Um método baseado em multirresolução para análise filoproteômica de venenos de serpentes

Aluno: Gustavo Mendes Maciel

Orientador: Marcelo da Silva Reis

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

Laboratório Especial de Ciclo Celular (LECC), Instituto Butantan, 15 de maio de 2019

Resumo

Venenos de serpentes são misturas complexas de proteínas e peptídeos. Uma estratégia analítica muito utilizada para estudá-los é a proteômica baseada em espectrometria de massas (EM). Trabalhos recentes utilizando *clusterings* para analisar dados proteômicos produzidos por EM mostraram que há uma relação entre o proteoma dos venenos de diferentes espécies do gênero *Bothrops* e a classificação filogenética de tais serpentes. Todavia, a superrepresentação de algumas espécies (e.g., *B. jararaca*) no banco de dados utilizado para identificação computacional de peptídeos potencialmente acarreta em vieses nessas análises. Para mitigar isso, recentemente foram geradas árvores filoprotêomicas a partir de sequenciamento *de novo*, isto é, sem o uso de banco de dados; porém, essa abordagem gera muitos peptídeos falso positivos, o que também introduz ruído nas análises. Neste projeto, propomos contornar esses problemas utilizando diretamente os dados brutos provenientes da EM para estimar as árvores filoproteômicas. Para isso, utilizaremos uma estratégia multirresolução para construção de matrizes a partir desses dados. Tais matrizes serão utilizadas para gerar árvores filoproteômicas por meio de uma abordagem de inferência Bayesiana, empregando métodos de Monte Carlo com Cadeias de Markov. Por fim, o teste estatístico CADM será aplicado para comparar os cladogramas obtidos pelo método proposto nesse projeto e os obtidos pelo uso de dados de DNA mitocondrial. Assim, são esperados resultados mais contundentes acerca da influência do perfil proteômico dos venenos na filogenia de diferentes espécies de serpentes do gênero *Bothrops*.

Sumário

1 Introdução

Venenos de serpentes são misturas complexas de proteínas e peptídeos, fundamentais para a sobrevivência das espécies venenosas, podendo ser utilizados como defesa ou para imobilizar e matar presas. Já há algum tempo que os proteomas dos venenos (i.e., conjunto de todas essas proteínas) são estudados pela proteômica, com o objetivo de identificar as composições dos venenos e assim entender a relação entre seus componentes e os efeitos que acarretam em um outro organismo. A espectrometria de massas (EM), uma técnica analítica que mede a massa/carga de moléculas ionizadas, vem sendo amplamente utilizada na identificação das proteínas presentes nos venenos de serpentes. Trabalhos recentes utilizando essa técnica mostraram que há uma relação muito forte entre o proteoma dos venenos das diferentes espécies do gênero *Bothrops* e a classificação filogenética de tais serpentes [1, 2, 3].

As proteínas que compõe os venenos podem sofrer glicosilações, que são modificações pós-traducionais nas quais um glicano (i.e., um polissacarídeo) é ligado a essa proteína, formando uma glicoproteína. Os dois trabalhos de Andrade-Silva et al. [1, 2] mostraram que as variações de glicanos e glicoproteínas nos venenos de diferentes espécies contribuem para a caracterização dos fenótipos desses venenos. Cladogramas obtidos por meio de aglomerações hierárquicas sobre os glicoproteomas se mostraram semelhantes aos obtidos com uso de DNA mitocondrial (mtDNA). Porém, não foram utilizadas informações de peptídeos na construção dos cladogramas e não houve nenhuma quantificação na comparação deles.

Com o objetivo de quantificar e melhorar esses resultados, o trabalho de Victor Raposo [3] fez o uso de métodos de inferência Bayesiana para gerar os cladogramas a partir de informações trazidas pelos peptídeos identificados na espectrometria de massas. Para identificar os peptídeos, os dados provenientes da EM foram comparados aos dados teóricos de um banco de dados. Todavia, embora esse método tenha apresentado progressos, ainda possui alguns problemas. Para mitigar a superrepresentação de algumas espécies no banco

de dados, foi utilizada uma estratégia de sequenciamento *de novo*, que dispensa o uso do banco. Porém, tal estratégia apresenta limitações, principalmente porque são gerados muitos peptídeos candidatos que são falsos positivos, o que impacta se não na topologia das árvores obtidas, no tamanho dos ramos dessas árvores.

Dessa forma, uma possibilidade de contornar os problemas apresentados acima seria fazer uso das informações adquiridas da espectrometria de massas sem a identificação de peptídeos. Uma maneira seria utilizar matrizes que são obtidas quando consideram-se as intensidades de detecções da EM como uma função do tempo de eluição (i.e., do tempo do experimento). Exemplos dessas matrizes são apresentados na Figura 1. Nesses exemplos, observamos que esse tipo de matriz é esparsa e com uma distribuição não-homogênea das informações. Para casos como esse, uma abordagem potencialmente adequada envolve o uso de classificadores multirresolução. Essa categoria de classificadores já se mostrou adequada para esse tipo de dado de treinamento no contexto de processamento de imagens [4]. Todavia, tal técnica nunca foi explorada em análises filoproteômicas.

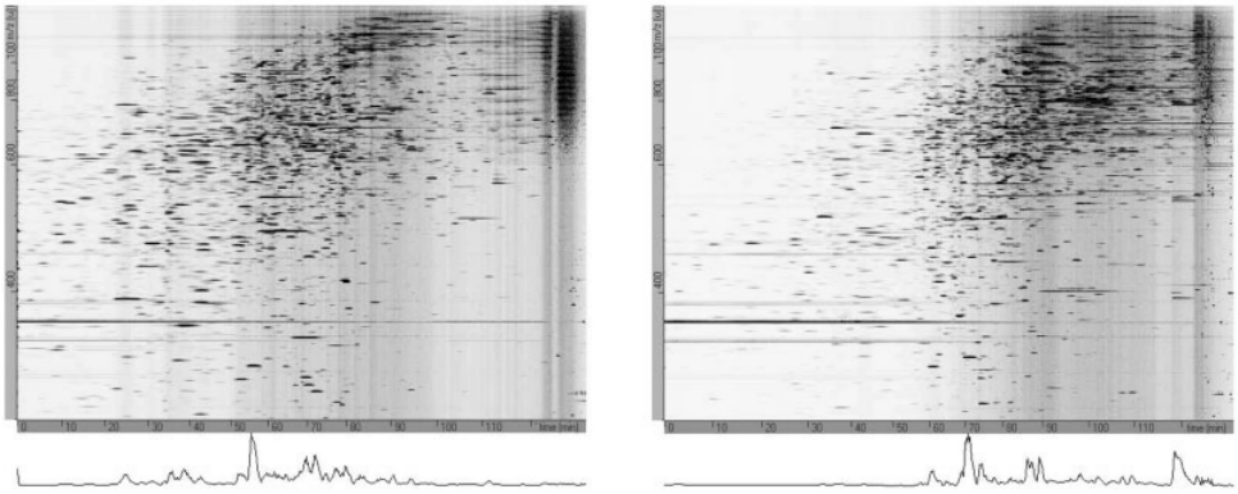

Figura 1: **Exemplo de visualização de um experimento de espectrometria de massas**. Neste gráfico são exibidas as intensidades de massa/carga como uma função do tempo de eluição. Figura extraída de Fox e Serrano [5].

2 Objetivos

Este projeto tem como objetivo geral desenvolver e implementar uma metodologia para construir árvores filoproteômicas a partir de dados proteômicos obtidos pela espectrometria de massas, sem a identificação de peptídeos e aplicando a estratégia de classificadores multirresolução.

Mais especificamente, essa metodologia será utilizada para construir cladogramas a partir de informações proteômicas adquiridas dos venenos de serpentes do gênero *Bothrops* por meio de EM, e compará-los com as árvores filogenéticas obtidas com o uso de dados genômicos. A ideia é que o método desenvolvido nesse trabalho apresente melhores resultados em relação aos trabalhos anteriores.

3 Metodologia

O encadeamento de processos que será utilizado para construir e analisar os cladogramas está apresentado na Figura 2, e será detalhado mais abaixo.

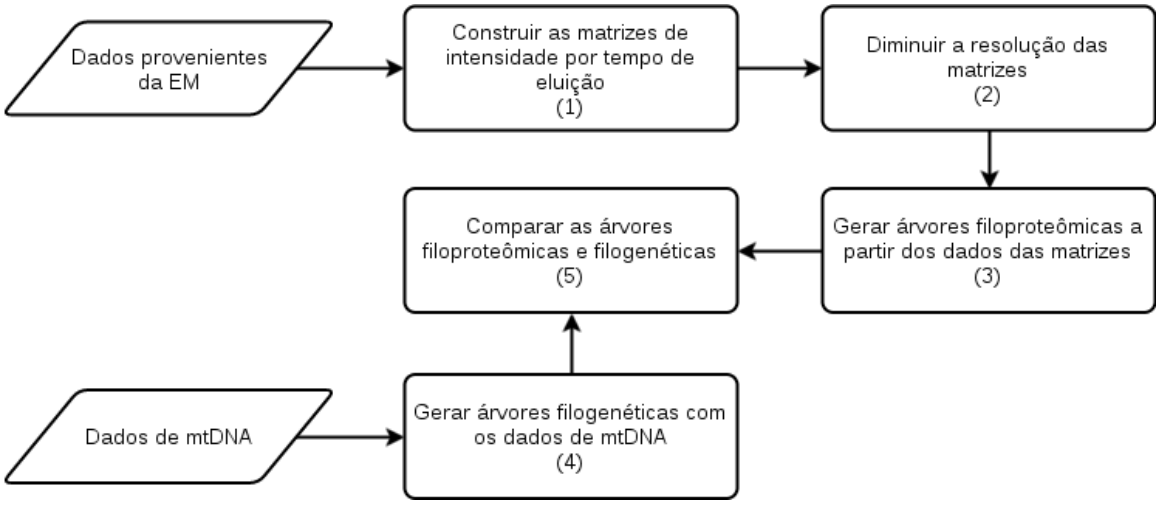

Figura 2: **Fluxograma do projeto proposto.** O retângulos representam os processos, que estão numerados de 1 a 5. Nos paralelogramos temos os dados que serão utilizados como entrada no encadeamento desses processos.

Construir as matrizes de intensidade por tempo de eluição (1). Para construir as matrizes de intensidade por tempo de eluição (MITEs) será necessário estudar os dados proteômicos que são adquiridos pela espectrometria de massas e utilizar algum *software* para gerar os arquivos que representam essas matrizes; uma possibilidade é o uso do Progenesis QI (Waters Corporation, Mitford, MA, Estados Unidos).

Diminuir a resolução das matrizes (2). As matrizes construídas no processo 1, que contêm informações dos peptídeos presentes nos venenos, serão utilizadas para gerar as árvores filoproteômicas. Porém, como já foi dito, essas matrizes são esparsas e têm uma distribuição não-homogênea das informações; além disso, suas dimensões podem ser extremamente grandes. Isso faz com que a comparação entre matrizes de venenos de diferentes espécies seja inviável. Para resolver esse problema será utilizada uma estratégia multirresolução [4], a fim de diminuir a resolução das MITEs, deixando a informação mais compactada. Daí surge um problema de otimização, pois deseja-se obter o melhor mapeamento de resolução possível sujeito à restrição de que o mapeamento precisa ser igual para todas as matrizes, para que comparações possam ser feitas entre elas. Uma possibilidade para resolver tal problema de otimização é utilizando o arcabouço featsel [6]. Esse arcabouço, codificado em C++, permite a inclusão de algoritmos de otimização discreta para seleção de características e também de funções custo. Dessa forma, adaptaremos ou criaremos dentro do featsel um algoritmo de otimização. Como função custo, empregaremos alguma das funções custo baseadas em informação mútua que já estão implementadas nesse arcabouço.

Gerar árvores filoproteômicas e filogenéticas (3 e 4). Para gerar as árvores filoproteômicas e filogenéticas por meio das MITEs e dos dados de mtDNA, respectivamente, a mesma abordagem de inferência Bayesiana usada no trabalho de Victor Raposo [3] será adotada aqui. Será necessário adaptar o código do programa que gera arquivos no formato NEXUS, desse mesmo trabalho, para gerá-los a partir dos dados obtidos no processo 2. Os ar-

quivos NEXUS servirão como entrada para o MrBayes [7, 8], programa que utiliza inferência Bayesiana para estimar árvores filogenéticas e/ou filoproteômicas. Para isso, o programa utiliza a estratégia de Monte Carlo com Cadeias de Markov, empregando o algoritmo de Metropolis-Hastings.

Comparar as árvores filoproteômicas e filogenéticas (5). Para comparar os cladogramas e avaliar os resultados obtidos, será aplicado o teste CADM, assim como no trabalho de Victor Raposo [3].

4 Plano de trabalho e cronograma de execução

Para a execução deste projeto proposto, foram listadas abaixo as principais atividades previstas. O diagrama de Gantt com o cronograma é apresentado na Tabela 1.

Atividade 1: Leitura inicial dos textos que servirão de base para o projeto, o que inclui a monografia do trabalho de Victor Raposo [3] e o artigo e dissertação que descrevem a abordagem multirresolução no projeto de W-operadores [4];

Atividade 2: Escrita do esboço do projeto de pesquisa;

Atividade 3: Estudo dos dados proteômicos provenientes da EM que serão usados para montar as MITEs e escolha do *software* que facilitará essa tarefa [5, 9]; conclusão do projeto de pesquisa;

Atividade 4: Estudo do MrBayes [7, 8] e do código legado do trabalho de Victor Raposo [3], incluindo o banco de dados e o programa que, a partir desse banco, gera os arquivos no formato NEXUS utilizados pelo MrBayes;

Atividade 5: Desenhar e implementar o programa que estima a classe de equivalência ótima a ser utilizada nas MITEs, baseando-se na abordagem multirresolução;

Atividade 6: Adaptar o código do trabalho de Victor Raposo [3] para gerar os arquivos no formato NEXUS a partir dos dados do item anterior;

Atividade 7: Rodar o MrBayes com os arquivos do item anterior e também com os dados de mtDNA que se encontram no banco de dados do trabalho de Victor Raposo [3];

Atividade 8: Comparar as árvores filoproteômicas (obtidas com o nosso método) com as filogenéticas (obtidas por meio de mtDNA) utilizando o teste CADM [10];

Atividade 9: Escrita da monografia do Trabalho de Formatura Supervisionado;

Atividade 10: Preparação e apresentação de pôsteres na Reunião Científica Anual do Instituto Butantan e na disciplina do Trabalho de Formatura Supervisionado.

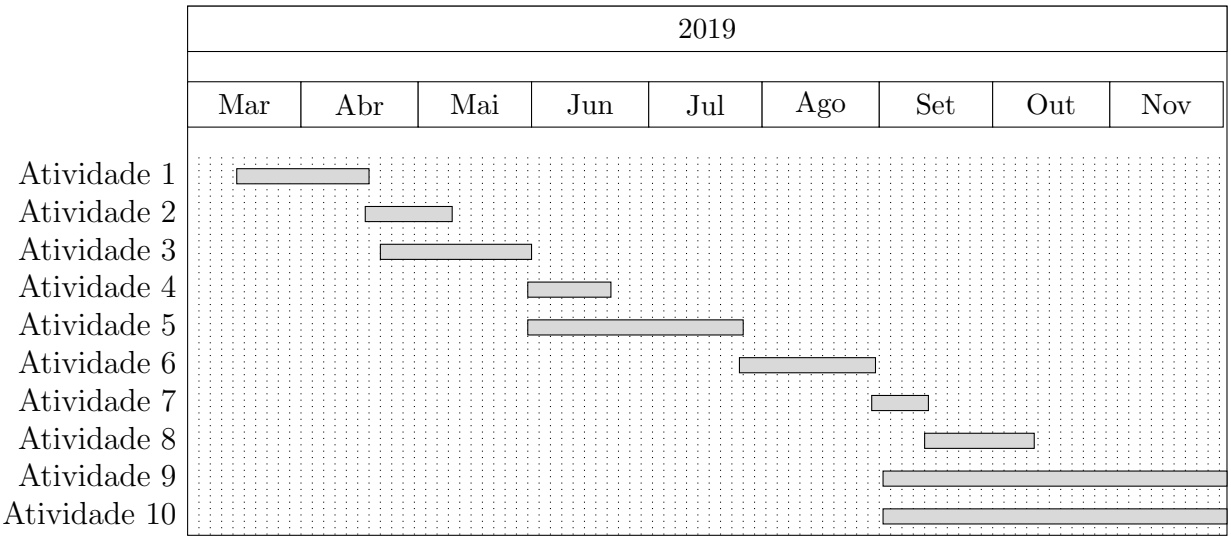

Tabela 1: Diagrama de Gantt contendo o cronograma de execução deste projeto proposto.

Referências

[1] 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 glyco-

proteomic profilings reveal that post-translational modifications of toxins contribute to venom phenotype in snakes. *Journal of proteome research*, 15(8):2658–2675, 2016.

[2] 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.

[3] Victor Wichmann Raposo. Análise filogenética computacional de serpentes do gênero *bothrops* a partir de proteomas de venenos. Technical report, Instituto de Matemática e Estatística, Universidade de São Paulo, 2018. Monografia de graduação em Computação.

[4] Daniel André Vaquero, Junior Barrera, and R Hirata. A maximum-likelihood approach for multiresolution w-operator design. In *XVIII Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI'05)*, pages 71–78. IEEE, 2005.

[5] Jay W Fox and Solange MT Serrano. Exploring snake venom proteomes: multifaceted analyses for complex toxin mixtures. *Proteomics*, 8(4):909–920, 2008.

[6] Marcelo S Reis, Gustavo Estrela, Carlos Eduardo Ferreira, and Junior Barrera. featsel: A framework for benchmarking of feature selection algorithms and cost functions. *SoftwareX*, 6:193–197, 2017.

[7] Fredrik Ronquist and John P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19(12):1572–1574, 2003.

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

[9] Andrew W Dowsey, Jane A English, Frederique Lisacek, Jeffrey S Morris, Guang-Zhong Yang, and Michael J Dunn. Image analysis tools and emerging algorithms for expression proteomics. *Proteomics*, 10(23):4226–4257, 2010.

[10] 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.