Análise filogenética computacional de serpentes do gênero *Bothrops* a partir de proteomas de venenos

Bolsista: Victor Wichmann Raposo

Orientador: Marcelo da Silva Reis

Centro de Toxinas, Imuno-resposta e Sinalização Celular (CeTICS)

Laboratório Especial de Ciclo Celular (LECC), Instituto Butantan

São Paulo, 4 de abril de 2018

Resumo

Venenos de serpentes são complexas misturas proteicas, cujas proteínas podem receber quantidades variadas de glicosilação. Existe variação inter-espécie tanto na composição da mistura (proteoma) quanto nos tipos de glicanos que se ligam a suas proteínas. Recentemente, foram demonstradas evidências de que, entre serpentes do gênero *Bothrops*, tanto um cladograma obtido a partir do proteoma quanto um gerado utilizando estruturas de N-glicanos se correlacionam com o cladograma filogenético produzido através de DNA mitocondrial (mtDNA) e/ou de características morfológicas. Todavia, não foram aplicadas nesses estudos métricas quantitativas para comparação entre os diferentes cladogramas. Além disso, não foi totalmente explorado o uso das informações fornecidas pelos peptídeos detectados nos ensaios de proteômica baseada em espectrometria de massas. Neste projeto, utilizando as mesmas informações biológicas de venenos de sete espécies de serpentes do gênero *Bothrops* apresentados em estudos anteriores, propomos o desenho de cladogramas filogenéticos que combinarão informações dos proteomas, incluindo os peptídeos utilizados na etapa de identificação proteica, com as de estruturas de N-glicanos. Para este fim, utilizaremos uma abordagem de inferência Bayesiana, empregando métodos de Monte Carlo com cadeias de Markov. Para analisar os resultados, implementaremos uma métrica de comparação entre cladogramas, para assim podermos quantificar a distância das novas árvores filogenéticas em relação a uma produzida com mtDNA e/ou características morfológicas. Dessa forma, esperamos testar a hipótese de que o perfil glicoproteômico dos venenos de serpentes do gênero *Bothrops* está altamente correlacionado com a sua filogenia.

Computational phylogenetic analysis of *Bothrops* snakes from venom proteomes

Student: Victor Wichmann Raposo

Advisor: Marcelo da Silva Reis

Center of Toxins, Immune-response and Cell Signaling (CeTICS)

Laboratório Especial de Ciclo Celular (LECC), Instituto Butantan

São Paulo, April 4, 2018

Abstract

Snake venoms are complex protein-based mixtures, whose proteins can undergo variable levels of glycosylation. There is interspecies variation both in the mixture composition (proteome) and in the types of glycan structures that bind to its proteins. Recently, it was presented evidences that, among *Bothrops* snakes, cladograms obtained using either proteome or N-glycan structures correlate with the phylogenetic cladogram produced through mitochondrial DNA (mtDNA) and/or morphological characters. However, in these studies, it was not applied quantitative metrics for comparison among different cladograms. Moreover, it was not totally exhausted the usage of information contained in the peptides detected during the mass spectrometry-based proteomics assays. In this project, using the same biological information presented in previous studies, which covers venoms from seven *Bothrops* snakes, we propose the designing of phylogenetic cladograms that will combine information from proteomes, including their peptides, with the one from N-glycan structures. To this end, we will make use of a Bayesian inference approach, using Markov chain Monte Carlo methods. To analyze the results, we will implement a comparison metric for cladograms, which will allow us to measure the distance of the new phylogenetic trees in respect to one produced with mtDNA and/or morphological characters. Therefore, we expect to test the hypothesis that the glycoproteomic profile of venoms from *Bothrops* snakes is highly correlated to their phylogeny.

Sumário

1 Introdução

Venenos de serpentes são misturas proteicas altamente complexas, usadas tanto para a defesa contra predadores quanto como meio de imobilização e digestão de presas. O conjunto dessas proteínas também é denominado proteoma. As proteínas que compõem esse conjunto podem sofrer mudanças pós-traducionais chamadas glicosilações, que são ligações de um glicano (i.e., um polissacarídeo) a um dos aminoácidos de uma dada proteína. Se essa ligação se dá especificamente no átomo de nitrogênio da amida de uma asparagina, então denominamos esse processo como N-glicosilação; já se a ligação ocorre no átomo de oxigênio de um dado aminoácido, então denominamos o processo como O-glicosilação. Glicosilações são reações de grande relevância biológica, por se tratarem de um dos tipos mais prevalentes de modificação pós-traducional de proteínas [1].

O proteoma dos venenos pode sofrer um nível variado de glicosilação, contribuindo, assim, para a complexidade dessa mistura e para uma diferenciação entre venenos de cada espécie de serpente. Após estudar a variabilidade entre esses venenos como uma função dos níveis de glicosilação das proteínas de seus respectivos proteomas, a pesquisadora Solange M.T. Serrano, do Laboratório Especial de Toxinologia Aplicada (LETA) do Instituto Butantan, reportou indícios de que existe um núcleo de glicoproteínas que define o perfil de cada veneno de serpentes do gênero *Bothrops* [2]. Além disso, tal perfil se correlaciona com a classificação filogenética feita com marcadores mais tradicionais, tais como o DNA mitocondrial (mtDNA) e características morfológicas (Figura 1). Mais recentemente, a mesma pesquisadora fez, em uma colaboração com Vernon Reinhold (Universidade de New Hampshire, EUA), uma análise comparativa das estruturas de N-glicanos presentes nos venenos das mesmas serpentes; estes novos resultados corroboraram as conclusões apresentadas no estudo anterior [3].

No entanto, nesses dois trabalhos foram feitas análises qualitativas das árvores filogenéticas obtidas (cladogramas). Isso significa que foram inspecionadas as relações de ordem dos cladogramas glicoproteômicos, gerados através de procedimento de aglomeração hierárquica sobre proteomas ou estruturas de glicanos, comparando-as com as de cladogramas obtidos com informações de mtDNA e/ou características morfológicas [4], sem utilizar métricas para fazer uma medida quantitativa das distâncias entre diferentes árvores. Além disso, não foi investigado o uso direto dos peptídeos dos proteomas, identificados por espectrometria de massas, para a construção dos cladogramas; isto é, após o uso desses mesmos peptídeos na identificação de proteínas através de busca em banco de dados, estas são utilizadas para produzir o cladograma, enquanto que aqueles são descartados. Aproveitar essa informação que é jogada fora poderia melhorar os resultados obtidos, além de possivelmente

mitigar o viés causado por espécies super-representadas nesses bancos de dados (e.g., *B. jararaca*). Por fim, também permanece como problema em aberto uma análise filogenética que combine as informações biológicas heterogêneas obtidas nos estudos anteriores [2, 3] e que, posteriormente, agregue também outros dados disponíveis na literatura.

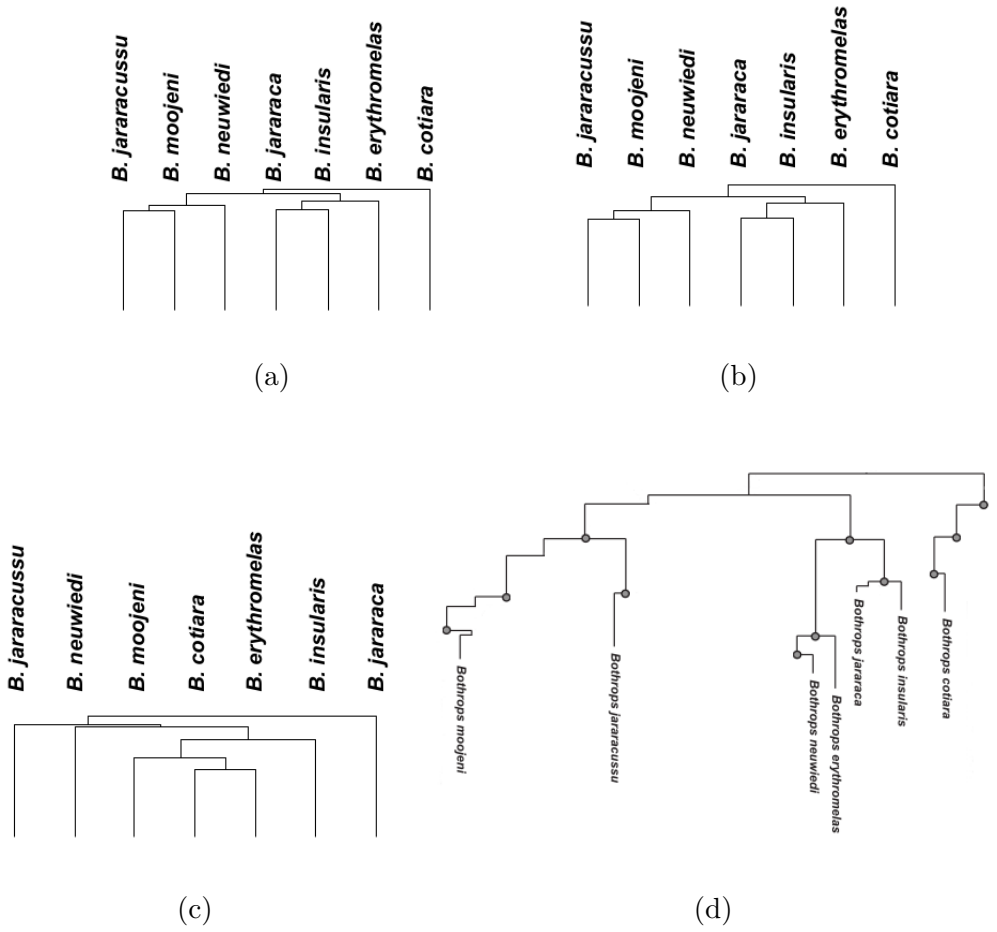

Figura 1: Cladogramas filogenéticos de sete espécies do gênero *Bothrops*, obtidos utilizando diferentes marcadores moleculares. São apresentados cladogramas construídos a partir de proteoma total de veneno (Fig. 1a), de glicoproteínas detectadas através de proteomas total e baseados em protocolos de enriquecimento por afinidade a lectinas (Fig. 1b), e de não-glicoproteínas detectadas nos mesmos ensaios de proteomas anteriores (Fig. 1c). Já na Fig. 1d é mostrada uma sub-árvore de um cladograma obtido através do uso de mtDNA e de características morfológicas. Observe que, à exceção de *B. neuwiedi*, os cladogramas das Figs. 1a e 1b apresentam a mesma hierarquia da subárvore da Fig. 1d, enquanto que o da Fig. 1c é dissimilar em comparação aos demais. As Figs. 1a–1c foram extraídas de Andrade-Silva et al. [2], enquanto que a Fig. 1d foi adaptada de Fenwick et al. [4].

2 Objetivos

Os objetivos geral e específico deste projeto proposto são:

- Geral: Montar um encadeamento (*pipeline*) de processos para desenho, comparação e visualização de análises filogenéticas computacionais a partir de informações biológicas heterogêneas.
- Específico: Aplicar o encadeamento desenvolvido para testar a hipótese de que o perfil glicoproteômico dos venenos de serpentes do gênero *Bothrops* está fortemente correlacionado com a filogenia observada em análises que empregam mtDNA e/ou características morfológicas dessas espécies.

3 Metodologia

O encadeamento proposto consiste em quatro processos principais, que estão organizados de acordo com o fluxograma apresentado na figura 2. Ao longo desta seção, descreveremos cada um desses processos.

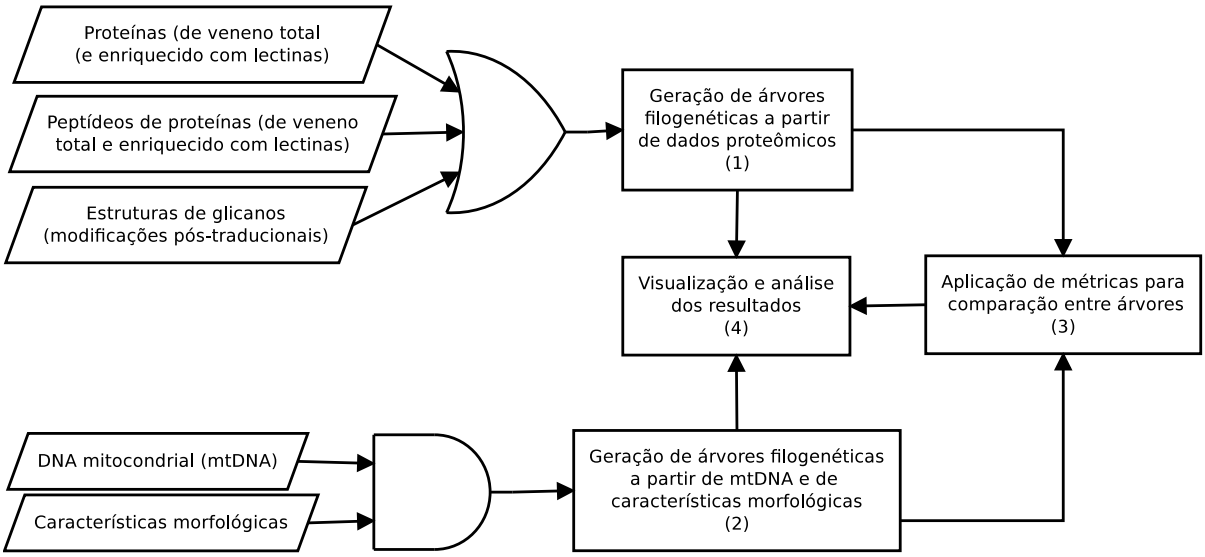

Figura 2: Fluxograma do encadeamento de processos que compõem a metodologia deste projeto proposto. As caixas retangulares numeradas de 1 a 4 representam processos, enquanto que os paralelogramos são as informações biológicas heterogêneas que são utilizadas como entrada do encadeamento. Essas informações podem ser utilizadas em diferentes combinações entre elas, conforme assinalam as duas portas lógicas presentes no fluxograma.

Geração de árvores filogenéticas (processos 1 e 2). Para implementar esses dois processos, adotaremos uma abordagem de inferência Bayesiana [5]. Em uma análise Bayesiana computamos a probabilidade *a posteriori* das árvores filogenéticas. Sejam $B(s)$ uma função que dado o número de espécies s devolve a quantidade de árvores possíveis, τ_i a i-ésima árvore filogenética (dentre todas as possíveis) e $\mathbf{X}$ um conjunto de informações biológicas (e.g., una matriz de ocorrências das estruturas de N-glicanos presentes em cada um dos venenos). A probabilidade posteriori de τ_i dado $\mathbf{X}$ é expressa por:

$$f(\tau_i|\mathbf{X}) = \frac{f(\mathbf{X}|\tau_i)\ f(\tau_i)}{\sum_{j=1}^{B(s)} f(\mathbf{X}|\tau_j)\ f(\tau_j)}, \tag{1}$$

onde a probabilidade *a priori* $f(\tau_i)$ normalmente segue uma uniforme com probabilidade $\frac{1}{B(s)}$. Já a função de verossimilhança $f(X|\tau_i)$ pode ser calculada usando a Lei da Probabilidade Total sobre os parâmetros que definem a árvore, e é dada por:

$$f(\mathbf{X}|\tau_i) = \int_{\upsilon}\int_{\theta} f(X|\tau_i, \upsilon, \theta)\ f(\upsilon, \theta)\ d\upsilon\ d\theta, \tag{2}$$

onde υ e θ são, respectivamente, tamanhos das ramificações e parâmetros de substituição, que têm probabilidade *a priori* $f(\upsilon, \theta)$. No entanto, como a integral da equação 2 não pode ser calculada analiticamente, são empregados métodos de aproximação; um dos mais utilizados é o Monte Carlo via cadeias de Markov (MCMC). A maioria dos métodos MCMC funciona da seguinte maneira:

1. Define aleatoriamente a posição atual no espaço dos parâmetros;
2. Começa o algoritmo na posição atual no espaço dos parâmetros;
3. Propõe uma nova posição no espaço;
4. Aceita ou rejeita a nova posição, utilizando informações *a priori* disponíveis;
5. Se a posição for aceita, então atualize a posição atual e volte para o passo 2;
6. Se a posição for rejeitada, então volte para o passo 2;
7. Após um número determinado de iterações, devolve todas as posições aceitas.

A principal diferença entre os diferentes métodos MCMC está nas técnicas empregadas para escolher novas posições e decidir se ela será aceita ou não. Em qualquer um desses métodos, a amostra obtida pela cadeia de Markov ao término da última iteração é uma aproximação da distribuição *a posteriori*. Um dos métodos MCMC mais relevantes é o algoritmo Metropolis–Hasting, indicado para situações em que o número de combinações de valores para υ e θ é muito grande. Dessa forma, para o desenvolvimento dos processos 1 e 2, utilizaremos a terceira versão do MrBayes [5], um programa que implementa o Metropolis–Hasting e também variantes paralelizáveis do mesmo. Além disso, esse programa tem a vantagem de permitir como entrada informações biológicas heterogêneas, o que vai ao encontro de nossas necessidades. Definido esse conjunto de dados heterogêneos $\mathbf{X}$ e escolhida uma distribuição *a priori* para υ e θ, faremos uma inferência Bayesiana com métodos MCMC, como foi descrito anteriormente. A saída desse programa será composta por arquivos do tipo NEXUS, formato que encapsula uma análise estatística dos parâmetros e a árvore filogenética mais provável.

Aplicação de métricas de comparação (processo 3). Para a comparação quantitativa entre os cladogramas iremos estudar e aplicar métricas tais como a desenvolvida por Lapointe e Legendre [6]. Nessa metodologia, cladogramas, que são equivalentes a árvores aditivas, são comparados entre si através de suas respectivas matrizes de distâncias (tamanhos de caminhos). Essas matrizes são produzidas através da decomposição das árvores aditivas; os componentes dessa decomposição são permutados e somados. As matrizes produzidas, por sua vez, são utilizadas para testar a hipótese nula de que as árvores aditivas que são comparadas entre si não são mais similares do que árvores aleatórias [6].

Visualização e análise dos resultados (processo 4). Para a visualização dos cladogramas utilizaremos uma ferramenta que trabalha com o formato NEXUS; dois exemplos são a MacClade e a Mesquite [7, 8]. Caso venha a ser necessário, também implementaremos um programa em Python para esse fim, utilizando a biblioteca matplotlib. Por fim, para análises estatísticas complementares dos resultados, utilizaremos o programa Tracer [9]; essa ferramenta permite, entre outras coisas, verificar a correlação existente entre os parâmetros utilizados para produzir os cladogramas, devolvendo os resultados em histogramas e gráficos.

4 Plano de trabalho e cronograma de execução

Listamos abaixo as principais atividades previstas neste projeto, cujo cronograma de execução é apresentado no diagrama de Gantt da tabela 1.

Atividade 1: Estudar os artigos que servirão de base para o projeto, o que inclui tanto os trabalhos que descrevem os dados biológicos [2, 3, 4] quanto os que descrevem filogenia por inferência Bayesiana [5, 10];

Atividade 2: Aprender todas as etapas envolvidas na identificação computacional de proteínas no contexto da proteômica baseada em espectrometria de massas. Para esse fim, estudaremos o tutorial de Colinge e Bennett [11];

Atividade 3: Organizar as informações biológicas em um banco de dados relacional, utilizando como gerenciador o MySQL, permitindo assim o que seja feito o relacionamento entre as informações (portas lógicas no fluxograma da figura 2) para gerar as entradas para os processos 1 e 2;

Atividade 4: Desenvolver um programa em Python para gerar instâncias para o MrBayes a partir do banco de dados. Esse programa também deverá fazer o cruzamento entre peptídeos de diferentes venenos, o que exigirá o uso de ferramentas de alinhamento de sequências, tanto dois-a-dois (e.g., BLAST [12]) quanto múltiplas (e.g., Clustal [13]);

Atividade 5: Definir distribuições *a priori* de v e θ para o método MCMC a ser utilizado no MrBayes. Para isso, combinaremos os estudos de distribuições de probabilidades e de inferência Bayesiana que faremos na Atividade 1 com as propriedades das informações biológicas que serão aprendidas nas Atividades 2 e 3;

Atividade 6: Gerar cladogramas com o MrBayes, para diferentes modelos e diferentes conjuntos de dados de entrada, conforme apresentado na figura 2;

Atividade 7: Estudar artigos sobre métricas de comparação entre cladogramas, iniciando pelo artigo de Lapointe e Legendre [6];

Atividade 8: Implementar em Python a métrica escolhida para comparação entre cladogramas; esse programa deve receber como entrada dois ou mais arquivos NEXUS, cada um contendo um cladograma, e devolver medidas de similaridade entre os cladogramas desses arquivos;

Atividade 9: Gerar as comparações e analisar os resultados obtidos, utilizando para esse fim ferramentas de visualização e também programa de análises estatísticas tal como Tracer [9];

Atividade 10: Escrever a monografia do Trabalho de Formatura Supervisionado (TFS) e o Relatório Científico;

Atividade 11: Elaborar pôsteres e apresentá-los em conferências científicas e também na disciplina do TFS.

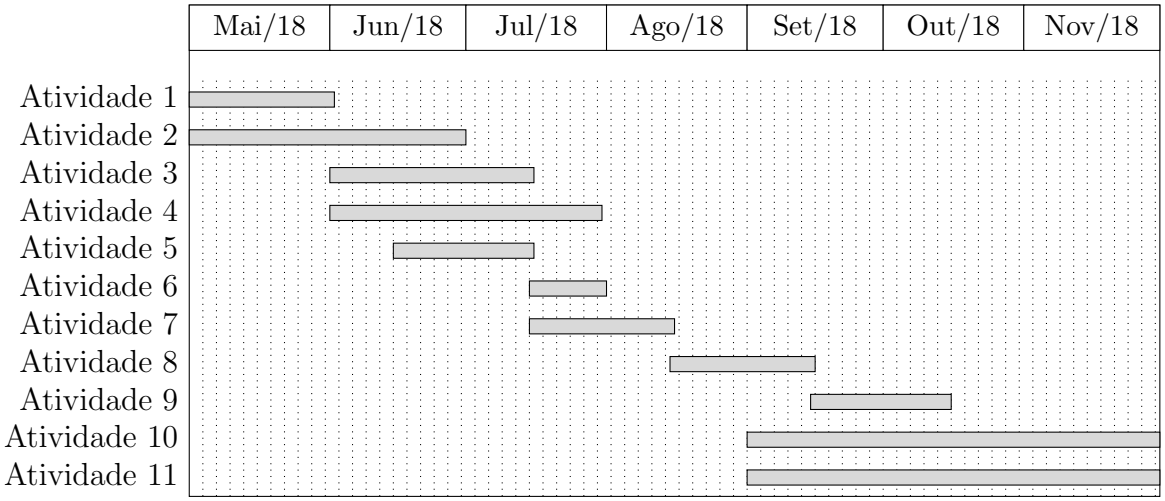

Tabela 1: Cronograma de execução das Atividades 1–10 deste projeto proposto.

5 Forma de análise e disseminação de resultados

A corretude dos resultados computacionais será verificada tanto com testes unitários automatizados quanto através de uma avaliação crítica que será feita pelo aluno e seu orientador. Já a validação experimental dos resultados científicos poderá ser feita no médio prazo, através de resultados que serão obtidos em dois projetos de doutorado sob responsabilidade de Solange M.T. Serrano: em um deles, Carolina Brás Costa produzirá três perfis glicoproteômicos de venenos, dois deles enriquecidos por lectinas com afinidade a ácido siálico e o terceiro enriquecido por uma lectina com afinidade a N-acetilglucosamina; esses novos perfis poderão ser utilizados tanto para verificação dos resultados anteriores num esquema de validação cruzada quanto para a produção de árvores filogenéticas mais precisas. Já no outro projeto,

Débora Andrade-Silva identificará sítios de N- e O-glicosilação das glicoproteínas presentes nos diversos venenos, informação biológica que também poderá ser utilizada para aperfeiçoar os cladogramas.

O aluno divulgará os resultados obtidos com pôsteres em duas conferências que serão realizadas entre o início de outubro e o final de novembro: o X-Meeting e a Reunião Científica Anual (RCA) do Instituto Butantan. Além disso, uma monografia será elaborada sobre as atividades deste projeto, dentro do contexto de MAC499 – Trabalho de Formatura Supervisionado, disciplina do Instituto de Matemática e Estatística da Universidade de São Paulo na qual o aluno está matriculado. Finalmente, os resultados científicos aqui produzidos poderão ser incorporados em futuros manuscritos, que serão elaborados em colaboração com a pesquisadora Solange M.T. Serrano.

Referências

[1] George A Khoury, Richard C Baliban, and Christodoulos A Floudas. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific reports*, 1:90, 2011.

[2] Débora Andrade-Silva, André Zelanis, Eduardo S Kitano, Inácio LM Junqueira-de Azevedo, Marcelo S Reis, Aline S Lopes, and Solange MT Serrano. Proteomic and glycoproteomic profilings reveal that post-translational modifications of toxins contribute to venom phenotype in snakes. *Journal of proteome research*, 15(8):2658–2675, 2016.

[3] Débora Andrade-Silva, David Ashline, Thuy Tran, Aline Lopes, Silvia Cardoso, Marcelo Reis, André Zelanis, Solange Serrano, and Vernon Reinhold. Structures of N-Glycans of *Bothrops* venoms revealed as molecular signatures that contribute to venom phenotype in viperid snakes. *Molecular and cellular proteomics*, 2018. In revision.

[4] Allyson M Fenwick, Ronald L Gutberlet, Jennafer A Evans, and Christopher L Parkinson. Morphological and molecular evidence for phylogeny and classification of South American pitvipers, genera *Bothrops*, *Bothriopsis*, and *Bothrocophias* (Serpentes: Viperidae). *Zoological Journal of the Linnean Society*, 156(3):617–640, 2009.

[5] John P Huelsenbeck and Fredrik Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

[6] Francois-Joseph Lapointe and Pierre Legendre. A statistical framework to test the consensus among additive trees (cladograms). *Systematic Biology*, 41(2):158–171, 1992.

[7] David R Maddison and Wayne P Maddison. MacClade 4: Analysis of phylogeny and character evolution, 2005.

[8] Wayne P Maddison. Mesquite: a modular system for evolutionary analysis. *Evolution*, 62:1103–1118, 2008.

[9] A Rambaut and A Drummond. Tracer: a program for analysing results from Bayesian MCMC programs such as BEAST & MrBayes. *University of Edinburgh, UK*, 2003.

[10] Johan AA Nylander, Fredrik Ronquist, John P Huelsenbeck, and Joséluis Nieves-Aldrey. Bayesian phylogenetic analysis of combined data. *Systematic biology*, 53(1):47–67, 2004.

[11] Jacques Colinge and Keiryn L Bennett. Introduction to computational proteomics. *PLoS computational biology*, 3(7):e114, 2007.

[12] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[13] Mark A Larkin, Gordon Blackshields, NP Brown, R Chenna, Paul A McGettigan, Hamish McWilliam, Franck Valentin, Iain M Wallace, Andreas Wilm, Rodrigo Lopez, et al. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.