



**HAL**  
open science

# Molecular Phylogeny, Classification and Evolution of Conopeptides

N. Puillandre, D. Koua, P. Favreau, B. Olivera, R. Stöcklin

► **To cite this version:**

N. Puillandre, D. Koua, P. Favreau, B. Olivera, R. Stöcklin. Molecular Phylogeny, Classification and Evolution of Conopeptides. *Journal of Molecular Evolution*, 2012, 74 (5-6), pp.297-309. hal-02002427

**HAL Id: hal-02002427**

**<https://hal.science/hal-02002427>**

Submitted on 31 Jan 2019

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## Molecular phylogeny, classification and evolution of conopeptides

Journal:	<i>Journal Of Molecular Evolution</i>
Manuscript ID:	JME-2012-0012.R1
Manuscript Type:	Manuscript
Date Submitted by the Author:	n/a
Complete List of Authors:	Puillandre, Nicolas; Atheris Laboratories, Koua, Dominique; Atheris Laboratories, ; Swiss Institute of Bioinformatics, Favreau, Philippe; Atheris Laboratories, Olivera, Baldomero; University of Utah, Biology Stöcklin, Reto; Atheris Laboratories,
Key Words:	cone snails, Conus, Conoidea, Cys-pattern, venom, molecular evolution

1  
2  
3 **Molecular phylogeny, classification and evolution of conopeptides**  
4  
5  
6

7  
8 N. Puillandre<sup>1,2</sup>, D. Koua<sup>1,3</sup>, P. Favreau<sup>1</sup>, B. M. Olivera<sup>2</sup>, R. Stöcklin<sup>1</sup>.  
9

10  
11  
12 <sup>1</sup> Atheris Laboratories, Case postale 314, CH-1233 Bernex-Geneva, Switzerland  
13

14  
15 <sup>2</sup> Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112.  
16

17  
18 <sup>3</sup> Swiss Institute of Bioinformatics, CH-1206 Geneva, Switzerland  
19

20  
21  
22 Corresponding Author: Nicolas Puillandre, Atheris Laboratories, Case postale 314, CH-1233 Bernex-  
23  
24 Geneva, Switzerland; nicolas.puillandre@atheris.ch  
25  
26  
27  
28  
29  
30  
31  
32  
33

34 **Running title:** Phylogenetic classification of conopeptides  
35  
36  
37  
38

39 **Keywords:** cone snails, *Conus*, Conoidea, Cys-pattern, venom, molecular evolution.  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Abstract**

Conopeptides are toxins expressed in the venom duct of cone snails (*Conoidea*, *Conus*). These are mostly well-structured peptides and mini-proteins with high potency and selectivity for a broad range of cellular targets. In view of these properties, they are widely used as pharmacological tools and many are candidates for innovative drugs. The conopeptides are primarily classified into superfamilies according to their peptide signal sequence, a classification that is thought to reflect the evolution of the multigenic system. However, this hypothesis has never been thoroughly tested. Here we present a phylogenetic analysis of 1,364 conopeptide signal sequences extracted from GenBank. The results validate the current conopeptide superfamily classification, but also reveal several important new features. The so-called “cysteine-poor” conopeptides are revealed to be closely related to “cysteine-rich” conopeptides; with some of them sharing very similar signal sequences, suggesting that a distinction based on cysteine content and configuration is not phylogenetically relevant and does not reflect the evolutionary history of conopeptides. A given cysteine pattern or pharmacological activity can be found across different superfamilies. Furthermore, a few conopeptides from GenBank do not cluster in any of the known superfamilies, and could represent yet-undefined superfamilies. A clear phylogenetically-based classification should help to disentangle the diversity of conopeptides, and could also serve as a rationale to understand the evolution of the toxins in the numerous other species of conoideans and venomous animals at large.

## Introduction

Cone snails of the genus *Conus* are predatory venomous marine mollusks feeding on fish, worm or snails. After decades of biological prospecting, conopeptides expressed in their venom duct have emerged as one of the richest and most promising marine sources of natural products (Blunt et al. 2012). The analysis of cone snail venoms has revealed a complex exogenome that is characterized by an extremely high level of diversity. With more than 600 described *Conus* species, each producing an estimated 100-200 venom components, the ensemble of cone snails were, until recently, estimated to produce between 50,000 and 100,000 different toxins (Menez et al. 2006; Olivera 2006). Recent studies, however, clearly demonstrate that this figure is an underestimation, probably by a factor of ten or so, with several new species described every year, more venom components detected in each sample using evolving technologies such as mass spectrometry (Biass et al. 2009; Ueberheide et al. 2009; unpublished results) and NextGen sequencing (Hu et al. 2011; Terrat et al. 2011) or combinations thereof (Violette et al. 2012), and marked intra-species and even intra-specimen variations in venom composition (Davis et al. 2009; Dutertre et al. 2010; Jakubowski et al. 2005). It is now estimated that the number of cone snail venom components exceeds one million.

An important characteristic of conopeptides, which makes them attractive for drug development is their high selectivity for molecular targets that span a broad range of therapeutic applications (Gayler et al. 2005; Leary et al. 2009; Molinski et al. 2009). So far, the conopeptide MVIIA (SNX-111, Prialt, or Ziconotide) from *Conus magus* (the magician cone) that selectively blocks Cav2.2 N-type voltage-gated calcium channels has been approved for the treatment of severe chronic pain (McGivern 2007; Miljanich 2004) and there are more promising drug candidates in the pipeline (*e.g.*, see Favreau et al. 2012; Han et al. 2008a; Lewis 2012). The potential of this rich source of pharmacological products has stimulated a race for the discovery of new toxins. From the traditional bioactivity-guided identification, lead discovery efforts have evolved towards modern

1  
2  
3 structure-driven characterization (venom peptidomics and proteomics, venom gland transcriptomics,  
4 targeted genomics, structure-function studies) and biocomputing-assisted analyses (proprietary  
5 databases and bioinformatic tools) (Daly and Craik 2009; Favreau and Stöcklin 2009; Koua et al.  
6 2012; Laht et al. 2011). In addition, phylogenetic approaches have recently emerged as an effective  
7 way to quickly identify divergent lineages that are likely to have evolved with different functional  
8 characteristics. This approach to identify these previously uncharacterized conopeptides is referred to  
9 as concerted discovery (Conticello et al. 2001; Duda and Remigio 2008; Olivera 2006; Puillandre and  
10 Holford 2010).

11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22 However, despite the effectiveness of phylogenetic approaches in concerted discovery, the  
23 technique is rarely used for the classification of conopeptides (but see Aguilar et al. 2009; Conticello  
24 et al. 2001; Wang et al. 2008; Zhangsun et al. 2006). Several statistical methods for conopeptide  
25 classification, such as Mahalanobis (Lin and Li 2007) or BLAST and Euclidian distances among  
26 others (Mondal et al. 2006) have been described; however, most of these approaches are primarily  
27 designed for classification of new sequences rather than for testing the current classification (*i.e.*,  
28 checking the validity of each known group by a blind-exploratory approach). Conopeptide precursors  
29 are characterized by a typical structural organization consisting of a highly conserved signal region,  
30 followed by a more variable pro-region and a hyper-variable mature toxin containing a few  
31 conserved amino acids such as the cysteine residues required for disulfide bonds. Conopeptides are  
32 mainly named and classified according to three properties: first, they are characterized by their signal  
33 sequence, this short sequence (~20 amino-acids) is highly conserved, and has been used to define  
34 superfamilies; second, mature toxins structural families are characterized depending on their pattern  
35 of cysteines (the Cys-pattern), for example, the mature toxin can include a variable number of  
36 cysteines (most commonly 4 or 6), and their respective position can vary (4 cysteines can be  
37 organized as C-C-C-C or CC-C-C where “-” represents a variable number of amino-acids); finally,  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 several conopeptides have also been characterized according to their molecular targets, referred to  
4 hereafter as “functional families”, and also previously termed “pharmacological families”.

5  
6  
7  
8 In a recent paper, Kaas *et al.* (2010) reviewed the structure, function and diversity of  
9  
10 conopeptides on the ConoServer database ([www.conoserver.org](http://www.conoserver.org)). In particular, they proposed that  
11  
12 “the ‘gene superfamily’ classification scheme focuses on evolutionary relationships between  
13  
14 conopeptides”, while the two other classification schemes (cysteine framework and function) do not.  
15  
16 Their underlying hypothesis was that similarities in the Cys-pattern or function might have arisen by  
17  
18 convergence. While we fully agree with this statement, we also argue that it could serve as a rationale  
19  
20 to assess the congruence between the current gene superfamily classification and the evolution of the  
21  
22 corresponding multigenic system, and to accurately demonstrate that convergence phenomena are  
23  
24 common in conopeptide structure and function.  
25  
26  
27  
28

29 Here we review the current superfamily classification of conopeptides by analysing all the  
30  
31 signal sequences available in GenBank using a phylogenetic approach to check: (i) if all the defined  
32  
33 superfamilies correspond to homogeneous groups; and (ii) if all the GenBank signal sequences  
34  
35 belong to a known superfamily. This study seeks to provide a "rationale" for a phylogenetic  
36  
37 classification of conopeptides and to clarify their current classification, thus complementing the work  
38  
39 initiated by Kaas *et al.* (2010).  
40  
41  
42  
43  
44  
45

## 46 **Materials and Methods**

### 47 *1. Sequences from GenBank*

48  
49 Since the signal sequences used for phylogenetic analyses (see below), are only found on  
50  
51 complete nucleotide precursors and are not known for conopeptide discovered using proteomic  
52  
53 approaches, all the nucleotide sequences associated with the genus *Conus* were downloaded from  
54  
55 GenBank ([www.ncbi.nlm.nih.gov](http://www.ncbi.nlm.nih.gov)). The sequences corresponding to non-coding regions, ribosomal  
56  
57  
58  
59  
60

1  
2  
3 genes, mitochondrial genes, and genes with a function that did not relate to toxin activity were  
4 removed from the dataset, thus keeping only coding genes with a potential toxin activity. Only  
5  
6 sequences obtained from *Conus* species belonging to the large major clade (Duda and Kohn 2005)  
7  
8 were conserved, as a large number of the conopeptides found in species from other clades (*e.g.*, *C.*  
9  
10 *californicus*) are highly divergent and do not match with any of the currently known superfamilies  
11  
12 (Biggs et al. 2010; [www.conoserver.org](http://www.conoserver.org)). Consequently, the classification in the present analysis is  
13  
14 relevant only for conopeptides of the large major clade species. Conopeptide superfamilies are  
15  
16 defined by a conserved signal sequence, thus we used the Signalp 3.0 server (Bendtsen et al. 2004) to  
17  
18 identify the signal sequence; all sequences that did not include at least 50% of the signal region were  
19  
20 removed, together with sequences including a stop codon. Only the signal region was used for  
21  
22 phylogenetic analyses, as only this part of the conopeptides can be aligned within and, to some  
23  
24 extent, between superfamilies.  
25  
26  
27  
28  
29  
30  
31  
32  
33

## 34 2. Phylogenetic analysis

35  
36 Aligning signal sequences between highly divergent conopeptides (*i.e.*, belonging to different  
37  
38 superfamilies) is arduous, and homology hypotheses are doubtful. Thus, sequences were translated to  
39  
40 amino acids and automatically aligned using two different algorithms: Muscle (Edgar 2004  
41  
42 [www.ebi.ac.uk/Tools/msa/muscle](http://www.ebi.ac.uk/Tools/msa/muscle)) and ClustalW (<http://clustalw.ddbj.nig.ac.jp/top-e.html>). Best  
43  
44 model of evolution for these two datasets was selected using Modelgenerator V.85 (Keane et al.  
45  
46 2006) following the corrected Akaike Information Criterion (with four discrete gamma categories)  
47  
48 and used to reconstruct phylogenetic trees. The best model of evolution identified by Modelgenerator  
49  
50 was JTT+G (Jones Taylor Thornton model, implemented under the name “Jones model” in MrBayes  
51  
52 – Jones et al. 1992) for both datasets. Bayesian analyses were performed by running two parallel  
53  
54 analyses in MrBayes (Huelsenbeck et al. 2001), each consisting of eight Markov chains of  
55  
56  
57  
58  
59  
60



1  
2  
3 30,000,000 generations each with a sampling frequency of one tree every ten thousand generations.  
4  
5 The number of swaps was set to 5, and the chain temperature at 0.02. A Neighbor-Joining tree  
6  
7 obtained with MEGA5 (Tamura et al. 2011) was used as starting tree. Convergence of the parameters  
8  
9 was evaluated using Tracer 1.4.1 (Rambaut and Drummond 2007), and analyses were terminated  
10  
11 when ESS values were all superior to 200. A consensus tree was then calculated after omitting the  
12  
13 first 25% trees as burn-in.  
14  
15

16  
17 As is the case for most multigenic families, the identification of an outgroup was highly  
18  
19 problematic. No gene phylogenetically related to, and proven to be an outgroup for, conopeptides has  
20  
21 been described. Furthermore, the use of toxins from other conoidean species was not possible, as it  
22  
23 would require that the toxins from cone snails all arose from duplication events that took place after  
24  
25 the divergence between the cone snails and other conoideans. Consequently, no outgroup was  
26  
27 included in the analysis. This absence of an outgroup did not allow us to infer ancestor/descendant  
28  
29 relationships.  
30  
31  
32

## 33 34 35 36 **Results**

37  
38 A total of 1,364 sequences potentially corresponding to conopeptides and with a signal  
39  
40 sequence were downloaded from GenBank (performed on 1<sup>st</sup> of July, 2011). Alignments were 34 and  
41  
42 30 amino-acids long with Muscle and Clustal W, respectively. To limit the time of calculation for  
43  
44 phylogenetic analysis, only one sequence per amino-acid haplotype was kept; finally, 585 sequences  
45  
46 were retained. Overall, the phylogenetic trees obtained from the Muscle and Clustal alignments were  
47  
48 congruent; discrepancies were not supported (Posterior Probabilities < 0.90) and concerned  
49  
50 phylogenetic relationships between the main clades and the position of a few highly divergent  
51  
52 sequences (see details below). For clarity, only the phylogenetic tree based on the Clustal alignment  
53  
54 is presented (Fig. 1), but results obtained from the Muscle alignment, when different, are discussed.  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Using information from GenBank and the literature, it was possible to link the clades defined with the bayesian analysis to known superfamilies. Most of the defined superfamilies (A, D, I1, I2, I3, J, L, O1, O3, P, S, T, V) corresponded to monophyletic groups, with some highly supported (Fig. 1). With the Muscle alignment, the O2 superfamily was included within the O1 superfamily; the superfamily Y was represented by a single sequence, and corresponded to a unique lineage in the tree. However, some superfamilies did not correspond to a monophyletic group, as they included other conopeptides (*e.g.*, O2 included sequences of contryphans, and M included conomarphin – a result already discussed by Han *et al.* 2008b). Several conopeptides from GenBank did not cluster in any of the known superfamilies. These corresponded to known cysteine-poor conopeptides, contulakin and conantokin, shown in Fig. 1 as the B and C superfamilies, respectively (the C superfamily has been previously defined by Jimenez *et al.* (2007)); two conoCAP sequences (FN868446.1 and FN868447.1 – named X1 in the Fig. 1 and appendix 1) described by Möller *et al.* (2010); and sequences putatively annotated (FJ237364.1, named X2) or without annotation in GenBank (DQ359922.1, EF493183.1/EF493184.1 and DQ359921.1, named respectively X3, X4 and X5). In the Clustal alignment, two other groups of sequences, FJ375238.1/FJ375239.1/FJ375240.1 and EF208033.1 clustered in the superfamily A and O1, respectively with long branches, but corresponded to independent lineages in the Muscle alignment (X6 and X7, respectively).

Function and cysteine pattern were not clade-specific; conopeptides with the same function or cysteine pattern were found in different clades. Additionally, sixteen new (*i.e.*, not numbered with Roman numbers) cysteine patterns were identified; however, most of them certainly correspond to anecdotic mutations of the canonical framework in a given family (*i.e.*, C-CC-C-C, C-C-CC-C-CC, and C-CC-C-C, found in the O1 superfamily, differ from the pattern VI/VII by only one mutation), while others may represent a new Cys-pattern number (*e.g.*, the Cys-pattern C-C-C-CC-C, found in

1  
2  
3 the three members of the X6 group). The results are summarized in Table 1 (full details are provided  
4  
5 in Appendix 1).  
6  
7

8 Table 2 lists the number of conopeptides found in each superfamily and their distribution  
9  
10 among the 71 *Conus* species. The superfamilies A, M and O1 were the largest, each containing at  
11  
12 least 39 species, followed by the superfamilies T and I2. *Conus characteristicus*, *C. imperialis* and *C.*  
13  
14 *litteratus* each express conopeptides belonging to more than 10 different superfamilies in their  
15  
16 venom; however, it was difficult to know if this result reflects a higher conopeptide diversity in  
17  
18 comparison to other species, or is due to a greater sampling effort in these species. All the  
19  
20 superfamilies present in more than 10 *Conus* species (A, B, I2, M, O1, O2 and T) were found in  
21  
22 mollusk, worm and fish-hunting species.  
23  
24  
25  
26  
27  
28

## 29 Discussion

### 30 31 32 1. An updated classification of conopeptides

33  
34 Overall, the molecular phylogeny, based on more than 1,300 conopeptides signal sequences  
35  
36 extracted from GenBank, strongly supports the current superfamily classification based on phenetic  
37  
38 resemblances, as established in ConoServer. But, this relative congruency between phylogenetic and  
39  
40 phenetic classifications is not surprising given the relative conservation of the signal sequence within  
41  
42 superfamilies compared with between superfamilies, and the phylogenetic tree reflects these  
43  
44 differences. However, the phylogenetic approach also revealed several new features, the most  
45  
46 striking of which is the presence of deeply divergent lineages that, until now, were not included in the  
47  
48 conotoxin superfamily classification. There are two main explanations for this result. First, the  
49  
50 conopeptide superfamily classification reviewed by Kaas *et al.* (2010) includes only what is  
51  
52 traditionally referred to as “cysteine-rich” conotoxins (*i.e.*, conopeptides with at least two disulfide  
53  
54 bridges in the mature sequence as defined by Norton and Olivera in 2006), thus excluding the  
55  
56  
57  
58  
59  
60

1  
2  
3 conopeptides with two cysteines and linear conopeptides also broadly present in the venom  
4 (unpublished results). However, although the authors noted that “in future, all disulfide-poor  
5 conopeptides will probably have to be attributed to a superfamily”, they refrained from doing so  
6 because of the low number of cysteine-poor conopeptides with precursor sequences in ConoServer  
7 (21). In GenBank we identified more than 50 such sequences and included them in the current  
8 analysis. The signal sequences of cysteine-poor conopeptides do not cluster separately from the  
9 conotoxins; some of them share highly similar signals with know superfamilies (contryphan with O2  
10 and conomarphin with M), therefore, their exclusion from the superfamily classification is not  
11 phylogenetically justified. We identified two additional superfamilies, B and C, for conantokins and  
12 contulakins, respectively, one of which (C) has been proposed previously (Jimenez et al. 2007).  
13  
14 Second, including non-annotated sequences from GenBank in the dataset helped to identify several  
15 independent lineages in the tree (X1–X7). The level of divergence of their respective signal  
16 sequences with the signals of other superfamilies was equivalent to the level of divergence between  
17 known superfamilies, and they thus deserve recognition as new superfamilies. However, as these  
18 independent lineages are represented by only one, two or three sequences, and because some of them  
19 may not exhibit toxin activity (even if they were all found in venom ducts of cone snails), we  
20 refrained from proposing new superfamily names, and only provided temporary names (X1-X7). It  
21 should also be borne in mind that many other conopeptides have been described in the literature,  
22 some of which have been given formal names (conkunitzin, conolysin, conomap, conophysin,  
23 conopressin, conorfamide and conorphan). Because their signal sequences are not represented as  
24 nucleotides in GenBank, they were not included in the analysis. However, a search in the protein  
25 database of GenBank retrieved two complete precursors of Conkunitzin, with highly similar signal  
26 sequences (P0C1X2.1 and P0CY85.1) and a local BLAST search (performed using BioEdit – Hall  
27 1999) of the dataset used for the phylogenetic analyses revealed that the conkunitzin signals were  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 unique, and probably represent a new superfamily. Finally, if most of the superfamily-level clades are  
4  
5 highly supported, most of the inter-superfamily nodes are not, preventing any reliable conclusion  
6  
7 concerning the phylogenetic relationships at this level.  
8  
9

10 The original results presented herein raise several issues concerning the classification and  
11  
12 nomenclature of the conopeptides and, more generally, of the genes that belong to multigenic  
13  
14 systems. The updated classification system we propose is based on a phylogenetic reconstruction that  
15  
16 guarantees the identification of sequences clusters that share a common ancestor. However, such  
17  
18 phylogenetic trees cannot help in deciding which clades deserve a superfamily-level ranking and  
19  
20 which ones do not. One common solution is to rely on a threshold of genetic distances, but the  
21  
22 analyses of the genetic distances (calculated as the number of differences) between all the  
23  
24 conopeptide signal sequences revealed that the distribution of genetic distances within superfamilies  
25  
26 of conopeptides largely overlaps with the distribution of genetic distances between superfamilies  
27  
28 (Fig. 2). This overlap can be linked to the high level of homoplasy found in conopeptides, making  
29  
30 two conopeptides from different clades having, by chance, a relatively low genetic distance, or to the  
31  
32 fact that two previously defined superfamilies would actually correspond to only one. This is the case  
33  
34 of the L and I3 superfamilies, separated by genetic distances comprised between 0.38 and 0.69 that  
35  
36 would, in most cases, correspond to within superfamily genetic distances.  
37  
38  
39  
40  
41  
42

43 Consequently, it is not possible to rely only on a genetic threshold to define superfamilies for  
44  
45 conotoxins. A threshold of 0.6, roughly corresponding to the gap between the two distributions of  
46  
47 genetic distances (Fig. 2), would lead to the division of the M-superfamily into numerous  
48  
49 superfamilies (indeed, Wang *et al.* 2008 proposed to divide the M superfamily in M1 and M2), and to  
50  
51 the grouping of the superfamilies I1, I3 and L in a single one. However, our approach is aimed at  
52  
53 offering a complementary guidance to help, in the future, deciding if a conotoxin or a group of  
54  
55 conotoxins deserve a superfamily name: (i) since the minimum genetic distance between conotoxins  
56  
57  
58  
59  
60

1  
2  
3 is 0.32, this distance should be the minimum distance between the potential new conotoxin(s) and all  
4  
5 the others; (ii) the new conotoxin(s) should correspond to an independent lineage, *i.e.* it should not  
6  
7 cluster in any of the superfamily clades previously defined; (iii) the molecular target of the new  
8  
9 conotoxin(s) should ideally be identified, to avoid naming conopeptides that would not be functional;  
10  
11 (iv) the structure (cysteine pattern) and/or function should be different from the most closely related  
12  
13 conotoxins in terms of genetic distances and/or phylogenetic relationships. All these criteria apply to  
14  
15 the B and C superfamilies (genetic distances with other superfamilies > 0.3, these two lineages are  
16  
17 independent and monophyletic, their molecular targets are identified – Mena et al. 1990, Craig et al.  
18  
19 1999 –, and their cysteine framework are different from their respective sister-groups), justifying the  
20  
21 attribution of new superfamily names. We followed the traditional nomenclature of conopeptide  
22  
23 superfamilies, *i.e.* a Roman capital letter. As the number of Roman letter is limited, some  
24  
25 superfamilies have been named with a Roman letter followed by an Arabic number (*e.g.* I1, I2, I3,  
26  
27 O1, O2, O3) when several superfamilies share a common cysteine framework or molecular target.  
28  
29 Because of the potentially high number of unknown superfamilies of conopeptides, we have no doubt  
30  
31 that the nomenclature based on both Roman letters and Arabic numbers will become the reference  
32  
33 rule.  
34  
35

36  
37  
38  
39  
40  
41 The first and fourth criteria also apply to the seven “X” lineages (Fig. 1), but the second  
42  
43 applies to only 5 of them (two clustered within the A and O1 superfamilies with the muscle  
44  
45 alignment) and the third to none of them. We propose to name such potential superfamilies of  
46  
47 conopeptides that currently do not meet all the criteria but could in the future with the X Roman  
48  
49 letter, followed by an Arabic number, waiting for either to be fully recognized as a separate  
50  
51 superfamily or as belonging to an existing one.  
52  
53  
54  
55  
56

## 57 2. Evolution of the conopeptides

58  
59  
60

1  
2  
3 The phylogenetic analysis clearly confirms that most of the defined superfamilies include  
4 conopeptides with different cysteine frameworks and functions. Conversely, similar cysteine  
5 frameworks and functions are found in different superfamilies, suggesting that a given cysteine  
6 framework or function can appear several times independently, probably as a result of convergent  
7 evolution. The multiple apparitions of the same framework and function during conotoxin evolution  
8 are probably linked to the extremely rapid diversification of the genes. Several molecular  
9 mechanisms have been proposed as being responsible for this high rate of diversification. Pi *et al.*  
10 (2006) suggested that alternative splicing, unequal crossing-over or exon shuffling could explain this  
11 diversity. Olivera *et al.* (1999) proposed two other mechanisms: the lack of a mismatch repair  
12 system, at least in the hypervariable part of the sequence (the mature toxin); and recombination  
13 mechanisms. Several other hypotheses, such as a high rate of duplication, followed by a strong  
14 diversifying selection on the newly created gene copies that could lead to the rapid appearance of  
15 several structurally and functionally highly divergent genes, have been also proposed and tested by  
16 different authors (Duda and Palumbi 1999; Conticello et al. 2000; Duda and Palumbi 2000;  
17 Conticello et al. 2001; Espiritu et al. 2001; Duda and Remigio 2008; Chang and Duda 2012). All  
18 these molecular mechanisms, together with observed differences in the expression pattern between  
19 species, maybe linked to episodes of gene silencing and reactivation (“Lazarotoxins”, Conticello et  
20 al. 2001; Duda and Palumbi 2004; Duda 2008), could favor the rapid diversification of *Conus*  
21 species, by allowing them to envenomate and feed on new prey and thus colonize new niches (Duda  
22 and Lee 2009).

23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
A phylogenetic approach could be very useful to identify divergent conopeptides with  
potentially different functions, even if they share a common structural framework. For example, the  
cysteine framework IV, found in the A-superfamily, is already linked to two different functions ( $\alpha$ A -  
Hopkins et al. 1995 and  $\kappa$ A - Craig et al. 1998). However, conotoxins, described by Conticello *et al.*

1  
2  
3 in 2001, with the same framework, belong to the M superfamily, suggesting that these IV-conotoxins  
4 that are structurally convergent with the IV-conotoxins in a different superfamily, could exhibit a  
5 completely different function. A similar strategy could also apply within each superfamily, where not  
6 only the signal sequence, but also the propeptide and mature regions can be aligned, and could reveal  
7 divergent lineages with as yet uncharacterized functions (*e.g.*, see Aguilar et al. 2009; Puillandre et  
8 al. 2010; Wang et al. 2008; Zhangsun et al. 2006).

9  
10  
11  
12  
13  
14  
15  
16  
17  
18 Furthermore, our identification of numerous new cysteine frameworks among the GenBank  
19 sequences was also surprising. Even if some of them may be non-functional genes (pseudogenes),  
20 others could correspond to novel protein structures. A few publications demonstrated that even toxins  
21 with odd numbers of cysteines can be functional, for example with two 5-Cys toxins forming a  
22 functional dimer or bioactive polymers of the 13-Cys “Con-ikot-ikot” peptide from *Conus striatus*  
23 (Quinton et al. 2009, Walker et al. 2009). Our findings challenge the traditional view where  
24 conotoxins are characterized by a limited number of cysteine frameworks: by exploring new  
25 evolutionary pathways, the apparition of novel cysteine frameworks may also participate in the  
26 hyper-diversification of the conotoxins. Additionally, this raises the question of the total number of  
27 cysteine patterns one could expect to find among cone snail toxins. It is possible to predict the  
28 theoretic number of cysteine patterns that could exist. If we limit the exercise to the 2, 4 and 6  
29 cysteine patterns and exclude those with more than two consecutive cysteines, 20 different  
30 frameworks can be proposed (C-C\*, CC, CC-C-C\*, CC-CC\*, C-CC-C, C-C-CC\*, C-C-C-C\*, CC-  
31 CC-CC, CC-CC-C-C, CC-C-CC-C, CC-C-C-CC\*, CC-C-C-C-C\*, C-CC-CC-C, C-CC-C-CC, C-CC-  
32 C-C-C\*, C-C-CC-CC, C-C-CC-C-C\*, C-C-C-CC-C\*, C-C-C-C-CC, C-C-C-C-C-C\*). Ten of these  
33 frameworks (marked with an \*) can be found in GenBank. Given the extreme capacity of the  
34 conopeptides to evolve and the apparent lack of evolutionary constraints (as illustrated by the  
35 multiple apparitions of identical frameworks during their evolution), there is no reason that all these  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 theoretical patterns will not be found in the future. It could be argued that mechanical constraints  
4 would prevent the existence of some cysteine patterns; for example, it could be unfavorable to have a  
5 disulfide bridge between two adjacent cysteines. However, despite this we found a short mature toxin  
6 in the venom of one cone snail with a disulfide bridge between adjacent cysteines (unpublished  
7 results). The peptide has been reproduced by protein synthesis, confirming this finding.  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17

### 18 3. *Conus* and *Conoidea* toxin diversity

19  
20 The diversity of conotoxins in the venom of several *Conus* species (Table 2) confirms that most  
21 species are able to express a variety of conotoxins, as widely reported in literature (e.g., Olivera  
22 2002). Furthermore, our results also suggest that *Conus* diet (fish, mollusk and worm) is not  
23 correlated with differences in venom composition at the superfamily level. If differences exist, as  
24 suggested in the literature (e.g., Conticello et al. 2001; Kaas et al. 2010), they most likely occur at the  
25 species and intra-superfamily levels. Furthermore, phylogenetic analyses suggest that, at least, the  
26 worm- and fish-hunting species are not monophyletic, as these two diets appeared independently  
27 several times during the *Conus* evolution (Duda and Palumbi 2004; Espiritu et al. 2001; Kraus et al.  
28 2011). Thus, differences in the venom composition should not be sought between the three diet  
29 groups, but between the monophyletic clades defined within these three groups (Duda and Palumbi  
30 2004).  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44

45  
46 Diversity of the marine snail toxins is not limited to species included in the large major clade  
47 of *Conus*. Recent analyses in other conoidean taxa suggest that toxin hyperdiversity is not the  
48 privilege of the *Conus* large major clade. *C. californicus*, which is highly divergent from all the other  
49 *Conus* species (Duda and Kohn 2005), showed a high diversity of toxins in its venom and several of  
50 them were thought to correspond to new superfamilies (Biggs et al. 2010;  
51 [www.conoserver.org/?page=classification&type=genesuperfamilies](http://www.conoserver.org/?page=classification&type=genesuperfamilies)). To a lesser extent, species in  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 the small major clade of *Conus*, may also contain several novel conotoxins, as suggested by an  
4 original Cys-pattern (XIII) found in the species *C. delessertii* (Aguilar et al. 2005). In addition to the  
5 family Conidae, original toxins have already been reported in several other species of Conoidea, such  
6 as *Polystira albida* (Lopez-Vera et al. 2004; Rojas et al. 2008), *Gemmula periscelida* (Lopez-Vera et  
7 al. 2004), *G. speciosa*, *G. sogodensis*, *G. diomedea*, *G. kieneri* (Heralde et al. 2008), *Lophiotoma*  
8 *olangoensis* (Watkins et al. 2006), *Terebra subulata* (Imperial et al. 2003), *Hastula hectica* (Imperial  
9 et al. 2007) and *Crassispira cerithina* (Cabang et al. 2011). Furthermore, taxonomic surveys  
10 (Bouchet et al. 2009) and phylogenetic analyses (Puillandre et al. 2011) suggest that the superfamily  
11 Conoidea actually comprises a number of deeply divergent clades, whose species diversity is  
12 currently largely underestimated. Presently, around 4,500 species have been described, but the group  
13 is believed to include more than 10,000 species (Bouchet et al. 2009). Even if the venom apparatus  
14 has been lost in several lineages of Conoidea (e.g., Fedosov 2007, Fedosov and Kantor 2008; Holford  
15 et al. 2009; Medinskaya and Sysoev 2003), these findings suggest that the conotoxin diversity  
16 characterized so far represents only a small part. If the level of diversity across all conoidean species  
17 is similar to that found in those already investigated, the number of toxins produced by this single  
18 superfamily could be as high as ten millions.

### 43 **Acknowledgments**

44 We are grateful to the European Commission for financial support. This study has been performed as  
45 part of the CONCO cone snail genome project for health ([www.conco.eu](http://www.conco.eu)) within the 6th Framework  
46 Program (LIFESCIHEALTH-6 Integrated Project LSHB-CT-2007, contract number 037592). We are  
47 also grateful to Frédérique Lisacek from the Swiss Institute of Bioinformatics for ongoing help. We  
48 would like to thank Dr Ron Hogg of OmniScience SA for editorial support.

### Conflict of interest

The authors declare that they have no conflict of interest.

### References

- Aguilar MB, Chan de la Rosa RA, Falcon A, Olivera BM, Heimer de la Cotera EP (2009) Peptide pal9a from the venom of the turrid snail *Polystira albida* from the Gulf of Mexico: Purification, characterization, and comparison with P-conotoxin-like (framework IX) conoidean peptides. *Peptides* 30:467-476
- Aguilar MB, Lopez-Vera E, Ortiz E, Becerril B, Possani LD, Olivera BM, Heimer de la Cotera EP (2005) A novel conotoxin from *Conus delessertii* with posttranslationally modified lysine residues. *Biochemistry* 44:11130-11136
- Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*:783-795
- Bias D, Dutertre S, Gerbault A, Menou J-L, Offord R, Favreau P, Stöcklin R (2009) Comparative proteomic study of the venom of the piscivorous cone snail *Conus consors*. *Journal of Proteomics* 72:210-218
- Biggs JS, Watkins M, Puillandre N, Ownby JP, Lopez-Vera E, Christensen S, Moreno KJ, Bernaldez J, Licea-Navarro A, Showers Corneli P, Olivera BM (2010) Evolution of *Conus* peptide toxins: analysis of *Conus californicus* Reeve, 1844. *Molecular Phylogenetics and Evolution* 56:1-12
- Blunt JW, Copp BR, Keyzers RA, Munro MH, Prinsep MR (2012) Marine natural products. *Natural Product Reports* 29:144-222
- Bouchet P, Lozouet P, Sysoev AV (2009) An inordinate fondness for turrids. *Deep-Sea Research II* 56:1724-1731
- Cabang AP, Imperial JS, Gajewiak J, Watkins M, Showers Corneli P, Olivera BM, Concepcion GP (2011) Characterization of a venom peptide from a crassispirid gastropod. *Toxicon* 58:672-680
- Chang C, Duda TF (2012) Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Molecular Biology and Evolution* Advance access
- Coticello SG, Gilad Y, Avidan N, Ben-Asher E, Levy Z, Fainzilber M (2001) Mechanisms for evolving hypervariability: the case of conopeptides. *Molecular Biology and Evolution* 18:120-131
- Coticello SG, Pilpel Y, Glusman G, Fainzilber M (2000) Position-specific codon conservation in hypervariable gene families. *Trends in Genetics* 16:57-59
- Craig AG, Norberg T, Griffin D, Hoeger C, Akhtar M, Schmidt K, Low W, Dykert J, Richelsoni E, Navarro V, Mazella J, Watkins M, Hillyard DR, Imperial J, Cruz LJ, Olivera BM (1999) Contulakin-G, an O-glycosylated invertebrate neurotensin. *Journal of Biological Chemistry* 274:13752-13759
- Craig AG, Zafaralla G, Cruz LJ, Santos AD, Hillyard DR, Dykert J, Rivier J, Gray WR, Imperial J, Delacruz RG, Sporning A, Terlau H, West PJ, Yoshikami D, Olivera BM (1998) An O-glycosylated neuroexcitatory *Conus* peptide. *Biochemistry* 37:16019-16025
- Daly NL, Craik DJ (2009) Structural studies of conotoxins. *IUBMB Life* 61:144-150

- 1  
2  
3 Davis J, Jones A, Lewis RJ (2009) Remarkable inter- and intra-species complexity of conotoxins  
4 revealed by LC/MS. *Peptides* 30:1222-1227
- 5 Duda JTF, Lee T (2009) Ecological release and venom evolution of a predatory marine Snail at  
6 Easter Island. *PLoS ONE* 4:e5558
- 7  
8 Duda TF (2008) Differentiation of venoms of predatory marine gastropods: divergence of  
9 orthologous toxin genes of closely related *Conus* species with different dietary  
10 specializations. *Journal of Molecular Evolution* 67:315-321
- 11 Duda TF, Kohn AJ (2005) Species-level phylogeography and evolutionary history of the  
12 hyperdiverse marine gastropod genus *Conus*. *Molecular Phylogenetics and Evolution* 34:257-  
13 272
- 14  
15 Duda TF, Palumbi SR (1999) Molecular genetics of ecological diversification: duplication and rapid  
16 evolution of toxin genes of the venomous gastropod *Conus*. *Proceedings of the National  
17 Academy of Sciences* 96:6820-6823
- 18  
19 Duda TF, Palumbi SR (2000) Evolutionary diversification of multigene families: allelic selection of  
20 toxins in predatory cone snails. *Molecular Biology and Evolution* 17:1286-1293
- 21 Duda TF, Palumbi SR (2004) Gene expression and feeding ecology: evolution of piscivory in the  
22 venomous gastropod genus *Conus*. *Proceedings of the Royal Society B* 271:1165-1174
- 23 Duda TF, Remigio A (2008) Variation and evolution of toxin gene expression patterns of six closely  
24 related venomous marine snails. *Molecular Ecology* 17:3018-3032
- 25  
26 Dutertre S, Biass D, Stöcklin R, Favreau P (2010) Dramatic intraspecimen variations within the  
27 injected venom of *Conus consors*: an unsuspected contribution to venom diversity. *Toxicon*  
28 55:1453-1462
- 29  
30 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput.  
31 *Nucleic Acids Research* 32:1792-1797
- 32  
33 Espiritu DJD, Watkins M, Dia-Monje V, Cartier GE, Cruz LE, Olivera BM (2001) Venomous cone  
34 snails: molecular phylogeny and the generation of toxin diversity. *Toxicon* 39:1899-1916
- 35 Favreau P, Benoit E, Hocking E, Carlier L, D'hoedt D, Leipold E, Markgraf D, Schlumberger S,  
36 Cordova M, Gaertner H, Paolini-Bertrand M, Hartley O, Tytgat J, Heinemann S, Bertrand D,  
37 Boelens R, Stöcklin R, Molgo J (2012) A novel mu-conopeptide, CnIIIC, exerts potent and  
38 preferential inhibition of NaV1.2/1.4 channels and blocks neuronal nicotinic acetylcholine  
39 receptors. *British Journal of Pharmacology* in press
- 40 Favreau P, Stöcklin R (2009) Marine snail venoms: use and trends in receptor and channel  
41 neuropharmacology. *Current Opinion in Pharmacology* 9:594-601
- 42 Fedosov A, Kantor Y (2008) Toxoglossan gastropods of the subfamily Crassispirinae (Turridae)  
43 lacking a radula, and a discussion of the status of the subfamily Zemaciinae. *Journal of  
44 Molluscan Studies* 74:27-35
- 45  
46 Fedosov AE (2007) Anatomy of accessory rhynchodeal organs of *Veprecula vepatica* and  
47 *Tritonoturris subbrissoides*: new types of foregut morphology in Raphitominae (Conoidea).  
48 *Ruthenica* 17:33-41
- 49  
50 Gayler K, Sandall D, Greening D, Keays D, Polidano M, Livett B, Down J, Satkunanathan N, Khalil  
51 Z (2005) Molecular prospecting for drugs from the sea. *IEEE Engineering in Medicine and  
52 Biology Magazine* 24:79-84
- 53  
54 Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program  
55 for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41:95-98
- 56  
57 Han TS, Teichert RW, Olivera BM, Bulaj G (2008a) *Conus* venoms - a rich source of peptide-based  
58 therapeutics. *Curr Pharm Des* 14:2462-2479
- 59  
60

- 1  
2  
3 Han Y, Huang F, Jiang H, Liu L, Wang Q, Wang Y, Shao X, Chi C, Du W, Wang C (2008b)  
4 Purification and structural characterization of a d-amino acid-containing conopeptide,  
5 conomarphin, from *Conus marmoreus*. FEBS Journal 275:1976-1987  
6  
7 Heralde FM, Imperial J, Bandyopadhyay P, Olivera BM, Concepcion GP, Santos AD (2008) A  
8 rapidly diverging superfamily of peptide toxins in venomous *Gemmula* species. Toxicon  
9 51:890-897  
10  
11 Holford M, Puillandre N, Terryn Y, Cruaud C, Olivera BM, Bouchet P (2009) Evolution of the  
12 Toxoglossa venom apparatus as inferred by molecular phylogeny of the Terebridae.  
13 Molecular Biology and Evolution 26:15-25  
14  
15 Hopkins C, Grilley M, Miller C, Shon K-J, Cruz LJ, Gray WR, Dykert J, Rivier J, Yoshikami D,  
16 Olivera BM (1995) A new family of *Conus* peptides targeted to the nicotinic acetylcholine  
17 receptor. Journal of Biological Chemistry 270:22361-22367  
18  
19 Hu H, Bandyopadhyay PK, Olivera BM, Yandell M (2011) Characterization of the *Conus bullatus*  
20 genome and its venom-duct transcriptome. BMC Genomics 12:60  
21  
22 Huelsenbeck JP, Ronquist F, Hall B (2001) MrBayes: bayesian inference of phylogeny.  
23 Bioinformatics 17:754-755  
24  
25 Imperial JS, Kantor Y, Watkins M, Heralde FM, Stevenson B, Chen P, Hansson K, Stenflo J, Ownby  
26 J-P, Bouchet P, Olivera BM (2007) Venomous auger snail *Hastula (Impages) hectica*  
27 (Linnaeus, 1758): molecular phylogeny, foregut anatomy and comparative toxinology.  
28 Journal of Experimental Zoology 308B:744-756  
29  
30 Imperial JS, Watkins M, Chen P, Hillyard DR, Cruz LJ, Olivera BM (2003) The augertoxins:  
31 biochemical characterization of venom components from the toxoglossate gastropod *Terebra*  
32 *subulata*. Toxicon 42:391-398  
33  
34 Jakubowski JA, Kelley WP, Sweedler JV, Gilly WF, Schulz JR (2005) Intraspecific variation of  
35 venom injected by fish-hunting *Conus* snails. Journal of Experimental Biology 208:2873-  
36 2883  
37  
38 Jimenez EC, Olivera BM, Teichert RW (2007)  $\alpha$ C-conotoxin PrXA: a new family of nicotinic  
39 acetylcholine receptor antagonists. Biochemistry 46:8717-8724  
40  
41 Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from  
42 protein sequences. CABIOS 8:275-282  
43  
44 Kaas Q, Westermann JC, Craik DJ (2010) Conopeptide characterization and classifications: An  
45 analysis using ConoServer. Toxicon 55:1491-1509  
46  
47 Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO (2006) Assessment of methods  
48 for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions  
49 for choice of matrix are not justified. BMC Evolutionary Biology 6:1-17  
50  
51 Koua D, Brauer A, Laht S, Kaplinski L, Favreau P, Remm M, Lisacek F, Stöcklin R (2012)  
52 ConoDictor: a tool for prediction of conopeptide superfamilies. Nucleic Acids Research in  
53 press  
54  
55 Kraus NJ, Showers Corneli P, Watkins M, Bandyopadhyay PK, Seger J, Olivera BM (2011) Against  
56 expectation: a short sequence with high signal elucidates cone snail phylogeny. Molecular  
57 Phylogenetics and Evolution 58:383-389  
58  
59 Laht S, Koua D, Kaplinski L, Lisacek F, Stöcklin R, Remm M (2011) Identification and  
60 classification of conopeptides using profile Hidden Markov Models. Biochimica et  
Biophysica Acta 1824:488-492  
Leary D, Vierros M, Hamon G, Arico S, Monagle C (2009) Marine genetic resources: A review of  
scientific and commercial interest. Marine Policy 33:183-194

- 1  
2  
3 Lewis RJ (2012) Discovery and development of the  $\chi$ -conopeptide class of analgesic peptides.  
4 Toxicon in press
- 5 Lin H, Li Q-Z (2007) Predicting conotoxin superfamily and family by using pseudo amino acid  
6 composition and modified Mahalanobis discriminant. Biochemical and Biophysical Research  
7 Communications 354 548-551
- 8 Lopez-Vera E, Heimer de la Cotera EP, Maillo M, Riesgo-Escovar JR, Olivera BM, Aguilar MB  
9 (2004) A novel structure class of toxins: the methionine-rich peptides from the venoms of  
10 turrid marine snails (Mollusca, Conoidea). Toxicon 43:365-374
- 11 McGivern JG (2007) Ziconotide: a review of its pharmacology and use in the treatment of pain.  
12 Neuropsychiatric Disease and Treatment 3:69-85
- 13 Medinskaya AI, Sysoev A (2003) The anatomy of *Zemacies excelsa*, with a description of a new  
14 subfamily of Turridae (Gastropoda, Conoidea). Ruthenica 13:81-87
- 15 Mena EE, Gullak MF, Pagnozzi MJ, Richter KE, Rivier J, Cruz LJ, Olivera BM (1990) Conantokin-  
16 G: a novel peptide antagonist to the N-methyl-D-aspartic acid (NMDA) receptor.  
17 Neuroscience Letters 118:241-244
- 18 Menez A, Stocklin R, Mebs D (2006) Venomics' or: the venomous systems genome project.  
19 Toxicon 47:255-259
- 20 Miljanich GP (2004) Ziconotide: neuronal calcium channel blocker for treating severe chronic pain.  
21 Current Medicinal Chemistry 11:3029-3040
- 22 Molinski TF, Dalisay DS, Lievens SL, Saludes JP (2009) Drug development from marine natural  
23 products. Nature Reviews Drug Discovery 8:69-85
- 24 Möller C, Melaun C, Castillo C, Díaz ME, Renzelman CM, Estrada O, Kuch U, Lokey S, Marí F  
25 (2010) Functional hypervariability and gene diversity of cardioactive neuropeptides. Journal  
26 of Biological Chemistry 285:40673-40680
- 27 Mondal S, Bhavna R, Babu RM, Ramakumar S (2006) Pseudo amino acid composition and multi-  
28 class support vector machines approach for conotoxin superfamily classification. Journal of  
29 Theoretical Biology 243 252-260
- 30 Norton RS, Olivera BM (2006) Conotoxins down under. Toxicon 48:780-798
- 31 Olivera BM (2002) *Conus* venom peptides: reflections from the biology of clades and species.  
32 Annual Review of Ecology and Systematics 33:25-47
- 33 Olivera BM (2006) *Conus* peptides: biodiversity-based discovery and exogenomics. Journal of  
34 Biological Chemistry 281:31173-31177
- 35 Olivera BM, Walker C, Cartier GE, Hooper D, Santos AD, Schoenfeld R, Shetty R, Watkins M,  
36 Bandyopadhyay PK, Hillyard DR (1999) Speciation of cone snails and interspecific  
37 hyperdivergence of their venom peptides. Potential evolutionary significance of introns.  
38 Annals of the New York Academy of Sciences 870:223-237
- 39 Pi C, Liu J, Peng C, Liu Y, Jiang X, Zhao Y, Tang S, Wang L, Dong M, Chen S, Xu A (2006)  
40 Diversity and evolution of conotoxins based on gene expression profiling of *Conus litteratus*.  
41 Genomics 88:809-819
- 42 Puillandre N, Holford M (2010) The Terebridae and teretoxins: combining phylogeny and anatomy  
43 for concerted discovery of bioactive compounds. BMC Chemical Biology 10:7
- 44 Puillandre N, Kantor Y, Sysoev A, Couloux A, Meyer C, Rawlings T, Todd JA, Bouchet P (2011)  
45 The dragon tamed? A molecular phylogeny of the Conoidea (Mollusca, Gastropoda). Journal  
46 of Molluscan Studies 77:259-272
- 47 Puillandre N, Watkins M, Olivera BM (2010) Evolution of *Conus* peptide genes: duplication and  
48 positive selection in the A-superfamily. Journal of Molecular Evolution 70:190-202
- 49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 Quinton L, Gilles N, De Pauw E (2009) TxXIIIa, an atypical homodimeric conotoxin found in the  
4 *Conus textile* venom. *Journal of Proteomics* 72:219-226  
5  
6 Rambaut A, Drummond AJ (2007) Tracer v1.4. Available from <http://beast.bio.ed.ac.uk/Tracer>  
7  
8 Rojas A, Feregrino A, Ibarra-Alvarado C, Aguilar MB, Falcon A, Heimer de la Cotera EP (2008)  
9 Pharmacological characterization of venoms obtained from mexican toxoglossate gastropods  
10 on isolated guinea pig ileum. *Journal of Venomous Animals and Toxins including Tropical*  
11 *Diseases* 14:497-513  
12  
13 Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular  
14 Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and  
15 Maximum Parsimony Methods. *Molecular Biology and Evolution* 28:2731-2739  
16  
17 Terrat Y, Biass D, Dutertre S, Favreau P, Remm M, Stöcklin R, Piquemal D, Ducancel F (2011)  
18 High-resolution picture of a venom gland transcriptome: Case study with the marine snail  
19 *Conus consors*. *Toxicon* 59:34-46  
20  
21 Ueberheide BM, Fenyo D, Alewood PF, Chait BT (2009) Rapid sensitive analysis of cysteine rich  
22 peptide venom components. *Proceedings of the National Academy of Sciences* 106:6910-  
23 6915  
24  
25 Violette A, Leonardi A, Piquemal D, Terrat Y, Biass D, Dutertre S, Noguier F, Ducancel F, Stöcklin  
26 R, Križaj I, Favreau P (2012) Recruitment of glycosyl hydrolase proteins in a cone snail  
27 venomous arsenal: further insights into biomolecular features of *Conus* venoms. *Marine*  
28 *Drugs* 10:258-280  
29  
30 Walker CS, Jensen S, Ellison M, Matta JA, Lee WY, Imperial JS, Duclos N, Brockie PJ, Madsen  
31 DM, Isaac JT, Olivera BM, Maricq AV (2009) A novel *Conus* snail polypeptide causes  
32 excitotoxicity by blocking desensitization of AMPA receptors. *Current Biology* 19:900-908  
33  
34 Wang Q, Jiang H, Hana Y-H, Yuan D-D, Chi C-W (2008) Two different groups of signal sequence  
35 in M-superfamily conotoxins. *Toxicon* 51:813-822  
36  
37 Watkins M, Hillyard DR, Olivera BM (2006) Genes expressed in a Turrid venom duct: divergence  
38 and similarity to conotoxins. *Journal of Molecular Evolution* 62:247-256  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Table 1: Number of sequences found in each superfamily, with list of cysteine patterns identified and known function in each superfamily.

Superfamily ID	# of sequences	ID	Cysteine pattern	# of sequences	Known function
A	153	I	CC-C-C	119	$\alpha$ , $\kappa$ , $\rho$
		II	CCC-C-C-C	3	
		IV	CC-C-C-C-C	25	
		VI/VII	C-C-CC-C-C	1	
		XIV	C-C-C-C	3	
			C	1	
B	41		0	38	conantokin
			C-C	3	
C	4		0	1	contulakin
			C-C	3	
D	13	XX	C-CC-C-CC-C-C-C-C	5	$\alpha$
			C-CC-C-CC-C-C-C	1	
			C-C-C-CC-C-C-C-C	7	
I1	6	XI	C-C-CC-CC-C-C	6	$\iota$
I2	45	XI	C-C-CC-CC-C-C	35	$\kappa$
		XII	C-C-C-C-CC-C-C	9	
			C-C-CC-CC-C	1	
I3	7	XI	C-C-CC-CC-C-C	7	
J	12	XIV	C-C-C-C	12	$\alpha + \kappa$
L	4	XIV	C-C-C-C	3	$\alpha$
			C-C-C	1	
M	193		0	3	$\alpha$ , $\kappa$ , $\mu$ , conomarphin
		II	CCC-C-C-C	1	
		III	CC-C-C-CC	172	
		IV	CC-C-C-C-C	4	
		IX	C-C-C-C-C-C	1	
		XVI	C-C-CC	1	
		XIX	C-C-C-CCC-C-C-C-C	1	
			C	1	
			C-C	2	
			CC-C-C-C	1	
			CC-C-C-CC-C	2	
			C-CC-C-C-C	4	
O1	625	VI/VII	0	4	$\delta$ , $\kappa$ , $\mu$ , $\omega$
			C-C-CC-C-C	613	
			C-C-C	1	
			C-C-CC-C	1	
			C-CC-C-C	4	
			C-C-C-C-C	1	
	C-C-CC-C-CC	1			
O2	67	VI/VII	C-C-CC-C-C	51	$\gamma$ , contryphan
			C-C	7	
		XV	C-C-CC-C-C-C-C	9	
O3	25	VI/VII	C-C-CC-C-C	25	bromosleeper
P	7	XIV	C-C-C-C	2	
			IX	C-C-C-C-C-C	
S	7	VIII	C-C-C-C-C-C-C-C-C-C	7	$\sigma$ , $\alpha$
T	140	X	0	12	$\varepsilon$ , $\chi$ , $\tau$
			CC-CXPC	4	



		V		CC-CC	121	
				C-C	2	
				CC-CCC	1	
	V	2	XV	C-C-CC-C-C-C-C	2	
	X1	2		C-C-C-C-C-C-C	2	conoCAP
	X2		III	CC-C-C-CC	1	
	X3	1		0	1	
	X4	2		C-C-C-C-C-C-C-C-C-CC-C-C-C-C-C-C-C-C-C	2	
	X5	1	VIII	C-C-C-C-C-C-C-C-C	1	
	X6	3		C-C-C-CC-C	3	
	X7	1	VIII	C-C-C-C-C-C-C-C-C	1	
	Y	1	XVII	C-C-CC-C-CC-C	1	

Table 2: Number of conopeptides in each superfamily and species. Feeding types: F: Fish hunting species; M: Mollusc-hunting species; W: Worm-hunting species

Species	Prey	A	B	C	D	I1	I2	I3	J	L	M	O1	O2	O3	P	S	T	V	X1	X2	X3	X4	X5	X6	X7	Y	Occurrence
<i>achatinus</i>	F	4														1											2
<i>aurisiacus</i>	F	2									5	1															3
<i>bullatus</i>	F	4									10	8															3
<i>catus</i>	F	4									29																2
<i>circumciscus</i>	F	4									1	4															3
<i>consors</i>	F	8	1								3	10															4
<i>ermineus</i>	F	4									1	4															3
<i>geographus</i>	F	3	5	1							2	5					1	2									7
<i>lynceus</i>	F										1																1
<i>magus</i>	F	5									5	13	1														4
<i>monachus</i>	F	4										3															2
<i>obscurus</i>	F	3	5																								2
<i>ochroleucus</i>	F		2																								1
<i>parius</i>	F			1																							1
<i>purpurascens</i>	F	4									1	9						1									4
<i>radiatus</i>	F	6	3			2					4	3		1		1											7
<i>stercusmuscarum</i>	F	6									3	6															3
<i>striatus</i>	F	10					1				3	50					1										5
<i>striolatus</i>	F	1										8															2
<i>sulcatus</i>	F	8	2																								2
<i>tulipa</i>	F	3									2	2															3
<i>ammiralis</i>	M	2									5	3															3
<i>aulicus</i>	M	2									3	3															3
<i>aureus</i>	M										2																1
<i>bandanus</i>	M	2																									1
<i>dalli</i>	M										3	8															2
<i>episcopatus</i>	M	1				1					3	2															4
<i>gloriamaris</i>	M		2								3	6	1		1		4										6
<i>marmoreus</i>	M	4					2				14	15	5				12				2						7
<i>omaria</i>	M										4	6															2
<i>pennaceus</i>	M	4									6	13	3	2			16										6
<i>textile</i>	M	5					2				15	29	24		1	1	18										8
<i>victoriae</i>	M	3																									1
<i>abbreviatus</i>	W											98															1
<i>arenatus</i>	W		1								2	31		8			8										5
<i>aristophanes</i>	W											8															1
<i>betulinus</i>	W	8				1	1				10	1	2				1										7
<i>capitaneus</i>	W						1				1	2	2														4
<i>characteristicus</i>	W	2	5					3			6	3	1	1		3	5								1	1	11
<i>coronatus</i>	W										1	7															2
<i>distans</i>	W	1									2	5															3
<i>ebraeus</i>	W											29															1
<i>eburneus</i>	W	2	1	2			4		1	1	6	2		1													9

1																													
2																													
3	<i>emaciatius</i>	W								3				1								1						3	
4	<i>ferrugineus</i>	W												2								1						2	
5	<i>figulinus</i>	W																				5						1	
6	<i>flavidus</i>	W																				1						1	
7	<i>generalis</i>	W																				2		2				2	
8	<i>imperialis</i>	W	4	3		1	2		6					2	7	1				1	2				1		1	3	13
9	<i>judaicus</i>	W																				2						1	
10	<i>leopardus</i>	W	8												5	9						4						4	
11	<i>litteratus</i>	W	4	2	2	7		6	1	3	3				11	8	5			3	16					1		1	15
12	<i>lividus</i>	W		3											1	84	2	1				3						6	
13	<i>miles</i>	W	1			1		1							6	2						1						6	
14	<i>miliaris</i>	W													2	18												2	
15	<i>musicus</i>	W													1													1	
16	<i>mustelinus</i>	W				2																						1	
17	<i>planorbis</i>	W									4				1													2	
18	<i>pulicarius</i>	W	4								3				3	6						6						5	
19	<i>quercinus</i>	W	6	3											4	3						2						5	
20	<i>rattus</i>	W													2	4												2	
21	<i>regius</i>	W																										1	
22	<i>sponsalis</i>	W	2														14											2	
23	<i>spurius</i>	W																										3	
24	<i>tessulatus</i>	W													10	1						8						5	
25	<i>ventricosus</i>	W													14	6	2	4				11						5	
26	<i>vexillum</i>	W													6	12	9	5				15						5	
27	<i>villepini</i>	W				2		2							1	8	3											5	
28	<i>viola</i>	W																				5						1	
29	<i>virgo</i>	W	2	2				4							5	3	2				4	1						8	
30	<i>vitulinus</i>	W	3	1			2		2						2							1						6	
31			39	16	4	5	4	14	3	5	2	50	51	17	9	5	5	21	2	1	1	1	1	2	1	1	1	1	
32	<b>Occurrence</b>																												

## Figure caption

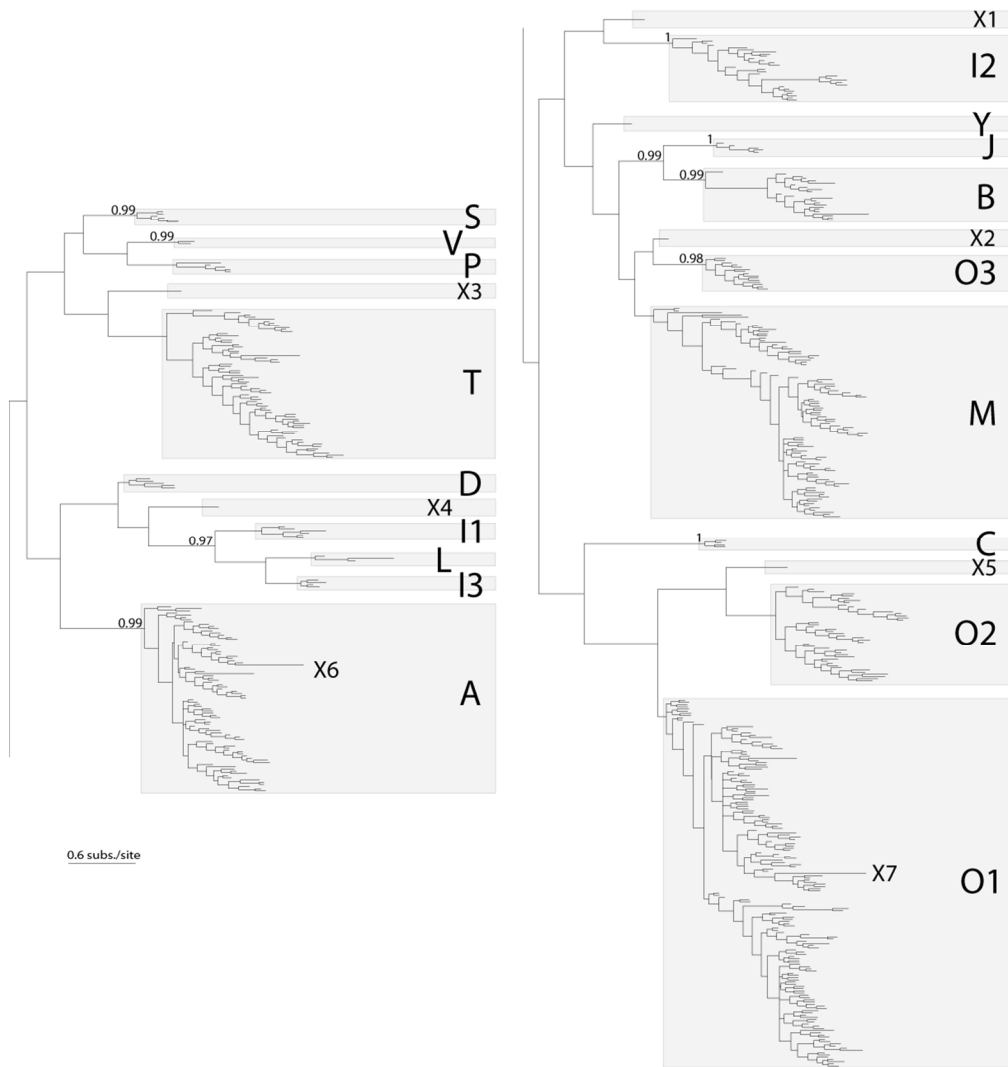
Figure 1: Bayesian phylogenetic tree (midpoint rooting) obtained from the Clustal alignment of the signal sequences of conopeptides from GenBank. Posterior Probabilities (when  $> 0.9$ ) are provided for each node. Grey boxes are used to visualize the superfamilies. The B and C superfamilies respectively correspond to the contulakins and conantokins. The lineages X1-X7 potentially correspond to previously unrecognized superfamilies (see details in the text).

Figure 2: Pairwise distribution of genetic distances (p-distances) calculated with MEGA5 using the Clust alignment. Genetic distances between sequences from the same superfamily are shown in grey, genetic distances between sequences from different superfamily in black.

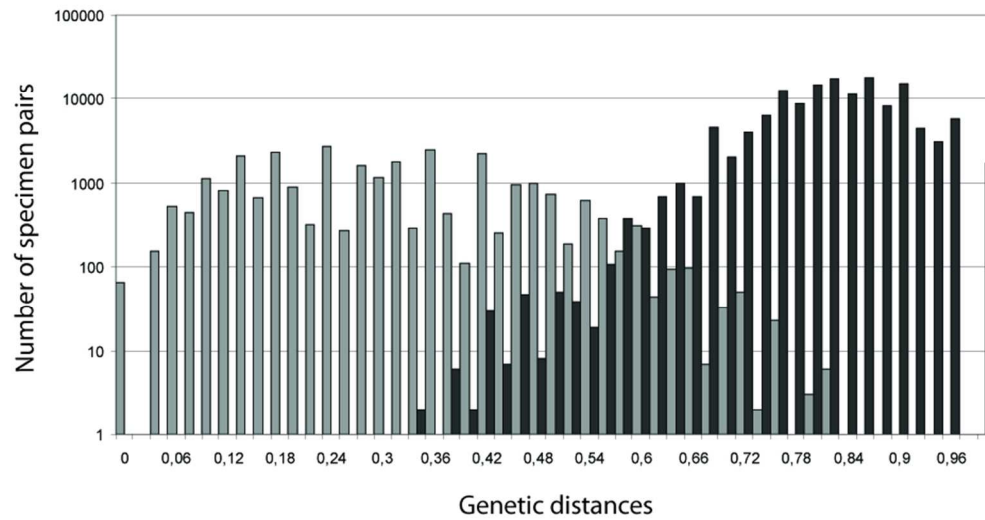
1  
2  
3 Appendix 1: List of analysed sequences with superfamily assignation, GenBank numbers, Cys-  
4  
5 pattern, species from which the sequence originated and corresponding feeding type F: Fish  
6  
7 hunting species; M: Mollusc-hunting species; W: Worm-hunting species).  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

For Peer Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



205x218mm (150 x 150 DPI)



159x85mm (150 x 150 DPI)

Peer Review

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60