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

Monografia apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
aprovação em MAC0499 – Trabalho
de
Formatura Supervisionado

Aluno: Victor Wichmann Raposo

Orientador: Marcelo da Silva Reis

Centro de Toxinas, Resposta-imune e Sinalização Celular
(CeTICS)

Laboratório Especial de Ciclo Celular, Instituto Butantan

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da FAPESP

(processo 18/06682-0)

São Paulo, 23 de dezembro de 2018

Agradecimentos

Agradeço aos pesquisadores e colegas do Instituto Butantan que colaboram no desenvolvimento deste trabalho:

- Dra. Solange M.T. Serrano (Laboratório Especial de Toxinologia Aplicada – LETA)
- Dr. Inácio L.M. Junqueira de Azevedo (LETA)
- Dr. Felipe Grazziotin (Laboratório de Coleções Zoológicas)
- Carolina Brás (LETA, doutoranda do IQ/USP)

Faço também meus agradecimentos ao Prof. Dr. François Joseph Lapointe (Universidade de Montreal, Canadá), que nos prestou assistência com o teste estatístico CADM, e a Olivia Tavares Cesar, que diagramou a imagem na figura 2.1.

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, desenhamos cladogramas gerados a parir de informações dos proteomas, incluindo os peptídeos utilizados na etapa de identificação proteica e peptídeos sequenciados pelo protocolo de novo, e de estruturas de N-glicanos. Para este fim, utilizamos uma abordagem de inferência Bayesiana, empregando métodos de Monte Carlo com cadeias de Markov. A análise dos resultados foi feita com uma métrica de comparação entre árvores, o teste CADM, que permite a quantificação da congruência topológica das novas árvores em relação a uma produzida com dados genômicos. Dessa forma, mostramos que o perfil peptidômico das proteínas de venenos de serpentes *Bothrops* está correlacionado com a sua filogenia, com exceção de uma pequena divergência da espécie *B. neuwiedi*.

Palavras-chave: Espectrometria de massas, Inferência Bayesiana, Venenos de serpentes, Análise filogenética, Glicoproteômica, Monte Carlo via Cadeias de Markov.

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 designed cladograms with the information from proteomes, including their peptides and peptides identified with de novo protocol, and the data from N-glycan structures. To this end, we used a Bayesian inference approach, using Markov chain Monte Carlo methods. The analysis of the results was done with a comparison metric for cladograms, the CADM test, which allowed us to measure the topological congruence of the new trees in respect to one produced with genomic data. Therefore, we showed that the peptidomic profile of proteins from venoms of *Bothrops* snakes is correlated to their phylogeny, except for a slight divergengence of the species *B. neuwiedi*.

Keywords: Mass spectrometry, Bayesian inference, Snake venom, Phylogenetic analysis, Glycoproteomics, Markov-Chain Monte Carlo.

Sumário

Capítulo 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 genes de DNA mitocondrial (mtDNA) e características morfológicas (Figura 1.1). 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 dos cladogramas obtidos. 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 genômicas e/ou características morfológicas [4], sem utilizar métricas para fazer uma me-

dida 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*).

1.1 Objetivos

Este projeto tem como objetivo geral montar um encadeamento (*pipeline*) de processos para desenho, comparação e visualização de árvores filoproteômicas a partir de informações biológicas heterogêneas.

Mais especificamente, esse trabalho visa aplicar o encadeamento desenvolvido para testar a hipótese de que o perfil proteômico e glicoproteômico dos venenos de serpentes do gênero *Bothrops* está fortemente correlacionado com a filogenia observada em análises que empregam dados genômicos. Para isso, pretendemos mitigar os vieses de nossos dados e utilizar um teste estatístico para ter uma métrica de comparação de cladogramas.

1.2 Organização do Trabalho

O restante desta monografia está organizada da seguinte maneira: no capítulo 2 (Revisão Bibliográfica) introduzimos conceitos biológicos necessários para o melhor entendimento deste trabalho. Fazemos também uma revisão da literatura, mais precisamente sobre como os dados que utilizamos são obtidos, o conceito matemático por trás da análise dos dados e o algoritmo de comparação de cladogramas.

No capítulo 3 (Materiais e Métodos) explicamos mais precisamente como armazenamos e tratamos os dados, assim como definimos os modelos e métodos utilizados. Informações complementares sobre os métodos, mais precisamente as documentações dos programas mais importantes, são disponibilizadas nos apêndices A e B.

Em seguida, no capítulo 4 (Resultados) apresentamos os resultados mais relevantes obtidos ao longo deste trabalho, tanto do ponto de vista tecnológico quanto científico.

Finalmente, no capítulo 5 (Conclusões) recapitulamos o conteúdo apresentado nesta monografia, destacando as principais contribuições e indicando as principais hipóteses biológicas geradas com a metologia desenvolvida neste trabalho. Por fim, listamos algumas possibilidades de continuidade nesta linha de pesquisa.

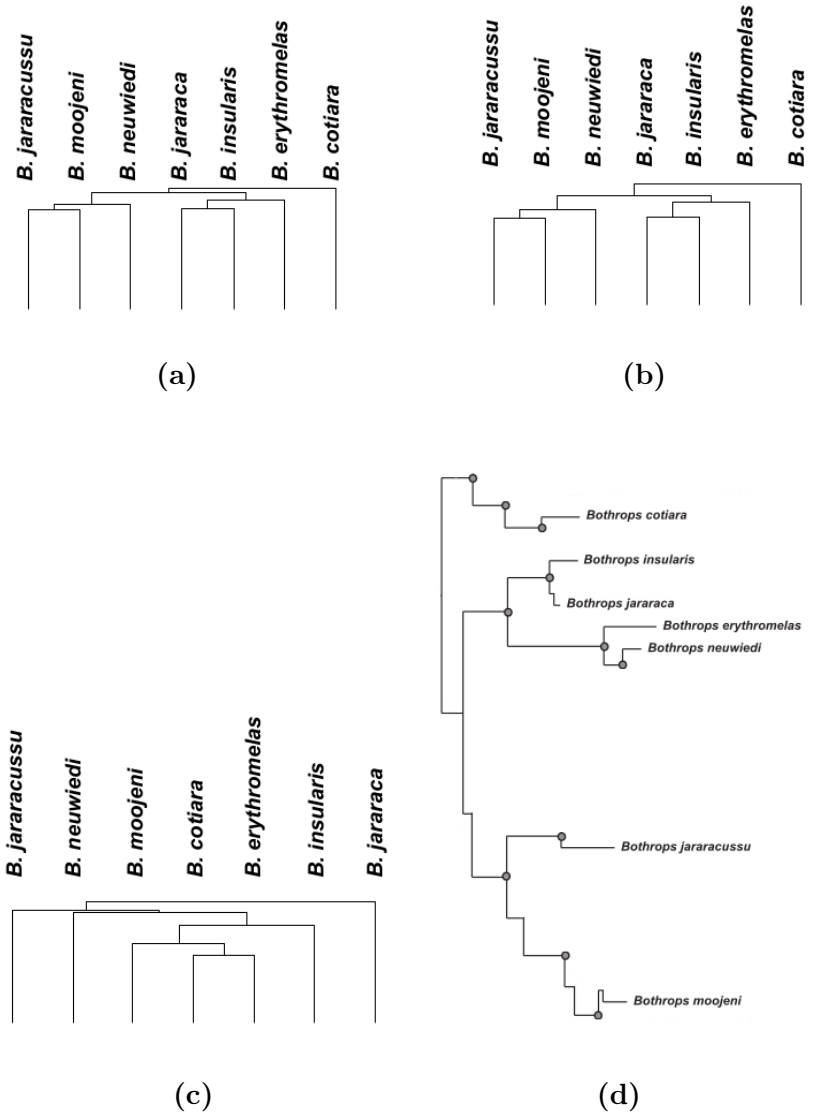


Figura 1.1: *Cladogramas filogenéticos de sete espécies do gênero* Bothrops*, obtidos utilizando diferentes marcadores moleculares. São apresentados cladogramas construídos a partir de proteoma total de veneno (Fig. 1.1a), de glicoproteínas detectadas através de proteomas total e baseados em protocolos de enriquecimento por afinidade a lectinas (Fig. 1.1b), e de não-glicoproteínas detectadas nos mesmos ensaios de proteomas anteriores (Fig. 1.1c). Já na Fig. 1.1d é mostrada uma sub-árvore de um cladograma obtido através do uso de genes de mtDNA e de características morfológicas. Observe que, à exceção de* B. neuwiedi*, os cladogramas das Figs. 1.1a e 1.1b apresentam a mesma hierarquia da subárvore da Fig. 1.1d, enquanto que o da Fig. 1.1c é dissimilar em comparação aos demais. As Figs. 1.1a–1.1c foram extraídas de Andrade-Silva et al. [2], enquanto que a Fig. 1.1d foi adaptada de Fenwick et al. [4].*

Capítulo 2

Revisão Bibliográfica

Neste capítulo faremos uma revisão bibliográfica de conceitos fundamentais para um bom entendimento deste trabalho. Iniciaremos com os principais conceitos de biologia presentes neste projeto. Descreveremos também cladogramas e a espectrometria de massas, estratégia analítica utilizada para a medição de proteomas. Por fim, apresentaremos os métodos estatísticos e computacionais que empregamos em nossa metodologia.

2.1 DNA, RNA, proteína e Dogma Central

As informações genéticas de um organismo, que foram herdadas de seus ancestrais e posteriormente serão passadas para seus descendentes, são armazenadas na molécula chamada DNA (*DeoxyriboNucleic Acid*). O DNA é uma molécula longa composta por nucleotídeos, subdividida em genes, que são segmentos que contêm instruções para a produção de proteínas. Proteínas são polipeptídeos, ou seja, uma cadeia de peptídeos, que, por sua vez, são estruturas formadas por aminoácidos.

O conjunto de moléculas de DNA (i.e., o material genético) fica num compartimento interno celular denominado núcleo, portanto isolado do restante da célula. Então, para as instruções presentes nos genes chegarem na estrutura responsável pela produção de proteínas, chamada de ribossomo, é necessária uma molécula intermediária para carregar a "mensagem"; tal molécula é conhecida como RNA (*RiboNucleic Acid*).

O transporte da mensagem é feito da seguinte maneira: primeiramente, as informações do DNA são transcritas para uma molécula de RNA, a RNA mensageira (mRNA), que por sua vez é transportada até o ribossomo. Dentro do ribossomo, a mRNA será traduzida: trata-se de um processo que transforma cada três nucleotídeos (i.e., um códon) em um aminoácido da proteína. Todo esse processo, conhecido como o caso geral do Dogma Central da Biologia Molecular, é ilustrado na figura 2.1.

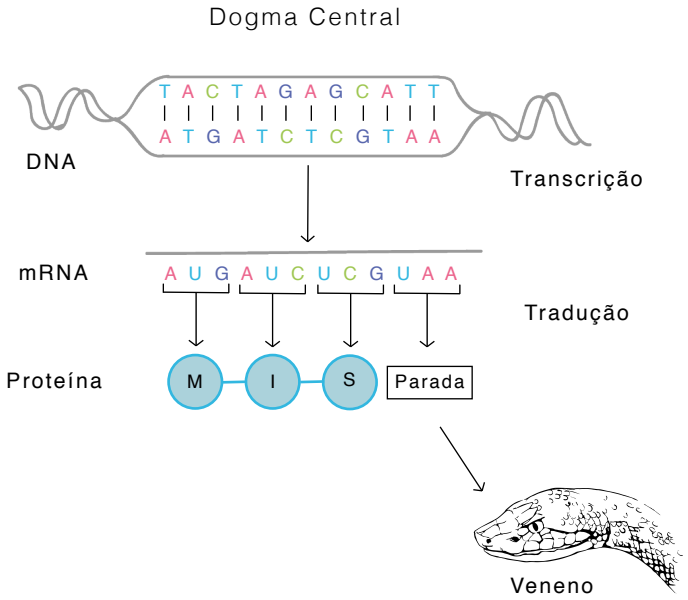


Figura 2.1: *Ilustração do caso geral do Dogma Central da Biologia Molecular, que modela a produção de proteínas de venenos de serpentes.*

2.1.1 DNA mitocondrial

As células animais possuem uma organela chamada de mitocôndria, que é extremamente importante para respiração celular. Além disso, ela é uma estrutura que possui o próprio material genético e esse material é passado diretamente da mãe para o filho. Por esse fato, o DNA mitocondrial é muito estável para analisar a evolução dos organismos.

Dessa forma, fragmentos de genes mitocontriais são utilizados como marcadores genéticos, isto é, como características moleculares que diferenciam indivíduos e são facilmente detectáveis. Dois dos marcadores comumente utilizados, e que também são empregados neste trabalho, são fragmentos dos genes NADH dehidrogenase – subunidade 4 (ND4) e citocromo b (cyt b).

2.1.2 Proteoma de veneno de serpentes

Venenos de serpentes são misturas proteicas complexas, utilizadas para a defesa contra predadores e para a caça de presas. Tais proteínas são sintetizadas em células endócrinas presentes nas glândulas de veneno. O veneno produzido fica armazenado nessa glândula até que o mesmo seja injetado em um alvo, através de canais existente em suas presas.

Em um proteoma de veneno de serpente do genêro *Bothrops*, estão presentes mais de uma centena de diferentes proteínas, com funções biológicas distintas: por exemplo, metaloproteinases, serino-proteinases e fosfo-lipases [2]. Conforme já mencionamos na introdução, essas proteínas podem sofrer diferentes níveis de N- e de O- glicosilações; tais modificações pós-traducionais (i.e., após a produção da sequência protéica através das etapas do Dogma

Central) podem contribuir para a estabilidade protéica e também com a atividade catalítica responsável pela ação fisiológica do veneno na presa ou alvo. Para os venenos de *Bothrops*, já foram reportados mais de 50 tipos diferentes de N-glicosilação, com variações inter-espécies que têm alguma correlação com a filogenia dessas serpentes [3].

2.2 Cladogramas

Cladogramas, ou árvores filogenéticas/filoproteômicas, são diagramas que representam relações evolutivas entre organismos. É uma árvore com a melhor hipótese sobre como esse conjunto de organismos evoluiu de um ancestral em comum. Em um cladograma, a raiz da árvore representa o ancestral comum a todos os organismos, enquanto que os nós internos denotam um evento que causou uma divergência, gerando grupos diferentes de descendentes. Cada aresta representa uma série de ancestrais que foram se modificando até chegar no extremo da aresta. Um exemplo de cladograma é apresentado na figura 2.2.

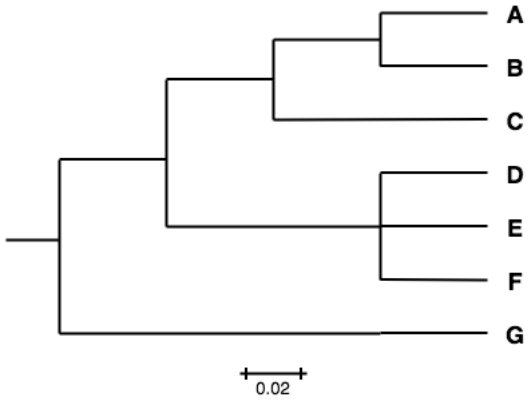


Figura 2.2: *Exemplo de árvore filogenética de sete organismos (A–G). A escala abaixo da árvore mostra a proporção de mudanças que os organismos sofrem em relação ao tamanho dos ramos; neste caso, o tamanho de segmento representa 0.02 substituições por campo.*

Cladogramas podem ser interpretados da seguinte maneira: dois organismos (folhas da árvore) são mais relacionados se seu menor ancestral comum é recente e menos caso contrário. Por exemplo, na figura 2.2 temos que A e B são mais relacionados do que A e G.

O cladograma pode conter uma politomia, que é quando um nó da árvore possui mais de dois filhos, como acontece no ancestral de D, E e F na figura 2.2. Isso pode significar que não temos informações suficientes para descobrir a relação exata entre aqueles organismos.

2.3 Espectrometria de Massas (EM)

Espectrometria de massas (em inglês, *Mass Spectrometry – MS*) é uma técnica analítica para medição da massa de uma amostra. A máquina que realiza essa análise, chamada de

espectrômetro de massa, é composta por três componentes independentes: uma fonte de ionização, uma célula de fragmentação e um analisador de massa.

Inicialmente cada molécula precisa ser ionizada, ou seja, deve ser carregada eletricamente, isto ocorre na fonte de ionização do espectrômetro. Portanto, esse processo não mede a massa da amostra diretamente, mas sim a taxa massa/carga da amostra.

Os espectrômetros mais atuais possuem células de fragmentação, um componente que utiliza um gás para quebrar as moléculas por dissociação induzida por colisão (CID). Tal fragmentação permite quebrar a amostra em fragmentos ainda menores, para facilitar a posterior identificação computacional de sua composição.

Os analisadores de massa mais comuns em laboratórios de proteômica são armadilha de íons (IT), nos quais os íons ejetados são detectados pela variação da frequência causada; e tempo-de-voo (TOF), aonde o tempo necessário para atravessar uma região sem campo elétrico é observada e correlacionada com a massa do íon.

2.3.1 Proteômica baseada em EM

Por meio da espectrometria de massas é possível identificar proteínas de uma amostra. Uma técnica bastante comum para isso consiste em primeiro digerir as proteínas com uma enzima chamada tripsina, obtendo assim um conjunto de peptídeos (moléculas que são segmentos de uma proteína). Em seguida é feita uma análise dos peptídeos por MS e os resultados permitem a identificação dos peptídeos e, por conseguinte, a identificação das proteínas.

Existem diferentes tipos de análises que fazem a identificação com essa técnica. A análise que estudamos é a espectrometria de massa tipo *Tandem* (MS/MS). Inicialmente neste procedimento, para separar os peptídeos da amostra é feita uma cromatografia líquida, uma técnica de separação de componentes de uma mistura entre uma corrente de fluído em movimento. Em seguida quebra-se os peptídeos em moléculas menores (fragmentos); é possível fazer isso de várias formas: uma delas é colidir as moléculas com um gás inerte. Então obtém-se a massa das novas moléculas e, como a fragmentação de peptídeos segue algumas regras, é possível determinar a massa do peptídeo a partir da massa dos fragmentos.

Os resultados do ensaio MS/MS (espectro experimental), que são espectros das massas dos fragmentos do peptídeo, são comparados com dados teóricos encontrados em um banco de dados. Utilizando-se de uma função de pontuação, que calcula a similaridade entre as informações, encontra-se o peptídeo mais provável. A identificação das proteínas a partir da pontuação dos peptídeos é uma questão em aberto e há vários problemas associados ao compartilhamento de peptídeos entre diferentes proteínas. Logo, há várias opções para fazer isso; uma abordagem padrão é somar a pontuação dos peptídeos e obter uma pontuação para as proteínas, assim considerando as proteínas mais prováveis aquelas com pontuações maiores. A escolha do banco de dados de proteínas que será utilizado para a identificação dos dados é extremamente importante, como vamos mostrar na seção 2.3.2.

Para casos em que o banco de dados é inacessível ou inapropriado é possível identificar os peptídeos diretamente do espectro obtido no ensaio MS/MS; esse tipo de sequenciamento é chamado de de novo. É um processo complicado, pois é inviável testar todas as possibilidades de combinação de aminoácidos, um método bem estabelecido envolve a construção de um grafo do espectro (mais detalhes em [5]).

Alguns aminoácidos da proteína, podem sofrer modificações pós-traducionais (PTMs), ou seja, mudanças na cadeia proteica como a adição de algum grupo ou remoção de outros. Essas mudanças causam alterações na massa da proteína, por exemplo a glicosilação aumenta a massa, pois há a adição de um glicano na proteína, enquanto desfosforilação diminui a massa, ao remover um fosfato de um dado aminoácido. Portanto, essas modificações devem ser levadas em conta no momento de fazer a espectrometria de massa.

É possível, também, identificar as PTMs por meio dessa análise; por exemplo, a identificação das estruturas de N-glicanos, que são glicanos que se ligam ao átomo de Nitrogênio da amina do aminoácido asparagina. Para a identificação de N-glicanos em Andrade et al [3], foram utilizados protocolos específicos para a remoção e isolamento dessas estruturas de suas respectivas proteínas e posterior resolução das estruturas por meio de ensaios de MS.

Para maiores informações sobre o processamento computacional de dados proteômicos baseados em espectrometria de massas, recomendamos a leitura do ótimo artigo escrito por Colinge e Bennett [5].

2.3.2 Limitações da proteômica baseada em EM

Como mencionado na seção anterior, a escolha do banco de dados para os ensaios de proteômica baseada em espectrometria de massas é extremamente importante. Uma das razões é que o banco de dados, ao representar boa parte ou mesmo todas as proteínas presentes na amostra, restringe dramaticamente o espaço de possibilidades de sequências de aminoácidos que precisam ser consideradas durante a identificação computacional de peptídeos. Quando se estuda organismos que não possuem o genoma completo sequenciado (ou seja, organismos não-modelos), tais banco de dados podem subrepresentar os tipos de proteínas presentes na amostra.

O problema da subrepresentação de proteínas no banco de dados pode ser crítico, em particular quando o(a) pesquisador(a) está interessado(a) em uma análise comparativa inter-espécies e diferentes espécies têm diferentes níveis de cobertura de identificação de suas proteínas. Por exemplo, no caso das serpentes do gênero *Bothrops*, vemos nas tabelas 2.1 e 2.2 que as informações em relação aos dados das serpentes que estudamos evidencia um claro viés em favor de *B. jararaca*.

Em todos os bancos de dados em que buscamos e para todos os tipos de dados, notamos que há muito mais informações referentes a *B. jararaca* em comparação com as outras espécies. Mais precisamente, em média as outras espécies possuem apenas $16,46\%$ da quantidade de sequências nucleotídicas, $9,74\%$ dos transcritos (ESTs), $16,58\%$ das sequências proteicas

Espécie	Nucleotídeos	ESTs	% Nucleotídeos	% ESTs
B. jararaca	244	1.158	100	100
B. jararacussu	43	0	17,62	0
B. cotiara	6	0	2,45	0
B. insularis	24	677	9,83	58,46
B. neuwiedi	98	0	40,16	0
B. erythromelas	32	0	13,11	0
B. moojeni	67	0	27,45	0

Tabela 2.1: *Quantidade de entradas de sequências nucleotídicas e de ESTs no banco de dados GenBank. A coluna de % tem a proporção de informação da espécie comparado com a quantidade encontrada da B. jararaca. As informações foram obtidas em outubro de 2018.*

Espécie	GenBank	%	UniProt	%
B. jararaca	348	100	125	100
B. jararacussu	102	29,31	35	28,00
B. cotiara	20	5,74	19	15,20
B. insularis	46	13,21	21	16,80
B. neuwiedi	105	30,17	46	36,80
B. erythromelas	36	10,34	21	16,80
B. moojeni	95	27,29	63	50,40

Tabela 2.2: *Quantidade de sequências proteicas encontradas em dois bancos de dados, GenBank e UniProt. A coluna de % tem a proporção de informação da espécie comparado com a quantidade encontrada da B. jararaca referente ao banco da coluna à esquerda. As informações foram obtidas em outubro de 2018.*

encontradas no GenBank e $27,33\%$ das sequências proteicas encontradas no UniProt, das quantidades observadas da *B. jararaca* (para informações específicas de cada espécie verificar as tabelas 2.1 e 2.2).

A super-representação da *B. jararaca* nos bancos de dados gera um viés sobre os dados que utilizamos. Existem formas de mitigar tais vieses: por exemplo, podemos usar diretamente os peptídeos identificados pelo MS/MS; uma segunda alternativa seria utilizar os resultados de sequenciamento de novo de peptídeos, já que essa abordagem não usa o banco de dados. Essas ideias serão exploradas nos próximos capítulos.

2.4 Inferência Bayesiana e MCMC

Para a geração de cladogramas com os dados coletados, adotaremos uma abordagem de inferência Bayesiana [6]. Em uma análise Bayesiana computamos a probabilidade *a posteriori* das árvores. Sejam $B(s)$ uma função que dado o número de espécies s devolve a quantidade de árvores possíveis, τ_i a i-ésima árvore (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:

$$f(\tau_i|\mathbf{X}) = \frac{f(\mathbf{X}|\tau_i)\ f(\tau_i)}{\sum_{j=1}^{B(s)} f(\mathbf{X}|\tau_j)\ f(\tau_j)}, \tag{2.1}$$

onde a probabilidade *a priori* $f(\tau_i)$ normalmente segue uma distribuição 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:

$$f(\mathbf{X}|\tau_i) = \int_{\upsilon}\int_{\theta} f(X|\tau_i, \upsilon, \theta)\ f(\upsilon, \theta)\ d\upsilon\ d\theta, \tag{2.2}$$

onde θ e υ são, respectivamente, parâmetros de substituição e de definição da forma da árvore (e.g., tamanho dos ramos); esses parâmetros têm probabilidade *a priori* $f(\upsilon, \theta)$. No entanto, a integral da equação 2.2 não pode ser computada analiticamente, já que, potencialmente, ela é calculada em um espaço de parâmetros cuja dimensão é muito alta. Logo, são empregados métodos de aproximação para seu cálculo; um dos mais utilizados é o Monte Carlo via cadeias de Markov (MCMC). A maioria dos métodos MCMC funciona da seguinte maneira:

1. Defina aleatoriamente a posição atual no espaço dos parâmetros;
2. Comece o algoritmo na posição atual no espaço dos parâmetros;
3. Proponha uma nova posição no espaço;
4. Aceite ou rejeite 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, devolva 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*, como mostrado pela figura 2.3. Note que quanto maior o número de amostras, ou seja, iterações, mais próxima é a aproximação.

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.

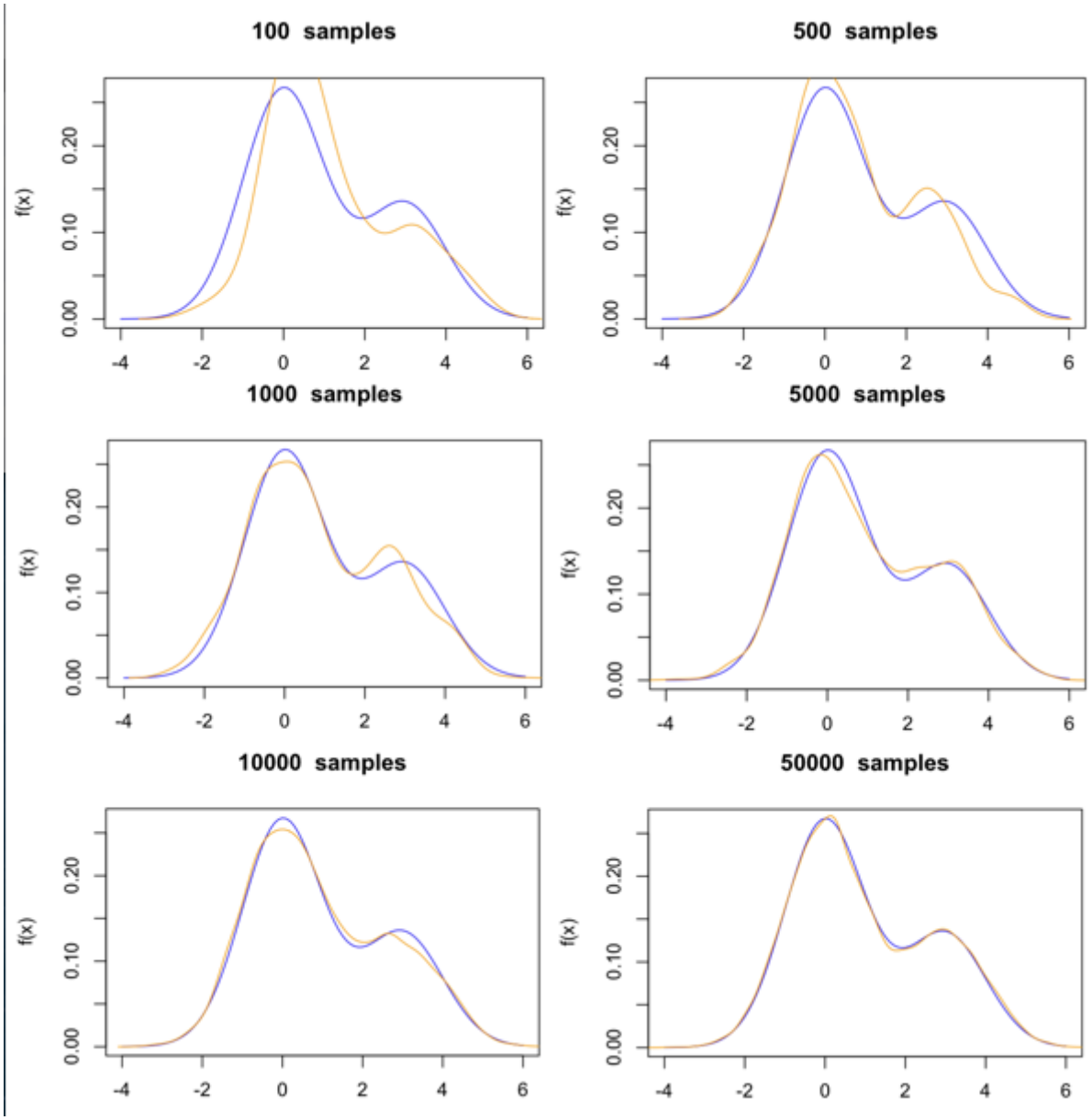


Figura 2.3: *Convergência do algoritmo Metropolis-Hastings. Conforme aumenta-se o número de amostras (*samples*) utilizadas pelo algoritmo, o resultado do método MCMC, representado pelo linha laranja, tende a se aproximar da distribuição real, indicada em azul. Figura pública sob a licença GNU Free Documentation License.*

2.5 teste CADM

O teste CADM, introduzido pelo artigo Legendre e Lapointe (2004) [7], é um teste para estimar a congruência entre matrizes de distância. Ao aplicar esse teste em matrizes de distância provenientes de cladogramas podemos pensar que *incongruência* se refere a árvores com diferentes topologias e/ou comprimento de ramos muito diferentes, o que sugere histórias evolutivas distintas. De forma contrária, entendemos que *congruência* se refere a duas ou mais árvores com uma história evolutiva idêntica.

Este teste verifica a hipótese nula de que todas as árvores são incongruentes, devol-

vendo um nível de congruência entre 0 e 1. Caso a hipótese nula seja rejeitada, é possível fazer um teste *a posteirori* para identificar quais matrizes são congruentes ou não. Dadas as matrizes de distância normalizadas, o procedimento do método CADM é descrito abaixo. O teste utiliza-se de alguns coeficientes estatísticos que serão descritos no final da seção 2.5.1.

1. Transforme a diagonal superior (ou inferior) de cada matriz em um vetor e adicione em uma linha de uma tabela.
2. Construa uma relação de ordem com os valores de cada linha dessa tabela.
3. Compute W = Coeficiente de Kendall (equação 2.4) de concordância entre as matrizes após os procedimentos 1 e 2. Transforme W na estatística χ^2 de Friedman (equação 2.6) que será usada como referência (χ^2_{ref}) para testes.
4. Permute as matrizes de distância e compute um χ^{2*} sobre a permutação.
   (a) para o teste global de congruência todas as matrizes são permutadas aleatoriamente e independentemente.
   (b) Em comparações *a posteirori* apenas uma matriz é permutada por vez. Isto é repetido para todas as matrizes.
5. Repita o passo 4 um grande número de vezes para estimar a distribuição de χ^2. Adicione o valor de referência χ^2_{ref} na distribuição
6. Calcule a probabilidade da hipótese nula ser válida (p-value) como a proporção dos valores de χ^{2*} que são maiores ou iguais que χ^2_{ref}.

   O teste indicará que o conjunto contém matrizes congruentes se χ^2_{ref} é maior ou igual que a maioria (digamos 95% para $\alpha = 0,05$) dos χ^{2*}. Caso a hipótese nula seja rejeitada, testes *a posteirori* podem determinar quais matrizes são congruentes.

O teste CADM é vantajoso pois permite a comparação simultânea de várias matrizes de distâncias e, no caso de análise filogenética, permite verificar a congruência tanto filogenética quanto topológica, ao fixar o tamanho dos ramos da árvore como 1. Além disso, em Campbell et al. (2011) [8] foi demonstrado que esse teste tem um erro tipo-1 adequado para análises filogenéticas.

2.5.1 Coeficientes estatísticos

Apresentaremos agora as definições do coeficiente de Kendall e da estatística de Friedmann; ambos possuem uma relação próxima, uma vez que o primeiro é uma normalização do segundo. A seguir, vamos formular exemplos das hipóteses nulas de cada coeficiente para observar as diferenças entre eles.

Considere p juízes (linhas na tabela) julgando n atletas (colunas) em uma competição. Temos as seguintes hipóteses nulas (H_0):

- H_0 de Friedmann: Os n objetos (atletas) são retirados da mesma população;
- H_0 de Kendall: Os p juízes produzem classificações independentes dos objetos.

No teste CADM, os juízes são as matrizes de distâncias e os atletas são os pares de objetos entre quais as distâncias são calculadas. Dada a soma das ordens R_j de cada coluna e a média $\bar{R}$ de todos R_j, calculamos a variância:

$$S = \sum_{j=1}^{n} (R_j - \bar{R})^2. \tag{2.3}$$

A partir disso computamos o coeficiente de Kendall:

$$W = \frac{12S}{p^2(n^3 - n) - pT}, \tag{2.4}$$

onde T é um fator de correção para ordens iguais dado pela seguinte formula:

$$T = \sum_{k=1}^{m} (t_k^3 - t_k), \tag{2.5}$$

onde t_k é o número de ordens iguais em cada k dos m grupos de empate.

Portanto, o coeficiente de Kendall é simplesmente a variância das somas das colunas dividido pelo maior valor possível que a variância pode atingir.

Por fim, o χ^2 de Friedmann é obtido a partir de W pela seguinte fórmula:

$$\chi^2 = p(n-1)W. \tag{2.6}$$

Capítulo 3

Materiais e Métodos

Neste capítulo descreveremos todas as informações (materiais) utilizadas ao longo deste trabalho, incluindo a organização das mesmas em um banco de dados relacional. Faremos também uma apresentação dos métodos empregados, incluindo ferramentas, programas e scripts desenvolvidos.

A figura Fig. 3.1 apresenta um diagrama que representa o encadeamento dos processos, desde a obtenção de dados até a geração e comparação de cladogramas construídos a partir desses dados. Cada etapa desse encadeamento será descrita nas próximas seções.

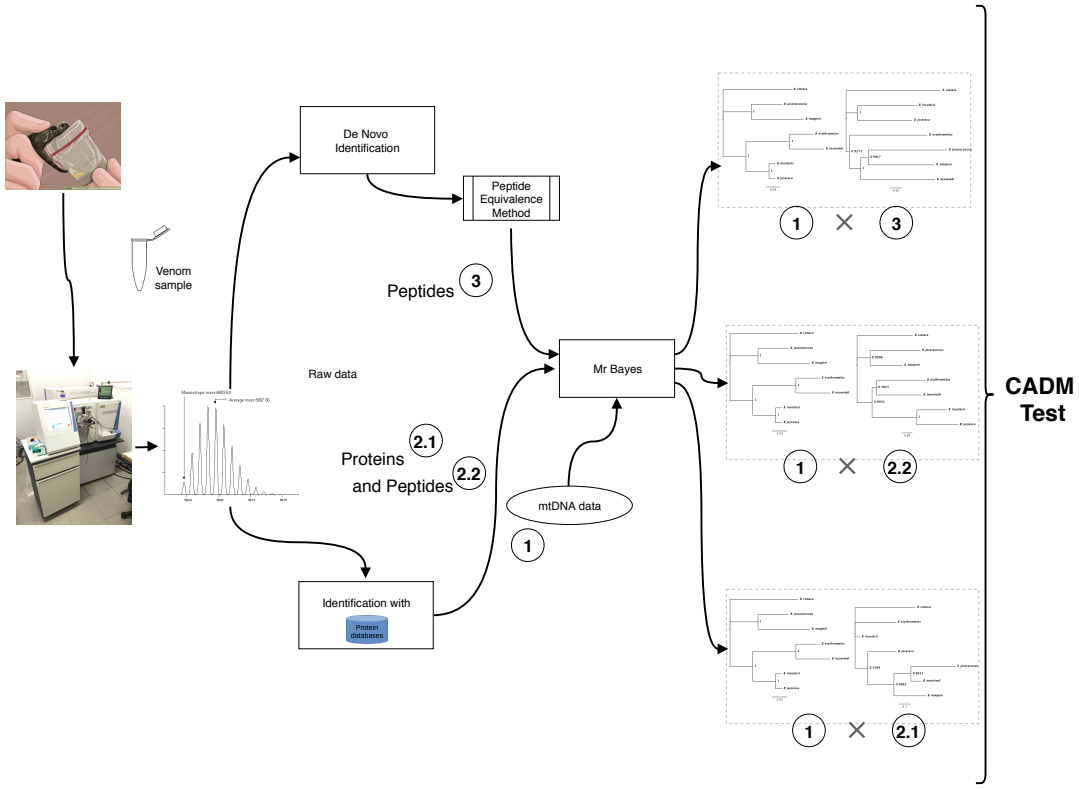


Figura 3.1: *Encadeamento do procedimento desde obtenção dos dados até a geração e comparação de cladogramas gerados com diferentes dados.*

3.1 Organização de informações em banco de dados

Para armazenar e organizar as informações dos venenos de serpentes estudados, incluindo informações proteômicas e de estruturas de glicanos, construímos um banco de dados relacional utilizando o gerenciador PostgresSQL. O banco de dados foi desenhado conforme o Modelo Entidade-Relacionamento (MER) apresentado na figura 3.2.

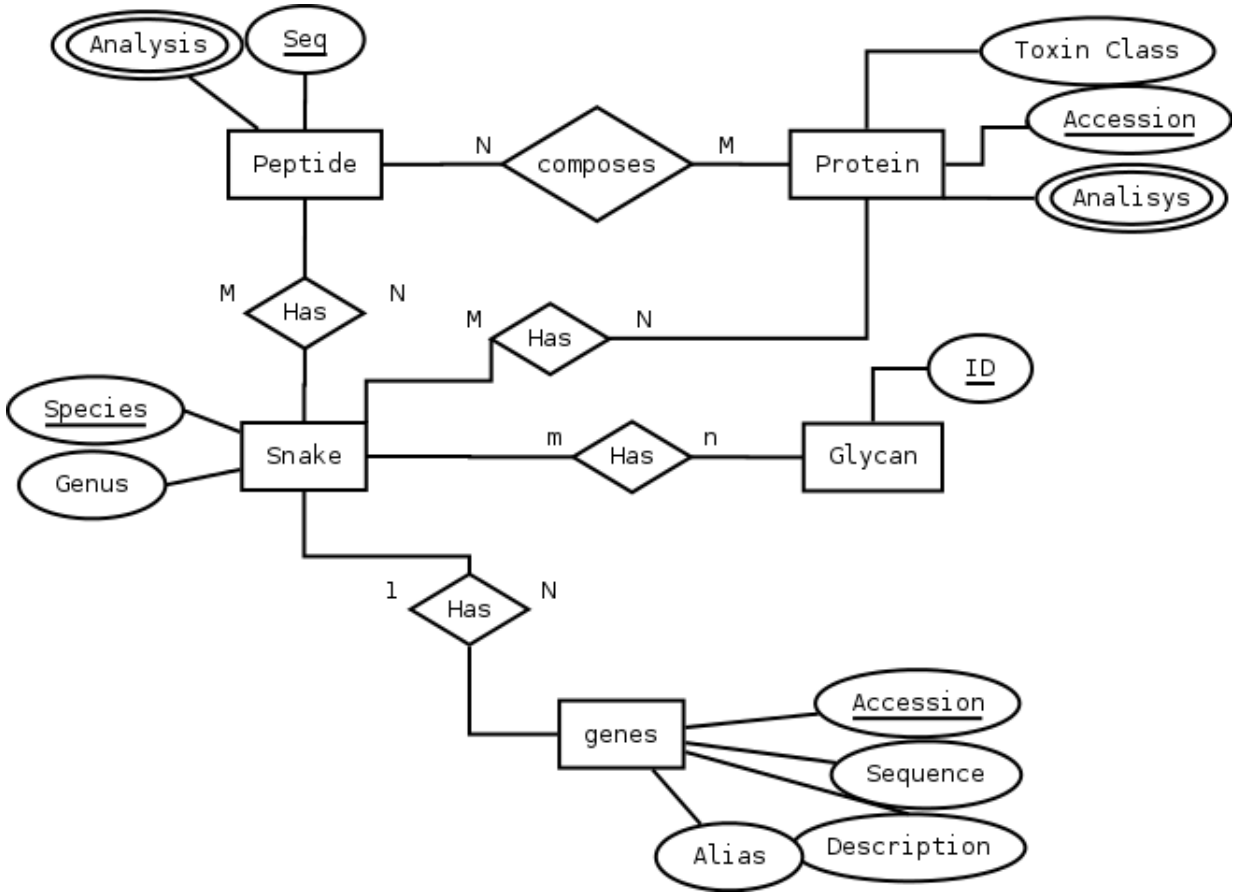


Figura 3.2: *Diagrama MER do banco de dados relacional no qual foram armazenadas as informações biológicas utilizadas neste trabalho.*

Populamos o banco com uma variedade de dados provenientes de fontes diferentes, como informações genéticas, proteicas, peptídicas e de N-glicanos. Manipulamos e inserimos os dados com programas descritos na sessão 3.3.

No artigo de Fenwick et al. [4] foi construída uma filogenia de 43 serpentes, dentre elas 31 eram serpentes do gênero *Bothrops*. Foram utilizados 85 dados morfológicos, as sequências dos genes ribossomais 12S e 16S, e as sequências dos genes mitocondriais NADH dehidrogenase subunidade 4 (ND4) e do cytochrome b (cyt b). A partir dos números de acesso fornecidos no artigo (tabela 3.1) buscamos as sequências no GenBank, obtendo as informações genômicas das 7 serpentes estudadas. O GenBank, mantido pelo National Center for Biotechnology Information (NCBI), é um banco de dados público de anotações de sequências de nucleotídeos (tanto DNA quanto RNA) e suas traduções de proteínas.

Adquirimos os dados proteicos e de peptídeos a partir do artigo de Andrade-Silva et

al. [2], no qual os autores caracterizaram o proteoma dos venenos das mesmas sete espécies de serpentes estudadas aqui. Nesse trabalho, foram identificadas, a partir de protômica baseada em EM, as proteínas dos venenos a partir de um protocolo de proteoma total e por meio de protocolos de enriquecimento por lectinas, que são proteínas que tem afinidade (se ligam a) a carboidratos. Mais precisamente, foram feitos três protocolos diferentes utilizando as lectinas: Concanavalina A (ConA), aglutinina de germe de trigo (WGA) e aglutinina de amendoim (PNA).

As proteínas e peptídeos do artigo [2] foram identificados com o auxílio de banco de dados proteicos, como explicado na seção 2.3. Então, para obter uma identificação de peptídeos pela técnica de novo, nossos colaboradores Dra. Solange Serrano e Carolina Brás utilizaram os dados brutos do trabalho anterior [2] no programa Peaks (versão 8.5). Dessa forma, obtemos listas de peptídeos identificados sem o auxílio de banco de dados, evitando assim o viés descrito no capítulo anterior.

As informações sobre N-glicanos presentes nas proteínas dos venenos foram obtidas do artigo de Andrade-Silva et al. [3], no qual foi reportada a identificação das diferentes estruturas de N-glicanos presentes nos venenos das sete serpentes estudadas neste trabalho.

Espécie	12S	16S	cytb	ND4
B. jararaca	EU867254.1	EU867266.1	EU867278.1	EU867290.1
B. jararacussu	AY223661.1	AY223674.1	AY223602.1	AY223643.1
B. cotiara	AF057217.1	AF057264.1	AY223597.1	AY223640.1
B. insularis	AF057216.1	AF057263.1	AY223596.1	AF188705.1
B. neuwiedi	EU867260.1	JQ627282.1	AF292586.1	AF292624.1
B. erythromelas	AF057219.1	AF057266.1	AY223600.1	AF292626.1
B. moojeni	EU867256.1	EU867268.1	EU867280.1	EU867292.1

Tabela 3.1: *Número de acesso das sequências de DNA mitocontrial utilizadas neste trabalho.*

3.2 Ferramenta de inferência Bayesiana

Para gerar cladogramas por inferência Bayesiana a partir de dados genômicos ou proteômicos, utilizamos a terceira versão do MrBayes [6]. Esse programa recebe como entrada um arquivo tipo NEXUS contendo informações biológicas, que podem ser heterogêneas, ou seja, no mesmo arquivo podem haver tanto sequências de nucleotídeos quanto dados discretos. Após escolhida distribuições *a priori* e os modelos de substituição, é feita uma inferência Bayesiana com métodos MCMC, como foi descrito anteriormente na seção 2.4. O MrBayes implementa o método do Metropolis–Hasting e também variantes paralelizáveis do mesmo. A saída desse programa é composta por arquivos do tipo NEXUS que encapsulam uma análise estatística dos parâmetros e também a árvore mais provável.

Cada análise foi feita duas vezes, começando com árvores aleatórias, por pelo menos 2×10^6 gerações (iterações), amostrando a cada 100 gerações e, de forma conservadora,

descartando o primeiro quarto das iterações como *burn-in*. Além disso, cada conjunto de informação foi tratada como uma partição com taxas e parâmetros independentes; as exceções foram os dados dos genes mitocondriais, ND4 e cytb, que foram divididos em três partições diferentes, uma para cada posição dos códons.

Para as partições com dados discretos como o de proteínas, glicanos e peptídeos, utilizamos o modelo de substituição padrão (MkModel) de Lewis (2001) [9], combinado com uma distribuição gama para as taxas de variação de características. Enquanto isso, para partições com dados genômicos o processo MCMC amostra sobre todos os modelos de substituição reversíveis no tempo combinado com uma uma distribuição gama com uma proporção de campos invariáveis para as taxas de variação de características.

Utilizamos uma distribuição uniforme *a priori* para a topologia, ou seja, todas as árvores possuem a mesma probabilidade de ocorrência. Para a distribuição *a priori* da árvore usamos a combinação de uma distribuição $Gama(1, 1)$ para o tamanho da árvore e uma distribuição $Dirichlet(1, 1)$ para os tamanhos do ramo. Enquanto a distribuição *a priori* usada para frequência dos nucleotídios foi uma $Dirichlet(1, 1, 1, 1)$. Todas essas distribuições não são muito informativas; dessa maneira, a distribuição *a posteriori* será mais baseada nos dados que temos.

Para escrever os arquivos NEXUS de entrada do MrBayes com os dados provenientes do nosso banco de dados utilizamos o programa descrito em A e foi necessária uma formatação diferente de acordo com o tipo de dado. Descreveremos agora as formatações utilizadas.

3.2.1 Formatações dos dados

Genes. Antes de usarmos os dados dos genes, as suas sequências nucleotídicas precisam ser alinhadas, ou seja, devem ser organizadas de tal forma que tenham o mesmo tamanho e as regiões similares estejam na mesma posição. Por esse motivo, utilizamos a ferramenta Clustal Omega, que nos permite fazer o alinhamento de múltiplas sequências contidas em um arquivo tipo FASTA. Depois de alinhadas as sequências, podemos inserir o resultado desse alinhamento num arquivo de entrada do MrBayes.

Proteoma total ou N-glicanos. Os dados de N-glicanos e proteicos (obtidos pelo ensaio de proteoma total) foram convertidos para valores discretos relacionados com a presença ou ausência dos mesmos em um dado veneno e inseridos em uma matriz de ocorrência binária, inserindo-a num arquivo de entrada do MrBayes.

Proteomas total e enriquecidos por lectinas. Combinando as informações do protocolo de proteoma total com as obtidas dos três ensaios feitos com enriquecimento por lectinas, também construímos matrizes de ocorrência proteica tal que cada entrada da matriz tenha um valor entre 0 e 4, equivalente ao número de ensaios em que o par $\langle veneno, proteina \rangle$ foi observado. A ideia dessa construção seria valorizar dados observados mais frequentemente,

mitigando assim falsos positivos.

Peptídeos. Entre as sequências peptídicas armazenadas em nosso banco, podemos ter casos que sequências diferentes se referem ao mesmo segmento de proteína. Isso pode acontecer em função de polimorfismo de nucleotídeo único (SNP), uma variação na sequência que afeta somente uma base. É possível, também, que durante a espectrometria de massa a fragmentação da proteína ocorreu em pontos diferentes, gerando sequências de aminoácidos do mesmo peptídeos com comprimentos diferentes ou de seções diferentes. A diferença pode ocorrer também na identificação computacional, um processo probabilístico.

Portanto, precisamos de uma forma de medir a similaridade de sequências, de tal forma que se duas sequências são muito similares, então consideramos que ambas se referem ao mesmo peptídeo. Dessa forma, podemos eliminar assim a redundância de duas ou mais sequências similares entre si, e que putativamente se referem ao mesmo segmento proteico. Portanto, para construção de matrizes de entrada do MrBayes com dados de peptídeos primeiro aplicamos o método de remoção dessas redundâncias, que será apresentado na próxima seção. Então, construímos uma matriz sobre as classes que pode ser binária se utilizamos apenas informações do proteoma total ou com valores entre 0 e 4, como no caso das proteínas, se acrescentarmos as informações dos protocolos com enriquecimento por lectinas.

3.2.2 Método de equivalência de peptídeos

Desenvolvemos uma metodologia, que vamos chamar de equivalência de peptídeos, que dado um conjunto de sequências peptídicas construímos a relação de equivalência entre elas, estabelecendo quais peptídeos são similares entre si, dentre todos os peptídeos detectados em todos os proteomas de venenos. Este método baseia-se no estabelecimento de uma relação de similaridade entre sequências peptídicas e do uso da ferramenta BLAST para alinhamento 2-a-2 das mesmas.

Critério de similaridade. Podemos considerar a similaridade entre sequências uma relação de equivalência; utilizaremos $\equiv$ para denotar essa relação. Então, sejam s_1, s_2 e s_3 três sequências quaisquer; a relação de equivalência satisfaz as seguintes propriedades:

$$s_1 \equiv s_1 \qquad \text{(Reflexividade)}$$

$$s_1 \equiv s_2 \Rightarrow s_2 \equiv s_1 \qquad \text{(Simetria)}$$

$$s_1 \equiv s_2, s_2 \equiv s_3 \Rightarrow s_1 \equiv s_3. \qquad \text{(Transitividade)}$$

O programa BLAST. Para verificar se duas sequências são similares entre si, utilizamos a ferramenta BLAST, um programa que encontra regiões de similaridade entre duas sequências biológicas [10]. Ele compara dois-a-dois uma dada sequência (nucleotídicas ou de

aminoácidos) com as sequências armazenadas em um banco de dados, calculando a significância estatística dos eventuais alinhamentos encontrados. Mais especificamente, o BLAST calcula uma pontuação, que é uma nota baseada no número de pareamentos perfeitos e imperfeitos entre as sequências; então, quanto maior a nota melhor o pareamento. O BLAST calcula também um valor estatístico (*e-value*) indicando se o alinhamento foi obtido por acaso.

O nosso método possui os seguintes parâmetros: MIN_EVALUE, MAX_HITS e MAX_DIFF. Os dois primeiros são usados para limitar o resultado do BLAST e o último determina a maior diferença de comprimento que duas sequências podem ter para serem consideradas equivalentes. Utilizamos a estrutura *Union Find* manter a relação de equivalência e obter as classes. O programa que implementa essa metodologia está documentado e explicado mais a fundo no Apêndice B.

3.3 Scripts e Programas

Para automatizar algumas tarefas e assim implementar o pipeline proposto, escrevemos diversos programas em Python (versão 3). O código-fonte de todos esse programas se encontra no repositório deste projeto, disponível de forma livre e gratuita sob a licença GNU GLP. No restante desta seção descreveremos tarefas feitas por esses scripts e as principais bibliotecas utilizadas.

3.3.1 Gerenciamento do banco de dados

A manipulação do banco de dados, como a inserção e edição de informação ou apenas a recuperação dos dados, foi feita por meio da biblioteca Psycopg. Que é um adaptador de um banco de dados gerenciado pelo PostgresSQL para a linguagem de programação Python.

3.3.2 Biopython

Como trabalhamos com muitos dados biológicos precisamos de uma biblioteca de Bioinformática; portanto, utilizamos o conjunto de ferramentas para Biologia Molecular Computacional chamada de Biopython. O projeto Biopyhton é extremamente completo, possui diversas ferramentas para diferentes funções. Neste trabalho usamos extensivamente as ferramentas de manipulação de sequências nucleotídicas e arquivos do tipo FASTA. Além disso, usamos os invólucros Biopython das ferramentas BLAST e Clustal Omega.

3.3.3 Testes automatizados

Para a criação de testes utilizamos o arcabouço Pytest, que nos permite fazer testes simples de nossos métodos e programas.

Fizemos testes de integridade sobre os dados de entradas e das saídas de nossos programas; isto é, verificamos se os dados de entrada estão no formato correto e que os dados de saída contém o resultado esperado: por exemplo, se a função que recebe como argumento o tipo de ensaio proteômico é esperado que a mesma devolva a respectiva matriz de ocorrências das proteínas. Além disso, fizemos testes unitários de corretude do gerador de arquivo NEXUS A.

3.3.4 Ferramentas de manipulação de arquivos e de visualização

Para podermos utilizar a implementação do teste CADM (seção 2.5), utilizamos o pacote de análise de filogenia e evolução ape para R. Para isso, empregamos a interface rpy2 para executar códigos escritos na linguagem de programação R embutidos em um programa em Python.

Outros módulos Python utilizados foram:

- csv um módulo para manipulação de planilhas no padrão CSV, que foi utilizada para inserir os dados provenientes dos artigos no banco de dados;
- o módulo os para a manipulação de arquivos e diretórios;
- o módulo tempfile para criação de arquivos temporários.

Além dos programas em Python, utilizamos o programa FigTree para a visualização e edição dos cladogramas em formato NEXUS gerados em nossas análises.

Capítulo 4

Resultados

Utilizando os materiais e métodos que apresentamos na seção 3.2, geramos diversas cladogramas com diferentes dados como entrada. Ao longo deste capítulo, apresentaremos as árvores mais importantes que obtivemos e faremos uma breve análise de cada uma delas.

4.1 Árvore genômica

Inicialmente repetimos o procedimento de Fenwick et al. [4], porém restrito às sete espécies que estudamos (figura 4.1a). Apesar de não termos utilizados os dados morfológicos como foi feito no artigo, os resultados obtidos são topologicamente equivalentes (compare figura 4.1a com 4.1b).

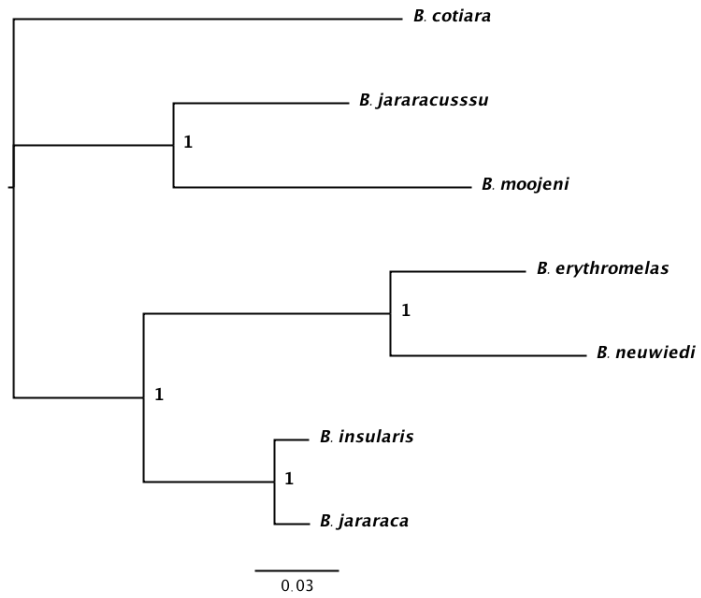
Além disso, geramos diversas árvores filogenéticas usando como entrada diferentes combinações das sequências e notamos que a árvore gerada com apenas os dados dos genes mitocondriais ND4 e ctyb mantém a equivalência topológica, portanto dispensando não somente o uso das informações morfológicas como também as trazidas pelos genes ribossomais.

Portanto, dada essa equivalência, utilizamos o cladograma da figura 4.1a como referência para todas as comparações que foram feitas com as árvores filoproteômicas geradas neste trabalho e que serão apresentadas nas próximas seções.

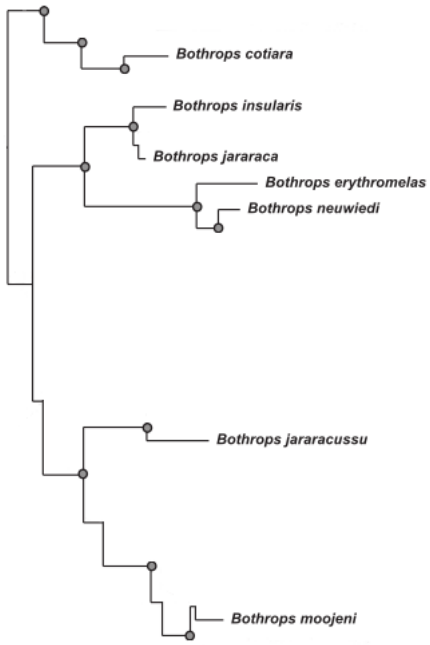
4.2 Árvores de proteínas

A partir do proteoma total das sete espécies de serpentes, fizemos uma inferência Bayesiana, da maneira descrita no capítulo 3. Na figura 4.2b mostramos a árvore filoproteômica obtida; observe a existência de uma grande discrepância topológica entre ela e a árvore obtida anteriormente por Andrade-Silva et al. [2] (compare figura 4.2b com 4.2a). Para verificar a distância dessa árvore obtida em relação à árvore de referência (figura 4.1a), utilizamos o teste CADM, cujo resultado, resumido na tabela 4.1, confirma de fato uma grande diferença topológica entre essas árvores.

Tendo em vista tal diferença entre a árvore referência e a filoproteômica da figura 4.2c,



(a)



(b)

Figura 4.1: *Cladogramas filogenéticos de sete espécies do gênero* Bothrops. *São apresentados dois cladogramas: a fig. 4.1a foi obtida por uma inferência Bayesiana com dados dos genes mitocondriais ND4 e cytb. as probabilidades a posteriori se encontram à direita dos nós. Já na fig. 4.1b é mostrada uma sub-árvore de um cladograma obtido através do uso de mtDNA e de características morfológicas - essa sub-árvore foi adaptada de Fenwick et al. [4].*

adicionamos as informações dos ensaios com enriquecimento das diferentes lectinas, imaginando que isso acarretaria em uma melhora no resultado (figura 4.2c). Todavia, apesar da semelhança visual entre as árvores referência e filoproteômica aumentar um pouco, a congruência entre elas não aumentou significativamente, o que foi comprovado pelo resultado do teste CADM.

4.3 Árvore de N-glicanos

No artigo de Andrade et al. (2018) [3] foi feito um clustering (aglomeração) hierárquico dos venenos de acordo com a composição dos N-glicanos, cujo resultado é ilustrado na figura 4.3a. Com os mesmos dados dessas estruturas fizemos uma inferência Bayesiana da

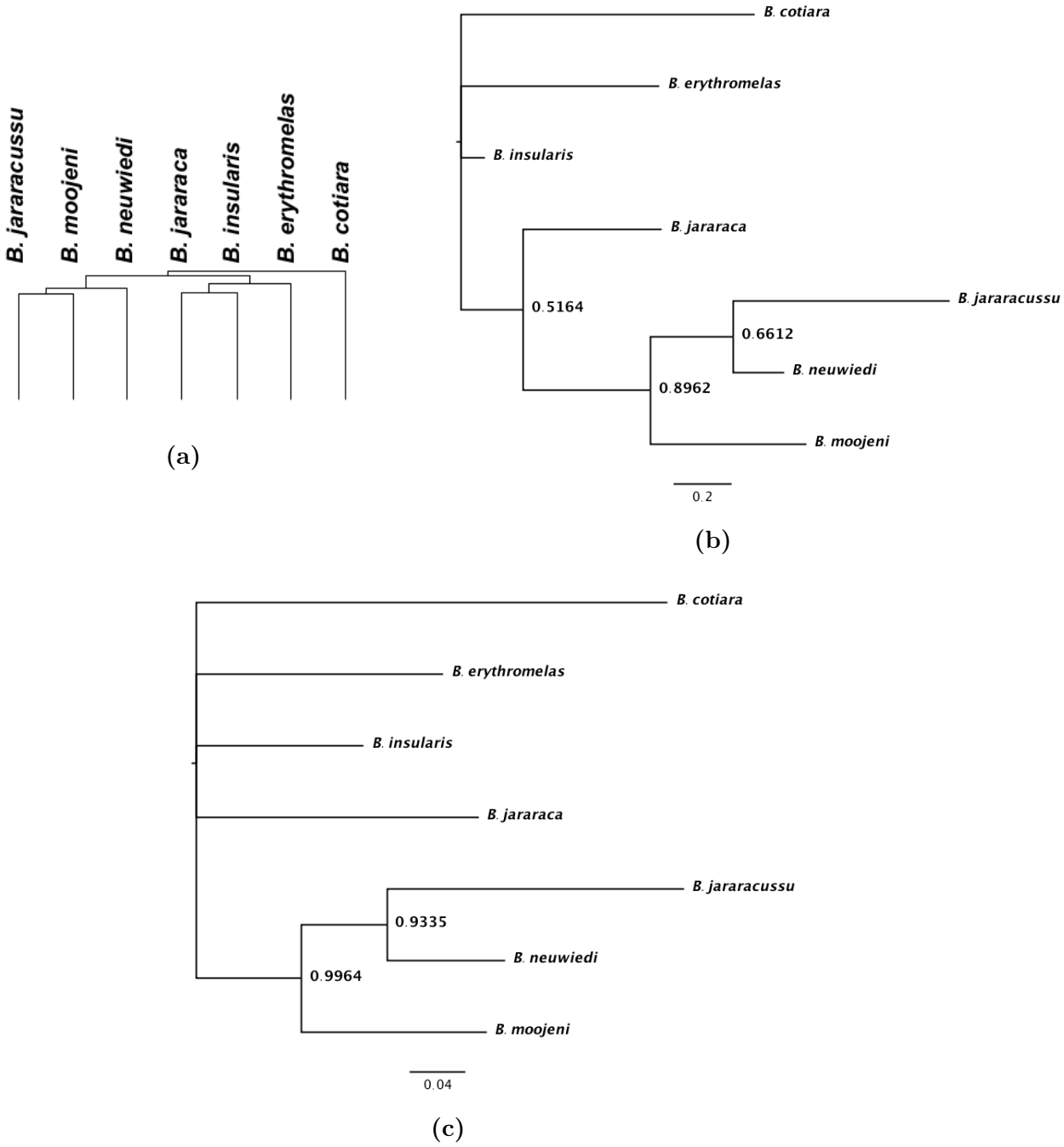


Figura 4.2: *Cladogramas filoproteômicos de sete serpentes* Bothrops. *São apresentados três cladogramas: a fig. 4.2a é um cladograma extraído de Andrade-Silva et al. [2], obtido através de uma aglomeração hierárquica sobre proteínas detectadas em um ensaio de proteoma total de veneno. A fig. 4.2b foi obtida por uma inferência Bayesiana com dados binários da presença ou ausência das proteínas observadas no proteoma total. Já a fig. 4.2c foi obtida por uma inferência Bayesiana com dados das proteínas observadas nos proteomas total e com enriquecimento com três diferentes lectinas. Para todas as árvores inferidas, probabilidades à posteriori se encontram à direita dos nós.*

Cladograma	W	χ^2
proteoma total	0.4942	19.77
proteoma total + enriquecimento	0.5058	20.23

Tabela 4.1: *Resultados de dois testes CADM, comparando a árvore de referência (figura 4.1a) com a árvore de proteoma total (figura 4.2b) ou de proteoma total com as informações dos ensaios com enriquecimento por lectina (figura 4.2c).*

maneira descrita no capítulo 3, obtendo assim um cladograma (figura 4.3b) com topologia idêntica a do artigo.

Portanto, concluímos que os procedimentos com esses dados não tem resolução suficiente para gerar um perfil filogenético refinado, como pode ser visto pela discrepância das espécies *B. neuwiedi* e *B. erythromelas*. Para confirmar essa discrepância estatisticamente, empregamos o teste CADM, comparando os cladogramas das figuras 4.1a e 4.3b e obtendo seguintes valores: $W = 0.4653$ e $\chi^2 = 18.61$; ou seja, verificamos assim a pequena congruência entre as árvores. Tampouco combinar as informações de estruturas de N-glicanos com as do proteoma proporcionou uma melhora significativa (dados não mostrados).

4.4 Árvores de peptídeos

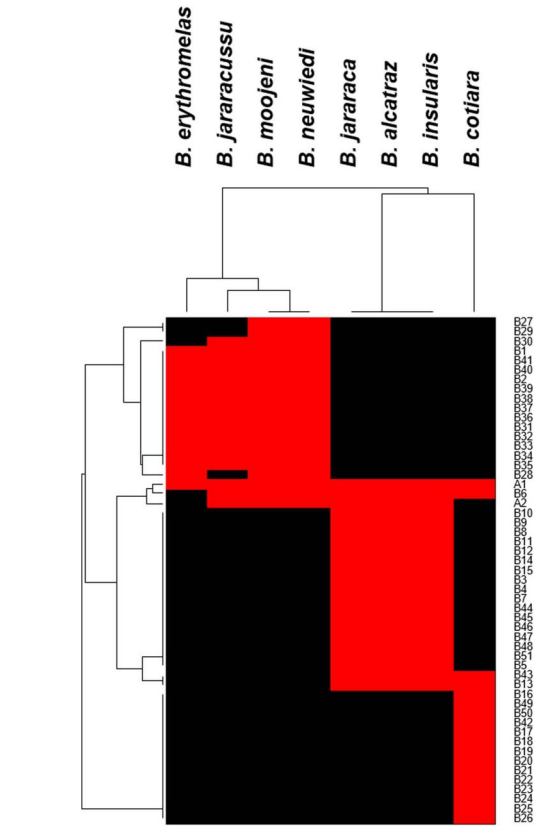
Neste trabalho geramos árvores de peptídeos identificados através de duas estratégias: com o auxílio de banco de dados de sequências e também com a abordagem de novo, que dispensa o uso de tal banco.

4.4.1 Peptídeos identificados com banco de dados de sequências

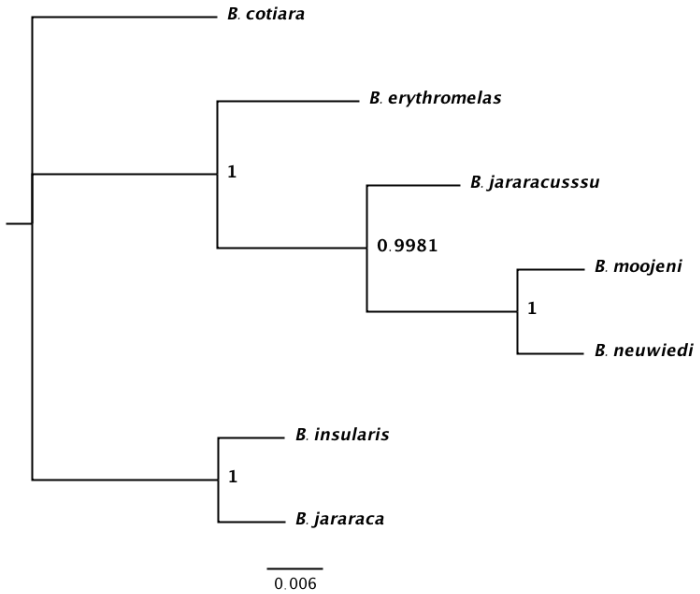
A partir dos peptídeos identificados pelo protocolo de proteoma total das sete espécies de serpentes, fizemos uma inferência Bayesiana, da maneira descrita no capítulo 3. Na figura 4.4a mostramos a árvore obtida; note uma topologia muito semelhante à do cladograma de referência, o que é corroborado pelo resultado do teste CADM apresentado na tabela 4.2. No entanto, há um grande grande incerteza na ramificação entre a *B. neuwiedi* e *B. erythromelas*, pois a mesma possui uma probabilidade *a posteriori* de apenas 0.5025.

Da mesma forma que fizemos com os dados proteicos, adicionamos as informações dos ensaios com enriquecimento das diferentes lectinas; neste caso, utilizamos os peptídeos identificados nesses experimentos. O resultado dessa análise é apresentado na figura 4.4b; observe que houve uma pequena piora com a grande divergência da espécie *B. neuwiedi* e uma pequena da *B. erythromelas*, comprovada pelo resultado inferior do teste CADM em relação ao obtido com a árvore da figura 4.4a. No entanto, ainda é uma árvore mais congruente à referência quando comparada com as geradas com dados de proteínas ou de N-glicanos.

Como descrito na seção 3.2, construímos uma metodologia de equivalência de peptídeos por meio da estrutura *Union Find*. Utilizamos ela na construção das matrizes que



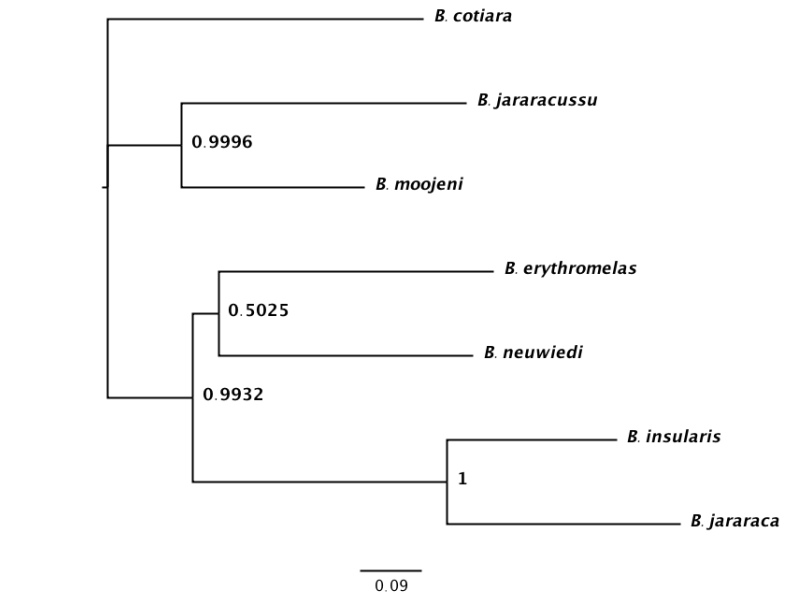
(a)



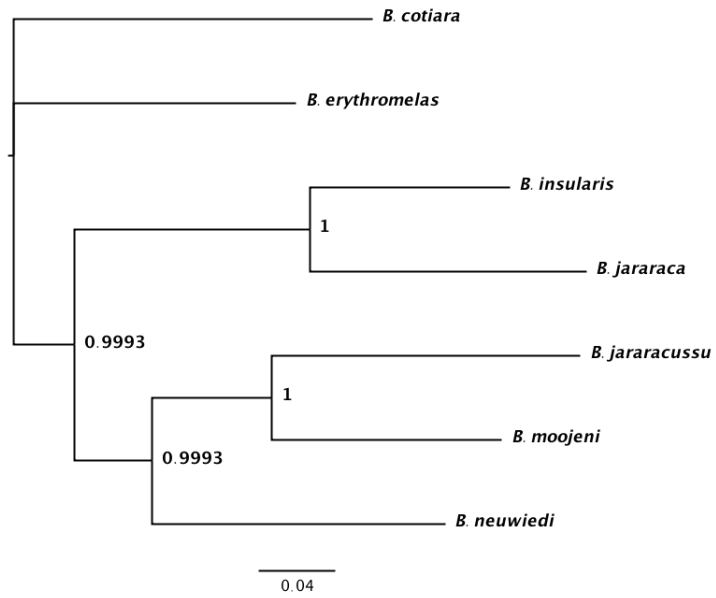
(b)

Figura 4.3: *Análises filoproteômicas de estruturas de N-glicanos. Na fig. 4.3a, extraída de Andrade-Silva et al. (2018) [3], apresentamos uma visualização gráfica de um clustering de acordo com a composição dos N-glicanos. Para cada veneno, uma dada estrutura está presente (vermelho) ou ausente (preto). Já na fig. 4.3b mostramos o cladograma obtido por uma inferência Bayesiana com dados das estruturas de N-glicanos; as probabilidades a posteriori se encontram à direita dos nós.*

geraram as árvores na figura 4.4, utilizando os seguintes parâmetros: MIN_EVALUE =



(a)



(b)

Figura 4.4: *Árvores filoproteômicas geradas com peptídeos identificados com banco de dados de sequências. Na fig. 4.4a exibimos um cladograma obtido por uma inferência Bayesiana com dados peptídicos binários provenientes de protocolos de proteoma total. Já na fig. 4.4b mostramos o cladograma obtido por uma inferência Bayesiana com dados peptídicos provenientes de protocolos de proteoma total e com enriquecimento por lectina. As probabilidades à posteriori se encontram à direita dos nós.*

10^{-20}, MAX_HITS = 3 e MAX_DIFF = 0. A árvore resultante da aplicação da equivalência de peptídeos manteve a topologia; além disso, os resultados do teste CADM foram idênticos comparado a árvore gerada sem a utilização esse método.

Cladograma	W	χ^2
peptídeos de proteoma total	0.7927	31.7096
peptídeos proteoma total + enriquecimento	0.6080	24.3206

Tabela 4.2: *Resultados do teste CADM comparando as árvores de peptídeos de proteoma total (figura 4.4a) e peptídeos de proteoma total com as informações dos ensaios com enriquecimento por lectina (figura 4.4b) com a árvore de referência (figura 4.1a).*

Todavia, durante nossos testes notamos que deve-se tomar cuidado na decisão dos parâmetros para a execução da equivalência de peptídeos. Se os parâmetros tornarem a relação de similaridade muito flexível pode-se perder informação e modificar o cladograma. Um exemplo que mostra essa possível divergência foi obtido com os seguintes parâmetros: MIN_EVALUE $= 10^{-5}$, MAX_HITS $= 5$ e MAX_DIFF $= 2$. Com isso geramos a matriz e fizemos a análise, obtendo uma árvore na qual a espécie *B. neuwiedi* divergiu quando comparada com a árvore genômica de referência, enquanto que a topologia das outras seis espécies se manteve.

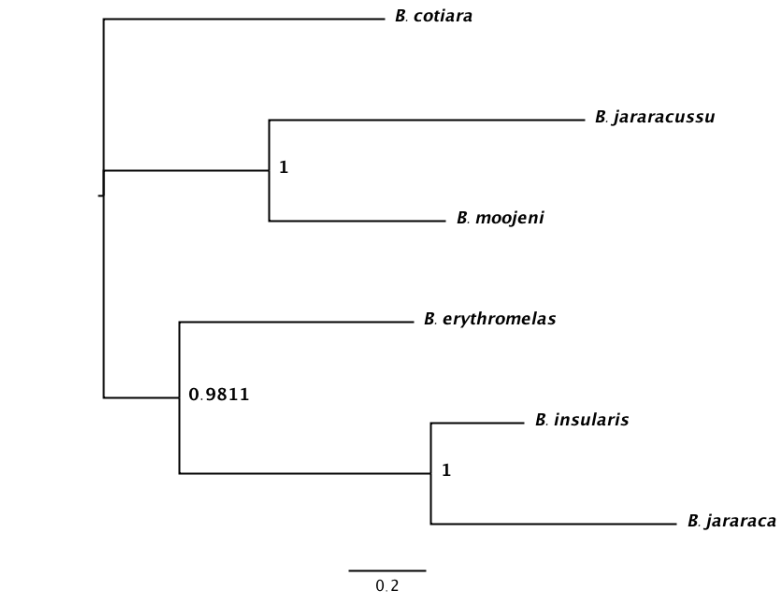
Para confirmar que tal discrepância afeta somente *B. neuwiedi*, geramos árvores genômicas e com dados peptídicos com apenas seis espécies de *Bothrops*, excluindo *B. neuwiedi* das análises. No caso da árvore gerada com dados peptídicos, utilizamos a equivalência de peptídeos com os parâmetros MIN_EVALUE $= 10^{-5}$, MAX_HITS $= 5$ e MAX_DIFF $= 2$. E, de fato, as árvores obtidas eram topologicamente equivalentes (figura 4.5).

4.4.2 Peptídeos identificados pelo protocolo de novo

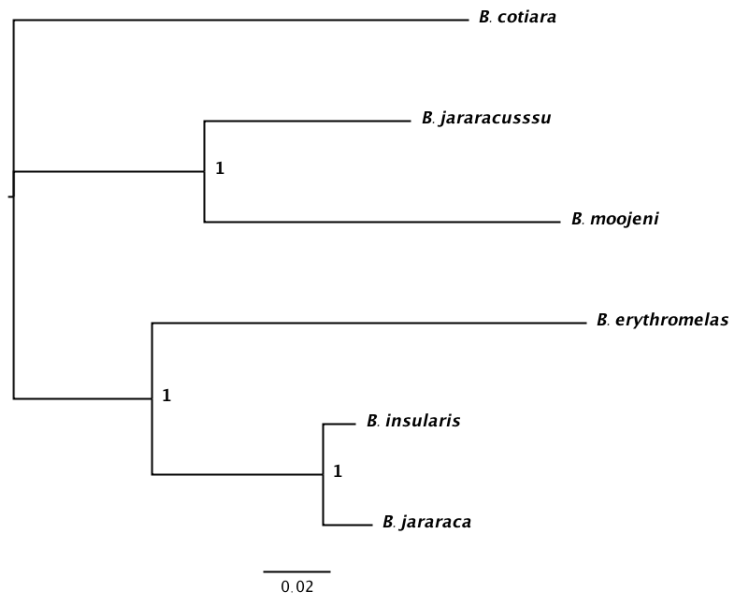
Uma hipótese para a discrepância no posicionamento de *B. neuwiedi* reportada nos resultados anteriores é que a mesma seria devido ao viés que a superrepresentação de *B. jararaca* (e a subrepresentação da própria *B. neuwiedi*), mostrada nas tabelas 2.1 e 2.2, acarreta no processo de identificação dos peptídeos por banco de dados de sequências.

Para testar essa hipótese, optamos por uma estratégia que elimina totalmente o viés dos dados anteriores causado pelo uso de um banco de dados de sequências. Tal estratégia utiliza o protocolo de novo, que foi aplicado nos dados brutos conforme descrito na seção 2.3. Como esse tipo de identificação não conta com a restrição imposta pelas sequências protéicas nas possibilidades de sequências de peptídeos, a lista de proteínas identificadas pela estratégia de novo é muito maior (um total de 5408, comparado com 1212 sequências obtidas pela identificação usando o banco de dados).

Dessa forma, com uma lista maior de peptídeos, pudemos testar de forma mais precisa nossa metodologia de equivalência de peptídeos do que nos experimentos anteriores. Nosso teste consistiu em variar a definição de similaridade, tornando-a mais flexível, e verificando os resultados; portanto fixamos os parâmetros MAX_HITS $= 5$ e MAX_DIFF $= 2$ e variamos o parâmetro MIN_EVALUE. Na tabela 4.3 apresentamos os resultados do teste CADM comparado a árvore genômica com árvores filoproteômicas obtidas com diferentes níveis de similaridade; observe um melhoramento da congruência conforme aumentamos a



(a)



(b)

Figura 4.5: *Exclusão de* B. neuwiedi *das análises leva a árvores topologicamente equivalentes. Na fig. 4.5a mostramos um cladograma obtido por uma inferência Bayesiana com dados peptídicos em forma de uma matriz binária, provenientes de protocolos de proteoma total após a aplicação da metodologia de equivalência de peptídeos. Já a fig. 4.5b contém uma árvore filogenética, obtida por uma inferência Bayesiana com dados dos genes mitocondriais ND4 e cytb. As probabilidades a posteriori se encontram à direita dos nós.*

flexibilidade da relação. Além disso, na figura 4.6, que mostra as respectivas árvores resultantes, nota-se um aumento das probabilidades *a posteriori* dos nós. Portanto, concluímos que nossa metodologia reduz o número de dados que serão analisados, mantendo a topologia ou até tornando-a mais precisa, cumprindo o papel do que é conhecido em Aprendizado de Máquina como seleção de características. Além disso, verificamos que a discrepância no

posicionamento de *B. neuwiedi* manteve-se também em todos esses experimentos (figura 4.6).

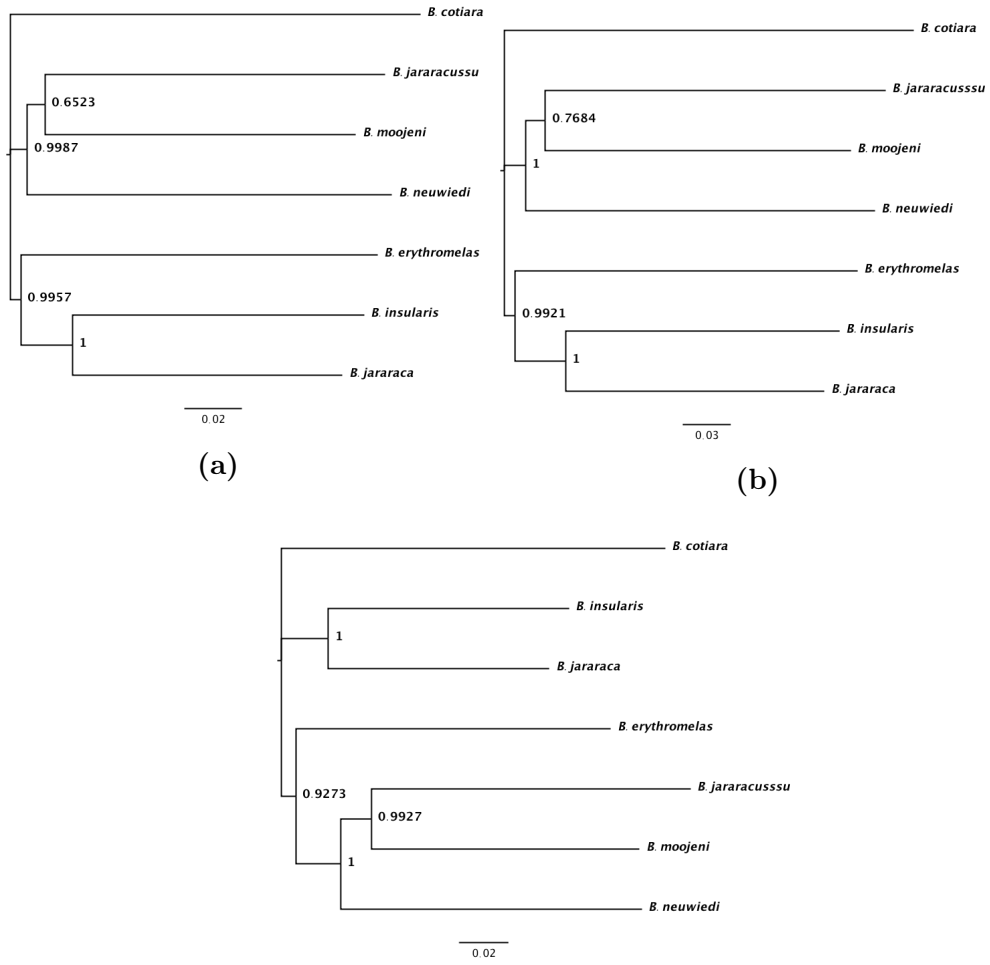


Figura 4.6: *Árvores filoproteômicas obtidas por uma inferência Bayesiana com dados peptídicos binários provenientes de protocolos de proteoma total e identificados por protocolo de novo. Todas as análises foram feitas após aplicada a metodologia de equivalência de peptideos com os parâmetros* MAX_HITS $= 5$, MAX_DIFF $= 2$ *e* MIN_EVALUE *igual a* 10^{-10} *(Fig. 4.6a),* 10^{-5} *(fig. 4.6b) e* 10^{-2} *(fig. 4.6c). As probabilidades a posteriori se encontram à direita dos nós.*

Árvore	(MIN_EVALUE)	W	χ^2	Número de Sequências
fig. 4.6a	(10^{-10})	0.6662	26.6493	5408
fig. 4.6b	(10^{-5})	0.7207	28.8311	4901
fig. 4.6c	(10^{-2})	0.7214	28.8571	3258

Tabela 4.3: *Resultados do teste CADM comparando cada uma das árvores de peptídeos de proteoma total identificados pelo protocolo de novo (figura 4.6), geradas após a aplicação da metodologia de equivalência de peptídeos variando o* MIN_EVALUE*, com a árvore de referência (figura 4.1a).*

Capítulo 5

Conclusões

Neste trabalho testamos a hipótese de que o perfil proteômico de venenos de serpentes do gênero *Bothrops* é altamente relacionado com a filogenia das espécies. Para isso, ao longo do projeto foram geradas diversas contribuições tecnológicas e científicas.

Tecnologicamente, o encadeamento desenvolvido para geração, comparação e visualização de cladogramas obtidos por análises evolucionárias com dados não tradicionais, foi uma contribuição importante. Além disso, o método de equivalência de peptídeos é um processo inovador que, pelos nossos resultados, sua aplicação gera uma remoção de dados redundantes e uma melhora da análise, num processo que em Aprendizado de Máquina é chamado de seleção de características. Como essas metodologias funcionam com peptídeos identificados por protocolo de novo, elas podem ser aplicadas em organismos não modelos, isto é, organismos que não possuem seu genoma completamente sequenciado. Futuramente, é possível estender essas metodologias para outras serpentes, ou mesmo para organismos em outros contextos biológicos, como por exemplo de modificações epigenéticas (i.e., mudanças que não envolvem alterações no DNA) em câncer, comparando a filogenia com a filoproteômica do tumor, ou então a filoproteômica entre dois ou mais tumores.

Cientificamente, a principal contribuição foi o teste da hipótese mencionada acima, levantada no início deste trabalho. Para este fim, fizemos diversas comparações entre a árvore filogenética com árvores obtidas por uma combinação de dados proteicos, glicoproteicos, de N-glicanos e de peptídeos. Verificamos, então, que as árvores geradas com apenas dados de peptídeos obtiveram um resultado mais topologicamente congruente com o cladograma genômico. No entanto, em todos os casos vimos uma divergência da serpente *B. neuwiedi*, já que ela ou fica topologicamente distante da posição da árvore filogenética ou se aparece na mesma posição com baixa a probabilidade *a posteriori*.

Pensamos em duas hipóteses para essa divergência. A primeira delas diz respeito à composição das amostras de venenos utilizadas nos ensaios proteômicos: cada veneno é composto de um *pool*, ou seja, uma mistura de extrações de venenos de ao menos 10 espécimes diferentes. No caso do *pool* de veneno de *B. neuwiedi* utilizado, existe o problema de recentemente a espécie *B. neuwiedi* ter sido particionada, ou seja, o que antes eram subespécies se tor-

naram espécies. Como o *pool* é muito antigo, não há registros de quais dessas subespécies contribuíram para o mesmo e em qual quantidade. Além disso, essas antigas subespécies estão muito espalhadas geograficamente e possuem dietas muito diferentes. Portanto, o *pool* provavelmente é uma combinação de venenos de espécies diferentes e isso poderia ser a causa da divergência observada; neste caso, poderia ser feita uma análise evolutiva com dados gerados a partir de novos *pools*, cada um deles gerado com espécimes de apenas uma das novas espécies derivadas da antiga *B. neuwiedi*.

Outra hipótese levantada seria uma discrepância causada por pressões evolucionárias: sabe-se que a evolução de genes é mais lenta do que modificações epigenéticas em um curto período de tempo; logo, se existirem pressões ambientais mais fortes sobre o veneno de *B. neuwiedi* em relação às demais espécies aqui estudadas, então o perfil de proteínas dos venenos poderia ter características não presentes no genoma. Logo, poderia ser aplicada a metodologia com dados de outros tecidos como controle (e.g., amostra de sangue, que não sofreria tanta pressão ambiental quanto o veneno) e comparar os resultados.

Apêndice A

Escritor de arquivo NEXUS

O programa `nexus.py`, escrito em Python3 implementa a classe `NexusWriter` que serve para a automatização da escrita de arquivos do tipo NEXUS para servir de entrada no MrBayes.

As distribuições *a priori* usadas não são muito informativas, assim a distribuição *a posteriori* será baseada nos dados de entrada. Então a topologia da árvore é distribuída de acordo com uma uniforme, os tamanho da árvore por uma combinação de $gama(1, 1)$ e $Dirchlet(1, 1)$ para o tamanho dos ramos e a frequência dos nucleotídeos é distribuida por uma $Dirchlet(1, 1, 1, 1)$. Os modelos de substituição são pré-definidos de acordo com o tipo de dado. Descrito em detalhes abaixo.

- **DNA**: o processo MCMC amostra sobre todos os modelos de substituição reversíveis no tempo e as taxas variam por uma distribuição *gama* com uma proporção de campos invariáveis para as taxas de variação de características.
- **Códon**: a sequência é particionada em 3 partições, uma para cada posição de códon. Para cada partição o processo MCMC amostra sobre todos os modelos de substituição reversíveis no tempo e as taxas variam por uma distribuição *gama* com uma proporção de campos invariáveis para as taxas de variação de características.
- **Discreto**: o modelo de substituição é o MkModel de Lewis (2001) [9] combinado com uma distribuição gama para as taxas de variação de características.

Além disso, cada conjunto de dados inseridos será uma partição da matriz com taxas independentes.

A.1 Métodos de NexusWriter

- `NexusWriter()`

  Construtor de uma instância da classe `NexusWriter`.

- `add(taxon, charset, datatype, seq)`

  Método para adicionar dados em um arquivo NEXUS

  - **taxon:** o taxon do organismo que possui os dados a serem inseridos.
  - **charset:** o nome do conjunto de dados, cada um se torna uma partição
  - **datatype:** tipo dos dados, pode ser DNA, Codon e Standard
  - **seq:** dados, que pode ser uma sequência de nucleotídeos (se o tipo for DNA ou Codon) ou uma sequência de números entre 0 e 9 (se o tipo for Standard).

- `setNgen(ngen)`

  Método para definir o número de gerações que a inferência Bayesiana vai executar, o valor padrão é 210^6.

  - **ngen: número de gerações**

- `setSampleFreq(samplefreq)` Método para definir a frequência de amostras na inferência Bayesiana, o valor padrão é 100 gerações.

  - **samplefreq: frequência de amostra**

- `writeFile(outfile)` Método para escrever o arquivo contendo os dados inseridos.

  - **outfile: arquivo destino**

A.2 Exemplo de Uso

```
1 from nexus import NexusWriter
2
3 nw = NexusWriter()
4
5 nw.add('insularis', '16S', 'DNA', 'GTATTAAAGGCG-CGCTGCCCAGTGAAAAATT')
6 nw.add('jararaca', '16S', 'DNA', 'GTATTA-AGGCGACGCTGCCCAGTGAAAA-TT')
7 nw.add('insularis', 'peptides' 'Standard', '110100001011011000010101')
8 nw.add('jararaca', 'peptides', 'Standard', '010100001010011101010110')
9
10 nw.writeFile("out.nex")
```

Apêndice B

Equivalência de Peptídeos

O programa `pep_equiv.py`, escrito em Python3 implementa a classe `PepEquiv` que é a implementação da metodologia desenvolvida explicada em 3.2.

B.1 Métodos de PepEquiv

- `PepEquiv(peptides)`

  Construtor de uma instância da classe PepEquiv.

  - **peptides:** lista de ids de peptídeos, os quais serão aplicados a metodologia. Os ids devem ser compatíveis com os do banco de dados do BLAST.

- `setParams(diff, hits, evalue)`

  Define os parâmetros que determinam a similaridade entre sequências de peptídeos

  - **diff:** diferença máxima de comprimento entre sequências, valor padrão é 2.
  - **hits:** número máxima de sequências devolvidas pelo BLAST, valor padrão é 5.
  - **evalue:** valor mínimo do E-value entre sequências do BLAST, valor padrão é $10^{-}5$.

- `run(db)`

  Método que computa as classes de equivalência da lista de peptídeos

  - **db:** nome do banco de dados do BLAST e do arquivo contendo a lista de peptídeos

- `getClasses`

  Devolve uma lista de peptídeos que são os representantes das classes de equivalência.

- `getRep(pep_id)`

  Devolve o representante da classe em que o peptídeos **pep_id** pertençe.

- `writeFile(filename)`

  Escreve o arquivo **filename** codificando a classe de equivalência. Para cada peptídeo na lista há uma linha `<peptideos>:<classe>`.

Referências Bibliográficas

[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. 1

[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. 1, 3, 6, 17, 23, 25

[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. 1, 7, 9, 17, 25, 27

[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. 1, 3, 16, 23, 24

[5] Jacques Colinge and Keiryn L Bennett. Introduction to computational proteomics. *PLoS computational biology*, 3(7):e114, 2007. 9

[6] John P Huelsenbeck and Fredrik Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001. 10, 17

[7] Pierre Legendre and François-Joseph Lapointe. Assessing congruenceamong distance matrices: Single-malt scotch whiskies revisited. *Australian & New Zealand Journal of Statistics*, 46(4):615–629, 2004. 12

[8] Véronique Campbell, Pierre Legendre, and François-Joseph Lapointe. The performance of the congruence among distance matrices (cadm) test in phylogenetic analysis. *BMC evolutionary biology*, 11(1):64, 2011. 13

[9] Paul O Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic biology*, 50(6):913–925, 2001. 18, 35

[10] AM Amaral, MS Reis, and FR Silva. Programa blast: guia prático de utilização. *Brasília: Embrapa*, 2007. 19