAAGTCAAGACTTTTGTGGAAAAAGATGAAGTTGTTGTCATTGGTTTCTTCAAGGATCAAGAATCCACAGGTGCTTTGGCCTTCAAAAAG
GCAGCTGCCGGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTACTGCTGAAGAAGTTT
GATGAGGGCCGTAATGACTTCGAGGGGAATTTGGAGGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGA
GTCTGCCCAGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAAAAGTTCAGAGGTGCAGCTG
AGGATTTCAAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCAGCTG
TGCGTCTCATCCAGCTGGCAGAGGACATGTCAAAGTACAAGCCCGAGTCCTCGGATTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGA
AACTGAAGCCCCATCTGATGTCTGAGGATGTGCCTGGTGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCGATGGACAAA
TCAAAGGCTGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGAC
ATTGTTGTTGCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGATGAGGCTGTG
GACTACAATGGCGAGAGAACCTTGGATGCTTTCGTCAAATTCCTCGAGAGCGGTGGCACGGAAGGTGCTGGAGTGCAAGAGGATGAGGAAGAGGAAGAGGAAG
ATGAGGAGGGTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>PDI_GH/GH_[Conus_bullatus]
ATGAAGTTTTCATCTTGTTTAGTTTTAACTCTTGTGGTTTTTGTATCTGCCGAAGATGTCAAACAGGAGGAAGGTGTCTACGTTTTGACGACGAAAAATTTTGACTC
CTTCATAGCTGATAATGAGTTCGTGCTTGTGGAATTTTATGCTCCCTGGTGTGGCCATTGCAAGGCATTGGCACCAGAATACGCCAAAGCTGCAACAAGTTTGGAG
GAAGAGAAGTCAAACATCAAGTTGGGCAAAGTGGATGCTACTGTGGAGGAGAAATTGGCCTCTACATTCGAAGTTCGTGGATACCCAACAATCAAGTTCTTCCGT
AAAGAGAAGCCTGATGGTCCAGCAGACTACACTGGTGGTCGCCAAGCTAAAGATATTGTTGACTGGCTGAAGAAGAAGACGGGACCACCTGCCAAGGAACTGAA
GGAGAAAGATGATGTCAAGTCTTTTGTGGAAAAAGACGAAGTTGTTGTCATTGGGTTCTTCAAGGATCAAGAATCCACAGGTGCTTTGGCCTTCAAAAAGGCAGC
TGCCGGCATTGATGACATTCCATTTGCCATCACTTCAGAGGATCATGTTTTCAAGGAGTACAAGATGGACAAAGATGGCATTGTACTGCTGAAGAAGTTTGATGA
GGGCCGTAATGACTTTGATGGGGAGTTTGAGGAGGAGGCCATCGTCAAGCACGTCAGGGAAAACCAACTGCCACTGGTTGTAGAGTTCACTCAAGAGTCTGCCC
AGAAGATCTTTGGAGGTGAGGTGAAGAACCACATTCTGCTGTTCCTGAAGAAGGAAGGTGGAGAAGACACAATTGAGAAGTTCAGAGGTGCAGCTGAGGATTTC
AAAGGAAAGGTCCTGTTTATCTACTTGGACACTGACAATGAGGAGAATGGACGTATCACAGAGTTCTTTGGCTTGAAGGATGATGAAATCCCCGCTGTGCGTCTC
ATTCAGCTGGCAGAGGATATGTCCAAGTACAAGCCTGAGTCCTCGGACTTGGAAACTGCCACCATCAAGAAATTTGTCCAGGATTTCCTGGATGGGAAACTGAAG
CCCCATCTGATGACTGAGGATGTGCCTGATGACTGGGATGCCAAGCCTGTGAAGGTCCTGGTGGGCAAGAACTTCAAGGAAGTGGCAATGGACAAATCAAAAGC
TGTCTTTGTGGAGTTCTATGCTCCCTGGTGTGGACACTGCAAGCAGCTGGCCCCTATCTGGGATGAGCTGGGTGAAAAGTACAAGGACAGCAAGGACATTGTTGT
TGCCAAGATGGATGCCACTGCCAATGAGATTGAAGAGGTCAAAGTGCAGAGCTTCCCCACCCTCAAGTACTTCCCCAAGGACAGCGAGGAGGCTGTGGACTACA
ATGGCGAGAGAACCTTGGATGCTTTCATCAAATTCCTCGAGAGCGGTGGCACGGAAGGTGCTGGAGTGCCAGAGGAAGAGGAAGAGGAAGAGGAAGATGAGGA
GGGTGATGATGAAGATCTGCCAAGAGATGAACTGTAG
>csPDI_GA/GH_[Conus_geographus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGATGAAAAAGTCTATGTTTTAAAGACAACAAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGTGGCGCATGCAAGAACCTGGCTCCCGTATACCACGAGGTTGCTGGGAAATTGAT
GGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTTACCGCAGAGACGGACTTGGCAGACAAGTTCAACATTACAAGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAATGGAATACACTGGAGGCAGGCAGACTTCCAACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATTGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCGGAA
ATTGAGGATGTTGCCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGAGACTTTGAGAAGGCTGCCATGAGCAAGTTCATCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGATCGCCGAGAAGCTC
TTCACTGGGGATGTGCAGAGCTACCTCATGCTGTTCGTCAAGAAGGAGGAAGCAAAGGACACATTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAACAAGGAGAGTGAACAGATCATGGAATTCTTCGGCCTCAAGGCTGCCGACACTCCGGCGATGCGCCTGATTTATCT
GGGCGAAGACTTGGCCAAGTACAAACCCGAGTCAGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTTCTGGATGGCAAGCTGAAGCCTTACCT
GAAGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAAGAGGTGGCCATGGACAAATCCAAGGCCGTCTTTG
TGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCAATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAG
ATGGACTCCATGGCCAATGAACTGGAAGAGATTCAAATCAAGGGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCAAATATGATGGCAT
GAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAA
GGATGAACTCTAA
>csPDI_GA/GH_variant_2_[Conus_geographus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGATGAAAAAGTCTATGTTTTAAAGACAACAAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGTGCATGCAAGAACCTGGCTCCCGTATACCACGAGGTTGCTGGGAAATTGAT
GGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTTACCGCGGAGACGGACTTGGCGGACAAGTTCAACATTACAGGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAATGGAATACACTGGAGGCAGGCAGACCTCCGACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATTGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATGGCCTCGAAA
ATCGAGGACGTTGTCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGTGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGATCGCCGAGAAGCTC
TTCACTGGGGATGTGCAGAGCCACCTCATGCTGTTCGTCAAGAAGGAAGAAACAAAGGACACATTGGACACCTTTAAGGCTGCTGCCGGCGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAGTGAACGGATCATGGAATTCTTTGGCCTCAAGGCTGCCGACACTCCGGCGATGCGCCTGATTTATCT
GGGCGAAGACTTGGCCAAGTACAAACCCGAGTCAGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTTCTGGATGGCAAGCTGAAGCCTTACCT
GAAGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAAGAGGTGGCCATGGACAAATCCAAGGCCGTCTTTG
TGGAGTTTTATGCCCCATGGTGTGGACACTGCAAGGAGTTGGCTCCAATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGA
TGGACTCCACAGCCAATGAACTGGAAGAGGTCCAAATCGAGGGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCAGATATGATGGCAAG
AGGACCCTGGAAGAACTGACTAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GA/GL_[Conus_geographus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGATGAAAAAGTCTATGTTTTAAAGACAACAAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGTGGCGCATGCAAGAACCTGGCTCCCGTATACCACGAGGTTGCTGGGAAATTGAT
GGACGAGGGTTCCAACATCAAGCTGGCCGAGGTTGATGTTACCGCAGAGACGGACTTGGCGGACAAGTTCAACATTACAGGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAATGGAATACACTGGAGGCAGGCAGACTTCCAACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGATGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATTGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCGGAA
ATTGAGGATGTTTCCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGAGACTTTGAGAAGGCTGCCATGAGCAAGTTTGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGATCGCTGAGAAGCTC
TTCACTGGGGATGTGCAGAACTACCTCATGCTGTTCGTCAAGAAGGAGGAAGCCAAGGACACATTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAACAAGGAGAGTGAACAGATCATGGAATTCTTCGGCCTCAAGGCTGCCGACACTCCGGCGATGCGCCTGATTTATCT
GGGCGAAGACTTGGCCAAGTACAAACCCGAGTCAGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTTCTGGATGGCAAGCTGAAGCCTTACCT
GAAGGCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAAGAGGTGGCCATGGGCAAATCCAAGGCTGTCTTTG
```

```
TGGAGTTTTATGCCCCATGGTGCGGACTCTGCAAGGAATTGGCTCCAATCTGGGATCAGCTGGGTGAGAAATTCAAGGACAGCAAGGATATCATCATCGCCAAGA
TGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAACCGAGGGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCAGATATGATGGCAAG
AGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GH/GH_[Conus_geographus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAAAGAAAAAAAATTTTGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGAAACTGGCTCCCATGTACAGCGAGGCTGCTGGGAAATTGA
TGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGACGGACTTGGCGGGCAAGTTTGAAGTGAAAGGTTTCCCCACCATCAAATTCTTCA
TCGATGGCGAATCAGTAGATTACACAGGAGGCAGGCAGACCTCCGACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACGTGAAGACTTCA
GAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATTGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGATGCTGCTGCCTTCAAGAAGACTGCCTCGAA
AATCGAGGACGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCG
CAATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACTGCCCAGAAGAT
CTTCGCAGGTGACATTCAGAGCCACCTCATGCTGTTCGTCAAGAAGGAGGAAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGA
AGGTCCTGTTTATCTACCTGGACACAACCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCTGACGCTCCCGCGATGCGCCTGATTCAGC
TGGGCGAAGACCTTGCCAAGTACAAACCCGAGTCAGACTCTCTGGACAAGTCCACCGTGACCAAGTTTGTTCAGGACTTTCTGGATGGCAAGCTGAAGCCTCACC
TTAAGTCGGAGGAGGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAAGAGGTGGCCATGGACAAATCCAAGGCCGTCTTT
GTGGAGTTTTATGCCCCATGGTGTGGACACTGCAAGAAGTTAGCTCCGATCTGGGATCAGTTGGGTGAGAAGTTCAAAGACAGCAAAGATATCATCATCGCCAAG
ATGGACTCCACAACCAATGAGCTGGAAGAGGTCCAAATCAAGAGTTTCCCAACCTTGAAATACTTCTCTAAGGGCAGCAACGAGATCATTGAATATGATGGTGA
GAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAA
GGATGAACTCTAA
>csPDI_DA/EF_[Conus_geographus]
ATGAAGTTTGCAACTGTTTTCTCTTTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAAAGACAACAAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACACACCATGGTGCGATGCATGCAAGAACCTGGCTCCCGTGTACCACGAGGTTGCTGGGAAATTGAT
GGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATATTATCGCGGAGATGGACTTGGCGGACATGTTCAGTGTTACAAGTGACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAATGGAATACACTGGAGGCAGGCAGACTTCCAACTTCATCAACTGGCTGAAGAAGAAGACCGGCCACCCTGCCAAGGACGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATTGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCGGAA
ATTGAGGATGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCCACGAGTATACACAGGAGATCGCCATGAAGCTC
TTGAAAGGGGATGTGCAGAGCTACCTCATGCTGTTCATCAAGAAGGAGGAAGCAAAGGACACATTCAACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTGTCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGAATTCTTCAGCCTCAAGGCTGCCGACATTCCAGCGATGCGCCTGATTTATCT
GGGCGAAGACCGGGCCAAGTACAAACCCGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTTCTGGATGGCAAGCTGAAGCCTCACCT
GAAGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAAGAGGTGGCCATGGACAAATCCAAGGCCGTCTTTG
TGGAGTTTTATGCCCCATGGTGCGAATTCTGCAAGGAGTTGGCTCCAATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGA
TGGACTCCATGGCCAATGAACTGGAAGAGATTCAAATCAAGGGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCAAATATGATGGCATG
AGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGATGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GA/GH_[Conus_textile]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCTTGTGAAGATGTTGAACGAGAGGAAAACGTCTATGTTTTAAAGGCAACAAACTTCGACA
ACTTCATTGATGAAAATGAATTTGTTCTTGTGGAGTTCTATGCACCATGGTGCGGTGCCTGCAAGAACCTGGCTCCCGTGTACAGTGAGGTTGCTAAGAAATTGAT
GGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTTACCGTGGAAAAGGACTTGGCGGCCAACTTCAACGTCACAGGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAATGGAATACACTGGAGGCAGGCAGACTTCCGACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGAGGTGAAGACTTGTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCTGCAGAA
ATTGACGACGTTGCTTTTGGCATCACGTCAGAGGACAACGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGGCTTCTCTGGCGACTTTGAGGAGGCTGCCATGAGCAAGTTCATCAAGGACAATTACCTCCCTCTGATCAACGAGTTTACACAGGATGCCGCCCAGAAGCTC
TTCACAGGGGATGTGAAGAGCTATCTCATGCTGTTCGTCAAGAAGGAGGGAGCAAAGGACATATTGGACACCTTTAAGGCTGCTGCTGGTGAATTCAAAGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCATAGAGGAGCATGAACGTGTCATGGAAGTCTTCGGCCTCAAGGCTGCTGACACTCCGGCGATGTGCCTGATTCAGCT
GGATGAAGACCTGGTCAAGTACAAACCCGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTTACCT
GAAGTCGCAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACCTCAAAGAGGTGGCCATGGACAAATCCAAGGCCGTTTTTG
TGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAG
ATGGACGCCACAGCCAATGAGCAGGAGGAGGTCAAAATCAAGAGTTACCCAACCTTCAAATACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGTGA
GAGGACCCTGGAAGAACTGATCAAATTCGTGGAGAGTGGCGGCAAGCAAGAACCTCCGAAGAGAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAA
GGATGAACTTTAA
>csPDI_GH/GH_[Conus_textile]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCTTGTGAAGAAGTTGAACAAGAGGAAAACGTCTATGTTTTAAAGGCAACAAACTTCGACA
ACTTCATTGATGAAAATGAATTTGTTCTTGTGGAGTTCTACGCACCATGGTGCGGCCACTGCAAGGAACTGGCTCCCGTGTACAGCGAGGCTGCTGGGAAATTGA
TGGATAAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCATGGAGAAGGACTTGGCGGAAAAGTTTGAAGTCAAAGGCTTCCCCACCATCAAATTCTTCA
TCAGTGGTGAACCAATGGATTACACAGGAGGCAGGCAGACTTCCGACATCATCAACTGGCTGAAGAAGAAGACCAGCCCCCCTGCCAAGGAGGTGAAGACTTGT
GAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTTAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGTGAA
AATCAAGGACGTTTCCTTTGGCATCACGTCAGAGGATAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCG
CAGTGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCATCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACGGCCCAGAAGCT
CTTCACAGGGGATGTGCAGAGCTATCTCATGCTGTTCGTCAGGAAGGAGGGAGCAAAAGACACACTGGACACCTTTAAGGCTGCTGCTGGTGAATTCAAGGGGA
AGGTCCTGTTTATCTACCTGGACACAGCCATAGAAAAGAATAAACGCTTCATGGAAGTCTTCGGCCTCAAGGCTGCTGACACTCCAGCAATGCGCCTGATTCAGC
TGGGCGAAGGCCTGGTCAAGTACAAACCCGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTGATC
TGAAGTCGGAGGAAGTGCCAGAGAACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAAAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTTTTC
GTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGGAGTTCAAGGACAGCAAGGATGTCATCATCGCCAA
GATGGACTCCACAGCCAATGACGTGGAAGAGGTCCCAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCAAATATGATGGTG
AGAGGACCCTGGAAGGATTGATCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAGAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAA
AGGATGAACTCTAA
>csPDI_GH/GH_variant_2_[Conus_textile]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAAAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAAAGACAAAAAATTTCGACA
ACTTCATTACGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGAAACTGGCTCCCGTGTACAGTGAGGCTGCTGGGCAATTGAT
GGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGGAAAAGTTTGAAGTCAAAGGTTTCCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAGTGGATTACACAGGAGGCAGGCAGACTTCCGACATCATCAACTGGCTGAATAAGAAGACCGGCCCCCCTGCCAAGGAGGTGAAGACTTGTG
ATCAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTTAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCTGCAGAAA
TTGACGACGTTGCTTTTGGCATCACGTCAGAGGACAACGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTGTGTTTAAAAAGTTTGACGAGGGTCGCA
ATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACGGCCCAGAAGATCT
TCTCAGGGGATGTGCAGAGCCACCTCATGCTGTTTGTCAAGAAGGAGGGAGCAAAAGACACACTGGACACCTTTAAGGCTGCTGCTGGTGAATTCAAGGGGAAG
GTCCTGTTTATCTACCTGGACACAACCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCTGACGCTCCAGCAATGCGCCTGATTCAGCTG
GGCGAAGACCTGGCCAAGTACAAACCCGAGTCGGACTCTCTGGACAAGTCCACCATTACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTCACCTG
AAGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTCTTTGT
GGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATGTCATCATCGCCAAGA
TGGACGCCACAGCCAATGAGCTGGAAGACCTCCAAATCAAGAGTTTCCCAACCTTGCAATACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGTGAG
```

```
AGGACCCTGAAAGAACTGATCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAGAGAGGAGGAAGAAAAAGAGGAAGATGACGTCAAGAAAAAG
GATGAACTCTAA
>csPDI_GH/GH_[Conus_bullatus]
ATGAAGTTTGCCACTGTTCTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTAAAACAAGAGGAAAAAGTCTATGTTTTAAAGGAAAAAAATTTCGAC
AACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGAGCCTGGCTCCTGTGTACAGCGAGACTGCTGGGAAATTG
ATGGAAGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGAGCAAGTTTGAAGTGAAAGGTTTCCCCACCATCAAATTCTTC
ATGAATGGCCAATCAGTGGATTACACAGGAGGCAGGCAGTCTTCCGACTTCATCAACTGGCTGAAGAAGAAAACCGGCCCCCCTGCCAAGGATGTGAAGACTTC
TGAGGAGGCCAAGACTTTCATTGACAGTGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCGG
GAATTGAGGACGTTGCCTTTGGCATCACATCAGAGGAGAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTGTGTTTAAAAAGTTTGACGAGGGT
CGCAGTGACTTCTCTGGAGATTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGGTCAACGAGTTTTCACAAGAGACTGCCCAGAAG
ATCTTCGCAGGAGACATTCAGAGCCACCTCATGCTGTTCGTCAAGAAGGAGGAAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGG
GAAGGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCTGACGCTCCAGCGATGCGCCTGATTCA
GCTGGGCGAAGACCTGGCCAAGTACAAACCTGAGTCGGACTCTCTGGACAAGTCCGTCATAACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTCA
CCTGAAGTCAGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAAGAGGTGGCCATGGACAAATCCAAGGCTGTCT
TTGTGGAGTTTTATGCCCCATGGTGTGGACACTGCAAGAAGTTGGCTCCGATCTGGGACCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCA
AGATGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGAACATCGAATATGATGGC
GAGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAA
AAGGATGAACTCTAA
>csPDI_GH/GH_variant_2_[Conus_bullatus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAAAGGCAACGAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGTGGCCACTGCAAGAGCCTGGCTCCTGTGTACAGCGAGACTGCTGGGAAATTGAT
GGAAGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGGGCAAGTTCAACGTCACAAGTTACCCCACCATCAAATTCTTCAT
CAGCGGCGAACCAATGGAATACACTGGAGGCAGGCAGACTTCCAACTTCATCAACTGGCTGAAGAAGAAGACCGGCTCCCCTGCCAAGGATGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGACGAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCGGAA
ATTGAGGACATTGCCTTTGGCATCACATCAGAGGAGAGCGTCTTTAAAGAGCATAAGATGAAGAAGGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AGTGACTTCTCTGGAGATTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGGTCAACGAGTTTTCACAAGAGATTGCCAAGACGCTC
TTCACAGGGGACGTGCAGAGCTACCTCATGCTGTTCATCAAGAAGGAGGAAGCAAAGGACACACTGGACACCTTTAAGGCTGTTGCCAGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCTGACGCTCCAGCAATGCGCCTGATTTATCT
GGATGAAGACCTGGCCAAGTATAAACCTGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGACTTACCT
GAAGTCTGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCTGTCTTTG
TGGAGTTTTATGCCCCATGGTGTGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGA
TGGACTCCACAGCCAATGAGCTGGAAGAGGTCGAAATCATGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGATGAGATCATCGAATATGATGGTGAG
AAGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GH/GH_variant_3_[Conus_bullatus]
ATGAAGTTTGCCACTGTTCTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTAAAACAAGAGGAAAAAGTCTATGTTTTAAAGGAAAAAAATTTCTACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGAGCCTGGCTCCTGTGTACAGCGAGACTGCTGGGAAATTGA
TGGAAGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGAGCAAGTTTGAAGTGAAAGGTTTCCCCACCATCAAATTCTTCG
TGAATGGCCAATCAGTGGATTACACAGGAGGCAGGCAGTCTTCCGACTTCTTCAACTGGCTGAAGAAGAAAACCGGCCCCCCTGCCAAGGATGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGTGATGAAGTCATCGTCTTGGGCTTCTTCAAGGACCAGGAAGGCGAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCGGGA
ATTGAGGACGTTGCCTTTGGCATCACATCAGAGGAGAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTGTGTTTAAAAAGTTTGACGAGGGTCGC
AGTGACTTCTCTGGAGATTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGGTCAACGAGTTTTCACAAGAGACTGCCCAGAAGATC
TTCGCAGGAGACATTCAGAGCCACCTCATGCTGTTCGTCAAGAAGGAGGAAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCTGACGCTCCGGCGATGCGCCTGATTCAGCT
GGGCGAAGATCTGGCCAAGTACAAACCTGAGTCGGACTCTCTGGACAAGTCCGTCATAACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTCACCT
GAAGTCAGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAAGAGGTGGCCATGGACAAATCCAAGGCTGTCTTTG
TGGAGTTTTATGCCCCATGGTGTGGACACTGCAAGAAGTTGGCTCCGATCTGGGACCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAG
ATGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGAACATCGAATATGATGGCGA
GAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAA
GGATGAACTCTAA
>csPDI_GH/GH_variant_4_[Conus_bullatus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACATTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGACGAAAAAGTCTATGTTTTAAAGGCAACGAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGAGCCTGGCTCTTGTGTACAGCGAGACTGCTGGGAAATTGAT
GGAAGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGGGCAAGTTCAACGTCACAAGTTACCCCACCATCAAATTCTTCAT
CAGCGGTGAACCAATGGAATACACTGGAGGCAGGCAGACTTCCAACTTCATCAACTGGCTGAAGAAGAAGACCGGCTCCCCTGCCAAGGATGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGACGAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCGGAA
ATTGAGGACATTGCCTTTGGCATCACATCAGAGGAGAGCGTCTTTAAAGAGCATAAGATGAAGAAGGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AGTGACTTCTCTGGAGATTTTGGGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGGTCAACGAGTTTTCACAAGAGATTGCCAAGACGCTC
TTCACAGGGGACGTGCAGAGCTACCTCATGCTGTTCATCAAGAAGGAGGAAGCAAAGGACACACTGGACACCTTTAAGGCTGTTGCCAGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCTGACGCTCCAGCAATGCGCCTGATTTATCT
GGATGAAGACCTGGCCAAGTATAAACCTGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGACTTACCT
GAAGTCTGAGGAAGTGCCAGAGGACTGGGACGCCAAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCTGTCTTTG
TGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATTATTGCCAAGA
TGGACTCCACAGCCAATGAGCTGGAAGAGGTCGAAATCATGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGCGAG
AAGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GA/GH_[Conus andremenezi]
ATGAAGTTTGCAACTGTTTTCTGTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACTAGAGGAAAATGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCATTGAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCGCCTGCAAGCATGTGGCTCCCTTGTACAGCAAGGCTGCTGGGAAATTGAT
GGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTCACCGTTGAGAAGGACTTGGCTGACAGGTTTGAAGTCACAGGCTACCCCACCATCAAATTCTTCAG
CAGTGGCGAACCAACGGAGTACACTGGAGGCAGGCAGCCTTCCGACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACCTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTACCTACAAGAAGATTGCCGCAGAA
ATTGAGGACGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAAAGTATAAGATGAAGAAAGATGGCGTTGTTCTTTTTAAAAAGTTTGACGAGGGTCGC
CATGACTTCTCTGGAGACTTTGAGGAGACTGCCATGAGAAAGTTCATCCAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGATCATC
TTTGCAGGGGACGTGAAGAACTACCTGATGCTGTTAGTCAAGAAGGAGGGAGCAAAGGACACACTAGACGCCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCATGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAAGACGTCATGAGATTCTTGGGCCTCAAGGCTGCCGATGCTCCAACGATGCGCCTGATTCAGCT
GGATGAAGAAAACGTGATCAAGTACAAACCTGAGTTTCTGGACAAGTCCACCATGACCAAGTTTGTTCATGACTACCTGGATGGCAAGCTGAAGCCTTACCTGAA
GACTGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCAGTCTTTGTGG
AGTTTTATGCCCCGTGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGATG
GACTCCACAGCCAATGAGCTGGAAGAGTTCCAAATCCAGGGTTACCCAACCTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGCGAGAA
GACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACCTCCGAAGGAAAAGGAGGAAGAAAAAGAGGAAGAGGAAGAAAAAGAGGAAG
ATGACGACAAGAAAAAGGACGAACTCTAA
>csPDI_GA/GH_[Conus praecellens]
```

```
ATGAAGTTTGCAACTGTTTTCTGTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACTAGAGGAAAATGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCATTGAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCGCCTGCAAGCATGTGGCTCCCTTGTACAGCAAGGCTGCTGGGAAATTGAT
GGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTCACCGTGGAGAAGGACTTGGCTGACAGGTTTGAAGTCACAGGCTACCCCACCATCAAATTCTTCAG
CAGTGGCGAACCAACGGAGTACACTGGAGGCAGGCAGCCTTCAGACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACCTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCATCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGCAGAA
ATCATGGACATTGTCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAAAGTATAAGATGAAGAAAGATGGCGTTGTTCTTTTTAAAAAGTTTGACGAGGGTCGC
CATGACTTCTCTGGAGACTTTGAGGAGACTGCCATGAGAAAGTTCATCCAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACGGGAGACCTCCCAGATCATC
TTTGCAGGGGACGTGAAGAACTACCTGATGCTGTTAGTCAAGAAGGAGGGAGCAAAGGACACACTGGACGCCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAAGACGTCATGAGATTCTTGGGCCTCAAGGCTGCCGATGCTCCGACGATGCGCCTGATTCAGCT
GGACGAAGAAAACGTGATCAAGTACAAACCTGAGTTTCTGGACAAGTCCACTATGACCAAGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTTACCTGAA
GACTGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCAGTCTTTGTGG
AGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGTTCCGATCTGGGATCAGTTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGATGG
ACTCCACAGCCAATGAGCTGGAAGAGTTCCAAATCCAGGGTTACCCAACTTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGCGAGAAG
ACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACCTCCGAAGAAAAAGGAGGAAGAAAAAGAGGAAGAGGAAGAAAAAGAGGAAGAT
GACGACAAGAAAAAGGATGAACTCTAA
>csPDI_GA/GH_variant_2_[Conus praecellens]
ATGAAGTTTGCAACTGTTTTCTGTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACTAGAGGAAAATGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCATTGAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCGCCTGCAAGCATGTGGCTCCCTTGTACAGCAAGGCTGCTGGGAAATTGAT
GGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTCACCGTGGAGAAGGACTTGGCTGACAGGTTTGAAGTCACAGGCTACCCCACCATCAAATTCTTCAG
CAGTGGCGAACCAACGGAGTACACTGGAGGCAGGCAGCCTTCAGACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACCTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCATCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGCAGAA
ATCATGGACATTGTCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAAAGTATAAGATGAAGAAAGATGGCGTTGTTCTTTTTAAAAAGTTTGACGAGGGTCGC
CATGACTTCTCTGGAGACTTTGAGGAGACTGCCATGAGAAAGTTCATCCAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACGGGAGACCTCCCAGATCATC
TTTGCAGGGGACGTGAAGAACTACCTGATGCTGTTAGTCAAGAAGGAGGGAGCAAAGGACACACTGGACGCCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGACGCTCCGGCGATGCGCCTGATTCATCT
GACCGAAGAAGACCTGACCAAGTACAAGCCCGAGTCAGACACTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTT
ACCTGAAGACTGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCAGTC
TTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGTTCCGATCTGGGATCAGTTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCC
AAGATGGACTCCACAGCCAATGAGCTGGAAGAGTTCCAAATCCAGGGTTACCCAACTTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGG
CGAGAAGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACTTCCGAAGAAAAAGGAGGAGGAAGAAAAAGAGGAAGACGACGACAA
GAAAAAGGATGAACTCTAA
>csPDI_GA/GH_Gvariant_3_[Conus praecellens]
ATGAAGTTTGCAACTGTTTTCTGTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACTAGAGGAAAATGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCATTGAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCGCCTGCAAGCATGTGGCTCCCTTGTACAGCAAGGCTGCTGGGAAATTGAT
GGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTCACCGTGGAGAAGGACTTGGCTGACAGGTTTGAAGTCACAGGCTACCCCACCATCAAATTCTTCAG
CAGTGGCGAACCAACGGAGTACACTGGAGGCAGGCAGCCTTCAGACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACCTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCATCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGCAGAA
ATCATGGACATTGTCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAAAGTATAAGATGAAGAAAGATGGCGTTGTTCTTTTTAAAAAGTTTGACGAGGGTCGC
CATGACTTCTCTGGAGACTTTGAGGAGACTGCCATGAGAAAGTTCATCCAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACGGGAGACCTCCCAGATCATC
TTTGCAGGGGACGTGAAGAACTACCTGATGCTGTTAGTCAAGAAGGAGGGAGCAAAGGACACACTGGACGCCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGACGCTCCGGCGATGCGCCTGATTCATCT
GACCGAAGAAGACCTGACCAAGTACAAGCCCGAGTCAGACACTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTT
ACCTGAAGACTGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCAGTC
TTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCC
AAGATGGACTCCACAGCCAATGAGCTGGAAGAGTTCCAAATCCAGGGTTACCCAACTTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGG
CGAGAAGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACTTCCGAAGAAAAAGGAGGAAGAAAAAGAGGAAGAGGAAGAAAAAGA
GGAAGATGACGACAAGAAAAAGGATGAACTCTAA
>csPDI_GA/GH_variant_2[Conus andremenezi]
ATGAAGTTTGCAACTGTTTTCTGTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACTAGAGGAAAATGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCATTGAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCGCCTGCAAGCATGTGGCTCCCTTGTACAGCAAGGCTGCTGGGAAATTGAT
GGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTCACCGTTGAGAAGGACTTGGCTGACAGGTTTGAAGTCACAGGCTACCCCACCATCAAATTCTTCAG
CAGTGGCGAACCAACGGAGTACACTGGAGGCAGGCAGCCTTCCGACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACCTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTACCTACAAGAAGATTGCCGCAGAA
ATTGAGGACGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAAAGTATAAGATGAAGAAAGATGGCGTTGTTCTTTTTAAAAAGTTTGACGAGGGTCGC
CATGACTTCTCTGGAGACTTTGAGGAGACTGCCATGAGAAAGTTCATCCAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGATCATC
TTTGCAGGGGACGTGAAGAACTACCTGATGCTGTTAGTCAAGAAGGAGGGAGCAAAGGACACACTAGACGCCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCATGTTTATCTACCTGGACACAGCCAAGAAGGAGAATGAAGACGTCATGGGACTCTTGGGCTTCAAGGCTGCTGACGCTCCGGCGATGCGCCTGATTCATCT
GACCGAAGAAGACCTGACCAAGTACAAGCCCGAGTCAGACACTCTGGACAAGTCCACCATGACCAAGTTTGTTCATGACTACCTGGATGGCAAGCTGAAGCCTT
ACCTGAAGACTGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCAGTC
TTTGTGGAGTTTTATGCCCCGTGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCC
AAGATGGACTCCACAGCCAATGAGCTGGAAGAGTTCCAAATCCAGGGTTACCCAACCTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGG
CGAGAAGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACCTCCGAAGGAAAAGGAGGAAGAAAAAGAGGAAGAGGAAAAAGAGGA
AGATGACGACAAGAAAAAGGACGAACTCTAA
>csPDI_GH/GH_[Conus andremenezi]
ATGAAGTTTGCAACTGTTTTCTGTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACTAGAGGAAAATGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCATTGAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCCACTGCAAGCAACTGGCTCCAGTGTACAGCGAGGTTGCAGGGAAATTGA
TGGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGGGCAAGTTTGAAGTCAAAGGCTACCCCACCATCAAATTCTTCC
GCAATGGGGAACCAATGGAGTACACTGGAGGCAGGCAGACTTCCGACTTCATCACCTGGCTGAAGAAGAAGACCGGCCCCCCTACCACGGACCTGAAGACTTCT
GAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGCGGA
AATCATGGACATTGTCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAAAGTATAAGATGAAGAAAGATGGCGTTGTTCTTTTTAAAAAGTTTGACGAGGGTCG
CCATGACTTCTCTGGAGACTTTGAGGAGACTGCCATGAGAAAGTTCATCCAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGATCAT
CTTTGCAGGGGACGTGAAGAACTACCTGATGCTGTTAGTCAAGAAGGAGGGAGCAAAGGACACACTAGACGCCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGA
AGGTCATGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAAGACGTCATGAGATTCTTGGGCCTCAAGGCTGCCGATGCTCCAACGATGCGCCTGATTCAGC
TGGATGAAGAAAACGTGATCAAGTACAAACCTGAGTTTCTGGACAAGTCCACCATGACCAAGTTTGTTCATGACTACCTGGATGGCAAGCTGAAGCCTTACCTGA
AGACTGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCAGTCTTTGTG
GAGTTTTATGCCCCGTGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGAT
GGACTCCACAGCCAATGAGCTGGAAGAGTTCCAAATCCAGGGTTACCCAACCTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGCGAGA
AGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACCTCCGAAGGAAAAGGAGGAAGAAAAAGAGGAAGAGGAAGAAAAAGAGGAA
GATGACGACAAGAAAAAGGACGAACTCTAA
>csPDI_GA/GH_[Conus bocki]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCAGTGAAGAAGTTGAACAAGAGGGAAAAGTCTATGTTTTAAAGGCAACAAACTTCGAC
AACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCATGGTGCGGCGCGTGCAAGCACCTGGCTCCCGTGTACAGCGAGGCTGCTGGGAAATTG
ATGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTTACCGTGGAGACGGACTTGGCGGACAAGTTCAACATCAGAAGTTACCCCACCATCAAATTCTTC
```

```
ATCAGTGGCGAACCAATAGAGTACACTGGAGGCAGGCAGACTTCCGACATCATCATCTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGATGTGAAGACTTC
TGAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATTGCCGTGG
AAATCGAGGACGTTTTCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTC
GCAATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCATCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGACGAGAAGC
TCTTCACAGGGGATGTGCACAGTTACCTCATACTGTTCGTCAAGGAGGGAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAAG
GTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACGCATCATGGAATTCTTCGGCCTCAAAGCTGCTGACACTCCGGCGATGCGCCTGATTTATCTG
GGCGAAGACCAGGCTAACTACAAACCTGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTTACCTG
AAGTCGGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGATTTTGGTCAGCAAGAACTTCAATGAGGTGGCCATGGACAAATCCAAGGCCGTCTTTGT
GGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGA
TGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGAAGGCGAG
AGGACCCTGGAAGAACTGACTAAATTCGTGGAAAGTGGCGGAAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAGGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GA/GH_variant_2_[Conus bocki]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCAGTGAAGAAGTTGAACAAGAGGGAAAAGTCTATGTTTTAAAGGCAACAAACTTCGAC
AACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCATGGTGCGGCGCGTGCAAGCACCTGGCTCCCGTGTACAGCGAGGCTGCTGGGAAATTG
ATGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTTACTGTGGAGAAGGACTTGGCGGACAAGTTCAACGTCACAAGTTACCCCACCATCAAATTCTTC
ATCAGTGGCGAACCAATGGAATACACTGGAGGCAGGCAGACTTCCGGCTTCATCATCTGGCTGAAGAAGAAGACCGGCCCCCCTGTCAAGGACGTGAAGACTTC
TGAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCAG
AAATCGAGGACGTTTCCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTC
GCAATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCATCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGACGAGAAGC
TCTTCACAGGGGATGTGCACAGTTACCTCATACTGTTCGTCAAGGAGGGAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAAG
GTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACGCATCATGGAATTCTTCGGCCTCAAAGCTGCTGACACTCCGGCGATGCGCCTGATTTATCTG
GGCGAAGACCAGGCTAACTACAAACCTGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTTACCTG
AAGTCGGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGATTTTGGTCAGCAAGAACTTCAATGAGGTGGCCATGGACAAATCCAAGGCCGTCTTTGT
GGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGA
TGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGAAGGCGAG
AGGACCCTGGAAGAACTGACTAAATTCGTGGAAAGTGGCGGAAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAGGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GH/GH_[Conus bocki]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCAGTGAAGAAGTTGAACAAGAGGGAAAAGTCTATGTTTTAAAGGCAACAAACTTCGAC
AACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGCACCTGGCTCCCGTGTACAGCGAGGCTGCTGGGAAATTG
ATGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACTGTGGAGAAGGACTTGGCGGGCAAGTTTGAAGTCAAAGGTTTCCCCACCATCAAATTCTTC
ATTAGTGGCAAACCAGTGGATTACACAGGAGGCAGGCAGACTTCCGACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGATGTGAAGACTTC
TGAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATTGCCGTGG
AAATCGAGGACGTTTTCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTC
GCAATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCATCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGACGAGAAGC
TCTTCACAGGGGATGTGCACAGTTACCTCATACTGTTCGTCAAGGAGGGAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCAGTGAATTCAAGGGGAAG
GTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACGCATCATGGAATTCTTCGGCCTCAAAGCTGCTGACACTCCGGCGATGCGCCTGATTTATCTG
GGCGAAGACCAGGCTAACTACAAACCTGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTTACCTG
AAGTCGGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGATTTTGGTCAGCAAGAACTTCAATGAGGTGGCCATGGACAAATCCAAGGCCGTCTTTGT
GGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGA
TGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGAAGGCGAG
AGGACCCTGGAAGAACTGACTAAATTCGTGGAAAGTGGCGGAAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAGGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GA/GH_[Conus eburneus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTATGTGGCCTGTGAAGAAGTTGAACAAGAAGAAAAAGTCTATGTTTTAAAGGCAACAAACTTCGAC
AAGTTCATTAATGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCATGGTGCGGCGCCTGCAAGAACCTGGCTCCGGTGTACAGCGAGGTTGCTGGGAAATTG
AAGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTCACTGTAGAGAAGGACTTGGCAGAAAAGTTCAACGTCTCAGGTTACCCCACCATCAAATTCTTC
AGCAGTGGCGAACCAACAGATTACACTGGAGGCAGGCAGACTTCCAACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGAGATGAAGACTTC
TGAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGATGCTGCTGCCTTCAAGAAGGTTGCCACAG
AAATTGACGATGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCATTGTTCTGTTTAAAAAGTTTGACGAGGGTC
GCAGTGACTTCTCTGGAGCCTTCGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTCTCACAGGAGACCGCCCAGAAGA
TCTTCTCAGGGGACGTGAAGACCCACCTCATGCTGTTCATCAAGAAGGAGGGAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGG
AAGGTGATGTTTATCTACCTGGACATAGCCAAGGAGGAGAGTGAACGCATCTTGGAATTCTTCGGCCTCAAGGCTGCTGACACTCCAGTGATGCGCCTGATTCAG
CTGGGCGAAGAGCTGGTCAAGTACAAACCCGAGTCGGACTGTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTCAC
CTGAAGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTCTT
TGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTAGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATGTCATCATCGCCAA
GATGGACTCCACAGCCAATGAGCTGGAAGAGGTCAAAATCCAAGGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCGTTGATTATGATGGTG
AGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCCAGAACCTCCGAAGAAAAAGGAGGAAGAAGAAGAGGAAGATGCCGACAAGAAAA
AGGATGAACTCTAA
>csPDI_GA/GH_[Conus tessulatus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTATGTGGCCTGTGAAGAAGTTGAACAAGAAGAAAAAGTCTATGTTTTAAAGGCAACAAACTTCGAC
AAGTTCATTAATGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCATGGTGCGGCGCCTGCAAGAACCTGGCTCCGGTGTACAGCGAGGTTGCTGGGAAATTG
AAGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTCACTGTAGAGAAGGACTTGGCAGAAAAGTTCAACGTCTCAGGTTACCCCACCATCAAATTCTTC
AGCCGTGGCGAACCAACAGATTACACTGGAGGCAGGCAGACTTCCAACTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGAGATGAAGACTTC
TGAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGATGCTGCTGCCTTCAAGAAGGTTGCCGCGA
AAATCGAGGACGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTCGTTCTGTTTAAAAAGTTTGACGAGGGT
CGCAGTGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGAAG
ATCTTCTCAGGGGACGTGCAGAGCCACCTCATGCTGTTCATCAAGAAGGAGGGAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGG
GAAGGTCCTGTTTATCTACCTGGACACGACCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGATGCTCCGGCGATGCGCCTCATTCA
GCTGGGCGAAGACCTGGCCAAGTACAGACCAGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTC
ACCTGAAGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTC
TTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATGTCATCATCGCC
AAGATGGACTCCACAGCCAATGAGCTGGAAGAGGTCAAAATCCAAGGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCGTTGATTATGATGG
TGAGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCCAGAACCTCCGAAGAAAGAGGAGGAAGAAGAAGAGGAAGATGCCGACAAGAA
AAAGGATGAACTCTAA
>csPDI_GH/GH_[Conus eburneus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAAAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAAAGACAAAAAATTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGCAACTGGCTCCCGTGTACAGCGAGGCTGCTGGGAAATTGA
AGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGACGCCACCGTGGAGAAGGACTTGGCGGAAAAGTTTGAAGTCAAAGGTTTCCCCACCATCAAATTCTTC
AGCAGTGGCGAACCAGTGGATTACACAGGAGGCAGGCAGACTTCTGACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGAGATGAAGACTTC
TGAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGATGCTGCTGCCTTCAAGAAGGTTGCCGCGA
AAATCGAGGACGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTC
```

```
GCACTGACTTCTCTGGAGACTTTGAGGAGGCTCCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGAAGA
TCTTCTCAGGGGACGTGCAGAGCCACCTCATGCTGTTCATCAAGAAGGAGGGAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGG
AAGGTCCTGTTTATCTACCTGGACACGACCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGATGCTCCGGCGATGCGTCTGATTCAG
CTGGGCGAAGACCTGGCCAAGTACAGACCAGAGTCGGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTCAC
CTGAAGTCAGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTCTT
TGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATGTCATCATCGCCAA
GATGGACTCCACAGCCAATGAGCTGGAAGAGGTCAAAATCCAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCGTTGATTATGATGGTG
AGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAAAAAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAA
AGGATGAACTCTAA
>csPDI_GH/GH_[Conus tessulatus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAAAAGTTGAACGAGAGGAAAAAGTCTATGTTTTAAAGACAAAAAATTTCGACA
ACTTCATAGAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGCAACTGGCTCCCGTGTACAGCGAGGCTGCTGGGAAATTGA
TGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGGAAAAGTTTGAAGTCAAAGGTTTCCCCACCATCAAATTCTTCA
GCAGTGGCGAACCAGTGGATTACACAGGAGGCAGGCAGACTTCTGACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGAGGTGAAGACTTCT
GAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGATGCTGCTGCCTTCAAGAAGGTTGCCACAGA
AATTGACGATGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCG
CAGTGACTTCTCTGGAGACTTTGAGGAGGCTCCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGAAGAT
CTTCTCAGGGGACGTGCAGAGCCACCTCATGCTGTTCATCAAGAAGGAGGGAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGA
AGGTGATGTTTATCTACCTGGACATAGCCAAGGAGGAGAATGAACGCATCTTGGAATTCTTCGGCCTCAAGGCTGCTGACACTCCAGTGATGCGCCTGATTCAGC
TGGGCGAAGACCTGGTCAAGTACAAACCCGAGTCAGACTGTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTCACC
TGAAGTCGGAGGAACTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTCTTT
GTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATGTCATCATCGCCAA
GATGGACTCCACAGCCAATGAGCTGGAAGAGGTCAAAATCCAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCGTTGATTATGATGGTG
AGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCCAGAACCTCCAAAGAAAAAGGAGGAAGAAAAAGAAGAAGATGACGACAAGAAAA
AGGAGGAACTCTAA
>csPDI_GA/GH_[Conus marmoreus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACGAGAGGAAAACGTCTATGTTTTAAAGGCAACAAACTTCGACA
ACTTCATTAATGAAAATGAATTTGTTCTTGTGGAGTTCTATGCACCATGGTGCGGTGCCTGCAAGAACCTGGCTCCCGTGTACAGTGAGGTTGCTGGGAAATTGAT
GGACGATGGTTCCAACATCAAGCTGGCCAAGGTTGATGTTACCGCGGAGACGGACTTGGCGGACAAGTTCAACGTCACAGGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAATGCAATACACTGGAAGCAGGCAGACTTCTGGCTTCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGATGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGTGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGACAAGGGCGCTGCTGCTTTCAAGAAGGTTGCTGCGGAA
ATTGACAACATTGCCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGATGAGGGTCGC
AATGACTTCACTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCATCAAGGACAATCACCTCCCTCTGATCAACGAGTTTACACGGGAGACTGCCGAGAAGCTC
GTCACAGGGGATGTGCGGAGCTTCCTCATGCTGTTTGTCAAGAAGGAGGGAGCAAAGGACATACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCATAGACGAGAATGAACGCATCATGGGAATCTTCGGCCTCAAGGCTGCTGACACTCCGGCAATGCGCCTGATTCAGCT
GGGCAAAGAAGACATGGTCAAGTACAAACCTGAGTCGGACTCTCTGGACAAGTCCACCATGACCGAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTC
ACCTGAAGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCCGTGAAGGTTTTGGTCACCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTT
TTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGACATCATCATCGCC
AAGATGGACTCCATAGCCAATGAGCAGGAAGATGTCAAAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGATGAGATCATCGAATATGATGG
TGAGAGGACCCTGGAAGAACTGATCAAATTTGTGGAAAGTGGCGGCAAGCAAGAACCTCTGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAA
AAAGGATGAACTCTAA
>csPDI_GA/GH_variant_2[Conus marmoreus]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACGAGAGGAAAACGTCTATGTTTTAAAGGCAACAAACTTCGACA
ACTTCATTAATGAAAATGAATTTGTTCTTGTGGAGTTCTATGCACCATGGTGCGGTGCCTGCAAGAACCTGGCTCCCGTGTACAGTGAGGTTGCTGGGAAATTGAT
GGACGATGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGGAAAAGTTTGAAGTCAAAGGTTTCCCCACCATCAAATTCTTCAT
CAGTGGTGAACCAATGGATTACACAGGAGGCAGGCAGACTTCCGACATCATCAGCTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGATGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTTTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGCGAAA
ATCGAGGACGTTGCCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGATGAGGGTCGC
AATGACTTCACTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCATCAAGGACAATCACCTCCCTCTGATCAACGAGTTTACACGGGAGACTGCCGAGAAGCTC
GTCACAGGGGATGTGCGGAGCTTCCTCATGCTGTTTGTCAAGAAGGAGGGAGCAAAGGACATACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCATAGACGAGAATGAACGCATCATGGGAATCTTCGGCCTCAAGGCTGCTGACACTCCGGCAATGCGCCTGATTCAGCT
GGGCAAAGAAGACATGGTCAAGTACAAACCTGAGTCGGACTCTCTGGACAAGTCCACCATGACCGAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCCC
ACCTGATGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACCTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTT
TTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGACATCATCATCGCC
AAGATGGACTCCATAGCCAATGAGCAGGAAGATGTCAAAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGATGAGATCATCGAATATGATGG
TGAGAGGACCCTGGAAGAACTGATCAAATTTGTGGAAAGTGGCGGCAAGCAAGAACCTCTGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAA
AAAGGATGAACTCTAA
>csPDI_GH/GH_[Conus marmoreus]
ATGAAGTTTGCGACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAAAAGTTGAAGAAGAGGAAAAAGTCTATGTTTTAAAGACAGAAAGTTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGAAACTGGCCCCCGTGTACAGTGAGGCTGCTGGGAAATTGA
TGGAGGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGGAAAAGTTTGAAGTCAAAGGTTTCCCCACCATCAAATTCTTCA
TCAGTGGTGAACCAATGGATTACACAGGAGGCAGGCAGACTTCCGACATCATCAGCTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGATGTGAAGACTTCT
GAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTTTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGCGAA
AATCGAGGACGTTGCCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCG
CAATGACTTCTCTGGAGACTTTGAGGAAGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGAAGAT
CTTCTCAGGGGATGTGCAGATCCACCTCATGCTGTTTGTCAAGAAGGAGGGAGCAAAGGACATACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGA
AGGTCCTGTTTATCTACCTGGACACAGCCATAGACGAGAATGAACGCATCATGGGAATCTTCGGCCTCAAGGCTGCTGACACTCCGGCAATGCGCCTGATTCAGC
TGGGCAAAGAAGACATGGTCAAGTACAAACCTGAGTCGGACTCTCTGGACAAGTCCACCATGACCGAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCCC
ACCTGATGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACCTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTT
TTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGACATCATCATCGCC
AAGATGGACTCCACAGCCAATGAGCTGGAAGAGGTCAAAATCAACAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGTGACGAGATTGTTGAATATGATGG
TGAGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGAAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAA
AAAGGATGAACTCTAA
>csPDI_GH/GH_variant_2[Conus marmoreus]
ATGAAGTTTGCGACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAAAAGTTGAAGAAGAGGAAAAAGTCTATGTTTTAAAGACAGAAAGTTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGAAACTGGCCCCCGTGTACAGTGAGGCTGCTGGGAAATTGA
TGGAGGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGACTTGGCGGAAAAGTTTGAAGTCAAAGGTTTCCCCACCATCAAATTCTTCA
TCAGTGGTGAACCAATGGATTACACAGGAGGCAGGCAGACTTCCGACATCATCAGCTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGATGTGAAGACTTCT
GAGGAGGCCAAGACTTTCATTGACAGTGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGACAAGGGCGCTGCTGCTTTCAAGAAGGTTGCTGCGGA
AATTGACAACATTGCCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCG
CAATGACTTCTCTGGAGACTTTGAGGAAGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGAAGAT
CTTCTCAGGGGATGTGCAGATCCACCTCATGCTGTTTGTCAAGAAGGAGGGAGCAAAGGACATACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGA
AGGTCCTGTTTATCTACCTGGACACAACCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCTGACGCTCCGGCGATGCGCCTGATTCAGC
```

```
TGGGCGAAGACCTGGCTAAGTACAAACCCGAGTCGGACTCTCTGGACAAGTCCACCATGACCGAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTCACC
TGAAGTCGGAGGAAGTGCCAGAGGACTGGGACGCCCAGCCCGTGAAGGTTTTGGTCACCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTTTTT
GTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGACATCATCATCGCCAA
GATGGACTCCACAGCCAATGAGCTGGAAGAGGTCAAAATCAACAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGTGACGAGATTGTTGAATATGATGGTG
AGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGAAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAA
AGGATGAACTCTAA
>csPDI_GH/GH_[Conus pulicarius]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAAAGGCAAAAAACTTCGACA
ACTTCGTTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCATGGTGCGGCCACTGCAAGAAACTGGCTCCTGAGTACAGCAAGGCTGCTGGGAAATTGA
TGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCATTCTGGAGAAGGACTTGGCGGGCAAGTTTGAAGTCAAAGGTTTCCCCACCATCAAATTCTTCA
TCAATGGTGAACCAGTGGATTACACAGGAGGCAGGCAGACTTCCGAAATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGAGGTGAAGACTTCT
GAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGACGCTGCTGCCTTCAAGAACGTTGCCATGAA
AATCGAGGACGTTGCCTTTGGCATCACATCAGACGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCG
CAATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTTTAATCAACGAGTTCTCACATGAGACTGCCCAGAAGCT
CTTCACAGGGGACGTGCAGATCCACCTCATGCTGTTCATCAACAAGGAGGAAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCAGTGAATTCAAGGGGA
AGGTCCTGTTTATCTACCTGGACACAGTCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGATGCTCCGGCGATGCGCCTGATTCAGC
TGGGCGAAGACCTGGCCAAGTACAAACCCGAGTCGGACTCTCTGGACAAGTCCACCATCACCAAGTTTGTTCAGGACTACTTGGATGGCAAGCTGAAGCCTCACC
TGAAGTCGGAGGAATTGCCCGAGGACTGGGACGCCAAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTCTTT
GTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAA
GATGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAAGAGTTTCCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCGTCGAATATGATGGCG
AGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGAGGCAAGCAAGAACCTCCGAAGAAAGAGGAGGAAGAAAAAGAGGAAGACGACGACAAGAAAA
AGGATGAACTCTAA
>csPDI_GA/GH_[Conus victoriae]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCTTGTGAAGAAGTTGAACAAGAGGAAAACGTCTATGTTTTAAAGGCAACAAACTTCGACA
ACTTCATTGATGAAAATGAATTTGTTCTTGTGGAGTTCTATGCACCATGGTGCGGTGCCTGCAAGAACCTGGCTCCCATGTACAGTGAGGTTGCTGGGAAATTGAT
GGATGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGTTACCGTGGAAAAGGACTTGGCGGCCAACTTCAACGTCACAGGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAATGGAATACACTGGAGGCAGGCAGACTTCCGACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCATGGACGTGAAGACTTGTG
AGGAGGCCAAGACTTTCATTGACAGCAATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGGTTGCTGCAGAA
ATTAATGATGTTGCTTTTGGCATCACGTCAGAGGACAACGTCTTTAAAGAGCATAAGATGAAGAAAGATGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGCGACTTTGAGGAAGCTGCCATGAGCAAGTTCATCAAGGACAATTACCTCCCTCTGATCAACGAGTTTACACAGGATGCCGCCCAGAAGCTG
TTCACAGGGGATGTGAAGAGCTATCTCATGCTGTTCGTCAAGAAGGAGGGAGCAAAGGACATATTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCATAGAGGAGCATGAACGTGTCATGGAAGTCTTCGGCCTCAAGGCTGCTGCCACTCCGGCGATGTGCCTGATTCAGCT
GGATGAAGACCTGGTCAAGTACAAACCCGAGTCAGACTCTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTTACCT
GAAGTCGGAGGAAGTACCAGAGGACTGGGACGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACCTCAAAGAGGTGGCCATGGACAAATCCAAGGCCGTTTTTG
TGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCTAAGA
TGGACTCCACAGCCAATGAGCAGGAGGAGGTCAAAATCAAGAGTTACCCAACCTTGAAATACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGTGAG
AGGACCCTGGAAGAACTGATCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCGAAGAGAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GH/EK_[Conus tribblei]
ATGAAGTTTACAACTGTTTTCTCTCTCACGCTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGATAAAGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCGTTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCCACTGCAAGCGCCTGGCTCCCGTCTACAGCGAGGTTGCTGGGAAGTTGAT
GGATGAGGGTTCCAACGTCAAGATGGCCAAGGTGGATGCCATCGTGGAGAAGGACTTGGCGGGCAAGTTTAAAGTCCAAGGCTTCCCCACCATGAAGTTCTTCAT
CAGTGGCACACCAACAGACTACACTGGAGGCAGGCAGACTTCGGACTTCATGAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGGTGCTGCTGCCTTCAAGAAGATTGCCGCGAAA
ATCGAGGACATTGCCTTTGGCATCACGTCCGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATAAGCAAGTTCATCAAGGACAATCGCCTCCCTCTGATCAACGAGGTTACACAGGAGACCATCCAGAAGATC
TTCACAGGGGATGTGAAGAGCCACCTCGTGCTGTTTGTCAAGAAGGAGGATGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCTGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGACGCTCCGGCGATGCGCGTGATTCAGAT
GGGCGAAGAAGACCTGGCCACGTACACACCCGAGTCGGATGCTCTGGACAAGTCCACCATAACCACGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTC
ACCTGATGTCGGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCGGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCGGTC
TTTGTGGAATTTTATTCCCCAAGGTGCGAAAAGTGCAAGCAGTTGGCTCCAATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCC
AAGATGGACTCCACAGCCAATGAGCTGGAAGAGGTCAAAATCAGGAGTAACCCAACTTTGACATACTTCCCCAAGGGCAGCGACGAGATCATCAATTATGATGG
CAAGTTGACCCTGGAAGAACTGACCAAATTTGTGGAAAGTGGCGGCAAACAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACAACAAGAA
AAAGGATGAACTCTAA
>csPDI_GH/GH_[Conus tribblei]
ATGAAGTTTACAACTGTTTTCTCTCTCACGCTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGATAAAGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCGTTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCCACTGCAAGCGCCTGGCTCCCGTCTACAGCGAGGTTGCTGGGAAATTGAT
GGACGAGGGTTCCAACGTCAAGCTGGCCAAGGTGGATGCCATCGTGGAGAAGGACTTGGCGGTCAAATTTGAAGTCCAAGGCTTCCCCACCATGAAGTTCTTCAT
CAGTGGCAAACCAACAGACTACACTGGAGGCAGGCAGGCTTCAGACTTCATGAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACGTGAAGACTTCTG
AGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGATCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCACGGAA
ATCGAGGATATTGCCTTTGGCATCACGTCAGAGGGCAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAGCGAGTTTACACAGGAGACTGCCCAGATCATC
TTCGCAGGGGATGTGAAGAGCCACCTCTTGCTGTTTGTCAGTAAGGAGGAAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCTGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACTTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCAGCCTCAAGGCTGCCGACGCTCCGGCGATGCGCCTGCTTCAGAT
GGGCGAAGACCTGGCCAAGTACAAACCCGAGTCAGACACTCTGGACAAGTCCACCATAACCACGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTCACC
TGAATTCGGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAAGTTTTGGTCGGCAAAAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCGGTCTTT
GTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGCAGTTGGCTCCAATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAG
ATGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAAGAGTTTCCCAACCTTGAAGTACTTCCCCAAGGGCAGCGATGAGATCATCGAATATAGTGGCCT
GAGGACCCTGGAAGAACTGACCAAATTTGTGGAAAGTGGCGGCAAACAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAA
GGATGAACTCTAA
>csPDI_GH/GH_[Conus varius]
ATGAAGTTTACAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTAAACAAGAGGAAAAAGTCTATGTTTTAACGACAAAAAACTTCGACA
ACTTCATTAAGGAAAATGAATATGTTCTTGTGGAGTTTTATGCACCGTGGTGTGGCCACTGCAAGCAGCTGGCTCCTGTGTACAGCGAGGTTGCAGGGAAATTGA
AGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACTGTGGAGAAGGACTTGGCGGGCAGGTTTGAAGTCAAAGGCTTCCCCACCATCAAATTTTTCC
ATAATGGCAATCCAACAGATTACATGGGAGGCAGACAGACTTCCGACTTCATCACATGGCTAAAGAAGAGGACCGGCCCCCCTGCCCAGGACTTGAAGTCTTCT
GAGGAGGCCAAGACTTTCATTGACAGCGATGAAGTCATCGTCGTGGGCTTCTTCAAGGACCAGGAAGGCAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGCAGA
AATCGAGGACGTCGTCTTTGGCATCACGTCAGAGGACAGCATCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTAGTTTTGTTTAAAAAGTTTGACGAGGGTC
GCAGTGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCATCAAGGACAATCGCCTCCCTCTAATCAATGAGTTTACACAGGAGACCGCCCAGAAAA
TCTTCACAGGGGATGTGAAGAGCCACCTCATGCTGTTTGTCAAGAAGGAGGGAGCAGAGGACACACTGGACACCTTTAAGGCTGCTGCTGGTGAATTCAAGGGG
AAGGTCCTGTTTGTCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTCGGTCTCAAGGCTGCCGATGCTCCGGCGATGCGCCTGATTCAG
ATGGGTGATGACCTGACCAAGTACAAACCTGAGTCAGACACTATGGACAAGTCCACCATAACCAAGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTCAC
CTGAAGTCGGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAAGTGGCCATGGACAAATCCAAGGCGGTCTT
TGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGCAGTTGGCTCCGATCTGGGATCAGCTGGGTGACAAGTTCAAGGACAGCAAGGATATCATCATCGCCAA
```

```
GATGGACTCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAGGAGTTTCCCAACCTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGCG
AGAGGACCCTGGAAGAACTGACCAAATTTGTGGAAAGTGGTGGCAAACAAGAACCTCCGAAGAAAAAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAA
AGGATGAACTCTAA
>csPDI_GH/GH_[Conus generalis]
ATGACGTTTGCAACTGTTTTCTCTCTCATGTGGCTGGCCTTTGTGGCCTGTCTAGTTGAACAAAAGGAAAAAGTCCATGATTTAAAGGCGGAAAATTTCGACAACT
TCATTAAGGAACATGAATTTGCTCTTGTGGAATTTTACGCACCGTGGTGTGGCCACTGCAAGCAGCTGGCTCCTGTGTACAACGAAGTTGCTGGGAAATTGATGG
ATGAGGGTTCAAACATCAAGCTGGCCAAGGTTGATGCCACTGTGGAGAAGGACTTGGCGGGCAGGTTTGAAGTCAAACGCTTCCCCACCATCAAATTCTTCCAAA
GTGGCAAACCAACAGATTACACTGGAGGAAAGCAGACTTCCGACTTCATCAGCTGGCTGAAGAAGAGGACCGGCCCCCTTGTCACAGACCTGAAGACTTCTGAG
GAGGCCAAAAATTTTATTGACAGCGATGAAGTCATCGTCATGGGCTTCTTCAAGGACCAGGAAGGTAAGGGCGCTGCTGCCTTCAAGAAGATTGCCGCAGAAAT
CGAGGACATTGCCTTTGGCATCACATCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCGTTGTTCTGTTTAAAAAGTTTGACGAGGGTCGCA
GTGACTTCTCTGGAGACTTTGAGGAGGATGCCATGAGCAAGTTCATCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACAGAGGAGACCGCCCAGAAAATCT
TCTCAGGGGATGTGAAGAGCCACCTCATGCTGTTTGTCAAGAAGGAGGAAGCAAAGGACACACTGGACACCTTTAACGCTGCTGCTGGTGAATTCAAGGGGAAG
GTCCTGTTTATCTACCTGGACACAGCCAAGGAGGAGAATGAACACATCATGGGATTCTTTGGCCTCAAGGCTACCCACCCTCCGGCGATGCGCCTGATTTATCTGG
GCGATGACCTGGCCAAGTACAAACCCGAGTCAGACACTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTCACCTG
AAGTCAGAGGAAGTGCCAGAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCGGTCTTTGT
GGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCAATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCCAAGA
TGGACGCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAGCAGTTTCCCAACCTTGAAGTACTTCCCCAAGGGCAGCGATGAGGTCATCGAATATGATGGTGAG
AGGACCCTGGAAGAACTGACCAAATTCATGGAAAGTGGCGGCAAACAAGAACCTCCAAAGAAAAAGGAGGAAGAAAAAGAGGAAGATGACGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GA/GH_[Conus virgo]
ATGAAGTTTGCAACTGTTTTCTCCCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAACGGCAACAAACTTCGACA
AATTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCATGGTGCGGTGCCTGCAAGCAACTGGCTCCTGTGTACAGCGAGGTTGCTGGAAAATTGAT
GGATGAGGATTCCAACATCAAGCTGGCCAAGGTTGATGTCACCGTGGAGAAGGACTTGGCAGACAAGTTCAACGTCACAGGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAACAGATTACACTGGAGGCAGACAGACTTCCAACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACGTGAAGACTTCTG
AGGAGGCCAAGACTTTGATTGACAGCGATGAAGTCATTGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGCTGCTGCTGCCTTCAAGAAAACTGCTGTGAAA
ATCGAGGACGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCATTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCATCGAGTTTACACAGGAGACTGCCCAGATCTTA
TTCACAGGGGACGTGCAGAGCCACCTCCTGCTGTTCATCAAGAAGGAGGGAGCAGAGGACATACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAACCAAGGAGGACAATGAACACGTCATGGAATACTTCGGCGTCAAGGCTGCTGACGTGCCGGCAATGCACCTGATTCATCT
GGGTGAAGAAGACCTGACCAAGTACAAGCCCGAGTCGGACACTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTC
ACCTGAAGTCAGAGGAATTGCCAGAGGACTGGGATGCCCAGCCCGTGAAGGTTTTGGTCAAAAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTC
TTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGGGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCC
AAGATGGACGCCACAGCCAATGAGCTGGAAGAGGTCCAAATCGCTAGTTTCCCAACCTTGAAATACTTCCCTAAGGGCAGCGATGAGATCATTGAATATGAAGG
CGAGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGACGACGACAAGAA
AAAGGATGAGCTCTAA
>csPDI_GA/GH_variant_2[Conus virgo]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAACGGCAACAAACTTCGACA
AATTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCATGGTGCGGTGCCTGCAAGCAACTGGCTCCTGTGTACAGCGAGGTTGCTGGAAAATTGAT
GGATGAGGATTCCAACATCAAGCTGGCCAAGGTTGATGTCACCGTGGAGAAGGACTTGGCAGACAAGTTCAACGTCACAGGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAACAGATTACACTGGAGGCAGACAGACTTCCAACATCATCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACGTGAAGACTTCTG
AGGAGGCCAAGACTTTGATTGACAGCGATGAAGTCATTGTCATGGGCTTCTTCAAGGACCAGGAAGGCAAGGCTGCTGCTGCCTTCAAGAAAACTGCTGTGAAA
ATCGAGGACGTTGCCTTTGGCATCACGTCAGAGGACAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCATTGTTCTGTTTAAAAAGTTTGACGAGGGTCGC
AATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCATCGAGTTTACACAGGAGACTGCCCAGATCTTA
TTCACAGGGGACGTGCAGAGCCACCTCCTGCTGTTCATCAAGAAGGAGGGAGCAGAGGACATACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGGAA
GGTCCTGTTTATCTACCTGGACACAACCAAGGAGGACAATGAACACGTCATGGAATACTTCGGCGTCAAGGCTGCTGACGTGCCGGCAATGCACCTGATTCATCT
GGGTGAAGAAGACCTGACCAAGTACAAGCCCGAGTCGGACACTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTC
ACCTGAAGTCAGAGGAATTGCCAGAGGACTGGGATGCCCAGCCCGTGAAGGTTTTGGTCAAAAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTC
TTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATCAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATCATCGCC
AAGATGGACGCCACAGCCAATGAGCTGGAAGAGGTCCAAATCAAGGGTTTCCCAACCTTGAAATACTTCCCTAAGGGCAGCGATGAGATCATTGAATATGAAGG
CGAGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGAAGAGGAAGACGA
CGACGATGAACTCTAA
>csPDI_GH/GH_[Conus virgo]
ATGAAGTTTGCAACTGTTTTCTCCCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGAAAAAGTCTATGTTTTAACGGCAACAAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCATGGTGCGGCCACTGCAAAAATCTGGCTCCTGTGTACAGCGAGGTTGCTGGAAAATTGAT
GGATGAGGATTCCAACATCAAGCTGGCCAAGGTTGATGCCATCGTGGAGAAGGACTTGGCTGGCAAGTTTGAAGTCAAAGGTTACCCCACCATCAAATTCTTCAT
CAGTGGCGAACCAGTAGAATACACAGGAGGCAGACAGACTGCCGACATCGTCAACTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACCTGAAGACTTCT
GAGGAGGCCAAGACTTTGATTGACAGCGATGAAGTCATTGTCGTGGGCTTCTTCAAGGACCAGGAAGGCAAGGCTGCTGCTGCCTTCAAGAAAACTGCTGTAGA
AATCGAGGACGTTGCCTTTGGCATCACGTCAGAGGAAAGCGTCTTTAAAGAGCATAAGATGAAGAAAGACGGCATTGTTCTGTTTAAAAAGTTTGACGAGGGTC
GCAATGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATCGCCTCCCTCTGATCAACGAGTTTACACAGGAGACCGCCCAGAAGA
TCTTCGCAGGGGACGTGCAGAGCCACCTCCTGCTGTTCATCAAGAAGGAGGGAGCAGAGGACATACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGG
AAGGTCCTGTTTATCTACCTGGACACAACCAAGGAGGACAATGAACACGTCATGGAATACTTCGGCGTCAAGGCTGCTGACGTGCCGGCAATGCACCTGATTCAT
CTGGGTGAAGAAGACCTGACCAAGTACAAGCCCGAGTCGGACACTCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCC
TCACCTGAAGTCGGAGGAATTGCCAGAGGACTGGGATGCCCAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCG
TCTTTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGGAGTTGGCTCCGATCTGGGATCAGCTGGGGGAGAAGTTCAAGGACAGCAAGGATATCATCATCG
CCAAGATGGACGCCACAGCCAATGAGCTGGAAGAGGTCCAAATCGCTAGTTTCCCAACCTTGAAATACTTCCCTAAGGGCAGCGATGAGATCGTTGAATATGATG
GTGAGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGCGGCAAGCAAGAACCTCCAAAGAAAGAGGAGGAAGAAAAAGAGGAAGAAGAGGAAGACG
ACGACAAGAAAAAGGATGAGCTCTAA
>csPDI_GH/GH_[Conus distans]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTTCTGGCCTTTGTGGCCTGTGAAGAAGTTGAACAAGAGGAAAATGTCCATGTTTTAAATTCAAAAAACTTTGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTATGCACCGTGGTGCGGCCACTGCAAGAAGCTGGCTCCCGTGTACAGCGAGGTTGCTGGGAAATTGA
TGGACGAGGGTTCCAACATCAAGCTGGCCAAGGTTGATGCCACCGTGGAGAAGGATTTGGCTGGCAAGCTTGAAGTCCATGGCTACCCCACCATCAAATTCTTCC
ACAGTGGCAAATCGGAGGAGTACACCGGAGGCAGGAAGACTTCCGACTTCATCATCTGGCTGAACAAGAAGACTGGTTCCCCTGCCAAGGACCTGAAGACTTCT
GAGGAGGCCAAGACTTTCATTGACAGCAGTGAAGTCATCGTTGTGGGCTTCTTCAAGGACCAGGAAGGCAAGGACGCTACAGCCTTCAAGAAGATTGCTGCGGA
AATCGAGGACATTGCCTTTGGCATCACATCAGAGGACAGTGTCTTCAAAGAGCATAAGATGGAGAAAGACGGCGTTGTTGTGTTTAAAAAGTTTGACGAGGGTC
GCAGTGACTTCTCTGGAGACTTTGAGGAGGCTGCCATGAGCAAGTTCGTCAAGGACAATAGCCTCCCTCTCATCAACGAGTTTACACAGGAGAATGCCCAGAAGA
TCATGTCAGGGGACGTGAAAAGCCACCTCATGCTGTTCGTCAAGAAGGAGGCAGCAAAGGACACACTGGACACCTTTAAGGCTGCTGCCGGTGAATTCAAGGGG
AAGGTCCTGTTTGTCTACCTGGACGCAGACAAGAAGGAGAATGAACACATCATGGGATTCTTCAGCCTCAAGGCTGACGACGCTCCAGCAATGCGCCTGGTTCAG
CTGGGTGAAAACCTGGCCAAGTACAGACCCGAGTCAGACAATCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTCA
CCTGAAGTCGGACGAAGTGCCAGAGGACTGGGATGCCAAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCCTGGACAAATCCAAGGCCGTCT
TTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCCCCAGTCTGGGATCAGCTGGGTGACAAGTTCAAGGACAGCGAGGATATCATCATCGCCA
AGGTGGACGCCACAGCCAATGAGCTGGAAGAGGTCCAAATTAACAGTTATCCAACCTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATAATGGT
GGGAGGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACCTCCGAAGAAAAAGGAGGAAGAAAAAGAGGAAGATGACAAGAAGAAA
AAGGAAGAACTCTAA
```

```
>csPDI_GH/AH_[Conus planorbis]
ATGAAGTTTGCAACTATTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAGCAAGAGGATAATGTCTATGTTTTAAACGAAAACAACTTCGACG
ACTTCATCAGGGGAAATGAATTTGTTCTTCTGGAGTTTTATGCACCGTGGTGCGGCCACTGCAAGCGCCTGGCTCCAGTGTACAGTGAAGCTGCAGGGAAATTGA
TGCAAGAGGGTTCCAAAGTCAAGCTGGCCAAGATTGATGCCACTGTCGAGAAGGGCTTGGCGAGCAGGTATGAAGTCAAAGGCTACCCCACCATCAAATTCTTC
AACAAGGGCATATCATCAGATTACACGGGAGGCAGGCAGACTTCCGAGATCATCAACTGGCTGAAGAAGAGGACCGGCCCCCCTGCCCAGGTCCTGAAGTCTTC
TCAGGAGACCAAGACTTTCATTGACAGTGATGAAGTCGTCGTCTTGGGCTTCTTCAAGGACCAGGAAGGCGAGGACGCTGCTGCCTTCAAGGAGGTTGCCACAAA
AATCGAGGACATTGCCTTTGGCATCACGTCAGAGGATAGCCTCTTTAAAGAGCATAAGATGAAGAAAGATGGCGTTGTTTTATTTAAAAAGTTTGACGAGGGTCG
CAGTGACTTCTCTGGAGACTTTAAGGAGGATGCCATTAGCAAGTTCATCAAGGAAAGTCGCCTCCCTCTGATCAACGAATTTACAATGGAGACCGCCCAGAAAAT
CTTCTCAGGGGAAGTGAATAACCACCTCATGCTCTTTGTCAAGAAGGACGATGAAAAGGACACACTGGACACCTTTAAGGCTTGTGCTGGTGAATTCAAGGGAAA
GGTCCTGTTTGTCTACCTGGACACAGCCCGGGAGGATAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGATGCTCCGGCGATGCGCCTGATTCAGAT
GGGGGATGACCTGACGAAGTACAAACCCGAGTCGGACTCTCTGGACAAGTCCACCATAACCAAGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTCACC
TGAAGTCAGAGGAAGTGCCAGAGGACTGGGATGCCAAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCGGTCTTT
GTGGAGTTTTATGCCCCATGGTGCGCACACTGCAAGGAGTTGGCTCCGATCTGGGATAAACTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATTATCGCCAAG
ATGGACGCCACGGCCAATGAGCTGGAAGAGTTCCAAGTCCCAAGTTTCCCAACCTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGCGA
GAGGACCCTAGAAGAAATGACCAAATTTGTGGAAAGTGGTGGCAAACAAGAACCTCCGAAGAAAAAGGAGGAAGAAAAAGAGGAAGATGATGACAAGAAAAA
GGATGAACTCTAA
>csPDI_GI/AH_[Conus planorbis]
ATGAAGTTTGCAACTATTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGAAGTTGAGCAAGAGGATAATGTCTATGTTTTAAACGAAAACAACTTCGACG
ACTTCATCAGGGGAAATGAATTTGTTCTTCTGGAGTTTTATGCACCGTGGTGCGGCATCTGCAAGCGCCTGGCTCCAGTGTACAGTGAAGCTGCAGGGAAATTGA
TGCAAGAGGGTTCCAAAGTCAAGCTGGCCAAGATTGATGCCACTGTCGAGAAGGGCTTGGCGAGCAGGTATGAAGTCAAAGGCTACCCCACCATCAAATTCTTC
AACAAGGGCATATCATCAGATTTCACGGGAGGCAGGCAGACTTCCGAGATCATCAACTGGCTGAAGAAGAGGACCGGCCCCCCTGCCCAGGTCCTGAAGTCTTC
TCAGGAGACCAAGACTTTCATTGACAGTGATGAAGTCGTCGTCTTGGGCTTCTTCAAGGACCAGGAAGGCGAGGACGCTGCTGCCTTCAAGGAGGTTGCCACAAA
AATCGAGGACATTGCCTTTGGCATCACGTCAGAGGATAGCCTCTTTAAAGAGCATAAGATGAAGAAAGATGGCGTTGTTTTATTTAAAAAGTTTGACGAGGGTCG
CAGTGACTTCTCTGGAGACTTTAAGGAGGATGCCATTAGCAAGTTCATCAAGGAAAGTCGCCTCCCTCTGATCAACGAATTTACAATGGAGACCGCCCAGAAAAT
CTTCTCAGGGGAAGTGAATAACCACCTCATGCTCTTTGTCAAGAAGGACGATGAAAAGGACACACTGGACACCTTTAAGGCTTGTGCTGGTGAATTCAAGGGAAA
GGTCCTGTTTGTCTACCTGGACACAGCCCGGGAGGATAATGAACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGATGCTCCGGCGATGCGCCTGATTCAGAT
GGGGGATGACCTGACGAAGTACAAACCCGAGTCGGACTCTCTGGACAAGTCCACCATAACCAAGTTTGTTCAGGACTACCTGGATGGCAAGCTGAAGCCTCACC
TGAAGTCAGAGGAAGTGCCAGAGGACTGGGATGCCAAGCCCGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCGGTCTTT
GTGGAGTTTTATGCCCCATGGTGCGCACACTGCAAGGAGTTGGCTCCGATCTGGGATAAACTGGGTGAGAAGTTCAAGGACAGCAAGGATATCATTATCGCCAAG
ATGGACGCCACGGCCAATGAGCTGGAAGAGTTCCAAGTCCCAAGTTTCCCAACCTTGAAGTACTTCCCCAAGGGCAGCGACGAGATCATCGAATATGATGGCGA
GAGGACCCTAGAAGAAATGACCAAATTTGTGGAAAGTGGTGGCAAACAAGAACCTCCGAAGAAAAAGGAGGAAGAAAAAGAGGAAGATGATGACAAGAAAAA
GGATGAACTCTAA
>csPDI_GA/GH_[Conus imperialis]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGATGTTGAACAAGAGGAAAATGTCTATGTTTTAAATGCAAAAAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCGTGGTGCGGCGCATGCAAGAGTTTTGCTCCCACATACAGCGAGGCTGCTGGGAAATTGAT
GGATGAGGATTCCGACATCAAGCTGGCCAAGGTTGATGCCATCGCAGAGAAGGAGTTGGCGAACAAATTTGAAGTCAGAGGCTACCCCACCATCAAATTCTTCC
GCAGTGGTGAAGCAACAGACTACACTGGAGACAGGCACAGCTCTTCTGACGTCATCAGCTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACCTGAAGACT
TCTGAGGAGGCCAAGACTTTCATTGACAGCGGTGAAGTCATCGTCATAGGCTTCTTCAAGGACCAGGGATGCAAGGGCGCTAAAGTCTTAAAGAATATTGCCCCG
GATATCGAGGACGTTGCCTTTGGTATCACGTCAGAGGACAGCGTCTTTAAAGAGTATAAGATGAAGGAAGACGGCGTTGTTGTTTTTAAAAAGTTTGATGAGGGT
CGCAGTGACTTCTCTGGAGACTTTGAGGAGGTTGCCATGAGCAAGTTCATCAAGAATAATCGGCTCCCTCTGATCATCGAGTTTACACAGGAGACCGCCCAGAAG
ATCACCGCAGGGGATGTGCAGAACTACCTCATGCTGTTCGTCGAGAAGGAGGGAGCAAAGGACACATTGGATAACTTCAAGGCTGCTGCCGGTGAATTCAAGGG
GAAGGTCCTGTTTATTTACGTGGACACAGCCAAGGAGGAAAGTGGACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGACGCTCCGGCAATGCGCCTGATTAA
GCTGGGCGATGACCTGGCCAAGTACAAGCCCGACTCGGACAATCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATGGCAAGCTGAAGCCTTA
CCTGAAGTCGGAGGAAGTGCCAGAGGGCTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTCT
TTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATGAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATTATCATCGCCA
AGATGGACTCCACAGCCAATGAGCTGGAAGAAGTCAGCATCCAGAGTTACCCAACCTTGAAGTATTACCCCAAGGACAGCGACGAGATCATCGAATATAATGGC
GAGAAGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACCTCCAGAGGAAAAGGAGGAGGAAGAAGAGGAAGATGACAAGAAAAAG
GATGAACTCTAA
>csPDI_GH/GH_305_[Conus imperialis]
ATGAAGTTTGCAACTGTTTTCTCTCTCACGTTGCTGGCCTTTGTGGCCTGTGAAGATGTTGAACAAGAGGAAAATGTCTATGTTTTAAAGGAAAAAAACTTCGACA
ACTTCATTAAGGAAAATGAATTTGTTCTTGTGGAGTTTTACGCACCGTGGTGCGGCCACTGCAAGCAATTGGCTCCCGCGTACAGCGAGGCTGCTGGGAAATTGA
TGGATGAGGATTCCGACATCAAGCTGGCCAAGGTTGATGCCATCGCAGAGAAGGAGTTGGCGAACAAATTTGAAGTCAGAGGCTACCCCACCATCAAATTCTTCC
GCAGTGGTGAAGCAACAGACTACACTGGAGACAGGCACAGCTCTTCTGACGTCATCAGCTGGCTGAAGAAGAAGACCGGCCCCCCTGCCAAGGACCTGAAGACT
TCTGAGGAGGCCAAGACTTTCATTGACAGCGGTGAAGTCATCGTCATGGGCTTCTTCAAGGACCAAGAAGGCAAGGGCGCTAAAGCCTTCAAGAATATTTCTGCA
GAATTCATAGACATTGCCTTTGGCATCACGTCAGAGGACAGCATCTTTAAAAAGTATAAGATGAAGAAAGACGGCGTTGTTCTTTTTAAAAAGTTTGACGAGGGT
CGCAGTGACTTCTCTGGAGACTTTGAGGAGGTTGCCATGAGCAAGTTCATCAAGAATAATCGGCTCCCTCTGATCATCGAGTTTACACAGGAGACCGCCCAGAAG
ATCACCGCAGGGGATGTGCAGAACTACCTCATGCTGTTCGTCGAGAAGGAGGGAGCAAAGGACACATTGGATAACTTCAAGGCTGCTGCCGGTGAATTCAAGGG
GAAGGTCCTGTTTATTTACGTGGACACAGCCAAGGAGGAAAGTGGACACATCATGGGATTCTTCGGCCTCAAGGCTGCCGACGCTCCGGCAATGCGCCTGATTAA
GCTGGGCGATGACCTGGCCAAGTACAAGCCCGACTCGGACAATCTGGACAAGTCCACCATGACCAAGTTTGTTCAGGACTTCCTGGATCGCAAGCTGAAGCGTCA
CCTTAAGTCGGAGGAAGTGCCAAAGGACTGGGATGCCCAGCCTGTGAAGGTTTTGGTCAGCAAGAACTTCAAGGAGGTGGCCATGGACAAATCCAAGGCCGTCT
TTGTGGAGTTTTATGCCCCATGGTGCGGACACTGCAAGAAGTTGGCTCCGATCTGGGATGAGCTGGGTGAGAAGTTCAAGGACAGCAAGGATATTATCATCGCCA
AGATGGACTCCACAGCCAATGAGCTGGAAGAAGTCAGCATCCAGAGTTACCCAACCTTGAAGTATTACCCCAAGGACAGCGACGAGATCATCGAATATAATGGC
GAGAAGACCCTGGAAGAACTGACCAAATTCGTGGAAAGTGGTGGCAAACAAGAACCTCCAGAGGAAAAGGAGGAGGAAGAAGAGGAAGATGACAAGAAAAAG
GATGAACTCTAA
```

CHAPTER 4

USING A LOGISTIC REGRESSION MODEL TO DISCOVER NOVEL CONOTOXINS

Abstract

Mining next generation RNA sequencing data for discovery purposes is especially difficult for two classes of transcripts: those that evolve very rapidly and hence are poorly conserved; and those that belong to large, highly conserved gene families for which identification of new members is complicated by their similarity to known members. Conotoxins present both challenges. In Chapter 3, I described how I have used my pipeline to identify new members of highly paralogous gene families (e.g, protein disulfide isomerases, or csPDIs) Here, I describe how I have created a novel search tool for identifying highly divergent conotoxins using a logistic regression model based upon three prominent characters of conotoxins. Using this tool, I have discovered 47 new conotoxin superfamilies.

Introduction

There are about 730 species in the genus of Conus, many with unusual life histories, and their venoms are largely uncharacterized (PUILLANDRE *et al.* 2014; OLIVERA *et al.* 2015). It is estimated that each Conus species produces ~100-200 distinct venom peptides with almost no overlap in repertoire between different species (OLIVERA 2002). Despite the tremendous drug discovery potential in Conus venoms, only ~ 1400 nucleotide sequences of conotoxin genes have been reported from 100 Conus species by traditional approaches over the past decades (KAAS *et al.* 2008; KAAS *et al.* 2010). Traditional methods like peptide isolation and sequencing are time consuming, of low sensitivity and limited by sample availability. In contrast, high throughput sequencing can achieve greater sequencing depth and larger coverage of the transcriptome while it only requires a moderate amount of the sample (PRASHANTH *et al.* 2012). Recent studies on the venom duct transcriptome of several Conus species, using next generation sequencing technologies (NGS), have discovered ~100 conotoxin genes per Conus species (HU *et al.* 2011; HU *et al.* 2012; LLUISMA *et al.* 2012; TERRAT *et al.* 2012; DUTERTRE *et al.* 2014; BARGHI *et al.* 2015; HIMAYA *et al.* 2015; LAVERGNE *et al.* 2015).

Conotoxins can be classified into different gene superfamilies based on their signal peptide sequence (BUCZEK *et al.* 2005). As of 2014, 30-40 conotoxin gene superfamilies had been identified (ROBINSON AND NORTON 2014). After NGS sequencing and de novo transcriptome assembly, candidate conotoxin genes are usually predicted into different superfamilies by BLASTX search and HMMER analysis against a local reference database of known conotoxins from the ConoServer databases, including ConoPrec, ConoDictor and Conosorter (KAAS *et al.* 2008; KOUA *et al.* 2012; LAVERGNE *et al.* 2013; WHEELER AND EDDY 2013). However, this approach to conotoxin discovery has an obvious pitfall: it only discovers the conotoxin genes with statistically significant similarity to sequences in the current reference database. The short lengths and rapid evolution (OLIVERA 2006) of conotoxins and paucity of known references maske them difficult targets for alignment-based and hidden Markov Models (HMM)-based approaches (Figure 4.1). Given these facts, it is very likely that many novel conotoxins, even entire superfamilies, may still await discovery.

Conotoxin transcripts have three prominent characters: (1) an N-terminal signal sequence for targeting to the endoplasmic reticulum; (2) a cysteine-rich mature toxin region at the C terminus; (3) and they are short (180-360 nt) compared to the >500 nt average typical for remainder of the *Conus* transcriptome) (BANDYOPADHYAY *et al.* 1998;

CONTICELLO *et al.* 2003; BUCZEK *et al.* 2004).

In response to these facts, I implemented the logistic regression model (COX 1958). and used it in my discovery pipeline. The model uses the three prominent characteristics of conotoxins described above (short length, high-cysteine content and presence of a secretion signal) to predict the probability that transcript in a Conus transcriptome with no discernable homology by BLAST or HMM to any member of the Conus reference database may in fact be a conotoxin. In cross validation experiments using known conotoxin and nonconotoxin sequences, my method had high specificity (97.5%) and sensitivity (97.8%). When applied to real RNAseq data from 22 Conus species, the pipeline discovered 47 new potential conotoxin superfamilies. My method is implemented in Perl as a complete conotoxin discovery package under the name ConusPipe at https://github.com/Yandell-Lab/ConusPipe.

Materials and methods

Our conotoxin discovery package proceeds by first using known conotoxin sequences (from ConoServer databases plus all the manually curated conotoxins identified in this and previous studies) together with all nonconotoxin Conus transcripts as training datasets to build the logistic regression model. Formula 4.1 shows a generic logistic regression model. p_i stands for probability of succeed, and the log odds of succeed is the responsible variable. X stands for different explanatory variables. α and β stand for model parameters. Next, Conus transcripts from 22 Conus species, which are not present in the current reference database (a combined ConoServer and UniProtKB database plus all the manually curated conotoxins identified in this and previous studies) are used as input data to run the built logistic regression model to predict the probability of a certain input transcript to be conotoxin. Cross validation with training data is then used to evaluate the model and make a Receiver Operating Characteristic (ROC) curve to find out the cutoff probability with highest sensitivity and specificity (FAWCETT 2006). Then the input transcripts with predicted probability greater than the cut off probability are output by the pipeline and manually checked and confirmed by conotoxin experts to be putative new conotoxins.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} \tag{4.1}$$

Logistic regression is a special type of regression where binary response variable is related to a number of explanatory variables using a logit function. The explanatory variables can be either discrete or continuous (COX 1958). I used the training set of conotoxin sequences to define the distributions of three explanatory variables: signal sequence, cysteine percentage and overall sequence length (4.2). The discovery pipeline first runs signalP on a training dataset (5243 known conotoxin/nonconotoxin sequences) to get the signalP D score for each known sequence (PETERSEN *et al.* 2011). This is explanatory variable 1. Next, the pipeline parses each known sequence to get the cysteine percentage and sequence length, explanatory variables 2 and 3, respectively. Using these data, I trained a logistic regression model using the R logit regression package (TEAM 2013). The coefficients of the explanatory variables (the betas in formula 4.1) are estimated via minimizing the cost function for the logistic regression model via stochastic gradient descent of the values of the three coefficients (Betas) (WILKS 1938). The significance of each regression coefficient is assessed by computing the Wald statistic and the significance of the overall model is assessed by the difference between the residual deviance for the model with predictors and the null model (MENARD 2002; ALLISON 2014). All the coefficients of explanatory variables and overall model fit are significant, thus all three were included in the final model.

Cross validation

5243 known conotoxin sequences (from ConoServer databases plus all the manually curated conotoxins identified in this and previous studies) and 5243 nonconotoxin Conus transcripts were first split into 10 equal subsets, and then sequentially resampled so that one decile was used as the test set, and the other were combined as the training set. The logistic regression model was trained with training set and evaluated with a test set in 10 iterations. The average results of the 10 iterations was reported.

For each iteration, all the coefficients of explanatory variables and overall model fit are significant. The specificity and sensitivity of conotoxin prediction under different probability cut offs were calculated, and a ROC curve was plotted using the true positive rate against the false positive rate for the different predicted probability cut points. In this way, the probability cut point with highest specificity and sensitivity was discovered.

Applying to real data to discover new conotoxin superfamilies

Paired end RNAseq data from 22 Conus species were generated by Illumina HiSeq 2000 platform (Table 4.1). RNAseq reads were assembled using best practice Trinity settings from the simulation pipeline (Chapter 2), annotated with BLASTX against our improved reference dataset, and all the Conus transcripts which do not have homologs in the current reference database (as judged by BLAST E < 1e^{-4}) were recovered and translated in all 6 frames. These peptide sequences were then used as the input dataset for the logistic regression model to predict their probability of being conotoxin. The pipeline output all input transcripts, which have a predicted probability greater than the optimal probability cut off defined by ROC curve (134 putative conotoxins). Then the output sequences were manually checked and confirmed as conotoxins by two internationally recognized experts for conotoxin classification, Sam Robinson, [Biology department, University of Utah], and Helena Safavi [Biology department, University of Utah], with the judging criteria of charge distribution, potential cleavage site, etc. Then all-by-all BLASTP was conducted among the new conotoxins, and the new conotoxins that have high homology (Blastp e-value <1e^{-10}) were designated to be in the same superfamily.

Results

Build and cross validate the logistic regression model

5243 known conotoxin sequences (from ConoServer databases plus all the manually curated conotoxins identified in this and previous studies) together with 5243 nonconotoxin Conus transcripts were used to build the logistic regression model shown in (4.3). The probability distribution of conotoxin/nonconotoxin was plotted (Figure 4.2 A, B). In order to assess the performance of the model, the same dataset were partitioned into training and testing sets to validate the model. The main measures of performance were sensitivity and specificity under different probability cut offs. The sensitivity was defined as the fraction of known conotoxins predicted as conotoxin over the number of known conotoxins in the test dataset. The specificity was defined as the fraction of known nonconotoxins predicted as nonconotoxin over the number of known nonconotoxins in test dataset.

$$\text{Conotoxin/Nonconotoxin} \sim 6.91\text{signalP D Value} + 26.31\text{cystein\%} - 11.9\text{transcripts length} \qquad (4.3)$$

In order to determine the probability cut point with highest sensitivity and specificity, the ROC curve was plotted using the true positive rate against the false positive rate for the different predicted probability cut points (Figure 4.3). The highest specificities and sensitivities were 97.5% and 97.8%, respectively, at a probability cut point 0.6.

The discovery of new conotoxin superfamilies

Potential new conotoxins were defined as those with a predicted probability greater than the probability cut off (p= 0.6) defined by the ROC curve and manually checked and confirmed by conotoxin experts. 104 out of 134 putative conotoxins passed the judge criteria. The 104 potential new conotoxin sequences belong to 47 new conotoxin superfamilies.

The family and superfamily relationships of these newly discovered conotoxins (putativeXXX) are shown in Figure 4.4, which also contains a representative number of known members of every known superfamily (Robinson and Norton 2014). As can be seen, the new conotoxins comprise cohesive superfamilies distinct from any known conotoxin superfamilies in the tree.

Discussion

I have devised a novel approach for discovery of new conotoxins. Compared to previous studies, which only attempted to find novel variations of known conotoxins by BLASTX search and HMMER analysis, my approach is the first one capable of recovering entirely new conotoxins. To my knowledge, this is the first time a logistic regression model has been used as a search tool in a bioinformatics discovery pipeline. As of 2014, only 30-40 conotoxin gene superfamilies had ever been identified (ROBINSON AND NORTON 2014), despite over 2 decades of intense effort from investigators around the globe. My pipeline has discovered 47 new potential conotoxin superfamilies. Though further protein mass spectrometry and injection/functional analysis are needed to confirm the novel superfamily discovery, when the newly discovered conotoxin sequences were used to build BLASTX database and searched against previously published Conus transcriptome data, which are in different Conus species, we found additional homologs in previously published data (HU *et al.* 2011; HU *et al.* 2012). Except for some highly conserved house keeping genes/enzymes, conotoxins from the same superfamily are the most conserved sequences across different Conus species, which indicates that our newly discovered conotoxin superfamilies are very likely to be genuine peptides (BANDYOPADHYAY *et al.* 1998; CONTICELLO *et al.* 2003; BUCZEK *et al.* 2004).

My approach took 2 hours 40 minutes on a single CPU core to run the discovery pipeline for 1,359,647 Conus transcripts from 22 Conus species, My results contribute significantly to pharmacological discovery efforts and will also help answer basic ecological and evolutionary questions regarding Conus and the evolution of its venom repertory.

References

Allison, P. D., 2014 Measures of fit for logistic regression. in Proceedings of the SAS Global Forum 2014 Conference.

Bandyopadhyay, P. K., C. J. Colledge, C. S. Walker, L.-M. Zhou, D. R. Hillyard *et al.*, 1998 Conantokin-G precursor and its role in γ-Carboxylation by a vitamin K-dependent carboxylase from a Conus snail. J. Biol. Chem. 273**:** 5447-5450.

Barghi, N., G. P. Concepcion, B. M. Olivera and A. O. Lluisma, 2015 Comparison of the venom peptides and their expression in closely related Conus species: insights into adaptive post-speciation evolution of Conus exogenomes. Genome Biol. Evol. 7**:** 1797-1814.

Buczek, O., G. Bulaj and B. M. Olivera, 2005 Conotoxins and the posttranslational modification of secreted gene products. Cellular and Molecular Life Sciences 62**:** 3067-3079.

Buczek, O., B. M. Olivera and G. Bulaj, 2004 Propeptide does not act as an intramolecular chaperone but facilitates protein disulfide isomerase-assisted folding of a conotoxin precursor. Biochemistry 43**:** 1093-1101.

Conticello, S. G., N. D. Kowalsman, C. Jacobsen, G. Yudkovsky, K. Sato *et al.*, 2003 The prodomain of a secreted hydrophobic mini-protein facilitates its export from the endoplasmic reticulum by hitchhiking on sorting receptors. J. Biol. Chem. 278**:** 26311-26314.

Cox, D. R., 1958 The regression-analysis of binary sequences. Journal of the Royal Statistical Society Series B-Statistical Methodology 20**:** 215-242.

Dutertre, S., A. H. Jin, I. Vetter, B. Hamilton, K. Sunagar *et al.*, 2014 Evolution of separate predation- and defence-evoked venoms in carnivorous cone snails. Nature Communications 5.

Fawcett, T., 2006 An introduction to ROC analysis. Pattern Recognition Letters 27**:** 861-874.

Himaya, S. W. A., A. H. Jin, S. Dutertre, J. Giacomotto, H. Mohialdeen *et al.*, 2015 Comparative venomics reveals the complex prey capture strategy of the piscivorous cone snail Conus catus. Journal of Proteome Research 14**:** 4372-4381.

Hu, H., P. K. Bandyopadhyay, B. M. Olivera and M. Yandell, 2011 Characterization of the Conus bullatus genome and its venom-duct transcriptome. Bmc Genomics 12.

Hu, H., P. K. Bandyopadhyay, B. M. Olivera and M. Yandell, 2012 Elucidation of the molecular envenomation strategy of the cone snail Conus geographus through transcriptome sequencing of its venom duct. Bmc Genomics 13.

Kaas, Q., J. C. Westermann and D. J. Craik, 2010 Conopeptide characterization and classifications: an analysis using ConoServer. Toxicon 55**:** 1491-1509.

Kaas, Q., J. C. Westermann, R. Halai, C. K. L. Wang and D. J. Craik, 2008 ConoServer, a database for conopeptide sequences and structures. Bioinformatics 24**:** 445-446.

Koua, D., A. Brauer, S. Laht, L. Kaplinski, P. Favreau *et al.*, 2012 ConoDictor: a tool for prediction of conopeptide superfamilies. Nucleic Acids Res 40**:** W238-241.

Lavergne, V., S. Dutertre, A. H. Jin, R. J. Lewis, R. J. Taft *et al.*, 2013 Systematic interrogation of the Conus marmoreus venom duct transcriptome with ConoSorter reveals 158 novel conotoxins and 13 new gene superfamilies. BMC Genomics 14**:** 708.

Lavergne, V., I. Harliwong, A. Jones, D. Miller, R. J. Taft *et al.*, 2015 Optimized deep-targeted proteotranscriptomic profiling reveals unexplored Conus toxin diversity and novel cysteine frameworks. Proceedings of the National Academy of Sciences of the United States of America 112**:** E3782-E3791.

Lluisma, A. O., B. A. Milash, B. Moore, B. M. Olivera and P. K. Bandyopadhyay, 2012 Novel venom peptides from the cone snail Conus pulicarius discovered through next-generation sequencing of its venom duct transcriptome. Marine Genomics 5**:** 43-51.

Menard, S., 2002 *Applied logistic regression analysis*. Sage.

Olivera, B. M., 2002 Conus venom peptides: reflections from the biology of clades and species. Annual Review of Ecology and Systematics 33**:** 25-47.

Olivera, B. M., 2006 Conus peptides: biodiversity-based discovery and exogenomics. J. Biol. Chem. 281**:** 31173-31177.

Olivera, B. M., J. Seger, M. P. Horvath and A. E. Fedosov, 2015 Prey-capture strategies of fish-hunting cone snails: behavior, neurobiology and evolution. Brain Behavior and Evolution 86**:** 58-74.

Petersen, T. N., S. Brunak, G. von Heijne and H. Nielsen, 2011 SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods 8**:** 785-786.

Prashanth, J. R., R. J. Lewis and S. Dutertre, 2012 Towards an integrated venomics approach for accelerated conopeptide discovery. Toxicon 60**:** 470-477.

Puillandre, N., P. Bouchet, T. F. Duda, S. Kauferstein, A. J. Kohn *et al.*, 2014 Molecular phylogeny and evolution of the cone snails (Gastropoda, Conoidea). Molecular Phylogenetics and Evolution 78**:** 290-303.

Robinson, S. D., and R. S. Norton, 2014 Conotoxin gene superfamilies. Marine Drugs 12**:** 6058-6101.

Team, R. C., 2013 R: a language and environment for statistical computing.

Terrat, Y., D. Biass, S. Dutertre, P. Favreau, M. Remm *et al.*, 2012 High-resolution picture of a venom gland transcriptome: case study with the marine snail Conus consors. Toxicon 59**:** 34-46.

Wheeler, T. J., and S. R. Eddy, 2013 nhmmer: DNA homology search with profile HMMs. Bioinformatics 29**:** 2487-2489.

Wilks, S. S., 1938 The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics 9**:** 60-62.

Table 4.1 RNAseq data sets used in the discovery pipeline

Conus Species	Illumina HiSeq 2000	
	Number of reads	Read length (nt)
C.andremenezi	63,598,020	101
C.bullatus	107,790,134	101
C. geographus	158,004,874	101
C. pracellens	76,907,910	101
C. distans	85,877,500	101
C. eburneus	101,170,402	101
C. sulcatus	53,901,510	125
C. marmoreus	50,652,396	101
C. tessulatus	52,530,032	101
C. rolani	67,969,964	125
C. obscurus	101,151,254	125
C. textile	63,365,620	125
C. ateralbus	28,783,428	125
C. omaria	23,843,856	125
C. proximus	30,784,548	125
C. ammiralis	30,038,902	125
C. magus	31,056,732	125
C. mucronatus	31,180,460	125
C. ochroleucus	27,927,952	125
C. crocatus	19,556,244	125
C. striolatus	24,569,932	125
C. imperialis	68,293,558	101

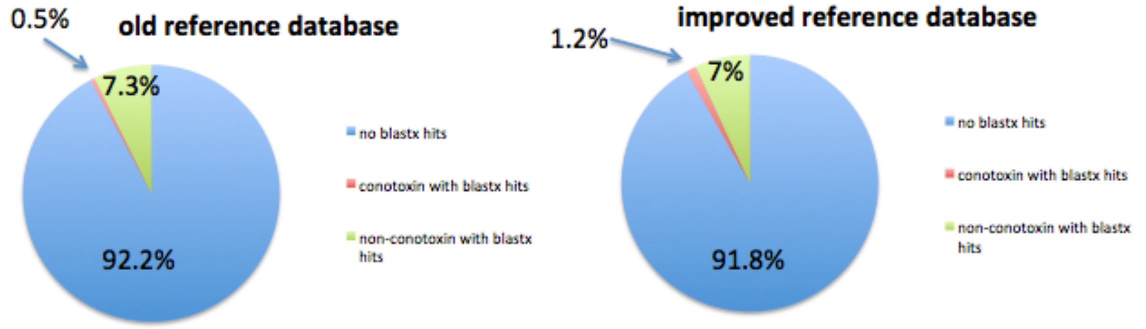


Figure 4.1 Reference database old vs. improved. Number of candidate conotoxin genes predicted in Conus *textile* transcriptome by BLASTX search increased when searching against an improved reference database. Old reference database: a combined ConoServer and UniProtKB database; Improved reference database: a combined ConoServer and UniProtKB database plus all the manually curated conotoxins identified in this and previous studies.

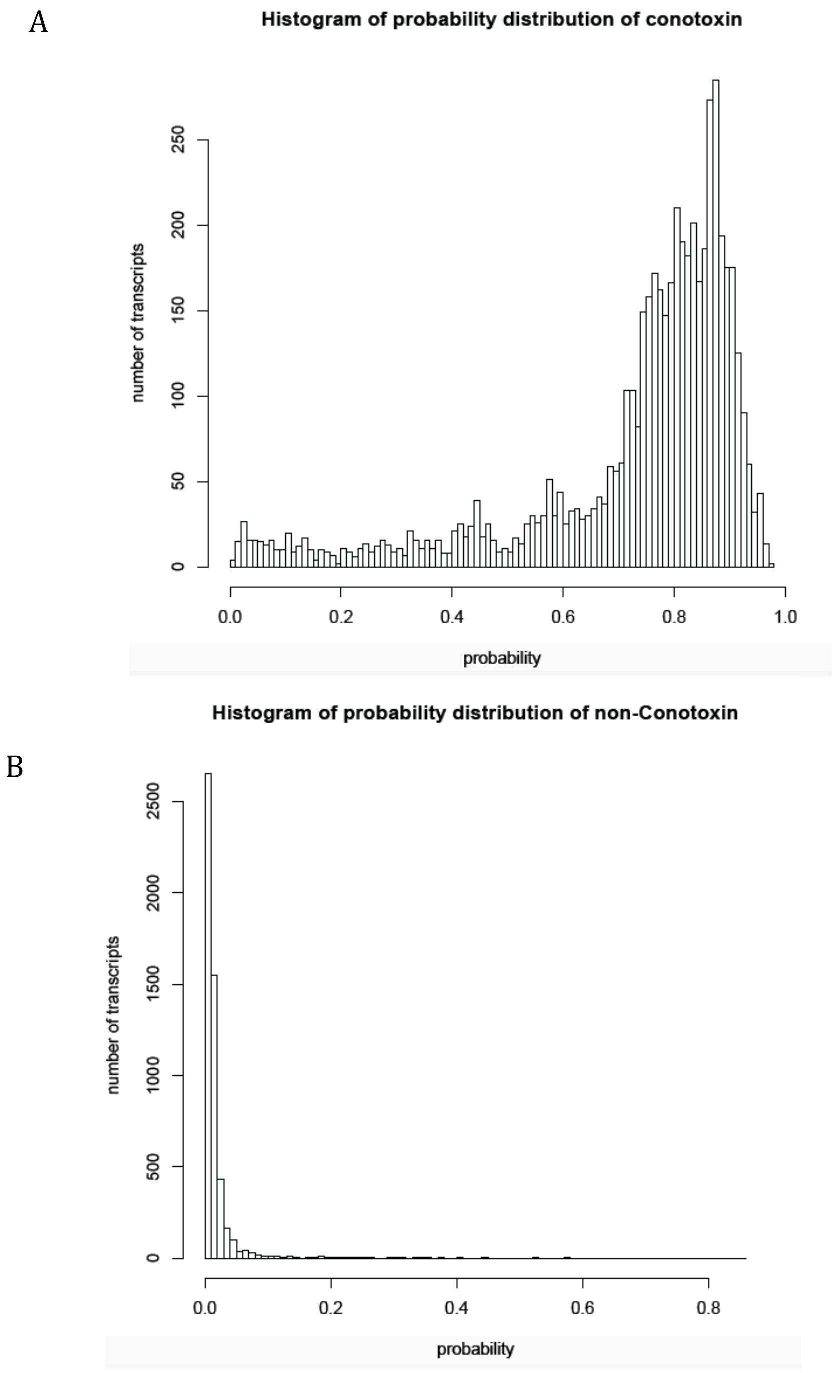


Figure 4.2. Probability distribution of conotoxin/nonconotoxin. (A) The probability distribution of known conotoxins predicted by the logistic regression model. (B) The probability distribution of known nonconotoxins predicted by the logistic regression model.

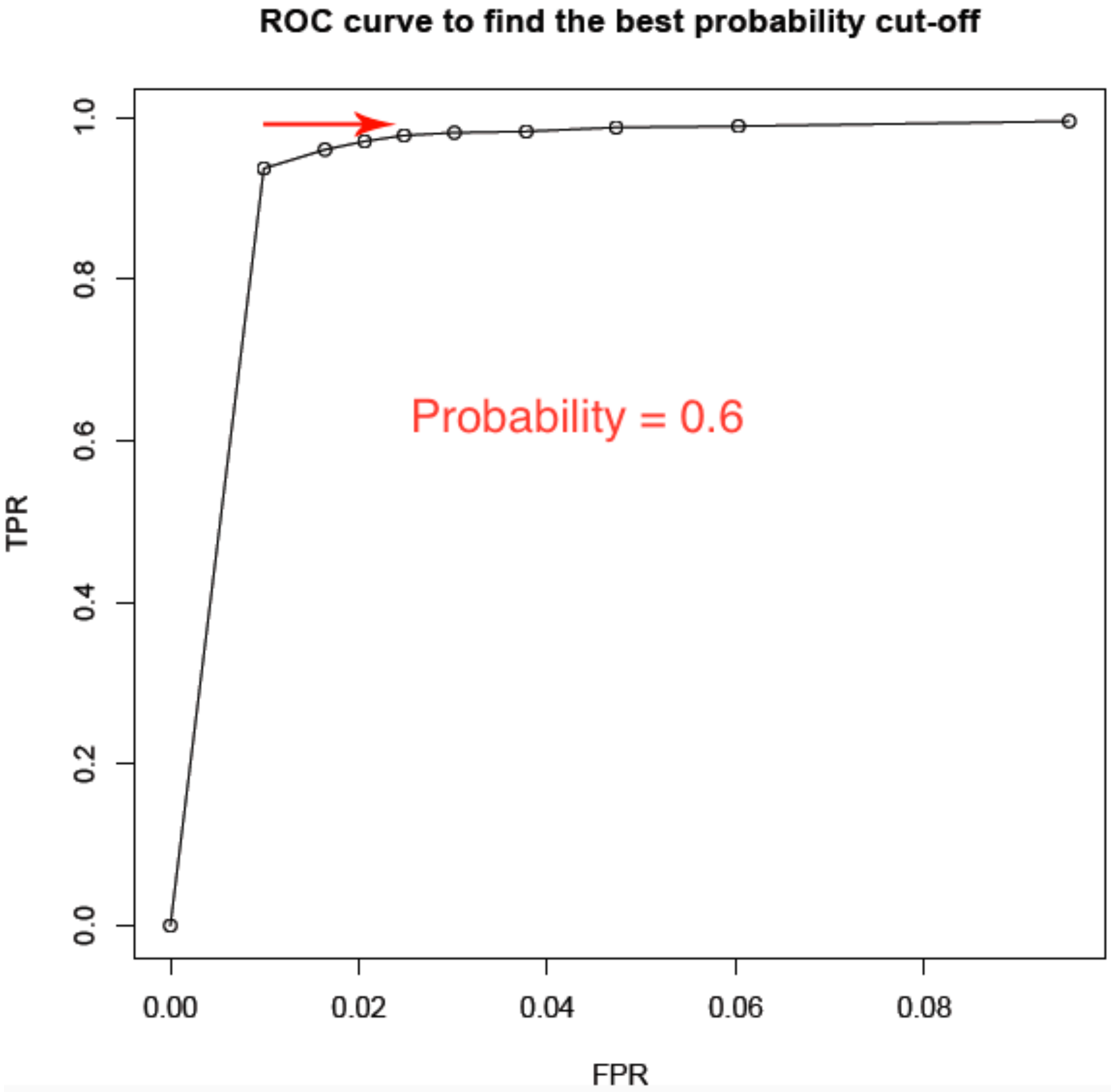


Figure 4.3. ROC curve based on predicted probability of known conotoxins and nonconotoxins. ROC curve finds the cutoff probability at the lowest false positive rate and the highest true positive rate.

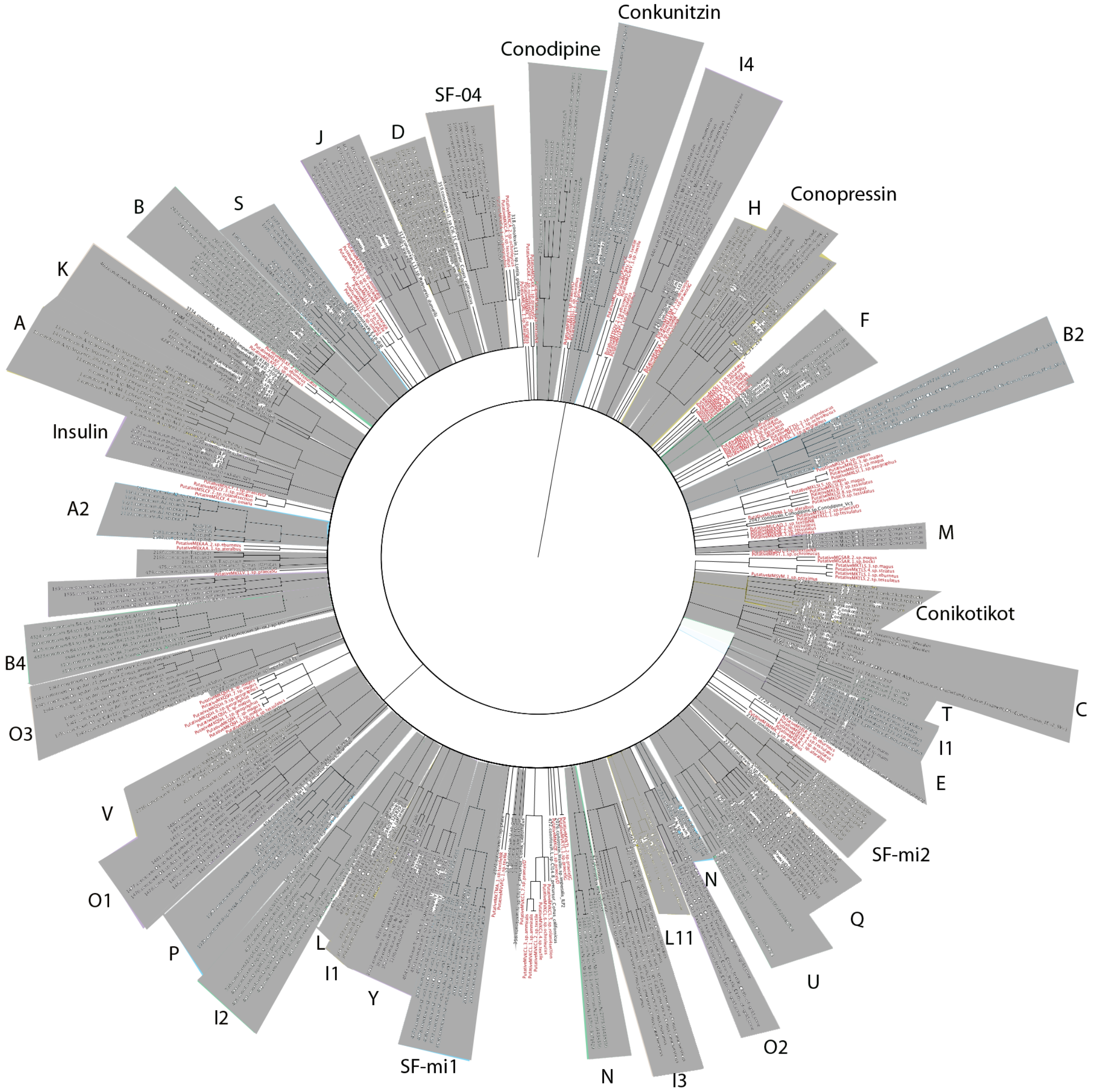


Figure 4.4 Conotoxin superfamily relationships. The superfamily relationships between newly discovered conotoxin superfamilies (putativeXXX, red text) and previously known conotoxin superfamilies (black text) shaded in grey shown in Figtree v1.4.2 software package (http://tree.bio.ed.ac.uk/software/figtree/). The tree file was generated from the multi fasta file of newly discovered and 36 known conotoxins (randomly picked 10 conotoxins/superfamily) using a kmer analysis tool – Gasoline, developed in our lab (Yandell et al., manuscript in preparation). Almost all known conotoxins from the same superfamily were clustered in one or two locations, and all the new conotoxins from the same superfamily were clustered in the same location. All the new conotoxin superfamilies stand alone by themselves without intersecting with any known conotoxin superfamilies.

CHAPTER 5

DISCUSSION

Summary

The work in the dissertation has advanced our discovery effort of the conotoxins. I developed a comprehensive pipeline of improved methods for NGS-based conotoxin discovery. Regarding improving the quality of conotoxin discovery, I developed a simulation pipeline to reduce chimeric conotoxins and a partial extension pipeline to extend truncated conotoxins; regarding improving quantity of conotoxin discovery, I developed a discovery pipeline to mine potential novel conotoxins from Conus transcripts that has no hits in the current reference database. I also automated the whole process in three steps to improve the efficiency of the pipeline by reducing I/O manipulation (Figure 5.1).

I also show that the partial extension pipeline can be repurposed to discover a conotoxin modification enzyme – conotoxin-specific protein disulfide isomerases, which were invisible using current standard transcriptome assembly methods in Chapter 3. Thus, I maximize the discovery potential of RNAseq-based conotoxin discovery. Actually, the application of the prototype of the pipeline already led to a series of novel discoveries and publications and many more are in preparation (ROBINSON *et al.* 2015; SAFAVI-HEMAMI *et al.* 2015; SAFAVI-HEMAMI *et al.* 2016a; SAFAVI-HEMAMI *et al.*

2016b).

Future direction

Since more and more conotoxins are discovered by my pipeline – we have processed RNAseq data from 22 Conus species, and much more sequence data are coming, there is a need to manage the conotoxin database we have for easy query and update. I am developing a new database management pipeline, which will build a SQL database for all the conotoxins we have, add future newly discovered conotoxins into the database if it is not redundant to conotoxins in the current database and allow the users to query the database easily either by superfamily or Conus species.

Moreover, we have also begun to use the output from my pipeline to explore how conotoxin repertory correlates with life history traits (Figure 5.2). The result shows that worm hunters (Amz and Pra) and fish hunters (Geo) form two well-separated clades by toxin composition. These preliminary results confirm the hypothesis that Conus species with different life history traits generally have different conotoxin expression profiles. Now we are extending the analysis to all the Conus species we have sequenced. Through doing this work, we will have a pretty clear picture that different Conus species utilize different venom peptides (with some overlap in their conotoxin repertoire). By sequencing many, we are able to observe this pattern now and uniquely look at the conotoxins that could be most interesting, such as those that are highly expressed in all species or those that are only expressed in fish hunters, etc. Also, once we find a conotoxin with an interesting activity, we can immediately look at what members of this family of peptides look like in other Conus species.

In summary, the work in this dissertation has the potential to have a strong impact as regards to how transcriptome data are mined, answer basic ecological and evolutionary questions regarding Conus life history and has great pharmacological discovery potential.

References

Robinson, S. D., Q. Li, P. K. Bandyopadhyay, J. Gajewiak, M. Yandell *et al.*, 2015 Hormone-like peptides in the venoms of marine cone snails. General and Comparative Endocrinology 244:11-18.

Safavi-Hemami, H., J. Gajewiak, S. Karanth, S. D. Robinson, B. Ueberheide *et al.*, 2015 Specialized insulin is used for chemical warfare by fish-hunting cone snails. Proceedings of the National Academy of Sciences of the United States of America 112**:** 1743-1748.

Safavi-Hemami, H., Q. Li, R. L. Jackson, A. S. Song, W. Boomsma *et al.*, 2016a Rapid expansion of the protein disulfide isomerase gene family facilitates the folding of venom peptides. Proceedings of the National Academy of Sciences of the United States of America 113**:** 3227-3232.

Safavi-Hemami, H., A. P. Lu, Q. Li, A. E. Fedosov, J. Biggs *et al.*, 2016b Venom insulins of Cone snails diversify rapidly and track prey taxa. Molecular Biology and Evolution 33**:** 2924-2934.

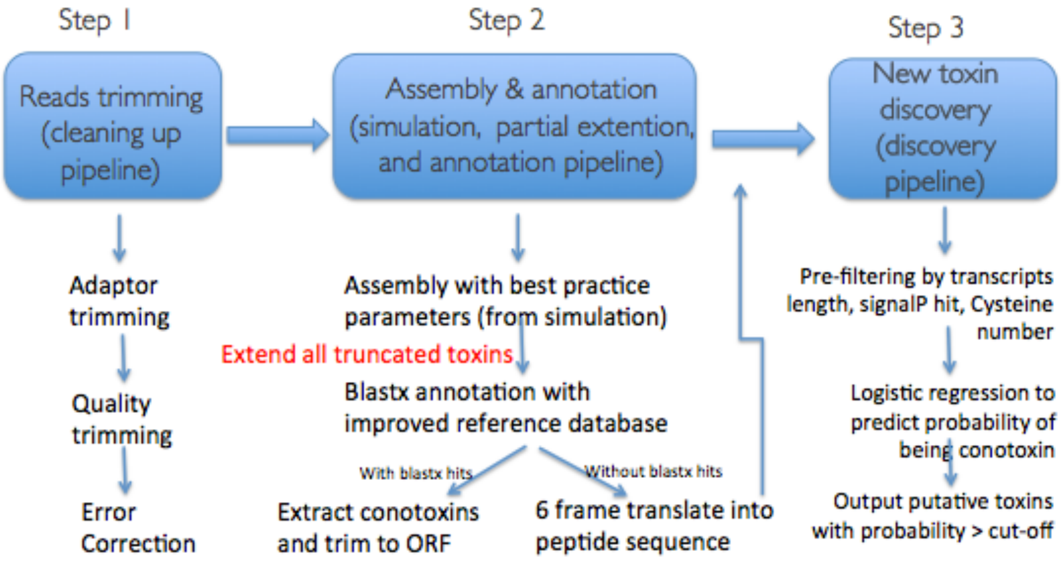

Step 1
Reads trimming (cleaning up pipeline)
Step 2
Assembly & annotation (simulation, partial extention, and annotation pipeline)
Step 3
New toxin discovery (discovery pipeline)
Adaptor trimming
Quality trimming
Error Correction
Assembly with best practice parameters (from simulation)
Extend all truncated toxins
Blastx annotation with improved reference database
With blastx hits
Without blastx hits
Extract conotoxins and trim to ORF
6 frame translate into peptide sequence
Pre-filtering by transcripts length, signalP hit, Cysteine number
Logistic regression to predict probability of being conotoxin
Output putative toxins with probability > cut-off

Figure 5.1.Conotoxin discovery in 3 steps. Improved efficiency of conotoxin discovery by automating the whole process in three steps

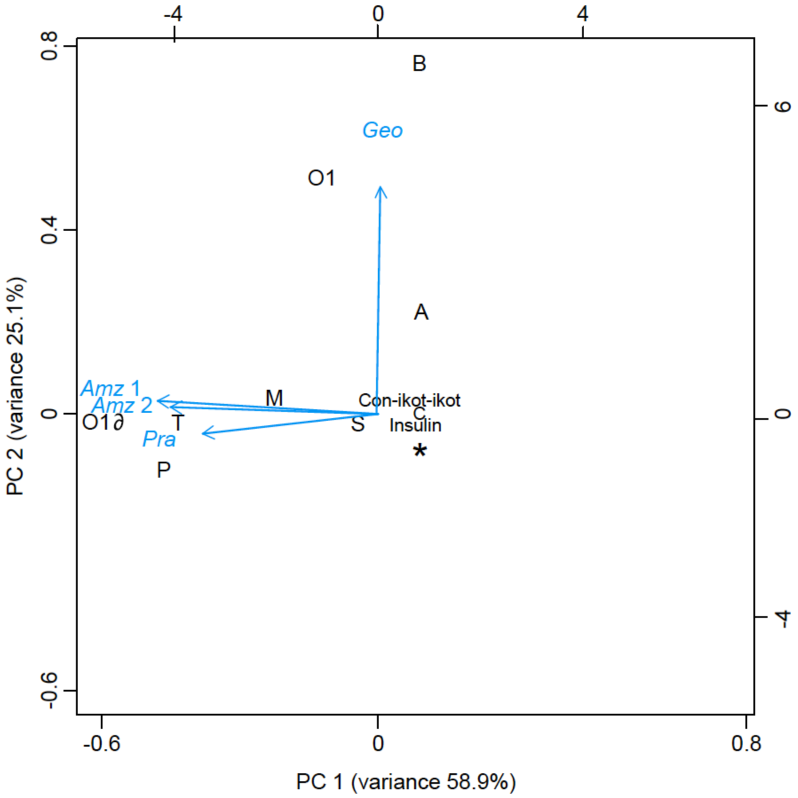


Figure 5.2 PCA analysis of conotoxin expression profiles. PCA analysis of relative expression of individual superfamilies. O1∂: ∂-like O1 superfamily; *: superfamilies that did not contribute to separation on PC1 and PC2 (conopressin, B, D, H, I1, I2, I3, I4, J, L, O2, O3, U, V).