Using phylogenetics to assess similarity in neurotoxic venom protein sequences from several snake species in Bangladesh

Varun Subramaniam | PUBH 6860 (Principles of Bioinformatics) | December 20th 2022

1 Introduction

In the words of former United Nations Secretary General Kofi Annan, snakebite is the "biggest public health crisis you've never heard of." Every year, tens of thousands of people die from snakebites and in regions like South Asia, rates of injuries and fatalities following a bite are disproportionately high. The Global Snakebite Initiative reports that the most common victim of a snakebite in Bangladesh is a "poor, young, and active individual" and that many patients cannot afford or access treatment following a bite (Ghose, 2021). Nevertheless, research into the biological elements of snake venom–a necessity for the development of effective treatments that can be produced at scale and made accessible to all–remains sparse and incomplete (Annan, 2018).

In the past decade, thanks to significant developments in sequencing technologies, scientists have begun to apply proteomics and phylogenetics to studies of snake venom. For example, Willard et al. applied proteomic analysis to mass spectrometry data to identify metabolic changes in gene expression among mice injected with hemotoxic venom from various rattlesnakes native to the Western United States (Willard *et al.*, 2021). Similarly, Tasoulis et al. used a variety of sequencing techniques and computational tools to identify a range of previously unknown protein families in three types of viper venom (Tasoulis *et al.*, 2021). The benefits of such deep-dives into venom cocktails are wide-reaching and immediate. The US Food and Drug Administration and the European Medicines Agency have collectively approved 11 venom-derived drugs with a variety of functions, from increasing blood clotting in patients with hemophilia to alleviating hypertension in people living with type 2 diabetes. This number is expected to rise significantly in the coming years, especially thanks to the creation of several open-access snake venom databases, which allow researchers to rapidly explore pressing gaps in our understanding of venom (von Reumont *et al.*, 2022).

One such gap, which leading herpetologists like Austin Stevens and Romulus Whitaker have highlighted as a critical area for research, is the lack of broad-spectrum antivenom (BSA) that can effectively neutralize the venoms of *multiple* species of snakes. BSA acts with the same mechanism as conventional monovalent antivenom (CMA); it is delivered via intravenous injection into snakebite victims and contains antibodies collected from host animals (usually horses or sheep) that have been exposed to low doses of venom proteins. Unlike CMA however, BSA contains a cocktail of antibodies designed to counter various proteins from different venom types. Key to the development of BSA is identifying proteins that are similar across diverse forms of venoms (Xiao *et al.*, 2017). BSA is particularly useful in low-income and rural contexts, where storage of several individual CMA vials is logistically difficult and costly. BSA is also applicable in cases where patients are unable to identify the exact snake by which they were envenomated and thus cannot be given any specific monovalent antivenom (Maduwage *et al.*, 2016).

Indian Polyvalent Antivenom (abbreviated as PAV) is an example of a recently developed BSA that has been found to successfully neutralize venoms of the Russell's viper, saw-scaled viper, Indian krait, and Indian cobra: the four species responsible for the most snakebite deaths in

India. PAV was developed by first comparing the protein profiles of each snake's venom to find commonalities; scientists then injected host animals with a mixture of highly similar proteins to collect antibodies that could potentially neutralize all four venoms (Madhushani *et al.*, 2021). Essentially, by applying proteomics to studies of snake venom, researchers were able to develop a contextually relevant BSA that has, in initial studies, proven effective.

Modeled after these preliminary steps of PAV production, our study aims to contribute to BSA development research in South Asia. The Indigenous Snakes of Bangladesh (ISOB) database is a publicly available tool containing mRNA sequences and GenBank accession numbers of functionally critical venom-related genes for multiple snake species in Bangladesh. ISOB contains over 100 neurotoxic protein sequences from several snake species native to Bangladesh, including those belonging to the *Bungarus* (krait) and *Naja* (cobra) genera. Sequences from these genera are particularly important to the development of BSA in this context; the Global Snakebite Initiative reports "neurotoxic envenoming by kraits and cobras is the principal cause of snake bite mortality in Bangladesh" (Ghose, 2021). In our study, we used various computational tools to assess regions of and overall similarity between several neurotoxic sequences from ISOB to provide initial insights into the development of a potential BSA curated for Bangladesh.

2 Materials and methods

2.1 Collecting Neurotoxic mRNA Sequence Data from ISOB

We used the `Advanced Search` tool in ISOB to filter for `Neurotoxins`. We then downloaded all of the sequences into a single FASTA file using the `Save as .txt` option (Roly *et al.*, 2015). The resulting FASTA file contained 111 mRNA sequences from 9 snake species: *Astrotia stokesii*, *Bungarus caruleus*, *Bungarus fasciatus*, *Hydrophis cyanocinctus*, *Laticauda colubrina*, *Laticauda laticaudata*, *Naja kaouthia*, *Naja naja*, and *Ophiophagus hannah.* This FASTA file (`VenomSeqs.fasta`) can be downloaded here.

2.2 Generating a Multiple Sequence Alignment in MEGA11

We generated a multiple sequence alignment (MSA) for the 111 sequences in `VenomSeqs.fasta` using the MUSCLE algorithm in MEGA11 (Tamura *et al.*, 2021). Since none of the default parameters for gap penalties, memory, iterations, or advanced options (including cluster methods) could be changed, we retained software-suggested penalties of –2.90 for "Gap Open" and 0.00 for "Gap Extend." Nevertheless, these penalties lend themselves to one main goal of our study–to identify regions of similarity within sequences–as they facilitate alignments with longer conserved regions and concentrated gapped regions, rather than those produced by inserting several shorter gaps throughout each sequence.

GAP PENALTIES		
Gap Open	☑	*-2.90*
Gap Extend	☑	*0.00*
Hydrophobicity Multiplier	☑	*1.20*
MEMORY/ITERATIONS		
Max Memory in MB	☑	*2048*
Max Iterations	☑	*16*
ADVANCED OPTIONS		
Cluster Method (Iterations 1,2)	☑	*UPGMA*
Cluster Method (Other Iterations)	☑	*UPGMA*
Min Diag Length (Lambda)	☑	*24*

Fig. 1. Parameters for MUSCLE MSA in MEGA11

2.3 Visualizing Conserved Regions from the MUSCLE MSA

We used the `Toggle Conserved Sites` tool under `Display` in MEGA11 to visualize regions of high amino-acid conservation across the MUSCLE-aligned neurotoxic mRNA sequences (Tamura *et al.*, 2021). We chose to use the `at 80% level` option to highlight sites at which 89 or more of the 111 sequences had identical amino acid compositions. A visual showing conserved regions of the sequences are shown below in Section 3.1.

2.4 Finding Best-Fit Evolutionary Model for MUSCLE MSA

We used ModelTest 0.1.6, the latest available version of the software with an integrated OS-compatible graphical user interface, to find the best-fit model for the MUSCLE MSA via a "Maximum Likelihood" tree. ModelTest was preferable to software such as MEGA11 for this step, as it tests alignments against several, more-complex models (Darriba *et al.*, 2020) and is computationally less intensive. The top ten models from ModelTest are shown below. Lower BIC scores (in the `score` column) indicate evolutionary models that better fit the input MSA (Stecher *et al.*, 2020).

For this MUSCLE MSA, the best-fit model of evolution is **WAG + G4**: a derivation of the Whelan and Goldman matrix, which uses the approximate maximum-likelihood method to estimate substitution rates for globular proteins, incorporating four gamma parameters (Whelan and Goldman, 2001). The exact 20 x 20 amino acid substitution matrix for WAG + G4 can be viewed here.

BIC	model	K	lnL	score	delta	weight
1	WAG+G4	1	-5029.5524	11102.9900	0.0000	0.9213
2	WAG+I+G4	2	-5029.6400	11107.9100	4.9200	0.0787
3	VT+G4	1	-5045.5295	11134.9441	31.9542	0.0000
4	VT+I+G4	2	-5045.4191	11139.4682	36.4782	0.0000
5	PMB+G4	1	-5054.7528	11153.3907	50.4007	0.0000
6	JTT-DCMUT+G4	1	-5054.9496	11153.7842	50.7942	0.0000
7	JTT-DCMUT+I+G4	2	-5054.6436	11157.9173	54.9273	0.0000
8	PMB+I+G4	2	-5054.7663	11158.1627	55.1727	0.0000
9	JTT+G4	1	-5057.1839	11158.2529	55.2630	0.0000
10	JTT+I+G4	2	-5056.9181	11162.4662	59.4762	0.0000

Fig. 2. ModelTest output for MUSCLE MSA sorted by BIC Score

2.5 Creating a Maximum Likelihood Tree for the MUSCLE-Aligned Sequences

We used "Phylogeny.fr", available publicly online at the Laboratoire d'Informatique de Robotique et de Microélectronique de Montpellier (LIRMM) website, to construct a maximum likelihood (ML) tree for the 111 MUSCLE-aligned neurotoxin sequences (Dereeper *et al.*, 2008). We used the `Advanced` phylogeny analysis tool, choosing only to include `Construction of phylogenetic tree: PhyML` and `Visualisation of phylogenetic tree: TreeDyn` in our Workflow Settings. Under parameters for PhyML, we used an `Approximate Likelihood-Ratio Test(aLRT)` for branch support and specified the `WAG (protein)` option with four substitution rate categories for substitution model.

Phylogeny.fr and our chosen parameters were appropriate choices for the purpose of our study. The in-built phylogenetic tree construction tool in MEGA11 is very slow on computers with M2 processing chips. MEGA11 estimated that generation of a ML tree for this dataset with 1,000 level bootstrapping would take over 500 hours on a 2022 MacBook Air. Phylogeny.fr processes MSA files much faster (about 20 minutes) and, unlike other online tools, also allows for users to specify a wide range of best-fit models of evolution (including WAG). Though Phyolgeny.fr also estimated over 100 hours of processing for bootstrapping, it offered aLRT as a much faster alternative statistical test for branch support. Anisimova et al. found that aLRT offers "not only speed advantages but also excellent levels of accuracy and power" for protein alignments when compared to bootstrapping and is therefore an efficient and accurate choice for this study (Anisimova *et al.*, 2011).

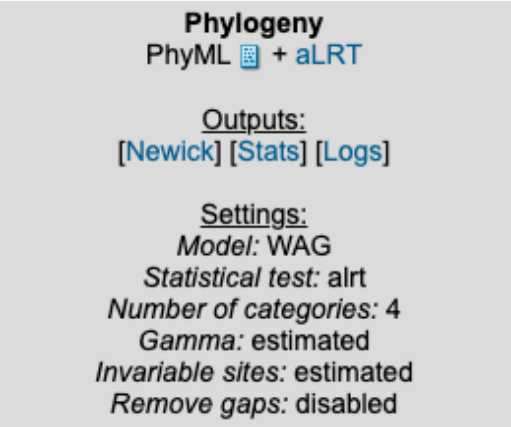

Fig. 3. Parameters for Phylogeny.fr ML Tree

2.6 Visualizing and Annotating the ML Tree in iTOL

The output tree from Phylogeny.fr was difficult to visualize and interpret. We saved this tree as a Newick file and uploaded it to iTOL for visualization and manual annotation (Letunic and Bork, 2007). These trees can be seen in Sections 3.2 and 3.3.

3 Results

3.1 Site-Level Amino Acid Conservation from MUSCLE MSA

Figure 4 below shows the result from toggling the displayed conservation level to 80% or more for all 111 sequences (first 15 sequences pictured) following alignment by MUSCLE. Sites shaded in black represent those at which 89 or more of the 111 sequences share a common amino acid (or gap). Sites with asterisks represent those where *all* 111 sequences share a common amino acid.

Many of these shaded sites are gapped regions or single amino acids. However, we observed the high conservation of G–C (sites 81-83) and CPXXK (sites 87-91) sub-sequences, highlighted in red on Figure 4. We also observed a highly conserved C–C sub-sequence (sites 98-100) in all but one neurotoxin–B2K5G3 from *B. fasciatus* being the exception–followed soon after by another highly conserved DXCN sub-sequence (sites 103-106), highlighted in green on Figure 4.

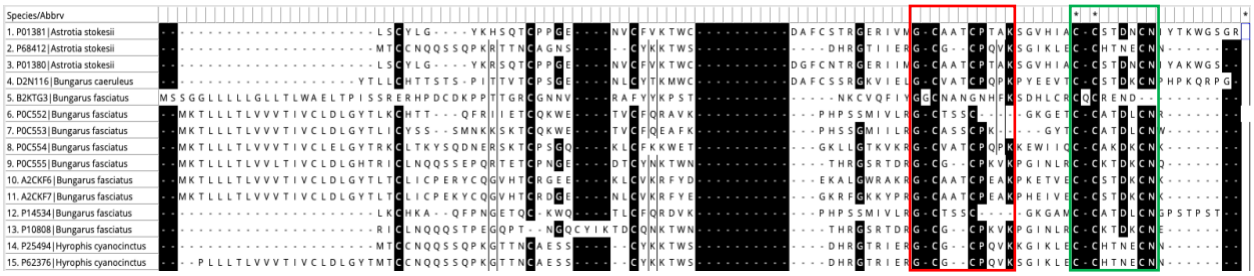

Fig. 4. Conserved Regions for MUSCLE MSA of neurotoxins. Black regions represent ≥80% conservation; red box contains G–C and CPXXK sites; green box contains C–C and DXCN sites.

3.2 Visualizing Circular ML Tree with aLRT Values

Figure 5 below shows the ML Tree for the MUSCLE MSA generated by Phylogeny.fr and visualized in circular form on iTOL (Dereeper *et al.*, 2008) (Letunic and Bork, 2007). Only branches with aLRT values greater than 50% are shown; blue circles on each node are sized proportionally to corresponding aLRT values. Branches and labels in red denote the sequences that are most similar to a sequence from a different species.

As can be seen, there are four pairs of neurotoxins in this ML tree that belong to the same sub-clade yet are derived from different snake species: CAB45156 (*N. naja*) and P0C554 (*B. fasciatus*); D2N116 (*B. caruleus*) and ACR55626 (*O. hannah*); P68412 (*A. stokesii*) and P62376 (*H. cyanocinctus*); and P10460 (*L. laticaudata*) and Q7T2I1 (*L. colubrina*). Of these pairs, the first three have aLRT values ≥ 50%, implying that this paired positioning is replicable in most ML tree iterations based on this MUSCLE MSA.

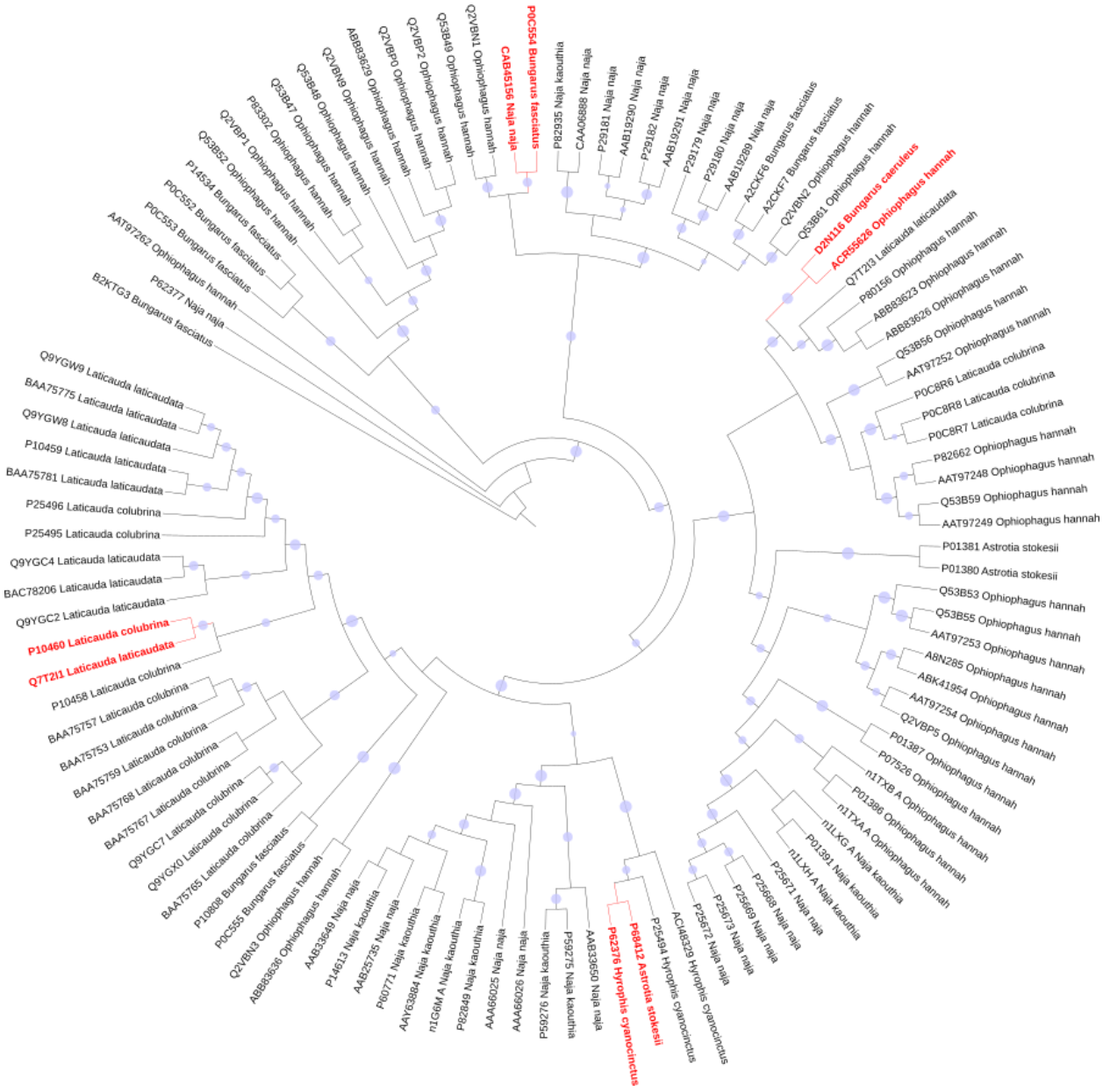

Fig. 5. Circular ML Tree Annotated in iTOL. Red clades denote those in which sequences are most closely related to a sequence from a different species. Blue circles are sized proportionate to aLRT values (only shown if ≥ 50%)

3.3 Further Annotations of Circular ML Tree

In Figure 6 below, we used the following criteria to select and highlight clades for interpretation:

1. Contains sequences from both *Bungarus* and *Naja* genera,
2. Contains sequences from at least three genera,
3. All branches have aLRT values ≥ 50%.

We selected these criteria to present statistically verified groups of proteins that can potentially be used in the development of a BSA that neutralizes the most common forms of envenomation in Bangladesh: snakebite from krait and cobra species. There are two such clades in this alignment, detailed below:

Clade A: Q53B49 and Q2VBN1 (*O. hannah*); CAB45156 (*N. naja*); P0C554 (*B. fasciatus*)

Clade B: P29179 and P29180 (*N. naja*); A2CKF6 and A2CKF7 (*B. fasciatus*); Q2VBN2 and Q53B61 (*O. hannah*)

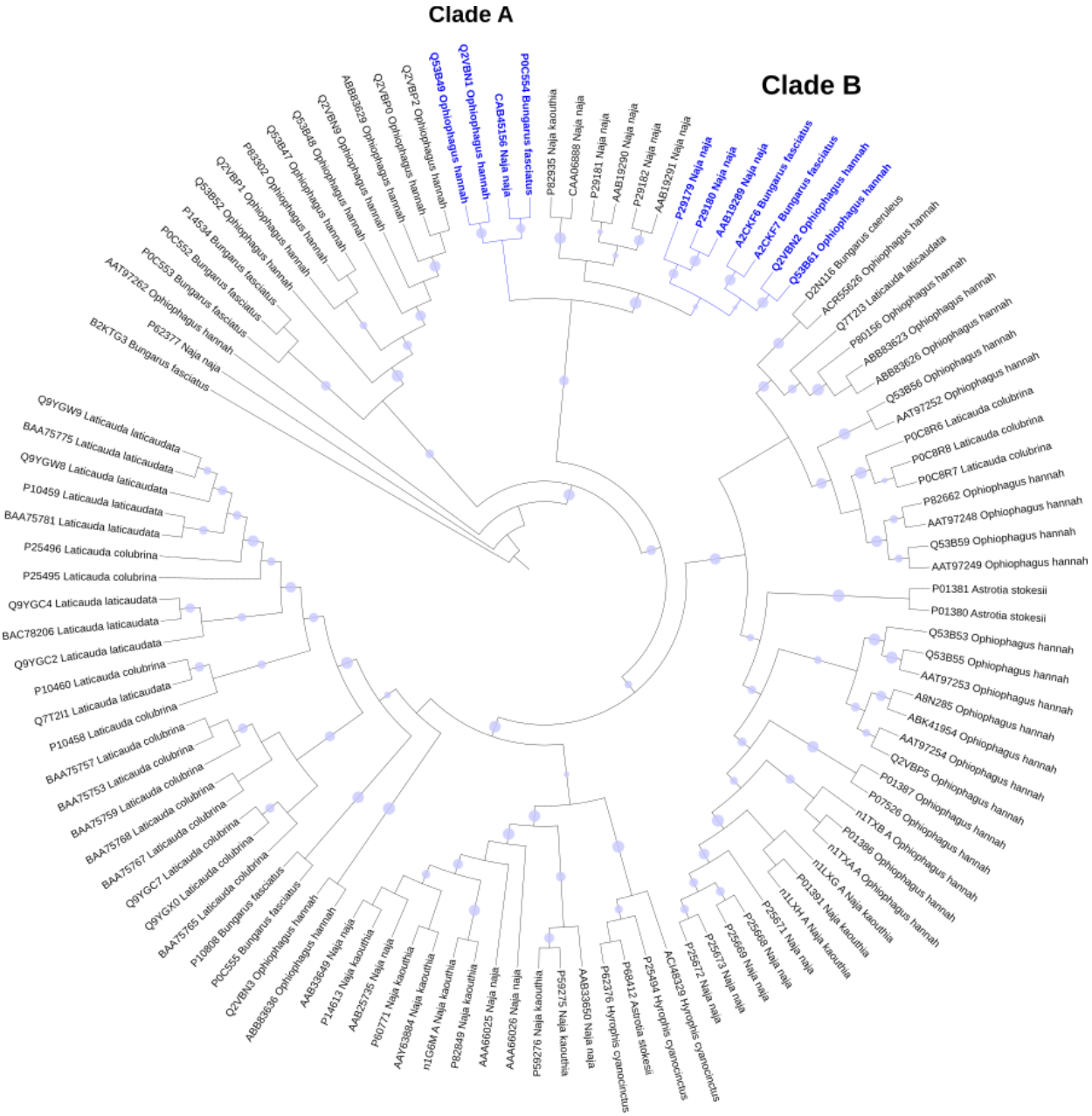

Fig. 6. Circular ML Tree with further annotations of clades containing sequences from *Bungarus* and *Naja* genera, 2+ genera total, and in which all branch aLRT values exceed 50%.

4 Discussion

4.1 Interpreting Observations from Site-Level Conservation Analysis

Following alignment by MUSCLE, we noted the existence of two highly conserved regions in `VenomSeqs.fasta`. These regions were common in well over 80% of our collected sequences, meaning that they are present in the vast majority of all neurotoxic venom proteins within the ISOB database. This is a novel yet preliminary finding and reveals a potential candidate sub-sequence for BSA manufacturers.

It is important to note that the development of PAV also began by finding granular similarities between snake venom proteins; however, *several* intermediate steps were needed to progress beyond this finding and to successfully develop polyvalent antivenom (Madhushani *et al.*, 2021). In this case, further research is needed to address the following questions (and beyond):

1. Can *in vitro* constructions of proteins composed of these specific sub-sequences be used to raise antibodies in host animals?
2. If so, can these antibodies effectively neutralize neurotoxic venom from several different snake species indigenous to Bangladesh (specifically, species within the *Bungarus* and *Naja* genera)?

4.2 Interpreting Patterns from Annotated ML Trees

Figure 5 showed four pairs of neurotoxin sequences more similar to each other than any other sequence, despite being derived from different species. Of these four pairs, only one contained sequences from both *Bungarus* and *Naja* genera: CAB45156 and P0C554. Given their close sequence similarity, this protein pair appears to be well suited for the production of a BSA that can neutralize the most common types of envenomation in Bangladesh. Further research is needed to determine if antibodies raised against one of these proteins can successfully neutralize the other, and whether this is enough to counter the overall physiological impacts of envenomation by both *B. fasciatus* and *N. naja*.

Figure 6 expands upon the findings of Figure 5, identifying two clades that contain highly similar sequences of various cobra and krait venom proteins, with statistical verification of each branch position. Both of these clades contain sequences from *B. fasciatus*, *N. naja*, and *O. hannah* and are therefore potential candidate proteins for the production of even broader spectrum antivenoms in Bangladesh. We note that Clade B is the best candidate for BSA production of all groups in Figures 5 and 6 due to the presence of *multiple* (rather than one) closely related proteins from each genus. Antibodies raised against these proteins could potentially neutralize more elements of neurotoxic venom cocktails and are thus more likely to limit overall physiological reactions to envenomation.

4.3 Conclusions from this study

We identified two sub-sequences common to several neurotoxic proteins, as well as a group of highly similar neurotoxic proteins from various cobra and krait species as potential candidates for BSA development. These results are preliminary and require further research to establish the viability of using these amino acid sequences and proteins in producing antivenom that can neutralize various neurotoxins. However, they are novel findings that mimic those from

the early stages of PAV development and, with the appropriate extensions of this study, could lay the groundwork for Bangladesh's first broad spectrum antivenom targeting neurotoxins.

References

Anisimova,M. *et al.* (2011) Survey of Branch Support Methods Demonstrates Accuracy, Power, and Robustness of Fast Likelihood-based Approximation Schemes. *Syst. Biol.*, **60**, 685–699.

Annan,K. (2018) Snakebite: The biggest public health crisis you've never heard of. *Kofi Annan Found.*

Darriba,D. *et al.* (2020) ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Mol. Biol. Evol.*, **37**, 291–294.

Dereeper,A. *et al.* (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465-469.

Ghose,A. (2021) Snakebite in Bangladesh |.

Letunic,I. and Bork,P. (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinforma. Oxf. Engl.*, **23**, 127–128.

Madhushani,U. *et al.* (2021) Effect of Indian Polyvalent Antivenom in the Prevention and Reversal of Local Myotoxicity Induced by Common Cobra (Naja naja) Venom from Sri Lanka In Vitro. *Toxins*, **13**, 308.

Maduwage,K. *et al.* (2016) Efficacy of Indian polyvalent snake antivenoms against Sri Lankan snake venoms: lethality studies or clinically focussed in vitro studies. *Sci. Rep.*, **6**, 26778.

Roly,Z.Y. *et al.* (2015) ISOB: A Database of Indigenous Snake Species of Bangladesh with respective known venom composition. *Bioinformation*, **11**, 107–114.

Stecher,G. *et al.* (2020) Molecular Evolutionary Genetics Analysis (MEGA) for macOS. *Mol. Biol. Evol.*, **37**, 1237–1239.

Tamura,K. *et al.* (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.*, **38**, 3022–3027.

Tasoulis,T. *et al.* (2021) Investigating Toxin Diversity and Abundance in Snake Venom Proteomes. *Front. Pharmacol.*, **12**, 768015.

von Reumont,B.M. *et al.* (2022) Modern venomics—Current insights, novel methods, and future perspectives in biological and applied animal venom research. *GigaScience*, **11**, giac048.

Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

Willard,N.K. *et al.* (2021) Proteomic Identification and Quantification of Snake Venom Biomarkers in Venom and Plasma Extracellular Vesicles. *Toxins*, **13**, 654.

Xiao,H. *et al.* (2017) Snake Venom PLA2, a Promising Target for Broad-Spectrum Antivenom Drug Development. *BioMed Res. Int.*, **2017**, 6592820.