Otimização baseada em multirresolução para análise filoproteômica de proteomas gerados por espectrometria de massas

Aluno: Guilherme Costa Vieira

Orientador: Marcelo da Silva Reis

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

Laboratório de Ciclo Celular (LCC), Instituto Butantan, 25 de maio de 2020

Resumo

Serpentes utilizam seus venenos tanto para defesa quanto para obtenção de presas. Essas substâncias tratam-se na verdade de complexas misturas proteicas, que normalmente são estudadas através de técnicas como a Proteômica baseada em espectrometria de massas (EM). Em trabalhos recentes, foram mostrados indícios de que os perfis proteômicos de serpentes do gênero *Bothrops* se correlacionam com a árvore filogenética desses mesmos organismos. Todavia, a superrepresentação de algumas espécies *Bothrops* no banco de dados utilizado para identificação de proteínas após o ensaio de EM introduziu viés nesses resultados. Para mitigar isso, foi proposto o uso de dados brutos de EM para construção de árvores filoproteômicas, utilizando como base matrizes de tempo de eluição por massa/carga e empregando uma grade para redução de dimensionalidade dessas matrizes. Neste projeto, propomos substituir tal grade por um procedimento de otimização inspirado na abordagem multirresolução no desenho de operadores morfológicos, no qual diferentes particionamentos seriam testados, tomando como função custo um teste estatístico CADM contra a árvore filogenética. Num segundo momento, investigaríamos ainda o uso de técnicas de Aprendizado Não-Supervisionado para escolher um particionamento, seja num procedimento isolado ou em combinação com a otimização mencionada acima. Esperamos com este trabalho introduzir uma nova abordagem de inferência de árvores filoproteômicas que seria computacionalmente tratável e que também dispensaria o uso de banco de dados.

Sumário

1 Introdução

Venenos de serpentes são uma complexa mistura de proteínas, que têm papel fundamental na sobrevivência das espécies que os produzem, pois são usados tanto na defesa contra predadores como na caça de presas, possibilitando imobilizá-los e/ou matá-los. Conjuntos de proteínas, também denominados proteomas, são estudados no campo de conhecimento chamado Proteômica. Existem técnicas na Proteômica que permitem identificar as composições dos venenos, possibilitando o estudo de como os compostos afetam os organismos envenenados; uma das técnicas mais relevantes é a espectrometria de massas (EM), que o faz através da fragmentação das mesmas por processos químicos e físicos, ionização de cada fragmento e posterior medição da razão massa/carga de íons - as proteínas que compõem o veneno podem então ser identificadas a partir de buscas em banco de dados para associar cada íon detectado a uma porção de uma dada proteína. A EM foi utilizada em trabalhos recentes para mostrar que há evidências de que o perfil proteômico dos venenos das espécies de serpentes do gênero *Bothrops* se correlaciona com a classificação filogenética obtida atráves de DNA mitocondrial (mtDNA) e/ou características morfológicas [1, 2, 3].

Todavia, os trabalhos mencionados acima fizeram uso de técnicas que limitavam uma investigação mais aprofundada dos experimentos gerados nesses estudos. Por exemplo, para gerar cladogramas de venenos, utilizou-se aglomeração hierárquica sobre as proteínas identificadas por EM com o auxílio de um banco de dados; porém, os venenos de algumas espécies de serpentes eram superrepresentados nesse banco, o que enviesou em algum grau os resultados. Visando mitigar esse problema, Victor Raposo, em seu Trabalho de Conclusão de Curso (TCC) realizado em 2018, utilizou a estratégia de identificação *de novo* de peptídeos (i.e., subsequências protéicas) para dispensar o uso de dados, além do uso de inferência Bayesiana de árvores filoproteômicas ao invés de aglomeração hierárquica [3]; entretanto, apesar dessa abordagem mostrar resultados promissores, uma limitação foi a alta quantidade de candida-

tos peptídeos falso positivos. Uma metodologia alternativa foi então iniciada por Gustavo Maciel em seu TCC realizado em 2019, na qual foram utilizados dados de EM brutos, sem passar pela etapa de identificação de proteínas e/ou peptídeos [4]. Esses dados brutos eram organizados em matrizes, uma por veneno, onde as linhas e colunas correspondiam, respectivamente, a tempo de eluição de um íon no espectrômetro e massa/carga. Na figura 1 observam-se exemplos dessas matrizes.

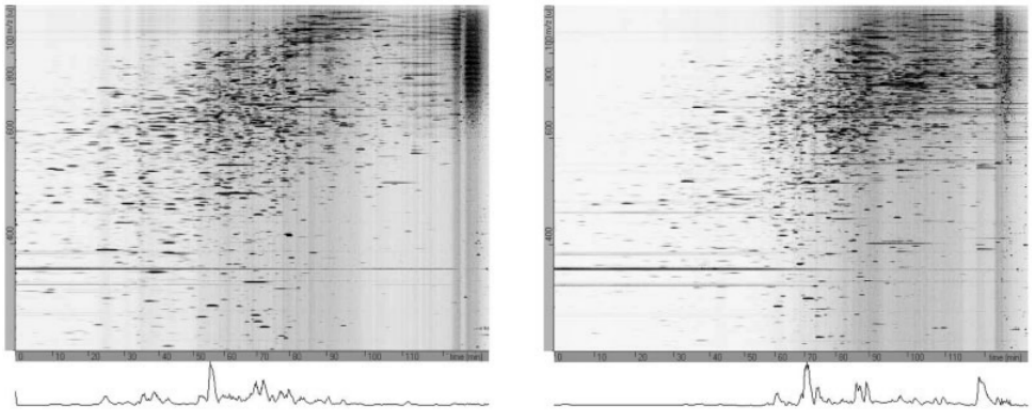

Figura 1: **Exemplo de visualização de um experimento de espectrometria de massas**. Neste gráfico são exibidas as intensidades de massa/carga como uma função do tempo de eluição. Figura extraída de Fox e Serrano [5].

Maciel então alinhou essas matrizes, gerou matrizes de ocorrência a partir delas e fez inferência Bayesiana de árvores filoproteômicas, de forma similar ao que foi feito por Raposo [4]. Todavia, a classificação topológica de duas espécies divergiram das árvores filogenéticas de referência de maneira mais significativa do que o esperado, possivelmente devido ao particionamento uniforme dos dados (i.e., para construir as matrizes de ocorrência, a dimensão das matrizes tempo de eluição por massa/carga foi reduzida através da utilização de uma grade); como os dados se espalham de maneira não-homogênea, sugere-se que o uso de uma grade para particionar esses dados para redução de dimensionalidade leva à atribuição de pesos iguais para áreas com pouca e muita informação (figura 2). Portanto, faz-se necessário o uso de técnicas de redução de dimensionalidade que levem em consideração a distribuição

não-homogênea dos dados brutos de venenos analisados em um espectrômetro de massas.

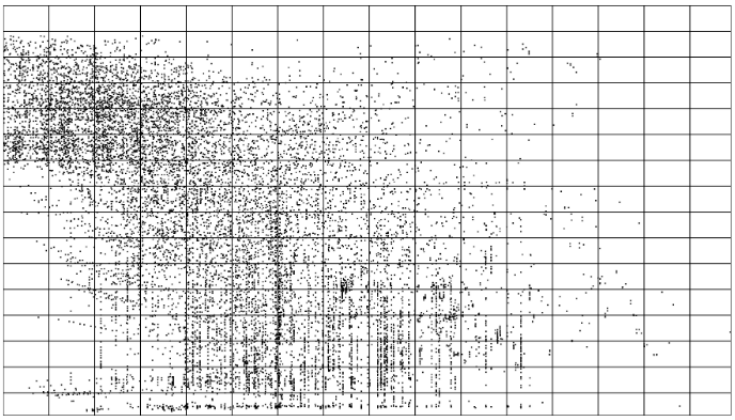

Figura 2: **Exemplo de visualização de um mapa de íon particionado de forma uniforme**. Neste exemplo, dividiu-se as linhas e as colunas por 16. Figura extraída do trabalho de Gustavo Maciel [4].

2 Objetivos

Este projeto tem como objetivo geral o desenvolvimento de uma metodologia de inferência Bayesiana de árvores filoproteômicas a partir de dados brutos de proteômica baseada em espectrometria de massas (EM). Como meta específica, propomos dar continuidade ao trabalho iniciado por Gustavo Maciel em seu TCC realizado em 2019 [4], substituindo a redução de dimensionalidade utilizando uma grade por uma busca de outros tipos de particionamento que melhorem o uso da informação contida nos dados de EM. O uso desse tipo de particionamento é inspirado no conceito de multirresolução, que é empregado no contexto de desenho de operadores morfológicos em Aprendizado Supervisionado [6]. Investigaríamos também o uso de Aprendizado Não-Supervisionado na redução de dimensionalidade, seja de forma isolada ou em combinação com essa otimização. Por fim, aplicaríamos, como estudo de caso, o novo método no desenho de árvores filoproteômicas utilizando os sete venenos de serpentes do gênero *Bothrops* já reportados anteriormente [1].

3 Metodologia

Como a busca por um particionamento ótimo dos dados brutos de EM é um problema muito difícil, lançaremos mão de algumas hipóteses de trabalho durante a execução deste projeto. Inicialmente suporemos que a árvore filoproteômica "correta" é a mais próxima, em termos de topologia, da respectiva árvore filogenética. Como podemos medir objetivamente a diferença de topologia entre duas árvores desse tipo utilizando um teste estatístico (e.g., o teste CADM [7]), logo podemos definir um problema de otimização no qual o espaço de busca é composto por todos os particionamentos possíveis e a função custo é a distância entre a árvore filogenética e a árvore filoproteômica gerada por um dado particionamento; ou seja, o objetivo aqui seria a minimização da distância entre as árvores filogenética e filoproteômica. Num segundo momento, abriríamos mão dessa hipótese e investigaremos técnicas de Aprendizado Não-Supervisionado para buscar o particionamento mais adequado para a inferência de árvores filoproteômicas. Em todas essas pesquisas, utilizaríamos os recursos computacionais que serão descritos no final desta seção.

3.1 Minimização da distância entre árvores filogenéticas e filoproteômicas

Como mencionado anteriormente, a metodologia nesta etapa consistiria numa generalização do encademento proposto por Maciel [4], porém acrescentando um *loop* ao encadeamento de processos original (Figura 3). As etapas envolvidas nessa generalização do encadeamento são as seguintes:

1. **Construir as matrizes de intensidade por tempo eluição.** Será adotada a mesma metodologia desenvolvida no trabalho de Maciel [4];

2. **Gerar um novo particionamento.** Nesta etapa será gerado um novo particionamento das matrizes. Caso este processo seja executado pela primeira vez (i.e., primeira iteração do encademento), um particionamento arbitrário será escolhido (e.g., aleatório, uniforme, etc.). Porém, para as execuções posteriores, a escolha do novo particionamento dependerá das escolhas anteriores, com critérios definidos pelo algoritmo de busca escolhido. Para este fim, propomos testar inicialmente os algoritmos

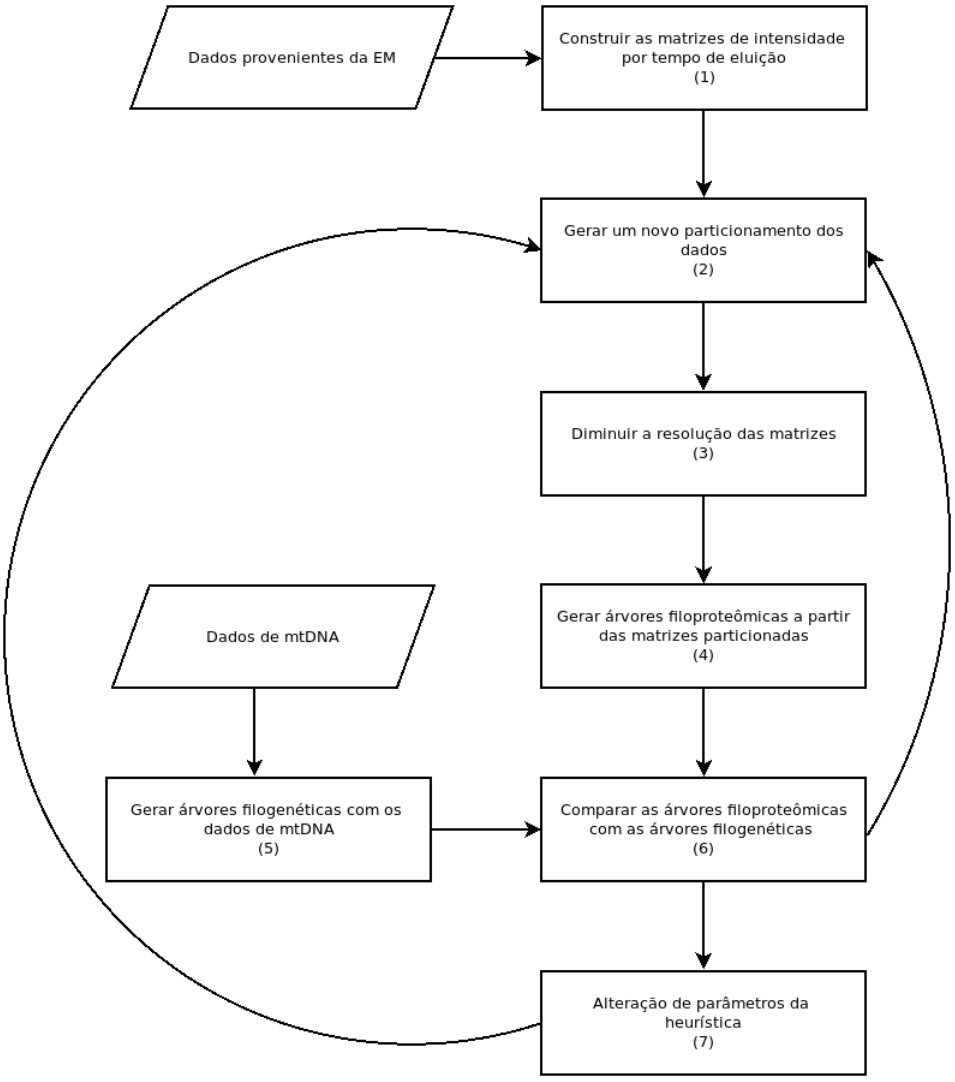

Figura 3: **Fluxograma do projeto proposto**. Os retângulos representam os processos, que estão numerados de 1 a 7. Nos paralelogramos temos os dados que serão utilizados como entrada no encadeamento desses processos.

de recozimento simulado (*simulated annealing*) [8] e de seleção sequencial [9].

3. **Diminuir a resolução das matrizes.** Será adotada a mesma metodologia desenvolvida no trabalho de Maciel [4];

4. **Gerar árvores filoproteômicas.** Para gerar as árvores filoproteômicas por meio das matrizes de tempo de eluição por massa/carga, a mesma abordagem de inferência Bayesiana usada por Raposo e Maciel será adotada aqui [3, 4];

5. **Gerar árvores filogenéticas.** Processo será similar ao anterior, porém substituindo as matrizes por dados de DNA mitocontrial (mtDNA);

6. **Comparar as árvores filoproteômicas e filogenéticas.** Para comparar os cladogramas e gerar uma medida de distância (i.e., uma realização da função custo de nosso procedimento de otimização), será aplicado o CADM, teste estatístico utilizado em Raposo [3]. Ao comparar as árvores, pode-se voltar para o processo (2) para a escolha de um novo particionamento, ou então ir para o processo (7), caso a abordagem utilizada no procedimento de otimização seja adaptativa;

7. **Alteração de parâmetros da heurística.** Caso o procedimento de otimização adotado seja adaptativo, mudanças seriam feitas nos parâmetros do algoritmo de busca utilizado e o procedimento retornaria para o processo (2).

3.2 Escolha de particionamento por Aprendizado Não-Supervisionado

Apesar do prevermos o uso da hipótese de que a árvore filoproteômica mais próxima da árvore filogenética seja a correta, não temos garantias da corretude dessa hipótese do ponto

de vista biológico. Por conta disso, numa segunda etapa também investigaremos a redução de dimensionalidade fazendo uso de técnicas de Aprendizado Não-Supervisionado. Em particular, investigaremos o uso para esse fim de mapas auto-organizáveis [10].

Poderíamos ainda utilizar a redução de dimensionalidade por Aprendizado Não-Supervisionado aqui proposta, como uma espécie de "burn-in" do procedimento de otimização da seção anterior, que proporcionaria uma redução do tamanho do espaço de busca e, consequentemente, da dificuldade computacional da otimização.

3.3 Recursos computacionais

Ao longo do projeto, será necessário testar diversas heurísticas para a resolução do problema de otimização deste trabalho [8, 9]; todavia, como o espaço de busca é muito grande e consequentemente computacionalmente custoso, logo propomos adaptar e executar o encadeamento atual em GPUs. Do ponto de vista *software*, o MrBayes, programa de inferência Bayesiana que utilizaremos, já oferece suporte a GPUS [11]. Já em termos de *hardware*, o Instituto Butantan dispõe de uma servidora com as seguintes especificações técnicas:

- 4 processadores Intel Xeon Platinum, 2.1 GHz, cada um com 28 núcleos (56 threads) Turbo e 38 MB de cache;
- 1 TB de memória RAM, RDIMM;
- Placa NVIDIA Tesla P40 24GB Passive GPU (3.840 núcleos).

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 as monografias de Victor Raposo [3] e Gustavo Maciel [4] e os trabalhos de Débora Andrade-Silva e colegas [1, 2];

Atividade 2: Escrita do esboço do projeto de pesquisa;

Atividade 3: Estudo do código legado de Gustavo Maciel, disponível em um repositório GitHub (https://github.com/msreis/MITE); testes com esse código, utilizando dados brutos de EM de venenos de serpentes, seriam realizados na servidora do Instituto Butantan;

Atividade 4: Implementação de heurísticas e do procedimento de otimização descrito na figura 3; testes iniciais do novo procedimento;

Atividade 5: Implementação e testes do uso de Aprendizado Não-Supervisionado (mapas auto-organizáveis) para redução de dimensionalidade das matrizes;

Atividade 6: Testes com o uso combinado do procedimento de otimização e de Aprendizado Não-Supervisionado; análise dos resultados;

Atividade 7: Escrita da monografia do Trabalho de Formatura Supervisionado;

Atividade 8: Preparação e apresentação de pôsteres na Reunião Científica Anual do Instituto Butantan e na disciplina do Trabalho de Formatura Supervisionado.

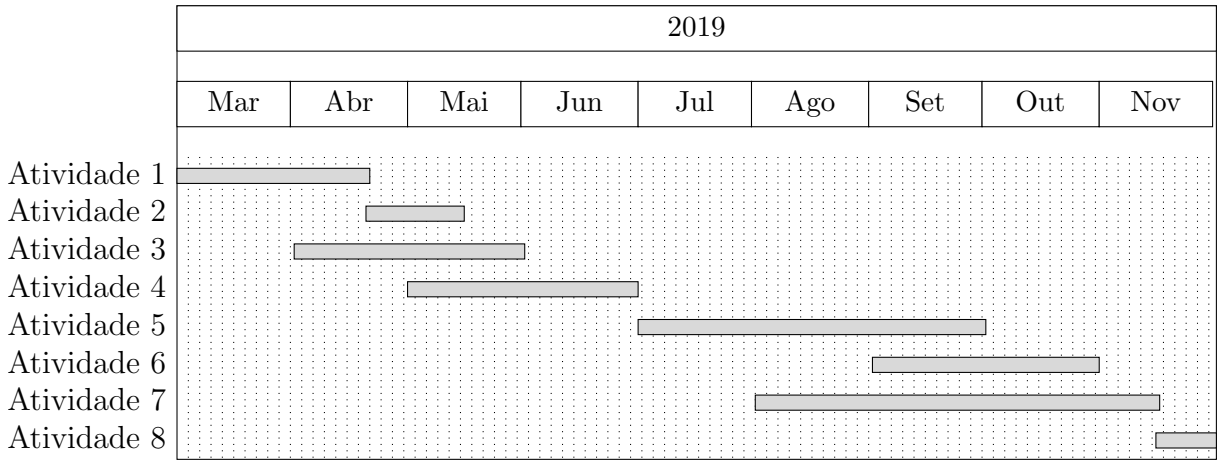

Tabela 1: Diagrama de Gantt contendo o cronograma de execução deste projeto proposto.

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

[2] D. Andrade-Silva, D. Ashline, T. Tran, A.S. Lopes, S.R. Travaglia Cardoso, M.S. Reis, A. Zelanis, S.M.T. Serrano, and V. Reinhold. Structures of N-Glycans of *Bothrops* venoms revealed as molecular signatures that contribute to venom phenotype in viperid snakes. *Molecular & Cellular Proteomics*, 17(7):1261–1284, 2018.

[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] Gustavo Mendes Maciel. Um método baseado em multirresolução para análise filoproteômica de venenos de serpentes. Technical report, Instituto de Matemática e Estatística, Universidade de São Paulo, 2019. Monografia de graduação em Computação.

[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] 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.

[7] 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.

[8] Scott Kirkpatrick, C Daniel Gelatt, and Mario P Vecchi. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[9] P. Pudil, J. Novovicová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[10] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, Jan 1982.

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