Evolution of Conus Peptide Genes: Duplication and Positive Selection in the A-Superfamily

Nicolas Puillandre, Maren Watkins, Baldomero M Olivera

▶ **To cite this version:**

Nicolas Puillandre, Maren Watkins, Baldomero M Olivera. Evolution of Conus Peptide Genes: Duplication and Positive Selection in the A-Superfamily. Journal of Molecular Evolution, 2010, 70 (2), pp.190-202. 10.1007/s00239-010-9321-7 . hal-02458071

HAL Id: hal-02458071
https://hal.science/hal-02458071

Submitted on 28 Jan 2020

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.

Evolution of *Conus* Peptide Genes: Duplication and Positive Selection in the A-superfamily.

Nicolas Puillandre, Maren Watkins and Baldomero M. Olivera

Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112.

Corresponding Author: Nicolas Puillandre, Department of Biology, University of Utah, 257 South 1400 East, Salt Lake City, UT 84112.Tel: 801-581-8370. fax: 801-585-5010.

puillandre@biology.utah.edu

ABSTRACT

A remarkable diversity of venom peptides is expressed in the genus *Conus* (known as "conotoxins" or "conopeptides"). Between 50 and 200 different venom peptides can be found in a single *Conus* species, each having its own complement of peptides. Conopeptides are encoded by a few gene superfamilies; here we analyze the evolution of the A-superfamily in a fish-hunting species clade, *Pionoconus*.

More than 90 conopeptide sequences from eleven different *Conus* species were used to build a phylogenetic tree. Comparison with a species tree based on standard genes reveals multiple gene duplication events, some of which took place before the *Pionoconus* radiation. By analysing several A-conopeptides from other *Conus* species recorded in GenBank, we date the major duplication events after the divergence between fish-hunting and non-fish-hunting species. Furthermore, likelihood approaches revealed strong positive selection; the magnitude depends on which A-conopeptide lineage and amino-acid locus is analyzed.

The four major A-conopeptide clades defined are consistent with the current division of the superfamily into families and subfamilies based on the Cys-pattern. The function of three of these clades (the κA-family, the α4/7-subfamily and α3/5-subfamily) has previously been characterized. The function of the remaining clade, corresponding to the α4/4-subfamily, has not been elucidated. This subfamily is also found in several other fish-hunting species clades within *Conus*. The analysis revealed a surprisingly diverse origin of α4/4 conopeptides from a single species, *Conus bullatus*. This phylogenetic approach that defines different genetic lineages of *Conus* venom peptides provides a guidepost for identifying conopeptides with potentially novel functions.

Key Words: Duplication, *Conus*, Positive selection, A-superfamily conotoxin, Concerted discovery, Molecular phylogeny.

INTRODUCTION

The predatory cone snails (genus *Conus*) paralyze prey, defend against predators and deter competitors using venoms that are complex mixtures of relatively small peptides (mostly 10-35 amino acids) with potent neuropharmacological activity. The 50-200 different peptides that can be expressed in the venom of a single *Conus* species are encoded by a relatively small number of gene superfamilies (Olivera 2006) that exhibit an unprecedented rate of accelerated evolution (e.g. Conticello et al. 2001; Duda and Remigio 2008).

A cone snail venom peptide can be assigned to a gene superfamily and to groups within a superfamily using several criteria. All *Conus* peptides are processed from prepropeptide precursors translated by ribosomes from mRNA transcripts expressed in epithelial cells of the venom duct of the cone snail (Woodward et al. 1990). The signal sequences of precursors of all members of a conopeptide superfamily are highly conserved, sharing considerable sequence identity. This conserved signal sequence is a signature sequence element that unequivocally identifies the gene superfamily to which a venom peptide belongs. In addition, the arrangement of cysteine residues within the primary structure of the mature peptide toxin (the "Cys pattern") is generally characteristic of the specific gene group (within a given superfamily) to which the peptide belongs (although this feature is not as conserved as the signal sequence). In the mature toxin region of most conopeptide superfamilies, the Cys codons at each locus are conserved (Conticello et al. 2001). The conopeptides that belong to the A-gene superfamily share a consensus signal sequence, and the Cys patterns in the mature peptides can be used to assign it to a group within the A-superfamily (Santos et al. 2004; Olivera et al. 2008).

In the venoms of several fish-hunting cone snails, the pharmacological mechanisms underlying the biological activities of several A-superfamily peptides are well understood. One extensively characterized group of A-superfamily peptides is the α-conotoxin group that targets the muscle subtype of nicotinic acetylcholine receptors (nAChR). These peptides

block neuromuscular transmission in their prey. A second group of A-superfamily peptides found in the venoms of fish-hunting cone snails, which is functionally very different, is a group of excitatory peptides called the κA-conotoxins. The precise molecular target of these peptides is still not definitively established, but instead of causing paralysis and a relaxation of the skeletal musculature, the κA- peptides cause nerves to fire uncontrollably. When injected into the mammalian CNS, these peptides elicit seizures. They are one component of a group of peptides that cause hyperexcitability of axons at the venom injection site, resulting in the almost instantaneous onset of a tetanic immobilization of the prey with extreme rigidity of the skeletal musculature (see Terlau et al. 1996).

The two different groups of conopeptides described above, though both in the A-superfamily, have quite different Cys-patterns. The first group, the α-conotoxins targeted to the muscle nAChR subtype have the following Cys pattern: [---**C**C(X_3)**C**(X_5)**C**---]. The arrangement of the Cys residues in the primary structure (i.e., —CC—C—C—) is characteristic of all α-conotoxins, but the paralytic *Conus* peptides found in fish-hunting cone snails comprise a subgroup of the α-conotoxin family called the α3/5 subfamily. These peptides all have 3AA between Cys2 and Cys3, and 5AA between Cys3 and Cys4. In contrast, the κA-conotoxins have 6 Cys residues with the following Cys pattern: [---**C**C(X_{6-7})**C**(X_2)**C**X**C**(X_3)**C**---].

Most fish-hunting *Conus* species express multiple A-superfamily peptides. Several (such as the striated cone, *Conus striatus*) are known to have several α3/5 subfamily peptides in their venoms (see Zafaralla et al. 1988). In addition, peptides that structurally do not belong to the two classes described above (the α3/5 subfamily and the κA-conotoxins) have been purified from the venoms of these species as well. The best characterized of these is α–conotoxin MII, a peptide specifically targeted to certain neuronal subtypes of nicotinic acetylcholine receptors. This peptide is not paralytic since it does not inhibit the muscle

nAChR subtype. It has been extremely useful for understanding the role of different neuronal nicotinic receptor subtypes in a variety of pathological conditions such as Parkinson's disease (see Olivera et al. 2008 for a review). In contrast to the α3/5 subfamily above that are all targeted to the muscle nicotinic acetylcholine receptor, α–conotoxin MII has a different structural motif: [---**CC(**X_4**)C(**X_7**)C**---]. The different spacing between cysteine residues indicates the subfamily to which this peptide belongs, the α4/7 subfamily.

Several conotoxin superfamily analyses are reported in literature (Duda and Palumbi 1999, 2000; Conticello et al. 2000, 2001; Espiritu et al. 2001; Wang et al. 2008; Duda and Remigio 2008; Aguilar et al. 2009). Evidence for duplication and positive selection (based on the estimation of the ratio between synonymous and non-synonymous mutations) has been found, particularly in the O-superfamily (e.g. Duda and Remigio 2008). However, the A-superfamily has never been thoroughly investigated. The analysis of a large dataset of conotoxins included in the A-superfamily from one group of fish-hunting cone snails that belong to the *Pionoconus* clade is reported below.

We chose the A-superfamily of conopeptides for this analysis for two reasons. First, sequences for peptides in this superfamily have been elucidated across a larger number of species. These were obtained not only using the standard analysis of cDNA libraries (which requires venom dissected from live specimens), but because a conserved intron sequence is present close to the mature toxin region, PCR of genomic DNA was also used to obtain A-superfamily sequences. A recent comprehensive and unbiased sampling of the transcriptome of a few *Conus* species indicates that the A-superfamily is a major gene family expressed in the venom ducts of *Conus* (F. Ducancel, unpublished results; A. Lluisma and P. Bandyopadhyay manuscript in preparation). The second reason for analysing this superfamily is that the diverse functional activity of these peptides is relatively well-defined, and more

structure/function information is available for A-superfamily than for any other *Conus* peptide gene superfamily.

A basic question that needs to be addressed is how the impressive diversity of peptide toxins in a single cone snail venom has been generated. Did all the sequences grouped according to their Cys-pattern evolve from a common ancestral sequence? When did different lineages within a gene superfamily first appear and how did they subsequently evolve? To address these questions, we superimposed the phylogeny of the fish-hunting *Conus* species included in the clade *Pionoconus* with the phylogeny of the A-superfamily conopeptides present in their venoms and assessed the potential effect of positive selection during A-superfamily evolution. We shown that the classic evolutionary pattern of multi-gene families (duplication followed by rapid diversification) was particularly pronounced in A gene superfamily of the *Pionoconus* clade.

MATERIAL AND METHODS

Species tree

Eleven species were included in the analysis: *Conus achatinus, C. aurisiacus*, *C. catus*, *C. circumcisus*, *C. consors*, *C. gauguini*, *C. magus*, *C. monachus*, *C. striatus*, *C. stercusmuscarum* and *C. striolatus*. All the species are included in the informal group *Pionoconus*, recognized as a clade by several independent analyses (e.g. Duda and Palumbi 1999; Espiritu et al. 2001).

To reconstruct the phylogenetic relationships between these species, partial sequences of four genes were used: COI, 16S, 12S and an intron of calmodulin (Duda and Palumbi 2004). These sequences were extracted from our own database or downloaded from GenBank (Table 1). Several outgroups were used: *C. textile* and *C. ebraeus*, both included in the "major" clade within *Conus* (Duda and Kohn 2005), *C. arcuatus* and *C. mahogani*, both included in the "minor" clade (Duda and Kohn 2005) and *C. californicus*, thought to be the

first *Conus* to diverge in the genus (Espiritu et al. 2001) and used to root the tree. Five other fish-hunting species that are not in the *Pionoconus* clade were also included in the tree (*C. ermineus*, *C. purpurascens*, *C. cervus*, *C. bullatus* and *C. kinoshitai*).

Sequences were aligned automatically using Bioedit (Hall 1999) and then modified manually. Substitution models were selected for each gene using Modelgenerator V.85 (Keane et al. 2006) following the Hierarchical Likelihood Ratio Test (with four discrete gamma categories). Trees were reconstructed using bayesian analyses, consisting of two independent analyses (six Markov chains, 10,000,000 generations, with a sampling frequency of one tree each thousand generations and three swaps at each sampling, and a temperature of 0.2 for each run) using Mr.Bayes 3.1.2 (Huelsenbeck et al. 2001). When the log-likelihood score was found to stabilize, a consensus tree was calculated after omitting the first 25% trees as burn-in.

A-conotoxin dataset

All the conotoxins from the A-superfamily available for these 11 species in our own database were added to the sequences available in GenBank. At the end, 92 sequences were included in the dataset, most of them (84) produced by our team (Table 2). One very short sequence (Cr4.3), too short to be correctly aligned, was removed from the dataset. All conotoxin sequences were obtained following the procedure described in Santos et al. (2004).

Sequences were temporarily translated to amino acids to facilitate the alignment, performed first automatically using Bioedit (Hall 1999) and then modified manually. The alignment of the highly conserved signal sequence, as well as the Cysteine pattern ([---**C**$\mathbf{C}(X_n)\mathbf{C}(X_n)\mathbf{C}$---] for the α and [---**C**$\mathbf{C}(X_{6\text{-}7})\mathbf{C}(X_2)\mathbf{C}\mathrm{X}\mathbf{C}(X_3)\mathbf{C}$---] for the κA), were used to drive the alignment for the remaining parts of the sequences. Amino acids before the beginning of the coding region and after the stop codon were removed from the analysis. The

same phylogenetic method applied to the species tree was used here with the A-conotoxin DNA sequences, except that the temperature of the chains was set to 0.02.

As is commonly found in multigenic families, one sequence from a given species can be phylogenetically more closely related to a sequence from another species than to a different sequence from the given species. Consequently, using an A-conotoxin sequence from a non-*Pionoconus* species as an outgroup was not possible as *Pionoconus* A-conotoxin sequences might not be exclusively monophyletic. Furthermore, it was not possible to use a conotoxin from another conotoxin superfamily as they were not alignable: even the signal sequences are highly divergent. Therefore, the tree are displayed unrooted, and several alternative rooted trees are discussed.

We used Notung 2.6 (Durand et al. 2006; Vernot et al. 2008) to reconcile the gene tree obtained with the A-conotoxin dataset with the species tree and quantify and locate gene duplications and losses. The position of the root in the gene tree was tentatively identified by minimizing the number of duplications and losses.

Tests for positive selection

The role of positive selection in the evolution of the A-conotoxin superfamily within *Pionoconus* has been assessed using likelihood approaches implemented in the program codeml of the PAML 4.2 package (Yang 2007). Several nucleotide substitution models are available, and the comparison between the likelihood of these models can be used to test for different alternative hypotheses. In all analyses, the tree topology obtained with the bayesian analysis was used, and branch lengths were estimated by PAML 4.2 (method = 1).

First, different site-models were compared to evaluate the effect of positive selection along the nucleotide sequences (Yang and Bielawski 2000; Yang 2002, 2006). Two pairs of model were compared: M1a vs M2a and M7 vs M8 (NSsites = 1, 2, 7, 8). The M1a and M7 models assume that the dN/dS ratio along the sequence ranges from 0 to 1 (purifying selection to

neutral drift), while the M2a and M8 assume that a few sites have a dN/dS ratio (*i.e.* ω) > 1 (positive selection). Furthermore, the M7 and M8 models assume a beta distribution of the ω classes. The likelihoods of these four models were compared using a likelihood ratio test (LRT) with a chi-square distribution (Yang et al. 1998, Wong et al. 2004). The Bayes Empirical Bayes approach (BEB – Yang et al. 2005) was used to calculate the posterior probabilities (PP) for site classes. A site was considered positively selected if PP > .95.

Second, a branch-model was used to test for different values of ω in the different lineages found in the tree (model = 2; Yang 1998; Yang and Nielsen 1998). Different likelihoods calculated under several models were compared: ω fixed in all branches, ω estimated but identical in all branches, ω estimated but different in all branches, ω fixed in some branches, ω identical in some branches and different in others. The comparison between these models allow to test for several hypotheses i.e. whether positive selection is more important in some branches than in others, or if ω is statistically superior to 1 in the whole tree or in specific branches. The likelihood of each of these models was calculated when using the whole sequence but also when considering only the mature toxin.

RESULTS

Species tree versus gene tree

The best models of evolution for the COI, 12S, 16S and calmodulin intron are respectively GTR+G (General Time Reversible model, with a gamma law parameter – α = 0.12), TVM+G (Transversional Model – α = 0.15), TVM+G (α = 0.11), and HKY+I (Hasegawa, Kishino and Yano, with invariant sites – I = 0.22). As no contradictions were found between independent analyses (results not shown), we combined the four gene fragments in a single dataset and ran a bayesian analysis where each gene was treated as a separate partition. The tree obtained for the fish-hunting species in *Pionoconus* support the monophyly of *Pionoconus* (PP = 1 [Fig. 1]). Most relationships within *Pionoconus* are

supported (PP > .95), except for the species *Conus consors*, either placed as the sister group of *C. magus* or of *C. achatinus* + *C. monachus*.

The best model of evolution for the toxin dataset is GTR+G ($\alpha = 0.61$). As predicted for a multi-gene family, the toxin tree does not match the species tree but actually includes several iterations of it (Fig. 2). By reconciling the gene and species trees, 53 duplications and 109 gene losses (D/L score = 188.5) were identified. As shown in Figure 2, three duplications gave rise to four major clades (PP > .98 for each), each including gene sequences found in the same species. For example, all four clades include at least one sequence from *C. striatus, C. stercusmuscarum* and *C. circumcisus*. The four clades are completely congruent with the differing Cys-patterns of the mature sequences. One clade includes sequences with a [---**C**$\mathbf{C}(X_{6\text{-}7})\mathbf{C}(X_2)\mathbf{C}X\mathbf{C}(X_3)\mathbf{C}$---] pattern, the next one sequences with a [---$\mathbf{C}\mathbf{C}(X_4)\mathbf{C}(X_7)\mathbf{C}$---] pattern, then sequences with a [---$\mathbf{C}\mathbf{C}(X_4)\mathbf{C}(X_4)\mathbf{C}$---] pattern, and the last sequences with a [---$\mathbf{C}\mathbf{C}(X_3)\mathbf{C}(X_5)\mathbf{C}$---] pattern. Furthermore, additional iterations of the species tree are embedded within three of the four lineages. As many nodes within each major clade were not supported (PP < .90), we used the rearrange mode implemented in Notung 2.6 to modify the topology of the tree at these nodes in order to minimize the number of duplications and losses in the tree. The new D/L score obtained was 101.5, with 41 duplications and 40 losses.

Several equally parsimonious roots, including the four main branches but also several intra-clades branches (22), were identified. The results are similar with the rearranged tree, even if the number of potential intra-clades roots is lower (only four).

As shown in Table 3, signal sequences are almost identical between the four major clades, pro-regions are more variable but still present several similar nucleotides, and mature regions are totally different and alignable only for the Cys sites. This pattern is also obvious when looking at the amino-acid sequences.

Positive Selection

The LRTs between different sites models (M1 vs M2 and M7 vs M8) support the presence of positively selected sites, a result found when all sequences are analysed together but also when each major clade is analysed separately (Table 4). These sites are identified (PP for BEB tests > .95) for each major clade as well as for the entire A-conotoxin dataset: most of them are located in the mature toxin and to a lesser extent, at the 3' end of the pro-region (Table 3). When the whole dataset is analysed, almost all the sites (except the cysteine sites) of the mature region have undergone positive selection.

Results obtained when comparing different branch models also confirm the presence of positive selection in these lineages for both datasets analysed (with the entire sequence and also with the mature region) (Tables 5 and 6). The likelihood value of the tree is significantly higher when the dN/dS ratio (ω) is not fixed at 0.5 or 1 (Table 6 – A-C and B-C comparisons). The estimated ω for the different lineages (corresponding to the four major clades, when analysed independently – Table 5, line D) range from 1.426 for the α3/5 subfamily to 3.656 for the α4/4 subfamily when the entire sequence is analysed, and from 1.846 for the α3/5 subfamily to ∞ for α4/4 and α4/7 subfamilies when considering only the mature toxin. However, the value obtained for the α4/4 subfamily is probably a biased estimation as only three sequences are included in this group. We performed the same analyses but removed the three α4/4 sequences from the dataset. The obtained results are highly similar: for example, the likelihood ratio statistic for the M1 and M2 models comparison is 136.2 (p-value=0). It should be noted that the results obtained for the entire dataset (Table 4, "A superfamily") might be questionable as the alignment of the mature sequence between each Cys is ambiguous between subfamilies.

We also tested if the strength of positive selection was different among the four major lineages. A model where all four ω are different is not better than a model where all ω are identical when the entire sequence is analysed, but better when only the mature toxin is

analysed (Tables 5 and 6). We also performed pairwise comparisons. To do so, we ranked all the subfamilies according to their estimated ω (from the highest to the lowest), and tested if the clade n has a significantly higher ω than the clade n+1: $\omega_{\alpha 4/7}$ and $\omega_{\alpha 4/4}$ are not significantly different, but superior to $\omega_{\kappa A}$ and $\omega_{\alpha 3/5}$ (using the mature sequence only). Both $\omega_{\kappa A}$ and $\omega_{\alpha 3/5}$ are significantly superior to 1 (except for $\omega_{\alpha 3/5}$ with the mature sequence).

The Pionoconus α4/4 clade and α4/4 conopeptides from other fish-hunting *Conus*

Three of the four major branches of the A-superfamily are well represented in all species of *Pionoconus* examined: the α3/5, κA, and α4/7 families. The α4/4 peptides comprise a small group, and these have not been extensively investigated. On the basis of molecular genetic criteria however, the α4/4 conopeptides clearly comprise a distinctive clade of A-superfamily gene sequences expressed in *Pionoconus*. We investigated whether this functionally undefined lineage is present in other fish-hunting *Conus* species outside the *Pionoconus* clade. All known α4/4 conopeptide sequences from fish-hunting species are shown in Table 7. Another phylogenetic tree was constructed (using the same methodology), comprising all the A-conotoxins from *Pionoconus,* but also including the A-conotoxins with α4/4 Cys pattern ([---**C**$\mathbf{C}(X_4)$**C**(X_4)**C**---]) shown in Table 7 (GenBank accession numbers: BD261436.1, BD261438.1, BD261439.1, BD261453.1, FB299972.1, FJ937346-FJ937350).

The non-*Pionoconus* sequences come from a closely related *Conus* clade (the "*Textilia* group" including *C. bullatus*, *C. cervus* and *C. kinoshitai*, see Figure 1) and from two more distant *Conus* species that belong to the *Chelyconus* clade (*C. ermineus* and *C. purpurascens*) (Espiritu et al. 2001; Duda and Palumbi 2004; Fig. 1). The results are shown in Figure 3. Most sequences from the *Textilia* group are closely related to the *Pionoconus* α4/4 (PP = .99), but sequences from *C. ermineus* and *C. purpurascens* are clustered in a non-related clade, whose relationship with other clades is not supported.

The most unexpected result was that one of the α4/4 conopeptides from *Conus bullatus*, Bu1.3 does not map with the other α4/4 sequences from *Pionoconus* and *Textilia*. Surprisingly, Bu1.3 is on the same branch as the α4/7 sequences from *Pionoconus*. Thus, on the basis of the mature peptide primary structure, Bu1.3 undoubtedly belongs to the α4/4 subfamily since it has the consensus Cys pattern of the subfamily. However, the phylogenetic analysis indicates that it belongs in the same branch as the α4/7 subfamily in *Pionoconus* (posterior probabilities = 1).

DISCUSSION

Congruency between phylogeny and cys-pattern

The data above demonstrate that there are four major groups of A-superfamily peptides found in *Pionoconus* species. These groups were previously recognized purely on the basis of the Cys pattern and the spacing between Cys residues in the mature toxin regions as families and subfamilies of A-conopeptides. We show that these are also coherent groups when evaluated using molecular phylogenetic criteria. The four classes of mature toxins belonging to the A-superfamily from *Pionoconus* (Table 3) are the κA-family, the α4/7-subfamily, α3/5-subfamily and the α4/4-subfamily. Each of these groups forms a distinctive clade on the phylogenetic tree of toxin sequences, with long branches between clades and high posterior probabilities. It is important to note that the four classes remain well defined (but slightly less supported) even when the mature toxin regions are deleted from the aligned sequences used to assemble the tree (results not shown). Our findings clearly demonstrate that all the *Pionoconus* A-superfamily sequences sharing a common Cys pattern are derived from a common ancestor (but see below for a discussion on the Bu1.3 sequence).

Furthermore, there is structural conservation within subfamilies, juxtaposed with significant divergence between subfamilies. The sequence conservation within subfamilies

can be useful if one needs to analyse only one subfamily and exclude the others. We can thus design a potential PCR primer sequence for each subfamily that should allow specific amplification (see Table 3).

Duplication

Our results also highlight two major evolutionary forces that shaped the pattern observed for the A-superfamily conotoxins from *Pionoconus*. First, the finding that in most of the branches of the tree shown in Figure 2, there are representatives from most species of *Pionoconus* suggests that three major duplication events gave rise to these discrete groups of peptide toxins. These major duplication events occurred before most speciation events that generated the different species in the *Pionoconus* clade. In order to more accurately date these duplication events, we analysed the A-conotoxin sequences from GenBank. Our survey revealed that the κA subfamily is restricted to the *Pionoconus* clade. The α3/5-subfamily is found almost exclusively in fish-hunting species (except one sequence found in *C. betulinus*, a worm-hunting species). Similarly, the α4/4-subfamily is restricted to fish-hunting species, except for one species (*C. quercinus*, another worm-hunting species). Only the α4/7-subfamily is found in numerous fish, worm and mollusc-hunting species. These findings would suggest that the duplication events that led to the appearance of the κA, α3/5 and α4/4 took place just after the separation between the fish-hunting species on one hand and the worm and mollusc-hunting species on the other hand. The sporadic presence of α3/5 and α4/4 in a worm-hunting species needs to be verified and explained; the possibility of contamination during the experiments (as only one non-fish-hunting-species' sequence was found so far in both cases) or hybridization between species need to be investigated.

However, other hypotheses can be proposed to explain the restriction of most of the A-conotoxin clades to the fish-hunting species. For example, the silencing of some genes

(Conticello et al. 2001; Duda 2008) in the non-fish-hunting species would explain why only α4/7 conotoxins were found in some of these species. It is also important to note that a significantly greater effort in the definition of A-conotoxins has been carried out in the *Pionoconus* clade compared to some of the other species clades in *Conus*: the lack of A-conotoxin subfamilies in non-fish-hunting clades could thus be an artifact due to biased sampling.

It is noteworthy that each major clade defined in the *Pionoconus* conotoxins' tree may have evolved a different function (Figure 2; see also the introduction), an observation potentially congruent with the classical model of duplication followed by neofunctionalization (Ohno 1970). Additional rounds of duplication certainly occurred within major branches, highlighted by the presence of several well supported clades, many of them including sequences from the same species. However, it is impossible to know if these different groups of paralogs within subfamilies results from common duplication events between subfamilies or if each subfamily has undergone a series of independent duplication events. Other analyses, such has genome mapping, in order to determine the relative position of the different loci should be helpful in distinguishing between the different hypotheses. However, based only on the well-supported clades, an estimate is obtained of about a dozen different A-superfamily genes in *Pionoconus*; when even the poorly supported clades are included, the estimate increases to >40, as determined by Notung 2.6.

What is not established from the analysis of the data is the order in which the duplication events may have taken place. Five alternative rooted trees can be proposed from the unrooted tree (Fig. 4), all of them being equally parsimonious regarding the number of duplications and gene losses. Thus, the results presented above do not allow us to favor one scheme over another. From the first four scenarios (Fig. 4A-D), we can infer that three duplication events occurred. From the last scenario (Fig. 4E), we can infer two or three

duplication events since the duplication of the ancestor of the κA-family and the α4/7-subfamily in one hand, and of the α3/5 and the α4/4-subfamilies in the other hand may correspond to only one duplication event. The latter hypothesis makes predictions regarding the respective position of these genes in the genome.

It has been argued elsewhere that because the α4/7-subfamily of peptides is the most widely distributed of all of the groups in the A-superfamily, it is likely to be the first group from which all other subfamilies are derived (Santos et al. 2004). If this were the case, then the scheme in Figure 4B would be preferred over the other alternatives; this would suggest that the first duplication event separated α4/7-subfamily from all the others, and a second duplication event gave rise to the κA-conotoxins and a gene that was subsequently further duplicated to ultimately generate the α4/4- and the α3/5-subfamilies.

Positive selection

The second major force that influenced the evolution of the A-superfamily conotoxins is positive selection, found in all subfamilies. The sites that have undergone positive selection are mostly located in the mature toxin. However, p-values obtained with the branch model are higher than those obtained by comparing different sites models. This result can be explained by the fact that branch models test for positive selection on all sites: as only some sites of the toxin sequence are positively selected, the presence of sites under neutral or purifying selection could decrease the significance of the branch models comparison even when only the mature sequence is analysed. On the other hand, several authors (e.g. Hughes and Friedman 2008) have shown that positive selection can be detected with likelihood approaches only because of stochastic mutations among branches.

In the A-superfamily, positive selection, although widespread, does not act equally in all subfamilies: the κA and α3/5 subfamilies present a significantly lower level of positive

selection than α4/4 and α4/7 subfamilies, especially for the mature toxin region. These differences between subfamilies could be characteristic of a dynamic system, where the appearance of new genes, followed by positive selection leads to the appearance of new function; conversely, some copies will retain the "ancestral" function and thus will not be subject to strong positive selection.

This is clearly a major force in the evolution of conotoxins, as has already been reported in other families (Duda and Palumbi 2000; Conticello et al. 2001; Duda 2008). As proposed previously, positive selection in conotoxins can of course be linked to the rapid diversification of the group: most species of *Conus* are included in the "major clade," whose diversification took place during the Miocene. Such diversification may be the result of species adaptation to new prey, enhanced by the rapid evolution through duplication and positive selection of conotoxins, as illustrated here by the analysis of the A-superfamily.

The α4/4 Cys-pattern

Finally, the clear implication of the analysis of the α4/4 sequences is that there are at least two different genes that give rise to the α4/4 conopeptides in *Conus bullatus*. An examination of the tree in Figure 3 suggests that there is likely to have been an additional duplication in *Conus bullatus*, and that three genes gave rise to the spectrum of *Conus bullatus* α4/4 sequences defined here. Several hypotheses can be proposed to explain the presence of α4/4 conopeptides in different clades. First, the α4/4 pattern could constitute the ancestral Cys-pattern that first diverged in several toxins with a α4/4 pattern but different signal and propeptide sequences, which then gave rise to all the other described subfamilies. A second hypothesis would involve recombination events between different genes, resulting in a sequence with a α4/4 pattern for the mature sequence but a signal and propeptide sequences similar to the α4/7 ones (as for the Bu1.3 toxin). Such events have already been

reported in literature for multi-gene families including toxin genes (e.g. Dolley 2008). Finally, convergent evolution, although not common between closely related genes within a single species, may also explain the results obtained. Toxins with similar signal and propeptide sequences could evolve different Cys-patterns, some of them being independently acquired in other clades.

Actually, although the discovery that Bu1.3 was in an entirely different branch of the phylogenetic tree was both unexpected and surprising, the available data on the mature peptide toxin is functionally consistent with this branch assignment. The mature peptide designated α-conotoxin Bu1A (identical to Bu1.3) has been extensively characterized. It is targeted to neuronal subtypes of the nicotinic acetylcholine receptor and has been productively used to differentiate between neuronal nicotinic receptors that have a β2 vs. a β4 subunit (Azam et al. 2005). Thus, the targeting of α-Bu1A is consistent with the only other peptide in this clade whose function is known, α-conotoxin MII, which is also targeted to neuronal nicotinic receptors (albeit with a different subtype preference from α-Bu1A). This raises the intriguing possibility that the entire α4/7 conopeptide subfamily in *Pionoconus* is targeted to various neuronal nicotinic acetylcholine receptor subtypes.

CONCLUSION

These findings illustrate how a phylogenetic perspective provides insights that are not at all obvious from primary amino acid sequences alone. The divergent evolutionary origins of different *Conus bullatus* α4/4 subfamily peptides could never have been discerned from the primary structure; it was the phylogenetic analysis that has elucidated the richness and diversity of the evolutionary history of the α4/4 subfamily peptides.

The evolutionary origine of most conopeptides superfamilies has not been determined, but the available data suggests that the A-gene superfamily is a relatively recent innovation compared to gene superfamilies of peptide toxins in most venomous animals. This perception rises from two factors: the genus *Conus* itself is evolutionarily more recent than other venomous lineages. The first adaptative radiation of *Conus* occured in the Eocene (Kohn 1990); scorpion, spiders and venomous snakes all appear earlier in the fossil record. Even within the family Conoidea however, there is evidence that the A-gene superfamily is a relatively recent innovation: other gene superfamilies (e.g., the I2 superfamily and the O-superfamily) are distributed more broadly across the superfamily Conoidea (Watkins et al. 2006). So far, the A-superfamily conopeptides have been found only within the genus *Conus*.

The phylogenetic approach that we employed to analyze the conotoxins of the A-superfamily within *Pionoconus* has provided insight into the pattern of evolution of this multigenic family. Several duplication events have resulted in the appearance of new gene copies that evolved different functions, each under positive selection. Furthermore, we have shown that the major lineages correspond to previously defined groups of toxins within the A-superfamily, sharing a particular Cys-pattern and general function. The function of one group, the α4/4 subfamily, has not been precisely characterized; its separate evolutionary history raises the possibility that a novel function has evolved that is divergent from the three other lineages. The use of molecular phylogenetic criteria for the identification of toxins with potentially novel functions was previously suggested (Olivera 2006; Olivera and Teichert, 2007) as one component of a phylogenetically based “concerted discovery” strategy.

ACKNOWLEDGMENTS

This work was supported by NIH Program Project grant GM48677 (to BMO). We are pleased to thank Jon Seger, Nicole Kraus, Naoko Takezaki and two anonymous reviewers for constructive comments on a previous version of the manuscript.

LITERATURE CITED

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

Azam L, Dowell C, Watkins M, Stitzel JA, Olivera BM, McIntosh JM (2005) Alpha-conotoxin BuIA, a novel peptide from *Conus bullatus*, distinguishes among neuronal nicotinic acetylcholine receptors. J Biol Chem 280:80-87

Conticello SG, Gilad Y, Avidan N, Ben-Asher E, Levy Z, Fainzilber M (2001) Mechanisms for Evolving Hypervariability: The Case of Conopeptides. Mol Biol Evol 18:120-131

Conticello SG, Pilpel Y, Glusman G, Fainzilber M (2000) Position-specific codon conservation in hypervariable gene families. Trends Genet 16:57-59

Doley R, Pahari S, Mackessy SP, Manjunatha Kini R (2008) Accelerated exchange of exon segments in Viperid three-finger toxin genes (*Sistrurus catenatus edwardsii*; Desert Massasauga). BMC Evol Biol 8:196

Duda TF (2008) Differentiation of Venoms of Predatory Marine Gastropods: Divergence of Orthologous Toxin Genes of Closely Related *Conus* Species with Different Dietary Specializations. J Mol Evol 67:315-321

Duda TF, Kohn AJ (2005) Species-level phylogeography and evolutionary history of the hyperdiverse marine gastropod genus *Conus*. Mol Phylogenet Evol 34:257-272

Duda TF, Palumbi SR (1999) Molecular genetics of ecological diversification: Duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. Proc Natl Acad Sci USA 96:6820-6823

Duda TF, Palumbi SR (2000) Evolutionary Diversification of Multigene Families: Allelic Selection of Toxins in Predatory Cone Snails. Mol Biol Evol 17:1286-1293

Duda TF, Palumbi SR (2004) Gene expression and feeding ecology: evolution of piscivory in the venomous gastropod genus Conus. Proc Roy Soc Lond Ser B Biol Scis 271:1165-1174

Duda TF, Remigio A (2008) Variation and evolution of toxin gene expression patterns of six closely related venomous marine snails. Mol Ecol 17:3018-3032

Durand D, Halldorsson BV, Vernot B (2006) A Hybrid Micro-Macroevolutionary Approach to Gene Tree Reconstruction. Journal of Computational Biology 13:320-335

Espiritu DJD, Watkins M, Dia-Monje V, Cartier GE, Cruz LE, Olivera BM (2001) Venomous cone snails: molecular phylogeny and the generation of toxin diversity. Toxicon 39:1899-1916

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41:95-98

Huelsenbeck JP, Ronquist F, Hall B (2001) MrBayes: bayesian inference of phylogeny. Bioinformatics 17:754-755

Hughes AL, Friedman R (2008) Codon-based tests of positive selection, branch lengths, and the evolution of mammalian immune system genes. Immunogenetics 60:495-506

Keane TM, Creevey CJ, Pentony MM, Naughton TJ, McInerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. BMC Evol Biol 6:1-17

Kohn AJ (1990) Tempo and mode of evolution in Conidae. Malacologia 32:55-67.

Ohno S (1970) Evolution by Gene Duplication. Springer-Verlag, Berlin

Olivera BM (2006) *Conus* Peptides: Biodiversity-based Discovery and Exogenomics. J Biol Chem 281:31173-31177

Olivera BM (2008) Venom peptides from *Conus* and other Conoideans : prospects and perspectives. In: Benoit E, Goudey-Perrière F, Marchot P, Servent D (eds) Toxines et fonctions cholinergiques neuronales et non neuronales. Librairie Lavoisier, Cachan

Olivera BM, Teichert RW (2007) Diversity of the neurotoxic *Conus* peptides: a model for concerted pharmacological discovery. Molecular Interventions 7:251-260

Santos AD, Mc Intosh JM, Hillyard DR, Cruz LE, Olivera BM (2004) The A-superfamily of conotoxins: structural and functional divergence. J Biol Chem 279:17596-17606

Terlau H, Shon KJ, Grilley M, Stocker M, Stühmer W, Olivera BM (1996) Strategy for rapid immobilization of prey by a fish-hunting cone snail. Nature 381:148-151

Vernot B, Stolzer M, Goldman A, Durand D (2008) Reconciliation with non-binary species trees. Journal of Computational Biology 15:981-1006

Wang Q, Jiang H, Hana Y-H, Yuan D-D, Chi C-W (2008) Two different groups of signal sequence in M-superfamily conotoxins. Toxicon 51 813-822

Watkins M, Hillyard DR, Olivera BM (2006) Genes Expressed in a Turrid Venom Duct: Divergence and Similarity to Conotoxins. Journal of Molecular Evolution 62:247-256.

Wong WSW, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168:1041-1051

Woodward SR, Cruz LJ, Olivera BM, Hillyard DR (1990) Constant and hypervariable regions in conotoxin propeptides. EMBO 9:1015-1020

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573

Yang Z (2002) Inference of selection from multiple species alignments. Curr Opin Genetics Dev 12:688-694

Yang Z (2006) Computational Molecular Evolution. Oxford University Press, Oxford

Yang Z (2007) PAML 4: a program package for phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496-503

Yang Z, Nielsen R (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46:409-418

Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. Mol Biol Evol 15:1600-1611

Yang Z, Wong WSW, Nielsen R (2005) Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. Mol Biol Evol 22:1107-1118

Zafaralla GC, Ramilo C, Gray WR, Karlstrom R, Olivera BM, Cruz LJ (1988) Phylogenetic specificity of cholinergic ligands: a-conotoxin SI. Biochemistry 27:7102-7105

Table 1: GenBank accession numbers for COI, 12S, 16S and Calmodulin (intron) gene.

		COI	12S	16S	Calmodulin
Pionoconus	***achatinus***	FJ868109	FJ868042	FJ868053	
	aurisiacus	FJ868111	EU682276.1	EU078943.1	
	catus	FJ868113	EU682278.1	FJ868055	AF113260.1
	circumcisus	FJ868114	FJ868045	EU078942.1	
	consors	FJ868115	EU682279.1	EU078940.1	AF113267.1
	gauguini	FJ868117	FJ868047	EU078944.1	
	magus	FJ868118	FJ868048	EU078939.1	AF113288.1
	monachus	FJ868120	FJ868050	EU078938.1	
	stercusmuscarum	EU733518	EU682294.1	EU078941.1	AF113310.1
	striatus	FJ868121	FJ868051	EU078945.1	AF113311.1
	striolatus	FJ868122	FJ868052	FJ868058	AF113312.1
Textilia	***bullatus***	FJ937338	FJ937334	FJ937342	
	cervus	FJ937339	FJ937335	FJ937343	
	kinoshitai	FJ937341	FJ937337	FJ937345	
Outgroups	***arcuatus***	FJ868110	FJ868043	FJ868054	AY382036.1
	californicus	FJ868112	FJ868044	AF036534	AY382040.1
	ebraeus	FJ868116	FJ868046	FJ868056	AF113272.1
	ermineus	FJ937340	FJ937336	FJ937344	
	mahogani	FJ868119	FJ868049	FJ868057	AY382050.1
	purpurascens			AF480308	AF480311
	textile	EU812758.1	EU682296.1	EU078936.1	AF113316.1

Table 2: List of A-conotoxin sequences included in the analysis.

Species	Toxin ID	Source	Genbank # 1	Clade
achatinus	Ac1.1	cDNA	BD394973.1	α3/5
achatinus	Ac1.2	cDNA	BD261435.1	α3/5
achatinus	Ac1.3	cDNA	BD394979.1	α3/5
achatinus	Ac1.4	cDNA	BD261468.1	α4/7
achatinus	Ac1.5	cDNA	BD261469.1	α4/7
achatinus	Ac1.6	cDNA	BD261470.1	α4/7
achatinus	Ac1.8	cDNA	BD394981.1	α3/5
achatinus	Ac4.1	cDNA	FJ868059	κ
achatinus	Ac4.2	cDNA	FJ868060	κ
achatinus			BD394972.1	α3/5
achatinus			BD394980.1	α3/5
achatinus			DQ311072.1	α3/5
achatinus			DQ359138.1	α3/5
achatinus			DQ359139.1	α3/5
aurisiacus	A1.1	cDNA	BD394982.1	α3/5
aurisiacus	A1.1a	cDNA	BD394983.1	α3/5
aurisiacus	A1.2	cDNA	BD261478.1	α4/7
aurisiacus	A1.3	cDNA	BD261479.1	α4/7
aurisiacus	A1.4	cDNA	FJ868061	α3/5
aurisiacus	A4.1	cDNA	FJ868062	κ
aurisiacus	A4.2	cDNA	FJ868063	κ
aurisiacus	A4.3	Genomic DNA	FJ868064	κ
aurisiacus	A4.4	Genomic DNA	FJ868065	κ
catus	C1.2	Genomic DNA	BD261484.1	α4/7
catus	C1.3	Genomic DNA	BD261485.1	α4/7
catus	C4.1a	cDNA	FJ868066	κ
catus	C4.1b	cDNA	FJ868068	κ
catus	C4.2	cDNA	FJ868067	κ
catus	Cl	cDNA	FJ868069	α3/5
circumcisus	Cr1.1	cDNA	BD394977.1	α3/5
circumcisus	Cr1.2	cDNA	BD261426.1	α4/7
circumcisus	Cr1.3	cDNA	BD261427.1	α4/4
circumcisus	Cr1.4	Genomic DNA	FJ868070	α4/7
circumcisus	Cr1.5	Genomic DNA	FJ868071	α4/7
circumcisus	Cr1.6	Genomic DNA	FJ868072	α4/7
circumcisus	Cr1.7	Genomic DNA	FJ868073	α4/7
circumcisus	Cr4.1	cDNA	FJ868074	κ
circumcisus	Cr4.2	Genomic DNA	FJ868075	κ
consors	Cn1.1	cDNA	BD261416.1	α3/5
consors	Cn1.2	cDNA	BD261442.1	α4/7
consors	Cn1.3	cDNA	FJ868076	α3/5
consors	Cn1.4	Genomic DNA	FJ868077	α3/5
consors	Cn1.5	Genomic DNA	FJ868078	α3/5
consors	Cn1.6	Genomic DNA	FJ868079	α4/7
consors	Cn4.1	cDNA	FJ868080	κ
consors	Cn4.2	cDNA	FJ868081	κ
consors	Cn4.3	Genomic DNA	FJ868082	κ
consors	Cn4.4	cDNA	FJ868083	κ
consors			BD394975.1	α3/5
gauguini	Ga1.1	cDNA	FJ868084	α4/7
gauguini	Ga1.2	cDNA	FJ868085	α3/5
magus	M1.1	cDNA	BD394984.1	α3/5
magus	M1.3	Genomic DNA	BD394985.1	α3/5

magus	M1.4	Genomic DNA	BD394986.1	α3/5
magus	M1.5	Genomic DNA	BD394987.1	α3/5
magus	M4.2	cDNA	FJ868086	κ
magus	M4.3	cDNA	FJ868087	κ
magus	Mg1	cDNA	BD261395.1	α4/7
magus	MVIII	cDNA	FJ868088	κ
monachus	Mn1.3	Genomic DNA	FJ868089	α3/5
monachus	Mn1.4	Genomic DNA	FJ868090	α3/5
monachus	Mn1.5	Genomic DNA	FJ868091	α3/5
monachus	Mn1.6	Genomic DNA	FJ868092	α4/7
monachus	Mn4.1	cDNA	FJ868093	κ
monachus	Mn4.2	cDNA	FJ868094	κ
monachus	Mn4.3	Genomic DNA	FJ868095	κ
monachus	MnI	cDNA	BD394976.1	α3/5
monachus	MnII	cDNA	BD394971.1	α3/5
stercusmuscarum	Sm1.1	cDNA	BD394966.1	α3/5
stercusmuscarum	Sm1.3	cDNA	BD261425.1	α4/4
stercusmuscarum	Sm1.5	Genomic DNA	BD261522.1	α4/7
stercusmuscarum	Sm4.2	cDNA	FJ868096	κ
stercusmuscarum	SmI	cDNA	BD261417.1	α4/7
stercusmuscarum	SmVIII	cDNA	FJ868097	κ
stercusmuscarum	SmVIIIA	cDNA	FJ868098	κ
striatus	S1.1	cDNA	BD261403.1	α4/4
striatus	S1.10a	Genomic DNA	FJ868099	α3/5
striatus	S1.10b	cDNA	FJ868100	α3/5
striatus	S1.11	cDNA	BD394962.1	α3/5
striatus	S1.12	cDNA	BD394967.1	α3/5
striatus	S1.3	Genomic DNA	BD394989.1	α3/5
striatus	S1.4	Genomic DNA	FJ868101	α3/5
striatus	S1.5	Genomic DNA	BD261523.1	α4/7
striatus	S1.6	Genomic DNA	FJ868102	α3/5
striatus	S1.7	Genomic DNA	FJ868103	α3/5
striatus	S1.8	cDNA	FJ868104	α3/5
striatus	S1.9	cDNA	FJ868105	α3/5
striatus	SVIII	cDNA	FJ868107	κ
striatus	SVIIIA	cDNA	FJ868106	κ
striatus			AY157497.1	α3/5
striatus			AY166873.1	κ
striolatus	Sx4.1	cDNA	FJ868108	κ

Provided for each sequence are the species, the toxin ID and the source from which the sequence was obtained (only for sequences from our laboratory), the GenBank accession number and the clade assignation (cf. Fig. 2).

Table 3: Comparisons of sequences from the four subfamilies of A-conotoxins.

Signal

```
κA     ATGGGCATGCGGATGATGTTCACCGTGTTTCTGTTGGTTGTCTTGGCAACCACTGTCGTTTCC
α4/7   ---------------.................................................
α4/4   ---------------.................................................
α3/5   .........................................A.......................
```

Pro-region

```
κA     ATCCCTTCAGATCGTGCATCTGATGTCAGGAATGCCGCAGTCCACGAGAGA
α4/7   T.....................G.............C.A....C.A.GCGTCTGACGTGATCACGCTGGCCCTCAAG
α4/4   T..............A........G.GC.....A..A..C..G.ACCGACGAGCCTGAGGAGCACGGACCGGACAGG
α3/5   T...........A............G....G...A..A..C.A.A..CGA.AGGTCTGACATGCACGAATCGGACCGGAATGGACGC
```

Mature toxin

```
κA     CAGAAGGAGCTGGTCGTTACGGCCACCACGACTTGCTGTGGTTATAATCCGATGTCAATGTGCCCTAAATGCATGTGCACTTATTCCTGTCCCCACCAAAAGAAGAAA
α4/7   ---------------------------GGA---......TCCA.CCC.GTC----------..T.ACTTGGAGCAT.CA.ACCT.---...GGTAGAAG.CGC------
α4/4   ------------------------.ATGGA---......A.GA..CC.G.C----------..T---G.GAG.CACA.A----------...GGT--------------
α3/5   ---------------------------GGATGC..T..CAA.CC.GCC------------..TGGCCC.AA.TATG.T----------...GG.AC.TC.TGCTCC.GG
κA     AGACCAGGCCGCAGAAACGAC
α4/7   ---------------------
α4/4   ---------------------
α3/5   ACCATC---------------
```

Entire amino-acid sequence

```
All
κA     MGMRMMFTVFLLVVLATTVVS  IPSDRASDVRNAAVHER-----------  QKELVVTATTTCCGYNPMSMCPKCMCTYSCPHQKKKRPGRRND
α4/7   -----.............  F.......G....ANDKASDVITLALK--  ---------G-..SNPV---.HLEHSNL-.GRRR---------
α4/4   -----.............  F....E..GA.DEARTDEPEEHGPDR---  ---------NG..RNPA---.ESHR-----.G------------
α3/5   .....T.....T.....  F...S...G.DDEAKDERSDMHESDRNGR  ---------GC..-NPA---.GPNYG---.GTSCSRTI-----
```

Sequences for κA, α4/7, α4/4 and α3/5 are respectively C4.2, Cn1.2, Sm1.3 and S1.10b. Cys pattern are in bold; positively selected sites are shaded (PP of BEB analysis > .95 with either M1/M2 or M7/M8 comparison); potential specific primers for each subfamily are underlined. "All": result of the site model analysis performed with all the sequences.

Table 4: Site model analyses.

	M1a/M2a	M7/M8
κA	85.9**	86.76**
α4/7	36.26**	35.72**
α3/5	39.34**	41.4**
α4/4	9.86**	9.86**
A superfamily	143.3**	153.2**

Likelihood ratio statistics comparison between M1/M2 and M7/M8, with degree of freedom (d.f.) = 2. A superfamily: analysis of the complete dataset. **: p-value < .01

Table 5: Branch model analyses.

		Entire sequence					Mature sequence only				
	Model	ℓ	$\omega_{\kappa A}$	$\omega_{\alpha 4/7}$	$\omega_{\alpha 4/4}$	$\omega_{\alpha 3/5}$	ℓ	$\omega_{\kappa A}$	$\omega_{\alpha 4/7}$	$\omega_{\alpha 4/4}$	$\omega_{\alpha 3/5}$
A	$\omega_{\kappa A} = \omega_{\alpha 4/7} = \omega_{\alpha 4/4} = \omega_{\alpha 3/5} = 0.5$	-3487.03	0.5	0.5	0.5	0.5	-2117.55	0.5	0.5	0.5	0.5
B	$\omega_{\kappa A} = \omega_{\alpha 4/7} = \omega_{\alpha 4/4} = \omega_{\alpha 3/5} = 1$	-3435.31	1	1	1	1	-2080.65	1	1	1	1
C	$\omega_{\kappa A} = \omega_{\alpha 4/7} = \omega_{\alpha 4/4} = \omega_{\alpha 3/5}$	-3424.44	1.734	1.734	1.734	1.734	-2068.62	2.208	2.208	2.208	2.208
D	$\omega_{\kappa A}, \omega_{\alpha 4/7}, \omega_{\alpha 4/4}, \omega_{\alpha 3/5}$	-3422.97	1.628	2.183	3.656	1.426	-2062.64	2.016	∞	∞	1.846
E	$\omega_{\kappa A} = \omega_{\alpha 3/5}, \omega_{\alpha 4/7} = \omega_{\alpha 4/4}$	-3423.34	1.555	2.492	2.492	1.555	-2062.66	1.966	∞	∞	1.966
F	$\omega_{\kappa A} = \omega_{\alpha 3/5}, \omega_{\alpha 4/7}, \omega_{\alpha 4/4}$	-3423.06	1.559	2.186	3.643	1.559	-2062.66	1.966	∞	∞	1.966
G	$\omega_{\alpha 4/4} = \omega_{\alpha 3/5}, \omega_{\kappa A} = \omega_{\alpha 4/7}$	-3424.44	1.734	1.734	1.732	1.732	-2067.81	2.358	2.358	1.833	1.833
H	$\omega_{\alpha 4/4} = \omega_{\alpha 3/5}, \omega_{\kappa A}, \omega_{\alpha 4/7}$	-3424.11	1.612	2.164	1.719	1.719	-2063.81	2.019	∞	1.824	1.824
I	$\omega_{\alpha 4/4} = \omega_{\alpha 4/7}, \omega_{\kappa A} = \omega_{\alpha 3/5}$	-3423.27	1.543	2.475	2.475	1.543	-2062.69	1.902	∞	∞	1.902
J	$\omega_{\alpha 4/4} = \omega_{\alpha 4/7}, \omega_{\kappa A}, \omega_{\alpha 3/5}$	-3423.18	1.607	2.482	2.482	1.416	-2062.51	2.020	∞	∞	1.601
K	$\omega_{\alpha 4/4} = \omega_{\alpha 4/7} = \omega_{\alpha 3/5}, \omega_{\kappa A} = 1$	-3428.45	1	1.861	1.861	1.861	-2074.51	1	2.593	2.593	2.593
L	$\omega_{\alpha 4/4} = \omega_{\alpha 4/7} = \omega_{\alpha 3/5}, \omega_{\kappa A}$	-3424.28	1.617	1.861	1.861	1.861	-2067.8	2.026	2.611	2.611	2.611
M	$\omega_{\alpha 4/4} = \omega_{\alpha 4/7} = \omega_{\kappa A}, \omega_{\alpha 3/5} = 1$	-3426.56	1.662	1.662	1.662	1	-2069.29	2.434	2.434	2.434	1
N	$\omega_{\alpha 4/4} = \omega_{\alpha 4/7} = \omega_{\kappa A}, \omega_{\alpha 3/5}$	-3423.52	1.661	1.661	1.661	3.628	-2067.73	2.4	2.4	2.4	1.749

Log likelihood values (ℓ) and dS/dN (ω) estimates under different models (lettered from A to N). An infinite value (∞) can be obtained when there are no synonymous substitutions.

Table 6: LRTs between the different models of the branch-model analysis.

Models compared	$2\Delta\ell$ (entire seq.)	$2\Delta\ell$ (mature seq.)	d.f.
A-C	125.18**	97.86**	1
B-C	21.74**	24.06**	1
C-D	2.94	11.96**	3
E-F	0.56	0	1
G-H	0.66	8**	1
I-J	0.18	0.36	1
K-L	8.34**	13.42**	1
M-N	6.08*	3.12	1

Likelihood ratio statistics ($2\Delta\ell$) for hypotheses testing. d.f. : degree of freedom. *: p-value < .05; **: p-value < .01.

Table 7: Amino-acid sequence for the mature toxin regions of all α4/4 sequences analysed.

bullatus_Bu1.1	PGCCNNPACVKHRCG
bullatus_Bu1.2	PGCCNNPACVKHRCGG
bullatus_Bu1.3	KGCCSTPPCAVLYCGRRR
bullatus_Bu1.4	NGCCWNPSCPRPRCTGRR
cervus_Cs1.2	PGCCNNPACGANRCG
circumcisus_Cr1.3*	NGCCGNPDCTSHSCD
kinoshitai_Kn1.2	PGCCNNPACVKHRCG
kinoshitai_Kn1.3	PGCCNNPACGKNRC
ermineus_E1.3A	PGCCWNPACVKNRCGRR
ermineus_E1.3B	PGCCWNPACVKNRCGRR
purpurascens_P1.7	PGCCRHPACGKNRCGR
stercusmuscarum_Sm1.3*	NGCCRNPACESHRCG
striatus_S1.1*	NGCCRNPACESHRCG

The sequences marked by an asterisk are the *Pionoconus* sequences. Although some of the mature toxin sequences shown are identical (e.g., Sm1.3 and S1.1; Bu1.1 and Kn1.2) there is significant divergence in the other precursor regions.

Figure caption:

Figure 1: Species tree obtained with the COI, 16S, 12S and calmodulin genes (Bayesian analysis). Posterior probabilities are indicated for each node. Each species of the *Pionoconus* clade is illustrated. I: first *Conus* radiation, Early Eocene, 55-45 MY (e.g. Espiritu 2001); II: second *Conus* radiation, Miocene, 20-10 MY (Duda and Palumbi 1999).

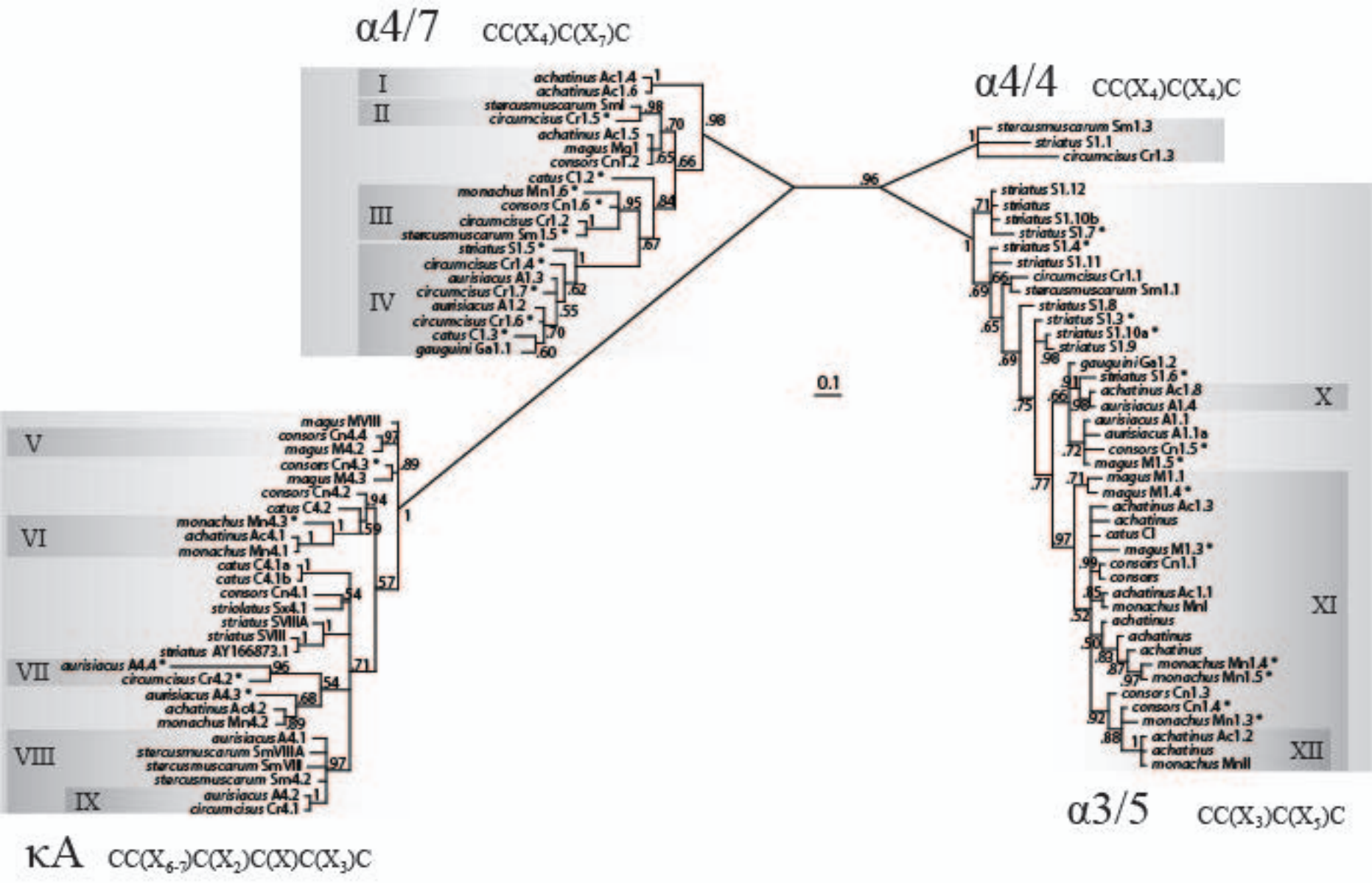
Figure 2: Unrooted bayesian tree inferred from the A-conotoxins. Posterior probabilities are indicated for each node. Four subfamilies are delimited corresponding to four different mature toxins (see text). Within each subfamily, several highly supported clades (PP > .95) are shaded, numbered from I to XII. *: sequences obtained from genomic sequencing; other sequences from our lab were obtained from cDNA libraries.

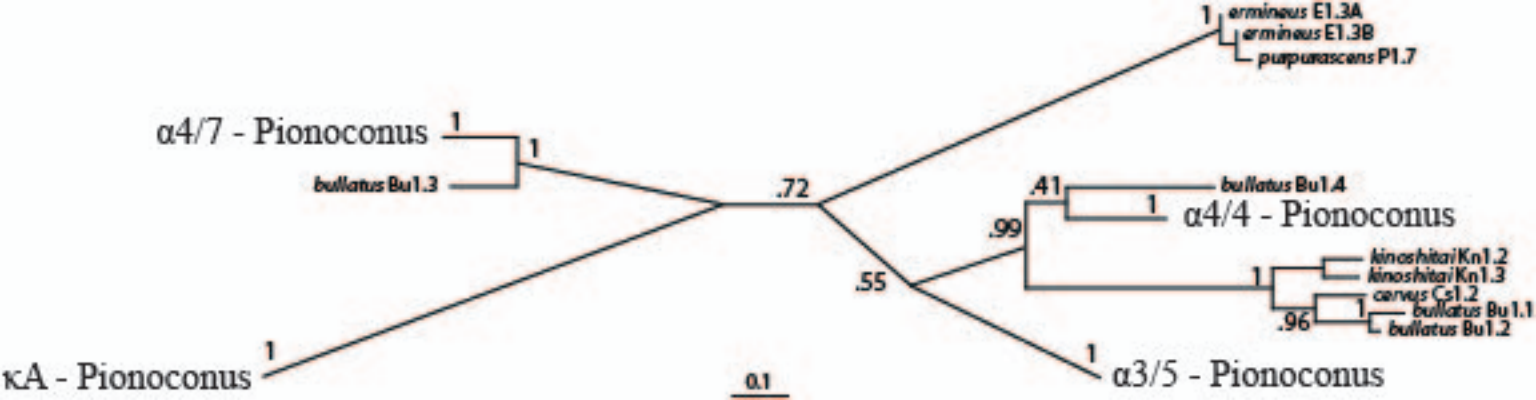
Figure 3: Unrooted bayesian tree including the A-conotoxins from the *Pionoconus* clade and the α4/4 from the *Textilia* group (*C. bullatus*, *cervus* and *kinoshitai*) and from the *Chelyconus* clade (*C. ermineus* and *C. purpurascens*). Posterior probabilities are indicated for each node. Details are not given for the four *Pionoconus* clades already described in Figure 2.

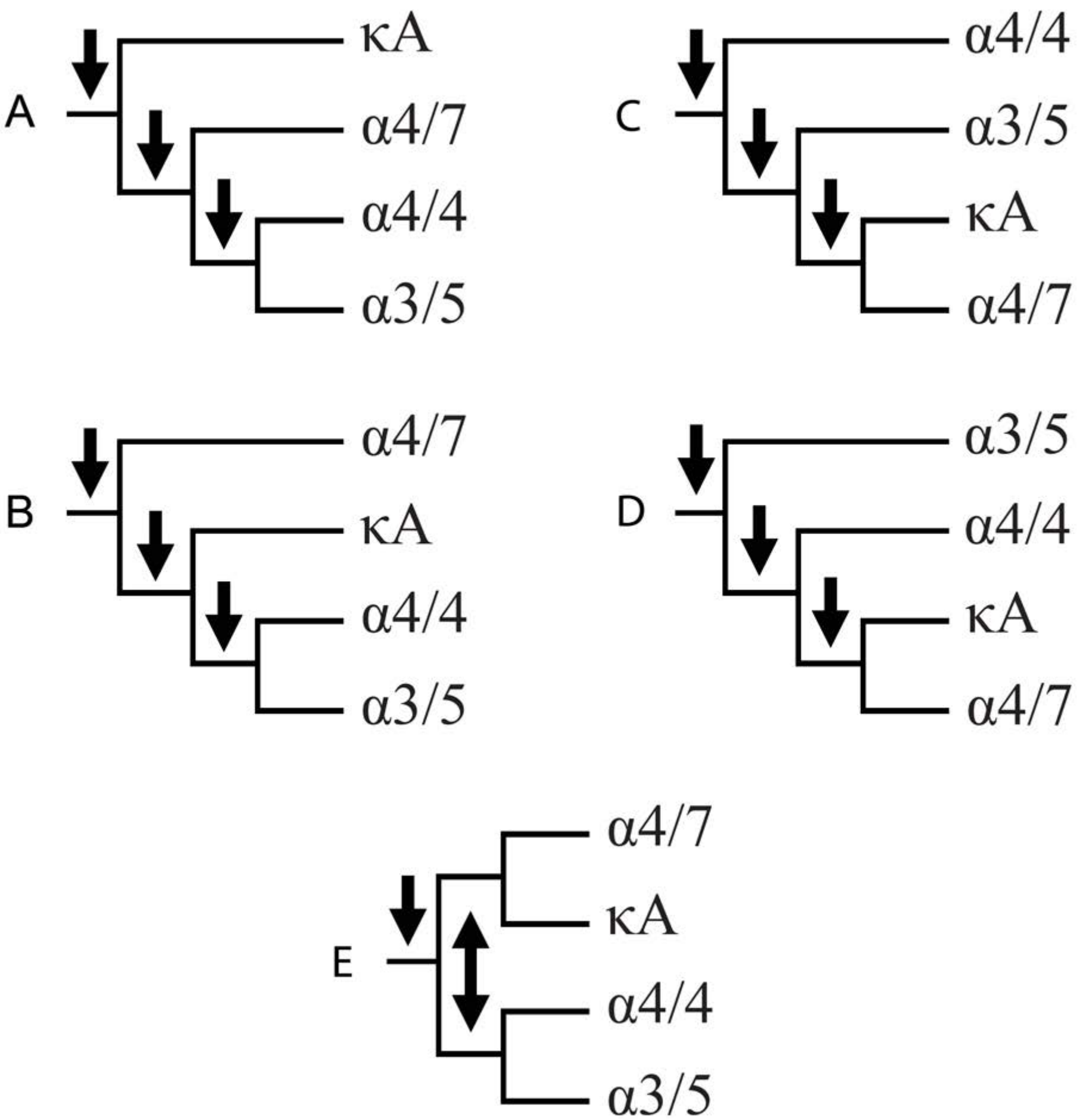
Figure 4: Alternative scenarios for the A superfamily evolution. Black arrow: potential duplication events. In the E scenarion, only two duplication events might have occurred.

californicus
arcuatus
mahogani
ebraeus
ermineus
purpurascens
textile
kinoshitai
bullatus
cervus
striatus
stercusmuscarum
circumcisus
aurisiacus
gauguini
striolatus
catus
magus
consors
achatinus
monachus
1
.95
.96
.92
.82
0.1
I
II
Textilia
Pionoconus

C. achatinus C. aurisiacus C. catus C. circumcisus
C. consors C. gauguini C. magus C. monachus
C. stercusmuscarum C. striatus C. striolatus


α4/7 CC(X4)C(X7)C
α4/4 CC(X4)C(X4)C
α3/5 CC(X3)C(X5)C
κA CC(X6-7)C(X2)C(X)C(X3)C
I
II
III
IV
V
VI
VII
VIII
IX
X
XI
XII
achatinus Ac1.4
achatinus Ac1.6
stercusmuscarum Sm1
circumcisus Cr1.5 *
achatinus Ac1.5
magus Mg1
consors Cn1.2
catus C1.2 *
monachus Mn1.6 *
consors Cn1.6 *
circumcisus Cr1.2
stercusmuscarum Sm1.5 *
striatus S1.5 *
circumcisus Cr1.4 *
aurisiacus A1.3
circumcisus Cr1.7 *
aurisiacus A1.2
circumcisus Cr1.6 *
catus C1.3 *
gauguini Ga1.1
stercusmuscarum Sm1.3
striatus S1.1
circumcisus Cr1.3
striatus S1.12
striatus
striatus S1.10b
striatus S1.7 *
striatus S1.4 *
striatus S1.11
circumcisus Cr1.1
stercusmuscarum Sm1.1
striatus S1.8
striatus S1.3 *
striatus S1.10a *
striatus S1.9
gauguini Ga1.2
striatus S1.6 *
achatinus Ac1.8
aurisiacus A1.4
aurisiacus A1.1
aurisiacus A1.1a
consors Cn1.5 *
magus M1.5 *
magus M1.1
magus M1.4 *
achatinus Ac1.3
achatinus
catus C1
magus M1.3 *
consors Cn1.1
consors
achatinus Ac1.1
monachus Mn1
achatinus
achatinus
achatinus
monachus Mn1.4 *
monachus Mn1.5 *
consors Cn1.3
consors Cn1.4 *
monachus Mn1.3 *
achatinus Ac1.2
achatinus
monachus Mn11
magus MVIII
consors Cn4.4
magus M4.2
consors Cn4.3 *
magus M4.3
consors Cn4.2
catus C4.2
monachus Mn4.3 *
achatinus Ac4.1
monachus Mn4.1
catus C4.1a
catus C4.1b
consors Cn4.1
striolatus Sx4.1
striatus SVIIIA
striatus SVIII
striatus AY166873.1
aurisiacus A4.4 *
circumcisus Cr4.2 *
aurisiacus A4.3 *
achatinus Ac4.2
monachus Mn4.2
aurisiacus A4.1
stercusmuscarum SmVIIIA
stercusmuscarum SmVII
stercusmuscarum Sm4.2
aurisiacus A4.2
circumcisus Cr4.1
0.1


α4/7 - Pionoconus
bullatus Bu1.3
κA - Pionoconus
.72
.55
.41
.99
.96
ermineus E1.3A
ermineus E1.3B
purpurascens P1.7
bullatus Bu1.4
α4/4 - Pionoconus
kinoshitai Kn1.2
kinoshitai Kn1.3
cervus Cs1.2
bullatus Bu1.1
bullatus Bu1.2
α3/5 - Pionoconus
0.1


A
κA
α4/7
α4/4
α3/5
C
α4/4
α3/5
κA
α4/7
B
α4/7
κA
α4/4
α3/5
D
α3/5
α4/4
κA
α4/7
E
α4/7
κA
α4/4
α3/5