UNIVERSIDADE FEDERAL DE ALAGOAS
CAMPUS A. C. SIMÕES
INSTITUTO DE COMPUTAÇÃO
CIÊNCIA DA COMPUTAÇÃO

LARISSA DA SILVA SANTOS

Trabalho de Conclusão de Curso

MACEIÓ-AL
2023

LARISSA DA SILVA SANTOS

Uma Metaheurística para Tratar a Parcimônia em Árvores Filogenéticas

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientadora:

Profa. Dra. Roberta Vilhena Vieira Lopes

MACEIÓ
2023

Folha de Aprovação

LARISSA DA SILVA SANTOS

Uma Metaheurística para Tratar a Parcimônia em Árvores Filogenéticas

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação do Instituto de Computação da Universidade Federal de Alagoas, aprovada pela comissão examinadora que abaixo assina.

Banca Examinadora

Documento assinado digitalmente
gov.br ROBERTA VILHENA VIEIRA LOPES
Data: 20/12/2023 16:50:28-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Roberta Vilhena Vieira Lopes - Orientador
(Universidade Federal de Alagoas - Campus A. C. Simões - Instituto de Computação)

Documento assinado digitalmente
gov.br ALMIR PEREIRA GUIMARAES
Data: 13/12/2023 11:49:08-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Almir Pereira Guimarães - Examinador
(Universidade Federal de Alagoas - Campus A. C. Simões - Instituto de Computação)

Documento assinado digitalmente
gov.br MARIA CRISTINA TENORIO CAVALCANTE ESCAR
Data: 17/12/2023 22:13:01-0300
Verifique em https://validar.iti.gov.br

Profa. Dra. Maria Cristina Tenório C. Escarpini - Examinador
(Universidade Federal de Alagoas - Unidade Penedo - Instituto de Computação)

Catalogação na Fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 - 1767

S237m Santos, Larissa da Silva.
Uma metaheurística para tratar a parcimônia em árvores filogenéticas / Larissa da Silva Santos. – 2023.
30 f. : il.

Orientadora: Roberta Vilhena Vieira Lopes.
Monografia (Trabalho de conclusão de curso em Ciência da Computação) – Universidade Federal de Alagoas, Instituto de Computação. Maceió, 2023.

Bibliografia: f. 28-30.

1. Parcimônia. 2. Filogenia. 3. Relações familiares. 4. Metaheurística. 5. Biotecnologia. 6. Algoritmo de Wagner. 7. Computação evolutiva. I. Título.

CDU: 004.023:575.86

Agradecimentos

Gostaria de agradecer aos meus pais, Sueli e Luciano, pelo suporte e demonstração de perseverança em todos esses anos.

Agradeço aos meus irmãos: Luciana, Lucas, Daniel e Ludmilla, por, nos momentos difíceis, me lembrarem de sorrir.

Agradeço com muito carinho e admiração à minha orientadora Profa. Dra. Roberta Vilhena, por sua orientação e paciência ao longo destes anos, e por representar modelo de inspiração de que é possível realizar diversas funções com otimismo.

Agradeço aos professores pelos ensinamentos passados que moldaram a profissional que sou hoje. Agradeço aos técnicos do Instituto de Computação pela cooperação no decorrer do curso.

Agradeço aos professores que aceitaram fazer parte de minha banca, Prof. Dr. Fábio e Profa. Dra. Maria Cristina.

Agradeço à Deus por colocar todas essas pessoas no meu caminho.

"Um ser humano deve transformar informação em inteligência ou conhecimento. Tendemos a esquecer que nenhum computador jamais fará uma nova pergunta."

– Hopper, *Grace*

Resumo

Na sociedade, existem diversos produtos com recursos alterados geneticamente, desde alimentos até tratamentos médicos. Seja para remover ou adicionar características em uma dada espécie, os pesquisadores precisam conhecer os graus de parentescos entre as espécies em estudo. Para compreender as relações de parentescos entre espécies é necessário representá-las na árvore filogenética, de acordo com suas características. Para construir uma árvore filogenética, são utilizados métodos por similaridade e por distância, ambos contém problemas de desempenho ou resultados conflitantes. No método por distância, é utilizado o Princípio de Parcimônia, este afirma que a árvore ideal teria a menor quantidade de modificações. Os algoritmos de construção de árvore filogenética apresentam problemas na parcimônia quando existem mais de uma espécie com a menor distância, o que limita o espaço de busca dos descendentes ou ascendentes das próximas espécies. É sugerida então uma metaheurística para solucionar este impasse, após modificações, definiu-se que classificando as espécies de três tipos de parcimônias por quantidade de vezes selecionadas junto com verossimilhança ajudam o Algoritmo de Wagner a sair de máximo globais ruins, tendo como resultado árvore mais realística.

Palavras-chave: parcimônia; árvore filogenética; relações de parentesco; metaheurística; biotecnologia; algoritmo de Wagner; computação evolucionária.

Abstract

In society, there are several products with genetically altered resources, from food to medical treatments. Whether to remove or add features in a given species, researchers need to know the degrees of relatedness between the species under study. To understand the kinship relationships between species, it is necessary to represent them in the phylogenetic tree, according to their characteristics. To build a phylogenetic tree, similarity and distance methods are used, both of which have performance problems or conflicting results. In the method by distance, the Principle of Parsimony is used, it is stated that the ideal tree would have the least amount of modifications. The phylogenetic tree construction algorithms present parsimony problems when there are more than one species with the shortest distance, which limits the search space for descendants or ancestors of close species. A metaheuristic is then suggested to solve this impasse, after modifications, to define that classifying the species of three types of parsimony by number of times selected together with likelihood helps Wagner's Algorithm to get out of maximum global ruins, resulting in a tree more realist.

Key-words: parsimony; phylogenetic tree; kinship relations; metaheuristics; biotechnology; Wagner's algorithm; evolutionary computation.

Lista de Figuras

Lista de Tabelas

Conteúdo

1

Introdução

1.1 Classificação das espécies

Para classificação de espécies, utiliza-se a premissa básica de que espécies com maior quantidade de características em comum, tendem a ser parentes próximos. Essa premissa, fomenta a hipótese de que organismos com maior relação de parentesco são mais semelhantes do que os demais seres.

A partir da hipótese de similaridade, tem-se dois métodos de classificação: Classificação Evolutiva e Classificação Taxonômica. A classificação evolutiva realiza agrupamento dos organismos conforme a distância no parentesco das espécies. O cálculo da distância é baseado na soma das distâncias existentes entre um conjunto de propriedades dos organismos selecionadas pelo biólogo.

Neste tipo de classificação, a seleção das características ocorre sem método objetivo, pois depende do pesquisador definir e evidenciar os aspectos das espécies em questão, o que torna improvável a replicação e comprovação. Além disso, esta classificação não permite a utilização de muitas características. Assim, como não existe metodologia nessa classificação e os resultados dependem de fatores subjetivos, as relações de parentescos resultantes podem não representar a evolução das espécies em estudo.

Seguindo as diretrizes da Teoria da Evolução das Espécies, Richard Owens definiu que caracteres com mesma origem embrionária e ancestralidade em comum são caracteres homólogos. Porém, os pesquisadores ainda precisavam identificar e organizar os caracteres homólogos.

Em 1960, o entomólogo alemão Willi Hennig propôs a classificação filogenética, esta teoria é a mais utilizada e, nesta classificação, entende-se que durante a evolução de determinados organismos, as características presentes poderão sofrer modificações e permanecer nas próximas gerações, tornando-se tais variações presentes no organismo e em seus descendentes. Essas variações nas características da linhagem ancestral são chamadas de condição derivada (Santos, 2012).

Atualmente, a classificação filogenética é utilizada para compreensão da evolução das espécies na Biotecnologia, tendo como resultados produtos alterados geneticamente nas mais diversas áreas. Nas subseções seguintes, serão apresentados alguns produtos transgênicos.

1.1.1 Arroz dourado

De acordo com a Organização Mundial de Saúde, cerca de 2,8 milhões de crianças em idade pré-escolar no mundo são clinicamente afetadas pela hipovitaminose A. Estima-se que de 250 a 500 mil crianças tornam-se cegas todos os anos, sendo que cerca da metade morrem antes de completar um ano de vida (Milagres, 2007).

O arroz dourado, nome técnico *Provitamin A Biofortified Rice Event GR2E*, foi produzido com objetivo de fornecer altos níveis de betacaroteno, precursor da vitamina A que é essencial para a manutenção dos tecidos, sendo estes o primeiro defensor do organismo contra infecções. Também atua na manutenção da visão e no funcionamento adequado do sistema imunológico (Stokstad, 2019).

O arroz convencional produz caroteno nas folhas, então foram selecionados dois genes pertencentes à via biossintética dos carotenóides para serem adicionados no genoma do arroz: o *psy1* de milho e o *ctr1* da bactéria *Pantoea ananatis*. Com esses dois elementos, o arroz transgênico consegue produzir a molécula que inicia os carotenóides em plantas.

1.1.2 Cana de açucar

O Brasil é o maior produtor de cana-de-açúcar do planeta. Os centros de pesquisa nacionais realizam melhoramento genético na cana-de-açúcar e utilizam a biotecnologia moderna para o desenvolvimento de variedades transgênicas. Esses centros procuram melhorar características como diferentes tipos de resistência, maior produção de açúcar, tolerância a estresse abiótico e adaptação em diferentes ambientes.

1.1.3 Feijão RMD

A produção do feijão no Brasil é afetada pelo mosaico-dourado, doença causada pelo vírus *Bean golden mosaic virus (BGMV)*, sendo transmitida pela mosca-branca (*Bemisia tabaci*), o que acarreta em implicações no abastecimento e no preço.

Devido aos danos causados pelo vírus mosaico-dourado, estima-se que haja uma redução de 90 a 280 mil toneladas na produção de feijão. Além dos danos econômicos, a virose inviabiliza a produção de feijão em sistemas de agricultura familiar em várias regiões do Brasil.

Este produto foi obtido a partir da inserção de um fragmento de DNA derivado do vírus no seu genoma nuclear, o que interfere no ciclo de replicação, eliminando o gene viral *rep*. Como consequência, a replicação viral é interrompida e o feijoeiro-comum torna-se resistente ao vírus (Barbosa, 2021). Essa modificação proporciona, além de manter a safra resistente, menos uso

de agrotóxicos e permite maior área de cultivo, neste cenário, pode-se recuperar rapidamente da perda de safra por razões climáticas.

Reconstruir relação de parentesco é um problema NP completo, sendo utilizado técnicas com resultados aceitáveis. Entre estas técnicas, está o método de construção por distância, no qual constrói-se a árvore pelo Princípio da Parcimônia, ou seja, infere que a próxima espécie gerada será aquela com menor distância em relação ao ser vivo atual. Nesta etapa, caso existam mais de uma espécies com a menor distância, é selecionado o primeiro clado apresentado. A proposta será utilizar o auxílio de três tipos de distâncias, escolhendo a espécie mais selecionada pelas distâncias.

1.2 Objetivo

O objetivo deste trabalho de conclusão de curso é propor uma meta heurística para lidar com o problema em que a distância de um espécie para duas outras espécies ancestrais são iguais, ou seja, há ocorrência de parcimônia entre as espécies estudadas.

1.3 Organização da dissertação

De forma a realizar os objetivos mencionados e compreender o contexto, este trabalho de conclusão organiza-se em cinco capítulos, incluindo o atual capítulo de introdução.

No capítulo dois, é apresentado breve relato do surgimento da Filogenia. Neste capítulo, compreende-se o que é árvore filogenética, quais são os tipos de matrizes usadas como dados para a construção da árvore e como é a inferência filogenética. Depois, é definido o Princípio da Parcimônias e suas variâncias.

No capítulo três é ressaltado a importância dos algoritmos para a construção das relações de parentescos, após são apresentadas as abordagens computacionais para a geração das árvores filogenéticas. Depois, de compreender o contexto, do ponto de vista computacional, é demonstrado o problema que ocorre durante a construção de árvores quando suas matrizes características têm distâncias conflitantes.

No capítulo quatro, é apresentada a arquitetura, elementos e funcionalidades do sistema proposto. Também é demonstrado a utilização do sistema com o caso de uso do gênero Edessa, sendo utilizada uma parte da base de características morfológicas desse gênero, visto que este gênero possui diversas espécies.

No capítulo 5, é exposta a conclusão deste trabalho de conclusão e as perspectivas de trabalhos futuros a serem desenvolvidos.

2

Fundamentos da Filogenia

2.1 Surgimento da Sistemática Filogenética

Até a Idade Moderna, o pensamento vigente era o essencialismo, este afirma que as espécies têm características imutáveis e semelhantes que as agrupam. Nesse contexto, Linneau projetou um sistema de classificação em que dividia as espécies em Reino das Plantas, Minerais e Animais (Klepka, 2018).

No início do século XIX, surgiram diversas teorias contrárias ao essencialismo. Dentre elas, o Lamarckismo do naturalista francês Jean-Baptiste de Lamarck o qual consiste em duas leis: *Lei do Uso e Desuso e a Lei da Herança dos Caracteres* (Frezzatti Júnior, 2011). A primeira lei afirma que um ser vivo fortalece ou atrofia certa característica a partir de sua utilização, enquanto a segunda lei afirma que as alterações adquiridas pelas espécies ancestrais seriam transferidas aos descendentes. Lamarck percebeu a influência do ambiente no comportamento e contribuiu para o pensamento evolucionista, no qual as espécies têm características mutáveis e compartilhadas com seus parentes.

Dentre os evolucionistas, o naturalista Charles Darwin explorou diversas espécies em vários estágios da vida de forma a compreender a evolução destas espécies. Após análise, criou a obra que o tornaria conhecido nas Ciências, *Origem das Espécies*.

Nesta obra, Darwin descreve a *Teoria da Evolução das Espécies*, a qual afirma que a evolução ocorre por meio da seleção natural em que o ambiente seleciona os indivíduos mais aptos. Ainda de acordo com a *Teoria da Evolução*, esses indivíduos selecionados reproduzem e transmitem suas características para seus descendentes, tornando essas características presentes nas próximas gerações (Darwin, 2018).

No início do século XIX, com a aceitação nas Ciências da *Teoria da Evolução*, foi possível trabalhar os aspectos responsáveis pela diversificação das linhagens e classificá-los de acordo com as hipóteses de parentescos (Mayr, 2005).

Com a Escola Taxonômica Filogenética, o parentesco entre clados (espécies) ocorre quando estes têm ancestral em comum e exclusivo denominados como grados, ou seja, as espécies são integrantes de mesma linhagem.

A proposta de Hennig, cientista da Escola Filogenética, *"é que as classificações biológicas devem ser reflexo inequívoca do conhecimento atual sobre as relações de parentescos entre táxons."*. Nesse contexto, a classificação filogenética definiu que as características evoluem através da filogenia (árvore filogenética), então a filogenia tornou-se uma fonte de informação para fisiólogos, ecólogos, bioquímicos, taxonômicos, etc (Sadava, 2019).

Com os conceitos de parentescos definidos, havia a necessidade de uma forma de representação eficiente destas relações, o que foi solucionado com utilização de grafo direcionado, mais especificamente árvore binária, que foi denominada de árvore filogenética.

Além disso, percebeu-se que conforme a quantidade de espécies e/ou características em estudo aumentam, a complexidade e o risco de resultados divergentes com a realidade crescem proporcionalmente, sendo essencial o uso de algoritmos para a construção das relações de parentescos (Wheeler, 2012).

Sendo a reconstrução das relações de parentesco um problema NP completo, pois as formas como a história evolutiva de N espécies pode ocorrer é igual a N!. Então, os algoritmos que garantem a árvore de solução ótima são problemas NP completos, por isto para resolvê-las utilizam-se técnicas com resultados aceitáveis (Dasgupta, 2000).

2.2 Árvore Filogenética

As relações evolutivas entre os seres vivos são representadas pela Árvore Filogenética. De acordo com Amorim (2002), *"Árvore Filogenética (no sentido da Sistemática Filogenética) que é um dendrograma que expressa relações filogenéticas tanto entre táxons terminais, quanto entre espécies ancestrais e espécies descendentes"* .

Esse dendograma, também conhecido como árvore, contém raiz, ramificações, nós e terminais; No qual a raiz (espécie ancestral), comum de todas as espécies consideradas, gera ramificações (relações de parentescos) para os nós (espécies descendentes), que por sua vez, têm ramificações para outros nós e assim por diante até se atingir nós que não tem filhos, denominados de folhas.

A representação de uma árvore filogenética na figura 2.1 pode ser analisada da seguinte maneira: A partir da espécie comum esp_1, a qual contém as características c_1, c_2 e c_3, deu origem às espécies esp_2 e esp_3, estas têm, respectivamente, as características (c_2, c_3, c_4) e (c_1, c_3, c_5). A espécie esp_3 gerou as espécies esp_4 e esp_5, as quais têm, respectivamente, as características (c_3, c_4, c_7) e (c_1, c_5, c_6). E a espécie esp_5 gerou as espécies esp_6 e esp_7 com as características (c_4, c_5, c_6) e (c_1, c_5, c_6, c_7) (Pacheco, 2011).

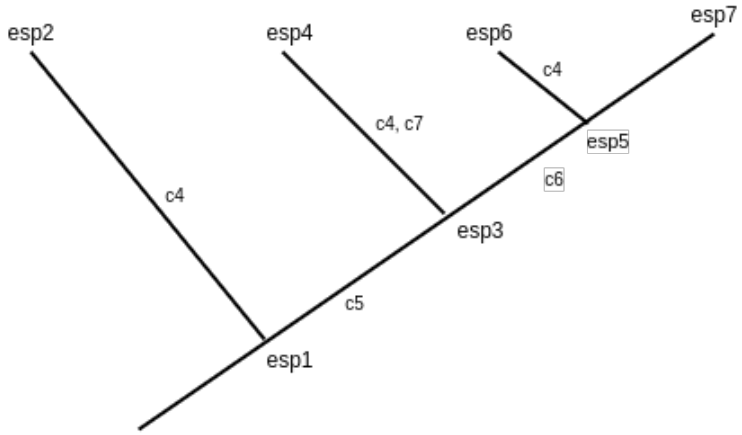

Figura 2.1: Exemplo de Árvore Filogenética

2.2.1 Definição Formal

Segundo a definição apresentada em Lopes (2003), uma árvore filogenética é uma estrutura F = (X, Y, φ , Γ) onde:

- X é um conjunto finito e não-vazio de espécies.
- Y é um conjunto finito das características apresentadas nos elementos de X.
- φ é uma relação de ordem sobre os elementos de X, na qual contém as seguintes propriedades:
  - Reflexiva: $\forall x \, \varepsilon \, X, \; x \, \varphi \, x$.
  - Assimétrica: $\forall x_1, x_2 \, \varepsilon \, X, \; se \; x_1 \, \varphi \, x_2, \; então \; not (x_2 \, \varphi \, x_1)$.
  - Transitiva: $\forall x_1, x_2 \; e \; x_3 \, \varepsilon \, X, \; se \; x_1 \, \varphi \, x_2 \; e \; x_2 \, \varphi \, x_3, então \; x_1 \, \varphi \, x_3$.
- $\Gamma : X \rightarrow P(Y)$: Em que é representada pelo arco G = (X, φ) com cada arco $a_k = (x_i, x_j)$ de G, rotulado pela sequência dos elementos do conjunto $\Sigma = \Gamma(x_i) - \Gamma(x_j)$.

Como exemplo de aplicação da formalização acima, temos uma árvore filogenética F = (X, Y, φ, Γ) , onde:

- $X = (esp_1, esp_2, esp_3, esp_4, esp_5, esp_6, esp_7)$;
- $Y = (c_1, c_2, c_3, c_4, c_5, c_6, c_7)$;
- $\varphi = \{(esp_1, esp_2), (esp_1, esp_3), (esp_3, esp_4), (esp_3, esp_5), (esp_5, esp_6), (esp_5, esp_7)\}$;
- As características presentes nas espécies são:
  - $\Gamma(esp_1) = (c_1, c_2, c_3)$;

- $\Gamma(esp_2) = (c_2, c_3, c_4)$;
- $\Gamma(esp_3) = (c_1, c_3, c_5)$;
- $\Gamma(esp_4) = (c_3, c_4, c_7)$;
- $\Gamma(esp_5) = (c_1, c_5, c_6)$;
- $\Gamma(esp_6) = (c_4, c_5, c_6)$;
- $\Gamma(esp_7) = (c_1, c_5, c_6, c_7)$.

2.2.2 Condições evolutivas

As características de uma espécie podem sofrer variações, podendo perdurar em seus descendentes, sendo esta modificação uma nova condição derivada presente na ancestralidade. A condição derivada é capaz de definir um novo grupo, sendo chamada de apomorfia (do grego, α*pó* ="longe de"e *morphé* ="forma"). Quando acontece em um grupo é denominada de autopormofia (do grego *autos* = “eu”, “mesmo” e *morphĕ* = “forma”). e ao ser compartilhada com mais de um grupo de sinapomorfia (do grego *sýnapsis* = “ação de juntar” e *morphĕ* = “forma”).

Nos casos em que não ocorrem mudanças nas características, é chamado de estado plesiomórfico (do grego *plesios* = “vizinho”, significando próximo e *morphé* = “forma”), e quando ocorre em mais de um grupo é conhecido como estado simplesiomórfico (do grego *syn* = “ação de juntar”, *plesios* = “vizinho”, significando próximo e *morphé* = “forma”).

As características semelhantes, análogas, devem ser discriminadas na análise filogenética para não gerar interpretações errôneas de que as espécies que as contém são parentes. Essas semelhanças estruturais não são decorrentes da evolução, mas acontecem devido ao paralelismo ou convergência evolutiva e são denominadas de homoplasia (do grego *homós* = “semelhante” e *plásis* = “formação”) (Simone, 2020).

Quando ocorre o desaparecimento de uma característica também é um fenômeno conhecido como homoplasia. Diversas características podem ser perdidas durante a evolução, e deve-se saber se ocorreu uma perda evolutiva ou se a característica nunca esteve presente na linhagem. Esses tipos de condições evolutivas são diferenciados pela sistemática filogenética.

2.3 Inferência Filogenética

Willi Hennig desenvolveu a Inferência Filogenética, método de reconstrução das relações de parentesco entre espécies com grupos de espécies. Estas relações de parentesco são obtidas da árvore filogenética (Cianciaruso, 2009).

No método da Inferência Filogenética, para construção da árvore filogenética é necessário selecionar o grupo de espécies e os caracteres mutáveis, produzir a matriz e observar o **Princípio de Parcimônia** para gerar a árvore com base na matriz (Teles Caldart, 2016).

2.4 Representação por matrizes

Com as espécies e as características selecionadas, constrói-se a matriz característica. A **matriz característica** contém valores binários e relaciona as espécies com as características terminais, se a característica estiver presente na espécie em questão, seu valor será 1 e, caso contrário, o valor será 0. Na Tabela 2.1 é exemplificada uma matriz característica.

Espécie	c1	c2	c3	c4	c5	c6	c7
esp1	1	1	1	0	0	0	0
esp2	0	1	1	1	0	0	0
esp3	1	0	1	0	1	0	0
esp4	0	0	1	1	0	0	1
esp5	1	0	0	0	1	1	0
esp6	0	0	0	1	1	1	0
esp7	1	0	0	0	1	1	1

Tabela 2.1: Matriz característica da árvore filogenética da Figura 2.1

Existem diversos tipos de matrizes para representar informações da evolução das características apresentadas por um conjunto de espécies. As principais são: matriz característica, matriz de similaridade, matriz de distância e matriz polarizada.

	esp1	esp2	esp3	esp4	esp5	esp6	esp7
esp1	0	5	5	3	3	1	2
esp2	5	7	3	5	1	3	0
esp3	5	3	7	3	5	3	4
esp4	3	5	3	7	1	3	2
esp5	3	1	5	1	7	5	6
esp6	1	3	3	3	5	7	4
esp7	2	0	4	2	6	4	7

Tabela 2.2: Matriz similaridade da árvore filogenética da Figura 2.1

	esp1	esp2	esp3	esp4	esp5	esp6	esp7
esp1	0	2	2	4	4	6	5
esp2	2	0	4	2	6	4	0
esp3	2	4	0	4	2	4	3
esp4	4	2	4	0	6	4	5
esp5	4	6	2	6	0	2	1
esp6	6	4	4	4	2	0	3
esp7	5	0	3	5	1	3	0

Tabela 2.3: Matriz distância da árvore filogenética da Figura 2.1

As **matrizes de similaridades e de distâncias** são obtidas a partir de análise da matriz característica. Estas matrizes informam a quantidade de vezes que a característica de uma espécie

é, respectivamente, igual ou diferente das demais espécies da matriz característica (Folkertsma, 2019).

Na Tabela 2.2 é demonstrada a matriz similaridade calculada com base na quantidade de caracteres semelhantes entre as espécies, enquanto na Tabela 2.3 é exibida a matriz de distância, sendo seus valores calculados com base nas diferenças entre os caracteres das espécies em questão. É perceptível que as matrizes de similaridade e de distância se complementam em seus valores.

Tome como exemplo as espécies esp_1 e esp_4 destacadas na Tabela 2.4. Para a matriz de similaridade são somados os caracteres com valores iguais, neste caso são os caracteres c_3, c_5 *e* c_6, sendo a similaridade entre esp_1 e esp_4 igual a 3, sendo a relação inversa de mesmo valor. No cálculo da distância entre as espécies esp_1 e esp_4, são considerados os caracteres diferentes, sendo estes c_1, c_2, c_4 *e* c_7, esse valor também é válido para a relação inversa.

Espécie	c1	c2	c3	c4	c5	c6	c7
esp1	1	1	1	0	0	0	0
esp4	0	0	1	1	0	0	1

Tabela 2.4: Matriz característica com foco nas espécies esp_1 e esp_4 da Tabela 2.1

Matriz polarizada quantifica a quantidade de vezes em que a transformação/evento ocorre nas espécies estudadas.

A partir da matriz característica, observa-se as distâncias entre as espécies, considerando que o objetivo do Princípio da Parcimônia é escolher a espécie com menor distância com relação à espécie atual, de forma a ter a árvore com menores transformações.

2.5 Parcimônia

"O termo Parcimônia (do latim parco, simples) foi utilizado para indicar a árvore que demandava o menor número de modificações, entre as árvores possíveis, pois corresponde, de um certo ponto de vista, à hipótese mais econômica." (Amorim, 2002).

Existem diversas formas de calcular as distâncias entre as espécies, a seguir serão descritas as distâncias de Hamming, P e Média.

Distância de Hamming

A Distância de Hamming (Kim and Lee, 1999), consiste em quantificar a quantidade de caracteres diferentes em duas sequências de mesmo alfabeto, conforme mostra a Equação 2.1.

$$d_{A,B} = \sum_{i=j=1}^{n} v_{i,j} \tag{2.1}$$

,onde $v_{i,j} = \{^{1\ se\ A \neq B}_{0\ caso\ contrrio}$ e n = Tamanho da sequência

A matriz de distância exemplificada na Tabela 2.3 foi realizada utilizando a Distância de Hamming.

Distância P

Este critério calcula a distância relativa entre duas espécies (Teles Caldart, 2016), pois calcula a quantidade de caracteres diferentes entre duas bases e depois divide pelo tamanho da base. Sendo representada pela Equação 2.2.

$$d_{A,B} = \frac{\sum_{i=j=1}^{n} v_{i,j}}{n} \tag{2.2}$$

,onde $v_{i,j} = \{^{1\ se\ A \neq B}_{0\ caso\ contrrio}$ e n = Tamanho da sequência

Distância Média

A Distância Média (Webb, 2000) calcula a distância filogenética média entre todas as combinações de pares de espécies, ou seja, realiza a média do somatório dos caracteres diferentes.Sendo representada pela equação 2.3.

$$d_{A,B} = \frac{\sum_{i=j=1}^{n} v_{i,j}}{2} \tag{2.3}$$

,onde $v_{i,j} = \{^{1\ se\ A \neq B}_{0\ caso\ contrrio}$ e n = Tamanho da sequência

Realizar a construção de Árvore Filogenética em várias espécies é relativamente exaustivo, pois a relação de quantidade de espécies e características aumentam, é proporcional ao tamanho da matriz característica. Dessa forma, existem algoritmos para a geração de parentesco entre espécies.

3

Métodos para Construção de Árvore Filogenética

3.1 Surgimento da Computação Evolucionária

Os primeiros sistemas de comparação foram baseados em um conjunto pequeno de características morfológicas e fisiológicas, selecionados subjetivamente. Como algumas características eram relevantes para o estabelecimento do agrupamento, as soluções resultantes não podiam ser aceitas de forma absoluta (Gomes, 1998).

Por volta da metade do século XX, cientistas perceberam que haviam problemas complexos ainda não resolvidos computacionalmente, por não ter descrição detalhada e solução aceitável em tempo hábil (Bittencourt, 1998).

Então, foi definido o conceito de classificação natural, o qual afirma que um grupo de espécies é formado quando estes elementos compartilham o maior número possível de características (Von Zuben, 2000).

Segundo Gomes (1998), os sistemas de classificação que se fundamentam nas similaridades das espécies, sem considerar a ancestralidade comum, são denominados fenéticos.

Conforme a Escola Filogenética definiu conceitos da sistemática, como espécie, clado e parentesco. Pesquisadores desenvolveram propostas para a geração da árvore filogenética de acordo com as ideias da Escola.

3.2 Classificação dos Métodos de Reconstrução

Métodos de reconstrução podem ser classificados em métodos não baseados em modelos e métodos baseados em modelos (Hillis, 1996).

Os métodos não baseados em modelos, seguem algoritmos para determinar a árvore desejada para o problema em questão. Esta abordagem é utilizada em casos que requerem tempo curto de resposta. Conforme a árvore é definida de acordo com os algoritmos, então a solução consiste em inferência, podendo ser aceita ou rejeitada.

Os métodos baseados em modelos, as árvores candidatas são avaliadas, de acordo os critérios definidos, até encontrar a árvore de maior verossimilhança, durante esse processo uma variedade de árvores são analisadas (Weir, 1996). Neste método, ao definir o critério, é possível operar com várias árvores, o que permite maior flexibilidade na escolha da árvore, porém, para casos genéricos, poderá produzir soluções não ótimas.

Estes métodos serão descritos nas seções 3.3 e 3.4.

3.3 Métodos não baseados em modelos

Segundo Hillis (1996), esta abordagem se baseia em critérios de otimização, sem precisar de modelos para elaboração. Assim, estes algoritmos têm desempenho maior do que os métodos baseados em modelos.

Os Métodos não baseados em modelos são subdivididos em algoritmos que usam a matriz distância e máxima parcimônia. A seguir, estes algoritmos serão descritos (Godini, 2019).

3.3.1 Matriz distância

Os algoritmos desta categoria partem da matriz de distância construída a partir do cálculo de distância entre os pares. A matriz é processada por algoritmos de agrupamento até produzir a árvore final. O resultado depende da forma que a matriz de distância foi desenvolvida, então é importante definir o critério do cálculo de distância (Saif, 2023).

Algoritmo da Média Aritmética Não-Ponderada - UPGMA

Algoritmo desenvolvido para construção de fenogramas, diagramas nos quais representam as características fenotípicas das espécies, porém pode ser utilizado para construção de árvores, contanto que as taxas de evolução entre as linhagens sejam constantes (Prosdocimi, 2001).

UPGMA constrói uma das árvores filogenéticas possíveis, a partir da matriz distância de um conjunto de espécies. Neste algoritmo, os ramos não contém rótulos. A seguir, são apresentados os passos deste algoritmo, segundo Andreatta (2002).

1. Tome o par de espécies com a menor distância entre si e agrupe-os em uma super espécie, este par terá uma espécie em comum S_1.

2. Recalcule a distância das demais espécies S_i para S_1, como sendo a média das distâncias de S_i para cada super espécie.

3. Enquanto houver duas ou mais super espécies não visitadas , volte ao passo 1.

Um dos problemas neste algoritmo é que considera que a distância entre quaisquer nós folhas até a raiz é a mesma e a árvore só é produzida corretamente caso a taxa de evolução seja ,aproximadamente, constante. Os algoritmos, a seguir, podem lidar com essas limitações.

Método dos Mínimos Quadrados

O método dos Mínimos Quadrados é uma técnica de otimização matemática que busca encontrar o melhor ajustamento possível para um conjunto de dados, de forma a minimizar a soma dos quadrados das diferenças entre os valores esperados e os valores obtidos (Simoncelli, 2003).

Aplicando esta técnica na construção de árvore filogenética, o problema se torna aproximar uma árvore candidata na árvore real. Cavalli-Sforza (1967) foi o primeiro a implementar esta ideia na geração de árvore filogenética. Os passos desse algoritmo são:

1. Calcular a matriz distância, será considerada como distância esperada
2. Aplicar a matriz distância em um algoritmo de agrupamento, para obtenção de uma árvore observada
3. As distâncias observadas serão a soma dos pesos nas arestas da árvore observada
4. É realizado o cálculo do quadrado mínimo com as distâncias esperada e observada
5. Retorna ao passo 2, até a obtenção de valor aceitável da equação dos quadrados mínimos.

É fácil observar que o desempenho será custoso, pois utiliza algoritmo auxiliar para geração da árvore observada. Além da importância da forma de definição da matriz distância para a solução.

Método de Evolução Mínima

Este é outro método muito difundido na literatura, o Método de Evolução Mínima parte do princípio de que quanto menor a quantidade de passos para explicar a evolução de um conjunto de espécies , mais próximo estará da árvore real (Catanzaro, 2022).

O Método de Evolução Mínima calcula a soma dos comprimentos de todas as árvores candidatas, de forma a selecionar a árvore com menor soma (Nei, 2000). Este algoritmo requer alto custo computacional, visto que avalia o somatório dos ramos de todas as soluções, para então escolher a melhor solução.

Método Neighbor-Joining

Como complementação ao Método de Evolução Mínima, (Saitou, 1987) propôs limitar o número de avaliações dessa topologia.

A ideia desse algoritmo é que, a partir da matriz distância, cria-se o nó central com as espécies na periferia em forma de estrela. Depois, são realizados agrupamentos até que todas as espécies estejam agrupadas.

É notável que a árvore resultante não contém raiz, ou seja, não tem parentesco comum a todas as espécies, assim as espécies se relacionam em termos de mudanças ocorridas e não com relação ao tempo evolutivo.

Algoritmo de Wagner

O algoritmo de Wagner (Farris, 1970) é capaz de construir árvores filogenéticas de um conjunto de espécies, a partir da matriz característica das espécies em questão. Diante disso, serão demonstrados os passos desse algoritmo, segundo Altshull (1997) e Vieira and Lopes (1999).

1. Especificar a espécie raiz da árvore
2. Construir a matriz distância das espécies fornecidas na matriz característica
3. Selecionar a espécie que tiver menor distância com relação a espécie raiz
4. Criar ramo da raiz até a espécie selecionada conforme a distância entre elas
5. Selecionar a próxima espécie de menor distância com relação a espécie raiz
6. Calcular a distância entre a espécie selecionada e as demais espécies
7. Selecione a espécie que tiver menor distância no passo anterior e atual espécie selecionada, esta será a espécie irmã selecionada
8. Criar ramo ligando a atual espécie selecionada S até ao meio do ramo da espécie irmã selecionada I, de comprimento calculado pela equação abaixo:

   $$Comp(S,I) = (Dist(S,I) + Dist(S,Ancestral(I))) - (Dist(I,Ancestral(I))$$

9. Determine o vetor característica do ancestral comum A, $caract_A = (c_{A,1},...,c_{A,n})$ entre a espécie S e a espécie I, onde $c_{A,i} = menor(c_{x,i},c_{y,i})$ e $c_{x,y}$ é o elemento da linha x e coluna y da matriz característica C
10. Enquanto existir uma espécie ainda não selecionada , volte ao passo 5

Os passos deste algoritmo são demonstrados na Figura 3.1 com o uso da tabela 3.1. Na Figura em questão, é possível visualizar que as espécies são agrupadas de acordo com a distância mínima em relação ao ancestral.

Espécie	c1	c2	c3	c4	c5	c6
A	0	0	0	0	0	0
B	1	0	0	0	0	0
C	1	1	1	0	1	0
D	1	1	1	1	0	1

Tabela 3.1: Matriz Característica

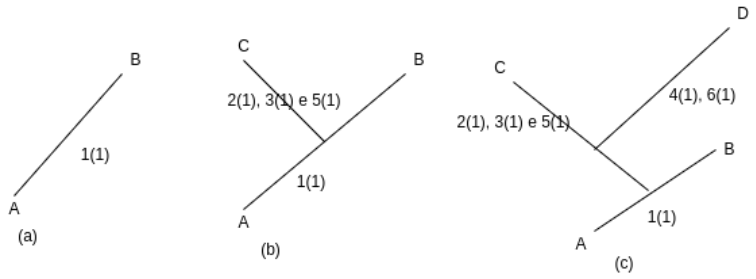

Figura 3.1: Árvores filogenéticas construídas pela primeira (a), segunda (b) e terceira execução do passo 5 do algoritmo de Wagner para a matriz característica 3.1

A regra da Inclusão e exclusão

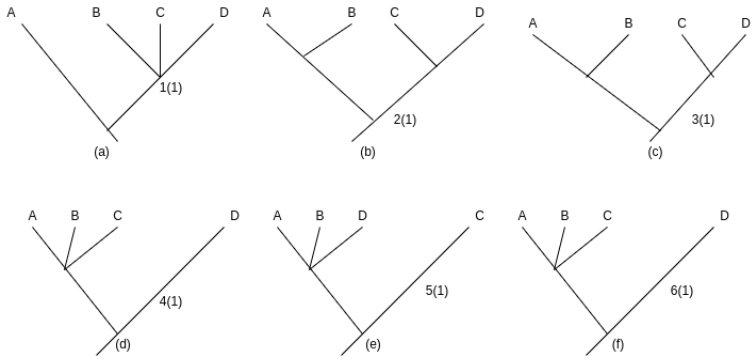

Figura 3.2: Árvores filogenéticas para as características terminais da tabela 2.1

Este algoritmo recebe a matriz característica de um conjunto de espécies e retorna o conjunto de todas as possíveis árvores filogenéticas dessas espécies. Serão apresentados ,a seguir, os passos desse algoritmo (Wiley, 1991).

1. Para cada característica terminal N considerada no estudo, será construída uma árvore filogenética AF_i , com $1 < i <$ número de características terminais consideradas. Na tabela 2.1 temos seis caraterísticas terminais , logo serão construídas seis árvores filogenéticas, conforme Figura 3.2.

2. Combine todas as relações de parentescos presentes nas AF_i construídas no passo anterior, de forma que, as relações de parentescos de AFi sejam preservadas pela adição das relações de parentescos presentes na AF_j com $i < j$, árvore gerada após este passo pode ser visualizada na Figura 3.3.

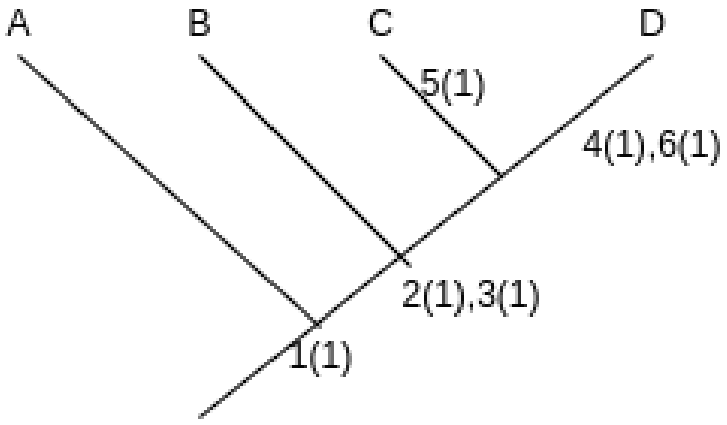

Figura 3.3: Árvore filogenética resultante da combinação das árvores acima

3.3.2 Máxima Parcimônia

Proposta por Cavalli-Sforza (1967), o segundo Método Não Baseado em Modelo, afirma que em contextos que têm um objetivo, a escolha deve ser a mais simples. Segundo Buso (2005), a parcimônia é um método baseado no estado do caráter, que realiza inferência com objetivo de minimizar o número total de passos evolutivos para explicar um conjunto de espécies.

3.4 Métodos baseados em modelos

3.4.1 Método da Verossimilhança Máxima

O método da Verossimilhança Máxima propõe uma forma de inferência que utiliza todas as informações disponíveis para a construção da árvore filogenética (Fisher, 1922). A inferência é realizada a partir de modelos probabilísticos (Felsenstein, 1981).

É definida uma topologia para a árvore, e a partir da topologia, selecionam os comprimentos dos ramos de maneira a maximizar a probabilidade dos dados analisados representarem o conjunto de espécies em estudo. Estas probabilidades são comparadas com as probabilidades de outras topologias, e aquela árvore que apresentar a maior probabilidade é considerada a mais verossímil (Nei, 2000).

O cálculo da Verossimilhança leva em consideração os caracteres e as possibilidades de mutações em todos os nós internos da árvore resultante (Weir, 1996).

3.4.2 Problema nos métodos de construção da árvore filogenética

Após descrição do funcionamento dos métodos baseados e não baseados em modelos, é perceptível gargalos de desempenho e resultados. Nos métodos baseados em modelos, percebe-se que o desempenho é prejudicado devido ao fato de avaliar todas as possibilidades de mutações possíveis (Buso, 2005).

Enquanto nos métodos não baseados em modelos, o parentesco depende do método de geração da matriz distância. Nestes algoritmos, principalmente o algoritmo de Wagner, quando seleciona a espécie de menor distância e existem mais de uma espécie com a menor distância, escolhe-se a primeira espécie na ordem disposta na matriz distância. Tal fato, não considera as demais espécies com distância menor e igual à espécie selecionada, o que pode resultar em árvore solução com parentescos errôneos (Godini, 2019).

No momento em que ocorre o Princípio da Parcimônia, o algoritmo utiliza a Distância de Hamming para definir a distância da atual espécie para as demais e, assim, selecionar a espécie mais próxima (Farris, 1970). Diante do contexto exposto, o Algoritmo de Wagner foi modificado para escolher a próxima espécie com base nas distâncias de Hamming, P e Média com auxílio da máxima similaridade.

Então, quando estiver no impasse de várias espécies com menor distância, será escolhida a espécie com menor distância em mais de uma forma de cálculo desta. E caso, existam mais de uma espécie com a menor distância durante a seleção por cada método, será consultada qual destas espécies têm maior semelhança com a espécie anterior.

No próximo capítulo será apresentado o sistema que faz uso do conceito de espécie vencedora em mais de uma parcimônia e máxima similaridade.

4

Metodologia

O sistema foi desenvolvido para construção de árvores filogenéticas, a partir de matriz característica de um conjunto de espécies. O código foi implementado utilizando o framework Django, que faz uso da linguagem Python, para as iterações do usuário com a matriz na tela foi implementada com JavaScript e os dados foram armazenados no SGBD (Sistema de Gerenciamento de Banco de Dados) SQLite3.

A leitura da matriz característica é realizada por meio da interface do sistema, dependendo da escolha do usuário se realizará upload de arquivo ou desenhar a matriz na própria interface. Com os dados necessários, a árvore será construída pelo Algoritmo de Wagner com o auxílio das 3 formas de decidir a Parcimônia e máxima similaridade, conforme mostra a Figura 4.1.

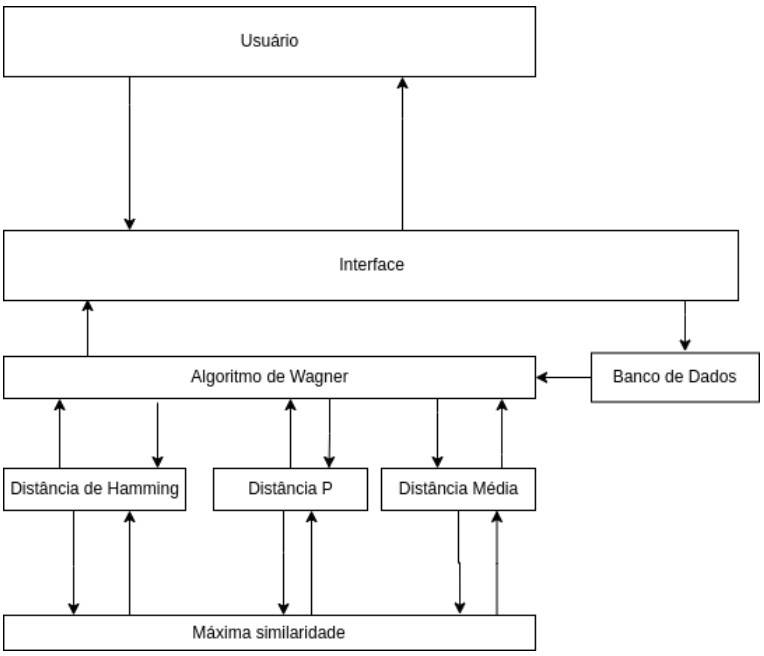

Figura 4.1: Arquitetura da solução implementada

4.1 Elementos

Os elementos da arquitetura apresentada na Figura 4.1 são:

- **Usuário**: O usuário escolhe a matriz característica a ser utilizada como base para a construção da árvore filogenética.
- **Interface**: Realiza leitura da matriz característica a partir do arquivo ou dados enviados pelo usuário, definindo as informações necessárias para o Algoritmo.
- **Algoritmo de Wagner**: É a abordagem que define a árvore filogenética, a partir das informações advindas da interface. Realiza sucessivos agrupamentos até que todas as espécies estejam agrupadas.
- **Distância de Hamming**: É o método mais utilizado, mede a distância entre duas espécies pelo somatório da diferença entre os itens de duas espécies.
- **Distância P**: Mede a distância relativa entre dois grupos, calculando o somatório da diferença entre os caracteres pertencentes a duas espécies dividido pela quantidade de caracteres das espécies. Nesse caso, as espécies precisam ter o mesmo tamanho de sequência de caracteres.
- **Distância Média**: Nesta abordagem , calcula-se o somatório da diferença entre as sequências de caracteres de duas espécies, sendo estes caracteres do mesmo alfabeto. Após, a distância é definida computando a média do somatório.
- **Máxima similaridade**: Elemento responsável por selecionar, dentre as espécies escolhidas pelo método de distância em questão, a espécie que tem maior semelhança com a espécie anterior. Tendo como espécie selecionada aquela que tem menor distância e maior semelhança com seu ancestral.

4.2 Funcionalidades

Como apresentado na Figura 4.2, o usuário seleciona qual método de entrada das informações utilizará, podendo ser inserção de arquivo contendo a matriz característica ou desenhar a matriz a partir da interface.

Com relação ao arquivo, Figura 4.3, este deve ser um arquivo no formato txt com as características das espécies, em cada linha, representada por 0's e 1's , separados por vírgula conforme exibido na Figura 4.4.

Para formar a matriz característica no sistema, é necessário definir a quantidade de espécies e características, depois clicar nos campos em que a característica existe para a espécie em questão, conforme mostrado na Imagem 4.5.

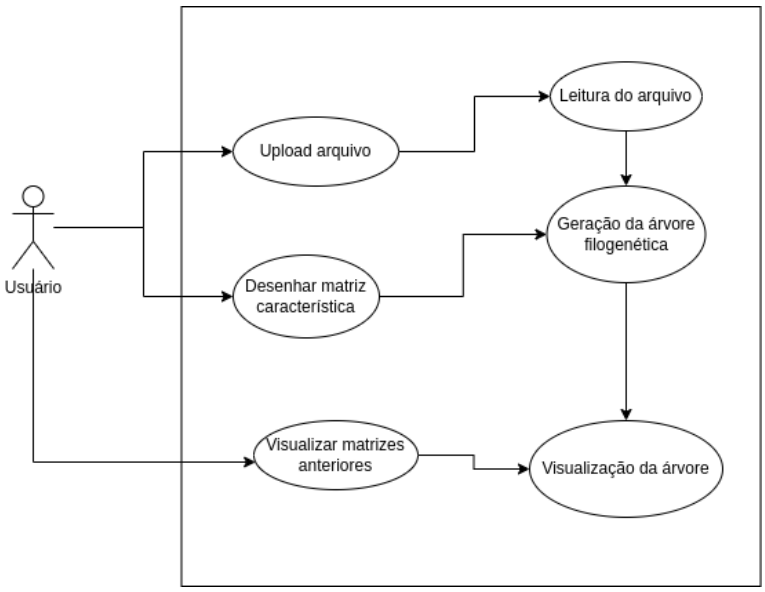

Figura 4.2: Casos de uso do sistema desenvolvido

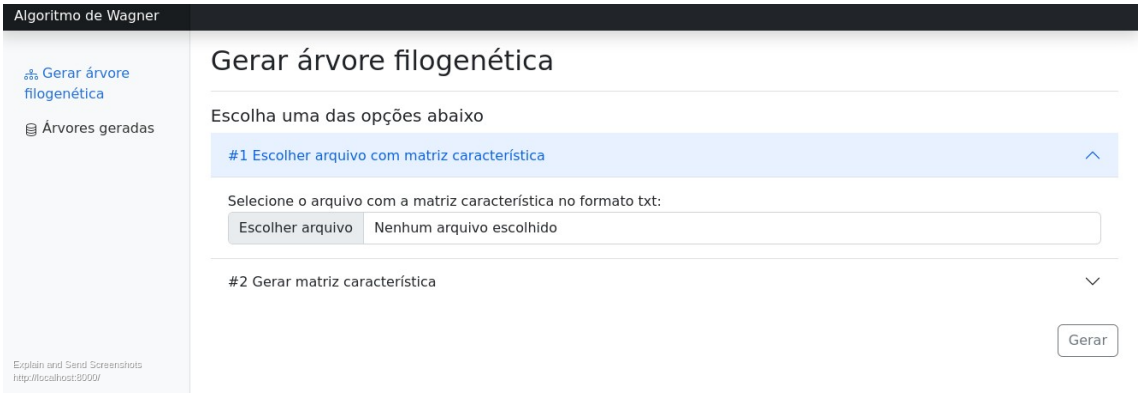

Figura 4.3: Opção de upload de arquivo contendo a matriz característica

Após, determinar a matriz, seja por arquivo ou pela interface, basta apertar em '*gerar*'. É realizada a leitura dos dados, determinando as quantidades de espécies e de características, definidas as matrizes de distância de Hamming e de similaridade. A matriz característica ou o caminho do arquivo que contém a matriz, quantidade de espécies e características são salvas no banco de dados para posterior solicitação de visualização destes dados.

O algoritmo recebe as informações, adiciona super espécie, esta é um ancestral em comum auxiliar, para relacionar as espécies. Em seguida, cada uma das distâncias utilizadas selecionará uma espécie de menor distância, caso existam várias espécies com a menor distância, é selecionada a espécie com maior semelhança com a espécie ancestral.

Após os métodos de distância escolherem suas respectivas espécies de menor distância, é realizada classificação por frequência destas espécies. Desta classificação, a espécie que teve maior frequência, ou seja, o clado que for mais selecionado pelos métodos de cálculo de distância, será o grupo escolhido. Tendo finalizada a escolha da Parcimônia neste passo, esta espécie será filha da super espécie adicionada anteriormente, depois é inserida outra super espécie e

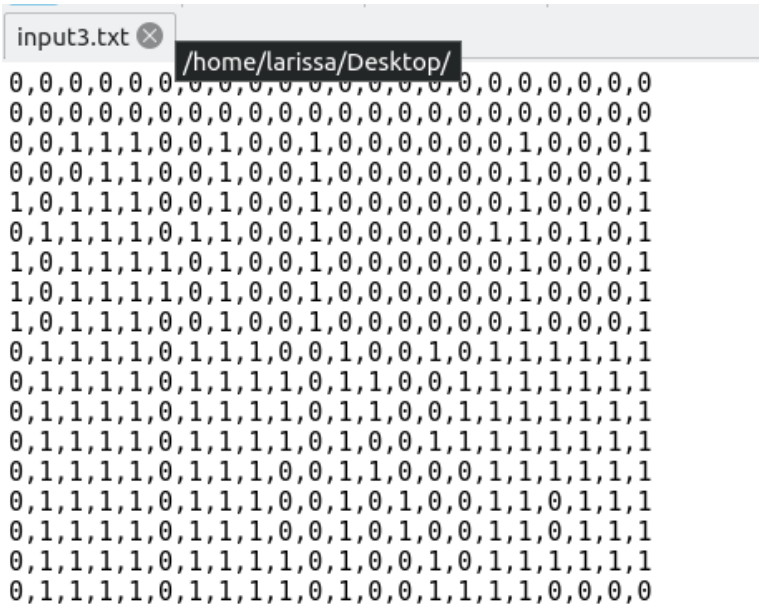

Figura 4.4: Exemplo de matriz característica em arquivo

retorna ao passo de escolher a espécie mais próxima. A repetição continua até não ter mais espécies a serem escolhidas.

Definida a árvore solução para o conjunto de espécies em questão, a solução e a matriz característica são enviadas para visualização do usuário, Figura 4.6.

4.3 Estudo de caso real: Gênero *Edessa*

Este gênero é um grande grupo de percevejos, contendo mais de 250 espécies descritas. No estudo de caso, foram utilizadas 18 espécies contendo 22 características na matriz característica. Na Figura 4.4, é exibida a matriz característica em questão no arquivo a qual foi gerada com os dados da Tabela 4.1 (Silva et al., 2012).

Neste caso, ocorrem alguns momentos em que existem várias espécies com a menor distância, tomemos como exemplo o momento que se escolhe o nó descendente ao nó *esp16*, durante a escolha pela Distância de Hamming, a distância entre o nó anterior e as espécies disponíveis *esp10, esp11* e *esp12* são todas 17, neste momento o algoritmo verifica qual destas espécies tem a maior semelhança com a *esp16* que é a *esp12*, conforme exibido na figura 4.9.

Caso, não houvesse a classificação e a consulta de máxima similaridade, o Algoritmo de Wagner escolheria a primeira espécie apenas pela verificação da distância, *esp10*.

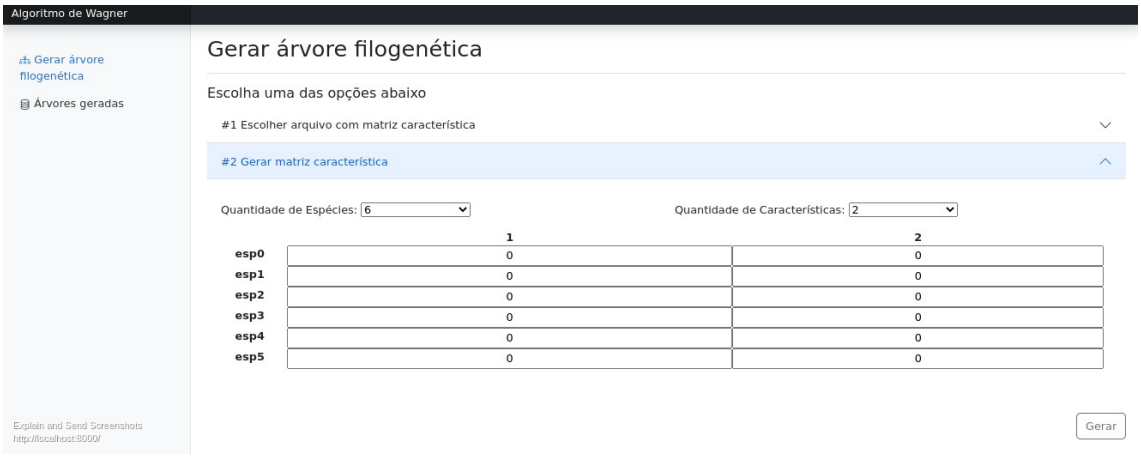

Figura 4.5: Opção de definir a matriz característica na tela

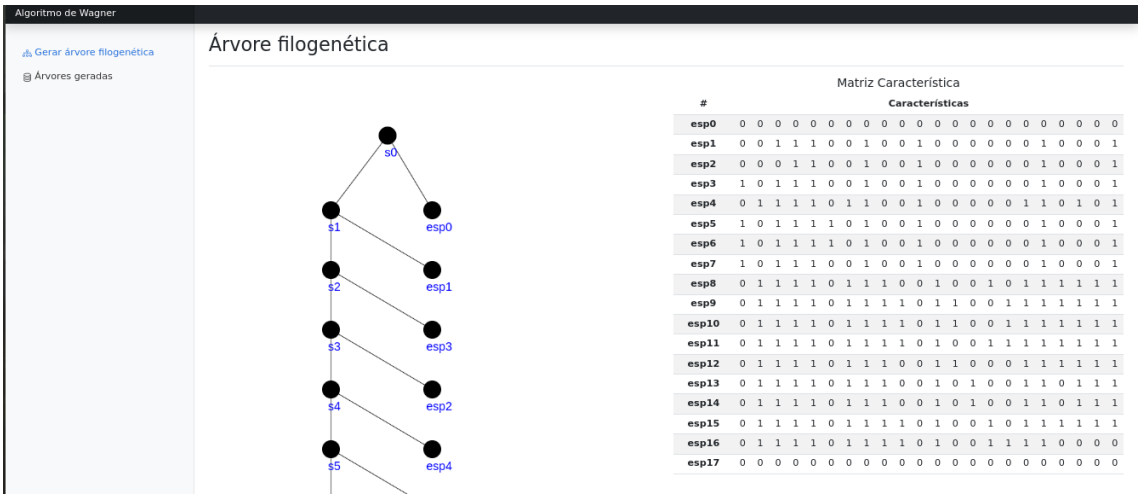

Figura 4.6: Visualização de geração de árvore filogenética da figura 4.4

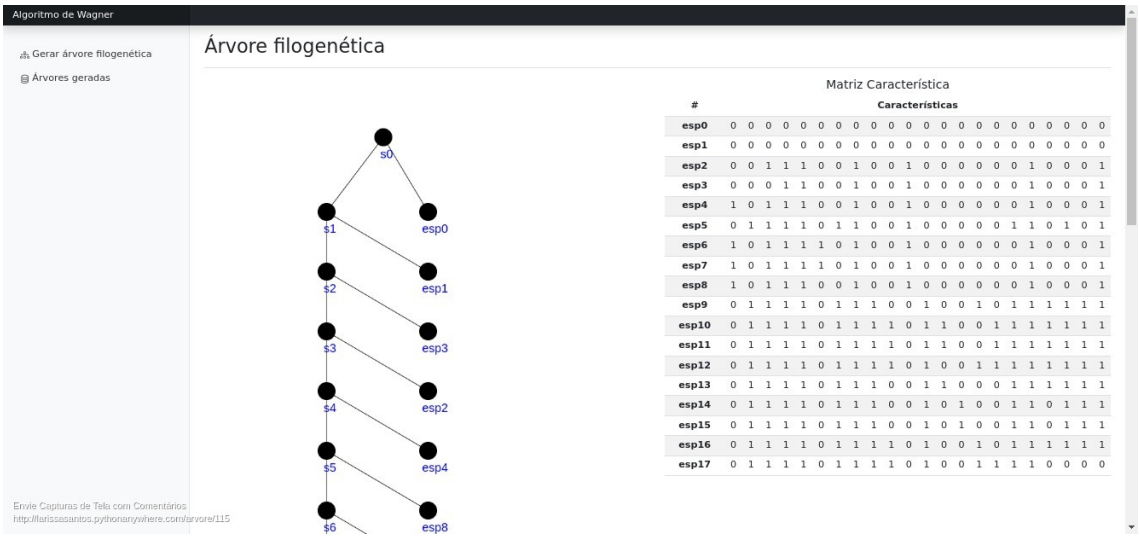

Figura 4.7: Árvore filogenética e matriz característica do gênero *Edessa*

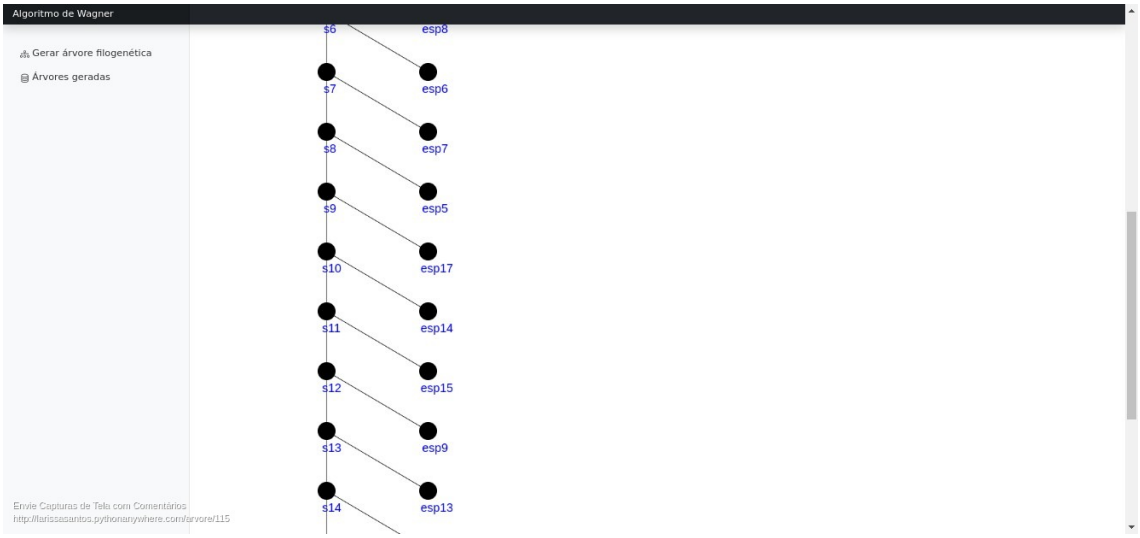

Figura 4.8: Continuação da árvore filogenética do gênero *Edessa*

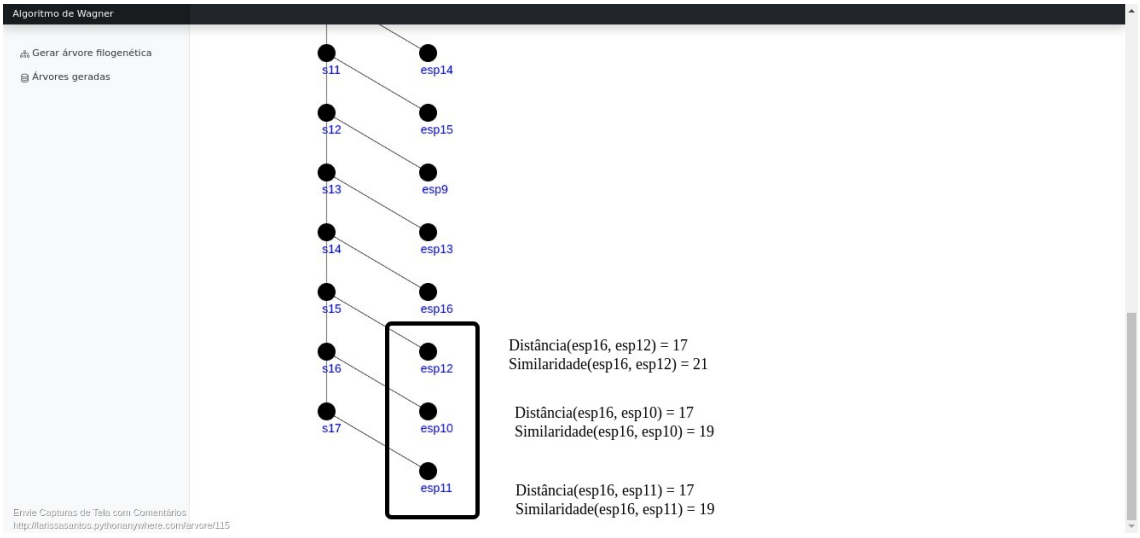

Figura 4.9: Continuação da árvore filogenética do gênero *Edessa*, com destaque para as distância de Hamming e similaridade de algumas espécies

Espécie	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Tibilis sp.	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Neotibilis fulvicornis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pantochlora vivida	0	0	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
Brachystethus cribrus	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
Edessa cervus	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
Edessa affinis	0	1	1	1	1	0	1	1	0	0	1	0	0	0	0	0	1	1	0	1	0	1
Peromatus sp.	1	0	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
Olbia elegans	1	0	1	1	1	1	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
Doesburgedessa elongatispina	1	0	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	1
Edessa stolida	0	1	1	1	1	0	1	1	1	0	0	1	0	0	1	0	1	1	1	1	1	1
Edessa verhoeffi	0	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1
Edessa heymonsi	0	1	1	1	1	0	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1
Edessa paravinula	0	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1	1	1	1	1	1	1
Edessa sp. nov 2	0	1	1	1	1	0	1	1	1	0	0	1	1	0	0	0	1	1	1	1	1	1
Edessa sp. nov 3	0	1	1	1	1	0	1	1	1	0	0	1	0	1	0	0	1	1	0	1	1	1
Edessa sp. nov 3a	0	1	1	1	1	0	1	1	1	0	0	1	0	1	0	0	1	1	0	1	1	1
Edessa sp. nov 4	0	1	1	1	1	0	1	1	1	1	0	1	0	0	1	0	1	1	1	1	1	1
Edessa sp. nov 5	0	1	1	1	1	0	1	1	1	1	0	1	0	0	1	1	1	1	0	0	0	0

Tabela 4.1: Matrix característica do gênero *Edessa*

5

Conclusão

A classificação dos seres vivos remonta desde a Grécia Antiga com teorias de filósofos como Aristóteles e Platão. Agrupar organismos começou de forma subjetiva, todavia diversos pesquisadores contribuíram para o desenvolvimento desta ação.

A medida que mais pesquisadores se envolviam na área, conceitos foram definidos, podendo assim, serem validados e aplicados em algoritmos.

Além disso, conforme a quantidade de espécies e caracteres em estudo, a probabilidade de resultado impreciso aumentava, então percebeu-se a importância do uso de algoritmos para classificar as espécies e otimizar o tempo dos pesquisadores envolvidos.

Os pesquisadores passaram a utilizar a árvore filogenética em diversas áreas, sendo Biotecnologia uma destas áreas. Este ramo da Biologia faz uso das relações de parentesco entre espécies para alterar os caracteres de um determinado produto.

Devido a complexidade de construção das árvores filogenéticas, os algoritmos produzem resultados aceitáveis. Dentre estes algoritmos, existe o Algoritmo de Wagner, no qual o parâmetro distância apresenta impasse quando existem mais de uma espécie com a menor distância, tendo como consequência a escolha da primeira espécie nessa situação.

5.1 Análise e discussão dos resultados

Foram realizados testes para algumas teorias até chegar na meta-heurística proposta. Inicialmente, observou-se que o problema das espécies com menor distância poderia ser detectado analisando a matriz distância antes da árvore ser construída, podendo-se neste caso, gerar as árvores de cada espécie no momento do impasse de distância. Depois, analisar qual árvore tem menor distância total, todavia o desempenho seria custoso e o resultado poderia ser incompatível com a realidade.

Na teoria seguinte, consistia em utilizar o Algoritmo de Wagner, pois este algoritmo usa a inferência filogenética e a matriz distância. A alteração proposta seria, quando for selecionar

a próxima espécie,ordenar a lista de espécies em ordem crescente de distância, escolhendo o primeiro grupo, essa teoria foi descartada, pois o impasse de espécies com menor distância iguais permanecia.

Então, a teoria anterior foi alterada para quando o Algoritmo de Wagner for selecionar a próxima espécie, utilizar 3 formas de parcimônias, ao invés de 1 parcimônia. Cada parcimônia seleciona as espécies candidatas e, depois, as candidatas são ranqueadas por quantidade de vezes selecionada pelas parcimônias. Dessa forma, o grupo selecionado seria o primeiro do ranking de espécies com a menor distância. Foram usadas 3 formas de parcimônias, porque era necessário uma quantidade ímpar para não gerar empate durante a classificação.

Durante os testes da metaheurística acima, observou-se que na escolha da parcimônia poderiam existir espécies com mesma distância e ser selecionada a primeira espécie disposta, para resolver tal impasse, acrescentou-se a verificação de similaridade das espécies com a menor distância, sendo selecionada o grupo com maior similaridade.

Foi demonstrado que aplicando os mecanismos de três parcimônias e máxima similaridade em espécies de menor e iguais distâncias, gerando ranking destas espécies, o que auxilia o Algoritmo de Wagner a escapar de máximos globais ruins, resultando numa árvore solução mais próxima da realidade.

5.2 Sugestões de trabalhos futuros

As seguintes funcionalidades não foram implementadas nesta versão do sistema:

- Explicação de cada passo tomado
- Formas de exportar a árvore solução
- Adicionar os nomes das espécies
- Adicionar nome dos conjuntos de espécie a serem analisados.

Estas ações podem auxiliar o pesquisador na usabilidade do sistema e tomada de decisões.

Nesta dissertação, foram propostos mecanismos de combinação para melhorar a tomada de decisão do Algoritmo de Wagner, quando a matriz característica apresenta conflitos, tendo como consequência a produção de árvores filogenéticas realistas.

5.3 Trabalhos Futuros

Com relação aos trabalhos futuros, pode-se destacar:

- Testes comparativos com implementações de outras abordagens para verificar desempenho e qualidade das soluções encontradas

- Comparar com a árvore gerada por outras técnicas
- Mesclar a geração de árvore filogenética com outras técnicas de programação.

Referências bibliográficas

Altshull, Stephen F e Madden, T. L. e. S. A. A. e. Z. J. e. Z. Z. e. M. W. e. L. D. J. (1997). Gapped blast and psiblast: a new generation of protein database searche programs. *Nucleic Acids Res*, 25(17):3.

Amorim, D. d. S. (2002). Fundamentos de sistemática filogenética. In *Fundamentos de sistemática filogenética*, pages 154–154.

Andreatta, Alexandre A e Ribeiro, C. C. (2002). Heuristics for the phylogeny problem. *Journal of Heuristics*, 8(4):429–447.

Barbosa, Flávia Rabelo e Souza, T. e. W. A. e. C. A. G. e. Q. E. D. e. A. F. J. L. e. F. J. (2021). Feijão resistente ao mosaico-dourado. *Santo Antônio de Goiás: Embrapa Arroz e Feijão*.

Bittencourt, G. (1998). *Inteligência artificial: ferramentas e teorias*. Série didática. Editora da UFSC.

Buso, G. (2005). Marcadores moleculares e análise filogenética.

Catanzaro, D. e. M. F. e. O. G. e. R. P. (2022). A tutorial on the balanced minimum evolution problem. *European Journal of Operational Research*, 300(1):1–19.

Cavalli-Sforza, Luigi L e Edwards, A. W. (1967). Phylogenetic analysis. models and estimation procedures. *American journal of human genetics*, 19(3 Pt 1):233.

Cianciaruso, Marcus Vinicius e Silva, I. A. e. B. M. A. (2009). Diversidades filogenética e funcional: novas abordagens para a ecologia de comunidades. *Biota Neotropica*, 9(3):93–103.

Darwin, C. e Pimenta, P. e. C. A. (2018). *Origem das espécies: ou A preservação das raças favorecidas na luta pela vida*. Ubu Editora.

Dasgupta, Bhaskar e He, X. e. J. T. e. L. M. e. T. J. e. Z. L. (2000). On distances between phylogenetic trees. *SODA*.

Farris, J. S. (1970). Methods for computing wagner trees. *Systematic Biology*, 19(1):83–92.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17(6):368–376.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368.

Folkertsma, H. e. M. e. A. (2018-2019). Comparing phylogenetic trees: an overview of state-of-the-art methods. *SC@ RUG*, 16th:14.

Frezzatti Júnior, W. A. (2011). A construção da oposição entre lamarck e darwin e a vinculação de nietzsche ao eugenismo. *Scientiae Studia*, 9(4):791–820.

Godini, R. e. H. F. (2019). A brief overview of the concepts, methods and computational tools used in phylogenetic tree construction and gene prediction. *Meta Gene*, 21:100586.

Gomes, C. d. A. (1998). A biologia molecular e os procariontes. *Acção de Formação"Biologia Molecular e a Humanidade"do Projecto"TRENDS*.

Hillis, DM e Moritz, C. e. M. B. (1996). Molecular systematics, sinauer. *Sunderland, Massachusetts*.

Kim, D. K. and Lee, Jee-Soo e Park, K. e. C. Y. (1999). Efficient algorithms for approximate string matching with swaps. *journal of complexity*, 15(1):128–147.

Klepka, Verônica e Corazza, M. (2018). O essencialismo na classificação de lineu e a repercussão dessa controvérsia na biologia. *História da Ciência e Ensino: construindo interfaces*, 18:73–110.

Lopes, R. V. V. (2003). Um algoritmo genético baseado em tipos abstratos de dados e sua especificação em z.

Mayr, E. (2005). *Biologia, ciência única*. Companhia das Letras.

Milagres, Regina Célia Rodrigues Miranda e Nunes, L. C. e. P.-S. H. M. (2007). A deficiência de vitamina a em crianças no brasil e no mundo. *Ciência & Saúde Coletiva*, 12:1253–1266.

Nei, Masatoshi e Kumar, S. (2000). *Molecular evolution and phylogenetics*. Oxford University Press, USA.

Pacheco, L. H. L. (2011). *Um sistema evolutivo para a construção da taxonomia dos seres vivos (SETAX)*. UFAL.

Prosdocimi, F. (2001). Medindo distâncias evolutivas. *Disponıvel em: http://www. icb. ufmg. br. Acessada em*, 13(11):2006.

Sadava, D. e Hillis, D. e. H. C. e. H. S. (2019). *Vida: A Ciência da Biologia - Vol II - 11.ed.: Evolução, Diversidade e Ecologia*. Artmed.

Saif, Rashid e Nadeem, S. e. K. A. e. Z. S. e. I. A. (2023). Mathematical understanding of sequence alignment and phylogenetic algorithms: A comprehensive review of computation of different methods. *Advancements in Life Sciences*, 9(4):401–411.

Saitou, Naruya e Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Santos, Charles Morphy Dias e Klassa, B. (2012). Sistemática filogenética hennigiana: revolução ou mudança no interior de um paradigma? *Scientiae Studia*, 10(3):593–612.

Silva, V. J. d. et al. (2012). Análise cladística e descrição de um grupo novo de espécies de edessa (heteroptera, pentatomidae, edessinae).

Simoncelli, Eero e Daw, N. (2003). Least squares optimization. *Lecture Notes, http://www. cns. nyu. edu/eero/teaching. html*.

Simone, L. d. (2020). *Bosquejos De Filogenia*.

Stokstad, E. (2019). Bangladesh could be the first to cultivate golden rice, genetically altered to fight blindness. *Science*, 366(6468):934.

Teles Caldart, Eloiza Mata, H. W. C. C. R. A. P. (2016). Análise filogenética: conceitos básicos e suas utilizações como ferramenta para virologia e epidemiologia molecular. *Acta Scientiae Veterinariae*.

Vieira, R. and Lopes, M. (1999). Uma abordagem teórica inicial para os algortimos genético através de tipos abstratos de dados. *Universidade Federal de Pernambuco, Recife, Relatório Técnico RT-D1/UFPE*, (002/99).

Von Zuben, F. J. (2000). Computação evolutiva: uma abordagem pragmática. *Anais da I Jornada de Estudos em Computação de Piracicaba e Região (1a JECOMP)*, 1:25–45.

Webb, C. O. (2000). Exploring the phylogenetic structure of ecological communities: an example for rain forest trees. *The American Naturalist*, 156(2):145–155.

Weir, B. S. (1996). Genetic data analysis ii sinauer, sunderland. *Massachusetts*.

Wheeler, W. (2012). *Systematics: A Course of Lectures*. Wiley.

Wiley, EO e Siegel-Causey, D. e. B. D. e. F. V. A. (1991). The compleat cladist: A primer of phylogeny procedures.