Characterizing the evolution and functions of the M-superfamily conotoxins

Maojun Zhou\textsuperscript{a,1}, Lei Wang\textsuperscript{a,1}, Yun Wu\textsuperscript{a,1}, Xiaoyan Zhu\textsuperscript{a}, Yuchao Feng\textsuperscript{a}, Zelin Chen\textsuperscript{a}, Yuxin Li\textsuperscript{a}, Dandan Sun\textsuperscript{a}, Zhenghua Ren\textsuperscript{a}, Anlong Xu\textsuperscript{a,b,*}

\textsuperscript{a} State Key Laboratory of Biocontrol, Guangdong Province Key Laboratory of Pharmaceutical Functional Genes, National Engineering Research Center of South China Sea Marine Biotechnology, Department of Biochemistry, College of Life Sciences, Sun Yat-sen University, Guangzhou 510275, People's Republic of China
\textsuperscript{b} Beijing University of Chinese Medicine, 11 Dong Shan Huan Road, Beijing 100029, People's Republic of China

\textbf{Abstract}

Conotoxins from cone snails are valuable in physiology research and therapeutic applications. Evolutionary mechanisms of conotoxins have been investigated in several superfamilies, but there is no phylogenetic analysis on M-superfamily conotoxins. In this study, we characterized identical sequences, gene structure, novel cysteine frameworks, functions and evolutionary mechanisms of M-superfamily conotoxins. Identical M-superfamily conotoxins can be found in different \textit{Conus} species from the analysis of novel 467 M-superfamily conotoxin sequences and other published M-superfamily conotoxin sequences. M-superfamily conotoxin genes consist of two introns and three exons from the results of genome walking. Eighteen cysteine frameworks were identified from the M-superfamily conotoxins, and 10 of the 18 may be generated from framework III. An analysis between diet types and phylogeny of the M-superfamily conotoxins indicate that M-superfamily conotoxins might not evolve in a concerted manner but were subject to birth-and-death evolution. Codon usage analysis shows that position-specific codon conservation is not restricted to cysteines, but also to other conserved residues. By analysing primary structures and physiological functions of M-superfamily conotoxins, we proposed a hypothesis that insertions and deletions, especially insertions in the third cysteine loop, are involved in the creation of new functions and structures of the M-superfamily conotoxins.

\textcopyright{} 2013 Elsevier Ltd. All rights reserved.

\textbf{1. Introduction}

Marine venomous cone snails (genus \textit{Conus}), which consist of 500–700 species, feed on polychaetes, molluscs, hemichordates and fish (Duda and Palumbi, 1999; Han et al., 2008a; Olivera, 2006). Cone snails use a potent cocktail of peptide toxins to hunt their prey, deter competitors and defend against predators (Norton and Olivera, 2006). Being extensively studied in the past three decades, each \textit{Conus} species is estimated to have a repertoire of 50–200 species–specific conotoxins (Mcintosh and Jones, 2001; Terlau and Olivera, 2004). Each conotoxin is encoded by a gene which contains 0–2 introns and processes from a prepropeptide precursor (typically 50–110 amino acids)
(Olivera et al., 1999; Yuan et al., 2007). The propeptide precursor includes an N-terminal hyperconserved signal peptide region (typically 19–27 amino acids), a rather conserved pro-region (typically 20–40 amino acids), and a C-terminal hypervariable mature toxin region (typically 10–40 amino acids) (Kaas et al., 2010; Puillandre et al., 2012). Conotoxins, with diverse post-translational modifications, target different ion channels, G-protein coupled receptors and neurotransmitter transporters with a high degree of specificity and affinity, which makes them useful tools in neuroscience research and potential therapeutic agents. Several conotoxins are on clinical trial and one (ω-Conotoxin MVIIA) was approved by FDA for the treatment of otherwise unmanageable severe chronic pain (Buczek et al., 2005; Han et al., 2008a; Lewis, 2009; Norton and Olivera, 2006; Olivera, 2002).

Conotoxins can be categorized into 25 superfamilies: A-, B-, C-, D-, E-, F-, G-, H-, I1-, I2-, J-, L-, M-, O1-, O2-, O3-, P-, R-, S-, T-, V-, X1-, X2-, X3-, X4-, X5-, X6-, X7 and Y- superfamilies up to now (Kaas et al., 2010; Puillandre et al., 2012; Wang and Chi, 2004). M-superfamily conotoxins presumably exist in all Conus species (Jacob and McDougal, 2010; Kaas et al., 2010), although there is argument regarding particular Conus species (Biggs et al., 2010). According to the signal region, M-superfamily conotoxins can be divided into the MLKM group and the MMSK group (Wang et al., 2008), according to the number of amino acids that exist between the fourth and fifth cysteine of framework III, M-superfamily conotoxins can be divided into the M-1, M-2, M-3, M-4 and M-5 branches (Corpuz et al., 2005). The physiological functions of some M-superfamily conotoxins are that they can block voltage-gated potassium channels (Al-Sabi et al., 2004), inhibit nicotinic acetylcholine receptors (Lluisma et al., 2008; Shon et al., 1997; Van Wagoner et al., 2003) or block voltage-gated sodium channels up to now (Fainzlizer et al., 1995; Han et al., 2008b; Imperial et al., 2008; Jiang et al., 2006; Pi et al., 2006; Wu et al., 2010).

Many mechanisms have been proposed to elucidate the evolution of conotoxins and other toxins. Sphingomyelinase D from spider venom, three-finger toxins and disintegrins from snake venom are subject to birth-and-death evolution (Binford et al., 2009; Fry et al., 2003; Juarez et al., 2008). Accelerated evolution and gene duplication followed by positive Darwinian selection contribute to the evolution of the venom compositions of Conus, spiders, snakes and scorpions (Binford et al., 2009; Lynch, 2007; Ohno et al., 2003; Puillandre et al., 2010; Zhu et al., 2004; Zuponksi et al., 2003). Decelerated evolution is also detected in snake venom toxins as a consequence of ecological niche shift (Li et al., 2005). Dietary specialization is associated with conotoxin allelic frequencies differentiation intraspecies (Duda et al., 2009) or reduction of conotoxins in Conus (Remigio and Duda, 2008). Insertions and deletions (indels) in toxin genes may lead to lower venom toxicity of Kangroo Island snakes (Doley et al., 2008).

Previous studies have indicated that Conus can feed on polychaetes, hemichordates, molluscs or fishes (Duda et al., 2001). Do the conotoxins vary greatly among Conus with different diet types? Are the M-superfamily conotoxin genes subject to birth-and-death evolution? Do indels take part in the generation of new genes? In this study, we attempt to address these questions and elucidate relationships among different diet types and mechanisms of interspecific conotoxin diversification. With a combination of 467 M-superfamily conotoxin sequences isolated from cDNAs and genomes of 18 Conus species with 180 published M-superfamily conotoxin genes from GenBank, we performed a large-scale analysis of M-superfamily conotoxins and their cysteine patterns. Subsequently, we performed phylogenetic analysis, codon usage analysis and indels analysis to elucidate the evolutionary mechanisms of conotoxins.

2. Material and methods

2.1. Specimens

Eighteen Conus species were collected from reef flats in Sanya, the south end of Hainan Island, China. For preparation of the venom duct total RNAs, eight individuals per species were used for the following 12 species: Conus textile, Conus lividus, Conus quercinus, Conus capitanus, Conus ebraeus, Conus tessulatus, Conus miles, Conus vexillum, Conus coronatus, Conus betulinus, Conus emaciatius and Conus caracteristicus. Three individuals per species were used for Conus litteratus, Conus rattus and Conus striatus; two individuals per species were used for Conus terebra and Conus varius. One individual was used for Conus planorbis.

2.2. Conotoxin sequences from cDNAs

SMARTer PCR cDNA Synthesis Kit (Clontech) was used to produce First-Strand cDNAs and Second-Strand cDNAs for 18 Conus species. The cDNAs were reverse transcribed from 50 ng poly A(+) mRNA following the manufacturer’s instructions, except for 3’ SMART CDS Primer II A and 5’ PCR Primer II A, which were modified as previously described (Beldade et al., 2006). The 3’RACE was performed on the first-strand cDNAs of 18 Conus species. Forward primers and reverse primers of 3’RACE were designed according to M-superfamily 5’-Untranslated Regions (UTR) (Conticello et al., 2001) and 3’ universal primers. To amplify M-superfamily conotoxin coding sequences, primers were designed according to conserved 5’- and 3’-UTR as previously described (Conticello et al., 2001; Holford et al., 2009). For each of the 18 Conus species, at least 20 positive clones from each reaction were sequenced.

2.3. Genomic structure

A mollusc DNA kit (OMEGA) was used for the preparation of genomic DNAs from the footpads of 11 Conus species: C. textile, C. capitanus, C. miles, C. betulinus, C. emaciatius, C. caracteristicus, C. rattus, C. terebra, C. varius, C. planorbis and C. striatus. We designed several pairs of primers based on the 5’-UTR, 3’-UTR and coding regions, but we failed to obtain any M-Superfamily conotoxin sequences from genomic DNA. Therefore, we deduced that there were one or more introns in the conotoxin gene. Subsequently, the GenomeWalker Universal Kit (Clontech) was used to construct genome walking libraries of C. terebra, C. textile and C. striatus. Gene-specific primers were
designed according to the conotoxin signal, pro-region and mature region sequences separately. To find unknown adjacent intron sequences, PCRs were performed towards 3′ for the signal region, 5′ for the mature region and both 3′ and 5′ for the pro-region. One conotoxin from the M-1, M-2, M-3, and M-5 branches was analysed separately for its genomic structure. To obtain mature region sequences, PCRs were performed on 11 Conus species genomic DNAs using primers corresponding to the conserved 3′ end of the second intron and 3′-UTR. For each of the 11 Conus species, at least 20 positive clones from each reaction were sequenced.

2.4. Conotoxin tree

Only complete coding sequences were used for the following phylogenetic analysis. The nucleotide alignments were translated to protein sequences using Codon Alignment. The corresponding translated amino acid sequences were aligned automatically using MUSCLE 3.8 (Edgar, 2004). According to the sequence alignment, we mapped the potential diversity of the M-superfamily cysteine patterns. Conotoxins with cysteine patterns 0C, 1C, 2C, 3C, XVI, M-3, and M-5 branches was analysed separately for its substitution models. The Bayesian conotoxin trees were constructed using MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001). Two unrooted Bayesian trees were constructed using prepropeptides of the MLKM group conotoxins and prepropeptides of the MMSK group conotoxins, respectively. We performed a MCMC analysis with four chains and ran the chains for a minimum of 2,000,000 cycles. TreeDyn (Chevenet et al., 2006) was used to edit and analyse the trees.

3. Results and discussion

3.1. Identical conotoxins in different Conus species

M-superfamily conotoxin sequences were obtained by PCR cloning from cDNAs of 18 Conus species and genomic DNAs of 11 Conus species (Table 1). After sequencing more than 4000 clones, 467 novel M-superfamily conotoxin sequences were isolated from 18 Conus species. The 467 novel M-superfamily conotoxin sequences, which mainly are 200–250 base pairs (bps) long, were submitted to GenBank [GenBank accession numbers: JF510538–JF511005]. All published M-superfamily conotoxin nucleotide and protein sequences were downloaded from NCBI and ConoServer (Kaas et al., 2008), including 180 individual nucleotide sequences from 61 Conus species. Subsequently, all of the sequences from different species were analysed, and identical sequences were recovered from different species. In order to verify whether this result was caused by the contamination in the lab, we collected six polychaete-hunting Conus from Hainan Island and repeat all the experiments in another lab separately. Identical conotoxins can still be found in different Conus species. Statistically, nearly 17% (79 out of 467) of M-superfamily conotoxins can be found in two or more Conus species. Although most researches on conotoxins indicate that each Conus species has species-specific conotoxins, identical conotoxins were found within different Conus species in A, O and I2 superfamily (Kauferstein et al., 2005; Liu et al., 2009; Nicke et al., 2003). Our findings indicate that identical conotoxin sequences were recovered from different Conus species is a common phenomenon for the conotoxin superfamilies. It is noteworthy that identical sequences were found among vermivorous, molluscivorous and piscivorous Conus species, which supporting the hypothesis that fish-hunting.

Table 1
Statistics of M-superfamily conotoxin sequences from the cDNAs and genomes of 18 Conus species.

<table>
<thead>
<tr>
<th>Conus species</th>
<th>RACE</th>
<th>cDNA PCR</th>
<th>Genome PCR</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLKM</td>
<td>MMSK</td>
<td>MLKM</td>
<td>MMSK</td>
</tr>
<tr>
<td>striatus</td>
<td>5</td>
<td>7</td>
<td>20</td>
<td>25</td>
</tr>
<tr>
<td>textile</td>
<td>6</td>
<td>7</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>betulinus</td>
<td>6</td>
<td>7</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>capitaneus</td>
<td>10</td>
<td>0</td>
<td>23</td>
<td>10</td>
</tr>
<tr>
<td>caracteristicus</td>
<td>10</td>
<td>3</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>coronatus</td>
<td>7</td>
<td>3</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>ebraeus</td>
<td>17</td>
<td>2</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>enaciatius</td>
<td>4</td>
<td>3</td>
<td>11</td>
<td>23</td>
</tr>
<tr>
<td>litteratus</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>lividus</td>
<td>7</td>
<td>2</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td>miles</td>
<td>3</td>
<td>3</td>
<td>9</td>
<td>17</td>
</tr>
<tr>
<td>quercinus</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>rattus</td>
<td>3</td>
<td>3</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>terebra</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>tessulatus</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>varius</td>
<td>8</td>
<td>2</td>
<td>13</td>
<td>25</td>
</tr>
<tr>
<td>vexillum</td>
<td>0</td>
<td>2</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>planorbis</td>
<td>2</td>
<td>3</td>
<td>13</td>
<td>21</td>
</tr>
</tbody>
</table>

Identification of the M-superfamily conotoxins was performed using ConoServer.
mollusc-hunting and polychaete-hunting Conus species all arose from a common ancestor.

### 3.2. Ten of the 18 cysteine frameworks may be generated from framework III

Eighteen cysteine frameworks were identified from the M-superfamily conotoxins (Fig. 1). We designated the frameworks 0C, 1C, 2C and 3C to distinguish cysteine patterns of non-disulfide-rich conotoxins. We defined the cysteine pattern C-CC-C-C-C as framework XXIII referring to the latest reports (Elliger et al., 2010; Moller and Mari, 2010). We designated the variants of framework III (CC-C-C-CC-C, CC-C-C-CC-C, CC-C-C-CC-C, and CCCC-CC-C-CC) as III-1, III-2, III-3, III-4, III-5, and III-6. Six cysteine frameworks were first reported in Conus: 3C, III-2, III-3, III-4, III-5, and III-6. One cysteine framework was first reported in the M-superfamily: XIV. Cysteine numbers of six cysteine frameworks are odd: 1C, 3C, III-1, III-2, III-4, and III-6. These unpaired cysteine conotoxins may form disulfide-crosslinked dimers similar to Vt3.1 and Vt3.2 (Wu et al., 2010), or a dimer of dimers, such as the conotoxin con-ikot-ikot with thirteen cysteines (Walker et al., 2009).

According to the sequence alignments (Fig. 1A), we mapped the potential evolution of the M-superfamily cysteine patterns (Fig. 1B). Ten cysteine frameworks (3C, II, IV, XIV, III-1, III-2, III-3, III-4, III-5, and III-6) may be generated from framework III by one to several nucleotides substitution or indels. Because two of them (3C, III-6) are only found in the genome, these two frameworks may be from pseudogenes. Variation mechanisms for the other seven frameworks that are changed beyond recognition from framework III are still unknown. It is noteworthy that framework 2C conotoxins show homology with framework III conotoxins in the signal and pro-region, while the other six of the seven framework type conotoxins show

---

### Table A

<table>
<thead>
<tr>
<th>Framework Type</th>
<th>Cysteine Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>3C</td>
<td>C-W-E-S-P-C-H-G-W-V-P-C</td>
</tr>
<tr>
<td>III</td>
<td>C-C-D-L-P-C-N-A-G-C-V-P-C</td>
</tr>
<tr>
<td>III-1</td>
<td>C-R-C-E-Q-T-G-C-T-V-P-C</td>
</tr>
<tr>
<td>III-2</td>
<td>C-R-C-E-Q-S-T-I-C-T-M-P-C</td>
</tr>
<tr>
<td>III-3</td>
<td>C-R-C-E-Q-S-T-I-C-T-Q-R-C-Q-R</td>
</tr>
<tr>
<td>III-4</td>
<td>S-C-C-A-E-G-M-C-H-S-G-C-E-C-C-C</td>
</tr>
<tr>
<td>III-6</td>
<td>C-C-C-F-V-C-C-H-Q-L-C-P-E-G</td>
</tr>
<tr>
<td>III-7</td>
<td>V-C-C-P-N-G-G-C-H-E-L-C-Q-E-G</td>
</tr>
</tbody>
</table>

### Diagram

![Diversity of the M-superfamily cysteine patterns.](image)

**Fig. 1.** Diversity of the M-superfamily cysteine patterns. (A) Alignments of Framework III conotoxins with other framework conotoxins. C represents cysteine and is marked in red. (B) Evolution of the M-superfamily cysteine patterns. C represents cysteine; the dash between C represents other amino acids. Roman numerals, Arabic numbers and Capital letters C in the brackets represent cysteine framework types. C and dash, which evolved from cysteine pattern III, are marked in black, and all others are marked in red. The top eight cysteine patterns indicated by solid arrowheads may be all generated from the middle cysteine pattern III. The bottom left cysteine patterns indicated by dash arrowheads are only found in the genome. The evolutionary mechanism of the remaining cysteine patterns indicated by blue arrowheads is still unknown. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)
homology with framework III conotoxins only in the signal region. We propose that recombination (Espiritu et al., 2001), premature stop (Remigio and Duda, 2008) or focal mutation are involved in this genetic event.

3.3. M-superfamily conotoxins consist of two introns and three exons

The gene structures of three MLKM group conotoxins and three MMSK group conotoxins from C. terebra, C. textile and C. striatus were analysed by genome walking. The corresponding cDNA sequences were used to determine the exon/intron boundaries, by following the GT-AG rule of splice-donor and splice-acceptor sites (Shepard et al., 2009). Genes of six conotoxins have identical architecture, with two introns and three exons, similar to the O1 conotoxin genes (Schoenfeld, 1999). The 5'UTR and 3'UTR are segregated from the exons, therefore, we assigned Exon1, Exon2 and Exon3 to correspond to the signal region, pro-region and mature region respectively, with several nucleotide differences at the boundaries. Alternative splicing was detected in two conotoxin transcripts, which were composed of Exon1 and Exon3, without Exon2.

The 3'-ends of Exon1 are interrupted after the first nucleotide of the last codon (phase1-intron), while the 3'-ends of Exon2 has its 3'-ends codon interrupted after the second nucleotide (phase2-intron) (Radis-Baptista et al., 2003). We only obtained the beginning and ending segments of the introns, from which we inferred the lengths of two introns to be more than 3000 bps. Introns show no homology between the MLKM and MMSK groups, while introns within the MLKM group or the MMSK group are conserved at the beginning and ending parts. This result is consistent with those obtained from O1 and A superfamilies (Olivera et al., 1999; Schoenfeld, 1999; Yuan et al., 2007).

3.4. Phylogeny of M-superfamily conotoxins and diet types

Because no sequence homology was detected between the MLKM group conotoxins and the MMSK group conotoxins, two unrooted Bayesian trees were constructed using prepropeptides of the MLKM group conotoxins and prepropeptides of the MMSK group conotoxins, respectively. Prepropeptides of M-superfamily conotoxins consist of Exon1, Exon2 and Exon3 based on their gene structure, so we manually checked all the branches of the phylogenetic tree to see whether the three exons are homologous. According to posterior probabilities and Exon3 homology, the MLKM group conotoxins are divided into 11 clades (MLKM-A, B, C, D, E, F, G, H, I, K clade in Fig. 2) and the MMSK group conotoxins are divided into 12 clades (MMSK-A, B, C, D, E, F, G, H, I, J, K clade in Fig. 2). Posterior probabilities (PP) of all the clades are >0.9 expect the two MMSK clades A and L. The MMSK-A (PP = 0.71) consist of more than ten robust clades (PP > 0.9), Exon3 of these clade conotoxins are homologous and different from other MMSK clade conotoxins, so we merged these clades into MMSK-A for phylogenetic analysis. The MMSK-L clade (PP = 0.67) was also retained for phylogenetic analysis because it is close to the clades H, I, J and K which contain conotoxins with definite pharmacological target. Then we focus on the relationship between diet types and phylogeny of the M-superfamily conotoxins. The diets of all Conus species were identified by ConoServer and the literature (Duda and Kohn, 2005; Duda et al., 2001; Kaas et al., 2008). The 61 Conus species of this article were classified into five diet types: mollusc-hunting (M), errant polychaete-hunting (E), sedentary polychaete-hunting (S), hemichordate-hunting (H) and fish-hunting (P). The diet types of the Conus species from which the clade conotoxins were identified were marked in red in Fig. 2.

From Fig. 2, we found that mollusc-hunting Conus contain conotoxins from three clades: MLKM-A, MMSK-A
and MMSK-J. Dietary-specific conotoxin clade is detected in polychaete-hunting Conus: MLKM-K and MMSK-B are monopolized by errant polychaete-hunting Conus, and MMSK-I is monopolized by sedentary polychaete-hunting Conus. Hemichordate-hunting Conus do not contain conotoxins from the common clade MLKM-A and MMSK-A of which other four diet types contain conotoxins, but contains conotoxins from four clades: MLKM-D, MLKM-H, MMSK-C and MMSK-L, which is less than other Conus, suggesting that diet shifts to hemichordate are associated with a reduction of conotoxins. This is coincident with a previous report on hemichordate-hunting Conus leopards (Remigio and Duda, 2008). Fish-hunting Conus monopolize two piscivorous-specific clades: MMSK-H and MMSK-K. In addition, fish-hunting Conus contain lots of conotoxins similar to molluscivorous and vermivorous conotoxins. One explanation is that fish-hunting Conus secrete these conotoxins to defend against mollusc-hunting Conus; another explanation is that fish-hunting Conus consume polychaetes during their juvenile stage (Duda and Palumbi, 2004); the third explanation is that the fish-hunting Conus, mollusc-hunting Conus and polychaete-hunting Conus all retain common conotoxins from their ancestral Conus species.

Besides the dietary-specific conotoxin clade, each Conus species seems to have its distinct conotoxin clade (Table 2). A typical species-specific conotoxin clade is MLKM-K, which is monopolized by C. varius. According to the relationship between diet types and conotoxin clades, we presumed that the M-superfamily conotoxins did not evolve in a concerted manner but were subject to birth- and-death evolution. In birth-and-death evolution, gene duplication generates new genes. Some duplicated genes are maintained in the genome and may evolve new functions, while others are inactivated to pseudogenes or deleted (Nei and Rooney, 2005). The birth-and-death evolution model may elucidate the phenomenon that different Conus species or Conus with different diet types have distinct composition of M-superfamily conotoxins.

3.5. Position-specific codon conservation is not restricted to cysteine, but also to other conserved residues

It has been observed that codons of specific cysteines are highly conserved according to previous literature (Conticello et al., 2000; Dewan, 2006). In this article, cysteine codon usage (TGC or TGT) analysis of M-superfamily conotoxins reveals codon conservation in specific cysteine positions (Fig. 3A). In order to find whether codon conservation exists on other residues except cysteines, codon usages of other conserved residues were analysed. By scanning all the M-superfamily conotoxins, we found an amino acid that is second conserved to cysteine in the M-2 branch conotoxins of the MMSK group. More than 90% of the 204 M-2 branch conotoxins of MMSK group possess proline which was located next to the fifth cysteine (Fig. 3B). Compared to codon usage in all Pro sites of M-2 branch conotoxins of the MMSK group, the conserved Pro located between the fourth and fifth cysteine has a preference to use codon CCT, and its codon usage is more conservative than the adjacent fourth and fifth cysteines (Fig. 3A), which indicates that position-specific codon conservation is not restricted to cysteines, but also to other conserved residues. Position-specific codon conservation was also found in other conotoxin clades containing several to forty conotoxins. For example, the codon usage of proline and glycine of the eight conotoxins from clade K of MLKM group is more conservative than the sixth cysteine (Fig. 3C). The phenomenon that position-specific codon conservation is not restricted to cysteines is also found in scorpion toxin (Kozminsky-Atias et al., 2008). Because position-specific codon conservation is not only restricted to cysteines, we may invalidate the hypothesis that protecting molecules bound to cysteine codons to protect them from

<table>
<thead>
<tr>
<th>Diet type</th>
<th>Province</th>
<th>Conus species</th>
<th>MLKM</th>
<th>MMSK</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mollusc (M)</td>
<td>IP</td>
<td>Conus varius (Vr)</td>
<td>A, G, I, K</td>
<td>A, C, D, E, G</td>
</tr>
<tr>
<td>Errant polychaete (E)</td>
<td>IP</td>
<td>Conus planorbid (Pl)</td>
<td>A, B, C, G, I</td>
<td>A, D, E</td>
</tr>
<tr>
<td>Sedentary polychaete (S)</td>
<td>IP</td>
<td>Conus capitanus (Cp)</td>
<td>A, D, F, G, H, I, J</td>
<td>A, B, E, G</td>
</tr>
<tr>
<td>Hemichordate (H)</td>
<td>IP</td>
<td>Conus miles (Mi)</td>
<td>A, C, G</td>
<td>A, C, D, C, G</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus coronatus (Co)</td>
<td>F, J</td>
<td>A, D, G</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus ebreaus (Eb)</td>
<td>A, J</td>
<td>A, D</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus betulinus (Bt)</td>
<td>C, G, H</td>
<td>D, E, I</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus caracteristicus (Ca)</td>
<td>A, E</td>
<td>A, D, F, G, L</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus emaciatu (Ec)</td>
<td>B, D</td>
<td>A, C, D, C, I, L</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus literatus (Lr)</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus terebra (Tr)</td>
<td>A, B</td>
<td>A, D, E</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus tessulatus (Ts)</td>
<td>I</td>
<td>A, D, I</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus lividus (Lv)</td>
<td>D</td>
<td>C, L</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus quercinus (Qc)</td>
<td>D</td>
<td>C, L</td>
</tr>
<tr>
<td></td>
<td>IP</td>
<td>Conus striatus (S)</td>
<td>A, C, G, H</td>
<td>A, D, E, H, K</td>
</tr>
</tbody>
</table>

IP, Indo-Pacific; EP, eastern Pacific. Abbreviations of Conus species were in the bracket. The diets of all Conus species were identified by ConoServer and the literatures (Duda and Kohn, 2005; Duda et al., 2001; Kaas et al., 2008).
mutation (Conticello et al., 2000), or that mutator complexes, such as DNA-Pol-V-like enzyme, target cysteine codons to induce mutagenic processes (Conticello et al., 2001). We prefer the hypothesis that secondary structure formations of conotoxin genes may play an important role in mediating variability (Dewan, 2006). According to this hypothesis, conserved residue codons form highly stabilized secondary structures to resist mutation, which rationally explain that not only cysteines but also other conserved residues maintain strict codon conservation.

3.6. Indels are involved in the creation of new functions and structures of the M-superfamily conotoxins

A great deal of insertions and deletions (indels) are found in M-superfamily conotoxins according to DNA sequence alignment. Single nucleotide indels that cause frame-shift are found in several transcripts. A 63-nucleotide insertion is found in Exon2 of the MMSK group. Plenty of triple nucleotide indels are found, especially in Exon3. Because six cysteines are under purifying selection or silent mutation in
framework III conotoxins, the numbers of amino acids existing in the three cysteine loops can be a good indication of indels. CC(X2-4)C(X2-4)C(X1-4)CC of the MLKM group and CC(X1-9)C(X1-9)C(X1-5)CC of the MMSK group suggest that indels are general genetic mechanisms in the evolution of conotoxin genes (Fig. 4). One representative example is that the number of residues in the first cysteine loop of sodium channel-blocking conotoxins varies from 1 to 9, while the number of residues in the third cysteine loop is 4 or 5 (Fig. 4B). This indicates that indels are more active in the first cysteine loop. This region is not essential for structure and function because deletion of the first disulfide bond only slightly affects KIIIA bioactivity (Han et al., 2009); thus, we propose that the function-unimportant region is associated with the relaxation of selective constraints on indels rates in conotoxin genes. By contrast, alteration of residue number in the third cysteine loop changes the disulfide bond connectivity patterns, three-dimensional structures and physiological functions of M-superfamily conotoxins (Han et al., 2006; Jacob and McDougal, 2010). Therefore, indels, especially indels in the third cysteine loop, may play an important role in creating new functions and structures for duplicated genes.

One proposed mechanism is that indels induce nucleotide mutations in the surrounding DNA according to indels-associated mutations (Tian et al., 2008). The rates of indels in conotoxins are in the following order: Exon3 > Exon2 > Exon1, which is the same as the nucleotide substitution rates. A further analysis on relationship between nucleotide divergence and the distance to indels is needed to determine whether indels induce nucleotide mutations in conotoxin genes.

4. Conclusions

In this study, we characterized identical sequences, gene structure, novel cysteine frameworks, functions and evolutionary mechanisms of M-superfamily conotoxins, and six novel features were found for the M-superfamily conotoxins: 1. Identical M-superfamily conotoxins can be found in different Conus species. 2. M-superfamily conotoxins consist of two introns and three exons. 3. Ten of the 18 cysteine frameworks may be generated from framework III. 4. Position-specific codon conservation is not restricted to cysteines, but also to other conserved residues. 5. M-superfamily conotoxins may not evolve in a concerted manner but were subject to birth-and-death evolution. 6. Indels are involved in the creation of new functions and structures of the M-superfamily conotoxins.

Acknowledgements

This work was supported by the State High-Tech Development Program (863 Program) from the Ministry of Science and Technology of China (No. 2006AA090504 and No. 2008AA09Z401), Science and Technology Support Program (2012BAD117802), National Natural Science Foundation of China (30871921), Project of Guangdong Science-Tech Program (No. 2006A36501001), Guangdong Natural Science Foundation (815102750100083).

Conflict of interest statement

The authors declare that there are no conflicts of interest.

References


