Centro Nacional de Pesquisa em Energia e Materiais – CNPEM
Ilum – Escola de Ciência

Trabalho final: Espaços normados

Construção de árvores filogenéticas a partir de métricas matemáticas distintas

Estudo acerca do uso das métricas Hamming, Levenshtain e Euclidiana para construção de matrizes de sequências gênicas e construção de árvores filogenéticas.

Ana Luiza Poletto Loss
Ilum - Escola de Ciência.
Rua Lauro Vannucci, 1.020,
bairro Santa Cândida,
Campinas (SP).

Eloísa Maria Amador Souza
Ilum - Escola de Ciência.
Rua Lauro Vannucci, 1.020,
bairro Santa Cândida,
Campinas (SP).

Giovana Martins Coelho
Ilum - Escola de Ciência.
Rua Lauro Vannucci, 1.020,
bairro Santa Cândida,
Campinas (SP).

Vinicius Francisco Wasques
Ilum - Escola de Ciência.
Rua Lauro Vannucci, 1.020,
bairro Santa Cândida,
Campinas (SP).

Campinas — SP
25 de novembro de 2025

Resumo

Este trabalho analisa como diferentes métricas de distância, Hamming, Euclidiana e Levenshtein influenciam a construção de árvores filogenéticas a partir do gene mitocondrial COI para vertebrados. Assim, formamos matrizes de distância e aplicamos o método *Neighbor-Joining* para analisar similaridade entre espécies. As métricas Hamming e Euclidiana produziram resultados biologicamente coerentes, enquanto Levenshtein apresentou resultados inconsistentes devido à forma como penaliza inserções e deleções. Os resultados destacam que métricas simples podem recuperar padrões evolutivos gerais, mas possuem limitações quando comparadas a modelos evolutivos mais robustos.

1 Introdução

Árvores filogenéticas são um forma gráfica e visual de representar as relações evolutivas entre os organismos. Esse tipo de representação apresenta a distância genética entre os indivíduos, bem como, marca o início da presença de uma característica marcante através dos nós presentes nas árvores. Além disso, esse tipo de representação pode passar ou não pelo processo de enraizamento dependendo do objetivo de estudo, ou seja, a raiz da árvore filogenética é o ponto mais antigo da árvore, correspondendo ao último ancestral comum das unidades taxonômicas presente na árvore. Esse tipo de indicação permite que a árvore tenha direcionalidade com relação à evolução [1]. Essa estrutura é essencial para o estudo de processos fundamentais da evolução biológica. Com isso, árvores filogenéticas permitem a compreensão das principais transições na evolução, como inferir a origem de novos genes, detectar adaptações moleculares, compreender a evolução de caracteres morfológicos e reconstruir mudanças demográficas [2].

A construção de uma árvore filogética consiste em algumas etapas comuns, como a coleta de dados das sequências que serão utilizadas, o alinhamento dessas sequências, a construção da árvore e a análise dessa [3]. Há, também, etapas mais específicas durante essa construção, como utilizar uma métrica para a análise das distânicas entre sequências genéticas. Nesse sentido, no presente trabalho, as árvores filogenéticas foram construídas para espécies distintas de vertebrados, dos quais utilizou-se o gene COI, uma vez que esse é pequeno, facilitando a análise das distâncias. Além disso, serão utilizados diferentes métricas para avaliar como estas impactam nas distâncias apresentadas na árvore. Dessa forma, serão utilizadas as métricas Euclidiana, *Hamming*, *Levenshtein*.

A métrica euclidiana é a distância em linha reta entre dois pontos, calculada por meio de um espaço euclidiano, sendo uma das métricas mais populares. O espaço eucliadiano n-dimensional $\mathbb{R}^n$ é o produto de n fatores iguais a $\mathbb{R}$:$\mathbb{R}^n$ =$\mathbb{R} \times \mathbb{R} \times ... \times \mathbb{R}$ [4]. Seus elementos são caracterizados por sequências de n números reais $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$. A métrica euclidiana é dada pela seguinte função d_e: $\mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$

$$d_e(x, y) = (\sum_{i=1}^{2}(x_i - y_i)^2)^{\frac{1}{2}} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2},$$

sendo x e y são pontos dentro do espaço euclidiano.

Já a distância *Hamming* é uma métrica em um conjunto de cadeais com comprimento definido. Nesse sentido a distância de *Hamming* entre dois sistemas de igual comprimento $x = (x_1, ..., x_n)$ e $y = (y_1, ..., y_n)$, é denotado como $d_H(x, y)$ é dado por:

$$d_H(x, y) = \sum_{i=1}^{n} \delta(x_1, y_i),$$

sendo,

$$\delta(x_i, y_i) = \begin{cases} 0 x_i = y_i \\ 1 x_i \neq y_i \end{cases},$$

ou seja, é dado pelo número de posições em que x_i difere de y_i [5].

Por fim, a métrica de *Levenshtein*, é utilizada para descrever o número mínimo de alterações (inserções, exclusões ou substituições) necessárias para trasnformar uma palavra em outra. A distância entre duas coordenadas a e b, de comprimento $|a|$ e $|b|$ respectivamente, é denotado $lev_{a,b}(i, j)$ e dado por [6]:

$$lav_{ab}(i,j) = \begin{cases} max(i,j) & \text{se} \quad min(i,j) = 0 \\ min \begin{cases} lev_{a,b}(i-1,j)+1 \\ lev_{a,b}(i,j-1)+1 \\ lev_{a,b}(i-1,j-1)+1_{a_i \neq b_j} \end{cases} & \text{caso contrário} \end{cases},$$

sendo $1_{a_i \neq b_i}$ a função indicadora igual a 0 quando $a_i \neq b_i$ e igual a 1 caso contrário, $b(i,j)$ é a distância entre os primeiros i caracteres de a e os primeiros j caracteres de b. Cada equação descrita serão responsáveis por realizar a eliminação de caracteres, e/ou sua inserção e/ou a análise de correspondência.

O algortimo *Neighbor-joining*, ou junção de vizinhos em português, é um método utilizado para construir árvores filogenéticas a partir de dados de distância evolutiva. Neste método, buscamos por pares de unidades taxonômicas operacionais (OTSU, do inglês *operational taxonomic units*) que minimizem o comprimento total dos ramos em cada etapa de agrupamento de vizinhos. Neste algortimo, os vizinhos são OTSUs conectados por um nó em uma árvore não enraizada [9]. Por exemplo, na **Figura** 1, os OTSUs 1 e 2 são vizinhos, visto que estão conectados pelo átomo interno A.

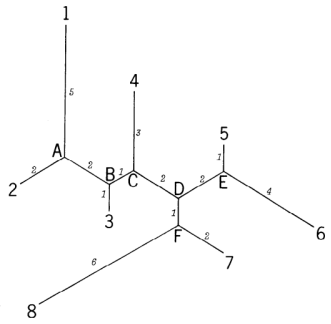


Figura 1: Árvore não enraizada para 8 OTSUs (numerados de 1 a 8). As letras (A-F) representam os nós internos do sistema. Em itálico, temos a distância entre os vizinhos. Imagem obtida da referência [9]

É possível definir a topologia de uma árvore utilizando este método por meio da união sucessiva de pares de vizinhos, gerando um novo vizinho. Para a **Figura** 1, por exemplo, iniciamos com os pares de vizinhos [1,2], [5,6] e [7,8]. Podemos agrupar os vizinhos $X_{1,2}$ = [1,2], e então teremos um novo conjunto [$X_{1,2}$, 3], [5,6] e [7,8]. Assim, relizaremos o agrupamento de vizinhos por Neighbor-joining para as matrizes de distâncias das sequências genéticas calculadas a partir de métricas distintas.

2 Metodologia

Para realizar este estudo, analisamos o gene da subunidade I da citocromo C oxidase mitocondrial (COI). Esta é a região gênica mais extensamente sequenciada do reino animal, sendo utilizado como um "código de barra"para identificação de espécies [8]. Nesse sentido, devido à ampla quantidade de dados disponíveis e seu tamanho reduzido (650 pares de base), o que diminui o custo computacional, este foi o gene escolhido para construção das árvores filogenéticas. As espécies escolhidas para comparação genética foram *Pan troglodytes* (chimpanzé), *Gorilla gorilla* (gorila), *Canis lupus familiaris* (cachorro doméstico), *Gallus gallus* (galo-banquiva), *Felis catus* (gato doméstico), *Bos taurus* (gado bovino) e *Danio rerio* (zebra-fish, ou paulistinha), os quais fazem parte do grupo dos vertebrados.

Assim, a primeira etapa para o desenvolvimento do projeto foi obter as sequências do gene COI para cada espécie a paritr do *National Center for Biotechnology Information* (NCBI). Foi selecionada uma sequência representativa do fragmento padrão de análise do COI para cada espécie estudada. As sequências foram armazenadas em um arquivo de texto, o qual posteriormente foi salvo no formato Fasta, que é amplamente utilizado em bioinformática para armazenar informação genética [10].

Em seguida, as sequências do gene COI foram importadas para o ambiente de análise utilizando Python. Para permitir o cálculo da distância de *Hamming*, que realiza comparações entre posições de sequências de mesmo comprimento, todas as sequências foram padronizadas para o mesmo tamanho. Para isso, cada sequência foi truncada para o comprimento mínimo observado entre as sequências analisadas.

As métricas de distância foram implementadas em Python. A métrica de *Hamming* foi codificada manualmente, contabilizando as discrepâncias posição a posição entre duas sequências de mesmo comprimento. A distância Euclidiana também

foi implementada manualmente, atribuindo valores inteiros a cada base nitrogenada e calculando a norma Euclidiana entre os vetores resultantes. Por fim, a distância de *Levenshtein* foi obtida por meio da implementação disponível na biblioteca *python-Levenshtein*, que calcula o número mínimo de operações de edição necessárias para transformar uma sequência em outra.

A métrica de *Hamming* requer que todas as sequências tenham o mesmo comprimento, pois sua definição se baseia na comparação posição a posição entre duas cadeias. Assim, durante a implementação computacional, para cada par de caracteres analisado, soma-se 1 à distância sempre que os nucleotídeos diferem naquela posição.

A métrica Euclidiana foi implementada convertendo cada nucleotídeo em um valor inteiro (A=0, C=1, G=2, T=3), o que permite representar cada sequência como um vetor numérico unidimensional. Em seguida, a distância Euclidiana entre duas sequências foi calculada como a norma L2 da diferença entre esses vetores. Esse procedimento quantifica diferenças base a base a partir da variação entre seus valores numéricos, sem utilização de codificação one-hot ou tratamento específico para bases degeneradas.

Além disso, a métrica *Levenshtein* calcula o custo mínimo necessário para transformar uma sequência na outra. São consideradas no processo as operações de remover, adicionar ou substituir um caractere, sendo que cada operação conta como um passo, e o algoritmo busca o caminho com menor custo total. Dessa forma, quanto menor o valor obtido após as operações, mais semelhantes são as duas sequências, pois poucas operações são necessárias para convertê-las.

Após definir as métricas, calculamos matrizes de distâncias entre as sequências utilizando a métrica desejada. Para cada par de espécies (i, j), a distância entre suas sequências completas é computada e armazenada em uma matriz simétrica, cuja diagonal é igual a zero. Assim, obtivemos uma matriz $N \times N$ que contém as distâncias entre todas as espécies analisadas, servindo como base para a construção das árvores filogenéticas.

A partir da matriz de distâncias obtida, construímos a árvore filogenética utilizando o método *Neighbor-Joining*, por meio da função `nj()` disponibilizada pela biblioteca `skbio.tree`. A árvore resultante foi visualizada em formato filogenético, enquanto a matriz de distâncias foi representada como um *heatmap*, permitindo comparar graficamente o grau de similaridade entre as espécies.

3 Resultados e Discussões

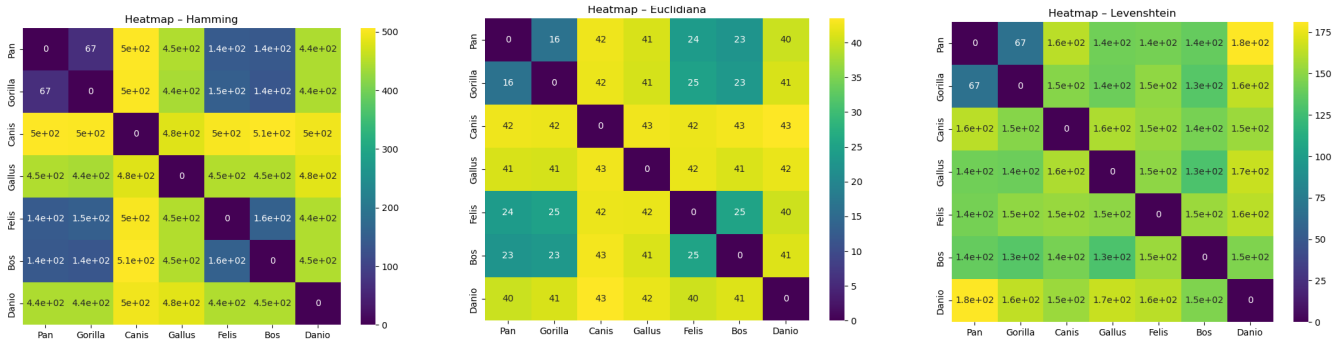
3.1 Análise Metodológica: O Impacto da Métrica de Distância

3.1.1 Matrizes de Distância em *Heatmap*

Após a normalização das sequências do gene COI, isto é, deixar todas do mesmo tamanho, foram geradas as matrizes de distância *(heatmaps)* para cada métrica analisada, permitindo avaliar como cada abordagem quantifica a similaridade genética entre os táxons.

Hamming e Euclidiana. As matrizes correspondentes (Figuras 2) exibem padrões de similaridade semelhantes. As menores distâncias (regiões escuras) ocorrem entre *Pan* e *Gorilla*, o que é consistente com a proximidade filogenética esperada entre esses primatas. As maiores distâncias (regiões amarelas) aparecem entre *Danio rerio* e os demais táxons (tetrápodes), refletindo adequadamente a divergência evolutiva mais profunda. O táxon é um grupo de organismos que são agrupados na classificação biológica com base em características compartilhadas.

Levenshtein. A matriz baseada na distância *Levenshtein* (Figura 2) apresenta valores em escala distinta e mais homogênea. Essa métrica mostrou-se inadequada para o gene COI, cujo comprimento é altamente conservado, pois penaliza inserções e deleções com o mesmo peso que substituições. A Inserção e a deleção são tipos de mutações genéticas que envolvem a adição (inserção) ou remoção (deleção) de um ou mais nucleotídeos em uma sequência de DNA. Em genes codificadores, onde inserções são eventos raros, essa penalização excessiva distorce a estimativa de distância evolutiva.



(a) Hamming *heatmap* (b) Euclidiana *heatmap* (c) Levenshtein *heatmap*

Figura 2: Matrizes de Distância do Gene COI, visualizadas em *Heatmap*. Os painéis representam as distâncias calculadas entre os táxons para as métricas Hamming , Euclidiana e Levenshtein. As cores escuras indicam alta similaridade (distância baixa), notavelmente entre Pan e Gorilla. As cores amarelas indicam alta divergência, consistentemente entre Danio e o restante dos táxons. Nota-se que a matriz Levenshtein possui uma escala de distância significativamente diferente devido à sua penalização por inserções e deleções além das substituições.

3.1.2 Necessidade de Normalização e a Falha da Métrica *Levenshtein*

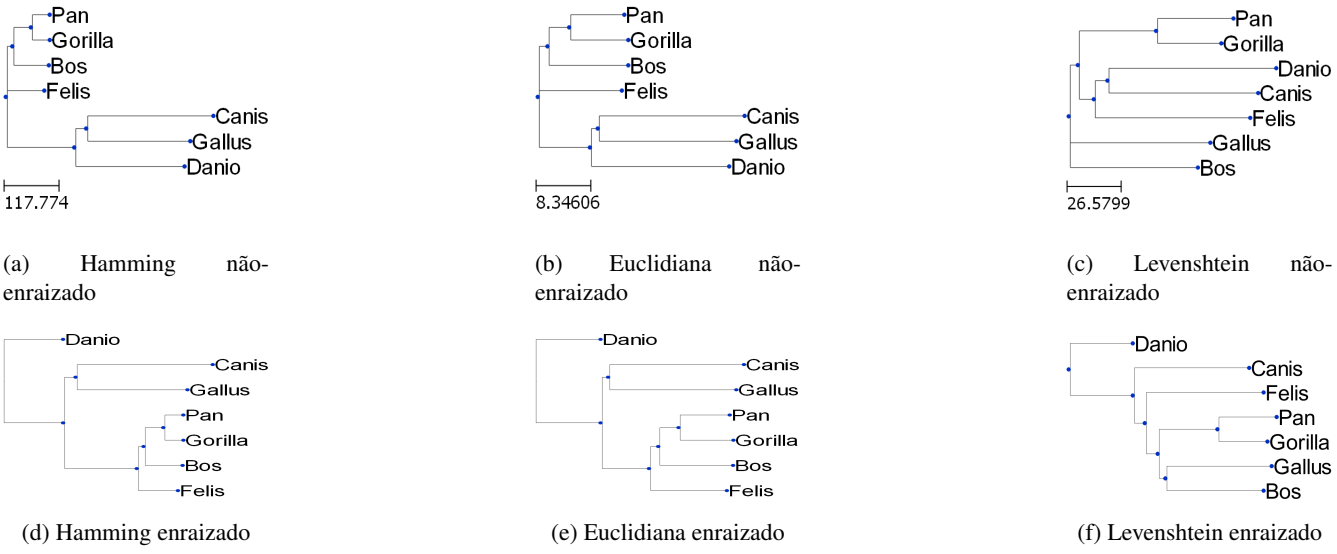
A normalização das sequências (via corte para tamanhos iguais) foi essencial para as métricas *Hamming* e Euclidiana, que dependem de vetores de mesma dimensão para o cálculo posição a posição. A falha na geração das matrizes quando a normalização foi omitida confirma essa limitação estrutural.

Por outro lado, a métrica Levenshtein produziu valores semelhantes independentemente da normalização, evidenciando sua insensibilidade ao alinhamento posicional e sua tendência a penalizar diferenças de comprimento, o que não necessariamente reflete a história evolutiva. Esse comportamento reforça que a distância *Levenshtein* não isola o sinal evolutivo relevante para sequências codificadoras como o gene COI.

3.2 Análise Filogenética e Topologias

As topologias iniciais (Figura 3), ainda sem enraizamento, representam apenas os agrupamentos relativos entre os táxons; a escala sob cada árvore corresponde ao comprimento total dos ramos (somatório das distâncias observadas).

O enraizamento foi realizado utilizando *Danio rerio* como grupo externo (*outgroup*), escolha filogeneticamente apropriada, uma vez que os peixes ósseos divergiram muito antes do ancestral comum dos tetrápodes (mamíferos e aves).



(a) Hamming não-enraizado (b) Euclidiana não-enraizado (c) Levenshtein não-enraizado

(d) Hamming enraizado (e) Euclidiana enraizado (f) Levenshtein enraizado

Figura 3: Topologias filogenéticas obtidas pelo método Neighbor-Joining (NJ) com diferentes métricas de distância. Os painéis superiores (a–c) mostram as árvores não enraizadas, enquanto os inferiores (d–f) apresentam as árvores enraizadas usando Danio rerio como *outgroup*. A topologia enraizada com distância Hamming aproxima-se da relação evolutiva esperada, ao passo que a obtida com Levenshtein exibe agrupamentos biologicamente inconsistentes.

Topologia baseada em *Hamming*. A árvore obtida com a métrica *Hammin* (Figura 3 (d)) foi a mais coerente com o conhecimento filogenético, recuperando o clado *Pan* + *Gorilla* e posicionando *Gallus* (Aves) como mais basal em relação aos mamíferos. Contudo, a proximidade entre *Canis* e *Gallus*, bem como entre *Bos* e *Felis*, representa uma anomalia; filogenias reais indicam que a separação entre aves e mamíferos deve ocorrer antes das divergências internas entre os mamíferos.

Topologia baseada em Levenshtein. A árvore construída a partir da métrica *Levenshtein* (Figura 3 (f) apresenta agrupamentos biologicamente improváveis, como o pareamento recente entre *Gallus* e *Bos*. Essa incongruência reflete diretamente a distorção introduzida pela matriz de distâncias, já discutida na Seção 3.1.2.

3.3 Comparação com Métodos Teóricos e Referência

A topologia obtida difere daquela esperada e também da árvore gerada como referência pelo NGPhylogeny [11] (Figura 4), principalmente devido ao uso de métricas de distância não corrigidas.

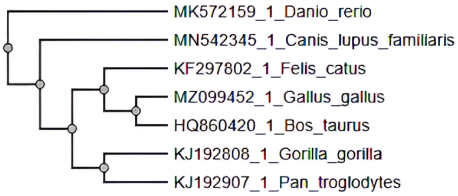


Figura 4: Topologia de Referência (Resultado Esperado) Gerada por Software Profissional (e.g., NGPhylogeny). Esta árvore representa a topologia ideal e estatisticamente mais robusta. Nota-se a correta separação entre o Danio (Outgroup), o clado dos Primatas (Pan/Gorilla) e o agrupamento basal do Gallus (Ave) antes da diversificação dos mamíferos.

O método Neighbor-Joining (NJ) foi utilizado por sua eficiência computacional e por não assumir ultrametricidade (taxas evolutivas constantes entre linhagens). De modo geral, os métodos filogenéticos clássicos podem ser divididos em três categorias:

Método	Base Principal	Princípio	Suposição Chave
UPGMA	Distância	Agrupamento hierárquico pela média.	Assume relógio molecular (taxas evolutivas constantes).
Neighbor-Joining (NJ)	Distância	Minimiza o comprimento total dos ramos.	Não assume relógio molecular; estatisticamente consistente.
Máxima Verossimilhança (MV)	Caractere	Maximiza a probabilidade dos dados dado um modelo evolutivo ($P(D \mid T)$).	Requer um modelo evolutivo explícito.

Tabela 1: Comparação entre métodos clássicos de inferência filogenética. [9, 12, 13]

A topologia ideal costuma ser obtida por métodos baseados em caractere (como MV), porém seu custo computacional é muito grande. Também pode ser feito por NJ quando associado a modelos de distância corrigidos, como JC69 ou Kimura 2-parâmetros.

A discrepância observada na árvore construída com a métrica Hamming decorre, sobretudo, da saturação do sinal evolutivo. Essa métrica não corrige para múltiplas substituições no mesmo sítio (por exemplo, $A \to T \to A$), registrando tais eventos como uma única diferença. Essa subestimação reduz a acurácia dos comprimentos de ramos e pode alterar a ordem de divergência entre os clados, especialmente entre os mamíferos, onde mais substituições tendem a se acumular ao longo do tempo.

4 Conclusão

O desenvolvimento do projeto permitiu o entendimento que as árvores filogenéticas são altamente sensíveis à escolha da métrica de distância, mesmo quando é utilizado um método estatisticamente consistente como o *Neighbor-Joining*. Também foi realizado o enraizamento utilizando o *Danio rerio* como grupo externo, isso confirmou a separação esperada entre Peixes e Tetrápodes, corroborando para a correção metodológica dessa escolha.

Entre as métricas avaliadas, a distância *Hamming* apresentou o desempenho mais coerente do ponto de vista biológico. Isso porque com o uso dessa métrica foi possível capturar adequadamente o principal processo evolutivo relacionado ao gene

mitocondrial COI, ou seja, as substituições nucleotídicas. No entanto, a métrica *Levenshtein* não foi adequada para este tipo de marcador, uma vez que sua penalização para substituições e *indels*, acarretando em topologias filogenéticas biologicamente incoerentes.

As diferenças apresentadas entre as árvores filogenéticas obtidas e a topologia filogenética já consolidada destacam a limitação de utilizar métricas de distância simples e não corrigidas para dados moleculares. Assim, desenvolvimentos futuros do projeto devem priorizar o uso de Modelos de Substituição (como K2P, JC69 ou GTR), capazes de corrigir múltiplas substituições em um mesmo sítio, bem como incorporar alinhamentos múltiplos de alta qualidade, fundamentais para estimar relações evolutivas com maior precisão.

Com isso, o trabalho reforça que a escolha metodológica adequada, especialmente para cálculo das distâncias evolutivas, é um fator determinante para a reconstrução adequada da história evolutiva dos organismos.

5 Agradecimentos

Agradecemos ao professor Vinicius Francisco Wasques pelas orientações durante a construção deste trabalho.

6 Referências

[1] KINENE, T.; WAINAINA, J.; MAINA, S.; BOYKIN, L. M. *Rooting Trees, Methods for*. In: KLIMAN, R. M. (Ed.). **Encyclopedia of Evolutionary Biology**. 2016. p. 489-493. DOI: 10.1016/B978-0-12-800049-6.00215-8.

[2] KAPLI, P.; YANG, Z.; TELFORD, M. J. *Phylogenetic tree building in the genomic era*. **Nature Reviews Genetics**, v. 21, p. 428-444, 2020. DOI: 10.1038/s41576-020-0233-0.

[3] ZOU, Y. et al. *Common methods for phylogenetic tree construction and their implementation in R*. **Bioengineering**, v. 11, n. 5, p. 480, 2024. DOI: 10.3390/bioengineering11050480.

[4] ELON LAGES LIMA. Analise real. 2, Funcoes de en variaveis. Rio De Janeiro: Impa, 2010.

[5]HIDEKI IMAI. Essentials of error-control coding techniques. San Diego: Academic Press, 1990.

[6]HARSH BINANI. The Levenshtein Algorithm. Disponível em: https://medium.com/@binaniharsj/the-levenshtein-algorithm-215567a3ab1f . Acesso em: 20 nov. 2025.

[7] SEGURA-ALABART, Natàlia; SERRATOSA, Francesc; GÓMEZ, Sergio; FERNÁNDEZ, Alberto. Nonunique UPGMA clusterings of microsatellite markers. Briefings in Bioinformatics, v. 23, n. 5, p. bbac312, 01 ago. 2022. DOI: 10.1093/bib/bbac312.

[8]PENTINSAARI, M. et al. Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. Scientific Reports, v. 6, n. 1, 13 out. 2016.

[9]SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution, v. 4, n. 4, p. 406–425, jul. 1987.

[10]FASTA Format for Nucleotide Sequences. Disponível em: https://www.ncbi.nlm.nih.gov/genbank/fastaformat/.

[11]NGPHYLOGENY.FR. NGPhylogeny.fr. Disponível em: https://ngphylogeny.fr. Acesso em: 20 nov. 2025.

[12] FELSENSTEIN, J. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *Journal of Molecular Evolution*, v. 17, n. 6, p. 368–376, 1981.

[13] SOKAL, R. R.; MICHENER, C. D. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, v. 38, p. 1409–1438, 1958.