J Mol Evol (2008) 66:151–166
DOI 10.1007/s00239-008-9067-7

Rapid Evolution by Positive Selection and Gene Gain and Loss: PLA_2 Venom Genes in Closely Related *Sistrurus* Rattlesnakes with Divergent Diets

H. Lisle Gibbs · Wayne Rossiter

Received: 14 June 2007 / Accepted: 24 December 2007 / Published online: 6 February 2008

Abstract Rapid evolution of snake venom genes by positive selection has been reported previously but key features of this process such as the targets of selection, rates of gene turnover, and functional diversity of toxins generated remain unclear. This is especially true for closely related species with divergent diets. We describe the evolution of PLA_2 gene sequences isolated from genomic DNA from four taxa of *Sistrurus* rattlesnakes which feed on different prey. We identified four to seven distinct PLA_2 sequences in each taxon and phylogenetic analyses suggest that these sequences represent a rapidly evolving gene family consisting of both paralogous and homologous loci with high rates of gene gain and loss. Strong positive selection was implicated as a driving force in the evolution of these protein coding sequences. Exons coding for amino acids that make up mature proteins have levels of variation two to three times greater than those of the surrounding noncoding intronic sequences. Maximum likelihood models of coding sequence evolution reveal that a high proportion ($\sim$30%) of all codons in the mature protein fall into a class of codons with an estimated d_N/d_S (ω) ratio of at least 2.8. An analysis of selection on individual codons identified nine residues as being under strong ($p < 0.01$) positive selection, with a disproportionately high proportion of these residues found in two functional regions of the PLA_2 protein (surface residues and putative anticoagulant region). This is direct evidence that diversifying selection has led to high levels of functional diversity due to structural differences in proteins among these snakes. Overall, our results demonstrate that both gene gain and loss and protein sequence evolution via positive selection are important evolutionary forces driving adaptive divergence in venom proteins in closely related species of venomous snakes.

Keywords *Sistrurus* rattlesnakes · Snake venom genes · PLA_2 venom proteins · Positive selection · Gene family evolution · Functional diversity in proteins

H. L. Gibbs (✉) · W. Rossiter
Department of Evolution, Ecology and Organismal Biology, Ohio State University, 318 West 12th Avenue, Columbus, OH 43210-1293, USA
e-mail: gibbs.128@osu.edu

Present Address:
W. Rossiter
Department of Ecology, Evolution, & Natural Resources, Rutgers University, New Brunswick, NJ 08901-8551, USA

Introduction

Snake venoms have been described as one of the most sophisticated integrated weapons delivery systems in existence (Fry 2005) and represent an example of an extraordinary predatory adaptation. Interactions between venomous snakes and their prey have all the hallmarks of an escalating arms race. First, a high level of functional diversity in venom proteins at both the species and the population level (Daltry et al. 1996; Fry et al 2003; Sanz et al. 2006) is generated through strong positive diversifying selection acting on venom genes (Kordis et al. 2002). This diversity can have prey-specific effects (Heatwole and Powell 1998; Jorge da Silva and Aird 2001). Due to their potential biomedical importance, there is substantial knowledge on the structural composition and function of many of the diverse set of proteins found in venom, as well as the genes coding for these proteins (e.g., Menez 2002). Thus, snake venoms represent a system in which it may be possible to realize several important goals in evolutionary

biology. We can identify the molecular basis of adaptations in natural populations. Furthermore, we can examine the connections between genotype and phenotype for fitness-related traits at the molecular level in "nonmodel" organisms (Golding and Dean 1998; Feder and Mitchell-Olds 2003).

To date, much work has focused on characterizing venom gene and protein evolution at broad taxonomic levels among phylogenetically and ecologically diverse species (e.g., Slowinski et al. 1997; Fry 2005; Fry et al. 2003, 2006; Fry and Wuster 2004; Lynch 2007). These macroevolutionary studies are useful in revealing broad evolutionary patterns in venom evolution. However, they lack the ability to inform us about the evolutionary processes acting at the species or population level, which are the ultimate causes of venom variation at higher taxonomic levels. To address this gap, there have been an increasing number of studies that focus on describing venom evolution among small numbers of closely related species which differ in ecological characteristics that may select for diversification in venom genes and proteins (e.g., Chijiwa et al. 2003; Creer et al. 2003; Li et al. 2005a, b). Here, we provide a detailed examination of evolution of the genes coding for phospholipase A_2 (PLA_2) venom proteins, which are an important component of the venom of a group of closely related *Sistrurus* rattlesnakes with variable diets (Sanz et al. 2006).

Sistrurus are small North American rattlesnakes (Conant and Collins 1998). We report on venom variation in four taxa: *Sistrurus miliarius barbouri* (pygmy rattlesnake), *S. catenatus catenatus*, *S. c. tergeminus*, and *S. c. edwardsii* (eastern, western, and desert massasauga rattlesnakes, respectively). Recent phylogenetic analyses based on mitochondrial and nuclear DNA indicate that *miliarius* is basal to all three *catenatus* subspecies, whereas the named *catenatus* subspecies fall into two distinct clades: one consisting of *S. c. catenatus* alone and the other consisting of both *tergeminus* and *edwardsii* (Gibbs et al., unpublished data; see Fig. 2b). Diet studies show that different taxa of *Sistrurus* rattlesnakes vary in the degree to which they specialize on endothermic versus ectothermic prey (Holycross and Mackessy, 2002). Specifically, there are snakes that largely specialize on mammals (*S. c. catenatus*) versus frogs and lizards (*S. m. barbouri*) as well as snakes that bridge this dietary transition by utilizing both mammals and ectotherms (*S. c. tergeminus* and *S. c. edwardsii*). If venom composition is strongly related to aspects of diet (Daltry et al. 1996), then variation in venom proteins and hence the genes coding for these proteins should be observed among these related taxa.

PLA_2 proteins are commonly found in viperid snakes and show a range of distinct toxic properties that are likely to have functional significance for subduing different prey (review by Kini 1997). Our proteomics-based analyses of *Sistrurus* venom has shown that PLA_2 proteins are a major component of the venom of these snakes, making up between 14% and 32% of the overall protein content, depending on the taxon (Sanz et al. 2006). Recently, Lynch (2007) reported the results of a large-scale analyses of PLA_2 venom genes across a taxonomically diverse set of species. This work showed that positive adaptive evolution was common among PLA_2 loci and that new members of this gene family frequently arose though the neofunctionalization of duplicated gene copies. He also confirmed the results of Kini and Chan (1999) by showing that residues on the surface of the mature PLA_2 molecules are under strong selection. This is likely due to their role as "pharmacological sites" (cf. Kini and Evans 1989) which play a key role in protein-protein interactions between venom proteins and prey tissues. There is also evidence that variation in PLA_2 venom proteins at inter- and intraspecific levels is the result of diversification in PLA_2 genes via positive selection (Nakashima et al. 1993; 1995; Nobushisa et al. 1996; Gubensek and Kordis 1997; Chijiwa et al. 2000; Tsai et al. 2004). Other work (summarized by Chijiwa et al. 2003) links PLA_2 genetic and protein variation to differences in diet among island populations of the habu snake (*Trimeresurus flavoviridis*) (see also Creer et al. 2003). Recent work by Chen et al. (2004) has characterized N6 (basic) PLA_2 genes and proteins in several *Sistrurus* taxa, and Tsai et al. (2003) have isolated E6-PLA_2 (acidic) genes and proteins from a closely related rattlesnake (*Crotalus viridis*). However, a comprehensive analysis of the evolution of PLA_2 genes in *Sistrurus* as a whole has yet to be attempted.

Here we analyze PLA_2 gene sequences isolated from four *Sistrurus* taxa. We use a variety of techniques to assess patterns of evolution of these loci in a fully sampled clade. We aim to understand the role that gene family evolution and diversifying selection have played in generating functional diversity in the PLA_2 proteins encoded by these genes. Our study is novel in that we simultaneously address questions about the targets of selection, rates of gene turnover, and functional diversity of toxins generated that are rarely considered for congeneric taxa despite being essential for a comprehensive understanding of venom evolution.

Methods

Samples

We amplified PLA_2 loci from genomic DNA of single individuals from four *Sistrurus* taxa: *S. c. catenatus* (Killbear Provincial Park, Ontario), *S.c. termgeminus*

(Russell County, Kansas), *S. c. edwardsii* (Lincoln County, Colorado [gift of S. Mackessy]), and *S. miliarius barbouri* (Volusia County, Florida). As described above, recent phylogenetic analyses indicate that *miliaris* is basal to all three named *catenatus* subspecies, which fall into two distinct clades: one consisting of *S. c. catenatus* alone and the other consisting of *tergeminus* and *edwardsii* (Gibbs et al., unpublished data; see Fig. 2b). Despite their lack of phylogenetic distinctiveness, we analyzed both of the later taxa because each has distinct PLA_2 proteins in its venom (Sanz et al. 2006). We analyzed sequences from a single individual because this increases our ability to estimate the number of loci among sequence variants due to the fact that single diploid individuals can have a maximum of only two distinct alleles at a given locus. Finally, we used genomic DNA as a template rather than mRNA (e.g., Chijiwa et al. 2003) for several reasons. Use of a DNA template allows the generation of sequence from noncoding regions (e.g., introns and untranslated regions) that can then be used to compare rates of evolution between coding and noncoding regions and help establish phylogenetic relationships among sequences. Second, it is difficult to obtain venom gland tissue for mRNA extractions for one of the species (*S. c. catenatus)* which is protected at state, provincial, and federal levels in North America.

Cloning and Characterization of Loci

We amplified PLA_2 loci containing the entire coding region using previously described primers (PLA5FLS and PLA3FLAS [Kordis et al 1998]) which are located in the 5' and 3' flanking noncoding regions of these genes. The sequences of these primers are highly conserved ($\sim$95% sequence similarity) across all available published PLA_2 loci from different genera of venomous snakes (Rossiter, unpublished data) and thus should amplify most members of this gene family within each *Sistrurus* taxon.

Amplicons ($\sim$2.3 kb) containing PLA_2 loci were amplified using a two-step PCR as described by Kordis et al. (1998) using a high-fidelity polymerase (Platinum Taq; Invitrogen) to minimize polymerase errors in cloned products. We used the following PCR conditions: initial denaturation of 5 min at 95°C, 35 cycles of 1 min at 95°C and 6 min at 69°C, and a final extension of 10 min at 72°C. Products of appropriate size ($\sim$2.3 kb) were excised from agarose gels using a QIAquick Gel Extraction Kit, and products were cloned into a plasmid vector using a TOPO TA cloning kit. Thirty positive clones from each taxon were then sequenced on an ABI 3100 DNA Analyzer using a series of primers that cover the entire gene. Sequence contigs were then manually constructed in Sequencher and aligned in Bioedit 7.0.5 (http://www.mbio.ncsu.edu/BioEdit/bioedit.html), and distinct sequences identified.

Because our goal was to exhaustively sample all distinct sequences amplified from each taxon, we needed to be certain that all distinct sequences had been identified from each cloning experiment based on results from 30 clones. To do this we used randomization to generate a rarefaction curve based on the distribution of unique sequences among 30 clones that had been sequenced. By examining the inflection point of a curve based on the numbers of unique samples detected in replicate (n=100) samples of N = 1, 2, 3,..., 30 clones, we determined the likelihood of detecting all unique sequences among 30 clones.

We also generated an independent estimate of the number of distinct PLA_2 loci present in the genomes of these snakes using Southern blotting. Specifically, we first used standard protocols to generate a membrane (Hybond-N) which consisted of lanes of gel-separated snake genomic DNA (5 μg) from all four *Sistrurus* taxa that had been digested with *Eco*RI, *Hin*dIII, and *Bam*HI (three lanes per taxon). We then used a Roche DIG kit to carry out the Southern blot procedure, in which we probed the membrane with a conserved 168-bp portion of the PLA_2 gene that was 99%–100% similar among all distinct cloned sequences. The blot was then washed at high (0.2× SSC, 0.5% SDS) stringency and exposed to x-ray film. Based on predicted hybridization kinetics, this level of wash stringency should detect sequences with >85% nucleotide similarity.

Finally, we explored links between venom genotypes (PLA_2 gene sequences described here) and venom phenotypes (expressed proteins found in mature venom) for each taxa by comparing the predicted molecular weights of venom proteins that would be generated from the PLA_2 gene sequences with detailed information on the weights (in daltons) of mature PLA_2 proteins found in the venom of these snakes (Sanz et al. 2006). We used the ProtParam program (Gastiger et al. 2005) available at the EXPASY web site (www.expasy.ch/tools/protparam.html) to estimate both the predicted molecular weights (MW; excluding signal peptides) and the theoretical pI isoelectric points (pI; see below) of mature proteins as encoded by the PLA_2 genomic sequences which we isolated. For each taxon, we then compared these MWs with values for mature PLA_2 proteins as given in Table 6 of Sanz et al. (2006). We inferred that two proteins had the same MW, hence were classified as the same protein, if the MW of the sequence-inferred protein fell within the 95% confidence interval of the MW of the mature protein (±0.4% of the total MW as described by Sanz et al. 2006). This is a conservative criterion since it may underestimate the accuracy with which MWs are inferred from proteomics-based analyses and hence overestimate the numbers of

proteins that are assumed to be the same (J. Calvete, personal communication).

Phylogenetic Analyses

Determining the genealogical relationships of distinct PLA_2 sequences is important for understanding how evolutionary change within the family proceeds over time. We estimated phylogenetic relationships among distinct sequences in the following way. First, we generated an alignment of all distinct cloned sequences using the CLUSTAL-W server at EBI-EMBL (http://www.ebi.ac.uk/clustalw), using the default parameter for gap penalties. We inspected this alignment by eye and deleted a small number of regions (all in introns) where repetitive DNA made alignment problematic. After this step we judged that the placement of gaps generated by CLUSTAL was reasonable and that no further adjustments were necessary. The final alignment of these sequences is available from the authors upon request.

Next we used the alignment containing all available data (coding and noncoding regions combined) to estimate relationships among sequences under the justification that both regions of the gene contain information on genealogical relationships among sequences and that we maximized the availability of this information by using the complete data set. Using the steps below we also explored the kinds of trees that were generated using different subsets of the data (noncoding regions or coding regions only). In general, the major features of the topologies were similar for the different data sets with the notable exception that topologies based on coding regions were about four times more divergent than those based on noncoding regions (results not shown).

To choose the substitution model which best fit the combined data set, we used the online version of MODELTEST 3.7 (http://www.darwin.uvigo.es/software/modeltest.html; see Posada and Crandall 1998) in combination with PAUP 4.0b10 (Swofford 2003) to choose among different models using an AIC criteria. We then used PAUP to estimate the phylogenetic relationships among gene sequences using heuristic searches (TBR [tree bisection-reconnection] branch swapping and random sequence addition) under a maximum likelihood (ML) criterion using the parameter values from the "best-fit" model of sequence evolution identified by MODELTEST. We rooted this tree using a single PLA_2 sequence isolated from *Agkistrodon contortrix* using the same procedure applied to the *Sistrurus* taxa (Gibbs and Rossister, unpublished data). We used a Kishino-Hasegawa test (Hasegawa and Kishino 1989) implemented in PAUP to judge whether the tree with the highest likelihood value fit the data significantly better than other trees. Finally, we used bootstrap analyses (100 replicates) to estimate support for specific nodes in the gene tree with the highest likelihood value.

Analysis of PLA_2 Gene Evolution

Following Fry et al. (2003) and Slowinski et al. (1997), we used the program GeneTree (http://www.taxonomy.zoology.gla.ac.uk/rod/genetree/genetree.html; see Page 1998) to estimate the minimum number of gene duplications and deletions required to make the topology of the gene tree based on total PLA_2 sequences consistent with the species tree for these taxa. In light of the recent analysis of phylogenetic relationships among *Sistrurus* taxa based on mitochondrial and nuclear DNA markers (see above and Fig. 2b), all PLA_2 sequences for *edwardsii* and *tergeminus* were coded as belonging to the same taxon.

Evidence for Positive Selection

To look for evidence of positive selection among PLA_2 genes, we first compared patterns of substitution in coding versus noncoding regions of these genes. We used DnaSP 4.10.9 (Rozas et al. 2003) to estimate the number of substitutions per site (π) averaged across all sequences using a 50-bp sliding window for the Clustal W-generated alignment of all PLA_2 sequences (noncoding and coding).

Next, we used the CODEML subroutine in the PAML (ver. 3.15) package (Yang 1997) to estimate the proportion of codons within these gene sequences which show a d_N/d_S (ω) ratio >1, estimate what the ω ratio for sites in this class was, and identify individual sites that show evidence for positive selection. This analysis was conditioned on the topology of the gene tree in Fig. 2. We carried out this analysis for three sets of nested models of varying complexity, with each set consisting of a "neutral" model with no selected sites and a "selection" model which allows for a certain portion of sites under positive selection. We compared the following model pairs: M0 and M3, M1 and M2, and M7 and M8. Models differ in the way they assume that selection operates on classes of codons. All the neutral models (M0, M1, and M7) only allow sites with ω between 0 and 1, whereas all the selection models (M3, M2, and M8) allow an additional category of sites with ω >1. For each model pair, we determined whether the "selection" models (M3, M2, and M8) provided a better fit to the data than did the "neutral" models (M0, M1, and M7) by comparing twice the log-likelihood difference between the models to a χ^2 distribution where the degrees of freedom is

equal to the difference in the number of parameters used in any two nested models (Yang et al. 2000).

For the "selection" models we also used the site-by-site subroutine in CODEML to identify which individual residues were under positive selection as a result of having ω values significantly >1 ($p < 0.05$) based on site-specific Bayes empirical Bayes (BEB) probabilities (Yang et al. 2005). To examine the distribution of positively selected sites among putative functional regions (see below), we focused on the results from the M8-based analyses because the smaller number of sites detected under M3 and M2 models limit the statistical power to detect associations between sites and regions. We divided sites into two non-overlapping groups: "strongly selected sites" (SS), which had site-specific p-values < 0.01; and "weakly selected sites" (WS), which had p-values < 0.05 but > 0.01). We note that the SS category is a robust categorization of which codons are under selection because it contained sites identified as being positively selected under all three selection models (albeit at different levels of significance; see below).

Functional Variation in PLA_2 Amino Acid Sequences

To assess potential functional variation in translated PLA_2 sequences, we identified three nonexclusive classes of residues in the mature protein where variation may alter the functional properties of the protein (for reviews see Kini 2003, 2005).

Surface Residues

Kini and coauthors (Kini and Evans 1989; reviews by Kini 2003, 2005) have proposed the "target" model to explain the functional specificity of different PLA_2 venom proteins on prey tissue. Briefly they argue that residues on the surface of the PLA_2 molecule ("pharmacological sites") play a key role in the effects of PLA_2 proteins on specific tissues through interactions between pharmacological sites and target sites on the surface of tissues. To identify surface residues in the *Sistrurus* sequences, we followed the approach of Kini and Chan (1999) and used the published three-dimensional structure of a known PLA_2 venom protein from another rattlesnake (*C. atrox*; [Brunie et al. 1985]), to determine the surface accessibility of individual residues using the rolling-solvent method (Samanta et al. 2002) as incorporated in the program Swiss-model PRO (Peitsch 1996) to assign residues as either lying on the surface of the protein or buried in the interior. In this program surface accessibility values are grouped into five classes depending on the ease with which a small solvent (such as a water molecule) can interact with each residue. For our analyses we evaluated those amino acids that are in the class of residues (categories 5) that lie on the surface of the molecule and hence are most accessible and likely to contact cell surfaces. We also considered other residue classes (e.g., category 4) but did not analyze these further because the high degree of overlap of residues in this category with those in the anticoagulant region (see below) would make the interpretation of detected associations problematic.

Anticoagulant Region (Residues 54–77)

Kini (2005) described work on the anticoagulant properties of Type IB PLA_2 venom proteins from *Naja naja* which suggested that strong versus weak anticoagulant effects were strongly influenced by the identity of residues at sites 54–77. We examined variation in amino acid composition of the same region in *Sistrurus* Type IIA PLA_2 proteins (as determined through alignment of our sequences with the region identified in Fig. 4a of Kini 2005) under the assumption that it also influences anticoagulant properties in these snakes. We feel this is justified because although Type IB and IIA PLA_2 proteins have different evolutionary origins (nonvenomous pancreatic and synovial PLA_2s, respectively), hence different structures and possibly different functional residues, three different studies (Carredano et al 1998; Zhao et al. 2000; Singh et al. 2001) have provided independent evidence that location of the anticoagulant region is conserved in mature Type IB and Type IIA PLA_2 venom proteins (Kini 2005).

C-terminal Region (Residues 115–122)

As reviewed by Kini (2003), there is evidence that residue composition in this region influences the neurotoxic effects of PLA_2 proteins in *Vipera ammodytes* (Prijatelj et al. 2000, 2002; Petan et al. 2002) and myotoxic effects of these proteins in *Bothrops asper* (Lomonte et al. 1994). On this basis, we focused on variation in this region in *Sistrurus*.

Finally, similarities in amino acid composition of these regions between genes was assessed using a UPGMA clustering algorithim as implemented in the program MEGA, version 3 (Kumar et al. 2003).

Results

Characterization of PLA_2 Sequences

We identified four to seven clones containing distinct sequences from each species: seven from *S. c. catenatus*

and four from each of *S. c. edwardsii*, *S. c. tergeminus*, and *S. miliarius* (Appendix). We are confident that we exhaustively sampled the variation present in the cloned sequences because rarefaction curves for *S. c. catenatus* and *S. miliarius* show that both curves converge on the number of unique sequences identified (seven and four, respectively) after >20 clones from each species were sequenced (data not shown). Sequences differed by 0.023% divergence (5.06 substitutions/2200 bp), which is greater than that expected based on polymerase error rates (1.8×10^{-6}/base, or <1 erroneous base/2200 bp). The single exception was the ter 1 and 2 sequences, which differed by only 0.007% divergence. This value is still higher than that expected based on polymerase error and careful inspection of the cloned sequences showed that the substitutions that differed between sequences were found in multiple clones.

Sequences were between 2103 and 2309 in total length (DNA sequences of regions from exons 2, 3, and 4 which code for residues in the mature protein are shown in the Appendix). All but one total gene sequence (see below) displayed the structure typical of PLA_2 Type IIA loci (Danse et al. 1997), consisting of four exons and three introns with terminal codons in each of the first three exons, all interrupted at precisely the same position in the amino acid sequence: exon 1 is interrupted at Val(15) as a single-nucleotide overhang; exon 2 is interrupted at Arg(59), resulting in a two-nucleotide overhang; and exon 3 is interrupted at Lys(92), leaving a single-nucleotide overhang. When translated from nucleotide sequence the mature PLA_2 peptide is 122 amino acids long, a size that matches all other known members of the PLA_2 IIA family (John et al. 1994; Danse et al 1997). All translated proteins had the amino acid Asp at site 49 of the mature protein. As such, all isolated sequences fall into the Asp(49) subgrouping of PLA_2 venom proteins, which have high levels of enzymatic activity relative to the Lys(49) subgroup (Kini 1997). All introns conform to the GT-AG rule, which requires that splice sites at the 5'of each exon are GT and those at the 3'end are AG (Senepathy et al. 1990). We also detected the putative TATA box beginning 32 bp upstream of the initiation site, the poly(A) signal beginning 70 bp downstream of the stop codon. Finally, a phylogentic analysis based on exon coding sequences shows that a number of the sequences appear homologous with *Sistrurus* PLA_2 sequences isolated from a cDNA library by Chen et al. (2004) (results not shown). In summary, the conserved structure, a lack of frame-shift mutations, and homology with previously isolated expressed loci argue that these sequences represent functional PLA_2 loci in these snakes.

Based on the Southern blot experiments, more distinct bands were detected in *Hin*dIII- and *Eco*RI-cut DNA (H, mean across species, 5.75±1.26 [SD]; E, 5.25±0.96) than in *Bam*HI-cut DNA (3.25±1.25), but BamHI fragments were larger, suggesting the possibility of multiple tandemly repeated loci within single fragments. Based on the *Hin*dIII results, these values suggest that our PCR assay has sampled 100% (*catenatus*: 7 sequences/7 bands), 100% (*miliarius*: 4/4), 60% (*tergeminus*: 4/6), and 60% (*edwardsii*: 4/6) of the PLA_2 loci present (mean across species: 80%), while for the *Eco*RI data the values are >100% (*catenatus*: 7/6), 80% (*miliarius*: 4/5), 60% (4/6), and 100% (*edwardsii*: 4/4) (mean: 85%). We conclude that our PCR sequences represent ~80% of the PLA_2 loci present in the genomes of these snakes, although for some taxa (*tergeminus*) we may have isolated a lower proportion of existing loci.

Figure 1 shows the mature protein sequences inferred from all genomic sequences. All sequences differ by at least one residue with the exception of the ter 1 and 2 and ter 3 and 4 protein sequences, in which the members of each pair have the same sequence. All but one sequence (edw 4) have either a glutamic acid (E) or an asparagine (N) residue at position 6, and with one exception (edw 3), all E6 PLA_2 proteins have acidic (<7.0) theoretical pI's (Table 1). Six of eight N6 proteins have basic (>7.0) pI's,

```
                10        20        30        40        50        60        70        80        90        100       110       120
                ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ....|....| ..
mil1    NLLQFNKMIK IMTKKNAIPS YTSYGCYCGW GGRGRPKDAT DRCCFVHDCC YEKLTDCSPK TDIYSYSWKS GVITCGEGTP CEKQICECDR AAAVCFGENL PTYKKRYMFY PDFLCTDPSE KC
mil2    .......... .........F .......... .......... .......... ...P...... .......... .......... .......... .......... .......... .......... ..
mil3    H.....N... FE.N.....F .AF....... .......... ..G....... .......... .......... .......... .......... .......... .......... .......... ..
mil4    H.I..ETL.M KIAGRSGVFW .SA....... ..Y...Q... .......... .G.V.G.N.. K.F...TEEN .E.V..GDD. .K........ V..I..RD.I ...DDK.WRF .TEN.QEEP. P.
cat1    .......... .......... .......... .......... .........F .G.V...D.. ..F....EEN .Q.V..GDD. .K.......K .......... .......... .......... T.
cat2    H......... FE.N.....F .AF....... .......... .......... .G..PN.DT. W......L.. .F....G..W .K.......K ...I.LR... D..N.E.Q.. GH.H.KEGPK ..
cat3    H......... FE.N.....F .AF....... .......... .......... .......... .......... ......K..W ..E....... V..E.LRRS. S...YG.... ..SR.KG... Q.
cat4    S.VE.ETL.M KIAGRSGVWY .S.......T ..Q.W.Q..S .....E.... .A...G.D.I ..V.T.RQED .E.V.RRED. .GT....... ...I..RDSM D..NHK.WGV SLEN.QGETQ P.
cat5    S.VE.ETL.M KIAGRSXVWY .S.......T ..Q.W.Q..S .....E.... .A...G.D.I ..V.T.RQED .E.V..GED. .GT......K ...I..RDSM D..NHK.WRF SLEN.QGE.Q P.
cat6    H.I..ETL.M KIAGR.GVFW .SA....... ..YA..Q.TS .......... .G.V.G.N.. K.F....EEN .E.V..GDD. .K........ V..I..RD.I ...DDK.WRF .TEN.QEEP. P.
cat7    H.I..ETL.M KIAGRSGVFW .SA....... ..Y...Q.P. .......... .G.V.G.N.. K.F....EEN .E.V..GED. .KN...... V..I..RD.I ...DDK.WRF .TEN.QEEP. P.
ter1    .......... .........F .S........ .......... .......... .......... ..T....... ...I..GND. .K.......K .......... S......... .......... T.
ter2    .......... .........F .S........ .......... .......... .......... ..T....... ...I..GND. .K.......K .......... S......... .......... T.
ter3    ..I..ETL.L KVA..SGMF. .SA....... ..H...Q... .......... .G.V...D.. L.T.T..EEN .E.I..GDD. .K.......K ...I..RD.K N..NNK.WRL .TEN.QEEP. P.
ter4    ..I..ETL.L KVA..SGMF. .SA....... ..H...Q... .......... .G.V...D.. L.T.T..EEN .E.I..GDD. .K.......K ...I..RD.K N..NNK.WRL .TEN.QEEP. P.
edw1    ..I..ETL.L KVA..SGMF. .SA....... ..H...Q... .......... .G.V...D.. L.T.T..EEN .E.I..GDDQ .K.......K ...I..RD.K N..NNK.WRL .TEN.QEEP. P.
edw2    ..I..ETLTL KVA..SGMF. .SA....... ..H...Q... .......... .G.V...D.. L.T.T..EEN .E.I..GDDQ .K.......K ...I..RD.K N..NNK.WRL .TEN.QEEP. P.
edw3    ..I..ETL.L KV....GM.. ..F......L ..Q...R... .......... ..N..G.KT. .NP.P..R.N .A.V..GDDQ .K.......K ...I..RD.K N..NNK.WRL .TEN.QEEP. P.
edw4    S.VE.GR... EE.G..PF.. ..F......L ..Q...R... .......... ..N..G.KT. .NP.P..R.N .A.V..G..W .K.......K ...I.LR... D..N.E.Q.. GH.H.KEGPK ..
```

Fig. 1 Mature protein sequence for *Sistrurus* PLA_2 sequences

Table 1 Predicted molecular masses (M) and theoretical pI values estimated using ProtParam for translated PLA_2 proteins shown in Fig. 1

Sequence name	Residue at position 6	Predicted M (Da)	Theoretical pI	Corresponding PLA_2 protein from Sanz et al. (M; Da)[a]
ter 1	N	13,886	8.36	Sct-18 (13,887)
ter 2	N	13,886	8.36	Sct 18 (13,887)
mil 1	N	13,907	8.19	?
mil 2	N	13,951	8.19	Smb-15 (13,953)
mil 3	N	13,978	6.84	Smb-19 (13,980)
cat 1	N	13,969	6.70	Scc-16 (13,967)
cat 2	N	14,110	8.62	?
cat 3	N	14,128	8.19	Scc-14, (14,120)
edw 1	E	13,874	4.97	Sce-25 (13,842)
edw 2	E	13,861	4.97	Sce-23 (13,856)
ter 3	E	13,843	4.97	Sct-20 (13,826)
ter 4	E	13,843	4.97	Sct-20 (13,826)
cat 6	E	14,039	4.83	Scc-12 (13,952)
cat 7	E	14,008	4.72	Scc-18 (13,941)
mil 4	E	13,996	4.83	Smb-18 (13,963)
cat 4	E	13,800	4.49	?
cat 5	E	13,761	4.39	?
edw 3	E	13,881	8.19	?
edw 4	G	13,818	8.50	?

Note. A "?" indicates that no matching protein was identified. Names (M; Da) of *Sistrurus* PLA_2 proteins from Table 6 of Sanz et al. (2006) with no matching gene sequence: ter—Sct-17(13,741) and Sct-19 (14,119); mil—Smb-18 (13,962); cat—Scc-13 (13,880) and Scc-17 (13,952). Two-chain PLA_2's (e.g., Scc-8 and Scc-9) are not included in this list

[a] Our "best guess" as to the corresponding PLA_2 protein (if any) from Table 6 of Sanz et al. (2006) that matches the protein inferred from the gene sequence

whereas the other two sequences (cat 1 and mil 1) have pI's close to neutral. The division of *Sistrurus* PLA_2s into these two groups based on the residue at the sixth position is typical of that observed for PLA_2 in other viperid snakes including rattlesnakes (Tsai et al. 2003; Chen et al. 2004).

A mean of just over half (55%: range, 44.4%–66.7%) of all the PLA_2 gene sequences in a given taxa could be matched on the basis of molecular weight with proteins identified in the proteomics analyses by Sanz et al. (2006) (Table 1). A substantial number of gene sequences (mean, 28.5%; range, 0%–50%) had no matching proteins and a mean of 18.9% (range, 0%–33.3%) of proteins did not have matching gene sequences (Table 1). Thus the link between genotype and phenotype for these proteins is only moderate and varies between species.

Phylogenetic Analyses and Gene Family Evolution

Based on MODELTEST, the substitution models with the highest AIC scores for the combined PLA_2 sequences were TVM+I+G model (AIC score: 19,211.625) and the GTR+I+G model (19,211.875). These models are similar in parameter complexity; we choose to use the more general GTR+I+G model for further analyses. Figure 2 shows the rooted ML tree generated for the PLA_2 sequences, with bootstrap support values of >50% shown for specific nodes. The likelihood of this "best-fit" tree was significantly greater ($p < 0.001$) than that of the tree with the second-highest likelihood.

The topology of the best-fit gene tree does not parallel that of the "species" tree for these snakes, but rather clades consist of either sequences from different combinations of taxa (e.g., clade containing mil 1–3, ter 1 and 2, and cat 1–3) or taxon-specific clades which only contain sequences from a single *Sistrurus* taxa (e.g., cat 4 and 5). This topology is consistent with a multigene family whose members are undergoing a birth-and-death process in which the appearance and disappearance of particular loci are being driven by selection (Fry et al. 2003; Nei and Rooney 2005).

We used GeneTree to quantify the number of gene duplications and losses required to make the topology of the PLA_2 gene tree consistent with the species tree under a parsimony criterion. As required, we assumed a monophyletic origin of all extant sequences and that each sequence represents a separate locus. Also, to be consistent with the current hypothesis about relationships among these snakes, we classified all ter and edw sequences as being from a single taxon. Under these assumptions, GeneTree estimated a minimum number of 23 gene duplication or loss events (12 duplications/11 losses) have occurred, as this gene family has evolved in these three species since the origin of this group $\sim$10 myr bp (Douglas et al. 2006). Based on these values we estimate that rates of gene duplication are 0.40/species/myr, while rates of gene loss are 0.37/species/myr.

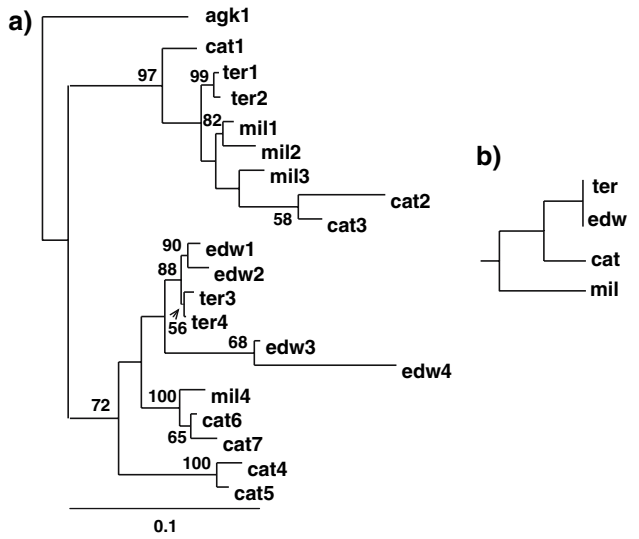

Fig. 2 **a** Rooted maximum likelihood tree showing phylogenetic relationships among overall *Sistrurus* PLA_2 gene sequences. Tree is rooted with *Agkistrodon contortrix* PLA_2 sequence (agk1). Taxon names are as in Fig. 1. Support values for clades with >50% bootstrap support are shown at the nodes of those clades. **b** Assumed phyologenetic relationships among *Sistrurus* taxa studied here (see above)

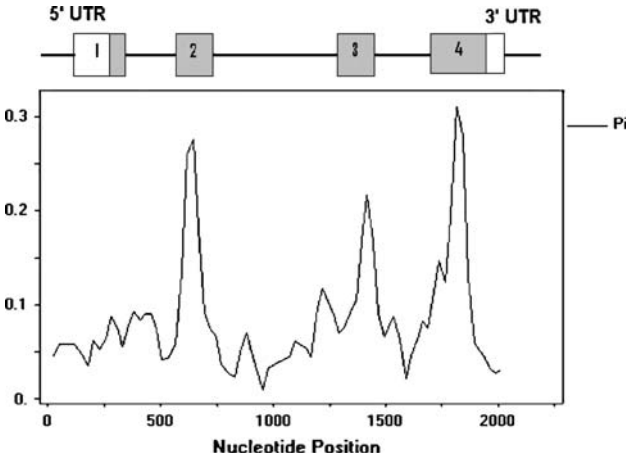

Fig. 3 Sliding window analysis of substitution rates per site (pI values) across all PLA_2 sequences. Note that peaks in substitution rates correspond to the exon regions that code for amino acids that make up the mature peptide

Evidence for Positive Selection on PLA_2 Loci

Both the increase in substitution rates in coding regions and the codon-based analysis of substitution patterns argue that positive selection drives the evolution of these genes. A sliding window analysis of substitutions patterns across all sequences shows that substitution rates are $\sim$2.5 times higher in exons 2–4 compared to untranslated regions (introns, 3' and 5' UTR regions) of these genes (Fig. 3). This is the expected pattern if positive selection acts to diversify functional variation in the coding regions of members of a gene family.

PAML analyses using CODEML show that for all three pairs of models, a model which incorporates a class of sites under positive selection fits the data significantly ($p < 0.001$) better than one which does not based on a log-likelihood ratio test (Table 1). All "selection" models infer that a high proportion of codons (range: 30%–32%) in these genes is under strong positive selection (range of estimated d_N/d_S ratio for positively selected sites [ω_2]: 2.84–2.98) (Table 2). A site-by-site analysis using the NSsites subroutine in CODEML under each "selection" model identified variable numbers of sites that were under significant ($p < 0.05$) positive selection, with more of these sites identified under the M8 selection model (25/138 codons; 18.1%] than the other two models (both, 11/138 codons; 7.9%) (Table 2). However, there is overlap in selected sites under the different models: all sites identified as under selection in M3 and M2 were also classified as under selection in M8; the M8-based analyses simply identified additional codons as being under selection (data not shown).

Table 3 shows that a significantly greater proportion of SS sites is found among class 5 (p= 0.039) residues compared to the rest of the molecule, whereas no such association was found for WS sites ($p > 0.19$) (Table 3). A similar pattern is found for positively selected sites in the putative anticoagulant region: significantly more SS were found in this region (p=0.012), but this was not the case for WS (p =0.531) (Table 3). The significant associations for the surface residue and anticoagulant residues are largely based on independent sets of sites under positive selection since there is only a single selected residue in common between these classes. Finally, there were no SS sites identified in the C-terminal region and no association between the numbers of WS present (p=0.152). Thus, strong positive diversifying selection on *Sistrurus* PLA_2 venom genes targets two largely nonoverlapping sets of residues: surface residues likely involved in the pharmacological targeting of specific prey tissues and a region which is known to determine hemorrhagic properties in other venomous snakes. Figure 4 shows the location of the nine strongly selected residues on a three-dimensional structure of the mature PLA_2 protein.

Functional Diversity in PLA_2 Venom Proteins

We characterized functional diversity in these proteins within and between taxa in two ways: by focusing on variation in surface residues and residues in the putative anticoagulant region and by assessing the diversity in proteins from functional classes of PLA_2 proteins identified

Table 2 Parameter estimates and likelihood scores under models of variable d_N/d_S ratios (ω) among sites for *Sistrurus* PLA_2 genes

Nested model pairs	ω_0	Parameter estimate(s)	PSS	Likelihood	Significance
M0: one-ratio (1)	0.935	$\omega = 0.935$	—	−2313.77	$p < 0.001$
M3: discrete (5)	1.240	$p_0 = 0.32$, $\omega_0 = 0$	WS = 8	−2235.39	
		$p_1 = 0.36$, $\omega_1 = 0.94$	SS = 3		
		$p_2 = 0.31$, $\omega_2 = 2.88$			
M1: neutral (1)	0.642	$p_0 = 0.36$, $\omega_0 = 0.013$	—	−2250.93	$p < 0.001$
		$p_1 = 0.64$, $\omega_1 = 1$			
M2: selection (3)	1.260	$p_0 = 0.32$, $\omega_0 = 0$	WS = 8	−2235.41	
		$p_1 = 0.38$, $\omega_1 = 1$	SS = 3		
		$p_2 = 0.30$, $\omega_2 = 2.98$			
M7: β (2)	0.700	$p_0 = 0.0158$, $q = 0.00635$	—	−2251.65	$p < 0.001$
M8:β & ω (4)	1.248	$p_0 = 0.0158$, $p = 0.0229$, $q = 0.026$	WS = 16[a]	−2235.48	
		$p_1 = 0.32$, $\omega_1 = 2.84$	SS = 9		

Note. The numbers of parameters estimated under each model is given in parentheses. The ω_0 value is the overall estimate of the d_N/d_S ratio across all sites for the specific model. Parameter estimates are the proportion of sites (p) falling into different classes of sites with different estimated ω values; for each selection model values for p_2 and ω_2 represent estimates of these parameters for sites under positive selection. PSS is the number of weakly (WS) and strongly (SS) selected sites (see above). Likelihood is the log-likelihood value for the model. Significance is the statistical significance of the more heavily parameterized selection model compared to its neutral partner under a chi-square distribution

[a] Residue number of selected sites in the mature protein: WS—3, 12, 20, 23, 30, 33, 52, 54, 74, 100, 109, 111, 112, 113, 119, and 121; SS—11, 14, 58, 61, 63, 68, 72, 80, and 101

on the basis of charge. These are acidic E6 proteins which show antiplatelet activities and basic N6 proteins which exhibit neurotoxic and myotoxic functions at least when tested on mammalian tissues (Tsai et al. 2003; Chen et al. 2004).

First, we sought to identify sets of sequences with similar amino acid composition in defined functional regions under the assumption that a similar composition would reflect similarity in function. Thus our analysis was explicitly not phylogenetic but rather was based only on similarity in amino acid composition in functionally defined regions. To do this we used an algorithm (UPGMA) which clusters sequences on the basis of the degree of their overall similarity in residue composition. Figure 5 shows UPGMA trees based on uncorrected percentage difference in amino acid residues for (a) all class 5 surface residues and (b) residues in the putative anticoagulant region (54–77). Both trees show similar features: first, a high diversity of residues within these regions (surface region, 13 of 19 sequences [68.4%] differ by one or more residue(s); anticoagulant region, 12 of 19 [63.2%]), implying a high level of functional diversity among sequences. In addition, the net charge of residues in the anticoagulant region which influences the anticoagulant properties of the protein varies from –5 to +4. Proteins which show net positive charges in this region have strongly anticoagulant properties, whereas proteins that are neutral or negatively charged exhibit weak anticoagulant properties (Kini 2005). Second, as for the gene tree in Fig. 2a, sequences fall either into taxon-specific clusters (e.g., edw 1 and 2 in Figs. 5a and b) which may identify taxon-specific venom components or into groups consisting of sequences from multiple taxa (e.g., cat 6 and 7 and mil 4 in Fig. 5a) which may identify more generic "housekeeping" venom proteins that serve common functions in different species. This type of "generic" protein is rare among the anticoagulant residues, as sets of proteins from different species that cluster together (e.g., cat 6 and 7 and mil 4) still have very different net charges, implying possible functional differences.

Finally, all taxa have both acidic and basic proteins in roughly equal proportion (Table 4). This implies that all venoms have both hemorrhagic and neurotoxic and myotoxic functions due to the presence of both types of PLA_2 proteins (Tsai et al. 2003; Chen et al. 2004).

Discussion

Our most important results are (1) our assessment of the link between venom genotype and phenotype for these proteins, (2) an explicit estimate of the high rate of turnover in this gene family over a relatively short evolutionary time span, (3) the use of phylogenetically informed codon-based models to identify specific amino acid residues which are targets of positive selection, and (4) the demonstration of links between sites under positive selection

Table 3 Relative frequency of two classes of positively selected sites (strongly and weakly selected) in three putative functional regions (class 5 surface residues, putative anticoagulant region [residues 54–77] and C-terminal region [residues 115–122]) relative to the rest of the mature PLA_2 protein

	Target region	Rest of protein	*p*-value[a]
a. Strongly selected sites			
Class 5 surface residues			
Sites under positive selection	4	5	0.039
Nonselected sites	17	96	
Putative anticoagulant region			
Sites under positive selection	5	4	0.012
Nonselected sites	18	95	
C-terminal region			
Sites under positive selection	0	9	0.531
Nonselected sites	9	105	
b. Weakly selected sites[b]			
Class 5 surface residues			
Sites under positive selection	4	15	0.185
Nonselected sites	13	81	
Putative anticoagulant region			
Sites under positive selection	0	9	0.531
Nonselected sites	9	105	
C-terminal region			
Sites under positive selection	3	20	0.152
Nonselected sites	5	85	

[a] For a Fisher's exact test comparing the relative number of selected vs. nonselected sites in each protein region

[b] Analysis excludes all sites under strong selection

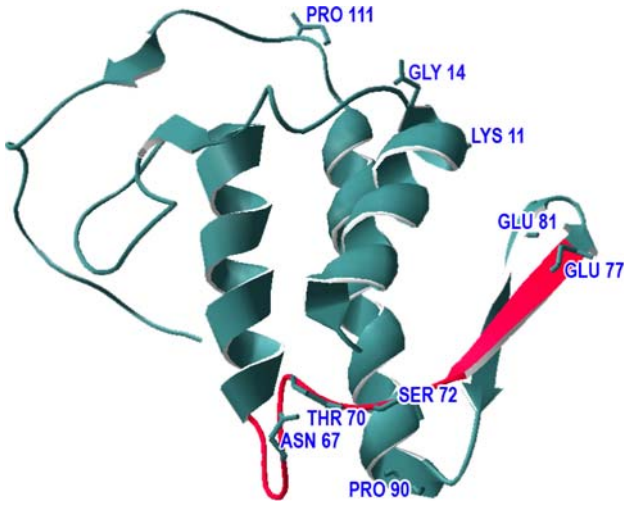

Fig. 4 Location of strongly selected residues in relation to the three-dimensional structure of the *Sistrurus* PLA_2 protein. The structure is represented by a ribbon, with disulfide bonds and catalytic residues shown as sticks. The number, consensus amino acid, and location of residues under strong selection are shown in blue lettering. Residues (54–77) which make up the putative anticoagulant region are in red

and functionally important regions of the mature PLA_2 protein. We discuss each of these findings below.

Venom Genotype and Phenotype

Assessing the nature and strength of the link between genotype (here, PLA_2 gene sequences) and phenotype (mature PLA_2 proteins found in venom) is an important goal of evolutionary biology (cf. Stern 2000; Feder and Mitchell-Olds 2003). This link is only moderate in strength for these genes and the proteins they encode since matching proteins were identified for only about half the gene sequences that were isolated.

A similar pattern occurs in other venomous snakes. In some cases, a tight relationship between PLA_2 sizes inferred from cDNA sequences and the MW of mature exists for other rattlesnakes including some *Sistrurus* species (Tsai et al. 2003; Chen et al. 2004). However, there are also well-known exceptions such as the PLA_2 two-chain proteins (see Sanz et al. 2006) found in *catenatus* and *tergeminus* venoms, which are much smaller than predicted based on gene sequences. As discussed by Sanz et al. (2006) these are two-chain PLA_2 proteins (e.g., Sistruxin A [see Chen et al. 2004]) which are encoded by genes similar to those which encode the larger PLA_2 proteins, but the precursor protein is posttranslationally proteolitically processed, resulting in a mature subunit protein ~9500 D in size (see Fig. 4 of Chen et al. 2004).

Other studies where broad comparisons are made between sequences isolated from venom gland cDNA libraries (venom gland "transcriptomes") and mature proteins found in expressed venom (venom "proteomes") transcripts are often found that have no matching expressed protein (cf. Junqueira-de-Azevedo et al. 2002, 2006). These observations argue that translational and/or regulatory effects may play an important role in how the venom genotype is transformed into the venom phenotype in these species. The opposite situation, in which venom proteins are found with no matching gene, may be due to incomplete sampling of all functional loci (suggested by our Southern blot results) and/or translational changes in expressed proteins.

Finally, we note that the strong link between genes and proteins for some PLA_2 venom proteins means that, at least for these loci, studies of how nonsynonymous substitutions in PLA_2 gene sequences evolve may provide direct insights into the adaptive evolution of venom proteins independent of translational and/or regulatory effects (cf. Hoekstra and Coyne 2007).

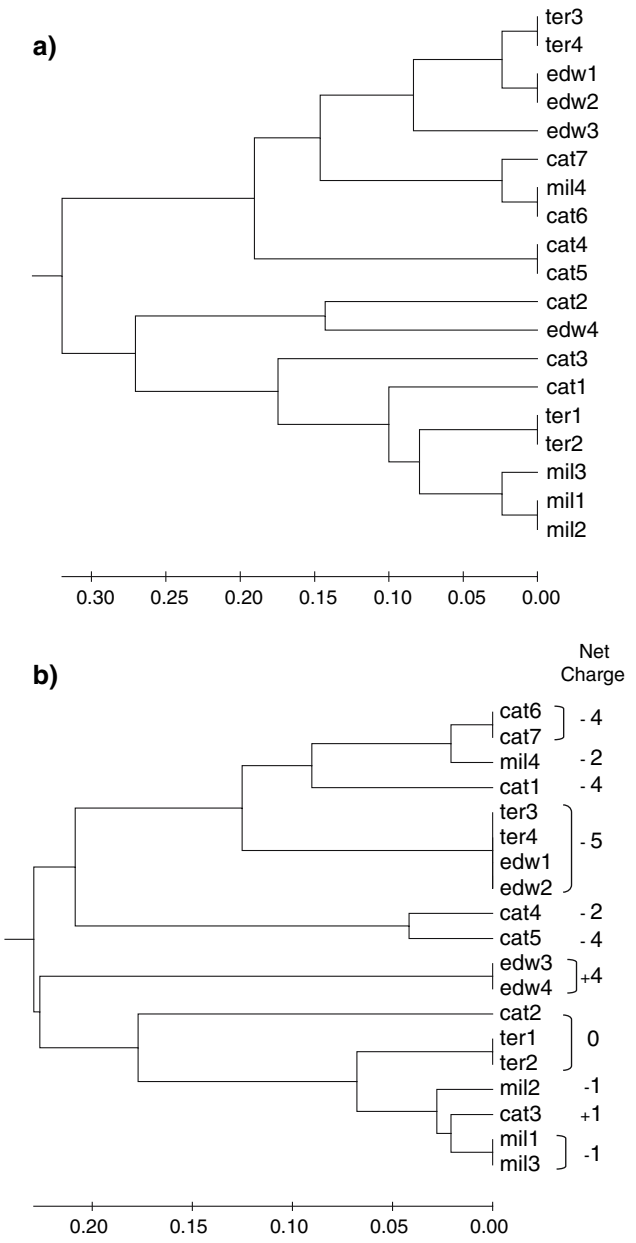

Fig. 5 UPGMA trees based on uncorrected *p*-distances showing similarity among PLA_2 sequences for two classes of amino acid residues. **a** Category 5 residues: fully exposed surface residues. **b** Residues 54–77: putative anticoagulant region. Estimated net charge of residues in region is shown to the right of the sequence name

Table 4 Relative numbers of acidic, basic, and neutral PLA_2 proteins in each *Sistrurus* taxon as inferred from theoretical pI values: *N* (% total)

Taxon	Acidic proteins	Basic proteins	Neutral proteins
miliarius	1 (25%)	2 (50%)	1 (25%)
catenatus	4 (57%)	2 (29%)	1 (14%)
tergeminus	2 (50%)	2 (50%)	0
edwardsii	3 (50%)	2 (50%)	0

PLA_2 Gene Family Evolution

Based on their studies of venom gene evolution across a phylogenetically diverse array of species, Fry et al. (2003) argued that a "birth-death" model was the mechanism that best described venom gene family evolution in snakes. Our results indicate that this mechanism also drives venom gene family evolution in closely related species over much shorter evolutionary time scales and provides the first estimates of the rates of gene gain (0.40 locus/species/myr) and loss (0.37 locus/species/myr) for a venom gene family. We have also found direct evidence for an "extinction" event in the form of a distinct PLA_2 DNA sequence that was identified via cloning from *edwardsii* and classified as a pseudogene due to the presence of a 37-bp frameshift insertion in an exon (data not shown). What is striking is that both rates are one to two orders of magnitude higher than mean estimates of rates of gene gain and loss based on studies of paralogous loci in humans. For example, Cotton and Page (2005) estimated mean rates of gene gain and loss to be 0.0012/gene/myr and 0.074/gene/myr, respectively, while earlier estimates by Lynch and Connery (2000, 2003) using different methods of analyses were higher (mean gain, 0.009/myr/gene; mean loss, 0.0924). One reason for this difference in rates is that at least one set of human estimates (Lynch and Connery 2000, 2003) explicitly avoided including large multigene families which may more closely resemble venom genes in their evolutionary dynamics.

Our estimates are based on several important assumptions. The first is that each of the sequences represents a distinct locus. If some of the distinct sequences represent allelic variation at a single locus, we would overestimate the per locus estimates of gains and losses. Second, we assume that our phylogenetic analysis has accurately recovered the true evolutionary history of these sequences in the form of the gene tree shown in Fig. 2. Finally, we assume that we have isolated all PLA_2 loci present in the genomes of these snakes. Our blotting results suggest that this is not true for all species which could lead to an underestimate of gene gains or an overestimate of gene losses. However, these results also suggest that the number of missing loci is small.

Overall, we feel it is unlikely that violations of these assumptions would alter our general conclusion that evolutionary turnover in venom gene families through a birth-and-death process may be at least an order of magnitude greater than that found in most other gene families. This implies that gene gain and loss are an important evolutionary force in the adaptive evolution of venom genes even among closely related species of venomous snakes (cf. Demuth et al. 2006). Further, the extremely high level of venom gene turnover hints that an unknown genetic

mechanism may allow unusually high rates of venom gene turnover to occur, leading to exceptionally rapid adaptive divergence through gene gain and loss when speciation occurs. Further estimates of venom gene duplication and extinction rates for other loci in other taxa are needed to assess whether these high rates are also found in other venomous snakes.

Positive Selection on Sistrurus PLA_2 Venom Genes

Several lines of evidence point to a significant role for positive selection in the evolutionary diversification of PLA_2 venom genes in *Sistrurus*. First nucleotide diversity is higher in exons 2–4 than in noncoding regions of these genes. Exon 1 shows less variation than the other three exons, likely because it contains the 5'UTR in PLA_2 genes and is not subject to the same level of selection as the other exons which code for residues that make up the mature protein. This pattern has been repeatedly found in other studies of PLA_2 snake venom loci (Nakashima et al. 1993, 1995; John et al. 1994; Kordis and Gubensek 1996; but see Kordis et al. 1998) and is interpreted to represent evidence for strong positive selection promoting functional diversification in coding regions of genes. This pattern is found in other gene families under positive selection (e.g., major histocompatibility complex genes in primates [Cereb et al. 1997; Tasuya et al. 2003] and sperm fertilization genes in abalone [Metz et al. 1998]) and differs from the pattern shown in many other genes, likely under purifying selection where synonymous substitution rates in introns and exons are similar (Hughes and Yeager 1997).

More compelling are the results of our codon-based analyses of positive selection, which have only rarely been used in analyses of the role of positive selection in venom gene evolution (however, see Lynch 2007). Previous studies of snake PLA_2 loci have detected evidence for positive diversifying selection based on d_N/d_S ratios greater than estimated over entire exons (e.g., Nakashima et al. 1993, 1995; for review see Ohno et al. 2003). However, as pointed out by Yang (2002) this approach has both limited resolution and limited power because it in effect averages substitution rates over all amino acids in a sequence. Because most amino acid sites are likely to be conserved, with adaptive evolution only affecting a few sites at any given time point (Gillespie 1994), the use of exon-wide d_N/d_S ratios will both fail to identify specific residues that are under selection and underestimate the strength of such selection. Our results support this claim. Based on three selection models which make different assumptions of how values for d_N/d_S are distributed across codons classes only one-third of sites as being under positive selection. However, these sites have an estimated overall d_N/d_S ratio of $\sim$2.8, which is one to two times higher than is typically observed in whole-exon comparisons (e.g., Ohno et al. 2003).

Further, site-by-site analysis reveals the possible functional consequences of this selection by showing that residues under strong (but not weak) positive selection are clustered in two functional regions of the protein. The first region includes residues on the surface of the mature protein which likely represent pharmacological sites involved in targeting the protein to specific tissues in prey. The second region includes residues which determine the anticoagulant properties of these proteins in other venomous snakes. Previous large-scale surveys of both class I and class II PLA_2 venom genes have also identified a high rate of evolutionary change in surface residues over a much more phylogenetically diverse set of species (Kini and Chan 1999; Lynch 2007). Our results refine this analysis by providing evidence that such selection also operates over much shorter evolutionary timescales involved in adaptive radiations consisting of small numbers of closely related species.

We are not aware of other analyses that have shown that the putative anticoagulant region is a target of positive selection in species other than *Naja naja*. Our interpretation of this pattern relies on the assumption that the functional characterization of this region which has been carried out for *Naja naja* (see Kini 2003, 2005) also applies to *Sistrurus.* This has not been demonstrated, but as described above, there is strong evidence that this region is conserved between the Type IB venom PLA_2's found in elapid snakes such as *Naja naja* and the Type IIA PLA_2's found in viperid snakes such as *Sistrurus* (Kini 2005). Our results provide a rare example of a direct link between the evolutionary and the functional diversification of venom proteins in these rattlesnakes: positive selection is acting to adjust which prey tissues and/or species are targeted and how the proteins act on those tissues, and this variation may be largely related to anticoagulant and/or hemolytic activity. To use the "weapons system" analogy (Fry and Wuster 2004), selection at the interspecific level has resulted in diversification in both the "targeting system" and the "payload" among PLA_2 venom proteins in these snakes. Future studies could test this with a series of in vitro and in vivo functional assays (cf. Tsai et al. 2003) in which the activity of PLA_2 proteins isolated from different *Sistrurus* taxa could be tested on platelets from different potential prey to see if there is variation in activity levels in relation to the abundance of prey in the diets of different snakes.

Our mapping of positively selected sites to functional regions of the PLA_2 protein relies on the assumption that such sites are accurately identified by the codon-based models in PAML. There have been suggestions that such codon-based approaches may result in the identification of

a significant number of false positives in terms of sites under positive selection (Suzuki and Nei 2004). We are confident that sites under positive selection were accurately identified for two reasons. First, the SS sites for which we found associations were identified as being under selection with a high degree of statistical significance ($p < 0.01$) under the M8 model and were also identified as being under selection in other models, albeit at lower levels of significance. Consistent identification of selected sites under different selection models argues that the identification of selected sites is robust (J. Bielawski, personal communication). Second, recent simulation analyses (Wong et al. 2004; Yang et al. 2005) argue that codon-based based methods of detecting selection, specifically the BEB method of detecting selected sites, are robust for both the numbers of sequences and the levels of divergence in this study.

Functional Divergence Among PLA_2 Venom Proteins

Figure 5 shows that there is substantial variation in the amino acid composition of two functional classes of residues among different loci in different snakes. However, as discussed above the precise functional significance of this variation remains unclear. In particular, the biochemical function of PLA_2 molecules with either a neutral or a negative charge in residues 54–77 remains obscure since research summarized by Kini (2005) suggests that, at least in *N. naja*, strong antiplatelet activity is only observed when the net charge in this region is positive. In fact only 2 of 12 distinct sequences show net positive charges, with the rest showing either no net charge or a net negative charge for residues in this region. Tsai et al. (2003) showed that some rattlesnake PLA_2 proteins induced edema in mice but exhibited no antiplatelet activity toward mouse platelets, so this represents a possible alternative function of neutral or negatively charged PLA_2's in *Sistrurus* toward mammalian prey. As suggested above, functional assays of the effect of isolated PLA_2's on different prey would help clarify the functional significance of the variation in this region and also whether the different acidic and basic PLA_2 proteins exhibit some form of prey-specific effects.

Acknowledgments We thank Steve Mackessy for generously providing the *S. c. edwardsii* DNA sample, Jose Diaz, Giancarlo Lopez-Martinez, and David Denlinger for assistance with the lab work, Doug Wynn and Dan Harvey for help in the field, Joe Bielawski, Juan Calvete, James Cotton, Matt Dean, Brian Golding, Jeff Good, Dusan Kordis, Michael Nachman, Tom Waite, and Tom Wilson for advice and discussion, Laura Kubatko for help with the phylogenetic analyses, and the Gibbs Lab Group, Greg Booton, Brian Fry, and three anonymous reviewers for comments. Funding for this study was provided by the Columbus Zoo and Ohio State University.

Appendix

a) Exon 2

```
                10        20        30        40        50        60        70        80        90
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|.
cat1      TCGAGGGGAACCTGCTGCAATTCAACAAAATGATCAAGATAATGACGAAGAAAAACGCTATTCCCTCCTATACCTCTTACGGCTGCTACTGTGGCT
cat2      ........C.................G.........T.TGA...A..C.................T....G...T.....................
cat3      ........C.................G.........T.TGA...A..C.................T....G...T.....................
cat4      .........G....G..G....TG.G.CGT......T..A...TG..GG..G..GT.G.G..TGG.A...C.G.........A............A
cat5      ..A......G....G..G....TG.G.CGT......T..A...TG..GG..G..GT.GWG..TGG.A...C.G.........A............A
cat6      ........C.....A.A.....TG.A.CGT......T..A...TG..GG..G.....G.G.GTTT.GG..C.G.G.......A........C....
cat7      ........C.....A.A.....TG.G.CGT......T..A...TG..GG..G..G..G.G.GTTT.GG..C.G.G.......A........C....
ter1      ..................................................................T.....G.......................
ter2      ..................................................A...............T.....G.......................
ter3      ..............A.A.....TG.G.CGT.....TT..A.G..G.........G..G...GTTT..G..C.G.G................C....
ter4      ..............A.A.....TG.G.CGT.....TT..A.G..G.........G..G...GTTT..G..C.G.G................C....
edw1      ..............A.A.....TG.G.CGT.....TT..A.G..G.........G..G...GTTT..G..C.G.G................C....
edw2      ..............A.A.....TG.G.CGT...C.TT..A.G..G.........G..G...GTTT..G..C.G.G................C....
edw3      ..............A.A.....TG.G.CGT.....TT..A.G..............GC..G.........C....T......A........C...C
edw4      ..A.A....G....G.CG.....GGG.G..........GA.GA....GG.......C.CT..........C....T......A........C...C
mil1      ...............................................................................................
mil2      ..................................................................T............................
mil3      ........C.................C.........T.TGA...A..C.................T....G...T.....................
mil4      ........C.....A.A.....TG.G.CGT......T..A...TG..GG..G..GT.G.G.GTTT.GG..C.G.G.......A........C....
```

Fig. A1 Nucleotide sequences from PLA_2 exons 2 (**a**), 3 (**b**), and 4 (**c**), which code for amino acid residues that make up the mature PLA_2 venom protein. Complete genomic sequences for each sequence type have been deposited in GenBank under accession numbers EU369741–EU369759

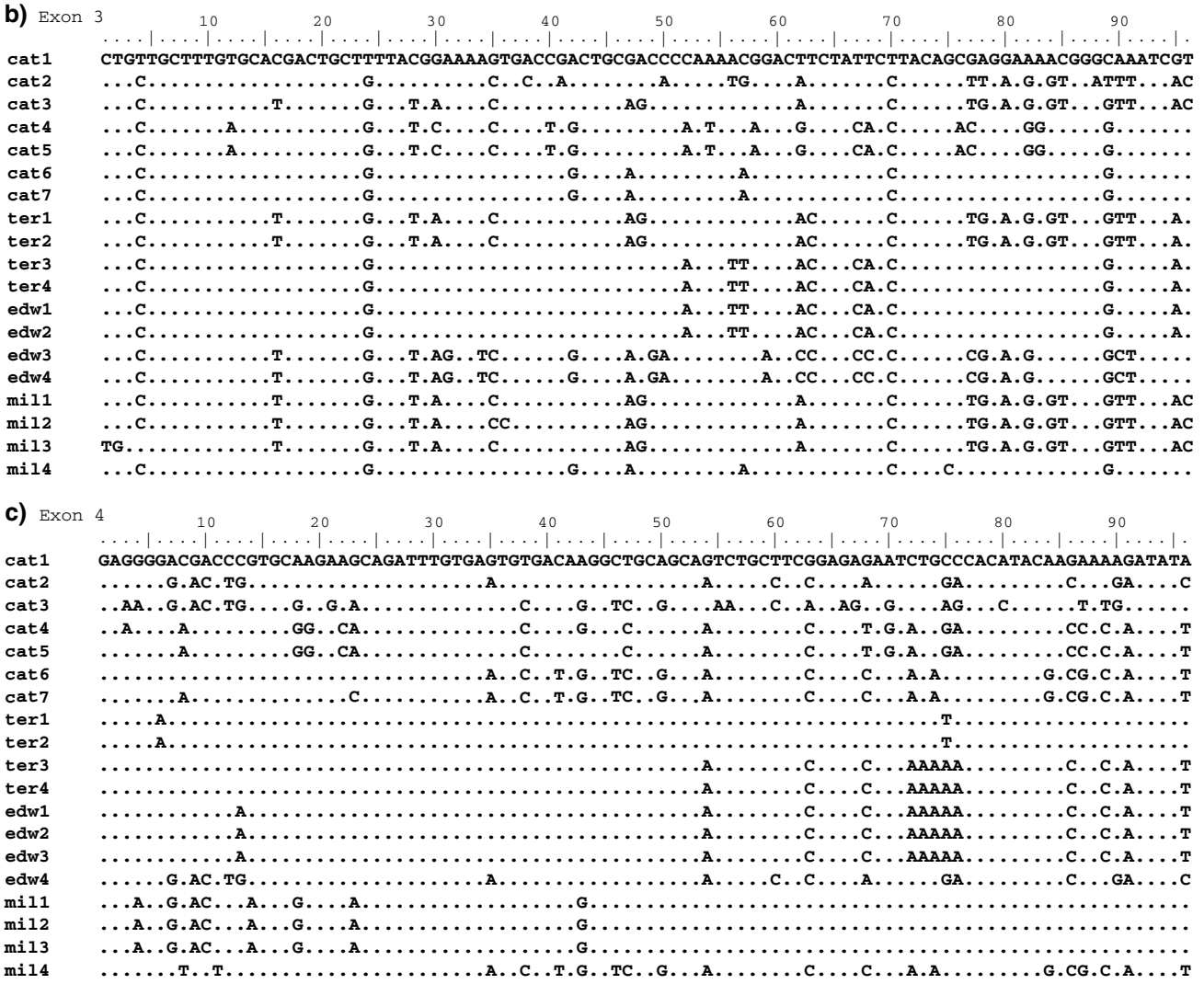

Fig. A1 continued

References

Brunie S, Bolin J, Gewirth D et al (1985) The refined crystal structure of dimeric phospholipase A_2 at 25 Å. J Biol Chem 260:9742–9749

Carredano E, Westerlund B, Persson B et al (1998) The three-dimensional structures of two toxins from snake venom throw light on the anticoagulant and neurotoxic sites of phospholipase A2. Toxicon 36:75–92

Cereb N, Hughes AL, Yang SY (1997) Locus-specific conservation of the HLA class I introns by intra-locus homogenization. Immunogenetics 47:30–36

Chen Y, Wang Y, Hseu M et al (2004) Molecular evolution and structure-function relationships of crototoxin-like and asparagines-6-containing phosphlolipase A_2 in pit viper venoms. Biochem J 381:25–34

Chijiwa T, Deshimaru M, Nobuhisha I et al (2000) Regional evolution of venom-gland phospholipase A_2 isoenzymes of *Trimeresurus flavoviridis* snakes in the southwestern islands of Japan. Biochem J 347:491–499

Chijiwa T, Yamaguchi Y, Ogawa T et al (2003) Interisland evolution of *Trimeresurus flavoviridis* venom phospholipase A_2 isozymes. J Mol Evol 56:286–293

Conant R, Collins JT (1998) A field guide to reptiles & amphibians of eastern & central North America. 3rd ed. Houghton Mifflin, New York

Cotton JA, Page RDM (2005) Rates and patterns of gene duplication and loss in the human genome. Proc Roy Soc Lond 272:277–283

Creer S, Malhotra A, Thorpe RS, et al (2003) Genetic and ecological correlates of intraspecific variation in pitviper venom composition detected using matrix-assisted laser desorption time-of-flight mass spectrometry (MALDI-TOF-MS) and isoelectric focusing. J Mol Evol 56:317–329

Daltry JC, Wuster W, Thorpe RS (1996) Diet and snake venom evolution. Nature 379:537–540

Danse JM, Gasparini S, Menez A (1997) Molecular biology of snake venom phospholipase A2. In: Kini RM (ed) Venom phospholipse A2 enzymes: structure, function, and mechanism. J. Wiley and Sons, Chichester, UK, pp 29–71

Demuth JP, Bie TD, Stajich JE et al (2006) The evolution of mammalian gene families. PLoS ONE 1:e85 doi:101371/journal.pone.0000085

Douglas ME, Douglas MR, Schuett GW et al (2006) Evolution of rattlesnakes (Viperidae; Crotalus) in the warm deserts of western North America shaped by Neogene vicariance and Quaternary climate change. Mol Ecol 15:3353–3374

Feder ME, Mitchell-Olds T (2003) Evolutionary and ecological functional genomics. Nature Rev Genet 4:649–655

Fry BG (2005) From genome to 'venome': molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. Genome Res 15:403–420

Fry BG, Wüster W (2004) Assembling an arsenal: origin and evolution ofthe snake venom proteome inferred from phylogenetic analysis of toxin sequences. Mol Biol Evol 21:870–883

Fry BG, Wüster W, Kini RM et al (2003) Molecular evolution of elapid snake venom three finger toxins. J Mol Evol 57:110–129

Fry BG, Vidal N, Norman JA et al (2006) Early evolution of the venom system in lizards and snakes. Nature 439:584–588

Gasteiger E, Hoogland C, Gattiker A et al (2005) Protein identification and analysis tools on the ExPASy server. In: Walker JM (ed) The proteomics protocols handbook. Humana Press, New York, pp 571–607

Gillespie JH (1994). The causes of molecular evolution. Oxford University Press, Oxford

Golding GB, Dean AM (1998) The structural basis of molecular adaptation. Mol Biol Evol 15:355–369

Gubensek F, Kordis D (1997) Venom phospholipase A2 genes and their molecular evolution. In: Kini RM (ed) Venom phospholipse A2 enzymes: structure, function, and mechanism. J. Wiley and Sons, Chichester, UK, pp 73–95

Hasegawa M, Kishino H (1989) Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. Evolution 43:627–677

Heatwole H, Powell J (1998) Resistance of eels (Gymnothorax) to the venom of sea kraits (*Laticauda colubrina*): a test of coevolution. Toxicon 36:619–625

Hoekstra HE, Coyne J (2007) The locus of evolution: evo devo and the genetics of adaptation. Evolution 61:995–1016

Holycross AT, Mackessey SP (2002) Variation in the diet of *Sistrurus catenatus* (massasauga) with emphasis on *Sistrurus catenatus edwardsii* (desert massasauga). J Herpetol 36:454–464

Hughes AL, Yeager M (1997) Comparative evolutionary rates of introns and exons in murine rodents. J Mol Evol 45:125–130

John TR, Smith L, Kaiser I (1994) Genomic sequences encoding the acidic and basic subunits of Mojave toxin:unusually high sequence identity of non-coding regions. Gene 139:229–234

Jorge da Silva NJ, Aird SD (2001) Prey specificity, comparative lethality and compositional differences of coral snake venoms. Comp Biochem Physiol C Toxicol Pharmacol 128:425–456

Junqueira-de-Azevedo ILM, Ho PL (2002) A survey of gene expression and diversity in the venom glands of the pitviper snake *Bothrops insularis* through the generation of expressed sequence tags (ESTs). Gene 299:279–291

Junqueira-de-Azevedo ILM, Ching ATC, Carvalho E et al (2006) *Lachesis muta* (Viperidae) cDNAs reveal diverging pit viper molecules and scaffolds typical of Cobra (Elapidae) venoms: implications for snake toxin repertoire evolution. Genetics 173:877–889

Kini RM, ed. (1997) Venom phospholipase A2 enzymes: structure, function, and mechanism. J. Wiley and Sons, Chichester, UK

Kini RM (2003) Excitement ahead: structure, function, and mechanism of snake venom phospholipase A_2 enzymes. Toxicon 42:827–840

Kini RM (2005) Structure-function relationships and mechanism of anticoagulant phospholipase A_2 enzymes from snake venoms. Toxicon 45:1147–1161

Kini RM, Chan YM (1999) Accelerated evolution and molecular surface of venom phospholipase A2 enzymes. J Mol Evol 48:125–132

Kini RM, Evans HJ (1989) A model to explain the pharmacological effects of snake venom phospholipases A2. Toxicon 27:613–635

Kordis D, Bdolah A, Gubensek F (1998) Positive Darwinian selection in *Vipera palestinae* phospholipase genes is unexpectantly limited to the third exon. Biochem Biophys Res Commun 251:613–619

Kordis D, Krizaj I, Gubensek F (2002) Funtional diversification of animal toxins by adaptive evolution. In: Menez A (ed) Perspectives in molecular toxinology. J. Wiley and Sons, New York, pp 401–419

Kumar S, Tamura K, Nei M (2004) MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5:150–163

Li M, Fry BG, Kini RM (2005) Eggs only diet: the shift in preferred prey by the marbled sea snake (*Aipysurus eydouxii*) resulting in a loss of postsynaptic neurotoxicity. J Mol Evol 60:81–9

Li M, Fry BG, Kini RM (2005) Putting the brakes on snake venom evolution: the unique molecular evolutionary patterns of Aipysurus eydouxii (marbled sea snake) phospholipase A2 toxins. Mol Biol Evol 22:934–941

Lomonte B, Moreno E, Tarkowski A et al (1994) Neutralizing interaction between heparins and myotoxin II, a lysine 49 phospholipase A2 from *Bothrops asper* snake venom. Identification of a heparin-binding and cytolytic toxin region by the use of synthetic peptides and molecular modeling. J Biol Chem 269:29867–29873

Lynch VJ (2007) Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. BMC Evol Biol 7:2

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 293:1151–1155

Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genom 3:35–44

Menez A (2002) Perspectives in molecular toxinology. J. Wiley & Sons, New York

Metz EC, Robles-Sikisaka R, Vacquier VD (1998) Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. Proc Natl Acad Sci USA 95:10676–10681

Nakashima K, Nobuhisa I, Deshimaru M et al (1993) Accelerated evolution of *Trimeresurus flavoviridis* venom gland phospholipase A_2 isozymes. Proc Natl Acad Sci USA 90:5964–5968

Nakashima K, Nobuhisa I, Deshimaru M et al (1995) Accelerated evolution in the protein-coding regions is universal in crotalinae snake venom gland phospholipase A_2 isozyme genes. Proc Natl Acad Sci USA 92:5605–5609

Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39:121–152

Nobuhisa I, Nakashima K, Deshimaru M et al (1996) Accelerated evolution of *Trimeresurus okinavensis* venom gland phospholipase A_2 isozyme-encoding genes. Gene 172:267–272

Ohno M, Chijiwa T, Oda-Ueda N et al (2003) Molecular evolution of myotixic phospholipases A_2 from snake venom. Toxicon 42:841–854

Page RDM (1998). GeneTree:comparing gene and species phylogenies using reconciled trees. Bioinformatics 14:819–820

Peitsch MC (1996) ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. Biochem Soc Trans 24:274–279

Petan T, Krizaj I, Gubensek F et al (2002) Phenylalanine-24 in the N-terminal region of ammodytoxins is important for both enzymic activity and presynaptic toxicity. Biochem. J 363:353–358

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. Bioinformatics 14:817–818

Prijatelj P, Copic A, Krizaj I et al (2000) Charge reversal of ammodytoxin A, a phospholipase A2-toxin, does not abolish its neurotoxicity. Biochem J 352:251–255

Prijatelj P, Krizaj I, Kralj B et al (2002) The C-terminal region of ammodytoxins is important but not sufficient for neurotoxicity. Eur J Biochem 269:5759–5764

Rozas J, Sánchez-DelBarrio J C, Messeguer X et al (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Samanta U, Bahadur RP, Chakrabarti P (2002) Quantifying the accessible surface area of protein residues in their local environment. Protein Eng 15:659–667

Sanz L, Gibbs HL, Mackessy SP et al (2006) Venom proteomes of closely related *Sistrurus* rattlesnakes with divergent diets. J Proteome Res 5:2098–2112

Senepathy P, Shapiro MB, Harris NL (1990) Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol 183:252–278

Singh G, Gourinath S, Sharma S et al (2001) Sequence and crystal structure determination of a basic phospholipase A2 from common krait (Bungarus caeruleus) at 24 Å resolution: identification and characterization of its pharmacological sites. J Mol Biol 307:1049–1059

Slowinski JB, Knight A, Rooney AP (1997) Inferring species trees from gene trees: a phylogenetic analysis of the elapidae (Serpentes) based on the amino acid sequences of venom proteins. Mol Phylogenet Evol 8:349–362

Stern DL (2000) Evolutionary developmental biology and the problem of variation. Evolution 54:1079–1091

Suzuki Y, Nei M (2004) False-positive selection identifed by ML-based methods: examples from the Sig1 gene of the diatom *Thalassiosira weissflogii* and the tax gene of a human T-cell lymphotropic virus. Mol Biol Evol 21:914–921

Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, MA

Tatsuya A, Shiina T, Kimura N et al (2003) Comparative sequencing of human and chimpanzee MHC class I regions unveils insertions/deletions as the major path to genomic divergence. Proc Natl Acad Sci USA 100:7708–7713

Tsai I-H, Wang Y-M, Chen Y-H et al (2003) Geographic variations, cloning, and functional analyses of the venom acidic phospholipases A2 of *Crotalus viridis viridis*. Arch Biochem Biophys 411:289–296

Tsai I, Wang Y, Chen Y et al (2004) Venom phospholipase A_2 of bamboo viper (*Trimeresurus stejnegeri*): molecular characterization, geographic variations and evidence for multiple ancestries. Biochem J 377:215–223

Wong WSW, Yang Z, Goldman N et al (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl BioSci 13:555–556

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

Yang Z, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118

Zhao K, Zhou Y, Lin Z (2000) Structure of basic phospholipase A2 from *Agkistrodon halys* Pallas: implications for its association, hemolytic and anticoagulant activities. Toxicon 38:901–916