Abstract:

Venomous animals have evolved specialized proteins that play a crucial role in their predatory and defensive strategies. Understanding the evolutionary relationships among venomous species and the proteins they produce is essential for unraveling the molecular basis of venom and its adaptive significance. This research proposal outlines a project aimed at estimating a phylogenetic tree based on venom protein sequences from diverse species. The study will utilize Bayesian phylogenetic analysis to uncover the evolutionary history of venomous lineages.

1. Introduction:

Venomous animals represent a fascinating and diverse group of organisms that have evolved an extraordinary arsenal of toxins. These toxins serve as multifaceted molecular weapons, enabling venomous species to subdue prey, deter predators, and protect themselves from threats. Within this intricate realm of venomous adaptations, venom proteins stand out as highly specialized molecules that have undergone extensive evolutionary changes. These proteins play a pivotal role in the delivery and efficacy of venom, making them a focal point of scientific inquiry.

Venomous animals occupy unique ecological niches and are often key players in their respective ecosystems. Venom is an evolutionary innovation that has arisen independently in multiple lineages, illustrating the remarkable power of adaptation in response to selective pressures. Understanding the evolution of venomous lineages and the molecular mechanisms driving this diversity is a scientific endeavor of paramount importance.

At the heart of venomous adaptations are venom proteins, the molecular architects of venom's potency. These proteins encompass a bewildering array of functions, from neurotoxicity and hemotoxins to cytotoxins and digestive enzymes. The astounding diversity of venom proteins is matched only by their capacity for rapid evolutionary change. This evolutionary dynamism reflects the ongoing arms race between venomous animals and their ecological counterparts.

While the study of venomous adaptations has yielded remarkable insights, a crucial research gap remains. Bayesian statistics, a powerful analytical framework, have seldom been employed to elucidate the evolutionary history of venomous species based on venom proteins. The integration of Bayesian methods offers the potential to provide robust estimates of phylogenetic trees and unravel the intricate web of relationships among venomous lineages. Bayesian approaches handle ambiguity and missing data in phylogenetic analysis through their probabilistic framework. They treat uncertain or missing sequence data as part of the model, allowing for the estimation of these unknowns alongside the phylogenetic tree. This approach involves assigning probabilities to different possible states of the ambiguous or missing data, based on the information available from other parts of the sequence and the model's parameters. By integrating over these uncertainties, Bayesian methods provide a more robust and comprehensive analysis, which can lead to more reliable phylogenetic inferences compared to methods that might simply ignore or inadequately handle such data. Bayesian analysis efficiently manages uncertainty in model parameters. It calculates probabilities of different phylogenetic trees, giving a more nuanced understanding of evolutionary relationships. This approach allows for the integration of prior knowledge (e.g., substitution models) into the phylogenetic analysis. It's versatile in reconstructing evolutionary relationships, especially when data is limited and complex. Furthermore, the application of deep learning techniques holds promise in streamlining the analysis of vast datasets of venom protein sequences, facilitating the exploration of sequence evolution. The primary objectives of this research project is to estimate the phylogenetic relationship between various species of venomous snakes across the Elapids and Viperids leveraging snake venom protein sequences, highlighting the central importance of venom evolution in the overall evolution of venomous snakes.

2. Related Work

Much work has been done on phylogenetic trees of various animalistic species to understand the evolutionary process resulting in the variance in species. Specifically, research on venomous animals has a longstanding history in the field of evolutionary biology and toxicology. Recently, Bayesian phylogenetic methods have gained prominence in recent years due to their ability to provide robust estimates of phylogenetic trees while accommodating uncertainty in model parameters. Bayesian approaches allow for the incorporation of prior information, such as substitution models, into the phylogenetic analysis, making them versatile tools for reconstructing evolutionary relationships.

(Mark Holder, 2003) delves into estimating phylogenetic trees using Bayesian statistics and concludes that Bayesian approaches can account for uncertainty in parameter estimates by marginalizing over parameters. In the same paper, it is argued that Markov Chain Monte Carlo can estimate posterior probability, making Bayesian phylogenetics feasible for more datasets. (Kelly) suggests that Bayesian approaches are as good as maximum likelihood approach in case of phylogenetics but can yield answers in less time. This study, however, focuses on mammalian nucleotide samples as the dataset hence does not consider other species. (Sebestian Hohna, 2017) introduces a software for Bayesian phylogenetics along with substitution models which can be incorporated into this study as the dataset used is that of primates in case of the mentioned research. (Suranse V, 2022) explores the phylogenetic relationship between snake venom based on the protein sequences found in the venom to provide insights into evolution of Phospholipase A_2 venom toxin along with the ecological factors that may have resulted in the evolution. However, the study only focuses on a specific toxin and does not incorporate Bayesian approaches in estimating the phylogenetics. (David Posada, 2004) addresses the field of molecular phylogenetics estimation using Bayesian methods and concludes that Bayesian approaches can compare multiple models, assess model selection uncertainty and allow for the estimation of phylogenies and model parameters using all available models. This is an instrumental study in the context of our research as it provides valuable insights into the methodology to be followed. However, this study, due to the constraints of the age in which it was conducted, could not use the power of available computation to address a specific application area like estimating phylogenetic trees using venom proteins, and hence serves as a technical aide instead of an application aide.

3. Methodology

Relevant and robust data is the bedrock of every AI application. The initial phase of our study entailed a meticulous process of data collection, where venom protein sequences from a diverse range of species were gathered. The dataset was obtained from Uniprot (Consortium, 2023), a comprehensive, high-quality, and freely accessible resource of protein sequence and functional information. The specific sequences included in our analysis were selected based on their relevance to rattlesnake venom and the study's phylogenetic scope.

Upon acquisition in the fasta file format, the sequences were processed to extract the protein sequences along with the species associated with the protein sequence. The sequences underwent a rigorous preprocessing regimen. This involved a detailed examination to identify and correct any ambiguities or potential sequencing errors present in the dataset. Such preprocessing is crucial as it directly impacts the reliability of the phylogenetic inferences drawn from the analysis. The integrity of the data was maintained by ensuring that only sequences with the highest quality were advanced to the alignment phase.

After the processing of the data, the next step is to conduct Multiple sequence alignment. The sequence alignment process is a pivotal step in phylogenetic analysis. It involves arranging the venom protein sequences in a way that identifies homologous (evolutionarily related) sequences. For this study, two primary tools were employed: Clustal Omega and the Hhalign algorithm.

Clustal Omega:

Clustal Omega (Sievers, 2018) performs multiple sequence alignments (MSA) by first creating pairwise alignments using a k-tuple method to identify similar regions across sequences. The sequences are then progressively aligned according to the guide tree, starting from the most similar pairs. It employs Hidden Markov Models (HMMs) for better alignment accuracy, especially with larger datasets. Hidden Markov Models (HMMs) are statistical models that represent probabilistic sequences of hidden (unobservable) states. In the context of protein alignments, each state corresponds to an amino acid or a gap. The mathematical structure of an HMM includes:

1. **States:** Represent different amino acids or gaps. The model transitions between these states with certain probabilities.
2. **Transition Probabilities:** Represent the likelihood of moving from one state to another. These are captured in a transition matrix *A*, where *Aij* is the probability of transitioning from state *i* to state *j*.
3. **Emission Probabilities:** Each state has an emission probability, indicating the likelihood of observing a particular amino acid when the model is in that state.
4. **Initial Probabilities:** Probabilities of the model starting in each state.

Mathematically, Clustal Omega employs a scoring matrix to calculate alignment scores, optimizing these scores to produce the most biologically relevant alignment. Once the HMM extraction is complete from Clustal Omega, the HMMs must be aligned to find areas of similarity across the sequences. For this purpose, the Hhalign Algorithm is used.

Hhalign Algorithm:

The Hhalign algorithm (Söding, 2005) is part of the HH-suite and operates on the principle of aligning Hidden Markov Models (HMMs) rather than the sequences themselves. An HMM is built for each sequence or sequence group, modeling the probability of observing certain amino acids and the transitions between different states (amino acids or gaps). The alignment is then computed using dynamic programming, maximizing the probability of the observed sequences given the HMMs. This approach is particularly powerful for aligning distantly related sequences as it considers both sequence and structural information.

Once the Sequence Alignment has been accomplished, the resultant sequences exhibit several key features:

- Homologous Regions Identification: The alignment identifies regions where sequences are similar or identical across different rattlesnake species. These homologous regions are essential as they suggest functional or structural conservation in the venom proteins. These similarities could be indicative of shared evolutionary histories or similar ecological adaptations among the species.
- Gaps in the Sequences: To facilitate the alignment of homologous regions, gaps are introduced into the sequences. Represented typically by dashes ('-'), these gaps denote insertions or deletions (indels) in the protein sequences among the species. The presence and distribution of gaps can provide insights into the evolutionary processes that have shaped these proteins.
- Length of the Sequences: As a result of alignment, all sequences will have the same length. This uniform length encompasses the original lengths of the sequences and the introduced gaps. The alignment length is critical for subsequent phylogenetic analysis, as it ensures that each position across the sequences can be directly compared.
- Variable and Conserved Regions: The alignment will reveal both variable and conserved regions across the sequences. Conserved regions, characterized by little to no variation, imply evolutionary conservation and potentially shared functional aspects among the species. Conversely, variable regions, where there are notable differences in amino acid sequences, indicate evolutionary divergence. These regions are particularly interesting for understanding the adaptive evolution of venom proteins.

Following the process of sequence alignment, our study transitions into the critical phase of Bayesian phylogenetic analysis using MrBayes. At this point, it is crucial to consider Substitution Models and our choice of them as they describe how sequences change over time due to evolutionary processes. Precisely, they provide a framework for understanding how one amino acid replaces another. For this study, The Jones-Taylor-Thornton substitution model (Jones, 1992) was chosen due to its specific utility with amino-acid based models. It is based on empirical data derived from observed changes in protein sequences. Mathematically, the model is represented by a 20x20 substitution matrix, where each element of the matrix corresponds to the relative probability of one amino acid being replaced by another over time. The model assumes that substitution rates are constant over time and across different lineages, but allows for varying rates of substitution between different amino acids. The JTT model was chosen for its suitability in analyzing protein sequences, offering a balance between biological realism and computational efficiency.

After selecting the Jones-Taylor-Thornton (JTT) substitution model based on its suitability for protein sequences, our analysis progresses to the computational phase where the theoretical framework is applied to the actual sequence data. This step is pivotal in translating the abstract probabilities and rates of the substitution model into a concrete phylogenetic tree. Since the approach being explored in this study is Bayesian in Nature, it is imperative to consider priors at this point. In the context of phylogenetic tree estimation, priors are defined as already known relationships between different species which can ensure greater accuracy in tree estimation. Relationships between some key species are already specified in the process before further processing. With the substitution model and priors in place, we introduce the MCMC method (Metropolis, 1953). MCMC is a computational technique used to sample from a probability distribution when direct sampling is challenging.

- In the context of phylogenetic analysis, MCMC is used to sample from the posterior distribution of trees. This involves calculating the probability of a tree given the data (the likelihood) and the prior probability of the tree.
- The Metropolis-Hastings algorithm, a common MCMC method, involves proposing a new tree, calculating the acceptance ratio based on the likelihood and the prior, and then deciding whether to accept or reject the new tree.
- Mathematically, the acceptance ratio α for a new state is given by α=min(1,P(current state∣data)P(new state∣data)), where P(state∣data) is the posterior probability of the state given the data.

Through the integration of the JTT model and MCMC methods, our study robustly inferred the phylogenetic relationships among snake species based on their venom protein sequences. This approach allowed us to incorporate both the empirical evidence from the sequences and the probabilistic framework of Bayesian inference. Using MrBayes, the MCMC process iteratively proposes new phylogenetic trees. At each step, a new tree is proposed, and its likelihood is calculated based on the JTT model. The Metropolis-Hastings algorithm (SiddharthaChib, 1995), a key component of MCMC, is then used to decide whether to accept this new tree. The decision is based on how probable the new tree is compared to the current one, considering the observed data. The MCMC chains are run for a pre-specified number of generations, ensuring adequate sampling from the posterior distribution. Convergence diagnostics are monitored to confirm that the chains have explored the tree space sufficiently and are representative of the posterior distribution. Upon completion of the MCMC run, the sampled trees are summarized. The consensus tree, which represents the most probable phylogenetic relationships given the data and the model, is derived from these samples.

4. Results

Our Bayesian phylogenetic analysis, utilizing the robust computational framework of MrBayes, has yielded a comprehensive phylogenetic tree, as visualized in the accompanying figure. The tree, derived from venom protein sequences of various snake species, encapsulates the inferred evolutionary relationships, grounded in the principles of Bayesian inference and the Jones-Taylor-Thornton (JTT) substitution model.

Evaluation Metrics and Tree Summary:

Burn-in Phase: The analysis began with the discarding of the initial 25% of the sampled trees from each run, deemed the burn-in phase, to ensure that only the trees sampled from the stationary distribution were considered for constructing the consensus tree.

Tree Sampling: A total of 802 trees were read from the two runs, with 301 trees sampled from each run after the burn-in phase, providing us with a comprehensive set of trees to analyze.

Convergence of Runs: The convergence diagnostics indicate that the MCMC runs have adequately explored the parameter space. This is evidenced by the average standard deviation of split frequencies approaching 0.024255, suggesting that the runs are converging towards a common distribution. The potential scale reduction factor (PSRF) values for most parameters approached 1, further confirming convergence.

Bipartition Analysis:

The analysis identified several bipartitions with high posterior probabilities, denoting them as robust groupings supported by the data. Bipartitions with a posterior probability of 1.0 were consistently observed across all runs, underpinning the stability and reliability of these clades. Variability in the posterior probabilities of some bipartitions across runs was minimal, indicating agreement between independent runs.

Branch Length Estimation:

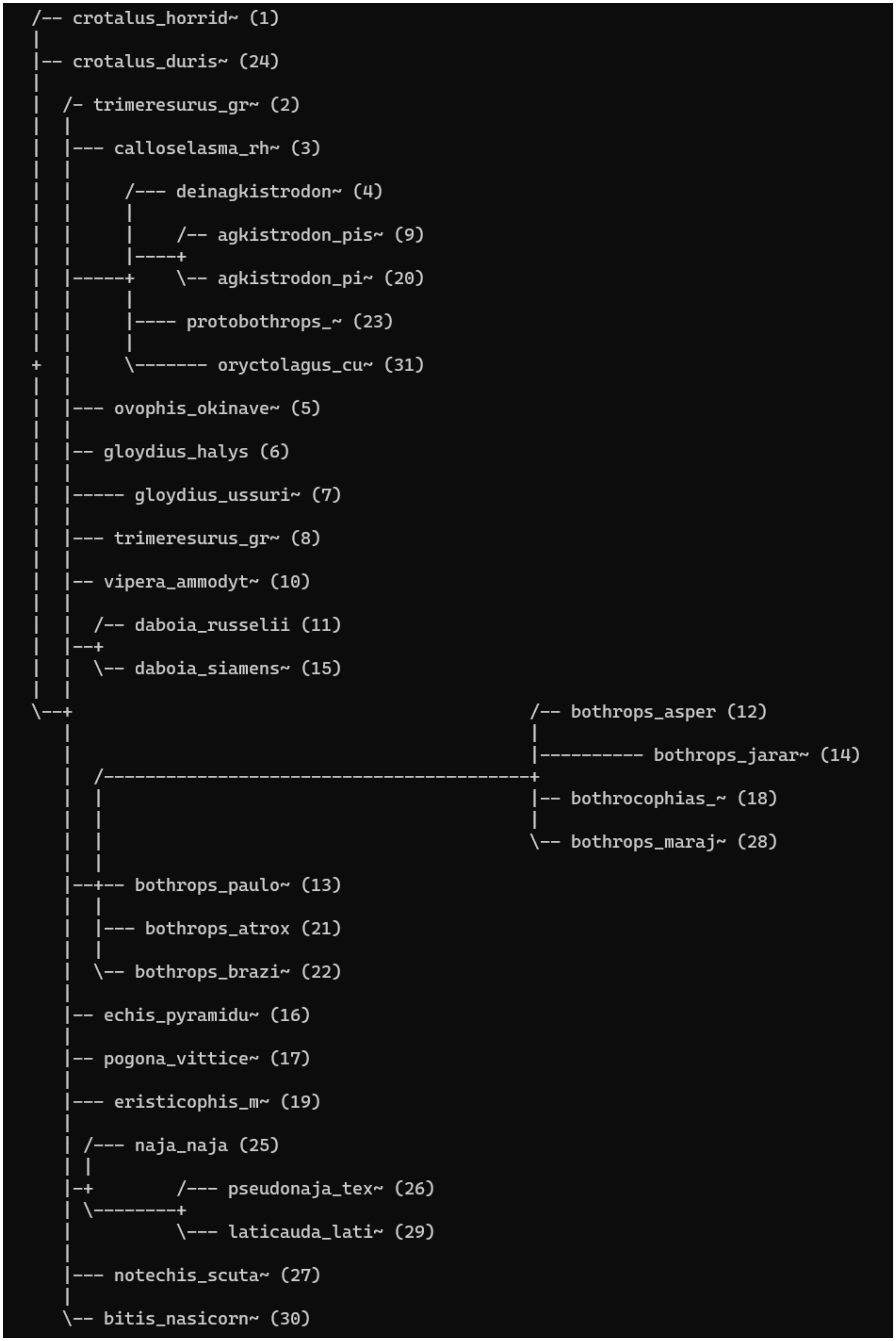
The summary statistics for branch lengths provided estimates for each parameter, including the mean, variance, and a 95% highest posterior density (HPD) interval. These metrics offer a probabilistic interpretation of the branch lengths, contributing to our understanding of the evolutionary distances between taxa.

Partition Frequency Analysis:

The maximum standard deviation of split frequencies reported was 0.093968, within an acceptable range, suggesting that the different MCMC runs yielded similar frequencies for the most part. The average PSRF for parameter values was 1.009, with the maximum PSRF observed being 1.123. While PSRF values ideally should be close to 1, slight deviations suggest that additional runs may be beneficial to confirm certain parameters.

Summary:

The consensus phylogenetic tree derived from our analysis provides a visual and statistical testament to the evolutionary narratives embedded within the venom protein sequences. The high posterior probabilities of the clades within the tree underscore the confidence in the inferred relationships. The convergence diagnostics and the close examination of bipartitions, branch lengths, and parameters bolster the robustness of our phylogenetic conclusions. This tree not only serves as a pivotal resource for understanding the evolutionary history of these species but also provides a solid foundation for future research into their evolutionary biology.



```
/-- crotalus_horrid~ (1)
|
|-- crotalus_duris~ (24)
|
|  /- trimeresurus_gr~ (2)
|  |
|  |--- calloselasma_rh~ (3)
|  |
|  |     /--- deinagkistrodon~ (4)
|  |     |
|  |     |    /-- agkistrodon_pis~ (9)
|  |     |----+
|  |-----+    \-- agkistrodon_pi~ (20)
|  |     |
|  |     |---- protobothrops_~ (23)
|  |     |
+  |     \------- oryctolagus_cu~ (31)
|  |
|  |--- ovophis_okinave~ (5)
|  |
|  |-- gloydius_halys (6)
|  |
|  |----- gloydius_ussuri~ (7)
|  |
|  |--- trimeresurus_gr~ (8)
|  |
|  |-- vipera_ammodyt~ (10)
|  |
|  |  /-- daboia_russelii (11)
|  |--+
|  |  \-- daboia_siamens~ (15)
|  |
\--+                                            /-- bothrops_asper (12)
   |                                            |
   |                                            |---------- bothrops_jarar~ (14)
   |  /-----------------------------------------+
   |  |                                         |-- bothrocophias_~ (18)
   |  |                                         |
   |  |                                         \-- bothrops_maraj~ (28)
   |  |
   |--+-- bothrops_paulo~ (13)
   |  |
   |  |--- bothrops_atrox (21)
   |  |
   |  \-- bothrops_brazi~ (22)
   |
   |-- echis_pyramidu~ (16)
   |
   |-- pogona_vittice~ (17)
   |
   |--- eristicophis_m~ (19)
   |
   | /--- naja_naja (25)
   | |
   |-+        /--- pseudonaja_tex~ (26)
   | \--------+
   |          \--- laticauda_lati~ (29)
   |
   |--- notechis_scuta~ (27)
   |
   \-- bitis_nasicorn~ (30)
```

Fig 1: Resulting Phylogenetic Tree

Bibliography

Consortium, T. U. (2023). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*.

David Posada, T. R. (2004). Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology*, 793-808.

Jones, D. T. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences (CABIOS)*, 275-282.

Kelly, C. D. (n.d.). Understanding mammalian evolution using Bayesian phylogenetic inference.

Mark Holder, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature Review Genetics*, 275-283.

Metropolis, N. R. (1953). Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, 1087–1092.

Sebestian Hohna, M. L. (2017). Phylogenetic Inference Using RevBayes.

SiddharthaChib, E. (1995). UnderstandingtheMetropolis-HastingsAlgorithm. *The American Statistician, Vol.49, No.4*, 327-335.

Sievers, F. &. (2018). Clustal Omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology*, 105-116.

Söding, J. (2005). Protein homology detection by HMM-HMM comparison. *Bioinformatics*, 951-960.

Suranse V, J. T. (2022). Contextual Constraints: Dynamic Evolution of Snake Venom Phospholipase A2. *doi: 10.3390/toxins14060420*.