Mechanisms for Evolving Hypervariability: The Case of Conopeptides

Silvestro G. Conticello, Yoav Gilad,† Nili Avidan,† Edna Ben-Asher,† Zehava Levy,* and Mike Fainzilber**

*Laboratory of Molecular Neurobiology, Department of Biological Chemistry, and †Crown Human Genome Center, Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

Hypervariability is a prominent feature of large gene families that mediate interactions between organisms, such as venom-derived toxins or immunoglobulins. In order to study mechanisms for evolution of hypervariability, we examined an EST-generated assemblage of 170 distinct conopeptide sequences from the venoms of five species of marine Conus snails. These sequences were assigned to eight gene families, defined by conserved elements in the signal domain and untranslated regions. Order-of-magnitude differences were observed in the expression levels of individual conopeptides, with five to seven transcripts typically comprising over 50% of the sequenced clones in a given species. The conopeptide precursor alignments revealed four striking features peculiar to the mature peptide domain: (1) an accelerated rate of nucleotide substitution, (2) a bias for transversions over transitions in nucleotide substitutions, (3) a position-specific conservation of cysteine codons within the hypervariable region, and (4) a preponderance of nonsynonymous substitutions over synonymous substitutions. We propose that the first three observations argue for a mutator mechanism targeted to mature domains in conopeptide genes, combining a protective activity specific for cysteine codons and a mutagenic polymerase that exhibits transversion bias, such as DNA polymerase V. The high D_n/D_s ratio is consistent with positive or diversifying selection, and further analyses by intraspecific/interspecific gene tree contingency tests weakly support recent diversifying selection in the evolution of conopeptides. Since only the most highly expressed transcripts segregate in gene trees according to the feeding specificity of the species, diversifying selection might be acting primarily on these sequences. The combination of a targeted mutator mechanism to generate high variability with the subsequent action of diversifying selection on highly expressed variants might explain both the hypervariability of conopeptides and the large number of unique sequences per species.

Introduction

Conopeptides are small (10–35 residues) neurotoxin gene products derived from Conus snail venoms, with solution structures defined and stabilized by a high density of cysteine residues (Norton and Pallaghy 1998). Disulfide bridges in conopeptides provide a structural scaffold that tolerates high variability in the intercysteine loops, thus enabling targeting of diverse receptors. For example, the ''six-cysteine, four-loop'' scaffold (C . . . C . . . CC . . . C . . . C) is shared by conopeptides targeting multiple subtypes of calcium channels and three different sites on sodium channels (McIntosh, Olivera, and Cruz 1999). The high selectivity of conopeptides has facilitated their use as pharmacological tools (McIntosh, Olivera, and Cruz 1999), and has provoked interest in their potential as drug leads (Bowersox and Luther 1998).

Although random design of conopeptide scaffolds has been attempted (Palmer et al. 1998), in practice the spacing of cysteine residues appears to be important for productive folding of such peptides (Drakopoulou et al. 1998). Correspondingly, the number of naturally occurring conopeptide scaffolds so far identified is limited (Fainzilber et al. 1995; McIntosh, Olivera, and Cruz 1999; Rigby et al. 1999). These scaffolds define large hypervariable families that may share a common evolutionary origin. Given the purported diversity of Conus venoms (an estimated 100 unique conopeptides per species for a genus of ca. 500 species), they present a unique opportunity for study of the evolution of large variable gene families. Although a number of speculations on conopeptide evolution have been advanced (Olivera et al. 1999), only Duda and Palumbi (1999) have so far addressed this topic in a quantitative manner. Their single study provided evidence for accelerated evolution in conopeptides; however, the analysis was based on a limited data set derived mainly from one species.

In order to obtain an unbiased overview of conopeptide precursor variability and evolution we applied an EST strategy to identify conopeptide-encoding transcripts. Conus venom systems seem ideally suited for such an approach, since conopeptide-encoding transcripts are relatively short ($\leq$0.5 kb) and highly expressed. Sequencing of over 2,000 cDNA clones and PCR products from five different Conus species provided a data set of 170 distinct conopeptide precursor sequences from eight gene families representing three cysteine scaffold superfamilies. The data suggest that current conopeptide diversity may reflect the combination of a targeted mutagenic mechanism to generate high variability with the subsequent action of diversifying selection on highly expressed variants.

Materials and Methods

The preparation of a *Conus textile* venom duct cDNA library has been described (Sasaki et al. 1999). Venom duct mRNA was prepared from 20–30 specimens each of the species *Conus arenatus*, *Conus pennaceus*, *Conus tessulatus*, and *Conus ventricosus.*

Key words: venom, conotoxin, mutagenesis, polymerase, gene family, codon bias.

Address for correspondence and reprints: Mike Fainzilber, Department of Biological Chemistry, Weizmann Institute of Science, 76100 Rehovot, Israel. E-mail: mike.fainzilber@weizmann.ac.il.

Mol. Biol. Evol. 18(2):120–131. 2001

cDNAs were prepared by oligo dT (*Not*I) priming, ligated to *Bst*XI adaptors, and cloned into pCDNA3/*Bst*XI/*Not*I. Multiple transformations for the *C. arenatus* and *C. pennaceus* ligations yielded libraries comprising approximately 2×10^5 primary clones each. Randomly picked clones from each library plating were sequenced by the dye terminator method on ABI 373 or ABI 377 automated sequencers.

Sequences were edited to discard vector and adaptor regions using Sequencher 3.0 (GeneCodes Corp., Ann Arbor, Mich.). Contigs were assembled automatically (identity cutoff 98%), followed by manual edition. Individual transcripts were aligned using CLUSTAL X (Thompson et al. 1997), and the alignments were refined manually. Phylogenetic trees were constructed using the neighbor-joining method (Saitou and Nei 1987) and then visualized with TreeView (Page 1996). Synonymous versus nonsynonymous substitution rates were analyzed using MEGA (Kumar, Tamura, and Nei 1993). A one-tailed *t*-test with infinite degrees of freedom was used to estimate the significance of differences in substitution rates in the different regions, both within and between species and gene families. Tip tests (Templeton 1996) were performed on the basis of alignments specific to the analyzed region (signal+propeptide, mature domain) in order to reduce the complexity of the cladogram (clades in signal+propeptide-based trees are quite different from the clades in mature-based trees and/or full-length–based trees). A Fisher 2×2 contingency test was performed on silent versus replacement substitutions in external and internal branches of the gene tree as suggested by Castelloe and Templeton (1994). Sequence logos were plotted according to Schneider and Stephens (1990). Codon variability profiles were drawn for graphic presentation of levels of variability along sequence stretches and were generated according to Pilpel and Lancet (1999), modified for nucleotide sequences by giving equal weight to all four nucleotides and gaps (program by S.G.C.).

RT-PCR was performed using primers on conserved elements in the 5′ and 3′ untranslated regions (UTRs) of each conopeptide family as follows: MEKL-5′1F, GACCCTGCCGTCATCTCAGC; MEKL-3′1R, AGCCTTGAAGTCTCTGAAGA; MKLT1–5′1F, CACTGTCGTCTTTCGCATCA; MKLT1–3′1R, TGTGCTGTCGCTTTATTTGG; MKLT2–5′1F, TGATCCCTGCACGGCGAATC; MKLT2–3′1R, TTGCCTAATTCGTCCATGCT; MSGL-5′1F, CTGTGATACCAGCCCAAACC; MSGL3′2R, GGACGAACGGATTTGAGATG; MRCL-5′1F, CCTGGCAGGTACTCAACGAA; MRCL-3′1R, AACAACACGCTGCCACTTGC; MLCL-5′1F, AAGCCATCAGCCCTCTTCAT; MLCL-3′1R, GASGACCTAGCGAWACGGAA; MMSK-5′1F, CGCCACAGCTAAGACAAGAA; MMSK-3′1R, CTTTGTTATCGCGGCCTCAT; MLKM-5′1F, TACGTGAAGAAGGGTGGAGA; MLKM-3′1R, ACGAACATGATTGCACTCTG. Conditions for RT-PCR were 50°C for 40 min and 94°C for 2 min, followed by 25 amplification cycles of 94°C for 30 s, 55–60°C for 30 s, and 68°C for 1 min. The resulting PCR fragments were ligated directly into a T-overhang vector, and at least 12 clones from each reaction were sequenced. Sequences were edited and assembled as described above, using an identity cutoff of 96% to avoid erroneous inclusion of PCR-generated mutants. RNAse protection assays (RPAs) were performed using the Multi-NPA system (Ambion Inc., Austin, Tex.) according to the manufacturer's instructions with the following oligonucleotide probes: TxMKLT1-031, CAGAATAAGCCAAAGCAGCAATACCTCGATCCAAAGAACGGTGCTaatataact (the lowercase letters are noncomplementary sequences, in order to distinguish full-length probes from protected fragments); TxVII, CAGACTCCAAGACATATGCCGGTGCAGCAATCAAGTGAAAtatgtaaaa; TxVIa+b, GCAGACAAATACTATGCAATAGCCGTCGCAGCAGaaaaccaaa; TxMEKL04111, GGGCATGTGAACAAGGCTCCGTGTAATCCGaaatccaa; TxA021+22, CCAGTACAGCATTGCGAGGCGTCCGTtttaag. Each RPA was repeated at least three times with different amounts of mRNA (40 to 800 ng).

Results

Sequencing a Few Hundred Venom Duct ESTs per Species Provides a Reliable Overview of the Most Prevalent Conopeptide Families

Sequencing of approximately 370 clones each from venom duct cDNA libraries of *C. textile*, *C. pennaceus*, and *C. arenatus* led to the identification of 365 individual transcripts, 38% of which represented putative conopeptide cDNAs. These were identified by manual inspection of a subset of the sequences for open reading frames enriched for cysteine in the predicted C-terminal segment followed by low-stringency alignments across the remaining sequences in order to cluster predicted conopeptide families sharing conserved elements in their precursor sequences. In order to verify comprehensive coverage of the identified families, PCR primers were designed on conserved 5′ and 3′ elements in each grouping and used to amplify related sequences from venom ducts. As shown in table 1, the majority of conopeptide EST sequences from each species belonged to one (in *C. textile* and *C. arenatus*) or two (in *C. pennaceus*) superfamilies. Strikingly, much of the available sequence diversity in the most prevalent families was identified in the ESTs, with only 7% (scaffold VI) or 25% (scaffold IX) additional transcripts added from sequencing of PCR products. Conversely, the scaffold III/IV family, which was poorly represented in our EST database (one sequence), was found to comprise a far larger range of diversity upon supplementary sequencing of PCR products (table 1). All conopeptide precursors revealed the canonical three-domain structure (signal, propeptide, mature) previously described for these genes (Olivera et al. 1999).

Two of the species analyzed above are molluscivorous (*C. pennaceus* and *C. textile*), whereas *C. arenatus* preys on worms. Since most Conidae are vermivorous, we enlarged our data set of sequences from worm-eating species by PCR on venom duct cDNA from *C. tessulatus* and *C. ventricosus.* The final data set comprised 170 distinct conopeptide precursor transcripts from five species (table 1). Evolutionary relationships

Table 1
Conopeptide Superfamilies by EST Sequencing and PCR

	Conopeptide Superfamily[a]		
	Scaffold VI/VII	Scaffold IX	Scaffold III/IV
Representation in ESTs (%)			
Conus textile	67.6	21.6	2.7
Conus pennaceus	24.5	34.7	10.2
Conus arenatus	60.4	27.1	2.1
No. of distinct transcripts			
C. textile			
EST	24	5	1
PCR	2	3	8
C. pennaceus			
EST	9	12	0
PCR	2	2	3
C. arenatus			
EST	26	3	0
PCR	0	0	2
Conus tessulatus PCR	7	7	11
Conus ventricosus PCR	26	11	6
Total no. of sequences	96	43	31

[a] Roman numerals denote different cysteine scaffolds (for nomenclature, see Fainzilber et al. 1995; McIntosh, Olivera, and Cruz 1999; Rigby et al. 1999).

between these cDNAs were examined by the construction of phylogenetic trees for each scaffold superfamily from alignments of both DNA (fig. 1*A*) and predicted protein sequences. Trees based on protein sequences were highly similar to those based on DNA (data not shown). Each superfamily is clearly divisible into two to four distinct families on the basis of conserved elements in the precursor sequences. These data suggest that conopeptide gene families can be defined primarily on the basis of their highly conserved signal domains, in combination with conserved elements in their 5′ and 3′ UTRs. Therefore, we propose that the nomenclature for conopeptide gene families be based on the first four residues of the signal peptide in combination with the Roman numeral defining their cysteine scaffolds, e.g., the MEKLT1 family of scaffold VI conotoxins (fig. 1*B*).

Conopeptide Transcripts Are Expressed at Vastly Differing Levels

In the course of the EST sequencing, we noted large apparent differences in the numbers of clones sequenced for specific transcripts. Indeed, upon plotting the number of sequences obtained for each conopeptide transcript, striking differences in the apparent levels of expression of different conopeptides could be observed (fig. 2*A*). For all three species analyzed, there were order-of-magnitude differences between relative expression levels of different transcripts from the same superfamily. Although comprehensive EST sequencing can measure relative expression levels of different transcripts, this typically requires sequencing of large numbers of clones. We therefore verified our estimates of transcript abundance for five selected scaffold VI conopeptides (indicated by numbers in fig. 2*A*) by independent quantification in RPAs. As shown in figure 2*B*, there was good correlation between the expression levels predicted from the EST data and those independently measured by RPA. Hence, related conopeptides may be expressed at vastly different levels in the venom ducts.

The Mature Domain in Conopeptides Is Undergoing Accelerated Mutation

Previous authors have suggested that venom-derived gene families, including conopeptides, are undergoing accelerated evolution (Ohno et al. 1998; Duda and Palumbi 1999; Froy et al. 1999). In order to find out if this hypothesis also holds for the eight gene families of conopeptides found in our data set, we first examined the relative mutation rates in different regions of the conopeptide transcripts. Since synonymous substitutions are apparently neutral, the fixation rate can be considered proportional to the mutation rate. Therefore, we can assume the number of synonymous substitutions per synonymous site (D_s) (Nei and Gojobori 1986) to be an adequate representation of the mutation rate. As seen in figure 3*A*, D_s in the mature peptide region is significantly higher than that observed for the signal domain, with the propeptide region in most families exhibiting an intermediate value. Nucleotide substitution rates in the signal peptide and noncoding UTRs (where available, calculated according to the Jukes and Cantor one-parameter model to be coherent with the D_s estimation) were for the most part similar to the value calculated for a Conus calmodulin gene fragment (fig. 3*A*) from the data of Duda and Palumbi (1999). The apparent mutation rates for the mature domain of conopeptides are up to one order of magnitude higher than this (fig. 3*A*). Moreover, the substitution rates for the mature domains are almost certainly underestimated due to the occurrence of multiple substitutions per site (apparent homoplasy) in the mature region, as shown in figure 3*B*. The difference in mutation rates between different regions in these conopeptide precursors is also reflected in their degree of divergence, as demonstrated by the unrooted tree diagrams of figure 3*C*. While segregation of the different gene families is clearly seen in trees from the untranslated regions and the signal domain, in the mature domain most transcripts diverge directly from the origin, and segregation of the different gene families is not preserved.

Transversion/Transition Ratios in the Mature Domain of Conopeptide Precursors Fit the Signal for Mutagenic Polymerase Action

The accelerated rates noted above might be due to strong positive selection, or they might reflect a targeted hypermutation mechanism. Recent studies have put forward the notion that error-prone DNA polymerases may act as ''mutases'' that are activated under specific conditions or in specific gene regions (Radman 1999). Very recent studies have further shown that mutagenic replication by DNA polymerase V is characterized by a bias for transversions over transitions (Maor-Shoshani et al. 2000; Tang et al. 2000). We therefore examined the transversion/transition (Tv/Ts) ratios in our conopeptide gene family alignments. The Tv/Ts ratio rises from sig-

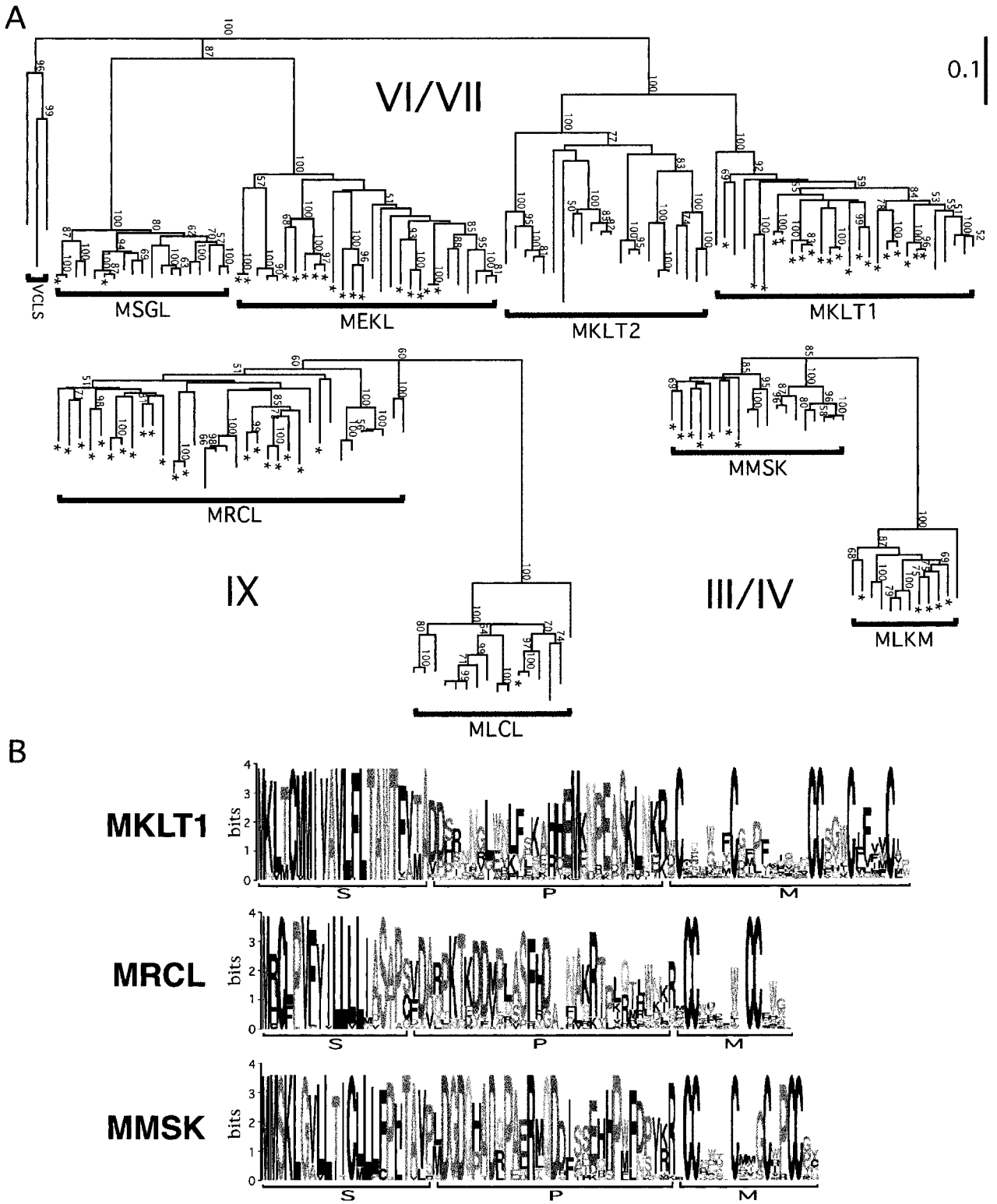

FIG. 1.—Phylogenetic trees and sequence logos of conopeptide families. *A,* Phylogenetic trees generated from DNA alignments of the main conopeptide superfamilies. Only full-length distinct sequences were used for building the trees. Each tree represents a single cysteine scaffold superfamily as indicated. Sequences derived from molluscivorous species are identified by stars. Bootstrap values >50% are shown adjacent to the branches. *B,* Sequence logos generated from protein alignments of representative conopeptide families. Note the hypervariability in the mature toxin domains, except for conserved cysteine residues.

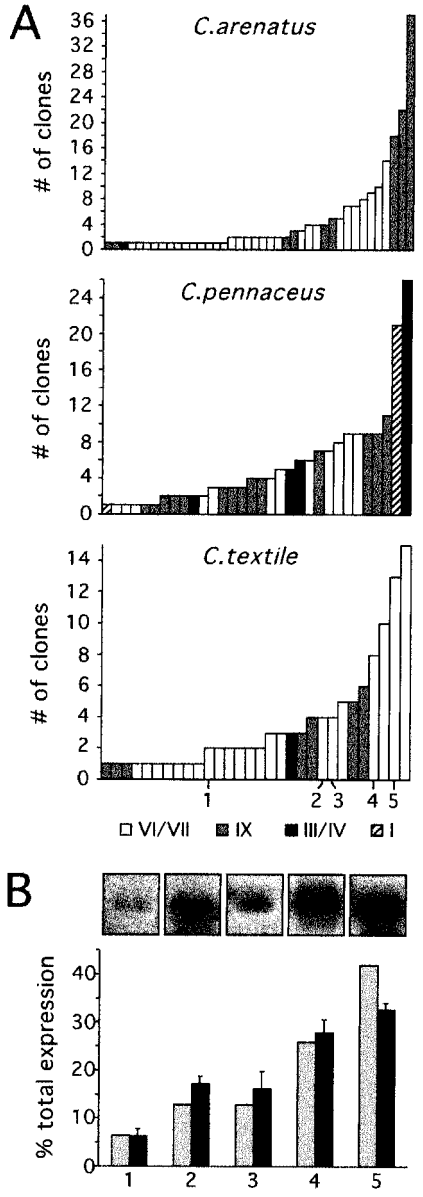

FIG. 2.—Conopeptide transcript expression levels. *A*, The number of clones obtained for each individual transcript by EST sequencing. The scaffold affiliation of each transcript is shown by shading of the bars. *Conus textile* transcripts analyzed independently by RPA are indicated by numbers: 1 = TxMKLT1-031; 2 = TxVII; 3 = TxMEKL-021/22; 4 = TxVIa/b; 5 = TxMEKL-04111. *B*, Expression levels of the five transcripts analyzed by RPA. Gray bars indicate predicted expression levels from the EST data, while black bars indicate the observed expression levels, calculated as percentages of total density (representative autoradiograms are shown above each bar). Three categories of expression levels (transcripts 1, 2+3, and 4+5) significantly differ from each other ($P < 0.01$, one-way ANOVA).

nal through propeptide to the mature domain in the alignments, with a twofold bias for transversions in the mature domain (fig. 4). This ratio is quite comparable to the in vitro measured transversion bias of DNA Pol V, as reported by Maor-Shoshani et al. (2000), and it is consistent with the possibility that mature domain hypervariability may result, in part, from actions of a similar polymerase.

Cysteine Codons Within the Hypervariable Mature Domain Are ''Hyperconserved''

Given the high rates of codon variability in the mature regions of conopeptide precursors (fig. 5*A*), it was striking to note almost no substitutions in the third position of each cysteine codon in the alignments (fig. 5*B* and table cited in *Supplementary Material*). Whereas the prediction would have been for silent variation of cysteine codons between TGT and TGC, there is an overwhelming bias for one or the other codon at different positions in the alignments. This is presented graphically for representative families in figure 5*B*, and the phenomenon is observed at a high level of statistical significance in all conopeptide gene families analyzed to date (table cited in *Supplementary Material*; Conticello et al. 2000). Since different codons are conserved at different positions, the observation cannot be attributed to simple codon bias, and it is especially striking in view of the extremely hypervariable environment in the mature domain. In order to verify that this phenomenon is restricted to the cysteine codons, we examined the degree of conservation in the nearest and only other conserved residue in these alignments, namely, the arginine residue at the consensus cleavage site between the pro and the mature domains. It should be noted that Olivera et al. (1999) have asserted that the C-terminal region of the pro domain (including the consensus arginine at the cleavage site) is encoded on the same exon as the mature domain in conopeptide genes. The conserved arginine residue in this position reveals a variability consistent with general molluscan codon usage (fig. 5). Finally, it is noteworthy that a number of frameshifted pseudotranscripts found in our EST database conserve the positional cysteine codons characteristic for the corresponding gene families (data not shown).

Conopeptide Gene Trees Show Evidence of Recent Diversifying Selection

In addition to molecular mechanisms that might be involved in generating hypervariability of conopeptides, positive selection has been suggested to play a decisive role in diversification of venom-expressed gene families (Ohno et al. 1998; Duda and Palumbi 1999). In order to examine this possibility, we determined the number of nonsynonymous substitutions per synonymous site (D_n) in the data set. D_n/D_s ratios greater than 1 are indicative of accelerated evolutionary change via positive selection (Endo, Ikeo, and Gojobori 1996), whereas $D_n/D_s < 0.3$ is usually indicative of strong purifying selection (Ophir et al. 1999), although exceptions to these rules do exist (Rooney, Zhang, and Nei 2000). As shown in figure 6, both D_n and D_s increase from the signal through the pro to the mature domain of conopeptide precursors. For the mature domain, $D_n/D_s > 1$ for most of the comparisons (fig. 6*C*), consistent with positive selection, although a plot of D_n/D_s versus D_s reveals saturation of the signal at D_s values greater than 1. Nonetheless, over 75% of the D_n/D_s ratios for the mature region are greater than 1, and their distribution is clearly different from that observed for the signal and propeptide regions (one-

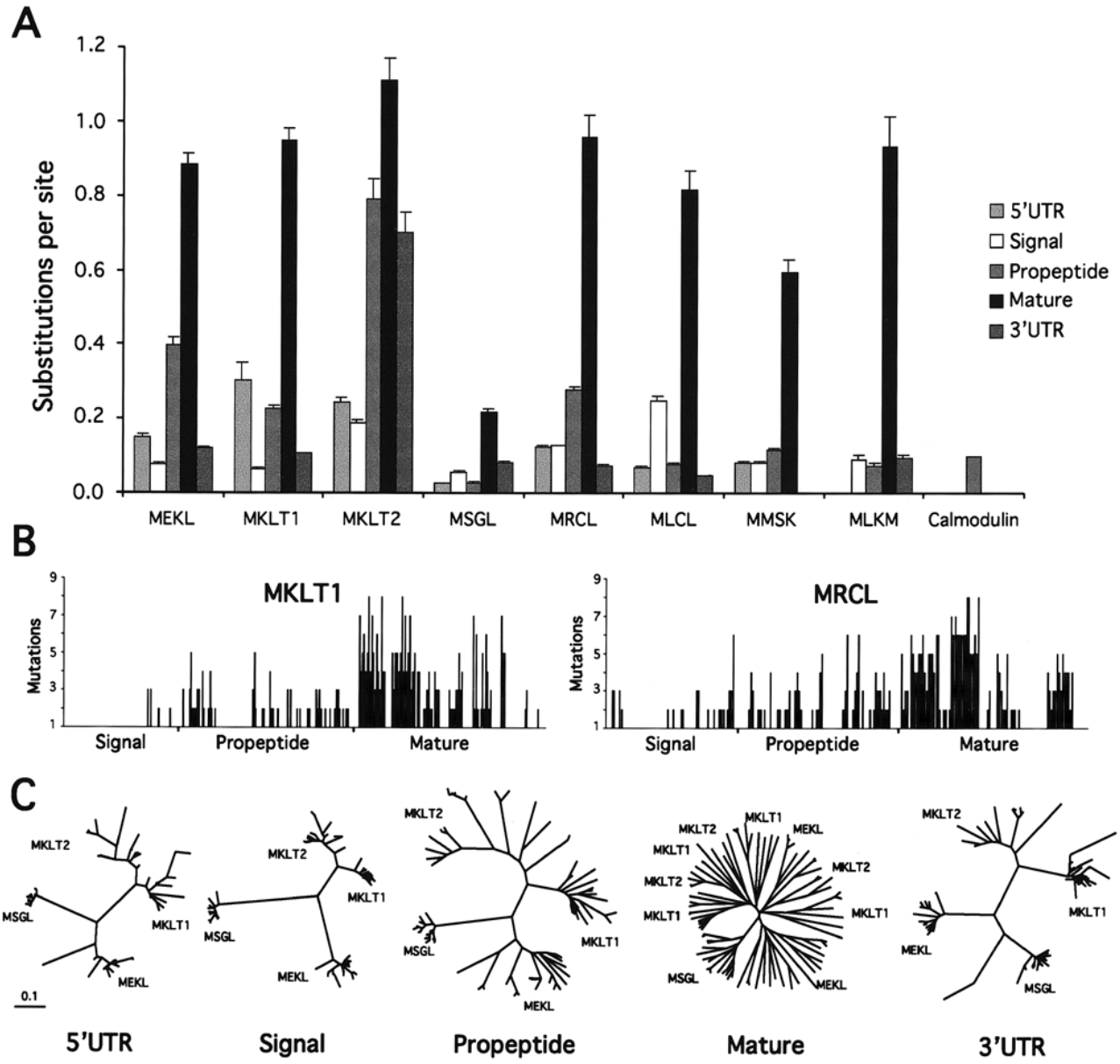

FIG. 3.—Different apparent nucleotide substitution rates for conopeptide precursor domains. *A,* Apparent substitution rates in the different domains of conopeptide gene families. The number of synonymous substitutions per synonymous site (D_s) in the coding regions is compared with the Jukes-Cantor corrected rate of nucleotide substitution in the noncoding regions. *B,* Apparent homoplasy plots for two representative gene families. Note the high levels of homoplasy in the propeptide and, especially, the mature region, suggesting that the apparent mutation rates shown in (*A*) are most likely underestimated for these domains. *C,* Unrooted trees generated using specific regions from all conopeptide transcripts belonging to the scaffold VI/VII superfamily. While in the untranslated regions and in the signal peptide domain, all the gene families (MEKL, MKLT1, MKLT2, and MSGL) segregate in clearly defined branches; in the mature domain, the transcripts start to diverge directly from the origin and gene family branches are mixed. The same pattern can be observed in the other superfamilies (not shown).

tailed *t*-test with infinite degrees of freedom; $P < 10^{-20}$ for both comparisons). An appreciable fraction of the D_n/D_s data points for the propeptide region are also greater than 1, although this bias is not statistically significant. A clear caveat for these analyses is that many of the comparisons in our data set will be between distantly related loci, and identification of orthologous pairs for more rigorous analysis is difficult due to the high variability. We therefore turned to a more rigorous test for positive or diversifying selection, specifically a gene tree homogeneity test based on the tip test of Templeton (1996). This test was originally developed as an alternative and more powerful contingency test for ratios of replacement versus silent mutations at polymorphic versus divergence sites. It also corresponds to a contrast of young and old variants according to the topology of the gene tree. We therefore used this test to examine the possibility of positive or diversifying selection affecting our sample under the assumption that its influence would be more evident in older variants (i.e., high value of replacement mutations adopted by such a selection mechanism). In order to apply the test, we had to un-

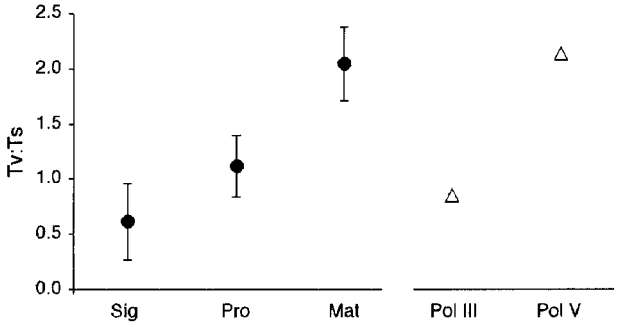

FIG. 4.—The ratio of nucleotide transversions to transitions (Tv/Ts) in alignments of conopeptide transcripts. Tv/Ts was calculated for each family alignment in the different regions, and the average ± SD for each region is shown. Transversion bias ratios determined by Maor-Shoshani et al. (2000) for DNA Pol III and DNA Pol V are shown for comparison.

cover all the mutations leading to a specific gene tree (see fig. 7*A* for a representative tree; for all others, see *Supplementary Material*). We than conducted a 2×2 contingency test of silent versus replacement substitutions in external and internal branches of the gene tree (Castelloe and Templeton 1994). In order to reduce the complexity of the trees and to highlight traces of positive selection, the analyses were carried out on separate cladograms for each region of the conopeptide precursors. The test on the signal+propeptide region, as expected from the D_n/D_s analysis, does not reach significance (table 2) in any but one of the gene families. The tip test was also performed on separate gene trees for signal and propeptide regions and on single-species trees, with similar results (data not shown). Although the null hypothesis of neutrality for these regions therefore cannot formally be rejected, this conclusion must be tempered with the fact that our gene trees could represent a very ancient history; therefore, it is possible that traces of positive selection are embedded in the entire tree. For the mature region, although high significance was obtained in only two gene families, the overall trend is consistent with recent strong diversifying selection (*P* values in table 2).

One possible driving force for diversifying selection in conotoxins is the prey specialization prevalent in this genus (Kohn and Nybakken 1975). If this is a dominant selective force, one might predict segregation of molluscivorous versus vermivorous branches in conopeptide gene trees. Although this is not the case for the trees of figure 1, the branching of these gene family trees is determined mainly by the relatively conserved signal and pro regions. We therefore plotted an unrooted dendrogram for only the mature region of the largest conopeptide superfamily in our data—the six-cysteine scaffold VI/VII grouping. As shown in figure 8*A*, there is no clear segregation of vermivorous versus molluscivorous branches also in this tree. A parallel plot restricted to those sequences shown to be highly expressed in the corresponding venoms (fig. 2) does, however, reveal clustering of most of the molluscivorous versus vermivorous sequences on different branches (fig. 8*B*). Unfor-

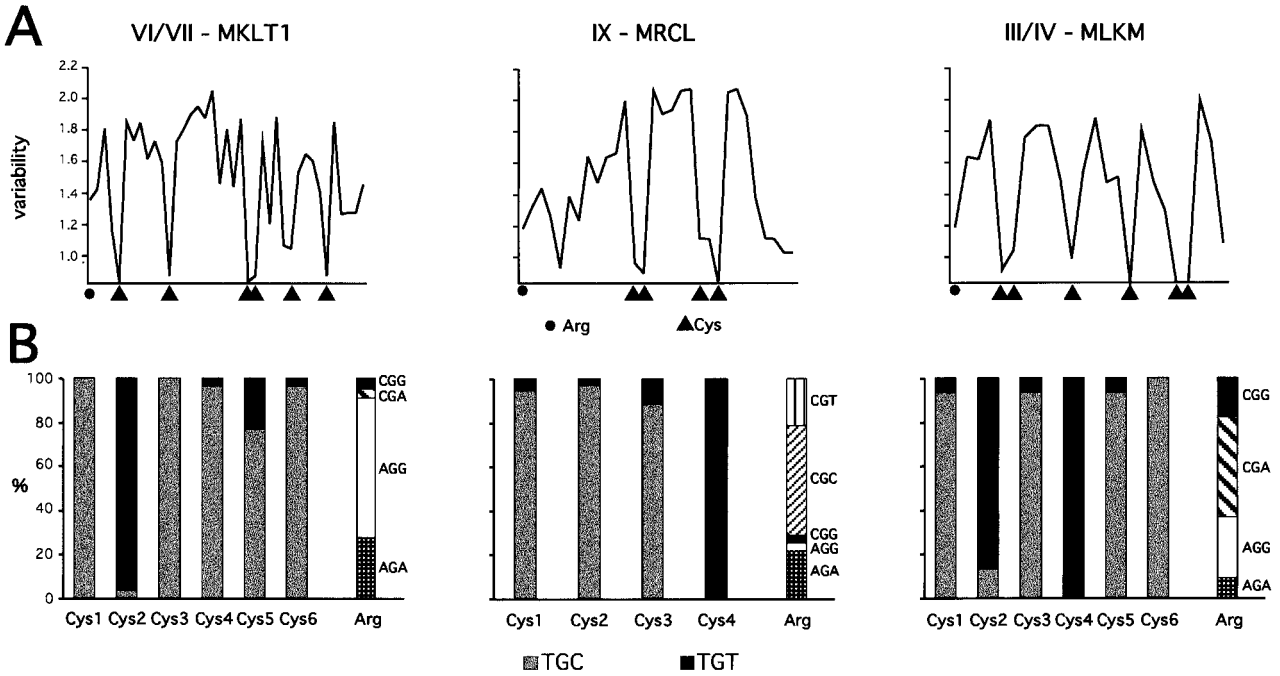

FIG. 5.—Position-specific cysteine codon conservation in conopeptides. *A,* Variability profiles (Pilpel and Lancet 1999) for the segment encoding the mature conopeptide region in alignments of three representative gene families. Note the variability minima represented by the cysteine codons. *B,* The percentages of cysteine codon representation are shown for the three representative families depicted in *A* compared with the codon usage for the conserved arginine residue that forms part of the propeptide cleavage site. The probabilities of obtaining the observed cysteine codon biases were estimated from a binomial distribution assuming a priori probabilities of 43.5% TGC versus 56.5% TGT, calculated from the codon bias tables for the five most sequenced molluscan species, and were highly significant in all cases ($P \ll 0.01$; see table cited in *Supplementary Material*).

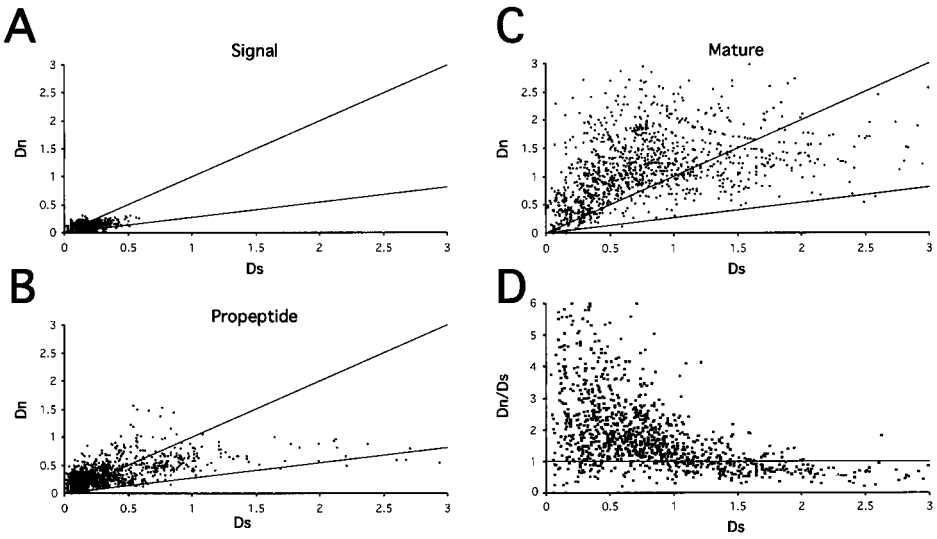

FIG. 6.—Nonsynonymous versus synonymous substitutions in conopeptides. D_n over D_s in intrafamilial comparisons of (*A*) signal, (*B*) propeptide, and (*C*) mature peptide encoding domains of the conopeptide gene families. Approximately 3% of the data points for the mature domain are outside the scale of the plot shown. Lines demarcate zones for $D_n/D_s > 1$ (positive or diversifying selection), $1 > D_n/D_s > 0.27$ (lack of constraints), and $D_n/D_s < 0.27$ (purifying selection). (*D*) D_n/D_s plotted against D_s for the mature domain only. Note that at $D_s > 1$, the signal drops, apparently due to saturation.

tunately, the data set of highly expressed transcripts did not include enough representatives from the same gene family to enable analysis by a separate Templeton tip test (the mature domain alignment for the unrooted tree in the right panel in fig. 8 is not sufficiently good to enable reliable analysis).

Discussion

Two views on the evolutionary origin of conopeptide hypervariability have been proposed. Duda and Palumbi (1999) suggested that gene duplication and diversifying selection underlie accelerated evolution of conopeptides. Similar hypotheses have been proposed for snake venom evolution (Ohno et al. 1998). In contrast, Olivera et al. (1999) have suggested that unconventional mechanisms, perhaps similar to recombination events in immunoglobulin loci, are required to explain conopeptide evolution. This debate parallels earlier controversies on evolution of hypervariable gene families in the immune system (Nei, Gu, and Sitnikova 1997; Hughes and Yeager 1998). In this context, it is striking that hypervariability in gene families appears most apparent in systems evolved to recognize foreign molecules, be they venom-derived toxins, gene families of the immune system, or antigenic parasite surface proteins (Field and Boothroyd 1996).

Our data suggest that in the case of venom-derived conopeptides, both a hypervariability-generating molecular mechanism and diversifying selection have contributed to the evolution of these large and hypervariable gene families. The putative hypermutation mechanism is postulated on the basis of two molecular signatures observed in our data set. The first of these is a clear bias for transversions over transitions in the mature domain (fig. 4). The selective generation of transversions has been reported by a number of groups for SOS stress response mutagenesis in *Escherichia coli* (Fijalkowska, Dunn, and Schaaper 1997; Watanabe-Akanuma, Woodgate, and Ohta 1997). This response is thought to have evolved as a means for increasing genetic diversity in order to accelerate adaptation of bacteria to hostile conditions (Radman 1999; Radman, Matic, and Taddei 1999). Two recent studies have shown that the SOS-inducible enzyme DNA polymerase V is highly mutagenic and has an obvious tendency to form transversions during gap-filling DNA replication (Maor-Shoshani et al. 2000; Tang et al. 2000). It is noteworthy that the low processivity of DNA Pol V, which adds six to eight bases before dissociating (Tang et al. 2000), would fit well with the short intercysteine stretches typical of conopeptide sequences (fig. 1*B*). Thus, the transversion bias and short hypervariable stretches observed in the mature domain of conopeptides are highly suggestive of a targeted mutagenic process involving a DNA-Pol-V-like enzyme.

If a targeted hypermutation process is indeed at work on conopeptide genes, clearly it would have been adaptive for Conidae to evolve a protective mechanism to conserve structurally crucial cysteine residues. We would like to argue that the striking positional conservation of cysteine codons in the mature region (fig. 5; Conticello et al. 2000) is the molecular signature of such a mechanism, for example, a macromolecule which would specifically bind to TGC or TGT triplets. Otherwise, the observed conservation is difficult to reconcile with the extremely high substitution rates in the immediate vicinity, especially since Olivera et al. (1999) have asserted that the mature domain in conopeptide genes is encoded on a single exon. It is not clear why selection processes would remove genes with silent mutations in the cysteine codons, which would surely be expected to arise in the course of hypermutation of the

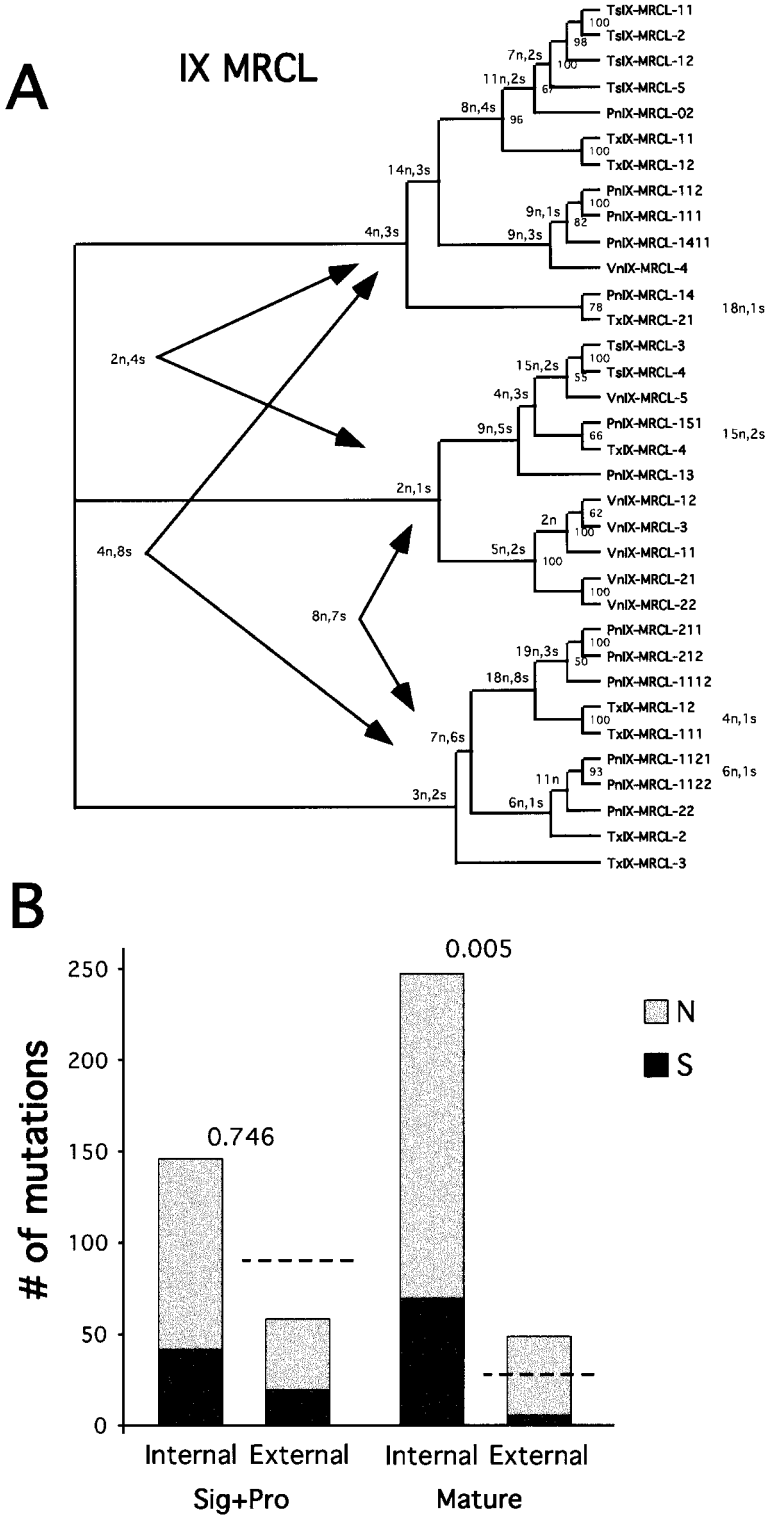

FIG. 7.—Intraspecific/interspecific gene tree contingency tests on conopeptide families. *A*, Representative cladogram used for the tip test. Separate cladograms based on alignments of the signal+propeptide or the mature region were built. The numbers beside the branches are the numbers of nonsynonymous (n) and synonymous (s) substitutions in the internal nodes; the substitutions in the tips are placed on the right of the transcript names, and the numbers near the arrows indicate the substitutions between the two main clades. *B*, Representative comparison of nonsynonymous/synonymous substitutions between external and internal nodes in the signal+propeptide and in the mature regions of a representative family (MRCL). The dotted line represents the theoretical number of nonsynonymous substitutions in the external nodes for $P < 0.05$. The numbers above the bars are the actual P values for the data.

Table 2
Intraspecific/Interspecific Gene Tree Contingency Tests on Conopeptide Families

		Internal Nodes		External Nodes		
Domain	Family	NS	S	NS	S	*P*
Signal and propeptide	MEKL	118	53	20	7	0.111
	MKLT1	98	34	24	10	0.745
	MKLT2	77	40	32	10	0.147
	MSGL	35	7	3	1	0.866
	MRCL	104	42	40	19	0.746
	MLCL	34	13	8	3	0.650
	MMSK	52	15	37	11	0.617
	MLKM	20	14	33	4	0.004*
Mature	MEKL	165	70	41	10	0.095
	MKLT1	137	54	87	22	0.078
	MKLT2	146	63	48	16	0.265
	MSGL	67	25	15	5	0.543
	MRCL	177	70	43	5	0.005*
	MLCL	117	56	12	2	0.132
	MMSK	44	18	19	4	0.210
	MLKM	34	20	44	3	0.0002*

Note.—Nonsynonymous (NS) and synonymous (S) substitutions in the internal and external nodes in all conopeptide gene families. Comparison of internal versus external nodes in the signal and propeptide domains does not reveal significant differences except for the MLKM family. The parallel analysis for the mature domain reveals a general trend for differences, which in most cases do not reach strong statistical significance.

* $P \leq 0.01$.

mature domains. Moreover, the cysteine codon conservation is also maintained in frameshifted pseudotranscripts, which represent loci that are not under selection pressure. Trivial explanations for the observed codon conservation, such as a global cysteine codon or tRNA bias, can be ruled out, since different cysteine codons are conserved at different positions (fig. 5) and the same tRNA recognizes both cysteine codons (Schimmel, Soll, and Abelson 1979). Although this putative macromolecule might not require cysteine codon bias at specific positions a priori, such bias may have arisen as a by-product of protecting these codons from mutagenic polymerases. Furthermore, it is noteworthy that a macromolecule binding to cysteine codons might actually serve to prime DNA Pol V mutagenic activity in the vicinity. This arises from the possibility that such a bound macromolecule would interrupt normal DNA replication, thus creating a single-stranded gap in the replicating strand. DNA Pol V or a similar enzyme would then be recruited as part of the damage response to repair the lesion (Maor-Shoshani et al. 2000), thus inserting a mutated stretch downstream of the protected codon. Finally, it should be noted that the available evidence for both components of this hypothetical mutagenic mechanism is correlative only, and other explanations for our observations cannot be ruled out. Future work on these lines should focus on obtaining direct biochemical evidence, including measurement of mutagenic DNA polymerase activities in Conus tissues and initiation of a search for the macromolecules involved.

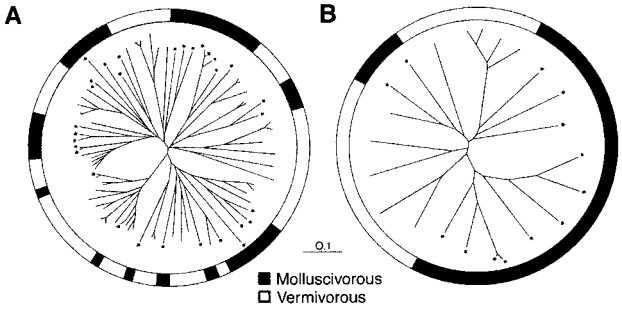

Fig. 8.—Segregation by feeding specificity of highly expressed conopeptide transcripts. Unrooted dendrograms for the mature region alone of the scaffold VI/VII superfamily are shown. *A,* In this tree, which incorporates all sequences in the superfamily, there is no clear segregation of vermivorous (*Conus arenatus*) versus molluscivorous (*Conus textile* and *Conus pennaceus*) branches. *B,* This tree incorporates only those sequences shown to be highly expressed in the corresponding venoms (fig. 2), and in this case clustering of most of the molluscivorous versus vermivorous sequences is apparent.

The second process manifested in our data is recent diversifying selection of the mature domain, as previously suggested by Duda and Palumbi (1999). Both the D_n/D_s ratios (fig. 6) and the Templeton tip tests for mature domain cladograms (fig. 7 and table 2) support this scenario, albeit not with strong statistical significance. Recent diversification of the genus Conus is also observed in the fossil record, with a very rapid increase in the number species from the Pliocene to the present of approximately 500 species (Kohn 1990). Reef assemblages of Conidae have been shown by extensive studies of Kohn and colleagues to exhibit extreme specialization in their feeding preferences (e.g., Kohn and Nybakken 1975); therefore, specialization for different prey species could provide one driving force for such diversifying selection on the toxins used to envenomate the prey. Although phylogenetic specificity is found in the actions of certain conotoxins (Fainzilber et al. 1994), and prey-specific morphological adaptations of the venom apparatus have been described (Kohn, Nishi, and Pernet 1999), the complete conopeptide trees shown in figure 1 do not appear to segregate based on divisions between molluscivorous and vermivorous species. However, we have shown (fig. 2) that order-of-magnitude differences

exist in expression levels for different conopeptide transcripts. These differences may be due to promoter strength differences, transcription levels, mRNA stability, or other factors. Whatever their source, differential expression levels of conopeptides are also observed in the protein content of the venom duct, and it is noteworthy that most conopeptides so far isolated on the basis of prey paralysis assays are quantitatively significant components of their respective venoms. The most striking example is that of conotoxin-GmVIA, which is the single predominant component of *Conus gloriamaris* venom (Shon et al. 1994). Thus, prey-driven diversifying selection may act primarily on the limited number of highly expressed toxins in each species. The unrooted dendrograms of figure 8 are consistent with this notion, in that only the dendrogram limited to highly expressed genes reveals clear segregation between toxins from species with different feeding specificities. This may also explain the borderline statistics regarding diversifying selection in our tip test analyses, since the signal for diversifying selection on highly expressed toxins may be diluted by nonselected mutants in the overall conopeptide set. This possibility should be tested in the future by examination of a wider range of Conus species, with a focus on the most highly expressed variants.

Given that the Conidae are arguably the largest known family of marine invertebrates and their venom ducts are thought to be the most diverse in toxin content, it is interesting to speculate how the two processes outlined above might have contributed to the recent evolutionary success of the genus. Evidence has accumulated in recent years to suggest that the spectrum and rates of mutation may be under genetic control in certain systems, and thus evolution of a capacity for enhanced mutation may be adaptive for organisms confronted with variable environments (Caporale 1999, 2000; Radman, Matic, and Taddei 1999; Metzgar and Wills 2000). The premier example is, of course, somatic hypermutation in the immune system. Other examples include the SOS mutagenesis response of *E. coli* (Radman, Matic, and Taddei 1999), the enrichment of mouse MHC genes with conversion-enhancing CpG doublets (Metzgar and Wills 2000), and the surface antigen-switching mechanisms of trypanosomes (Metzgar and Wills 2000). The biology of offensive venom systems is such that evolution of a hypermutation mechanism targeted to venom-derived toxins would not incur a high fitness cost. Toxin genes are expressed only in the venom apparatus and do not directly participate in the organism's general physiology or metabolism; thus, there is very limited scope for deleterious effects upon the emergence of a new variant. Indeed, once gene duplication has produced a number of copies of a given toxin, the fitness cost incurred by mutating one copy away from the original function may be negligible. Increased expression of those variants important for prey capture will be adaptive and selected for (as, indeed, is observed in our study); thus, a ''background'' of lowly expressed sequence variants may accumulate in the genome. This battery of ''lazarotoxins'' (by analogy with the well-known phenomenon of rare Lazarus taxa that vanish and reappear in the fossil record; Wignall and Benton 1999) would then provide a bank of toxin variants available for further diversifying selection in times of environmental change, thus facilitating rapid speciation of the genus. This may hold also for other venomous taxa. The scenario is attractive in that it provides explanations for both the high number of toxin variants in Conus venoms and the high number of Conus species in shallow marine ecosystems.

Supplementary Material

GenBank accession numbers for sequences reported in this manuscript are AF193254–AF193272, AF214923–AF215131, and AF217316. All of the alignments are available in the PopSet section of GenBank. A table of cysteine codon conservation and a full set of tip test cladograms are available as supplementary information on the MBE web site.

Acknowledgments

We thank Doron Lancet and Edward Trifonov for stimulating discussions, and Alan R. Templeton for invaluable advice. M.F. is an Allon fellow and the incumbent of the Daniel E. Koshland Sr. Career Development Chair. This work was supported by the Biotechnology Infrastructure Program of the Israeli Ministry of Science and the Crown Genome Center at the Weizmann Institute.

LITERATURE CITED

BOWERSOX, S. S., and R. LUTHER. 1998. Pharmacotherapeutic potential of omega-conotoxin MVIIA (SNX-111), an N-type neuronal calcium channel blocker found in the venom of Conus magus. Toxicon **36**:1651–1658.

CAPORALE, L. H. 1999. Chance favors the prepared genome. Ann. N.Y. Acad. Sci. **870**:1–21.

———. 2000. Mutation is modulated: implications for evolution. Bioessays **22**:388–395.

CASTELLOE, J., and A. TEMPLETON. 1994. Root probabilities for intraspecific gene trees under neutral coalescent theory. Mol. Phylogenet. Evol. **3**:102–113.

CONTICELLO, S. G., Y. PILPEL, G. GLUSMAN, and M. FAINZILBER. 2000. Position-specific codon conservation in hypervariable gene families. Trends Genet. **16**:57–59.

DRAKOPOULOU, E., J. VIZZAVONA, J. NEYTON, V. ANIORT, F. BOUET, H. VIRELIZIER, A. MENEZ, and C. VITA. 1998. Consequence of the removal of evolutionary conserved disulfide bridges on the structure and function of charybdotoxin and evidence that particular cysteine spacings govern specific disulfide bond formation. Biochemistry **37**:1292–1301.

DUDA, T. F., and S. R. PALUMBI. 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.

ENDO, T., K. IKEO, and T. GOJOBORI. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

FAINZILBER, M., O. KOFMAN, E. ZLOTKIN, and D. GORDON. 1994. A new neurotoxin receptor site on sodium channels is identified by a conotoxin that affects sodium channel inactivation in molluscs and acts as an antagonist in rat brain. J. Biol. Chem. **269**:2574–2580.

FAINZILBER, M., T. NAKAMURA, A. GAATHON, J. C. LODDER, K. S. KITS, A. L. BURLINGAME, and E. ZLOTKIN. 1995. A new cysteine framework in sodium channel blocking conotoxins. Biochemistry **34**:8649–8656.

FIELD, M. C., and J. C. BOOTHROYD. 1996. Sequence divergence in a family of variant surface glycoprotein genes from trypanosomes: coding region hypervariability and downstream recombinogenic repeats. J. Mol. Evol. **42**:500–511.

FIJALKOWSKA, I. J., R. L. DUNN, and R. M. SCHAAPER. 1997. Genetic requirements and mutational specificity of the Escherichia coli SOS mutator activity. J. Bacteriol. **179**: 7435–7445.

FROY, O., T. SAGIV, M. POREH, D. URBACH, N. ZILBERBERG, and M. GUREVITZ. 1999. Dynamic diversification from a putative common ancestor of scorpion toxins affecting sodium, potassium, and chloride channels. J. Mol. Evol. **48**: 187–196.

HUGHES, A. L., and M. YEAGER. 1998. Natural selection and the evolutionary history of major histocompatibility complex loci. Front. Biosci. **3**:d509–d516.

KOHN, A. J. 1990. Tempo and mode of evolution in Conidae. Malacologia **32**:55–67.

KOHN, A. J., M. NISHI, and B. PERNET. 1999. Snail spears and scimitars: a character analysis of Conus radular teeth. J. Molluscan Stud. **65**:461–481.

KOHN, A. J., and J. W. NYBAKKEN. 1975. Ecology of Conus on eastern Indian Ocean fringing reefs: diversity of species and resource utilization. Mar. Biol. **29**:211–234.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis. Pennsylvania State University, University Park.

MCINTOSH, J. M., B. M. OLIVERA, and L. J. CRUZ. 1999. Conus peptides as probes for ion channels. Meth. Enzymol. **294**: 605–624.

MAOR-SHOSHANI, A., N. B. REUVEN, G. TOMER, and Z. LIVNEH. 2000. Highly mutagenic replication by DNA polymerase V (UmuC) provides a mechanistic basis for SOS untargeted mutagenesis. Proc. Natl. Acad. Sci. USA **97**: 565–570.

METZGAR, D., and C. WILLS. 2000. Evidence for the adaptive evolution of mutation rates. Cell **101**:581–584.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

NEI, M., X. GU, and T. SITNIKOVA. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc. Natl. Acad. Sci. USA **94**:7799–7806.

NORTON, R. S., and P. K. PALLAGHY. 1998. The cystine knot structure of ion channel toxins and related polypeptides. Toxicon **36**:1573–1583.

OHNO, M., R. MENEZ, T. OGAWA et al. (12 co-authors). 1998. Molecular evolution of snake toxins: is the functional diversity of snake toxins associated with a mechanism of accelerated evolution? Prog. Nucleic Acid Res. Mol. Biol. **59**: 307–364.

OLIVERA, B. M., C. WALKER, G. E. CARTIER, D. HOOPER, A. D. SANTOS, R. SCHOENFELD, R. SHETTY, M. WATKINS, P. BANDYOPADHYAY, and D. R. HILLYARD. 1999. Speciation of cone snails and interspecific hyperdivergence of their venom peptides. Potential evolutionary significance of introns. Ann. N.Y. Acad. Sci. **870**:223–237.

OPHIR, R., T. ITOH, D. GRAUR, and T. GOJOBORI. 1999. A simple method for estimating the intensity of purifying selection in protein-coding genes. Mol. Biol. Evol. **16**:49–53.

PAGE, R. D. 1996. TreeView: an application to display phylogenetic trees on personal computers. Comput. Appl. Biosci. **12**:357–358.

PALMER, S. J., M. R. REDFERN, G. C. SMITH, and J. P. L. COX. 1998. Sticky Egyptians: a technique for assembling genes encoding constrained peptides of variable length. Nucleic Acids Res. **26**:2560–2565.

PILPEL, Y., and D. LANCET. 1999. The variable and conserved interfaces of modeled olfactory receptor proteins. Protein Sci. **8**:969–977.

RADMAN, M. 1999. Enzymes of evolutionary change. Nature **401**:866–869.

RADMAN, M., I. MATIC, and F. TADDEI. 1999. Evolution of evolvability. Ann. N.Y. Acad. Sci. **870**:146–155.

RIGBY, A. C., E. LUCAS-MEUNIER, D. E. KALUME et al. (13 co-authors). 1999. A conotoxin from Conus textile with unusual posttranslational modifications reduces presynaptic Ca2+ influx. Proc. Natl. Acad. Sci. USA **96**:5758–5763.

ROONEY, A. P., J. ZHANG, and M. NEI. 2000. An unusual form of purifying selection in a sperm protein. Mol. Biol. Evol. **17**:278–283.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SASAKI, T., Z. P. FENG, R. SCOTT, N. GRIGORIEV, N. I. SYED, M. FAINZILBER, and K. SATO. 1999. Synthesis, bioactivity and cloning of the L-type calcium channel blocker ω-Conotoxin TxVII. Biochemistry **38**:12876–12884.

SCHIMMEL, P. R., D. SOLL, and T. ABELSON. 1979. Transfer RNA. Cold Spring Harbor Laboratory Press, New York.

SCHNEIDER, T. D., and E. M. STEPHENS. 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. **18**:6097–6100.

SHON, K. J., A. HASSON, M. E. SPIRA, L. J. CRUZ, W. R. GRAY, and B. M. OLIVERA. 1994. Delta-conotoxin GmVIA, a novel peptide from the venom of Conus gloriamaris. Biochemistry **33**:11420–11425.

TANG, M., P. PHAM, X. SHEN, J. S. TAYLOR, M. O'DONNELL, R. WOODGATE, and M. F. GOODMAN. 2000. Roles of E. coli DNA polymerases IV and V in lesion-targeted and untargeted SOS mutagenesis. Nature **404**:1014–1018.

TEMPLETON, A. 1996. Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. Genetics **144**:1263–1270.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, and D. G. HIGGINS. 1997. The CLUSTAL-X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**: 4876–4882.

WATANABE-AKANUMA, M., R. WOODGATE, and T. OHTA. 1997. Enhanced generation of A:T→T:A transversions in a recA730 lexA51(Def) mutant of Escherichia coli. Mutat. Res. **373**:61–66.

WIGNALL, P. B., and M. J. BENTON. 1999. Lazarus taxa and fossil abundance at times of biotic crisis. J. Geol. Soc. **156**: 453–456.

STEPHEN PALUMBI, reviewing editor

Accepted September 24, 2000