Convergent molecular evolution of toxins in the venom of advanced snakes (Colubroidea)

Xie, B.

Chapter 4. Evolution of C-type Lectin Toxins

This chapter is published as part of:

Bing Xie, Daniel Dashevsky, Darin Rokyta, Parviz Ghezellou, Behzad Fathinia, Qiong Shi, Michael K. Richardson and Bryan G. Fry. Dynamic genetic differentiation drives the widespread structural and functional convergent evolution of snake venom proteinaceous toxins. *BMC Biology*, 2022, 20:4. https://doi.org/10.1186/s12915-021-01208-9

Abstract

C-type lectins are one of the largest protein families in mammals and reptiles and they have shown various functions including defence mechanisms against predators. Since the Viperidae split off from the remaining caenophidian snakes, a novel heterodimeric lectin type evolved through loss of the carbohydrate-binding activity. The detailed evolutionary history of these toxins is unknown. We therefore conducted large-scale transcriptome sequencing, phylogenetic analysis and selection analysis to address this issue. Our results showed that the heterodimeric lectins form three clusters: two in the Viperidae and one in the Colubridae. All three clusters are under strong selection. We discuss the evolutionary implications of our findings. We also argue that these toxins may have considerable potential in drug design and development, in biochemical assays and in drug discovery.

Keywords: C-type lectins, basal/ancestral form, dimeric form, evolution

Introduction

The first major evolutionary modification in snake-venom C-type lectin was the mutation of the EPN motif in the loop to QPD, which led to a change in specificity from mannose to galactose and thus to mediation of the alternative erythrocyte-agglutination pathway (Drickamer 1992). Apotypic heterodimeric lectin forms from snake venom (snaclec) that lack carbohydrate-binding ability but instead possess numerous neofunctionalizationed activities. They are covalently linked by a single disulfide bond, and each subunit contains three internal disulfide bonds (Andrews, et al. 1989; Usami, et al. 1993). Heterodimeric lectins represent an extreme structural derivation and form a complex with P-IIId type SVMP (Chapter 5). Examples of this apotypic form are RVV-X and carinactivase-I, which activate factor X and prothrombin, respectively (Morita 2005). The linkage between the snaclec domain and the metalloprotease/cysteine-rich and disintegrin domains involves an additional disulfide bond in the chain homologous with the alpha chain of other snaclecs.

A previous study showed two forms that lie basal to the alpha and beta viperid venom sequences in the phylogenetic tree (Fry, Jackson, et al. 2015). While viper venoms are extremely rich in heterodimeric forms, it remains unclear if those forms are present in other snake lineages. Heterodimeric lectins have been sequenced from the transcriptome of the RFS snake *Philodryas olfersii* (Ching, et al. 2006). These may represent the plesiotypic state of the heterodimers. As these sequences are known only from transcriptome sequencing, both their bioactivity and their heterodimer structure remain to be elucidated. Conversely, while a heterodimer has in fact been isolated from the venom of *Ophiophagus hannah* (ophioluxin)(Du, et al. 2002), it is known only from very small sequence fragments, and thus the phylogenetic affinity of each chain remains enigmatic.

Although structurally identical, snake venom C-type lectins proteins have a range of physiological roles that are reliant on binding to platelet cell surface receptors or coagulation factors, as mentioned above. It would be interesting to learn more about the evolutionary mechanisms that led to this functional difference. cDNA sequences of different C-type lectins and C-type lectin-like proteins from snake venoms were used to create a phylogenetic tree (Fry, et al. 2008) using the Bayesian approach. The tree revealed that these proteins split into three groups: C-type lectins, C-type lectin-like proteins, and C-type lectin-like proteins' A (or a) and B (or b) chains. The A and B chains clearly evolved from an ancient C-type lectin before snake species diversification. The evolutionary tree built using amino acid sequences shows a similar branching structure (Tani, et al. 2002).

In this chapter, we used phylogenetic reconstruction and sequence analysis to reconstruct the evolutionary history of snake venom C-type lectin toxins. Particular emphasis was given to the heterodimeric types. Analysis of selection was also utilized to evaluate the selection pressure on different phylogenetic clusters of the toxins. We argue that C-type lectins may be potential new leads in drug development.

Materials and Methods

Sequence Alignments and Phylogenetic Reconstruction

Protein sequences for all toxin sequences were retrieved from the UniProt database (https://www.uniprot.org) and NCBI database (http://www.ncbi.nlm.nih.gov), then combined with the toxin transcripts from our assembly and annotation (Chapter 2). Partial sequences, sequences with suspect assembling errors were excluded. For the blocks of sequence in between these sites, the sequences were aligned using a mix of manual alignment of the conserved cysteine locations and alignment using the Multiple Sequence Comparison by Log-Expectation (MUSCLE) method (Edgar 2004) implemented in AliView (Larsson 2014). Manual refinement of the alignment was also involved because there are structural differences within different toxin families. The phylogenetic trees for different toxin families were reconstructed with MrBayes 3.2 (Ronquist, et al. 2012) based on the amino acid sequence alignment. The settings for MrBayes can be found in the Supplementary File 2. The output trees from MrBayes were midpoint rooting, then further edited and annotated with iTol (Letunic and Bork 2007).

Tests for Selection

Coding DNA sequences, which are corresponding to the toxin sequences used for phylogenetic analysis, were retrieved from GenBank (Benson, et al. 2012) and our assembly. Using AliView and the MUSCLE method, the sequences were trimmed to only containing codons that translate to the mature protein, then translated, aligned, and reverse translated. Clades were created based on taxonomy and structural differences (functional domains/motifs, for example). The resultant codon alignments were used to create phylogenetic trees for each clade using the same methods outlined in the 'Phylogenetic Reconstruction' section. All following studies were conducted using these tree topologies.

Calculating the ratio of nonsynonymous nucleotide substitutions per nonsynonymous site (dN) to synonymous substitutions per synonymous sites (dS) (ω=dN/dS) for each codon in the alignment might reveal if a gene is undergoing rapid evolution or stays functionally restricted. Codons developing with $\omega > 1$ are thought to have evolved under positive selection (functional diversity), whereas codons evolving with $\omega < 1$ are thought to have evolved under purifying selection. Sites with a value of $\omega = 1$ are believed to evolve in a neutral manner. In order to find the most likely groups on which positive selection has been working, we conducted a series of experiments integrating data from site-based and lineage-specific studies.

Due to their various emphases, we employed many of the selection tests developed in HyPhy v 2.220150316 beta (Pond and Muse 2005) to study the patterns of selection acting on distinct toxin families. The Analyze Codon Data analysis in HyPhy produces overall alignment values, whereas the FUBAR technique assesses the intensity of persistent positive or negative selection on individual amino acids (Murrell, et al. 2013). The Mixed Effects Model of Evolution (MEME) approach, on the other

hand, finds particular locations that have been exposed to positive selection in the past (Murrell, et al. 2012).

Protein Modelling

To map residues evolving under positive selection in three-dimensional (3D) structures, sample sequences from the RCSB PDB database (Rose, et al. 2010) were used to create bespoke models for each clade belonging to various toxin families (Table 1). Alignments of each clade were trimmed to match these PDB structures. To render and colour the 3D structure of the proteins, we utilized the UCSF Chimera program v 1.10.2 with attribute files generated from FUBAR and MEME results. For FUBAR, we used the value from the beta-alpha column which is a measure of the difference between the rates of non-synonymous (beta) and synonymous (alpha) mutations. For MEME, since MEME estimates two rates of positive selection and gives each a probability, we take the weighted average of those two and then subtract alpha to arrive at a similar value to the one we used for FUBAR.

Table 1: Custom models for protein modelling.

Toxin Groups	Sequence for 3D modeling	PDB ID
ancestral	Erythrolamprus_poecilogyrus_A7X3Z7	5f2q
viper dimeric alpha	Bothrops_jararaca_Q56EB1	5f2q
non-viper dimeric	Philodryas_olfersii_Q09GK0	1v7p
viper dimeric beta	Echis_multisquamatus_Q7T2Q0	1fvu

Results and Discussion

The basal form of lectin toxins in reptile venoms is a single-chain form which may form non-covalently linked complexes and contains diagnostic tripeptide functional motif (Figures 1 and 2) (Walker, et al. 2004; Arlinghaus, et al. 2015). Consistent with previous analyses (Fry, et al. 2008), our phylogenetic results suggest that the earliest functional motif is the amino acids EPN (glutamic acid + proline + asparagine). Our results also indicate that the QPD (glutamine + proline + aspartic acid) motif has arisen on two convergent occasions, once in the last common ancestor of the advanced snakes, and again in the last common ancestor of the Australian radiation of elapids.

In addition, other mutations in the functional motif were documented across a myriad of lineages: EPG (glutamic acid + proline + glycine) in *Parasuta nigriceps* within the Elapidae; EPK (glutamic acid + proline + lysine) in *Heterodon nasicus* within the Colubridae; KPK (lysine + proline + lysine) in *Tropidolaemus subannulatus* within the Viperidae; KPN (lysine + proline + asparagine) in *Homalopsis buccata* within the Homalopsidae; KPS (lysine + proline + serine) in *Micrurus corallinus* within the Elapidae; KRN (lysine + arginine + asparagine) in *Leioheterodon madagascarensis* within the Lamprophiidae; LTD (leucine + threonine + aspartic acid) in *Bitis gabonica* within the Viperidae; and QPN (glutamine + proline + asparagine) in *Vipera transcaucasiana* within the Viperidae (Figures 1 and 2).

To-date only viperid venom variants of the QPD form have had their bioactivity tested. They were shown agglutinate erythrocytes and promote edema by increasing vascular permeability (Guimarães-Gomes, et al. 2004; Panunto, et al. 2006; Lin, et al. 2007). The impacts of the extreme diversifications of the key functional motif shown in this study upon the functions of the toxins are entirely unknown, and require further research. The overall ω value for these toxins was only 0.72, but there were 14 sites identified as positively selected by FUBAR & MEME, which is an indication that the variation which occurs in these toxins is tightly constrained and only occurs at a relatively small subset of positions (Table 2).

In addition to the ancestral single-chain form, a disulphide-linked dimer composed of two different lectins has long been known from viperid venoms (sometimes referred to as snaclecs, or C-type lectins), with variants producing a wide diversity of coagulotoxic effects including inhibition of the clotting factor vWF and clotting enzymes such as Factors IXa and XIa (Arlinghaus, et al. 2015). In addition to the diagnostic newly evolved cysteines that facilitate the inter-chain disulphide bond leading to the dimeric tertiary structure, this type is also molecularly distinct because they have lost the functional motif (Figure 1). The α and β chains have the interchain cysteine in the same position which suggests that either these toxins evolved from a single gene that produced a homomeric ancestral toxin and subsequently underwent duplication and sub-functionalization or that structural constraints led to the novel cysteine mutation occurring at the same location in two different genes. The α chain is readily distinguished from the β chain by a characteristic glutamine motif present immediately before this diagnostic cysteine (Figure 2).

Not only did we recover multiple variants from non-viperid snakes that possessed the diagnostic cysteine of the dimeric lectin form, but also forms with and without the glutamine motif form were present (Figure 2). This suggests that the evolution of the dimeric lectin toxins preceded the divergence of Viperidae from other advanced snakes. However, the rates of evolution are reflective of the explosive diversification of this toxin type within the viperids (Figure 3 and Table 2). A previous study found differential rates of evolution between the α-subunit and β-subunits based on one species (*Crotalus helleri*) (Sunagar, et al. 2014). Consistent with that study, we found that the α-subunit and β-subunit were subject to different selection pressures when analysed across all viperid species in this study. The α-subunit had an overall ω value of 1.40 with 49 sites shown to be positively selected by FUBAR & MEME, while the β-subunit had an overall ω value of 1.27 with 38 sites shown to be positively selected by FUBAR & MEME (Table 2). In contrast, the non-viperid dimeric forms (excluding the unique diversification within *Helicops*) had an overall neutral ω of 0.97 but with 10 sites shown as positively selected by FUBAR & MEME (Table 2).

The comparative analysis of 3D modelling (Figure 3) on different clades showed that those residues under positive selection are located on different position of the surface for the viperid α-subunit versus β-subunits, and the other forms as well, indicative different selective forces and the potential for the discovery of novel activities within the non-viperid dimeric and monomeric forms. Structure-function studies on these toxins may therefore be particularly interesting. The unique form present in the venom of *Helicops leopardinus,* which has novel insertions in the key functional region, including the evolution of novel cysteines may also be of particular interest for these future research efforts (Figure 1).

Table 2: Molecular evolutionary rates of lectins (See Figure 3 for modelling).

Toxin group	ω	FUBAR(-)[a]	FUBAR(+)[b]	MEME[c]	FUBAR & MEME[d]
Ancestral (monomeric)	0.72	44	14	31	14
Derived (dimeric)--Non-vipers excluding *Helicops*	0.97	6	16	21	10
Derived (dimeric)--*Helicops*	1.16	0	6	7	3
Derived (dimeric)--Viperidae-alpha	1.40	23	54	60	49
Derived (dimeric)--Viperidae-beta	1.27	17	45	54	38

[a] Number of codons under negative selection according to FUBAR

[b] Number of codons under positive selection according to FUBAR

[c] Number of codons under episodic diversifying selection according to MEME

[d] Number of codons that fit criteria [b] and [c]

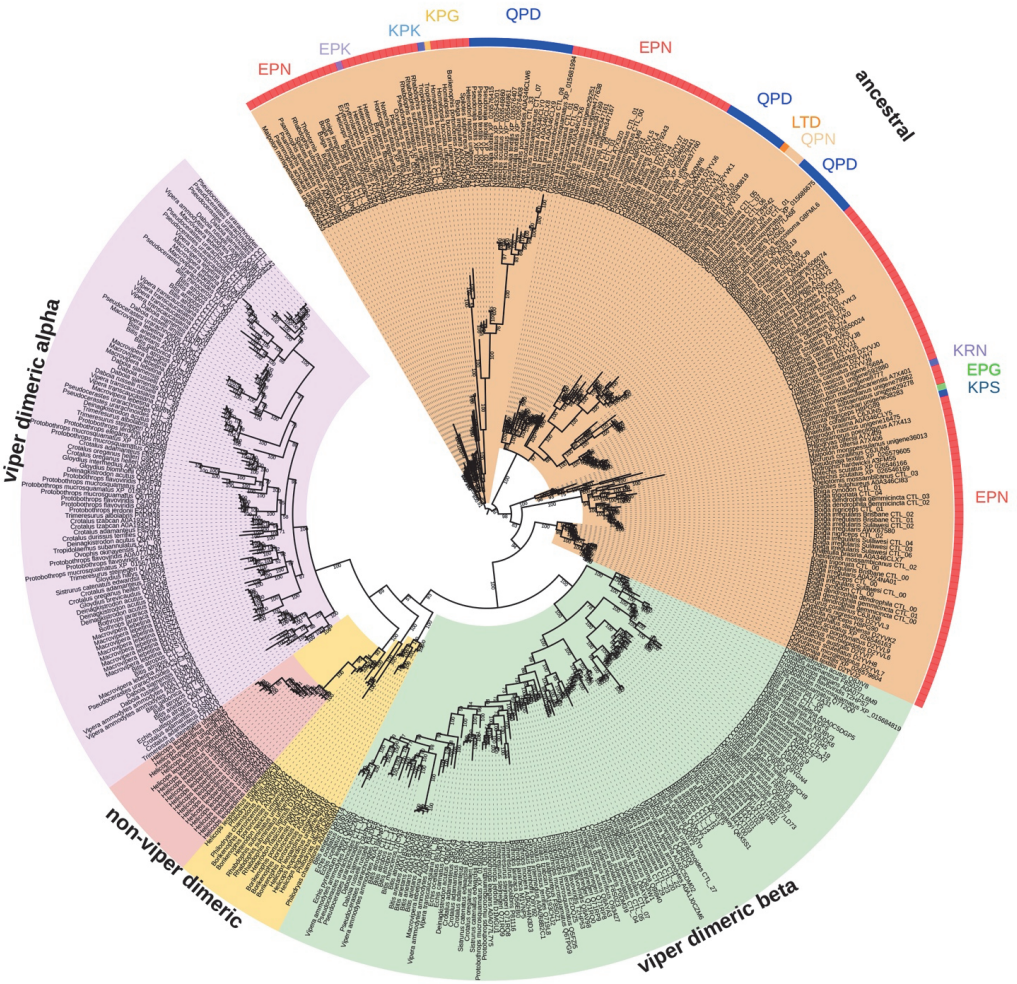

Figure 1: Molecular phylogenetic reconstruction of C-type lectin toxins. The ancestral monomeric form is shaded annotated with variations in the key functional amino acid triad. Sequence alignment for constructing phylogenetic tree can be viewed in Supplementary File 5. For tree output file for lectin toxins, see Supplementary File 6.

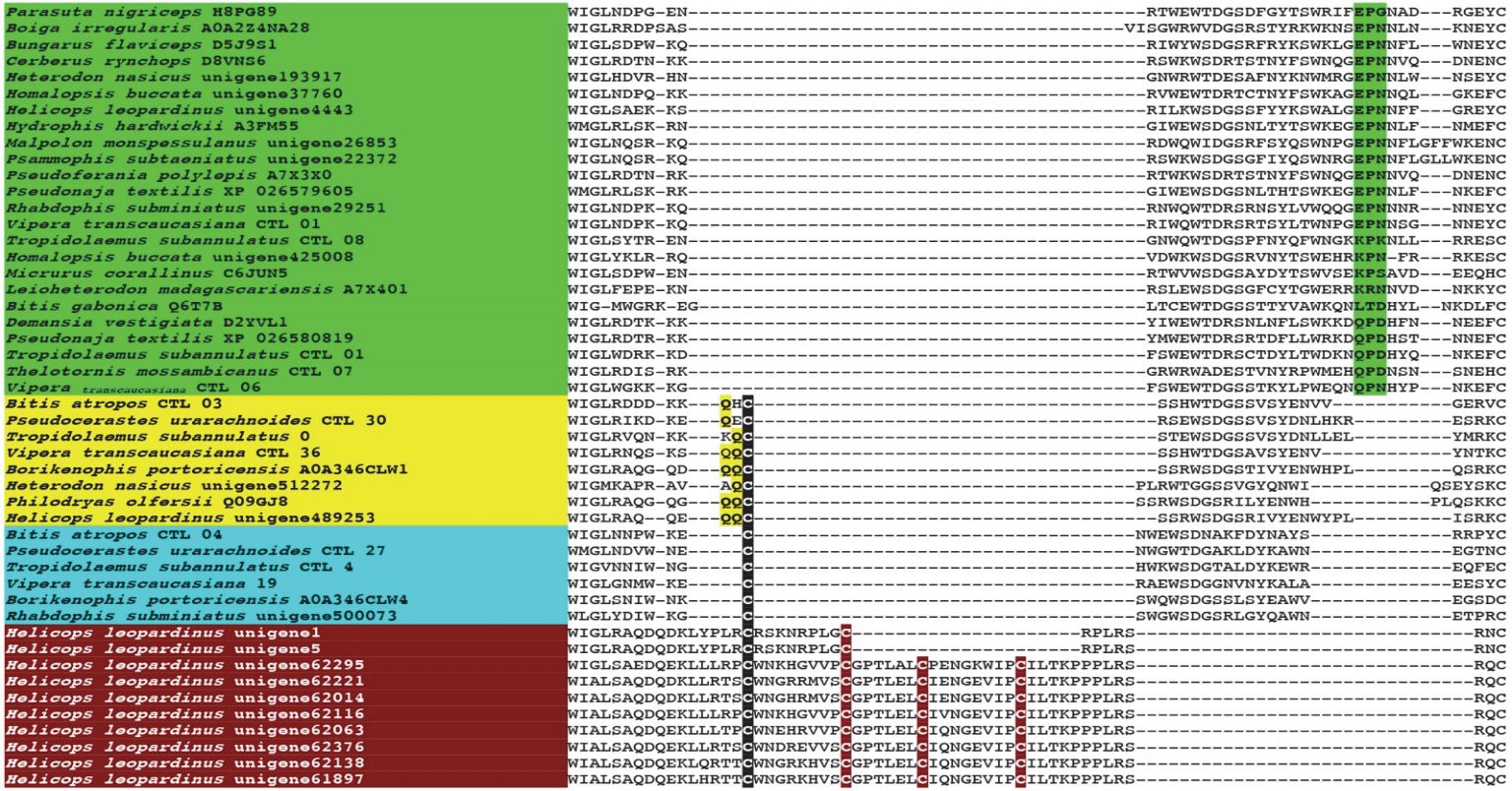

Figure 2: Partial amino acid sequence alignment of representative C-type lectin toxins in the region that contains the key functional amino acid triad (shaded in green). The alpha chain of the derived cysteine-linked heterodimeric form is shaded in yellow, including the glutamine motif that diagnoses the alpha chain, while the beta chain is highlighted in blue. The newly evolved dimer-forming cysteine is shaded in black. The sequence alignment also shows the insertion characteristic of an extremely derived form known currently only from *Helicops leopardinus,* which contains unique cysteines shaded in burgundy.

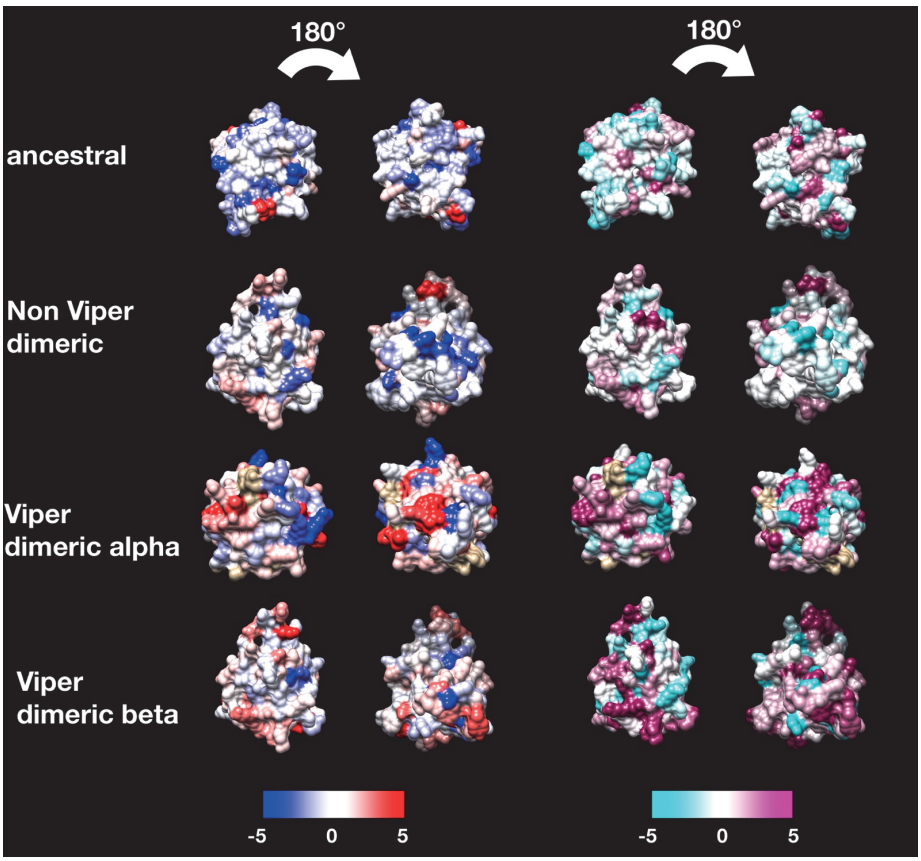

Figure 3: Molecular modelling of C-type lectin toxins showing sites under selection by FUBAR (left) and MEME (right) colour coded to show sites that are negatively, neutrally, or positively selected. See Table 2 for values. Protein models show front and back views colored according to FUBAR's estimated strength of selection (β-α, left) and MEME's significance levels (right). Table 1 contains the information regarding template choice for each toxin subclass.

References

Andrews RK, Booth WJ, Gorman JJ, Castaldi PA, Berndt MC. 1989. Purification of botrocetin from Bothrops jararaca venom. Analysis of the botrocetin-mediated interaction between von Willebrand factor and the human platelet membrane glycoprotein Ib-IX complex. Biochemistry 28:8317-8326.

Arlinghaus FT, Fry BG, Sunagar KK, Jackson TNW, Eble JA, Reeks T, Clemetson KJ. 2015. Lectin proteins. In. Venomous reptiles and their toxins: evolution, pathophysiology and biodiscovery. New York: Oxford University Press. p. 299-311.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2012. GenBank. Nucleic Acids Research 41:D36-D42.

Ching AT, Rocha MM, Leme AFP, Pimenta DC, Maria de Fátima DF, Serrano SM, Ho PL, Junqueira-de-Azevedo IL. 2006. Some aspects of the venom proteome of the Colubridae snake *Philodryas olfersii* revealed from a Duvernoy's (venom) gland transcriptome. FEBS letters 580:4417-4422.

Drickamer K. 1992. Engineering galactose-binding activity into a C-type mannose-binding protein. Nature 360:183-186.

Du X-Y, Clemetson JM, Navdaev A, Magnenat EM, Wells TN, Clemetson KJ. 2002. Ophioluxin, a convulxin-like C-type lectin from *Ophiophagus hannah* (King cobra) is a powerful platelet activator via glycoprotein VI. Journal of Biological Chemistry 277:35124-35132.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797.

Fry BG, Jackson TNW, Takacs Z, Reeks T, Sunagar K. 2015. C-type natriuretic peptides. In: Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and Biodiscovery. New York: Oxford University Press. p. 318-326.

Fry BG, Scheib H, van der Weerd L, Young B, McNaughtan J, Ramjan SR, Vidal N, Poelmann RE, Norman JA. 2008. Evolution of an arsenal: structural and functional diversification of the venom system in the advanced snakes (Caenophidia). Molecular & Cellular Proteomics 7:215-246.

Guimarães-Gomes V, Oliveira-Carvalho AL, de LM Junqueira-de-Azevedo I, Dutra DL, Pujol-Luz M, Castro HC, Ho PL, Zingali RB. 2004. Cloning, characterization, and structural analysis of a C-type lectin from *Bothrops insularis* (BiL) venom. Archives of Biochemistry and Biophysics 432:1-11.

Larsson A. 2014. AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics 30:3276-3278.

Letunic I, Bork P. 2007. Interactive Tree of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23:127-128.

Lin L-P, Lin Q, Wang Y-Q. 2007. Cloning, expression and characterization of two C-type lectins from the venom gland of *Bungarus multicinctus*. Toxicon 50:411-419.

Morita T. 2005. Structures and functions of snake venom CLPs (C-type lectin-like proteins) with anticoagulant-, procoagulant-, and platelet-modulating activities. Toxicon 45:1099-1114.

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K. 2013. FUBAR: a fast, unconstrained bayesian approximation for inferring selection. Molecular Biology and Evolution 30:1196-1205.

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Pond SLK. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genetic 8:e1002764.

Panunto PC, Da Silva MA, Linardi A, Buzin MP, Melo SE, Mello SM, Prado-Franceschi J, Hyslop S. 2006. Biological activities of a lectin from *Bothrops jararacussu* snake venom. Toxicon 47:21-31.

Pond SLK, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. In. Statistical Methods in Molecular Evolution: Springer. p. 125-181.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539-542.

Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlić A, Quesada M, Quinn GB, Westbrook JD. 2010. The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Research 39:D392-D401.

Sunagar K, Undheim EA, Scheib H, Gren EC, Cochran C, Person CE, Koludarov I, Kelln W, Hayes WK, King GF. 2014. Intraspecific venom variation in the medically significant Southern Pacific Rattlesnake (*Crotalus oreganus helleri*): biodiscovery, clinical and evolutionary implications. Journal of Proteomics 99:68-83.

Tani A, Ogawa T, Nose T, Nikandrov NN, Deshimaru M, Chijiwa T, Chang C-C, Fukumaki Y, Ohno M. 2002. Characterization, primary structure and molecular evolution of anticoagulant protein from *Agkistrodon actus* venom. Toxicon 40:803-813.

Usami Y, Fujimura Y, Suzuki M, Ozeki Y, Nishio K, Fukui H, Titani K. 1993. Primary structure of two-chain botrocetin, a von Willebrand factor modulator purified from the venom of *Bothrops jararaca*. Proceedings of the National Academy of Sciences 90:928-932.

Walker JR, Nagar B, Young NM, Hirama T, Rini JM. 2004. X-ray crystal structure of a galactose-specific C-type lectin possessing a novel decameric quaternary structure. Biochemistry 43:3783-3792.