Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Genomics 88 (2006) 809–819
GENOMICS
www.elsevier.com/locate/ygeno

Diversity and evolution of conotoxins based on gene expression profiling of *Conus litteratus* ☆

Canhui Pi [1], Junliang Liu [1], Can Peng, Yun Liu, Xiuhua Jiang, Yu Zhao [2], Shaojun Tang, Lei Wang, Meiling Dong, Shangwu Chen, Anlong Xu *

State Key Laboratory of Biocontrol, Guangdong Province Key Laboratory of Therapeutic Functional Genes,
The Open Laboratory for Marine Functional Genomics of the State High-Tech Development Program, Department of Biochemistry,
College of Life Sciences, Sun Yat-sen (Zhongshan) University, 135 Xingangxi Road, Guangzhou 510275, People's Republic of China

Received 22 April 2006; accepted 22 June 2006
Available online 14 August 2006

Abstract

Cone snails are attracting increasing scientific attention due to their unprecedented diversity of invaluable channel-targeted peptides. As arguably the largest and most successful evolutionary genus of invertebrates, *Conus* also may become the model system to study the evolution of multigene families and biodiversity. Here, a set of 897 expressed sequence tags (ESTs) derived from a *Conus litteratus* venom duct was analyzed to illuminate the diversity and evolution mechanism of conotoxins. Nearly half of these ESTs represent the coding sequences of conotoxins, which were grouped into 42 novel conotoxin cDNA sequences (seven superfamilies), with T-superfamily conotoxins being the dominant component. The gene expression profile of conotoxin revealed that transcripts are expressed with order-of-magnitude differences, sequence divergence within a superfamily increases from the N to the C terminus of the open reading frame, and even multiple scaffold-different mature peptides exist in a conotoxin gene superfamily. Most excitingly, we identified a novel conotoxin superfamily and three novel cysteine scaffolds. These results give an initial insight into the *C. litteratus* transcriptome that will contribute to a better understanding of conotoxin evolution and the study of the cone snail genome in the near future.

Keywords: Cone snail; Conotoxin; Expressed sequence tag; Diversity; Evolution; Exon shuffling

Conus is an extraordinary genus of marine gastropods. As members of one of the largest genera of marine invertebrates, almost all cone snails known to date are venomous and predatory. They utilize a cocktail of *Conus* peptides (commonly known as conopeptides or conotoxins) as a sophisticated arsenal to capture prey, defend against foes, and execute other biological purposes as well [1]. Up to the present, nowhere is found a more apparent venom biodiversity than that within the conopeptides present in the venom of cone snails [2]. It is estimated that there are 500–700 species of cone snails around the world; each species can express 50–200 different conopeptides, with little interspecies overlap [3,4]. These conopeptides demonstrate unprecedented excellent specificity and affinity for a wide range of voltage- and ligand-gated ion channels and receptors [5,6], which makes them not only effective tools for unraveling the nervous system of targeted biology but also a rich resource that provides us with neuroscience research tools and novel therapeutic agents.

Generally, cone snails are placed into three groups depending on the prey they subdue: the piscivorous species that can quickly immobilize and gulp fish, the molluscivorous species that hunt other gastropods, and the vermivorous cone snails that feed on polychaetes and other worms. Among the three groups of cone snails, vermivorous species are predominant, accounting for about 75% of all cone snails [7]. However, people have long focused on fish- and mollusc-hunting cone snails owing to the relatively close relationship between these cone snails and human beings. In contrast, the vermivorous species have been

☆ Sequence data from this article have been deposited with the GenBank Data Library under Accession Nos. ABA39796, DQ205654, DQ345351–DQ345391, DQ359921–DQ359922.

* Corresponding author. Fax: +86 20 84038377.
E-mail address: lssxal@mail.sysu.edu.cn (A. Xu).

[1] These two authors contributed equally to this work.

[2] Current address: Guangzhou Institute of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510663, People's Republic of China.

0888-7543/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2006.06.014

rarely studied. Recent studies have revealed a variety of novel conotoxin families with potent neuropharmaceutical activity existing solely in the worm-hunting cone snails [8–11]. It is reasonable to postulate that the vermivorous species might be equally promising pharmaceutical treasuries.

Present studies have divided conopeptides into two general groups, the non-disulfide-rich peptides that lack multiple disulfide bridges and the disulfide-rich conotoxins with two to five disulfide cross-links; the latter group was further sorted into at least seven superfamilies [1]. Every superfamily has its highly conserved signal peptide and characteristic pattern of cysteine residues in the mature peptide region. Up to today, a total of 14 patterns of cysteine scaffold (some are identical) have been defined [12]. In the mature peptide regions, aside from the structure-defining cysteine residues, other amino acids are hypervariable, which results in the structural and functional diversity of conotoxins.

Considering the unprecedented diversity of conopeptides, cone snail venom presents a unique opportunity for studying the evolution of large variable gene families. Recently, a few studies have tried to illuminate the origin of *Conus* peptide evolution and diversification. Duda and Palumbi [13,14] have investigated several hundreds of O-superfamily conotoxin mRNAs to understand the evolution of conotoxin multigene families; Conticello et al. [15] have further addressed this topic in a quantitative manner. On the other hand, Santos et al. [16] focused on the detailed analysis of A-superfamily conotoxin genes from several species to try to uncover the genetic events involved in the origin of superfamily branching. However, the total ESTs from a single species have seldom been detailed, while these data are of vital importance for identification of novel conotoxin genes, profiling the total expression spectrum of functional genes, and uncovering some possibly valuable molecular clues of conotoxin evolution. In this study, we focused on the detailed analysis of expressed sequence tags (ESTs) from a single species, *Conus litteratus*. Our work delineated a general expression picture of the venom duct from a typical vermivorous cone snail and uncovered some novel molecular genetic clues and evolutionary events.

Results

A total analysis of the cDNA library

After cDNA size filtration was performed in the process of library construction, the average length of cloned cDNA in the library was about 550 bp, ranging from 0.3 to 2 kb. The average meaningful readable sequence size (after subtracting vector, primer sequences, poly(A) tails, and poor-quality segments) was approximately 420 bp. The *C. litteratus* venom duct library contained 5.0×10^6 independent clones based on methods provided by the user's manual. A total of 916 sequenced high-quality ESTs were screened for homology to ribosomal RNA, mitochondrial genomic DNA, and *Escherichia coli* genomic sequences by a local server after the failed and poor-quality reads were eliminated. Among these sequences, 19 background sequences, mainly rRNA, were found and subtracted, resulting in a final set of 897 effective ESTs, on which the following analyses were based.

The trimmed ESTs from the *C. litteratus* library were assembled into clusters (or singletons) that were deposited with GenBank (by BLASTX and BLASTN) as queries for homology analysis. After being manually modified and rectified, 83 clusters were ascribed to known functional genes, which correspond to 500 clones. A total of 15 clusters (27 clones) showed high homology with unknown functional genes and the remaining 207 clusters (370 clones) had no significant similarity.

The known functional sequences were separated into two general classes: proteins involved in common cellular functions and toxic peptides (conopeptides or conotoxins). Among these matched sequences, other than 88 clones (corresponding to 41 clusters) belonging to common cellular proteins, 412 clones (42 clusters) were highly matched with conotoxin sequences or were ascribed to be conotoxins by further analysis, with T-superfamily conotoxins being the predominant components, accounting for 53.9% (222/412) of the toxin sequences. Several other typical conotoxin superfamilies such as A-, O-, and M-superfamily and non-cysteine-rich conopeptides such as contryphan and contulakin were also isolated. Several novel toxin sequences with less similarity to the present conotoxins were also identified and preliminarily verified through other experiments (unpublished data). One of the newly found conotoxins, owing to the new signal peptide and mature peptide sequences, was named as a new superfamily, the "L-superfamily." The total functional classifications and frequency distributions of different functional ESTs are shown in Table 1 and Fig. 1.

Common cellular protein ESTs

In contrast to the large amount of toxin sequences, common cellular proteins constitute only a small part of the known

Table 1
Classification of conotoxin gene transcripts from *C. litteratus* venom duct

Superfamily	Scaffold	No. of ESTs	No. of clusters	GenBank Accession No.
A	I	12	3	DQ345364–DQ345366
L	XIV	61	2	DQ205654, DQ345367
T	V	222	13	DQ345351–DQ345363
O	VI/VII	37	6	DQ345368–DQ345373
	XV	5	1	DQ345376
	0	7	2	DQ345374–DQ345375
M	III	33	5	DQ345377–DQ345381
	XVI	7	1	DQ345384
	0	8	3	DQ345382–DQ345383, DQ345385
P	IX	9	3	DQ345386–DQ345388
Non-C-rich	Contryphan	4	1	DQ345389
	Contulakin	7	2	DQ345390–DQ345391
Total		412	42	

A novel superfamily, the L-superfamily, and two newly identified conotoxin scaffolds (scaffolds XV and XVI) were first defined in this work. The scaffold XV conotoxin and two toxin clusters with no cysteine residues (scaffold 0) embrace the conserved signal peptide of the O-superfamily conotoxins, so we classified these members into the O-superfamily. Likewise, the scaffold XVI conotoxin and three members with no cysteine residues were classified into the M-superfamily.

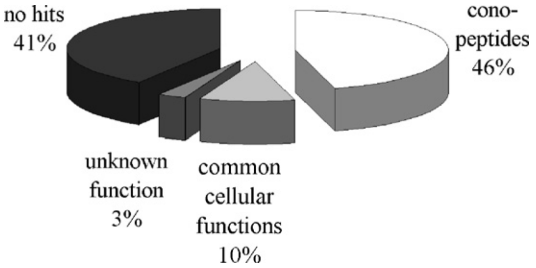

Fig. 1. General functional classification of ESTs from the *C. litteratus* venom duct.

functional proteins. These proteins were categorized as shown in Fig. 2A. From this histogram, we can see that proteins involved in defense and protein synthesis are the most abundant. Most of the immune-related proteins are many kinds of C-type lectins, such as macrophage mannose receptor, Fc receptor-like protein precursor, and brevican. The molecules related to protein synthesis are mainly diverse kinds of ribosomal proteins.

It is unclear whether these lectins are necessary components of the conotoxin cocktail or play only as members of the defensive system. Although the C-type lectins were also found

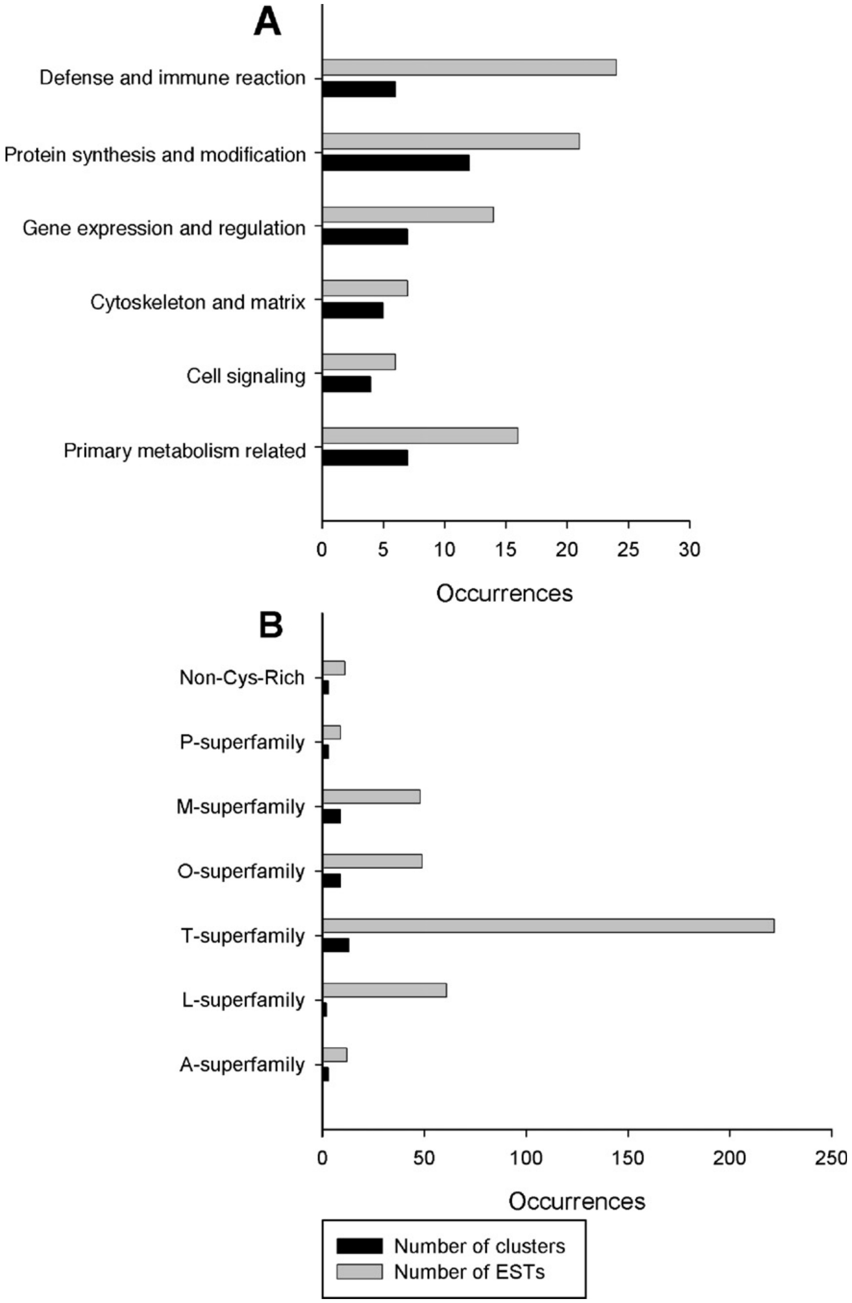

Fig. 2. Frequency distribution of the different functional categories of ESTs and the corresponding gene clusters. (A) Prevalence distribution among the common cellular proteins. (B) Prevalence distribution among the conotoxin superfamilies. The solid bars indicate the frequency distribution of clusters, whereas the gray bars indicate that of expressed sequence tags. ESTs matched with unknown proteins and with no significant similarities were not included in this analysis.

to be abundant in snake venom as components interfering with the hemostatic system [17], they appeared widely in almost all clades of life forms involving defense and immune reaction [18]. In this respect, the high prevalence of these lectins hints to us that other than the well-known function of producing offensive toxin peptides, the venom apparatus of cone snails may also play an active role in defense or even innate immune reaction.

A notable finding in our database was the unexpected high expression of transposon transposase. We identified four clones of this enzyme. As is known today, mobile elements and their remnants account for large proportions of most eukaryotic genomes, in which they have had central roles in genome evolution and hypervariation. Particularly in plants and mammals, retrotransposons have accumulated to constitute a large fraction of the genome and have shaped both genes and the entire genome [19]. The highly expressed transposase indicated that mobile elements might partly contribute to the diversification of conotoxin.

An analysis of the prevalence of each cluster revealed two genes with strikingly high frequency that have no significant matches in the available databases, including GenBank, EMBL, and DDBJ. These two genes, with 79 and 24 clones, respectively, were deposited with GenBank with the corresponding Accession Nos. DQ359921 andDQ359922.

The A-superfamily

Three kinds of cDNA sequences with the canonical organization of the A-superfamily were identified in this library.

Table 2
Comparison of conopeptide precursor sequences

A. Three A-superfamily conotoxin precursors	
Lt1.1	MGMRMMFIMFMLVVLATTVVTFTSDRALDAMNAAASNKASRLIALAVRGCCARAACAGIHQELCGGGR
Lt1.2	MGMRMMFIMFMLVVLATTVDTFTSDRALDAMNAAASNKASRLIALAVRGCCARAACAGIHQELCGGRR
Lt1.3	MGMRMMFTMFLLVVLTTTVVSFNLDRESNHENRRTSNQITRG---MWDECCDDPPCRQNNMEHCPAS▲GCCSRP
B. Three different scaffolds from O-superfamily	
Lt7.1	MEKLTILLLVAALLMSTQGLIQSGGENRPKEKIKFLSKRKTVAESWWEGECLGWSNYCTSHSICCS---GECILS-YCDIW-----
Lt15.1	MEKLTILILVATVLLAIQVLVQSDGENPVKGRVKHYAAKRFSALFRGPRE**C**TTKHRR**C**EKDEE**CC**PNLE**C**K**C**LTSPD**C**QSGYK**C**KP
Lt0.1	MKLTCMMIVAVLFLTAWTFVTADDPRDGLEDRGGWGQAGGWGKLFSKARNEMKNPKASKLDNTERRI--------------------------
Lt0.2	MKLTCMMIVAVLFLTAWTFATADDPRDGLENRGLTGEAGMLEGLSSKARDEMKNSEASKLDNTERRI--------------------------
δ-PVIA	MKLTCVMIVAVLFLTAWTFVTADDSKNGLENH------------FWKARDEMKNREASKLDKKEACYAPGTFCGIKPGLCCSEFCLPGVCFGG
C. Three different scaffolds from M-superfamily	
Lt16.1	MPKLGVSLFIFLVLFPLATLQLDGDQSAGRHAQERGE-DLFKMYQYLRRALERRRTGEDFLEE**C**MGG**C**AFDF**CC**KRSLRDTTSD
Lt3.2	MLKIGVVLFTFLVLFPLATLQLDADQPVERYAENKQDLNPNERMKMIMSALGQRR--CCISPACHEEC---YCCQ---------
Lt0.3	MMSKLGVVLFIFLVLFPMATLQLDGD--------EKD--LTQQYLNLRRVLQRGLVCAHAS PYHNAVWS--
Lt0.4	MMSKLGVVLFIFLVLFPMATLQLDGDQPADRRADEKD--LTQQYLNLRRVLQRGLVCAHAS PYHNAVWS--
Lt0.5	MMSKLGVVLFIFLVLFPMATLQLDGDQPADRRADEKDQDLTQQYLNLRRVLQRGLVCAHVR PYHNSMWS--
Lt3.2	MLKIGVVLFTFLVLFPLATLQLDADQPVERYAENKQDLNPNERMKMIMSALGQRRCCISP ACHEECYCCQ
D. Contryphan	
Contryphan-Lt	MGKLTILLLVAAALLSTQVMVQGGGDQPAARNAVPRDDNPDGMSGQFMNVLRRSGCPWEPWCG
Contryphan-R/Tx	MGKLTILVLVAAVLLSTQAMAQGDGDQPAARNAVPRDDNPDGPSAKFMNVQRRSGCPWEPWCG
E. Contulakin	
Contulakin-Lt1	MQMAYWVMVMMMVGITAPLSEGRKLNDAIRGLVPNDLTPQLLQ-SLVSRRHRVFHLDNTYLKIPICAWKVCPPTPWRRRDLKKRNK
Contulakin-Lt2	MQTAYWVMVMMMVGITAPLSEGRKLNDAIRGLVADYLTPQLLQ-SLVSAPYPEFQLDDPNLEIPVCIWKVCPPIPWRRRDLKKRNK
Contulakin-G	MQTAYWVMVMMMVWIAAPLSEGGKLNDVIRGLVPDDITPQLILGSLISRRQSEEGGSN----------ATKKPYILRASDQVASGP

Predicted signal peptides are shaded and mature peptide underlined. The hypothetical mature regions were in dashed line. In (A) ▲ means the stop codon. Regardless of this stop codon, the cysteine scaffold characteristic of Lt1.3 is similar to that of M-superfamily conotoxin. In (B) and (C) both superfamilies embrace three different mature toxin disulfide bond frameworks. The structure defined cysteine residues of novel scaffold XV and XVI were in bold. In (D) the mature peptide region of contryphan-Lt is equal to contryphan-R/Tx from *Conus textile,* and that the prepro region is conserved, indicating that these conotoxins may be derived a common ancestor or undergoing the convergent evolution. In (E) the hypothetical mature peptides shown by dashed line have two cysteine residues which characteristic is different from the typical contulakins, indicating possible divergent evolution.

We named them Lt1.1, Lt1.2, and Lt1.3. Lt1.1 has previously been cloned in *Conus leopardus,* a species closely related to *C. litteratus.* Lt1.2 appears to be a close relative of Lt1.1. The nucleotide sequence is identical except for a point mutation in the second to the last amino acid, as shown in Table 2A, with Gly being replaced by Arg. Because these two genes have slightly different sequences of 3′UTR, and both have five clones, we exclude the possibility of an error in the sequencing. Given the sequence identity between them, Lt1.1 and Lt1.2 may be allelic variants.

Another conotoxin belonging to the A-superfamily, named Lt1.3, is apparently different from the above two conotoxins. We can find the evolutionary trace of this conotoxin only by the relatively conserved signal peptide. The propeptide and mature peptide are infrequently homologous to Lt1.1 and Lt1.2; therefore, we can speculate that the A-superfamily conotoxin in this species might have at least two different branches that may derive from two distinct gene loci. It is notable that just behind the stop codon of Lt1.3, two cysteines and a stop codon can be found. Regardless of the former stop codon, this cysteine scaffold characteristic is similar to that of M-superfamily conotoxins (Table 2A).

The L-superfamily

A new conotoxin coding sequence, named Lt14.1, has been identified in this work. This kind of conotoxin is characterized by the novel cysteine scaffold "C-C-C-C". We report the full-length coding sequence of this type of conotoxin member for the first time and denote the conotoxins with similar signal peptides and characteristic pattern of cysteine residues as a new superfamily, the L-superfamily. This gene was named L-superfamily Lt14.1 (GenBank Accession No. DQ205654) according to the nomenclature proposed by McIntosh [20] and Olivera [21]. The deduced toxin (lt14a, GenBank Accession No. ABA39796) has been synthesized and the preliminary function identified [22].

It is noteworthy that Moller et al. [12] have recently reported a new family of four-cysteine, three-loop conotoxins (designated framework 14). Four peptides of this family were isolated from two worm-hunting Western Atlantic cone snail species. Comparing these framework 14 conotoxins with lt14a, we found the spacing of the loops to be quite different. Particularly, the number of intervening amino acid residues between the second and the third cysteine is 11 vs

Table 3
The cDNA and predicted translation product of Lt5.3, representing a typical T-superfamily conotoxin full-length gene

```
                                                                   atttca gcgattacac
       ctggcaggta ctcaacgaac ttcaggacac attcttttca cctgtgcacg ggaaactgac aacaagcaga

  1 ATGCGCTGTCTCCTAGTCTTCATCATTCTTCTACTGCTGATTCCATCTGCACCCAGCGTTGATGCACAACCGATG  75

  1  M  R  C  L  L  V  F  I  I  L  L  L  L  I  P  S  A  P  S  V  D  A  Q  P  M   25

 76 ACCAAAGATGATGTTCCCCTGTCATCTCTCCATGATAATGCAAAGCGAGCCCTACAAATGTTTTGGAACAAACGC 150

 26  T  K  D  D  V  P  L  S  S  L  H  D  N  A  K  R  A  L  Q  M  F  W  N  K  R   50

151 GATTGCTGCCCAGCAAAAATGTTCTGCTGTCAATGGTGAagggaaatgactttggatgagacccctgcgatgtcc 225

 51  D  C  C  P  A  K  M  F  C  C  Q  W  *                                        75

ctggatgtga gatttggaaa gcagactgtt cctttcgcgc gtgttcgtgg aatttcgaac agtcgttaac  aacacagttg
ccacttgcaa gctactatct ctttgtcatt tcatctttgg atctggatga cttaacaact gaaatatcat  gaaaagtttt
caatgggtat acactatgac catgtagtca gtaattacat cgtttggact ttttgaaaca tttttcaaaa  tgtaagtttt
ttctttggag agatcctttt tagttaaata tttcattatg ttattctttg cacacaagct atagaatgca  atctttcttt
ttgttaccac atcaatgatg ggaccagaaa ttattgggtt ttggtctatg taattatgac ctgcattaag  tgctatatgg
attgcatttt tcagggttga atgtgaatct gcaaagagaa agtggttgat cgactaataa aaacttgcat  ggcacagtaa
aaa
```

The signal sequence is shaded and mature toxin region is underlined. The coding sequence is in uppercase while untranslated region in lowercase letters. The polyadenylation signal "aataaa" is double underlined.

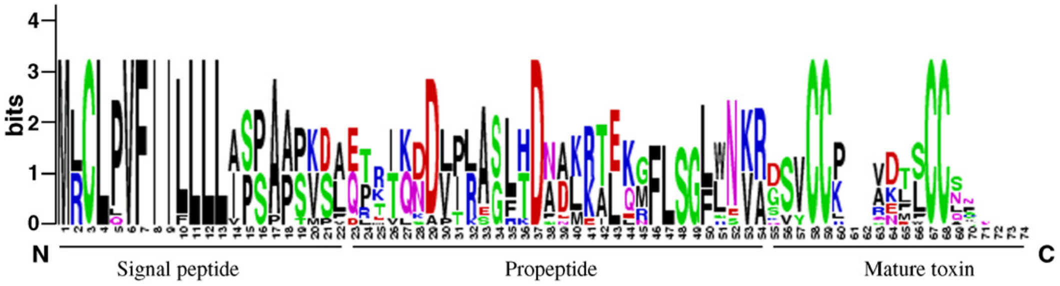

Fig. 3. Sequence logos of T-superfamily conotoxin precursors. Sequence logos were generated from protein alignments of 13 representative T-superfamily conotoxin precursors. The predicted signal peptide sequences are conserved, while the conservation dropped in the propeptide regions, and the mature toxin segments are hypervaried. However, the cysteines in the mature peptides are conserved.

3. Because the interval of the loops has enormous structural connotation, conotoxin lt14a may represent a novel conotoxin framework.

The T-superfamily

We obtained 222 T-superfamily conotoxin ESTs from this library, accounting for 53.9% (222/412) of all the toxin sequences. These sequences were ascribed to 13 clusters according to the difference in predicted coding sequences of mature peptides (each cluster represents a conotoxin coding sequence). Table 3 presents a typical full-length cDNA sequence of the T-superfamily, Lt5.3, which codes for a member of this superfamily, lt5c. All the predicted protein precursors of these clusters were listed in Fig. 3. As expected from previous studies of most conotoxin superfamilies, the predicted amino acids of these sequences revealed strong conservation in the N-terminal signal peptide regions, which dropped in the middle part of the propeptide, and the least conservation in the mature toxin segment except for the structure-defining cysteine residues.

We selected a representative full-length cDNA sequence from each cluster to analyze the molecular phylogeny. These sequences form two to four clades (clades II–IV may be ascribed to one general clade) with strong bootstrap test support (Fig. 4) and may be derived from two to four loci in light of the sequence divergence.

Meticulously comparing these clusters in the range of full-length nucleic acid sequences, we were able to tell few base differences in many clusters. Table 4 exemplifies this variety by aligning the coding regions of six similar clusters. In this case, Lt5.4.1 and Lt5.4.2 are the representative minor variants of Lt5.4. This phenomenon was also widely observed in other superfamilies. These ubiquitous (minor) variants are all the more remarkable because these sequences were from an individual cone snail.

In the mature peptide regions, the conserved cysteine can be encoded by two codons (TGC or TGT). C/T mutations in the third nucleotide of the triplet are expected to be silent, bearing no selective forces. However, an analysis of the codons of the key cysteine residues revealed that the sequences belonging to clades I and II are highly position-specific conservative (Table 4). In almost all these sequences (except one sequence belonging to Lt5.6), the first three cysteine residues are coded by TGC, and the last cysteine is coded by TGT. Given the fact that these hyperconserved codons appeared within the hypervariable mature domain, some specific mechanisms might have arisen to ensure the

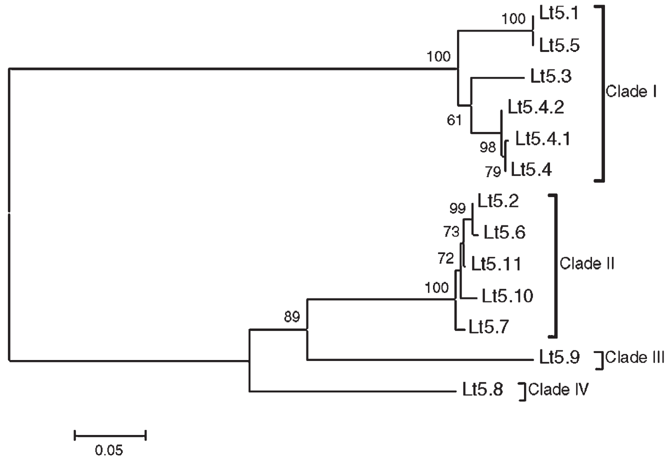

Fig. 4. Phylogenetic tree generated from full-length cDNA of *C. litteratus* T-superfamily conotoxins. A bootstrap analysis was performed using the neighbor-joining method with Kimura-2-parameter and shows that the T-superfamily conotoxins were evolved from at least two apparent clades.

Table 4
Precursor coding sequences of six T-superfamily conotoxins

Lt5.4	ATGCGCTGTCTCCCAGTCTTCATCATTCTTCTGCTGCTGATTCCATCTGCACCCAGCGTTGATGCCCAACCGACGACCAAAGATGATGTGCCCCTGGCATCT	
Lt5.4.1	....	
Lt5.4.2	....	
Lt5.3	.............T..................A................................A.......T...............T......T.....	
Lt5.5	.............A.............T....................C.....................G..A............................	
Lt5.1	....G..A............................	
Lt5.4	CTACATGATAATGCAAAGCGAGCCCTACAAATGTTTTGGAACAAACGCGATTGCTGCCCAGCAAAACTTTTATGCTGTAATCCATGATGAGGCAAATGA	32
Lt5.4.1	....T......................................	1
Lt5.4.2	....C.................................	1
Lt5.3	..C...A.G..C......C.ATGG...--...G......	12
Lt5.5	T.C..................A....GA...GAC..............TCG.........CG.G..T...........CTAAG....--.A.G......	3
Lt5.1	T.C..................A....GA...GAC..............TCG.........CA.G..T...........CTAT-.CCT--G.TG......	4

These sequences were classified into clade I by Phylogenesis analysis. Conserved nucleotides are represented by dots and gaps by dashes. End of every sequences are the number of clones. The signal peptide coding sequences are in gray and the mature peptide coding sequences underlined. The highly conserved cysteine residues codons are boxed. Note these codons are in the context of hyper–variable mature regions.

conservation of structurally crucial cysteines, which hypothesis has been proposed by Conticello et al. [23]. The highly position-specific conservative cysteine codons also reflect that these sequences may have the same origin, as speculated for scorpion toxin evolution [24,25].

The O-superfamily

A total of 49 ESTs (nine clusters) were categorized into O-superfamily conotoxins according to the conserved signal peptide sequences. Previous studies have defined a superfamily usually by two sequence elements: a highly conserved signal sequence and a characteristic arrangement of Cys residues in the primary amino acid sequence of the mature peptide region. In the case of the O-superfamily, the normal arrangement of cysteine residues is C-C-CC-C-C. However, as shown in Table 2B, the toxin sequences with the characteristic signal peptide of the O-superfamily in our library exhibit highly divergent mature regions. We can roughly divide these members into three groups according to the variation of mature peptides: the canonical O-superfamily conotoxins with six cysteine residues in the mature region represented by Lt6.1–6.5 and Lt7.1; one variant conotoxin with a brand new cysteine scaffold, C-C-CC-C-C-C-C (we named Lt15.1); and two odd members (Lt0.1 and Lt0.2) that have no cysteine residue in the mature region or probably even have no mature peptide.

This highly divergent event has rarely been found previously in conotoxins. Sequence comparison shows that the three branches of mature peptides are completely different not only in general sequence homology, but also in the arrangement of cysteine residues. The huge sequence difference shows that the three different branches of mature peptide belonging to this superfamily may be encoded by different exons.

The M-superfamily

The overall characteristics of M-superfamily conotoxin clusters are similar to those of the O-superfamily. A total of 48 ESTs with the highly homologous signal peptide sequences of this superfamily can be clustered into nine consensus sequences, among which five embrace the canonical mature peptide of this superfamily identified previously, characterized by the typical cysteine scaffold of the M-superfamily, CC-C-C-CC. As appeared in the O-superfamily, one cluster with an atypical cysteine scaffold, C-C-CC, which has never been found previously in conotoxins (we gave this member a new name, Lt16.1), and three other clusters without even an apparent mature peptide (Table 2C) have been identified.

The P-superfamily

We identified nine ESTs whose predicted translation products have the typical cysteine arrangement characteristic of P-superfamily members in the mature region. However, the coding signal peptide of these ESTs is scarcely homologous to those that appeared in P-superfamily conotoxin precursors reported previously [26,27]. These nine clones were grouped into three variants, one of which was verified by biochemical extraction and N-terminal peptide sequencing combined with matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (unpublished results). These three genes of the P-superfamily were deposited with GenBank under Accession Nos. DQ345386–DQ345388.

Non-cysteine-rich conopeptides

A total of 11 ESTs were identified with high similarity to non-cysteine-rich conopeptides, including contryphan (four ESTs) and contulakin (seven ESTs). These sequences were classified into three putative toxin gene sequences, contryphan-Lt, contulakin-Lt1, and contulakin-Lt2. However, to our surprise, the conceptual translation product of the contryphan-Lt mature region is identical to that of Contryphan-R/Tx [28] (Table 2D). Furthermore, both the predicted mature peptides of two contulakin-Lt genes have one disulfide cross-link, while previously identified contulakin-G has not been found to contain disulfide bonds [29,30] (Table 2E).

Discussion

Overview of Conus study strategies

The cone snails have come to human attention mainly in that they can be deadly to humans. The high frequency of fatalities caused by stings of *C. geographus* has been well documented in the medical literature [31]. The recognition that snails could kill people led to the initial scientific investigations, which focused on physiological and pharmacological properties of whole venom. Subsequently, with the development of modern protein chemistry techniques, the individual venom components became of increasing interest and were broadly studied. These studies have discovered that the venom of cone snails is a rich resource of pharmacologically active peptides and as a result opened a Pandora's box of potential drugs for commercial development as clinical pharmaceuticals [32]. Now, the conotoxin MVIIA (Ziconotide) has been approved by the FDA as a therapeutic for severe resistant pain; more candidates are under clinical trial [33].

These routine studies by protein chemistry always aim at the isolation of specific active components, subsequently cloning the genes by the PCR method with the primers deduced from protein sequences. Several recent studies also report the direct cDNA cloning of novel members of a superfamily based on the reported conserved gene sequences (usually the signal peptide-coding region) of this superfamily by RACE methods [34,35]. However, these studies cannot find new members without conserved sequences and hardly present the general picture of gene expression even within a specific conotoxin superfamily.

An analysis of conotoxins by both protein chemistry and specific primer PCR has raised the notion of the diversity and complexity of conotoxins. To examine this diversity in terms of gene expression and further project the genetics and evolution of conotoxins, we adopted the comprehensive molecular approach of generation and analysis of ESTs from the *C. litteratus* venom duct. This EST database showed the relative abundance of toxin transcripts over the total cellular transcriptome. It also allowed the identification of diverse classes of toxins and variants and demonstrated the relative proportion of each type of conotoxin. But undoubtedly EST sequences are somewhat independent of function. Especially when a large number of conotoxins contain complicated posttranslational modifications, only combined with biochemical methods can EST methods play an active role in addressing the relationships of conotoxin molecular genetics, structure, and function.

Conotoxin transcripts are expressed with order-of-magnitude differences

Upon plotting the number of ESTs for each conotoxin transcript, we observed the vast differences in the expression levels of different toxins. First of all, ESTs belonging to T-superfamily conotoxins are apparently the most abundant, accounting for 53.9%. Other superfamilies of conotoxin ESTs are presented at a relatively even distribution (Table 1). The differentiation of conotoxin expression patterns in different diet types of cone snails was also partly supported by the present research. Most conotoxins identified from piscivorous cone snails to date belong to the A- and O-superfamilies, which has recently been detailed elsewhere [36,37], whereas T-superfamily conotoxins are derived mainly from the vermivorous species [38]. Nowadays, the definite molecular targets of the T-superfamily have not been identified although many efforts have been made [38]. One possible reason is that the T-superfamily conotoxins might target specifically the lower invertebrate ion channels and receptors, which have rarely been isolated and tested.

These differences exist not only between different superfamilies, but also between the various members of the same superfamily as well. As shown in Fig. 2, for the T-superfamily with multiple members, order-of-magnitude differences between diverse transcripts were identified. These differences were also observed during purification of the protein components of this species (unpublished data). Whether there is a certain regulatory mechanism existing in these quantitative variations or a long-term adaptive evolution result in this situation is an open question. Conticello et al. [15] have also described the huge expression differences in conotoxins in some other *Conus* species. They reported that the unrooted dendrograms for highly expressed toxin sequences, but not for total sequences, revealed clear segregation between toxins from molluscivorous and vermivorous species. Therefore, it is possible that the highly expressed toxin sequences were strongly selected by different prey types, while the large number of unselected variants provides the reservoir available for further diversifying selection in times of seasonal and ecological environmental change.

A high abundance of minor variants was detected in conotoxin sequences

A notable observation on conotoxin transcripts is that many of the highly expressed toxin sequences possess mutant sequences with differences in only a few bases, or even one. We can basically exclude the possibility that these minor variants are derived from artifacts of original amplification and sequencing because the common cellular proteins from the same library do not present these corresponding variants. Furthermore, the variation was most prominent within the region that encodes the mature conotoxin (Table 4). Duda and Palumbi [14] have also reported this phenomenon. They verified the four-loop conotoxin mRNA sequences and their multiforms of minor variants are alleles of a single locus, rather than from many loci. A similar situation also happened in the process of snake venom expression [17].

Because a large number of variants we obtained were from a single locus in one individual, we can hardly explain it only by gene duplication and mutation in different specimens. Perhaps some unconventional mechanisms, such as overdominant selection making for an excess of heterozygotes, which established an explanation of the diversification of the major histocompatibility complex, are required to illuminate partially the origin of conotoxin hypervariability [39].

One conserved signal peptide does not always couple one cysteine scaffold-type of mature peptide. What is the mechanism?

Conotoxins are initially translated as prepropeptides, which usually have three canonical segments: the signal peptide, the propeptide, and the mature region. The precursors are then cleaved by a specific proteolytic enzyme to generate mature toxins. Generally, within the same superfamily, the signal sequences at the N-terminal are highly conserved, whereas the mature peptides at the C-terminal are diverged remarkably aside from the structural defined cysteine residues. This means that the conserved signal peptide sequences always have a conserved pattern of cysteine arrangement in the mature peptide region. For example, the O-superfamily conotoxins are characterized by a specific cysteine framework, "C-C-CC-C-C," defined as scaffolds VI/VII, leading to a four-loop structural configuration, whereas M-superfamily conotoxins are characterized by the specific cysteine framework "CC-C-C-CC," defined as scaffold III. However, in the present work, we found that the conotoxin precursors with the conserved O-superfamily signal peptide link separately with two novel frameworks in the mature peptide region: one of these frameworks, possessing eight cysteine residues, we defined as scaffold XV; another appears to have no cysteine residues. Accordingly, this situation also happened in the M-superfamily (Table 2). Furthermore, the predicted non-cysteine-rich conopeptides (contulakin-Lt1 and contulakin-Lt2) with one disulfide cross-link also show the different mature peptide scaffold from previously described contulakin-G.

Such diversified clones have rarely been reported in the study of conotoxins. They may be generated via selective splicing at the transcription level or even have already taken place on the genomic level by unequal crossing over, site-specific recombination, or exon shuffling. Among all the possible reasons, DNA-based exon shuffling is noteworthy. Lee et al. [40] have identified 56 different antimicrobial peptide cDNA sequences, each of which encodes a precursor that can give rise to two kinds of antimicrobial peptides, maximin and maximin H. They notice that in different genes, certain maximin sequences could link with various maximin H, and vice versa. The genomic structure revealed that domain shuffling among these genes might have happened frequently. Considering that the three domains of the conotoxin precursors may be encoded separately by three exons [41], exon shuffling events may have played an important role during the evolution of these proteins, yet this hypothesis requires further experimental validation.

In this study, we have presented a global view of the cone snail venom-secreting tissue. Our EST database uncovered the extraordinarily high diversity and huge exploitation potential of conotoxins in a typical vermivorous cone snail. Analysis of conotoxin sequence polymorphisms revealed three novel scaffold members and detected some novel genetic events (multiform scaffolds of a mature peptide may link with a kind of signal peptide). This initial work provided some clues for deciphering the genetic and evolutionary mechanisms of the unprecedented diversity of conotoxins and took the first step toward a better understanding of venom duct tissue on a genomic basis. In addition, some highly expressed ESTs with no significant similarity to the present data may open up new avenues for further exploration.

Materials and methods

cDNA library construction

A single specimen of *C. litteratus* was collected from Yalong Bay, the south end of Hainan Island, China. The venom duct was dissected from the living snail and then immediately frozen in liquid nitrogen. The frozen tissues were homogenized and dissolved in TRIzol. The following steps for total RNA isolation and cDNA library construction were performed as previously described [42]. cDNA size fractionation was performed with a Chroma Spin-400 provided with library construction kits and the fractions longer than 400 bp were collected. The presence and size of inserts were tested by PCR of randomly chosen colonies using vector primers (T7 forward and SP6 reverse primers). Further verification was performed by digestion with the restriction enzyme *Sfi*III of total cloned plasmid and analysis by electrophoresis.

Expressed sequenced tags sequencing

cDNA clones were randomly picked and cultured in individual wells of 96-well plates containing appropriate 2×YT medium with 100 mg/L ampicillin. After overnight culture, plasmids were prepared by using AxyGen 96-Easy plasmid miniprep kits (AxyGen Biochemical Technique). Purified plasmid DNAs were single-pass sequenced from the 5′ end in an automated ABI Prism 3730 sequencer (Perkin–Elmer) using the T7 promoter primer and ABI Prism Big Dye Terminator v3.1 ready reaction cycle sequencing kits (Applied Biosystems).

Sequence data analysis

Prior to further analysis, sequencing outputs were trimmed by removal of vector, primer sequences, and poly(A) tails with ABI Prism DNA Sequencing Analysis software v3.3. Low-quality segments and inserts less than 100 bp were also subjected to removal. All valid sequences were assembled into clusters (also called consensus sequences) with a minimal score of 95 using the software. The consensus sequences of each cluster and singleton were further filtered by screening for homology to ribosomal RNA, mitochondrial genomic DNA, and *E. coli* genomic sequences [43,44]. After matches were deleted, the remaining sequences were used as a query to search the protein database available from NCBI by the BLASTX algorithm with an E-value cutoff at 10^{-6} [45]. The unmatched ESTs were further searched in the nucleotide sequence database by the BLASTX algorithm with an E-value cutoff at 10^{-4}.

After BLAST search, the possible toxin ESTs were further identified by manual inspection of unmatched clusters for open reading frames using the software Seqtools 8.3 (http://www.seqtools.dk/); those enriched for cysteines in the predicted C-terminal segment were further analyzed and verified. In the following analysis, annotations of possible protein-coding genes were carried out in detail. Finally, the ESTs, clustering results, and corresponding annotations were deposited with a local searchable database named MOSAS (Marine Organism Sequence Analysis System: http://cbi.sysu.edu.cn/extend/datarelease.htm). The ESTs were also deposited with the database of GenBank under Accession Nos. DQ205654, DQ345351–DQ345391, and DQ359921–DQ359922.

Phylogenetic analysis and signal peptide prediction

Sequences were initially aligned by CLUSTALX and then manually adjusted. The phylogenetic analysis was performed by the neighbor-joining method using MEGA v3.1 software [46]. Bootstrap values were estimated from 500 replicates. Sequence logos were plotted according to Crooks et al. [47].

The signal peptide sequences and cleavage sites of the novel conotoxin precursors were predicted online with the SignalP 3.0 server [48] (http://www.cbs.dtu.dk/services/SignalP/).

Acknowledgments

The authors thank Dongmei Chen and her staff for EST sequencing and Yang Shen and Hongkai Wu for their help with sequence analysis. This work was supported by grants from the State High-Tech Research and Development Project (863) of the Ministry of Science and Technology of China (2003AA626010; 2004AA621030) and the Project of the National Natural Science Foundation of China (30300264).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2006.06.014.

References

[1] H. Terlau, B.M. Olivera, Conus venoms: a rich source of novel ion channel-targeted peptides, Physiol. Rev. 84 (2004) 41–68.

[2] R.J. Lewis, M.L. Garcia, Therapeutic potential of venom peptides, Nat. Rev. Drug Discovery 2 (2003) 790–802.

[3] B.M. Olivera, E.E. Just Lecture, 1996: Conus venom peptides, receptor and ion channel targets, and drug design: 50 million years of neuropharmacology, Mol. Biol. Cell 8 (1997) 2101–2109.

[4] B.G. Livett, K.R. Gayler, Z. Khalil, Drugs from the sea—Conopeptides as potential therapeutics, Curr. Med. Chem. 11 (2004) 1715–1723.

[5] L.J. England, et al., Inactivation of a serotonin-gated ion channel by a polypeptide toxin from marine snails, Science 281 (1998) 575–578.

[6] I.A. Sharpe, et al., Two new classes of conopeptides inhibit the alpha1-adrenoceptor and noradrenaline transporter, Nat. Neurosci. 4 (2001) 902–907.

[7] A.J. Kohn, F.E. Perron, Life History and Biogeography: Patterns in *Conus*, Oxford Univ. Press, Oxford, 1994.

[8] S. Kauferstein, et al., Novel conopeptides of the I-superfamily occur in several clades of cone snails, Toxicon 44 (2004) 539–548.

[9] A. Zugasti-Cruz, et al., Amino acid sequence and biological activity of a gamma-conotoxin-like peptide from the worm-hunting snail Conus austini, Peptides 27 (2006) 506–511.

[10] M.B. Aguilar, et al., A novel conotoxin from Conus delessertii with posttranslationally modified lysine residues, Biochemistry 44 (2005) 11130–11136.

[11] H. Jiang, et al., A novel M-superfamily conotoxin with a unique motif from Conus vexillum, Peptides 27 (2006) 682–689.

[12] C. Moller, et al., A novel conotoxin framework with a helix–loop–helix (Cs alpha/alpha) fold, Biochemistry 44 (2005) 15986–15996.

[13] T.F. Duda, S.R. Palumbi, Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod Conus, Proc. Natl. Acad. Sci. U. S. A. 96 (1999) 6820–6823.

[14] T.F. Duda, S.R. Palumbi, Evolutionary diversification of multigene families: allelic selection of toxins in predatory cone snails, Mol. Biol. Evol. 17 (2000) 1286–1293.

[15] S.G. Conticello, et al., Mechanisms for evolving hypervariability: the case of conopeptides, Mol. Biol. Evol. 18 (2001) 120–131.

[16] A.D. Santos, J.M. McIntosh, D.R. Hillyard, L.J. Cruz, B.M. Olivera, The A-superfamily of conotoxins: structural and functional divergence, J. Biol. Chem. 279 (2004) 17596–17606.

[17] L. Junqueira-de-Azevedo Ide, P.L. Ho, A survey of gene expression and diversity in the venom glands of the pitviper snake Bothrops insularis through the generation of expressed sequence tags (ESTs), Gene 299 (2002) 279–291.

[18] N. Zhang, et al., Molecular profile of the unique species of traditional Chinese medicine, Chinese seahorse (Hippocampus kuda Bleeker), FEBS Lett. 550 (2003) 124–134.

[19] H.H. Kazazian, Mobile elements: drivers of genome evolution, Science 303 (2004) 1626–1632.

[20] J.M. McIntosh, A.D. Santos, B.M. Olivera, Conus peptides targeted to specific nicotinic acetylcholine receptor subtypes, Annu. Rev. Biochem. 68 (1999) 59–88.

[21] B.M. Olivera, Conus venom peptides: reflections from the biology of clades and species, Annu. Rev. Ecol. Syst. 33 (2002) 25–47.

[22] C. Peng, et al., Discovery of a novel class of conotoxin from *Conus litteratus,* lt14a, with a unique cysteine pattern, Peptides (in press).

[23] S.G. Conticello, Y. Pilpel, G. Glusman, M. Fainzilber, Position-specific codon conservation in hypervariable gene families, Trends Genet. 16 (2000) 57–59.

[24] O. Froy, et al., Dynamic diversification from a putative common ancestor of scorpion toxins affecting sodium, potassium, and chloride channels, J. Mol. Evol. 48 (1999) 187–196.

[25] S. Zhu, H. Darbon, K. Dyason, F. Verdonck, J. Tytgat, Evolutionary origin of inhibitor cystine knot peptides, FASEB J. 17 (2003) 1765–1767.

[26] M.B. Lirazan, et al., The spasmodic peptide defines a new conotoxin superfamily, Biochemistry 39 (2000) 1583–1588.

[27] L.A. Miles, et al., Structure of a novel P-superfamily spasmodic conotoxin reveals an inhibitory cystine knot motif, J. Biol. Chem. 277 (2002) 43033–43040.

[28] E.C. Jimenez, M. Watkins, L.J. Juszczak, L.J. Cruz, BM. Olivera, Contryphans from Conus textile venom ducts, Toxicon 39 (2001) 803–808.

[29] A.G. Craig, et al., Contulakin-G, an O-glycosylated invertebrate neurotensin, J. Biol. Chem. 274 (1999) 13752–13759.

[30] A.G. Craig, et al., Enzymatic glycosylation of contulakin-G, a glycopeptide isolated from Conus venom, with a mammalian ppGalNAc-transferase, Toxicon 39 (2001) 809–815.

[31] L.J. Cruz, J. White, Clinical toxicology of *Conus* snail stings, in: J. Meier, J. White (Eds.), Clinical Toxicology of Animal Venoms, CRC Press, Boca Raton, FL, 1995.

[32] B.M. Olivera, et al., Neuronal calcium channel antagonists: discrimination between calcium channel subtypes using omega-conotoxin from Conus magus venom, Biochemistry 26 (1987) 2086–2090.

[33] D.P. Wermeling, Ziconotide, an intrathecally administered N-type calcium channel antagonist for the treatment of chronic pain, Pharmacotherapy 25 (2005) 1084–1094.

[34] B.S. Lu, F. Yu, D. Zhao, P.T. Huang, C.F. Huang, Conopeptides from Conus striatus and Conus textile by cDNA cloning, Peptides 20 (1999) 1139–1144.

[35] S. Kauferstein, C. Melaun, D. Mebs, Direct cDNA cloning of novel conopeptide precursors of the O-superfamily, Peptides 26 (2005) 361–367.

[36] C. Pi, et al., Analysis of expressed sequence tags from the venom ducts of Conus striatus: focusing on the expression profile of conotoxins, Biochimie 88 (2006) 131–140.

[37] H. Terlau, et al., Strategy for rapid immobilization of prey by a fish-hunting cone snail, Nature 381 (1996) 148–151.

[38] C.S. Walker, et al., The T-superfamily of conotoxins, J. Biol. Chem. 274 (1999) 30664–30671.

[39] A.L. Hughes, M. Nei, Maintenance of MHC polymorphism, Nature 355 (1992) 402–403.

[40] W.H. Lee, et al., Variety of antimicrobial peptides in the Bombina maxima toad and evidence of their rapid diversification, Eur. J. Immunol. 35 (2005) 1220–1229.

[41] B.M. Olivera, et al., Speciation of cone snails and interspecific hyperdivergence of their venom peptides: potential evolutionary significance of introns, Ann. N. Y. Acad. Sci. 870 (1999) 223–237.

[42] Y. Yang, et al., EST analysis of gene expression in the tentacle of cyanea capillata, FEBS Lett. 538 (2003) 183–191.

[43] R. Sorek, H.M. Safer, A novel algorithm for computational identification of contaminated EST libraries, Nucleic Acids Res. 31 (2003) 1067–1074.

[44] G. Weber, J. Shendure, D.M. Tanenbaum, G.M. Church, M. Meyerson, Identification of foreign gene sequences by transcript filtering against the human genome, Nat. Genet. 30 (2002) 141–142.

[45] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler, GenBank, Nucleic Acids Res. 33 (2005) 34–38.

[46] S. Kumar, K. Tamura, M. Nei, MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment, Brief. Bioinformat. 5 (2004) 150–163.

[47] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190.

[48] J.D. Bendtsen, H. Nielsen, G. von Heijne, S. Brunak, Improved prediction of signal peptides: SignalP 3.0, J. Mol. Biol. 340 (2004) 783–795.