

# Phylogenetics Trees

João Tomás Mota Cunha

Relatório no âmbito da UC: Estruturas de Dados para Bioinformática

**Mestrado Bioinformática e Biologia Computacional**  
**2021/2022**

Porto, junho de 2022

# Resumo

Algoritmos informáticos são uma grande ferramenta de ajuda aos bioinformáticos, estando cada vez mais presentes na área da genética. Exemplo disso, é o seu uso em análises filogenéticas e na criação de árvores filogenéticas. Este trabalho apresenta uma proposta para a construção personalizada de três árvores filogenéticas através do uso dos algoritmos computacionais: *Unweighted Pair Group Method with Arithmetic Mean*(UPGMA), *Neighbor Joining*(NJ) e *Fitch-Margoliash*. De forma a obter comparações fundamentadas na árvore obtida por cada algoritmo usando os mesmos dados de *input*. Como *input* sequencias genéticas de 40 espécies de víboras, características da Índia peninsular. Com base nos resultados obtidos foi possível evidenciar algumas diferenças e semelhanças em relação aos três métodos usados, revelando que para construções de árvores relativamente simples é pertinente usar métodos computacionalmente mais triviais como é o exemplo dos algoritmos UPGMA ou NJ; Enquanto que, para construções mais complexas é mais fiável usar métodos de computação mais avançados como é o exemplo do algoritmo Fitch-Margoliash, revelando obter resultados mais precisos e reais apesar de ser um processo mais demorado. Trabalhos como este retratam cada vez mais a importância do uso de este tipo de algoritmos na vida de bioinformáticos, biólogos ou investigadores, sendo um instrumento cada vez mais essencial na sua vida, facilitando e otimizando o seu trabalho e tornando possível construções de árvores filogenéticas cada vez mais precisas e, reduzindo o seu tempo de execução, minimizando potenciais erros e possibilitando cada vez melhores resultados neste campo.

# Introdução

Hoje em dia, a bioinformática é uma ferramenta muito utilizada para descobrir relações funcionais entre genes e proteínas. A construção de árvores filogenéticas é uma das técnicas mais usadas, atualmente, para examinar a evolução de genes e espécies, apenas através do uso e comparação de sequências genômicas inteiras e de vários genomas diferentes.[1]

As árvores filogenéticas pode estudar a história evolutiva das espécies tornou-se uma parte indispensável nas das análises biológicas. Historicamente, as árvores filogenéticas foram construídas comparando as características morfológicas dos organismos estudados. Hoje em dia, os estudos filogenéticos são realizados principalmente em sequências de DNA ou proteínas de organismos. A partir de um conjunto de sequências de entrada, é possível construir um alinhamento múltiplo de sequências (MSA) que serve como *input* padrão na na construção de árvores filogenéticas. O MSA pode delinear regiões de sequência que sofrem mudanças rápidas durante a evolução ou revelar resíduos que mostram evidências de serem moldados pela seleção natural. [2]

A análise filogenética é uma ferramenta importante para investigadores que procuram a anotação estrutural e funcional de um conjunto de sequências de entrada, onde as sequências podem estar intimamente relacionadas ou não. Dada a grande variedade de métodos filogenéticos, é necessário realizar uma seleção acertada dos métodos computacionais a utilizar, adaptando essa escolha às diferentes situações. Os métodos baseados em distância sequencial são computacionalmente menos exigentes tornando-os mais adequados para construir árvores filogenéticas para um grande número de sequências.[3]

Neste trabalho o objetivo será a construção personalizada de árvores filogenéticas, recorrendo à linguagem de programação python, através do uso dos algoritmos: *Unweighted Pair Group Method with Arithmetic Mean*(UPGMA), *Neighbor Joining*(NJ) e *Fitch-Margoliash*. De forma a obter comparações fundamentadas na árvore obtida por cada algoritmo usando os mesmos dados de *input*.

# Materiais e Métodos

## Python e Análise Filogenética

Python, é o exemplo de uma linguagem de programação amplamente utilizada na construção de árvores filogenéticas. Esta provou ser muito eficaz e é usada em muitas aplicações de computação científica,[4] usando algoritmos pré-empacotados é possível utilizá-la em processos de construção de árvores filogenéticas, usados para resolver tarefas de grande escala numa velocidade rápida e têm requisitos de memória e capacidade de computação.[5]

A Análise filogenética é um estudo estatístico com o qual é possível analisar as relações evolutivas entre diferentes espécies e organismos por métodos matemáticos. Neste sentido, uma das principais ferramentas para representar estas relações é o diagrama conhecido como árvore filogenética.[6]

Uma árvore filogenética, também conhecida como filogenia, é um diagrama que representa as linhas de descendência evolutiva de diferentes espécies, organismos ou genes de um ancestral comum. As filogenias são úteis para organizar o conhecimento da diversidade biológica, estruturar classificações e fornecer informações sobre eventos que ocorreram durante a evolução. Além disso, porque essas árvores mostram descendência de um ancestral comum, e porque grande parte da evidência mais forte para a evolução vem na forma de ancestralidade comum, é preciso entender as filogenias para apreciar plenamente as evidências esmagadoras que sustentam a teoria da evolução.[7]

## Matrizes de Distância Filogenética

A utilização de matrizes de distância filogenética é uma das estratégias mais comuns para introduzir informação filogenética num quadro estatístico. Muitas vezes, essas matrizes precisam ser euclidianas para um tratamento padrão de distâncias que facilita análises estatísticas posteriores.[8]

Matrizes de distância são usadas em filogenia como métodos de distância não paramétricos, sendo originalmente aplicados a dados fenéticos usando uma matriz de distâncias pareadas. A matriz de distância pode vir de várias fontes diferentes, incluindo distância medida ou análise morfométrica, várias fórmulas de distância pareada (como distância euclidiana) aplicadas a caracteres morfológicos discretos ou distância genética da sequência, fragmento de restrição e até dados de alozimas, entre outros.[9]

Os métodos usados para a criação de matrizes de distância, na análise filogenética, dependem explicitamente de uma medida de "distância genética" entre as sequências avaliadas e, portanto, exigem um *MSA* (*Multiple Sequence Alignment* ou Alinhamento Múltiplo de Sequências) como *input*. Uma *MSA* é uma abordagem de comparação de sequências que consiste na identificação das posições de maior similaridade entre múltiplas (3 ou mais) sequências de

DNA ou proteínas. Clustal, Clustal Omega e MUSCLE são exemplos de programas que realizam este tipo de alinhamento. Um MSA permite identificar a conservação de uma determinada proteína ou família de proteínas, fornecendo assim indícios sobre a sua evolução.[2]

A distância entre sequências é frequentemente definida como a fração de incompatibilidades em posições alinhadas, com lacunas ignoradas ou contadas como incompatibilidades.[10] Os métodos de distância tentam construir uma matriz a partir da distância entre cada par de sequência. Após isso, é construída uma árvore filogenética que coloca sequências intimamente relacionadas sob o mesmo nó e cujos comprimentos de ramificação reproduzem as distâncias observadas entre as sequências. Os métodos de matriz de distância podem produzir árvores enraizadas ou não enraizadas, dependendo do algoritmo usado para calculá-las. Eles são frequentemente usados como base para tipos progressivos e iterativos de alinhamento de múltiplas sequências.[11]

## **Métodos de Construção de Árvores Filogenéticas**

***Unweighted Pair Group Method with Arithmetic mean (UPGMA):*** Um método de agrupamento simples que assume uma taxa de evolução constante (hipótese do relógio molecular). Necessita de uma matriz de distância de espécies analisadas que possa ser calculada a partir de um alinhamento múltiplo (MSA).[12] UPGMA sempre produz uma árvore ultramétrica (ou dendrograma), este método constrói a árvore correta com probabilidade razoavelmente alta quando a hipótese do “relógio molecular” se aplica e a distância evolutiva é grande para todos os pares de sequências. O mesmo pode ser muito útil para biólogos interessados em construir árvores de espécies.[13]

***Neighbor Join (NJ):*** Este é sem dúvida o mais popular entre os métodos baseados em distância.[13] Caracterizado como o um método de agrupamento “de baixo para cima” que também precisa de uma matriz de distância. NJ é uma abordagem heurística que não garante encontrar o resultado perfeito, mas em condições normais tem uma probabilidade muito alta de fazê-lo. Tem uma eficiência computacional muito boa, tornando-o adequado para grandes conjuntos de dados.[12]

***Fitch-Margoliash:*** Este usa um método de mínimos quadrados ponderados para agrupamento com base na distância genética.[14] O algoritmo estima o comprimento total do ramo (distância) e *clusters* de acordo com o par de espécies para determinar a árvore não enraizada com distância mínima. Sequências intimamente relacionadas recebem mais peso no processo de construção da árvore para corrigir o aumento da imprecisão na medição de distâncias entre sequências relacionadas distantes. Na prática, a correção da distância só é necessária quando as taxas de evolução diferem entre os ramos.[11] Este não assume uma taxa de evolução constante, o que é bastante realista.[15]

## Modelo proposto

O modelo proposto para construção personalizada de árvores filogenéticas através do uso de três dos algoritmos computacionais: Unweighted Pair Group Method with Arithmetic Mean(UPGMA), Neighbor Joining(NJ) e Fitch-Margoliash.

Nesta construção o *input* utilizado foi o PopSet:1940952900, fornecido pela base de dados do NCBI [16]. Este PopSet é composto por 40 espécies de víboras divididas em 3 géneros (palavra usada em taxonomia para um grupo de espécies intimamente relacionadas): *Reptilia*, *Viperidae* e *Craspedocephalus*. Sendo estas cobras características da Índia peninsular.[17]

## Resultados e Discussão

Neste trabalho são apresentados 3 métodos computacionais diferentes tendo como objetivo a criação de uma árvore filogenética para cada um, com recurso a sequências genéticas de víboras, naturais da Índia peninsular.

Após realizar o *download* do ficheiro “.fasta” de todas as sequências foi necessário alinhá-las com recurso a alinhador *online* de Alinhamento Múltiplo de Sequências **MUSCLE** (*Multiple Sequence Comparison by Log- Expectation*), este passo está descrito na figura 1. Esta ferramenta fornece como output um ficheiro ClustalW “.clw” com a informação de todas as 40 sequências alinhadas.

Nota: É possível, neste alinhador, seleccionar como output outro tipo de ficheiros como “.fasta”, “.html”, etc.

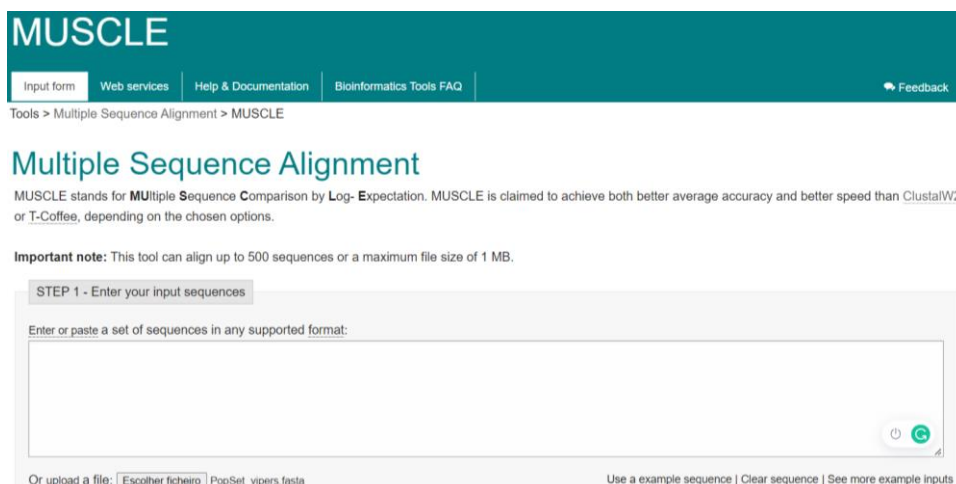
The image shows the web interface of the MUSCLE (Multiple Sequence Comparison by Log-Expectation) tool. At the top, there's a teal header with the 'MUSCLE' logo and navigation links: 'Input form', 'Web services', 'Help & Documentation', and 'Bioinformatics Tools FAQ'. Below the header, a breadcrumb trail reads 'Tools > Multiple Sequence Alignment > MUSCLE'. The main heading is 'Multiple Sequence Alignment'. A descriptive paragraph states: 'MUSCLE stands for Multiple Sequence Comparison by Log- Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.' An 'Important note' specifies: 'This tool can align up to 500 sequences or a maximum file size of 1 MB.' The interface is divided into 'STEP 1 - Enter your input sequences'. It contains a large text area with the instruction 'Enter or paste a set of sequences in any supported format:'. To the right of the text area are 'power' and 'refresh' icons. At the bottom, there are links for 'Or upload a file:', 'Escolher ficheiro', 'PopSet\_vipers.fasta', 'Use a example sequence', 'Clear sequence', and 'See more example inputs'.

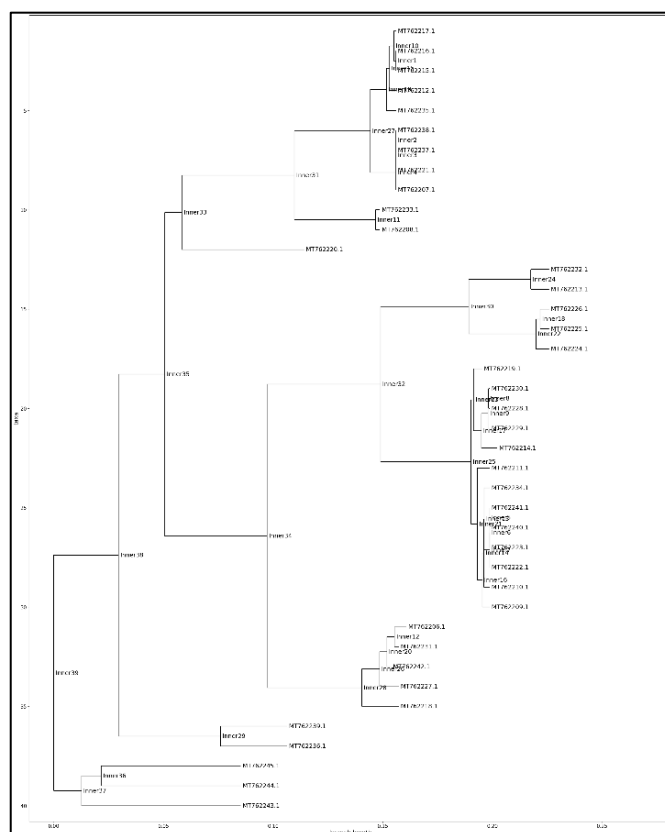
Figura 1: Alinhamento Múltiplo de Sequências no alinhador online MUSCLE.

Antes de iniciar a construção das árvores filogenéticas foi também necessário construir a matriz de distância, para isso usei a função DistanceCalculator() para gerar a matriz com uma *string* “identity”, que é o nome do modelo (matriz de pontuação) para calcular a distância. O modelo

Após estes passos foi possível construir a primeira árvore filogenética, usando o método UPGMA, com recurso ao *string parameter* ‘upgma’ na função DistanceTreeConstructor() e, em seguida, à função build\_tree(). Para conseguir visualizar toda a árvore tive de recorrer ao uso da biblioteca matplotlib, onde foi possível obter a figura 2.

Para a terceira e última árvore já foi necessário realizar uma abordagem diferente. Neste caso, foi usado o algoritmo ParsimonyScorer o qual se caracteriza por ser uma combinação do algoritmo *Fitch-Margoliash* e do algoritmo *Sankoff*. Neste caso funcionou como algoritmo *Fitch-Margoliash* por padrão pois nenhum parâmetro foi fornecido. Teria funcionado como algoritmo Sankoff se uma matriz de pontuação de parcimônia fosse fornecida, que neste caso não aconteceu. Para conseguir visualizar toda a árvore tive de recorrer ao uso da biblioteca matplotlib, onde foi possível obter a figura 4.





Fazendo uma avaliação final, foi possível perceber que todos os métodos computacionais usados para construção das árvores filogenéticas obtiveram com êxito exatamente os mesmos resultados, no entanto em tempos diferentes. Enquanto os métodos de UPGMA e NJ conseguiram um rápido processo de construção, realizados em cerca de 1 segundo, o mesmo não aconteceu ao método de *Fitch-Margoliash*, este mostrou ser um processo um pouco mais demorado que os restantes, tardando cerca de 9 segundos. Computacionalmente, a geração de árvores pelo método NJ é semelhante ao método UPGMA, obtendo também um tempo muito parecido na sua execução.

O método de *Fitch-Margoliash* é caracterizado por ser é mais preciso que os restantes métodos, mas menos eficiente. Pois o uso do método de mínimos quadrados ponderados para agrupamento com base na distância genética leva a um custo computacional maior. Sendo que para árvores como esta com relativamente poucas espécies não seria o método mais correto de executar pois seria mais pertinente usar métodos mais simples como UPGMA ou NJ, obtendo os mesmos resultados num menor espaço de tempo, sem recurso a mecanismos computacionais mais exigentes.



# Conclusão

Tem havido um grande aumento na quantidade de trabalhos de pesquisa realizados na área de construção de árvores genéticas através do uso de algoritmos computacionais como UPGMA, NJ ou Fitch-Margoliash, entre outros. No entanto, para chegar à sua máxima eficiência e desempenho terá um longo caminho de melhorias pela frente.

Este trabalho teve como objetivo a construção personalizada de árvores filogenéticas, recorrendo à linguagem de programação python, através do uso dos algoritmos: UPGMA, NJ e Fitch-Margoliash. Realizando comparações fundamentadas na árvore obtida por cada algoritmo usando os mesmos dados de *input*. Para que fosse possível atingir este objetivo foi necessário programar um código capaz de realizar estas árvores, considerando como input sequências genéticas de 40 espécies de víboras, características da Índia peninsular.

Este trabalho permitiu evidenciar algumas diferenças (sendo a mais evidente o tempo de processamento) e semelhanças em relação aos três métodos usados, revelando que para construções de árvores relativamente simples é pertinente usar métodos computacionalmente mais triviais como é o exemplo dos algoritmos UPGMA ou NJ; Enquanto que, para construções mais complexas é mais fiável usar métodos de computação mais avançados como é o exemplo do algoritmo Fitch-Margoliash, revelando obter resultados mais precisos e reais apesar de ser um processo mais demorado.

Estes resultados como este são muito relevantes no papel de um bioinformático analisar em detalhe alguns processos de criação de árvores filogenéticas, tornando mais acessível o estudo e melhoria de aplicações de genética computacional capazes de fazer construções cada vez mais precisas e eficientes de árvores filogenéticas, reduzindo o seu tempo de execução, minimizando potenciais erros e possibilitando cada vez melhores resultados neste campo.

# Referências bibliográficas

- [1] E. R. Ingham, T. P. Holtsford, and J. C. Walker, "Bioinformatics: Using phylogenetics and databases to investigate plant protein phosphorylation," *Adv. Bot. Res.*, vol. 32, pp. 45-65, Jan. 2000, doi: 10.1016/S0065-2296(00)32021-3.
- [2] "Introdução ao BioPython (parte IV): MSA e filogenia | by Frederico Schmitt Kremer | omixdata | Medium." <https://medium.com/omixdata/introdução-ao-biopython-parte-iv-msa-e-filogenia-ca825a0e84d3> (accessed Jun. 23, 2022).
- [3] P. Bawono and J. Heringa, "Phylogenetic Analyses," *Compr. Biomed. Phys.*, vol. 6, pp. 93-110, Jan. 2014, doi: 10.1016/B978-0-444-53632-7.01108-4.
- [4] E. Serrano, J. G. Blas, J. Carretero, M. Abella, and M. Desco, "Medical Imaging Processing on a Big Data Platform Using Python: Experiences with Heterogeneous and Homogeneous Architectures," *Proc. - 2017 17th IEEE/ACM Int. Symp. Clust. Cloud Grid Comput. CCGRID 2017*, pp. 830-837, Jul. 2017, doi: 10.1109/CCGRID.2017.56.
- [5] K. Gao, G. Mei, F. Piccialli, S. Cuomo, J. Tu, and Z. Huo, "Julia Language in Machine Learning: Algorithms, Applications, and Open Issues," *Comput. Sci. Rev.*, vol. 37, Mar. 2020, doi: 10.1016/j.cosrev.2020.100254.
- [6] "Árvores filogenéticas (artigo) | Filogenia | Khan Academy." <https://pt.khanacademy.org/science/ap-biology/natural-selection/phylogeny/a/phylogenetic-trees> (accessed Jun. 23, 2022).
- [7] "Phylogenetic Trees and Monophyletic Groups | Learn Science at Scitable." <https://www.nature.com/scitable/topicpage/reading-a-phylogenetic-tree-the-meaning-of-41956/> (accessed Jun. 19, 2022).
- [8] D. M. De Vienne, G. Aguileta, and S. Ollier, "Euclidean Nature of Phylogenetic Distance Matrices," *Syst. Biol.*, vol. 60, no. 6, pp. 826-832, Dec. 2011, doi: 10.1093/SYSBIO/SYR066.
- [9] "Distance matrices in phylogeny - Wikipedia." [https://en.wikipedia.org/wiki/Distance\\_matrices\\_in\\_phylogeny](https://en.wikipedia.org/wiki/Distance_matrices_in_phylogeny) (accessed Jun. 23, 2022).
- [10] D. W. Mount, "Bioinformatics: Sequence and Structural Analysis," vol. 1, p. 665, 2004.
- [11] D. Penny, "Inferring Phylogenies.—Joseph Felsenstein. 2003. Sinauer Associates, Sunderland, Massachusetts," *Syst. Biol.*, vol. 53, no. 4, pp. 669-670, Aug. 2004, doi:

10.1080/10635150490468530.

[12] R. Borriss, C. Rueckert, J. Blom, O. Bezuidt, O. Reva, and H. P. Klenk, "Whole Genome Sequence Comparisons in Taxonomy," *Methods Microbiol.*, vol. 38, pp. 409-436, Jan. 2011, doi: 10.1016/B978-0-12-387730-7.00018-8.

[13] V. Makarenkov, D. Kevorkov, and P. Legendre, "Phylogenetic Network Construction Approaches," *Appl. Mycol. Biotechnol.*, vol. 6, no. C, pp. 61-97, Jan. 2006, doi: 10.1016/S1874-5334(06)80006-7.

[14] W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees.," *Science (80-. )*, vol. 155, no. 760, pp. 279-284, Jan. 1967, doi: 10.1126/SCIENCE.155.3760.279/ASSET/D2E24AD2-D8C2-4193-9548-046C303E8AEB/ASSETS/SCIENCE.155.3760.279.FP.PNG.

[15] "CHAPTER 22: Construction of Phylogenetic Tree: Fitch Margoliash (FM) Algorithm - Basic Applied Bioinformatics [Book]." <https://www.oreilly.com/library/view/basic-applied-bioinformatics/9781119244332/c22.xhtml> (accessed Jun. 23, 2022).

[16] "PubChem." <https://pubchem.ncbi.nlm.nih.gov/> (accessed Jun. 09, 2022).

[17] "Crotalinae cytochrome b (cytb) gene, partial cds; mitochondrial. - PopSet - NCBI." <https://www.ncbi.nlm.nih.gov/popset/1940952900> (accessed Jun. 24, 2022).