UNIVERSIDADE ESTADUAL PAULISTA “Júlio de Mesquita Filho”

Implementação de Análise Filogenética na Metodologia SEQUENCE SLIDER

Leonardo Ferreira de Andrade

Trabalho de Conclusão de Curso apresentado ao Instituto de Biociências, Universidade Estadual Paulista, “Júlio de Mesquita Filho” - UNESP, Campusde Botucatu, para a obtenção do título de Bacharel emFísica Médica.

BOTUCATU
2022

Leonardo Ferreira de Andrade

Implementação de Análise Filogenética na Metodologia SEQUENCE SLIDER

Monografia apresentada ao Instituto de Biociências da Universidade Estadual Paulista "Júlio de Mesquita Filho",Campus de Botucatu, para obtenção do título de Bacharel em Física Médica.

Orientador: Prof. Dr. Marcos Roberto de Mattos Fontes

Co-orientador: Dr. Rafael Junqueira Borges

Departamento de Biofísica e Farmacologia

BOTUCATU
2022

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Andrade, Leonardo Ferreira de.
Implementação de análise filogenética na metodologia Sequence Slider / Leonardo Ferreira de Andrade. - Botucatu, 2022

Trabalho de conclusão de curso (bacharelado - Física Médica) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Marcos Roberto de Mattos Fontes
Coorientador: Rafael Junqueira Borges
Capes: 20901003

1. Alinhamento de sequências (Bioinformática). 2. Biologia molecular. 3. *Python* (Linguagem de programação de computador)

Palavras-chave: Alinhamento de sequência; Biologia molecular estrutural; *Python*; Sequence Slider.

Agradecimentos

Gostaria de começar agradecendo ao universo por reger o meu caminho até aqui, a sanidade da minha cabeça, apesar de pouca, para trilhar tais rotas e por me permitir vislumbrar alguns futuros possíveis. Por me reservar surpresas boas e ruins que fazem com que eu cresça diariamente. Espero que nunca falte experiência, resiliência e empatia para que eu possa continuar lidando com o futuro desconhecido.

A minha família, que me possibilitou estudar todo esse tempo de maneira tranquila, sem que eu tivesse outras grandes preocupações, como estudar e trabalhar ao mesmo tempo. Em especial a minha mãe, Silvia Helena, quem me ajudou e me ajuda até hoje. Sem essa base eu poderia nem ter chego onde estou hoje, por isso sou grato.

As minhas amigas, que dividiram as dores e seguraram a barra quando eu achei que não ia conseguir. Que me ouviram, me aconselharam, me abrigaram e estiveram presentes nos melhores e piores momentos da graduação. Obrigado pelas risadas, pelos choros, pelos abraços, pelas festas e pelos almoços. Eu jamais conseguirei retribuir a altura. Queria dedicar esse trabalho especialmente a Laíza Brugnerotto por ser sempre meu porto-seguro e minha fiel escudeira por todos esses anos de curso.

Ao meu orientador e co-orientador, Marcos Fontes e Rafael Borges, respectivamente, por ter me cedido a vaga para estagiar junto deles e acreditado em mim quando eu não via mais saída. Obrigado pela mentoria e pela experiência de trabalhar com um projeto interessante.

Ao meu amigo, quem eu sei que importunei muitas vezes durante toda a graduação, Nivaldo Ferreira, por ter me ajudado com todas as questões burocráticas que envolvem a graduação como um todo e pelos vários puxões de orelha.

Por fim, mas não menos importante, ao meu namorado, Igor Neves, por todos os conselhos, todo o tempo disponível pra me ouvir falar sobre esse trabalho e toda visão tida por outra ótica que ele me apresentou. Sou muito grato mesmo.

RESUMO

A análise filogenética – também chamada de filogenia – é a relação evolutiva de um grupo de espécies, e pode ser determinada a partir de um conjunto de dados moleculares ou morfológicos. Com o tempo e com o desenvolvimento da tecnologia, a análise filogenética tornou-se uma ferramenta importante para coisas como a preferência por proteínas. Num grau molecular, o objeto de interesse são as proteínas, tais como metaloproteases, fosfolipases-A2 (PLA2) e porinas, algumas destas proteínas são sintetizadas em células endócrinas, por exemplo nas glândulas de veneno de cobras de diferentes famílias. Refletindo sobre a análise filogenética dentro da metodologia do software SEQUENCE SLIDER, sendo este capaz de organizar a integração de informações estruturais e genéticas sob uma nova forma de estrutura, sob esta ótica, a análise filogenética se mostra um mecanismo essencial para as análises feitas pelo software, sendo aliada importante que trona possível analisar uma amostra desconhecida e pelas suas características proteicas predizer quais serão seus possíveis aminoácidos correspondentes dentro de seus resíduos. Dito isso, o objetivo central do trabalho é integrar a análise filogenética de proteínas ao SEQUENCE SLIDER escrevendo um script na linguagem de programação python, cuja função será ler as sequências homólogas com a sequência da proteína padrão e calcular a frequência de cada aminoácido dentro de cada resíduo para todas as cadeias proteicas analisadas.

Palavras-chave: Biologia molecular estrutural; Alinhamento de Sequência; SEQUENCE SLIDER; Python.

ABSTRACT

Phylogenetic analysis – also called phylogeny – is the evolutionary relationship of a group of species, and can be determined from a set of molecular or morphological data. Over time and with the development of technology, phylogenetic analysis has become an important tool for things like protein preferences. On a molecular level, the object of interest are proteins, such as metalloproteases, phospholipases-A2 (PLA2) and porins, some of these proteins are synthesized in endocrine cells, for example in the venom glands of snakes of different families. Reflecting on the phylogenetic analysis within the methodology of the SEQUENCE SLIDER software, which is capable of organizing the integration of structural and genetic information under a new form of structure, from this perspective, the phylogenetic analysis proves to be an essential mechanism for the analyzes made by the software , being an important ally that makes it possible to analyze an unknown sample and, due to its protein characteristics, predict which will be its possible corresponding amino acids within its residues. That said, the main objective of the work is to integrate the phylogenetic analysis of proteins to the SEQUENCE SLIDER by writing a script in the python programming language, whose function will be to read the homologous sequences with the standard protein sequence and calculate the frequency of each amino acid within each residue for all analyzed protein chains.

Keywords: Structural molecular biology; Sequence Alignment; SEQUENCE SLIDER; Python.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

1.1 Análise Filogenética

A filogenia é uma área da biologia que estuda as relações evolutivas dos organismos (espécies) (VIANA, 2007). Graur e Li (2000) mostram que estudos filogenéticos apontam 3 objetivos, sendo estes: a correta reconstrução das semelhanças genéticas entre entidades biológicas; a estimação do período de divergência entre organismos (a determinação do tempo de formação dessas espécies após compartilharem um antepassado em comum) e o detalhamento da sequência de eventos entre as diferentes linhagens evolutivas.

A partir da inferência filogenética pode-se determinar, através de um conjunto de dados moleculares ou morfológicos, quais as relações evolutivas de um conjunto de espécies. Estas relações evolutivas são apresentadas usualmente em uma forma e árvore, conhecida como árvore filogenética ou cladograma (TICONA, 2008) (figura 1).

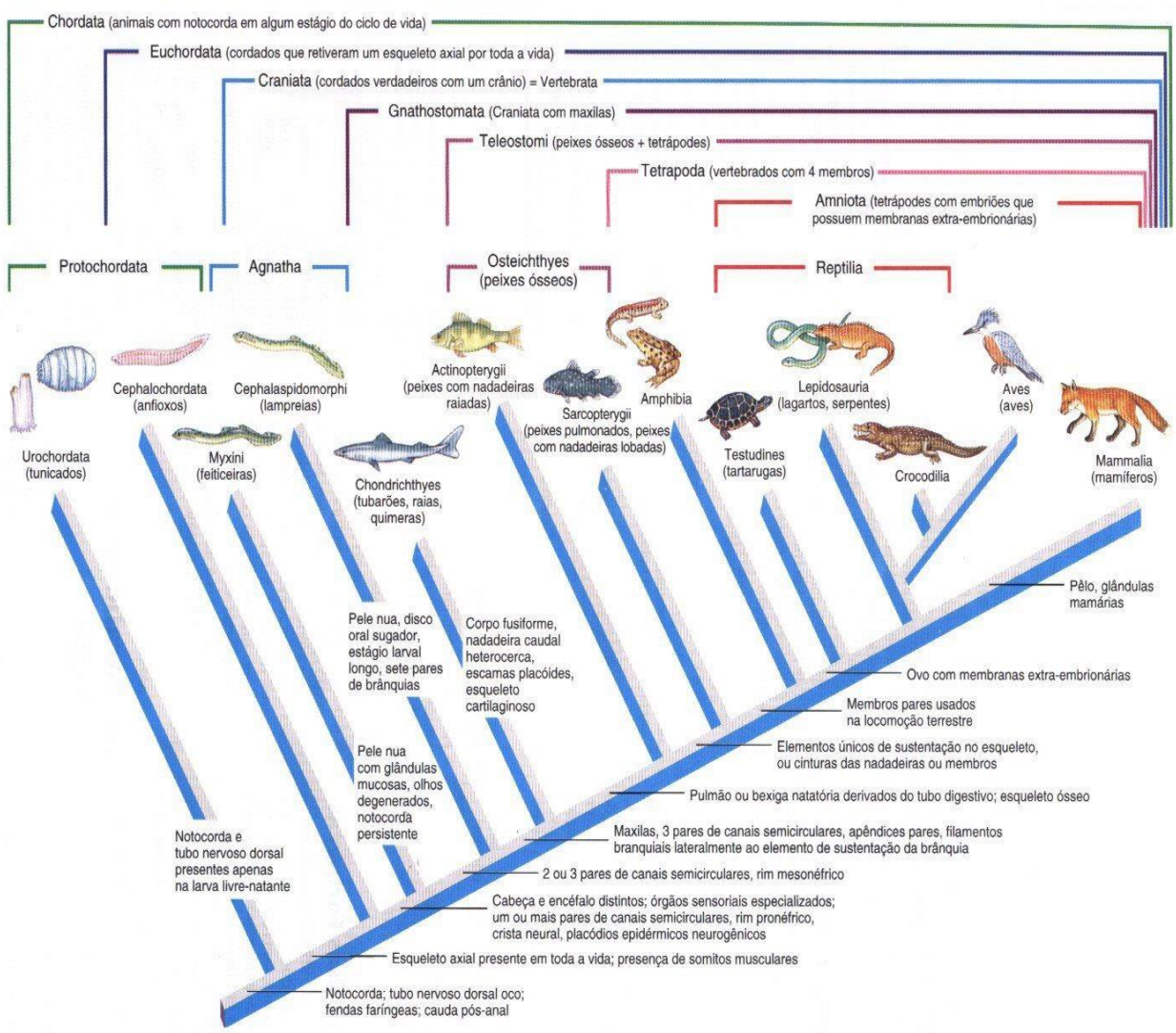

Figura 1 - Cladograma que evidencia características primitivas e derivadas em diferentes grupos

Na análise filogenética, os cladogramas são feitos para representar a história evolutiva ou as relações entre diferentes espécies, organismos ou características biológicas, como genes, proteínas, órgãos, que se desenvolveram a partir de um ancestral comum (KIMBALL, 2013). Utilizando as árvores filogenéticas é possível se estudar os possíveis relacionamentos entre as espécies. Tais árvores são inferidas a partir dos dados disponíveis de tal forma que espécies "mais próximas" nos tenham correspondentes mais próximos na árvore (VIANA, 2007).

Em uma árvore filogenética, folhas representam espécies, populações, indivíduos ou genes, os quais podem ser conectados a nós por ramos (ramificações externas) (SCOTT & BAUM, 2016). Os ramos representam a transmissão de informações genéticas entre os

descendentes, e os comprimentos dos ramos representam as mudanças ou divergências genéticas. O grau de divergência é geralmente estimado usando o número médio de substituições de nucleotídeos por sítio (VAN DE PEER & SALEMI, 2009).

O DNA como material genético pode atualmente ser sequenciado de forma fácil, rápida e barata, e os dados obtidos do sequenciamento genético são muito ricos e específicos. Além disso, estimativas morfológicas podem ser utilizadas para inferir o desenvolvimento evolutivo, especialmente na ausência de material genético, como em fósseis (SCOTT & BAUM, 2016). No caso deste trabalho, o uso da análise filogenética é focado no sequenciamento de estruturas proteicas em diferentes organismos. Analisando as mutações das diferentes estruturas biológicas é possível explicitar onde existe a preservação da sequência primitiva, por meio da utilização de técnicas, como a espectrometria de massa, por exemplo.

1.2 Venenos e a Filogenia

Frequentemente a composição dos venenos tem sido considerada com um reflexo da filogenia das serpentes, com importantes implicações para a produção de antivenenosos (SOUZA, 2014). Diversos estudos tentam explicar o processo evolutivo da função poeçonhenta nas serpentes, porém, ainda é um assunto para o qual faltam muitas respostas.

Estudos filogenéticos mostraram que a cerca de 80 a 60 milhões de anos atrás foi quando surgiu a glândula de veneno, local de produção das toxinas, e possui uma extensa evolução paralela com os sistemas de distribuição e composição de venenos (PAHARI et al., 2007). A capacidade de injetar veneno parece ter surgido na Superfamília Colubrídae (CADLE, 1983 apud SOUZA, 2014), onde a presença dos componentes tóxicos passíveis de inoculação, permitiram às serpentes colubrídaes uma excepcional diversificação, conferindo-lhes uma vantagem adaptativa para a captura de presas e contribuindo também para o aumento de suas taxas de especiação (PYRON *et al.*, 2011).

A partir do surgimento da função venenosa, as serpentes peçonhentas teriam divergido no início da evolução em diferentes linhagens, seguindo caminhos distintos na evolução do aparelho de veneno (KOCHVA, 1987;), o que explica a diversidade de padrões que pode ser observada entre os membros das famílias Elapidae, Viperidae e Atractaspididae, se tratando do modo de inoculação dos venenos.

A composição dos venenos de serpentes pode variar bastante em por uma série de fatores tais quais, as características filogenéticas da espécie, sazonalidade, dieta, sexo, distribuição geográfica e estágio ontogenético das serpentes (FURTADO, et al., 2010). 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.

Com o passar dos tempos e o desenvolvimento tecnológico, a análise filogenética se tornou uma ferramenta importante na predileção, entre outras coisas, das proteínas. Algumas proteínas como os venenos foram amplamente beneficiadas devido à escassez de informações mais aprofundadas dos diferentes tipos de venenos (FRY, 2005).

As proteínas escolhidas para este trabalho foram as metaloproteases, as fosfolipases-A_2 e as porinas pelas suas características naturais. As metaloproteases, ou metaloproteinases, que são abreviadas como (MMPs), são uma família de íons endopeptidases dependentes que promovem a degradação da matriz extracelular, também conhecidas como proteínas da matriz. As MMPs estão envolvidas em uma variedade de eventos biológicos porque podem potencialmente afetar o comportamento celular por meio de ações como a clivagem de proteínas que permitem a adesão intercelular, liberação de moléculas bioativas na superfície celular ou pela clivagem de moléculas presentes na superfície celular, em sinal no ambiente externo (ARAÚJO, *et al.*, 2011). Essas são as principais proteínas analisadas no trabalho. As

fosfolipases A_2 desempenham um papel importante em uma variedade de funções celulares, incluindo a manutenção de fosfolipídios celulares, produção de prostaglandinas e leucotrienos, tradução de sinais, proliferação celular e contração muscular (SANTOS FILHO, 2009). Por fim, as porinas, são proteínas transmembranares, responsáveis pela difusão de pequenos metabolitos como açucares, aminoácidos e íons. Possuem folhas betas antiparalelas formando um cilindro. Algumas porinas têm especificidade de substrato além de propriedades de difusão, como por exemplo as maltoporinas e OprB.

1.3 SEQUENCE SLIDER

O SEQUENCE SLIDER (SLIDER) é um método de avaliação da cadeia lateral, capaz de fornecer uma nova estrutura para a tarefa abstrusa de organizar a integração de informações estruturais e genéticas (BORGES, *et al*., 2022). O SLIDER foi primeiramente utilizado por Borges (2020) dentro da etapa *ab initio* pelo método ARCIMBOLDO, o qual utiliza substituição molecular para identificar pequenos fragmentos em unidades assimétricas (MCCOY, *et al*., 2007), revelando assim o resto da estrutura pela modificação de densidade e

interpretação automatizada de mapas (THORN & SHELDRICK, 2013). Na figura 1 a seguir é possível a evidenciar como são feitas as entradas e saídas disponibilizadas pelo método.

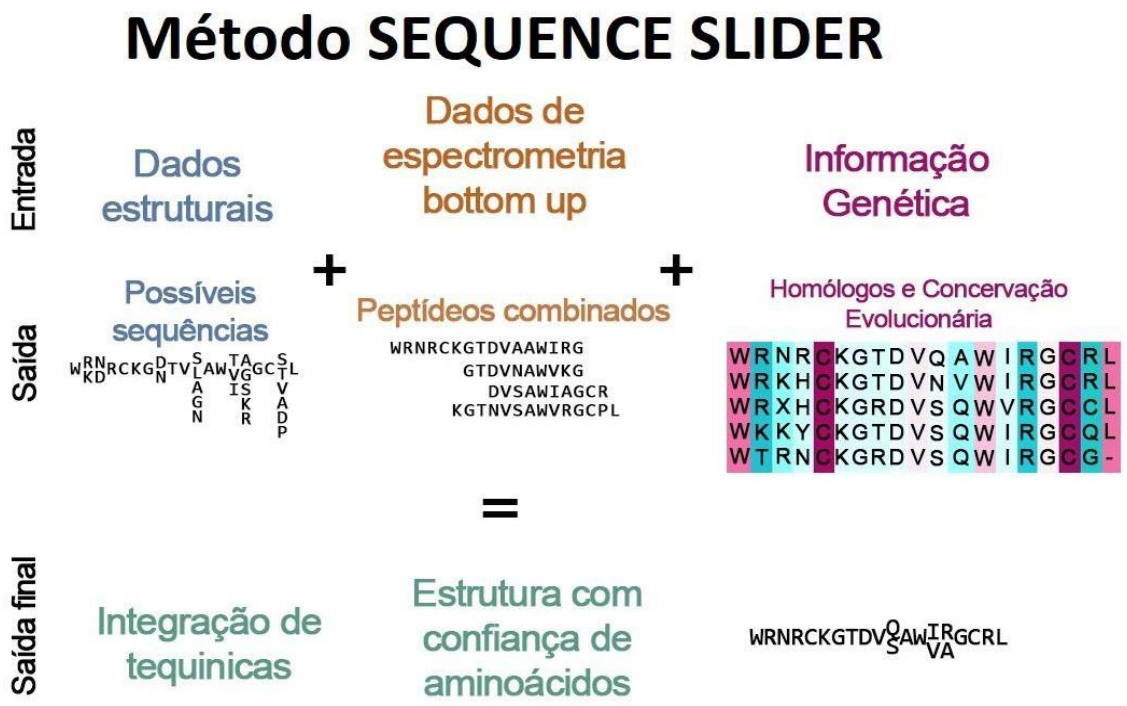

Figura 2 - Equação de como o SLIDER SEQUENCIAL utiliza as entradas e disponibiliza as informações de saída.

O SLIDER utiliza informações disponíveis de sequência, previsão de estrutura secundária e alinhamento entre homólogos remotos para construir os átomos de cadeia lateral. A discriminação nas estatísticas globais de concordância entre dados e modelos pode revelar hipóteses de sequência verdadeiras (BORGES, 2020).

O SLIDER tem a capacidade de equilibrar diferentes fontes de informação para atribuir um único aminoácido o qual quando identificado indubitavelmente, gera múltiplas possibilidades ou até mesmo desconhecidas, quando a evidência não é conclusiva, com uma abordagem que deriva hipóteses sequenciais de dados e impede a propagação de viés a partir de informações conhecidas *a priori* (BORGES, et *al.*, 2022). Na figura 2 a baixo, podemos ver

como o SLIDER integra dados estruturais, espectrometria de massa e filogenética para atribuição de sequência e sua heterogeneidade em amostras complexas.

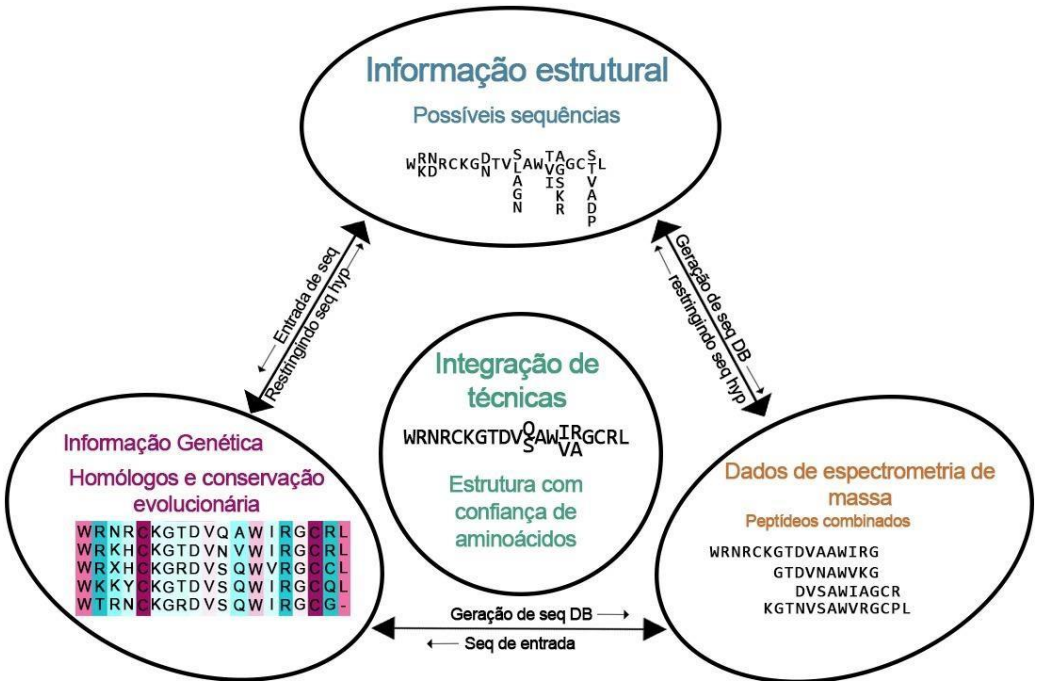

Figura 3 - Esquema de como o SLIDER SEQUENCIAL integra dados estruturais, espectrometria de massa e filogenética para atribuição de sequência.

Borges (*et al*., 2022) menciona que para rodar o SLIDER os requisitos computacionais dependem da resolução dos dados cristalográficos e do número de resíduos na unidade assimétrica, os cálculos podem variar de 5 e 20 minutos por resíduo, porém, com o método de aproveitamento no multiprocessamento, esse tempo é dividido pelo número de núcleos físicos disponíveis, a partir de teste inicial pode-se estimar o uso da memória dado o número de núcleos, os quais podem ser reduzidos para evitar sobrecarga de memória RAM, as análises em espectroscopia de massa e filogenéticas levam cerca de dez minutos por conjunto de dados. A implementação do método no trabalho em questão foi utilizada para integrar determinação estrutural, espectrometria de massa e informações genéticas para resolver a heterogeneidade em amostras complexas purificadas de fontes naturais construindo probabilidades de aminoácidos para cada base de resíduo (LIU, *et al.*, 2020).

2. JUSTIFICATIVA E HIPÓTESE

Uma vez que podem existir várias isoformas de uma determinada proteína purificadas a partir de fontes naturais, estes dados podem ser muito importantes em um estudo estrutural. Diversos estudos têm ignorado esta informação supondo que se trata de proteínas idênticas. Isso é especialmente comum no caso do veneno de serpentes, onde há uma mistura complexa de componentes proteínas. Portanto, a implementação de uma ferramenta de análise filogenética, que estima a conservação e variabilidade de cada resíduo na proteína de interesse, pode ajudar a calcular a probabilidade de atribuição correta dos aminoácidos, conforme mostrado pelo SEQUENCE SLIDER.

3. OBJETIVOS

3.1 Objetivos Gerais

Integrar análises filogenéticas ao SEQUENCE SLIDER.

3.2 Objetivos Específicos

Escrever um *script* em linguagem *Python* que leia como entrada um arquivo de alinhamento múltiplo de sequências de proteínas homólogas a uma sequência padrão gerada pela ferramenta CONSURF e calcular frequências de aminoácidos em cada resíduo da proteína analisada e essas sejam exibidas em tabelas.

4. MATERIAIS E MÉTODOS

4.1 Aplicação do SEQUENCE SLIDER para designação da sequência da lisozima

SEQUENCE SLIDER foi usado para compostos naturais para atribuir a sequência de lisozima. A lisozima de *Gallus gallus* foi adquirida pela liofilização de Sigma-Aldrich. Os cristais foram obtidos pelo método de difusão de vapor em gota suspensa a 18°C

(MCPHERSON, 2014). Cristais de lisozima foram crescidos em gotas de cristalização compostas por 1µl de solução de proteína (50 mg/ml) e 1µl de solução precipitante, equilibrada contra a solução reservatório de 1,2 M MgCl2 e 0,1 M Tris HCl, pH 7,6.

Os cristais foram montados em alças de nylon e resfriados rapidamente em nitrogênio líquido. Os dados de difração foram coletados na linha de luz MX2, Laboratório Nacional de Luz Síncrotron (LNLS, Campinas, Brasil). O conjunto de dados foi indexado, integrado e dimensionado usando XDS (KABSCH, 2010). A estrutura foi resolvida por substituição molecular com PHASER (MCCOY, et al., 2007) usando coordenadas de lisozima (6G8A).

Um modelo de proteína com boa concordância com o mapa experimental baseado em indicadores globais em conformidade com a norma, R-Free abaixo de 25% e score de Molprobity abaixo da resolução dos dados, foi usado como ponto de partida. Os rotâmeros para cada um dos 20 aminoácidos possíveis foram gerados usando a função coot 'auto-fit-best rotâmero' para cada resíduo no modelo de proteína. Moléculas de água dentro de 2,5Å foram removidas com o comando coot ‘delete_atoms’.

Todos os átomos dentro de 5Å de distância do resíduo de foco foram refinados usando coot 'refinar resíduos' (EMSLEY & COWTAN, 2004). Os átomos da cadeia lateral tiveram seus fatores B ajustados para 30, e seu coeficiente de correlação no espaço real (RSCC) foi calculado em relação aos mapas de omissão de phenix.polder (LIEBSCHNER, *et al.*, 2017). Essas etapas foram realizadas pelo coorientador Dr. Rafael Junqueira Borges.

As imagens foram geradas usando COOT (EMSLEY & COWTAN, 2004) e Raster3D (MERRITT; BACON, 1997). O refinamento foi realizado com phenix.refine (AFONINE, *et al.*, 2012), e o modelo foi construído manualmente com Coot (EMSLEY, *et al.*, 2010) usando os mapas σA-weighted Fobs-Fcalc e 2Fobs-Fcalc.

4.2 Diferentes parametrizações do CONSURF

Variações da parametrização do CONSURF foram realizadas sobre as sequências das

toxinas 1, 2 e 3 de origem recombinante.

4.3 Sequências selecionadas para análise filogenética

Junto com o aprendizado da nova linguagem foram reunidos 209 códigos, os quais podem ser encontrados no anexo A ao fim do trabalho, do Protein Data Bank (PDB). Este banco de dados foi estabelecido como o primeiro recurso de dados digitais de acesso aberto em biologia e medicina e hoje é o principal recurso do mundo para dados experimentais vitais para a descoberta científica (BERMAN, *et al.,* 2000), de diferentes origens, tais como metaloproteases, fosfolipases A_2 e porinas.

Depois de reunidos, esses códigos foram inseridos no software online CONSURF, o qual é um servidor web bem estabelecido para calcular a conservação evolutiva de posições de aminoácidos em proteínas usando inferência Bayesiana empírica de suas estruturas e sequências (ASHKENAZY, *et al.*, 2016). O CONSURF então promove o alinhamento de sequências múltiplas para que seja possível, a partir do dado código PDB, obter várias outras sequências de proteínas homólogas a primeira que foi inserida, que no caso é o código PDB.

E assim foi feito para as 209 estruturas analisadas. Os parâmetros inseridos no *software* para a caracterização dos alinhamentos múltiplos foram os seguintes: o algoritmo de pesquisa homólogo foi feito em HMMER, o número de iterações foi de 1 iteração, o valor residual foi de 0.0001, o banco de dados proteico utilizado foi o UNIREF-90, cujo banco de dados agrupa sequências e subfragmentos com 11 ou mais resíduos e pelo menos 90% de identidade de sequência entre si (de qualquer organismo) em uma única entrada UniRef, mostrando sequências representativas.

Assim, a análise dos homólogos pelo CONSURF foi feita automaticamente por eles, utilizando as seguintes requisições: foram demandadas 300 sequências proteicas homólogas que são embaralhadas utilizando o valor residual inserido com o máximo de identidade entre as sequencias de 100% e o mínimo de 35%, para maior identidade com a sequência padrão.

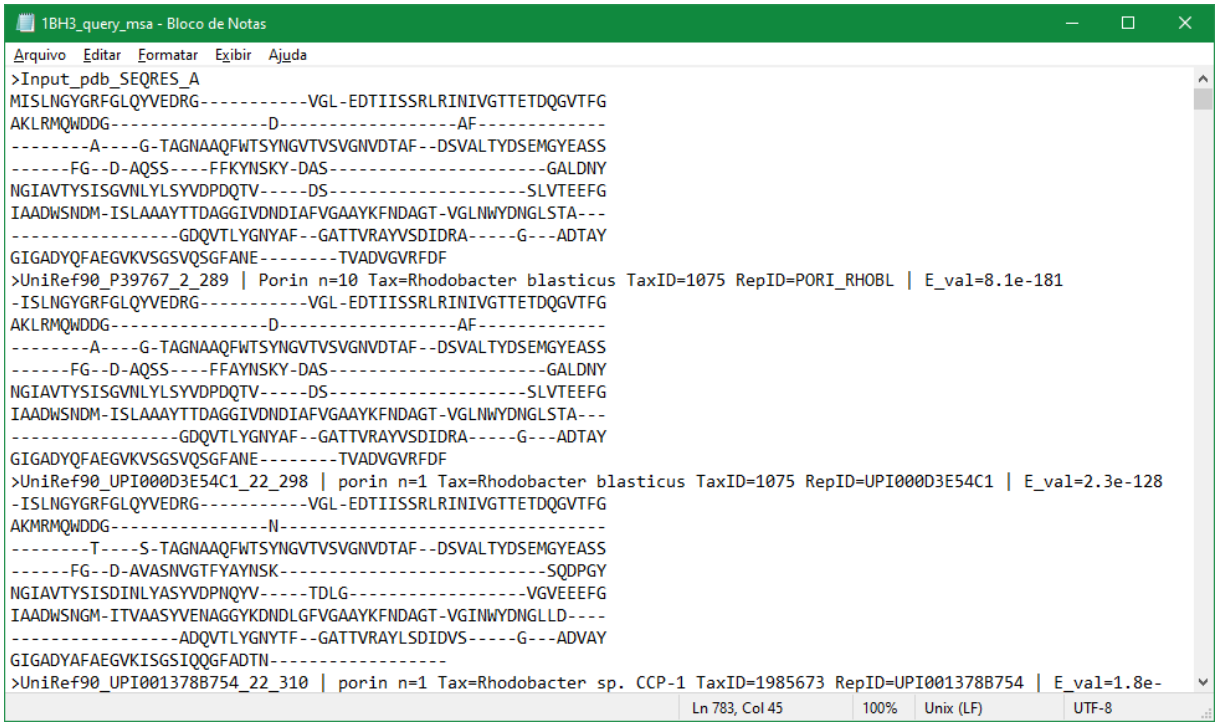

Figura 4 - Exemplo de arquivo gerado pelo CONSURF para a proteína porina 1BH3

5. RESULTADOS

5.1 Descrição do Método

Com o início do projeto foram sendo apresentados algumas ferramentas as quais iriam ser utilizadas para o trabalho. Ao primeiro contato, tiveram espaço os softwares COOT (Crystallographic Object-Oriented Toolkit) (EMSLEY, *et al.*, 2010) e o PHENIX (Python-based Hierarchical Environment for Integrated Xtallography) (LIEBSCHNER & AFONINE, *et al*., 2019), os quais foram utilizados para aprender mais sobre a estrutura das sequências e a aplicação do método SEQUENCE SLIDER (ASHKENAZY, *et al.*, 2016) ao processo.

Primeiro foi feita uma modelagem da estrutura da lisozima previamente elucidada em nosso laboratório no programa COOT para que houvesse a revisão de alguns aminoácidos presentes na sequência. Após o rearranjo dos aminoácidos toda a estrutura foi refinada com o auxílio do PHENIX. Após o refino o método foi reproduzido novamente e os dados refinados uma última vez.

Com a compreensão de algumas ferramentas de ambos os aplicativos, a segunda parte foi preencher uma cadeia de aminoácidos do zero, baseado também da sequência de aminoácidos da lisozima, utilizando uma lista de frequências geradas previamente pelo método SEQUENCE SLIDER.

Com o uso da lista e a nuvem eletrônica gerada pelos mapas de densidade eletrônica no software, a construção da sequência foi realizada. O resultado foi processado no PHENIX para que houvesse novamente o refinamento dos dados.

Após finalizada a parte de reconhecimento das técnicas e ferramentas, iniciou-se o aprendizado da linguagem de programação PYTHON para que pudessem ser desenvolvidos alguns *scripts* que facilitassem o trabalho computacional.

Com todos os parâmetros escolhidos foram feitos os downloads das sequências de cada código PDB analisado. O SLIDER foi primeiro cuidadosamente calibrado usando um conjunto de dados cristalográficos bem conhecido de lisozima de clara de ovo e posteriormente refinado usando a proteína central proteoglicana de sulfato de heparano (BaM) específica da membrana basal.

Posteriormente, aplicamos para elucidar uma metaloprotease (BmooMP-I) (SALVADOR, *et al.*, 2020.), uma proteína tipo PLA_2 de *Bothrops moojeni* (MjTX-I) e uma toxina da serpente *Crotalus durissus collineatus* (CBCol), revelando sequências ainda não vistas na literatura. Fornecemos à comunidade um método para atribuir consistentemente aminoácidos integrando informações disponíveis de macromoléculas purificadas de fontes naturais.

Assim, foi escrito um *script*, em linguagem PYTHON, que lê os arquivos da seguinte forma: analisa a sequência padrão e armazena-a. Posteriormente lê as outras sequências, padronizando-as de forma que todas fiquem uniformes e possam ser analisadas.

Verificar em cima de cada resíduo de cada sequência e compara a presença do

aminoácido presente com o aminoácido que reside na sequência padrão. Depois de finalizado o processo total de comparação, o *script* cataloga as frequências de cada aminoácido existente com cada resíduo analisado.

Ou seja, armazena a informação de quantas vezes cada aminoácido apareceu em cada resíduo com base na sequência padrão. Por fim, é gerada uma tabela com essas frequências e depois uma lista com as maiores frequências para cada resíduo da amostra.

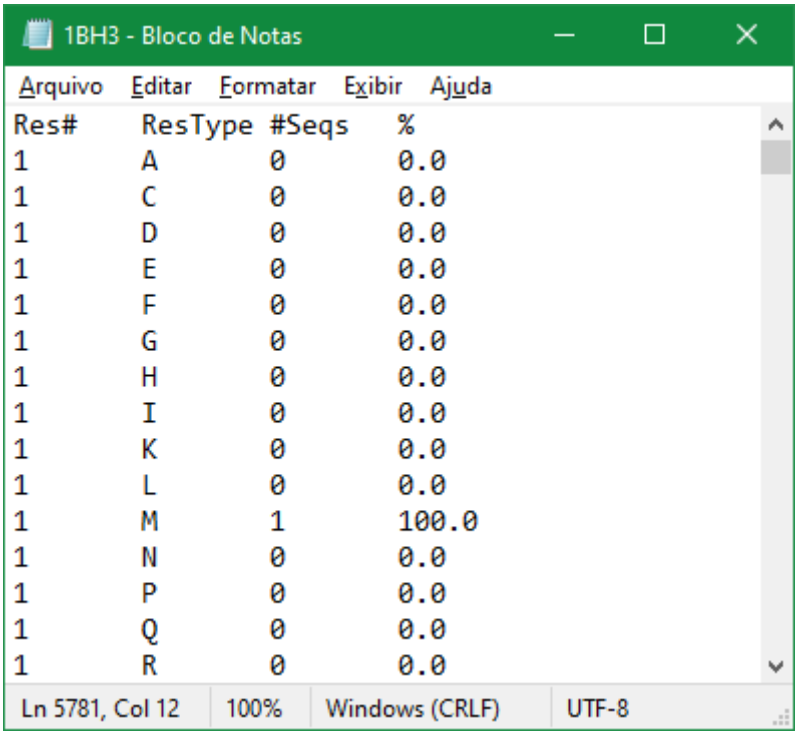

Figura 5 - Exemplo de arquivo de frequências primário para a protína porina 1BH3

5.2 Resultados da Análise do Coot+Phenix

Depois de construir e ordenar todo o sistema no COOT e incluir o mapa de nuvem eletrônica, a sequência da lisozima foi refinada para checar se o alinhamento dos aminoácidos estava o mais próximo do esperado. Para isso foram utilizando os arquivos .pdb e o .mtz originais junto ao software PHENIX. Seguindo a orientação do aplicativo, a ordem seguida a partir do menu foi clicar no *Calculate*, depois no *Scripting* e por fim no *Scheme* para que todo

o processo fosse concluído. Os resultados já foram gerados pelo software em formato de tabela e apresentados a seguir.

Tabela 1 - Primeiros resultados de refinamento obtidos pelo software PHENIX

	Starting	Final
R-Work	0,2521	0,2005
R-Free	0,2720	0,2564
Bonds	0,038	0,007
Angles	3,003	0,901

Após uma análise dos resultados, foi visto que esses poderiam ser melhorados para que ficassem mais próximos do esperado e o processo foi refeito. Com os aminoácidos realocados em seus lugares respectivos dessa vez, os dados foram introduzidos ao software PHENIX para que fossem processados. Por fim, depois de refinados, os resultados finais foram descritos da seguinte forma.

Tabela 2 - Resultados finais do obtidos pelo software PHENIX

	Starting	Final
R-Work	0,2618	0,2029
R-Free	0,2807	0,2578
Bonds	0,042	0,005
Angles	3,013	0,907

5.3 Resultados da Programação

Depois de obtidos os 209 arquivos Consurf, com mais de 300 sequências em cada, esses

foram analisados por um script em linguagem PYTHON, onde a finalidade era priorizar as maiores frequências de cada aminoácido em sua respectiva posição. Assim, para cada posição das diferentes sequências foram escolhidos os aminoácidos de maiores frequências como é possível observar na imagem a seguir.

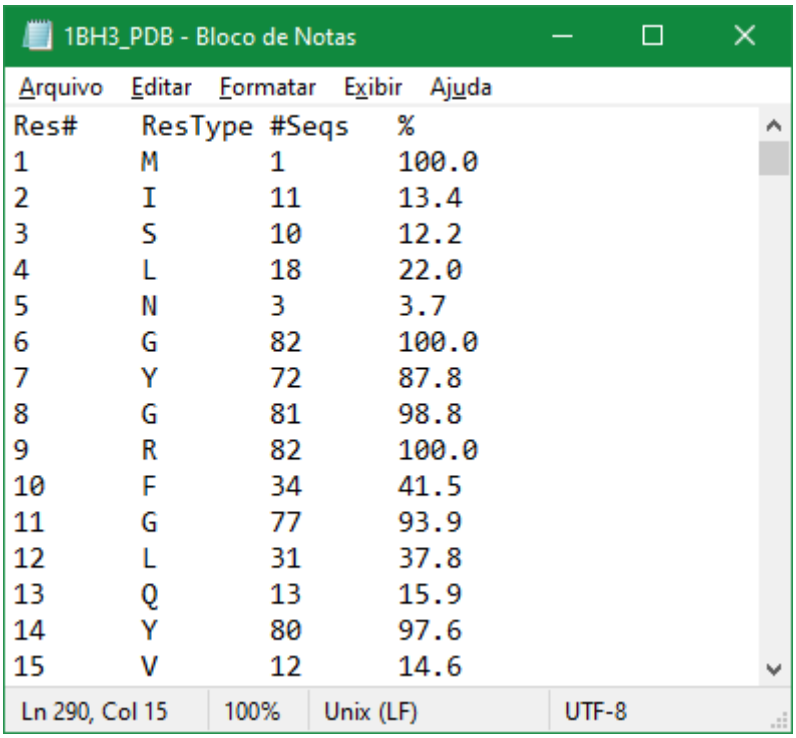

Figura 6 - Exemplo de tabela de frequência final para a proteína porina 1BH3

Com esses arquivos finais contendo as principais frequências de cada sequência proteica, uma segunda parte do script contou cada frequência de cada posição de cada sequência dentre todos os arquivos e destinou-as para diferentes ranges dentro de um gráfico de colunas. O gráfico 1 a seguir foi utilizado para visualizar a contagem de todos os aminoácidos e suas frequências totais, sendo cada uma dessas contabilizadas e destinadas para cada coluna característica. Os gráficos foram gerados pelo software do Anaconda (ANON, 2020) na ferramenta do Jupyter Notebook (KLUYVER, *et al.*, 2016)

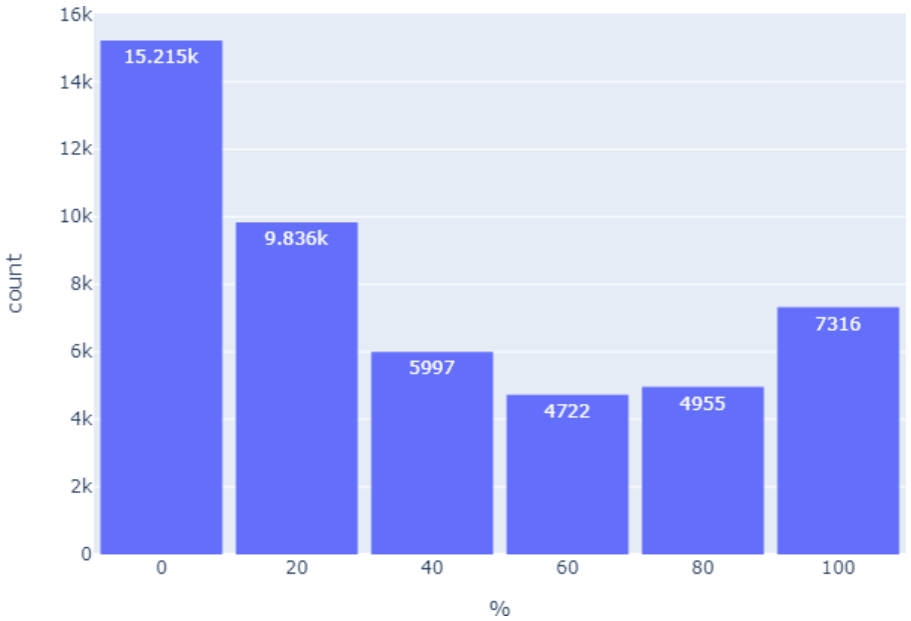

Figura 7 - Gráfico inicial com todos o total de frequências obtidas e a representação de cada uma dentro do intervalo correspondente

Depois da visualização cada range foi rearranjado e distribuído em uma proporção de dez em dez porcento entre si, portanto cada ponto dentro das colunas representa uma frequência específica de uma posição pertencente a uma sequência de aminoácidos.

Por fim, foi gerado um gráfico para exemplificar a média e o desvio padrão dos resultados de cada um dos 210 arquivos analisados. O gráfico a seguir, Figura 6, é um gráfico de pontos flutuantes que foi usado para exemplificar a média e desvio padrão de cada arquivo final analisado. A média é calculada somando todos os valores em um conjunto de dados e dividindo pelo número de elementos desse conjunto. Já o desvio padrão é uma medida de quão espalhado é um conjunto de dados. Ou seja, o desvio padrão indica quão uniforme é o conjunto de dados.

Figura 8 - Gráfico de pontos flutuantes de média e desvio padrão das frequências obtidas para cada arquivo analisado

6. DISCUSSÃO

Os resultados do PHENIX podem ser interpretados de forma onde alguns resíduos podem estar com as cadeias laterais flexíveis, ou seja, em parte das moléculas do cristal estão em uma conformação, em outras moléculas estariam em outra conformação, isso é conhecido como dupla ocupação.

Ao invés das cadeias laterais, poderia ser a cadeia principal também, podendo estar tão flexível que seria impossível de modelá-las. As regiões próximas das extremidades, N e C-terminal, continuam a ser mais flexíveis por conta de terem suas extremidades sem ligações. Outra explicação se daria pelos grupos de aminoácidos Treonina (Thr)/Valina (Val), Aspartato (Asp)/Asparagina (Asn)/Leucina (Leu), Glutamina (Gln)/Glutamato (Glu) são parecidos, sendo a única diferença sendo o Carbono (C)/Oxigênio (O)/Nitrogênio (N).

Como na densidade eletrônica observamos os elétrons dos átomos, a diferença entre o

número de elétrons desses átomos é muito pequena, portanto, quando não temos dados da resolução atômica (1$\mathring{A}$) não é possível diferenciá-las.

Em relação ao método de criação de alinhamento de sequência usado pelo CONSURF, é usado o MAFFT-L-INS-i (JOHN, *et. al*, 2019), que também é muito preciso. O método de cálculo utilizado é Bayesiano. Estes métodos demonstraram melhorar significativamente a precisão das estimativas de pontuação de conservação em comparação com os métodos de máxima verossimilhança, especialmente quando calculados com um pequeno número de sequências.

Outra vantagem da abordagem Bayesiana é que os intervalos de confiança são atribuídos a cada pontuação de conservação evolutiva inferida (LANDAU, *et. al*, 2005). O modelo de substituição evolucionária, que foi o último parâmetro para iniciar o download, foi o melhor modelo para cada conjunto de alinhamentos, sendo escolhido automaticamente pelo próprio CONSURF. É o modelo padrão para tal análise.

Com relação ao *script*, a escrita e a lógica deste foi a parte mais difícil de executar. A linguagem PYTHON apesar de parecer simples tornou esta parte mais trabalhosa, mas apesar das dificuldades a programação se mostrou positiva uma vez que o *script* se fez funcional.

Todos os arquivos foram lidos, suas frequências aferidas e seus resultados dispostos como foi visto anteriormente. Foram aproximadamente 48.041 resíduos analisados, dentre todos os arquivos. Acredita-se que a maioria destes se localizou no range de 0~10% pois há algumas lacunas dentro das sequências que são preenchidas em algumas sequências homólogas, porém não em todas.

Uma grande parte das frequências ficou dentro do range de 90~100% mostrando que há veracidade na estrutura ancestral da proteína. Os valores das frequências abaixo desta linha evidenciam a variedade presente na linha temporal proteica, mas mesmo assim tornando possível predizer qual em qual resíduo está cada aminoácido para diferentes proteínas.

7. CONCLUSÃO

Esse trabalho pretendeu entender a implementação da análise filogenética na metodologia SEQUENCE SLIDER para elucidar a existência de várias isoformas de proteínas purificadas a partir de fontes naturais, as quais muitas delas têm sido ignoradas na suposição de que são proteínas previamente caracterizadas, especialmente comum no caso do veneno de serpentes, a partir das análises de 209 arquivos de séries de proteínas padrão comparados com 300 sequências proteicas homólogas a estes para evidenciar as suas ancestralidades.

Para se atingir uma melhor compreensão para integrar as análises filogenéticas ao SEQUENCE SLIDER, definiu-se dois objetivos específicos. O primeiro foi aprender como as ligações operavam numa sequência em um aspecto molecular. Verificou-se que as ligações se fortificaram melhor na conformação final com o R-Free mais próximo do esperado. Depois, foram captados os arquivos que seriam usados, o código que seria usado para análise das sequências foi escrito e essas foram ponderados. A análise permitiu concluir que o experimento se mostrou factível de modo que os aminoácidos principais na cadeia sempre se repetem e evidenciam a frequência daqueles mais comuns para os menos comuns.

Com isso, a hipótese do trabalho de uma ferramenta de análise filogenética, que estima a conservação e variabilidade de cada resíduo na proteína de interesse, pode ajudar a calcular a probabilidade de atribuição correta dos aminoácidos, conforme mostrado pelo SEQUENCE SLIDER se confirmou devido os resultados obtidos pelo software.

Sendo assim, a implementação da análise filogenética é um mecanismo essencial para as análises de amostras naturais desenvolvidas pelo SEQUENCE SLIDER e trazem certeza maior para alguns tipos de sequências de proteínas que ainda estão à espera de serem descritas.

Os instrumentos de coleta de dados permitiram um primeiro contato e aprendizado sobre o funcionamento de algumas ferramentas, como o COOT, PHENIX e CONSURF, plataformas estas muito importantes para o desenvolvimento da pesquisa na área da biologia molecular.

Finalmente, a linguagem PYTHON trouxe uma nova ótica, mais facilitada de análise de dados e despertou o interesse de continuar desenvolvendo as habilidades adquiridas no processo.

O processo autônomo de aferição dos dados torna as características citadas no trabalho ligadas de forma intrínseca ao aprendizado de máquinas utilizado pelo João Bruno em seu trabalho de mestrado com o co-orientador deste projeto, Dr. Rafael Borges. O aprendizado de máquinas, que é proveniente do inglês *machine learning*, é um método de análise de dados que cria automaticamente modelos de análise. É um ramo da inteligência artificial baseado na ideia de que os sistemas podem aprender com os dados, reconhecer padrões e tomar decisões com o mínimo de intervenção humana.

Em pesquisas futuras, espera-se melhorar a captação de dados do CONSURF, que demorou muito em relação ao tempo desta coleta, por serem muitos arquivos e o processo ser feito de um em um. Um direcionamento útil é ir escrevendo bloco por bloco do *script* a ser desenvolvido para que os comandos funcionem adequadamente e os erros possam ser corrigidos rapidamente.

8. REFERÊNCIAS BIBLIOGRÁFICAS

AFONINE, P. V. *et al.* **Towards Automated Crystallographic Structure Refinement with Phenix**. refine. Acta Crystallographica Section D: Biological Crystallography, v. 68, n. 4, p. 352-367, 2012.

ANON. **Anaconda Software Distribution**, Anaconda Inc. Available at: https://docs.anaconda.com/, 2020.

ARAÚJO, R. V. de S.; SILVA, F. O.; MELO-JÚNIOR, M. R.; PORTO, A. L. F. **Metaloproteinases: aspectos fisiopatológicos sistêmicos e sua importância na cicatrização.** Revista de Ciências Médicas e Biológicas, [S. l.], v. 10, n. 1, p. 82–88, 2011. DOI: 10.9771/cmbio.v10i1.5470. Disponível em: https://periodicos.ufba.br/index.php/cmbio/article/view/5470.

ASHKENAZY, H.; ABADI, S.; MARTZ, E.; CHAY, O.; MAYROSE, I.; PUPKO, T.; BEN-TAL, N. **ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules.** Nucl. Acids Res, 2016

BERMAN, H. M.; WESTBROOK, J.; FENG, Z.; GILLILAND, G.; BHAT, T. N.; WEISSIG, H.; SHINDYALOV, I. N.; BOURNE, P.E. **The Protein Data Bank**. Nucleic Acids Research, 235-242, 2000.

BORGES, R. J.; MEINDL, K.; TRIVIÑO, J.; SAMMITO, M.; MEDINA, A.; MILLÁN C.; ALCORLO, M.; HERMOSO, J.A.; FONTES, M. R. M; USÓN, L; **SEQUENCE SLIDER: Expanding Polyalanine Fragments for Phasing with Multiple Side-Chain Hypotheses**. Acta Crystallogr D. Struct. Biol. 2020.

BORGES, R. J.; SALVADOR, G.H.M.; PIMENTA D.C.; DOS SANTOS L.D.; FONTES M.R.M.; USÓN, I. **SEQUENCE SLIDER: Integração de Dados Estruturais e Genéticos para Caracterizar Isoformes de Fontes Naturais**. SEQUENCE SLIDER: integration of structural and genetic data to characterize isoforms from natural sources. Nucleic Acids Res. 2022.

CADLE, J. E. **Problems and Approaches in the Evolutionary History of Venomous Snakes**. Mem. Inst. Butantan, v. 46, p. 255-274, 1983.

EMSLEY, P.; COWTAN, K., **Coot: Model-Building Tools for Molecular Graphics**. Acta crystallographica section D: biological crystallography, v. 60, n. 12, p. 2126-2132, 2004.

EMSLEY, P.; LOHKAMP, B.; SCOTT, W. G. *et al.* **Features and development of Coot**. *In*: Acta Crystallographica. Section D, Biological Crystallography, Vol. 66, No. 4. pp. 486-501, 2010.

FRY, B. G. **From Genome to "Venome": Molecular Origin and Evolution of the Snake Venom Proteome Inferred from Phylogenetic Analysis of Toxin Sequences and Related Body Proteins.** Genome Res, 2005.

FURTADO, M. F. D.; CARDOSO, S. T.; SOARES, O. E.; PEREIRA, A. P.; FERNANDES, D. S.; TAMBOURGI, D. V.; SANT'ANNA, O. A. **Antigenic Cross Reactivity and Immnunogenicity of Bothrops Venoms from Snakes of the Amazon Region.** Toxicon, v. 55, n. 4, p. 881-887, 2010.

GRAUR, D.; LI, W. H. **Fundamentaçs of Molecular Evolution**, 2ed. Sinauer, 2000.

KABSCH, W. **XDS**. Acta Crystallogr D Biol Crystallogr, 125–132, 2010.

KIMBALL, J. W. **Taxonomy: Classifying Life, Kimball's Biology Pages**, http://www.biology-pages.info/T/Taxonomy.html, 2013.

KLUYVER, T. *et al.* **Jupyter Notebooks – a publishing format for reproducible computational workflows**. *In* LOIZIDES, F. & SCHMIDT, B., eds. Positioning and Power in Academic Publishing: Players, Agents and Agendas. pp. 87–90, 2016.

KOCHVA, E. **The Origin of Snakes and Evolution of the Venom Apparatus.** Toxicon, v.

25, n. 1, p. 65-106, 1987.

JOHN, R.; SONGLING, L.; KARLOU, M. A.; DARON, M. S.; KAZUTAKA, K. **MAFFT-DASH: integrated protein sequence and structural alignment**, *Nucleic Acids Research*, Volume 47, Issue W1, 02 July 2019, Pages W5–W10, https://doi.org/10.1093/nar/gkz342

LANDAU, M.; MAYROSE, I.; ROSENBERG, Y.; GLASER, F.; MARTZ, E.; PUPKO, T.; BEN-TAL, N. **ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures**. Nucleic Acids Res, 2005.

LIEBSCHNER, D.; AFONINE, P.V.; *et al*. **Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix.** Acta Cryst, D75, 861-877, 2019.

LIEBSCHNER, D. *et al.* **Polder Maps: Improving OMIT Maps by Excluding Bulk Solvent**. Acta Crystallographica Section D: Structural Biology, v. 73, n. 2, p. 148-157, 2017.

LIU, X. R.; ZHANG, M. M.; GROSS, M. L. **Mass Spectrometry-Based Protein Footprinting for Higher-Order Structure Analysis: Fundamentals and Applications.** Chem Rev, 2020

MERRITT, E. A.; BACON D. J. **Raster3D: photorealistic molecular graphics**. Methods Enzymol, pp. 505-524, 1997.

MCCOY A. J.; GROSSE-KUNSTLEVE R. W., ADAMS P. D., WINN M. D., STORONI L. C., READ R. J. **Phaser Crystallographic Software**. J. Appl. Crystallogr, v. 40, n. 4, p. 658-674, 2007.

MCPHERSON A.; CUDNEY B. **Optimization of Crystallization Conditions for Biological Macromolecules**. Acta Crystallogr F Struct Biol Commun. 2014 Nov;70(Pt 11):1445-67, 2014.

PAHARI, S.; MACHESSY, S. P.; KINI, R. M**. The Venom Gland Transcriptome of the Desert Massasauga rattlesnake (Sistrusus catenatus edwardsii): Towards an Understanding of Venom Composition Among Advanced Snakes (Superfamily Colubroidea).** BMC Mol. Biol., v. 8, n. 11, p. 98-115, 2007.

PYRON, R. A.; BURBRINK, F. T.; COLLI, G. R.; DE-OCA, A. N.; VITT, C. A.; KUCZYNSKI, C. A.; WIENS, J. J. **The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and composition of support methods for likelihood trees.** Mol. Phylog. Evol., v. 58, n. 2, p. 329-342, 2011.

SALVADOR, G. H. M.; BORGES, R. J.; EULALIO, M. M. C.; SANTOS, L. D.; FONTES, M. R. M. **Biochemical, pharmacological and structural characterization of BmooMP-I, a new P–I metalloproteinase from bothrops moojeni venom**. Biochimie, 179, 54–64, 2020.

SCOTT, A. D.; BAUM, D. A. Phylogenetic Tree. *In*: KLIMAN, R. M. **Encyclopedia of Evolutionary Biology.** Oxford: Academic Press, p. 270-276, 2016.

SOUSA, L. F.de. **Estudo de venenos de serpentes do complexo Bothrops, uma comparação da filogenia e da reatividade com antivenenos**. 2014.

THORN A.; SHELDRICK G.M.; **Extending molecular-replacement solutions with SHELXE.** Acta Crystallography. D Biol. Crystallogr. 2013.

TICONA, W. G. C. **Algoritmos evolutivos multi-objetivo para a reconstrução de árvores filogenéticas**. Tese de Doutorado. Universidade de São Paulo. 2008.

VAN DE PEER, Y.; SALEMI, M. **Phylogenetic inference based on distance methods**. *In*: LEMEY P., SALEMI M.; VANDAMME A. **The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing**, Cambridge: Cambridge University Press, pp. 142-180, 2009.

VIANA, G. V. R**. Técnicas para construção de árvores filogenéticas**. 2007.

Anexo A: Códigos referente ao PDB das proteínas usadas para a elaboração do trabalho

Metaloproteases:
1HFS
1I1I
1I73
1I76
1JIW
1JIZ
1OS9
1R54
1R55
1RMZ
1ROS
1S4B
1SLM
1UTT
1UTZ
1Y93
1ZP5
1ZS0
1ZVX
2CLT
2D1N
2D1O
2DDF
2FV5
2HU6
2J83
2O3E
2O36
2OVX
2OW9
2OXU
2OXW
2OY2
2V4B
2X3A
2X3C
2X7M
2XS4
2Y6C
2YB9
2YIG
3B8Z
3DL1
3DPE
3DPF
3E8R
3EDZ
3EHX
3EHY
3ELM
3EWJ
3F1A
3F15
3F16
3F18
3F19
3LIR
3LJG
3LJT
3LJZ
3LK8
3LQ0
3LQB
3N2V
3NX7
3O64
3Q2G
3G42
3HB2
3HBU
3HBV
3HDA
3HY7
3HYG
3KEC
3KEJ
3KEK
3KHI
3KMC
3KME
3KRY
3L0T
3L0V
3LE9
3LEA
3LGP
3LIL
3TS4
3V96
3WV1
3WV3
3ZVS
3ZXH
4A3W
4A7B
4AXQ
4CTH
4DPE
4DV8
4EFS
4FVL
4FXQ
4GF1
4GQL
4GR0
4GR3
4GR8
4H1Q
4H3X
4H30
4H76
4H84
4I03
4IJO
4ILW
4JA1
4JIJ
4JP4
4JPA
4JQG
4L6T
4L63
4ON1
4PKQ
4PKR
4PKS
4PKU
4PKW
4R3V
4WF6
4WK7
4WKE
4WKI
4WZV
4XM6
4XM7
5B5O
5B5P
5BOY
5CUH
5CXA
5CZM
5D1S
5D1T
5D1U
5D2B
5D3C
5D7W
6GID
Fosfolipase-A2:
1FX9
1HN4
1LE6
1LE7
1O2E
1O3W
2AZY
2AZZ
2B00
2B01
2B03
2B04
2B96
2BAX
2BCH
2BD1
2WG7
2WG8
2WG9
2ZP3
2ZP4
2ZP5
3ELO
3FVJ
3G8F
3G8G
3U8B
3U8H
3U8I
4UY1
5G3M
5OW8
5OWC
5VFM
5WZM
5WZO
5WZS
5WZT
5WZU
5WZV
5WZW
5Y5E
6AL3
6G5J
Porinas:
1BH3
2ZFG
3A2S
3SY7
3SYB
3SZV
3T0S
3T24
3VZT
4GCP
4LSI
5DL7
5O77
5PRN
6PRN
7PRN