**Universidade de Aveiro** Departamento de Matemática
**2009**

**Olga Margarida Fajarda Oliveira**

# Árvores filogenéticas e o problema da evolução mínima

**Universidade de Aveiro** Departamento de Matemática
**2009**

**Olga Margarida Fajarda Oliveira**

# Árvores filogenéticas e o problema da evolução mínima

dissertação apresentada à Universidade de Aveiro para cumprimento dos requesitos necessários à obtenção do grau de Mestre em Matemática e Aplicações, realizada sob a orientação científica da Dra. Maria Cristina Saraiva Requejo Agra, Professora Auxiliar do Departamento de Matemática da Universidade de Aveiro.

*À memória do meu tio José Manuel Rodrigues Oliveira*

**o júri**

presidente

**Professor Domingos Moreira Cardoso**
Professor Catedrático na Universidade de Aveiro

vogais

**Professora Maria Adelaide da Cruz Cerveira**
Professora Auxiliar na Universidade de Trás-os-Montes e Alto Douro

**Professora Maria Cristina Saraiva Requejo Agra**
Professora Auxiliar na Universidade de Aveiro

**agradecimentos**

Em primeiro lugar agradeço à minha orientadora, a Professora Maria Cristina Saraiva Requejo Agra, sem a qual este trabalho não teria sido possível. Agradeço a sua dedicação, paciência, compreensão, os seus ensinamentos, as suas sugestões e a sua ajuda.

Aos meus pais e irmão Dani agradeço o apoio que sempre me deram apesar de se encontrarem longe. Agradeço, também, o apoio dos meus tios, Maria do Carmo e Vítor Ramos, da minha avó Cidália Grou e do meu primo Vítor Miguel Ramos.

Aos meus amigos, Ana Jordão, Bruno Patrício, Maria Madalena Fonseca, Sofia Silva, Sónia Figueiredo e Tânia Rodrigues agradeço o carinho, a amizade e o apoio, que foram de grande importância ao longo deste trabalho.
Agradeço, ainda, ao meu amigo Adérito Valentim que na parte prática deste trabalho me ajudou com os seus conhecimentos informáticos.

Aos Professores Daniele Catanzaro e Juan José Salazar-Gonzales agradeço o envio dos ficheiros com dados de distâncias evolutivas a serem usados na implementação dos modelos.

Por fim, agradeço a todos que de alguma forma me ajudaram a concluir este trabalho.

**palavras-chave** árvores filogenéticas, problema da evolução mínima, matriz de distâncias

**resumo** As árvores filogenéticas permite compreender a história evolutiva das espécies e pode ajudar no desenvolvimento de vacinas e no estudo da biodiversidade. Existem vários critérios para seleccionar uma árvore filogenética de entre as muitas possíveis, sendo um deles o da evolução mínima.

Nesta dissertação estudam-se vários métodos para a construção das árvores filogenéticas e várias formulações para a resolução do problema da evolução mínima. Ainda, se apresenta uma formulação alternativa que foi implementada em XPRESS.

**keywords** phylogenetic trees, the minimum evolution problem, distance matrix

**abstract** The phylogenetic trees permits to understand the evolutionary history of species and can assist in the development of vaccines and the study of biodiversity. There are several criteria to select a phylogenetic tree among the many possible, one being the evolution of the minimum.

In this thesis we study various methods for the construction of phylogenetic trees and various formulations to solve the problem of minimum evolution. It, also, presents an alternative formulation that was implemented in XPRESS.

# Conteúdo

# Lista de Tabelas

# Lista de Figuras

# Capítulo 1

# Introdução

Desde o tempo de Darwin que se procura reconstruir a história evolutiva de todas as espécies existentes na Terra. Essa história pode ser representada por uma árvore filogenética. Uma árvore filogenética é um grafo conexo, acíclico, não orientado, em que os vértices externos representam as espécies em estudo, os vértices internos representam os antepassados comuns e as arestas representam as relações evolutivas entre pares de vértices, podendo ter um peso associado ou não. Além da reconstrução da história evolutiva das espécies, as árvores filogenéticas também podem ajudar no desenvolvimento de vacinas e no estudo da biodiversidade. A inferência de árvores filogenéticas é feita, hoje em dia, através das sequências de ADN disponíveis em grandes bases de dados.

A construção das árvores filogenéticas pode ser vista como um problema de optimização. Para seleccionar uma árvore filogenética, de entre as muitas possíveis, podem usar-se vários critérios, sendo um deles o da evolução mínima. Este critério assume que a árvore filogenética óptima é aquela cuja soma dos pesos associados às arestas é mínima.

A construção de árvores filogenéticas é um problema NP-Difícil e, consoante o critério de selecção que se usa, existem vários métodos que permitem a sua construção. Os principais métodos são baseados na verossimilhança máxima, na máxima parcimónia e na matriz de distâncias.

Nesta dissertação estudamos vários métodos para construir árvores filogenéticas e apresentamos várias formulações para o problema da evolução mínima.

No Capítulo 2 apresentamos alguns aspectos ligados às árvores filogenéticas: uma possível classificação dos vários tipos de árvores, uma definição formal e a fórmula para a determinação do número de árvores possíveis.

No Capítulo 3 apresentamos alguns métodos para a construção das árvores filogenéticas, nomeadamente, métodos de máxima parcimónia, métodos de verossimilhança máxima e alguns métodos de distância. Relativamente aos métodos de distância, serão apresentados o método da média aritmética não ponderada e o método dos mínimos quadrados.

No Capítulo 4 analisamos o problema da evolução mínima, apresentamos o método neighbor-joining para o resolver, bem como duas formulações em programação linear encontradas na literatura.

No Capítulo 5 apresentamos uma formulação alternativa, os resultados obtidos quando usamos o software XPRESS na resolução de alguns problemas e a decomposição de Dantzig-Wolfe relativa a essa formulação.

Finalmente, no Capítulo 6 apresentamos as conclusões.

# Capítulo 2

# Árvores Filogenéticas

A filogenia é a história da evolução de uma espécie ou de qualquer grupo hierarquicamente reconhecido[1]. O termo *filogenética* terá sido usado pela primeira vez em 1866 por Ernest Haeckel para descrever a evolução das espécies vegetais e animais ao longo do tempo. Charles Darwin expôs, no seu livro *Origem das Espécies*, em 1872, a ideia de que todos os seres vivos se transformam ao longo do tempo, descendendo todos dum ancestral comum e definiu a filogenia como sendo "as linhas genealógicas de todos os seres organizados". Foi, também, neste livro que estas relações evolutivas entre espécies foram representadas, pela primeira vez, por meio de árvores filogenéticas [7].

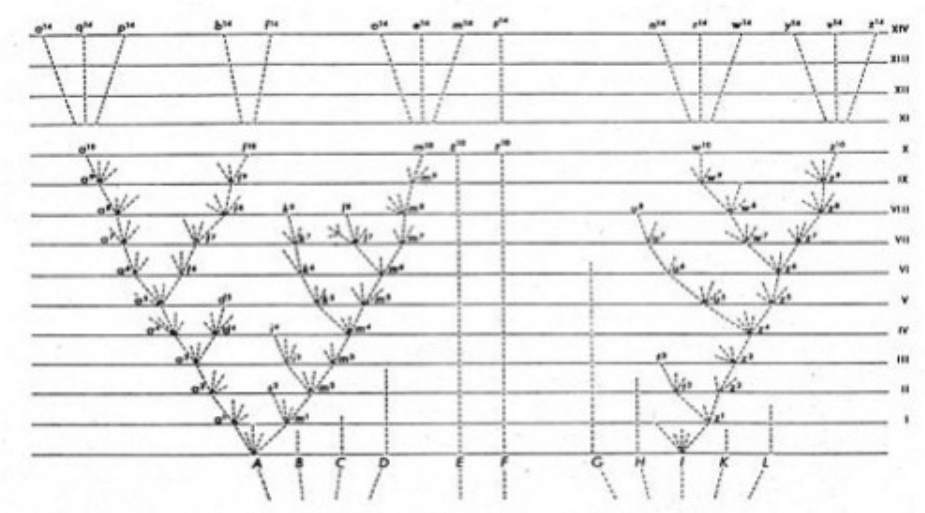


Figura 2.1: Árvores filogenéticas apresentadas no livro *Origem das Espécies*.

A análise filogenética pode ser feita por comparação de características paleontológicas (através do estudo de fósseis), morfológicas (através do estudo da forma exterior que os seres vivos podem tomar) e/ou fisiológicas (através do estudo das funções

[1]Dicionário Editora da Língua Portuguesa 2010

dos diferentes órgãos dos seres vivos) das várias espécies. Hoje em dia, com os avanços na biologia molecular, as árvores filogenéticas são construídas através da análise do ADN dos vários organismos, comparando famílias de ácidos nucleicos ou sequências de proteínas [21].

Muitos autores assumem que ao longo da evolução uma espécie se desenvolve em duas e apenas duas espécies diferentes. Consequentemente, ao recriar-se a história evolutiva de três espécies, espera-se que duas delas tenham um antepassado comum e que este por sua vez tenha um antepassado comum com a terceira espécie em estudo [20].

Uma árvore filogenética é um grafo conexo, acíclico, não orientado, em que os vértices têm grau um ou três e no máximo um dos vértices tem grau dois. Os vértices de grau um, designados por folhas ou vértices externos, representam os objectos em estudo que podem ser organismos, genes ou táxons[2]. Os vértices de grau três, designados por vértices internos, representam os ancestrais intermédios das folhas. Quando existir, o vértice de grau dois designa-se por raiz e representa o ancestral comum a todos os objectos estudados. As arestas representam as relações evolutivas entre os vértices e podem ter associadas um peso que representa a quantificação dessa relação evolutiva. As arestas incidentes em vértices externos designam-se por arestas externas e as outras arestas por arestas internas. Na Figura 2.2 vemos representada uma árvore filogenética de quatro organismos, com o mesmo antepassado comum e, com dois antepassados intermédios.

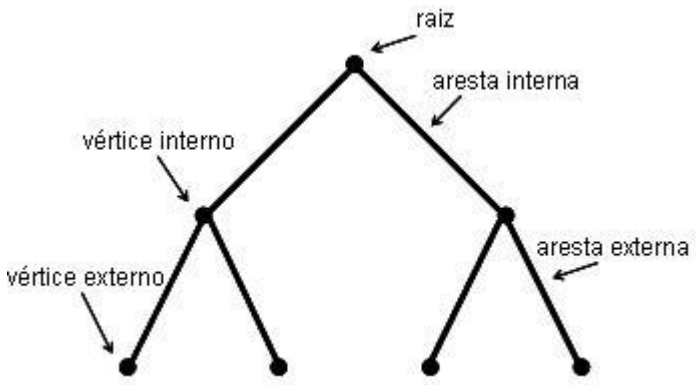


Figura 2.2: Exemplo de uma árvore filogenética com raiz

## 2.1 Tipos de árvores

As árvores filogenéticas podem ser classificadas usando vários critérios. Quando se estuda a evolução histórica de um grupo de espécies, pretende-se construir uma *árvore*

[2]Táxon é um organismo ou um conjunto de organismos devidamente classificados.

*de espécies*. Neste tipo de árvores a distância entre dois vértices externos, representa o tempo decorrido desde que uma dada espécie antecedente se separou e originou estas duas espécies isoladas [21]. Muitas vezes, usam-se os genes de cada espécie para fazer esse estudo e a árvore que se obtém pode não corresponder à árvore de espécie, por causa da retenção de polimorfismos ancestrais. Para diferenciar estes dois tipos de árvores, designam-se estas últimas por *árvores de genes* [16].

Usando as ligações existentes nas árvores filogenéticas podemos classificá-las da seguinte forma:

- O *dendograma* é uma árvore em que o peso de cada aresta é o mesmo, sendo, por isso, omitido. Neste tipo de árvores a relação entre vértices externos é exprimida através de uma sucessão de conexões.

- O *cladograma* é um dendograma construído a partir da análise cladística[3] e em que os vértices internos são obtidos através de sinapomorfias[4]. Neste tipo de árvores as arestas, também, não têm pesos.

- O *filograma* é um dendograma construído a partir da análise caldística e onde o peso de cada aresta exprime o grau de divergência entre vértices adjacentes.

- O *fenograma* é um dendograma obtido através da taxonomia numérica e onde as relações entre vértices externos indicam o grau de similaridade global, representado pelo peso atribuído às arestas.

Já foi referido que as árvores filogenéticas podem ou não ter uma raiz, sendo assim, podemos usar esse critério para as classificar:

- As *árvores sem raiz* apresentam apenas a noção de distância entre os vértices e não apresentam as noções de ancestralidade.

- Nas *árvores com raiz* está implícita a noção de tempo e as arestas podem ser orientadas de forma unívoca, pois a raiz é o antepassado comum a todos os outros vértices [7].

Na Figura 2.3 representamos estes dois tipos de árvores.

Como já foi referido, assume-se que uma sequência de ADN evolui em duas sequência descendentes. As árvores filogenéticas são por isso também designadas por *árvores bifurcadas*. Contudo, ao considerar uma sequência muito curta, poderão não aparecer diferenças entre várias sequências e, consequentemente, poderão aparecer vértices com grau superior a três. Esse tipo de árvore designa-se por *árvore multifurcada*. Ás vezes,

[3]Análise cladística é uma forma de análise filogenética através da análise de caracteres ancestrais (pleisiomórficos) e caracteres derivados (apomórficos).

[4]Sinapomorfias são caracteres homólogos apomórficos (derivados) compartilhados por dois ou mais táxons.

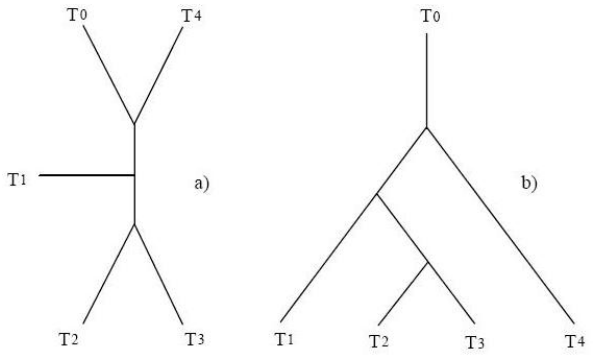


Figura 2.3: Exemplo de (a) uma árvore sem raiz e de (b) uma árvore com raiz [7].

as árvores bifurcadas podem ser reduzidas a árvores multifurcadas ao eliminar as arestas com peso igual a zero [21]. Na Figura 2.4 apresentamos um exemplo de uma árvore multifurcada.

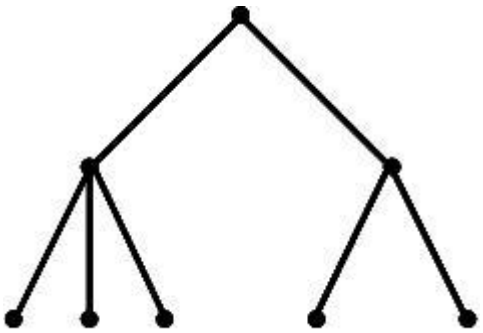

Figura 2.4: Exemplo de uma árvore multifurcada.

Podemos, ainda, distinguir entre *árvores esperadas* e *árvores realizadas*. Estudar as sequências completas de ADN das várias espécies torna-se muito demorado e complicado, preferindo os investigadores estudar apenas um segmento de ADN. Ao fazer isso, corre-se o risco de a árvore filogenética obtida ser diferente da árvore que se obteria usando a sequência completa de ADN. A árvore filogenética que se obtém usando sequências completas de ADN designa-se por *árvore esperada* e a árvore filogenética que se obtém usando apenas um segmento de ADN designa-se por *árvore realizada* [21].

## 2.2 Definição formal de árvore filogenética

Uma árvore filogenética é representada por um grafo $G(V, E)$, onde $V$ é o conjunto dos vértices e $E$ é o conjunto das arestas.

O conjunto dos vértices, $V$, pode ser dividido em dois subconjuntos disjuntos: o conjunto dos vértices externos, $V_{ext}$, e o conjunto dos vértices internos, $V_{int}$, sendo que

nas árvores filogenéticas com raiz, a raiz pertence ao conjunto $V_{int}$. Da mesma forma, o conjunto das arestas, $E$, pode ser dividido em dois subconjuntos disjuntos: o conjunto das arestas externas, $E_{ext}$, e o conjunto das arestas internas, $E_{int}$. As arestas externas ligam vértices externos a vértices internos e as arestas internas ligam vértices internos entre si.

Numa árvore filogenética com $n$ taxos, os conjuntos $V_{ext}$ e $E_{ext}$ têm $n$ elementos.

Nas árvores filogenéticas sem raiz, o conjunto $E_{int}$ tem $(n-3)$ elementos, o conjunto $E$ tem $(2n-3)$ elementos, o conjunto $V_{int}$ tem $(n-2)$ elementos e o conjunto $V$ tem $(2n-2)$ elementos.

Nas árvores filogenéticas com raiz, o conjunto $E_{int}$ tem $(n-2)$ elementos, o conjunto $E$ tem $(2n-2)$ elementos, o conjunto $V_{int}$ tem $(n-1)$ elementos e o conjunto $V$ tem $(2n-1)$ elementos.

As árvores filogenéticas com $n$ vértices externos podem ser representadas por uma matriz de incidência aresta-caminho [3, 4]. Entre cada par de vértices existe apenas um caminho. As $\frac{n(n-1)}{2}$ linhas desta matriz de incidência, $X = \{x_{ij,e} : i, j \in V_{ext}, i < j, e \in E\}$, representam os caminhos entre dois quaisquer vértices externos e as $(2n-3)$ colunas, no caso das árvores filogenéticas sem raiz, ou as $(2n-2)$ colunas, no caso das árvores filogenéticas com raiz, representam as arestas. As entradas $x_{ij,e}$ tomam o valor 1 se $e$ pertence ao caminho $p_{ij}$ que liga o vértice externo $i$ ao vértice externo $j$ e tomam o valor 0 caso contrário.

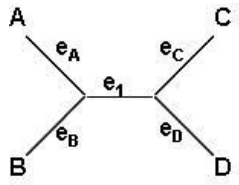


Figura 2.5: Árvore filogenética sem raiz com quatro vértices externos.

Considerando a árvore filogenética com quatro vértices externos representada na Figura 2.5, obtêm-se a seguinte matriz de incidência aresta-caminho [3]:

| | $e_A$ | $e_B$ | $e_C$ | $e_D$ | $e_1$ |
|---|---|---|---|---|---|
| $AB$ | 1 | 1 | 0 | 0 | 0 |
| $AC$ | 1 | 0 | 1 | 0 | 1 |
| $AD$ | 1 | 0 | 0 | 1 | 1 |
| $BC$ | 0 | 1 | 1 | 0 | 1 |
| $BD$ | 0 | 1 | 0 | 1 | 1 |
| $CD$ | 0 | 0 | 1 | 1 | 0 |

Tabela 2.1: Matriz de incidência aresta-caminho para a árvore filogenética representada na Figura 2.5.

## 2.3 Número de árvores filogenéticas possíveis

Uma mesma árvore pode ter várias formas equivalentes de ser representada (árvores isomorfas). Em cada uma das formas equivalentes as ligações entre os vértices é a mesma. A esse conjunto de formas equivalentes chama-se topologia da árvore. Para um conjunto de taxons existem várias topologias de árvores com raiz e sem raiz. Com quatro taxons podem-se construir quinze topologias diferentes de árvores com raiz e três topologias diferentes de árvores sem raiz. Na Figura 2.6 estão representadas as várias topologias de árvores que se podem construir com quatro taxons. Com o aumento do número de taxons, o número de topologias aumenta rapidamente.

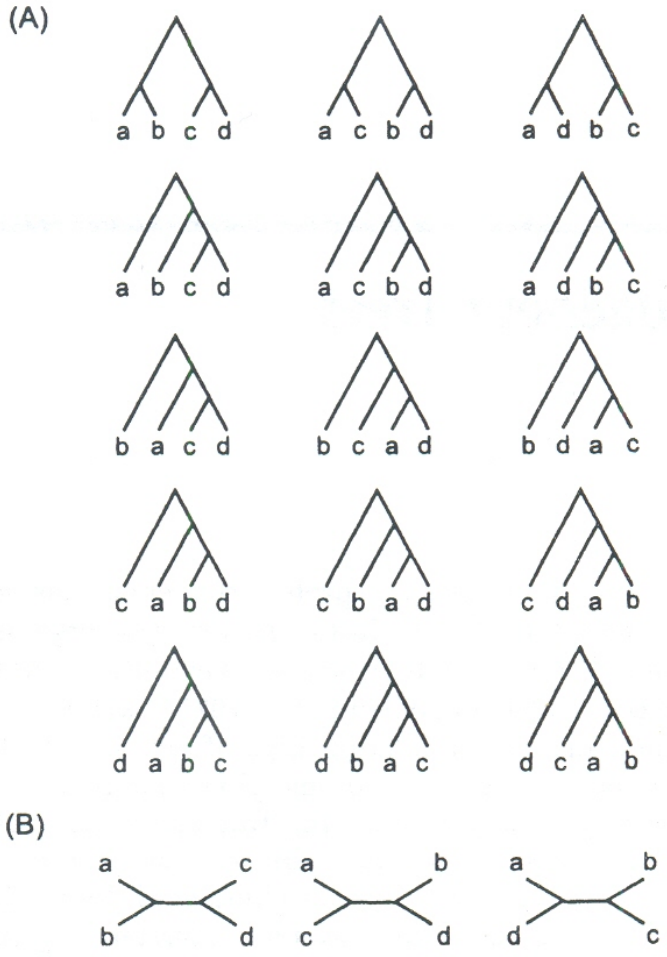


Figura 2.6: (A) As 15 topologias de árvores com raiz e (B) as 3 topologias de árvores sem raiz para quatro taxons [21].

A expressão que permite determinar o número de possíveis topologias de árvores com raiz para $n$ taxos (com $n \geq 2$) é dada por:

$$1 \times 3 \times 5 \times \cdots \times (2n-3) = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{2.1}$$

Para determinar o número de possíveis topologias de árvores sem raiz para $n$ taxos (com n $\geq 2$), usa-se a seguinte expressão[21]:

$$\frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{2.2}$$

Como se pode verificar o número de possíveis árvores aumenta rapidamente à medida que o número de taxons, $n$, aumenta, tornando a procura da árvore filogenética de forma exaustiva impossível para $n$ grande. Por exemplo, se $n = 15$ existem 213458046676875 topologias de árvores com raiz e 7905853580625 topologias de árvores sem raiz.

# Capítulo 3

# Métodos para reconstruir árvores filogenéticas

Segundo Zuben, Reis e Prado [25] os vários métodos para reconstruir árvores filogenéticas podem dividir-se em dois grupos:

- Métodos baseados em modelos. Estes métodos analisam e avaliam um conjunto de árvores filogenéticas e determinam, consoante o critério escolhido, a melhor ou uma boa árvore. Usualmente, esta análise é feita usando modelos probabilísticos.

- Métodos não baseados em modelos. Nestes métodos existe um algoritmo que permite determinar a melhor (ou uma boa) árvore filogenética.

Dos métodos baseados em modelos o mais conhecido é o método que usa como critério a verossimilhança máxima.

Os métodos não baseados em modelos podem ser divididos em dois grupos:

- Métodos de máxima parcimónia;

- Métodos de distâncias.

Enquanto os métodos baseados em modelos devolvem, geralmente, como resultado final do problema em análise um conjunto de boas árvores, os métodos não baseados em modelos devolvem, em princípio, apenas uma árvore, considerada a melhor segundo os critérios usados. O método não baseado em modelos "*parsimony ratchet method*" [22] pode devolver um conjunto de boas árvores.

Todos os métodos referidos são baseados num conjunto de critérios, que podem, usualmente, expressar-se através de uma função objectivo. Assim, ao optimizar-se essa função objectivo, obtemos como solução uma árvore óptima cuja topologia se pretende que seja a mais próxima da árvore esperada. Um critério, ou conjunto de critérios, diz-se estatisticamente consistente, se for de tal forma que a árvore realizada óptima se

aproxime da árvore esperada à medida que aumentamos os dados moleculares (i.e., o tamanho das sequências moleculares consideradas) [3].

Quando o método utilizado fornece um conjunto de boas árvores, para evitar apresentar todas as soluções, costuma apresentar-se uma árvore de consenso que é uma composição de todas as boas árvores. Existem vários tipos de árvores de consenso, mas as mais usadas são as árvores de consenso estrito e as árvores de consenso de maior regra [21].

De seguida apresentam-se, de uma forma muito geral, o método de máxima parcimónia, o método de verossimilhança máxima e alguns métodos de distância. Os métodos baseados no critério de evolução mínima são métodos de distâncias que serão estudados no capítulo seguinte.

## 3.1 Métodos de máxima parcimónia

Originalmente, os métodos de máxima parcimónia foram pensados para serem usados com características morfológicas. Os primeiros a usarem um método de máxima parcimónia com dados moleculares foram Eck e Dayhoff [9] em 1966. Em 1971, Fitch [12] e, depois em 1973, Hartigan [15] desenvolveram algoritmos de máxima parcimónia mais rigorosos [21]. Na ciência a ideia de parcimónia consiste em preferir hipóteses simples em detrimento das mais complicadas e que hipóteses ad hoc devem ser evitadas sempre que possível. Assim, os métodos de máxima parcimónia assumem que atributos compartilhados entre espécies deve-se ao facto de terem sido herdados por um antepassado comum [16]. Todos estes métodos baseiam-se na ideia filosófica de William de Ockham's que considera que das muitas explicações para um fenómeno devem-se escolher aquelas que requerem o menor número de hipóteses [21].

Os métodos de parcimónia procuram a árvore, que no menor número de passos, muda qualquer sequência nas outras todas e neste sentido contabilizam, para cada topologia de árvore, esse número de mudanças de sequências. Depois, comparam as várias topologias e escolhem a que minimiza o comprimento total da árvore. Por vezes, existe mais do que uma topologia com o comprimento mínimo e considera-se que cada uma delas pode ser a topologia certa. Os métodos de máxima parcimónia usam, geralmente, árvores sem raiz, uma vez que estes não determinam a raiz da árvore [21].

De seguida, ilustra-se a ideia dos métodos de máxima parcimónia através de um exemplo [20].

Consideram-se quatro taxons, cada um com nove caracteres (posições) como se mostra na tabela seguinte.

As regras para uma análise de máxima parcimónia neste exemplo são as seguintes:

1. Como existem quatro taxons apenas existem três topologias de árvores sem raiz.

| | Posição das sequências (sítios) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Taxons | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| I | A | A | G | A | G | T | G | C | A |
| II | A | G | C | C | G | T | G | C | G |
| III | A | G | A | T | A | T | C | C | A |
| IV | A | G | A | G | A | T | C | C | G |

Tabela 3.1: Sequências de nucleótidos para quatro taxos [20].

2. Alguns sítios são informativos permitindo diferenciar uma árvore em relação a outra (por exemplo o sítio 5 é informativo, mas os sítios 1, 6 e 8 não são).

3. Para ser informativo, um sítio tem de ter a mesma sequência de caracteres em pelo menos dois taxons (por exemplo, os sítios 1, 2, 3, 4, 6 e 8 não são informativos; os sítios 5, 7 e 9 são informativos).

4. Apenas aos sítios informativos precisam de ser analisados.

Nas Figuras 3.1, 3.2 e 3.3 apresentamos a análise desse três sítios.

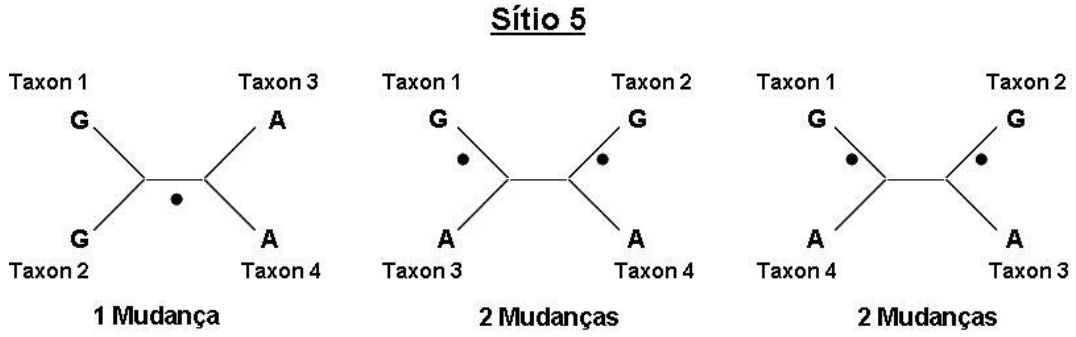


Figura 3.1: Contagem das mudanças dos caracteres relativamente ao sítio 5 nas três topologias.

Feita esta contagem para os sítios informativos basta contabilizar para cada topologia as mudanças obtidas e escolher a topologia com o menor número de mudanças. Na Figura 3.4 apresentamos essa contabilidade e podemos ver que a topologia de máxima parcimónia para este exemplo é a primeira.

Em termos matemáticos podemos definir o problema de máxima parcimónia da seguinte forma: do conjunto de todas as árvores possíveis, encontrar as árvores $\tau$ para as quais

$$L(\tau) = \sum_{k=1}^{B} \sum_{j=1}^{N} w_j * diff(x_{k'j}, x_{k''j}) \tag{3.1}$$

seja mínimo, onde $L(\tau)$ é o comprimento da árvore $\tau$, $B$ é o número de arestas, $N$ é o número de caracteres, $k'$ e $k''$ são dois vértices incidentes em cada aresta $k$, $x_{k'j}$

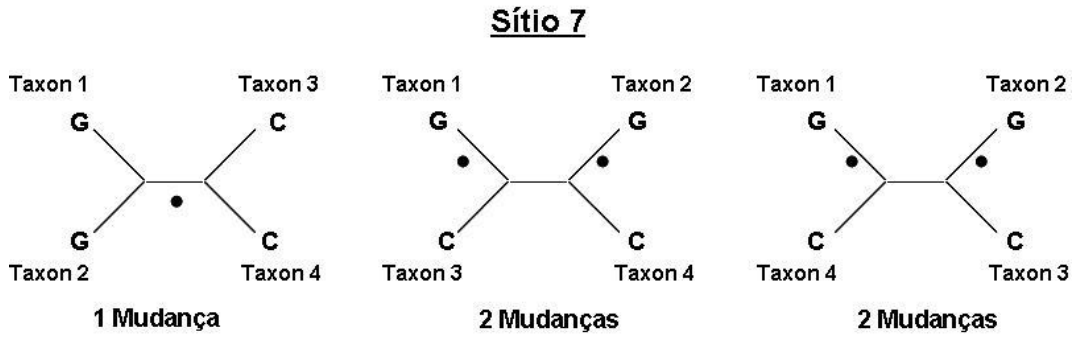


Figura 3.2: Contagem das mudanças dos caracteres relativamente ao sítio 7 nas três topologias.

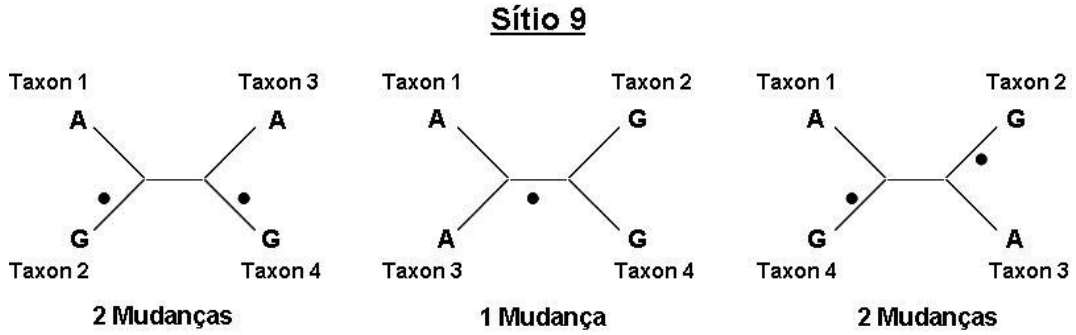


Figura 3.3: Contagem das mudanças dos caracteres relativamente ao sítio 9 nas três topologias.

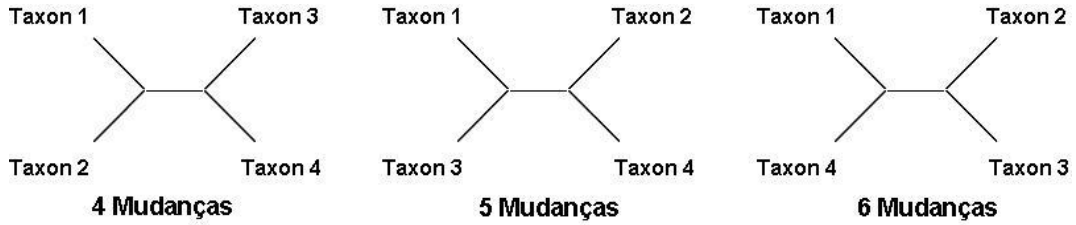


Figura 3.4: Contabilização das mudanças para cada uma das topologias

e $x_{k''j}$ podem representar elementos da matriz de dados de entrada ou um estado de carácter óptimo atribuído aos nós internos, $diff(y, z)$ é a função que indica o custo de transformar o estado $y$ no estado $z$ ao logo de cada aresta e os coeficientes $w_j$ atribuem um peso a cada caractér [16].

A partir de dez taxons estes métodos de procura exaustiva da árvore tornam-se pouco viáveis. Assim, se o número de taxons for maior do que dez existem duas formas de obter a árvore de máxima parcimónia. A primeira é usar o método *branch-and-bound*, que começa por ignorar as topologias que têm um comprimento maior que as topologias já estudadas e que a seguir, avalia os comprimentos dum conjunto de topologias que têm o comprimento menor para, finalmente, se escolher a de máxima parcimónia. Este método também se torna pouco viável se o número de taxons for maior que vinte. Nestes casos usam-se heurísticas, que apenas estudam uma parte das topologias possíveis sem garantia de que a árvore escolhida seja a de máxima parcimónia [21].

Encontrar uma árvore de máxima parcimónia é um problema NP-Difícil e em certas circunstâncias este critério é estatisticamente inconsistente [4].

## 3.2 Métodos de verossimilhança máxima

Cavalli-Sforza e Edwards [6] foram os primeiros, em 1967, a usarem um método baseado na verossimilhança máxima para inferir árvores filogenéticas. Contudo, tiveram muitos problemas ao tentarem implementar o método. Já em 1981, Felsenstein [10] desenvolveu um algoritmo para construir árvores filogenéticas usando métodos de verossimilhança máxima através de sequências de nucleótidos [14]. Estes métodos encontram a árvore que melhor representa a variação num conjunto de dados através de modelos probabilísticos. Para cada topologia os pesos das arestas são escolhidos de modo a maximizar a probabilidade dessa árvore ter gerado as sequências observadas através de modelos de substituição específicos. Depois comparam-se as várias topologias e a topologia com a maior probabilidade (maior verossimilhança) é considerada como sendo a árvore filogenética óptima [21].

Para se poder aplicar os métodos de verossimilhança máxima deve-se primeiro especificar um modelo concreto do processo de evolução, que contabiliza a conversão de uma sequência noutra. Este modelo pode ser totalmente definido ou pode conter parâmetros que serão estimados através dos dados. Neste contexto, o método de verossimilhança máxima determina a probabilidade do modelo escolhido gerar evolutivamente as sequências observadas [16] .

Os princípios básicos envolvidos no cálculo das probabilidades são apresentados, de seguida, através de um exemplo.

Consideremos quatro taxons e as respectivas sequências de nucleótidos (ver Figura 3.5).

Depois de escolher uma topologia sem raiz (ver Figura 3.6 (A)) pretende-se determi-


Figura 3.5: Sequências de nucleótidos de quatro taxons.

nar a probabilidade dessa topologia ter gerado as sequências de nucleótidos observados sob o modelo escolhido. Geralmente, a verossimilhança de uma árvore é independente da posição da raiz, uma vez que a grande parte dos modelos são de tempo reversível. Assim, coloca-se a raiz num nó interno (ver Figura 3.6 (B)).

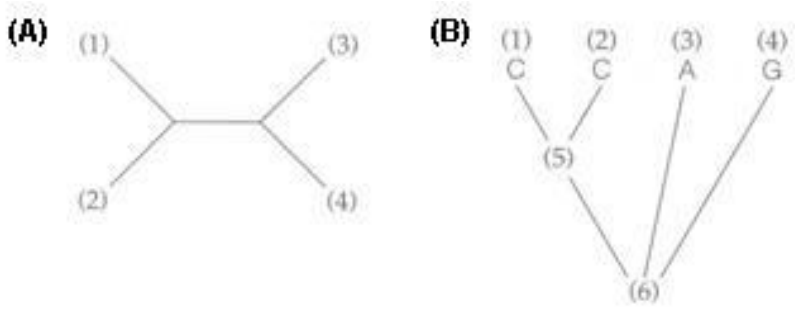


Figura 3.6: (A) topologia de árvore sem raiz; (B) a respectiva topologia de árvore já com a raiz.

Como se supõe que cada sítio da sequência evolui independentemente de outro, podemos analisar a probabilidade de cada um, separadamente, e depois combinar as probabilidades obtidas.

Para calcular a verossimilhança de um sítio temos de considerar todas as possibilidades através das quais os vértices externos poderão ser obtidos. Qualquer cenário tem alguma probabilidade de ter gerado o conjunto de nucleótidos observados, apesar de alguns terem maior probabilidade. Assim, um vértice externo pode ter como antecedente um A, um C, um G ou um T e para cada uma destas possibilidades o outro vértice interno também pode ser qualquer um desses quatro nucleótidos. Consequentemente, temos de considerar $4 \times 4 = 16$ possibilidades. Para cada sítio $j$ temos de calcular a probabilidade de cada cenário e somá-las de forma a obter a probabilidade desse sítio $j$ (ver Figura 3.7).

Depois de calculadas as verossimilhanças de cada sítio a probabilidade final daquela topologia é obtida fazendo o produto das verossimmilhanças de cada sítio, ou seja,

$$L = L_{(1)} \bullet L_{(2)} \bullet \cdots \bullet L_{(N)} = \prod_{j=1}^{N} L_{(j)}. \tag{3.2}$$

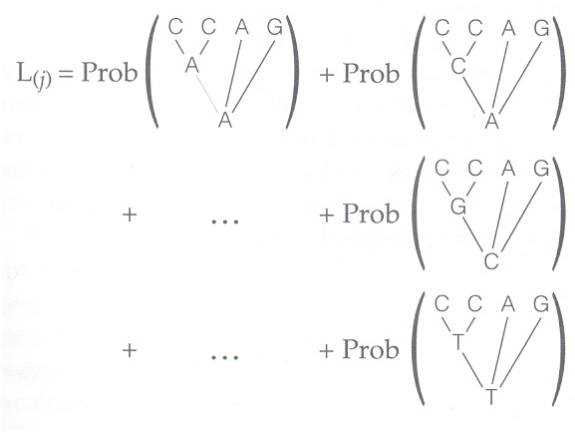


Figura 3.7: Cálculo da probabilidade do sítio $j$.

Uma vez que as probabilidade calculadas são valores muito pequenos costuma-se usar o logaritmo neperiano das probabilidades, ou seja,

$$\ln L = \ln L_{(1)} + \ln L_{(2)} + \cdots + \ln L_{(N)} = \sum_{j=1}^{N} \ln L_{(j)}. \tag{3.3}$$

Falta ainda definir como calcular as probabilidades das diferentes mudanças. Essas probabilidades dependem do modelo de substituição escolhido. Um dos mais usados é o modelo de Markov. Este modelo supõe que a mudança de uma base $i$ para uma base $j$ numa aresta é independente das mudanças que ocorrem nas outras arestas e supõe ainda que as probabilidades não mudam nas várias partes da árvore.

A expressão matemática de um modelo de substituição é uma tabela de variação (substituição por sítio, por unidade de distância evolucionária) onde cada nucleótido é substituído por cada nucleótido alternativo. Para uma sequência de ADN essas variações podem ser expressas por uma matriz $4 \times 4$, $Q$, onde cada elemento $q_{ij}$ representa a taxa de variação da base $i$ para a base $j$ num período de tempo infinitesimal $dt$. A Figura 3.8 apresenta essa matriz.

$$\mathbf{Q} = \begin{pmatrix} -\mu(a\pi_C + b\pi_G + c\pi_T) & \mu a\pi_C & \mu b\pi_G & \mu c\pi_T \\ \mu g\pi_A & -\mu(g\pi_A + d\pi_G + e\pi_T) & \mu d\pi_G & \mu e\pi_T \\ \mu h\pi_A & \mu j\pi_C & -\mu(h\pi_A + j\pi_C + f\pi_T) & \mu f\pi_T \\ \mu i\pi_A & \mu k\pi_C & \mu l\pi_G & -\mu(i\pi_A + k\pi_C + l\pi_G) \end{pmatrix}$$

Figura 3.8: Matriz de variação.

Nessa matriz as linhas (e as colunas) correspondem aos nucleótidos $A$, $C$, $G$ e $T$, respectivamente. O factor $\mu$ representa a taxa média de substituição instantânea. Esta

taxa média é modificada pelos parâmetros $a, b, c, \ldots, l$ que correspondem às possíveis transformações de uma base para outra diferente. Os parâmetros $\pi_A, \pi_C, \pi_G$ e $\pi_T$ representam as frequências dos nucleótidos $A$, $C$, $G$ e $T$, respectivamente, e assume-se que estas frequências se mantém constantes ao longo do tempo. Cada elemento da diagonal principal da matriz $Q$ é escolhido de forma a que a soma dos elementos dessa linha seja zero.

A probabilidade de mudança de um qualquer estado para qualquer outro ao longo de uma aresta de comprimento $t$ é calculado através da matriz $Q$ da seguinte forma:

$$P(t) = \exp^{Qt} \tag{3.4}$$

O exponencial pode ser calculado pela decomposição da matriz $Q$ nos seus valores próprios e vectores próprios.

Encontrar uma árvore de verossimilhança máxima é um problema NP-Difícil mas é estatisticamente consistente [4]. Estes métodos, também, têm a particularidade de serem os que menos são afectados por erros de amostragem e são robustos à violação dos pressupostos utilizados nos seus modelos [16].

## 3.3 Métodos de distâncias

Os métodos de distâncias tentam posicionar correctamente os vizinhos na árvore e encontrar os pesos das arestas que melhor se ajustam aos dados originais representados com a ajuda de uma matriz de distâncias evolutivas entre pares de dados moleculares [20]. Esse ajuste pode ser feito de várias formas e cada ajuste dá origem a um método ou critério para encontrar a árvore filogenética [4]. As distâncias evolutivas são, geralmente, calculadas tendo por base o número de alterações no ADN das várias espécies e, naturalmente, a qualidade da árvore filogenética construída depende da qualidade dessas distâncias e por isso, também, da forma como elas forem calculadas.

De seguida iremos indicar, apenas, algumas formas de calcular as distâncias.

### 3.3.1 Determinação das distâncias

Existem várias formas de calcular as distâncias, sendo que as mais simples de determinar são as chamadas *distâncias de Hamming*. Estas distâncias contabilizam o número de nucleótidos diferentes existentes em duas sequências alinhadas [14].

Considerando, por exemplo, as seguintes sequências:

(X) ATGCGTCGTT
(Y) ATCCGCGATC

tem-se que a distância de Hamming é 5.

Outra forma de calcular as distâncias é através da razão (ou percentagem) de similaridade ou não similaridade. Na forma mais simples, essa razão é calculada dividindo o número de nucleótidos diferentes nas duas sequências alinhadas ($n_p$) pelo número total de nucleótidos analisados ($n$), ou seja,

$$d = \frac{n_p}{n}. \tag{3.5}$$

As distâncias assim calculadas designam-se por *distâncias p* [21].

No exemplo anterior a distância $p$ é $d = \frac{5}{10} = 0.5$.

Por vezes também é útil conhecer a frequência relativa de cada par de nucleótidos na comparação de duas sequências $X$ e $Y$. Uma vez que existem quatro nucleótidos $A$, $C$, $G$ e $T$, existem 16 pares de nucleótidos diferentes. As frequências relativas dos diferentes pares de nucleótidos na dada comparação de pares de duas sequências $X$ e $Y$ são, frequentemente, apresentados através da seguinte matriz:

$$F_{XY} = \begin{bmatrix} \frac{n_{AA}}{n} & \frac{n_{AC}}{n} & \frac{n_{AG}}{n} & \frac{n_{AT}}{n} \\ \frac{n_{CA}}{n} & \frac{n_{CC}}{n} & \frac{n_{CG}}{n} & \frac{n_{CT}}{n} \\ \frac{n_{GA}}{n} & \frac{n_{GC}}{n} & \frac{n_{GG}}{n} & \frac{n_{GT}}{n} \\ \frac{n_{TA}}{n} & \frac{n_{TC}}{n} & \frac{n_{TG}}{n} & \frac{n_{TT}}{n} \end{bmatrix} \tag{3.6}$$

onde $n_{ij}$ é o número de vezes que o estado $i$ na sequência $X$ está alinhado com o estado $j$ na sequência $Y$ e $n$ é o número total de nucleótidos em cada sequência [16].

Voltando ao exemplo anterior obtém-se a seguinte matriz:

$$F_{XY} = \begin{bmatrix} \frac{1}{10} & 0 & 0 & 0 \\ 0 & \frac{1}{10} & \frac{1}{10} & 0 \\ \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & 0 \\ 0 & \frac{1}{5} & 0 & \frac{1}{5} \end{bmatrix} \tag{3.7}$$

O modelo de Junkes-Cantor assume que a substituição de nucleótidos ocorre com a mesma frequência em cada sítio e calcula as distâncias da seguinte forma:

$$d_{xy} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D\right) \quad com \quad D = 1 - \frac{n_{AA} + n_{CC} + n_{GG} + n_{TT}}{n} \tag{3.8}$$

onde $D$ representa a razão de diferença entre os nucleótidos de $X$ e $Y$ e não pode ser maior ou igual a 0,75. Isto significa que a razão de semelhança entre duas sequências deverá ser maior que 0,25 para podermos calcular as distâncias usando este modelo.

Uma vez que as substituições de nucleótidos podem ocorrer com frequências diferentes, Felsenstein (1981) [10] e Tajia e Nei (1982) [30] propuseram calcular as distâncias

da seguinte forma:

$$d_{xy} = -B \ln\left(1 - \frac{D}{B}\right) \tag{3.9}$$

onde $D = 1 - \frac{n_{AA}+n_{CC}+n_{GG}+n_{TT}}{n}$, $B = 1 - \pi_A^2 + \pi_C^2 + \pi_G^2 + \pi_T^2$ e os parâmetros $\pi_A, \pi_C, \pi_G$ e $\pi_T$ representam as frequências dos nucleótidos $A$, $C$, $G$ e $T$, respectivamente.

Quando $B = \frac{3}{4}$ obtemos as distâncias calculadas pelo modelo de Jukes-Cantor. As distâncias para o modelo de Poisson são obtidas fazendo $B = \frac{19}{20}$.

Pode-se ainda usar o modelo de Kimura com dois parâmetros. Este modelo calcula as distâncias da seguinte forma:

$$d_{xy} = \frac{1}{2} \ln\left(\frac{1}{1-2P-Q}\right) + \frac{1}{4} \ln\left(\frac{1}{1-2Q}\right) \tag{3.10}$$

onde $P = \frac{n_{AG}+n_{CT}+n_{GA}+n_{TC}}{n}$ e $Q = \frac{n_{AC}+n_{AT}+n_{CA}+n_{CG}+n_{GC}+n_{GT}+n_{TA}+n_{TG}}{n}$ [21, 16].

Usando a matriz $F_{XY}$ anteriormente definida podemos calcular as distâncias através da seguinte fórmula:

$$d_{xy} = -\ln(\det F_{XY}) \tag{3.11}$$

onde $\det F_{XY}$ é o determinante da matriz $F_{XY}$ [31].

### 3.3.2 Matriz de distâncias

A matriz de distância é uma matriz quadrada $n$, sendo $n$ o número de espécies em estudo. Mais especificamente, a matriz é triangular superior com diagonal zero e cada entrada $d_{ij}$ representa o valor da distância entre a espécie $i$ e a espécie $j$. Pressupõe-se assim que:

- $d_{ii} = 0 \quad \forall i = 1 \ldots n$
- $d_{ij} = d_{ji} \quad \forall i, j = 1 \ldots n,\ i \neq j$
- $d_{ij} > 0 \quad \forall i, j = 1 \ldots n,\ i \neq j$

e supõe-se, ainda, que estas condições são válidas para todos os pesos das arestas da árvore filogenética.

Se além disso a matriz verificar a desigualdade triangular,

$$d_{ij} \leq d_{ik} + d_{kj} \quad \forall i, j, k = 1 \ldots n \tag{3.12}$$

a matriz diz-se *metrica* [4]. Algumas fórmulas de distâncias calculadas a partir de dados biológicos, bem como distâncias derivadas de procedimentos experimentais podem não satisfazer a desigualdade triangular [11].

Se pudéssemos determinar de forma exacta as distâncias evolutivas entre os diferentes taxons em estudo essas distâncias teriam a propriedade da aditividade de árvore, ou seja, a distância evolutiva entre qualquer par de taxons seria igual à soma dos pesos das arestas que se encontram no caminho que ligam os vértices que representam esses dois taxons [16].

Uma matriz de distâncias de ordem $n$ diz-se *aditiva* se verificar a condição dos quatro pontos [4]:

$$d_{zi} + d_{kj} \leq d_{zj} + d_{ik} = d_{kz} + d_{ij} \quad \forall i, j, k, z = 1 \ldots n \tag{3.13}$$

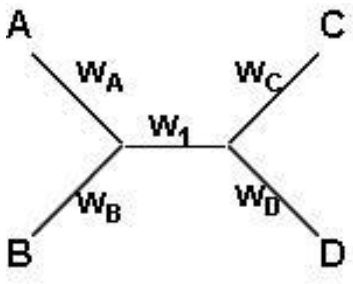


Figura 3.9: Exemplo de uma árvore filogenética cuja matriz de distâncias é aditiva.

Por exemplo, se considerarmos quatro taxons, $A, B, C$ e $D$ e considerarmos a árvore representada na Figura 3.9 tem-se

$d_{AB} = w_A + w_B$
$d_{AC} = w_A + w_1 + w_C$
$d_{AD} = w_A + w_1 + w_D$
$d_{BC} = w_B + w_1 + w_C$
$d_{BD} = w_B + w_1 + w_D$
$d_{CD} = w_C + w_D$
Assim,
$d_{AB} + d_{CD} = w_A + w_B + w_C + w_D$
$d_{AD} + d_{BC} = w_A + w_D + w_B + w_C + 2w_1$
$d_{CA} + d_{BD} = w_C + w_A + w_B + w_D + 2w_1$
como $w_1$ representa o peso de uma aresta tem-se $w_1 \geq 0$ e obtém-se a condição dos quatro pontos:

$$d_{AB} + d_{CD} \leq d_{AD} + d_{BC} = d_{CA} + d_{BD} \tag{3.14}$$

A condição de aditividade pode ser relaxada para

$$d_{AB} + d_{CD} \leq d_{AD} + d_{BC} \quad e \quad d_{AB} + d_{CD} \leq d_{CA} + d_{BD} \tag{3.15}$$

de forma a que ainda seja válida para sequências em que as mudanças nas sequências não sejam totalmente aditivas [16].

Uma matriz aditiva $D$ de ordem $n$ na qual se verificam também as seguintes desigualdades

$$d_{ij} \leq \max\{d_{ik}, d_{kj}\} \quad \forall i, j, k = 1 \ldots n \tag{3.16}$$

diz-se *ultramétrica* [4].

Se a matriz for ultramétrica, as distâncias entre dois taxons e o seu antepassado comum é a mesma [20].

Seguidamente, apresentam-se dois métodos de construção da árvore filogenética usando a matriz distância, o método da média aritmética não ponderada (UPGMA) e o método dos mínimos quadrados. No próximo capítulo serão apresentados outros métodos que, também, usam a matriz de distâncias para a construção de uma árvore filogenética mas que se podem agrupar em métodos da evolução mínima.

### 3.3.3 Método da média aritmética não ponderada

O método da média aritmética não ponderada ou UPGMA (*unweighted pair-group method using an arithmetic average*), cujas versões mais antigas datam de 1958, constrói fenogramas e era usado originalmente para representar a similaridade global entre um grupo de espécies na taxonomia numérica. Contudo, este método, também, pode ser usado para construir árvores filogenéticas a partir de dados moleculares quando a taxa de substituições de gene for mais ou menos constante [21]. Este método também assume que as distâncias são aproximadamente ultramétricas [20].

Este método junta os dois vértices externos cuja distância na matriz é a menor. De seguida, agrupa esses dois taxons e forma um novo conjunto. Depois volta a determinar a matriz de distâncias de forma a que a distância entre esse novo conjunto e os outros seja calculada através da média aritmética. Continua desta forma até serem incluídos todos os taxons e por fim o método prevê uma posição para a raiz da árvore [17, 20].

O algoritmo para a construção de árvores filogenéticas usando o método da média aritmética não ponderada será descrito através de um exemplo.

Suponhamos que queremos determinar a árvore filogenética para 5 taxons. A matriz de distâncias é dada por:

| $Taxon$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 0 | $d_{12}$ | $d_{13}$ | $d_{14}$ | $d_{15}$ |
| 2 | | 0 | $d_{23}$ | $d_{24}$ | $d_{25}$ |
| 3 | | | 0 | $d_{34}$ | $d_{35}$ |
| 4 | | | | 0 | $d_{45}$ |
| 5 | | | | | 0 |

(3.17)

onde $d_{ij}$ representa a distância entre o taxon $i$ e o taxon $j$.

Este método começa por escolher os dois taxons cuja distância na matriz é a menor. Suponhamos que $d_{12}$ é a menor de todas as distâncias apresentadas na matriz. Os taxons 1 e 2 são ligados por duas arestas com peso $\frac{d_{12}}{2}$ a um vértice interno $(1e2)$ que representa o seu antepassado comum. Assume-se que o peso das arestas que ligam estes dois vértices externos ao interno é o mesmo. Determinam-se agora as distâncias entre esse novo taxon composto $i = (1e2)$ e os restantes taxons $k(k \neq 1 \text{ e } k \neq 2)$ da seguinte forma:

$$d_{ik} = \frac{d_{1k} + d_{2k}}{2} \quad \forall k = 3, 4, 5 \tag{3.18}$$

obtendo-se a seguinte matriz de distâncias:

| $Taxon$ | $i = (1e2)$ | 3 | 4 | 5 |
|---|---|---|---|---|
| $i = (1e2)$ | 0 | $d_{i3}$ | $d_{i4}$ | $d_{i5}$ |
| 3 | | 0 | $d_{34}$ | $d_{35}$ |
| 4 | | | 0 | $d_{45}$ |
| 5 | | | | 0 |

(3.19)

Escolhe-se novamente a menor distância nesta nova matriz. Suponhamos que $d_{i3}$ é a menor de todas as distância apresentadas nesta nova matriz. Os taxons i e 3 são ligados por duas arestas com peso $\frac{d_{i3}}{2} = \frac{d_{13}+d_{23}}{2\times 2}$ a um vértice interno $(1e2e3)$, que representa o seu antepassado comum. A distância entre este novo vértice $j = (1e2e3)$, que resulta do agrupamentos dos vértices i e 3, e os restantes vértices é calculado da seguinte forma:

$$d_{jk} = \frac{d_{1k} + d_{2k} + d_{3k}}{3} \quad \forall k = 4, 5 \tag{3.20}$$

obtendo-se a seguinte matriz de distâncias:

| $Taxon$ | $j = (1e2e3)$ | 4 | 5 |
|---|---|---|---|
| $j = (1e2e3)$ | 0 | $d_{j4}$ | $d_{j5}$ |
| 4 | | 0 | $d_{45}$ |
| 5 | | | 0 |

(3.21)

Novamente suponhamos que $d_{j4}$ é a menor distância apresentada na matriz anterior. Ligam-se os taxons $j = (1e2e3)$ e 4 por arestas com peso $\frac{d_{j4}}{2} = \frac{d_{14}+d_{24}+d_{34}}{3\times 2}$. Uma vez que agora só sobra o taxon 5 este será ligado ao vértice resultante da ligação anterior e cujo peso das arestas é $\frac{d_{15}+d_{25}+d_{35}+d_{45}}{4\times 2}$.

Suponhamos que não era $d_{i3}$ a menor distância e sim $d_{45}$ (ou outra qualquer). Neste caso agrupavam-se os taxons 4 e 5 num novo taxon $l = (4e5)$ e o peso das arestas seria $\frac{d_{45}}{2}$. As distâncias entre este novo taxon e os outros seriam, $d_{3l} = \frac{d_{1k}+d_{2k}}{2}$ e

$d_{il} = \frac{d_{14}+d_{15}+d_{24}+d_{25}}{4}$. A matriz ficaria:

| $Taxon$ | $i = (1e2)$ | $3$ | $l = (4e5)$ |
|---|---|---|---|
| $i = (1e2)$ | $0$ | $d_{i3}$ | $d_{il}$ |
| $3$ | | $0$ | $d_{3l}$ |
| $l = (4e5)$ | | | $0$ |

(3.22)

Se nesta matriz a distância $d_{3l}$ é a menor ligar-se-iam os vértices 3 e $l$ e o vértice $i$ seria ligado por último. Caso seja a distância $d_{il}$ a menor ligar-se-iam primeiro os vértices $l$ e $i$ e por fim ligar-se-iam o vértice 3.

Os pesos das arestas quando se agrupam os vértices A e B são dados por:

$$\sum_{ij} \frac{d_{ij}}{rs} \tag{3.23}$$

sendo o somatório efectuado para os taxons $i$ e $j$, sendo que o taxon $i$ pertence ao agrupamento para chegar ao vértice A e o taxon $j$ pertence ao agrupamento para chegar ao vértice B e onde $r$ e $s$ representam os número de vértices agrupados até obter o vértice A e o número de vértices agrupados até obter o vértice B, respectivamente, e $dij$ é a distância entre o taxon $i$ e do taxon $j$. Os pesos das arestas que ligam o vértice A ao vértice B é dado por $\frac{d_{AB}}{2}$ [21].

Tendo em conta a forma como este método constrói a árvore filogenética, a árvore obtida é sempre uma árvore com raiz. Por vezes, torna-se mais conveniente ter uma árvore sem raiz, por exemplo, para poder comparar a árvore obtida com outras obtidas através de outros métodos. Neste caso pode não se considerar a raiz dada pelo método [21]. A desvantagem do método UPGMA é o facto de este supor que as distâncias de qualquer vértice externo à raiz é a mesma. Se as taxas de substituição são constantes, ou seja, se existe o pressuposto dum "relógio molecular", os resultados obtidos pelo método são aceitáveis, mas se a taxa de substituição variar o método UPGMA obtém maus resultados [17]. Quando não existe o pressuposto dum "relógio molecular" devem usar-se métodos que possibilitem diferenciar as taxas de substituição dos nucleótidos, sendo o método dos mínimos quadrados um desses métodos.

### 3.3.4 Método dos mínimos quadrados

Existem várias versões do método dos mínimos quadrados, sendo o método dos mínimos quadrados ordinário e o método dos mínimos quadrados ponderado os mais usados [21]. Os primeiros a introduzirem o método dos mínimos quadrados ordinário foram Cavalli-Sforza e Edwards [6] em 1967. Neste método consideram-se dois tipos de distâncias: as distâncias observadas, $d_{ij}$, que correspondem às distâncias da matriz de distâncias e as estimativas das distâncias esperadas $e_{ij}$ que correspondem, numa dada

topologia, à soma dos pesos das arestas que ligam o vértice externo $i$ ao vértice externo $j$ [2]. No método dos mínimos quadrados ordinários determina-se para cada topologia possível a seguinte soma residual:

$$R_S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{ij} - e_{ij})^2 . \tag{3.24}$$

A topologia com o menor valor de $R_S$ será escolhida como sendo a árvore filogenética óptima [21].

Cavalli-Sforza e Edwards [6] consideram que a matriz de distâncias é aditiva e que essas distância $d_{ij}$ entre pares de dados moleculares são variáveis aleatórias, uniformemente distribuídas e independentes. Fitch e Margoliash [13] discordaram com esse ponto de vista, considerando que devido à história evolutiva comum das espécies analisadas e à presença de erros de amostragem nos dados moleculares, as distâncias $d_{ij}$ podem não ser variáveis aleatórias, uniformemente distribuídas e independentes [2]. Assim estes autores propuseram a seguinte soma residual:

$$R_S = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(d_{ij} - e_{ij})^2}{d_{ij}}. \tag{3.25}$$

O método dos mínimos quadrados que usa esta soma residual é conhecido como método dos mínimos quadrados ponderado. Na prática os dois métodos acima referidos obtém a mesma topologia ou topologias muito parecidas [21].

Makarenkov e Lapointe [19] provaram que encontrar uma árvore filogenética usando o método dos mínimos quadrados ordinário e o método dos mínimos quadrados ponderado são problemas NP-Difícil.

Antes de calcular a soma residual é necessário determinar as estimativas das distâncias esperadas $e_{ij}$. Rzhetsky e Nei [26] desenvolveram um algoritmo (método dos mínimos quadrados) que permite determinar as estimativas das distâncias esperadas para qualquer topologia. Iremos usar um exemplo para ilustrar esse algoritmo. Consideremos a topologia com cinco vértice externos que se encontra na Figura 3.10.

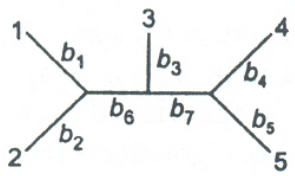


Figura 3.10: Topologia com cinco vértices externos.

Pretende-se determinar os pesos das arestas $b_1, b_2, \ldots$ e $b_7$. As distâncias da matriz de distâncias $d_{ij}$ ($i = 1, \ldots, 4$ e $j = i + 1, \ldots, 5$) podem ser reescritas, usando os pesos

das arestas da topologia dada, da seguinte forma:

$$\begin{array}{lllllllllllllllll}
d_{12} & = & b_1 & + & b_2 & & & & & & & & & & & + & \epsilon_{12} \\
d_{13} & = & b_1 & & & + & b_3 & & & & & + & b_6 & & & + & \epsilon_{13} \\
d_{14} & = & b_1 & & & & & + & b_4 & & & + & b_6 & + & b_7 & + & \epsilon_{14} \\
d_{15} & = & b_1 & & & & & & & + & b_5 & + & b_6 & + & b_7 & + & \epsilon_{15} \\
d_{23} & = & & & b_2 & + & b_3 & & & & & + & b_6 & & & + & \epsilon_{23} \\
d_{24} & = & & & b_2 & & & + & b_4 & & & + & b_6 & + & b_7 & + & \epsilon_{24} \\
d_{25} & = & & & b_2 & & & & & + & b_5 & + & b_6 & + & b_7 & + & \epsilon_{25} \\
d_{34} & = & & & & & b_3 & + & b_4 & & & & & + & b_7 & + & \epsilon_{34} \\
d_{35} & = & & & & & b_3 & & & + & b_5 & & & + & b_7 & + & \epsilon_{35} \\
d_{45} & = & & & & & & & b_4 & + & b_5 & & & & & + & \epsilon_{45}
\end{array}$$

onde $\epsilon_{ij}$ ($i = 1, \ldots, 4$ e $j = i + 1, \ldots, 5$) são erros de amostragem. Usando a notação matricial o conjunto das equações terá a seguinte representação:

$$d = Ab + \epsilon \tag{3.26}$$

onde $d^T = [d_{12}, d_{13}, \ldots, d_{45}]$, $b^T = [b_1, b_2, \ldots, b_7]$ e $\epsilon^T = [\epsilon_{12}, \epsilon_{13}, \ldots, \epsilon_{45}]$ são vectores coluna. Sendo $n$ o número de sequências em estudo, os vectores $d$ e $\epsilon$ têm $\frac{n(n-1)}{2}$ elementos e $b$ tem $(2n - 3)$ elementos. $A$ é a matriz que representa a topologia cujos elementos $a_{ij}$ tomam valor 1 se a aresta $b_j$ estiver no caminho entre o nodo $i_a$ e $i_b$ (em que $i = 1, \ldots, 10$ corresponde aos nodos $(i_a, i_b)$ com $i_a = 1, \ldots, 4$ e $i_b = i_a + 1, \ldots, 5$) e 0 caso contrário (ver as equações acima). Neste exemplo a matriz $A$ é dada por,

$$A = \begin{bmatrix}
1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 0 & 1 & 1 \\
1 & 0 & 0 & 0 & 1 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 0 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0
\end{bmatrix}$$

As estimativas dos pesos, $b$, são determinadas usando a seguinte fórmula:

$$\hat{b} = \left(A^T A\right)^{-1} A^T d = Ld \tag{3.27}$$

onde $Ld = \left(A^T A\right)^{-1} A^T$.

Assim, obtém-se neste exemplo:

$$\begin{aligned}
\hat{b_1} &= \tfrac{1}{2}d_{12} + \tfrac{1}{6}\left(d_{13} - d_{23} + d_{14} - d_{24} + d_{15} - d_{25}\right)\\
\hat{b_2} &= \tfrac{1}{2}d_{12} + \tfrac{1}{6}\left(d_{13} - d_{23} + d_{14} - d_{24} + d_{15} - d_{25}\right)\\
\hat{b_3} &= \tfrac{1}{4}\left(d_{13} + d_{23} + d_{34} + d_{35}\right) + \tfrac{1}{8}\left(d_{14} + d_{24} + d_{15} + d_{25}\right)\\
\hat{b_4} &= \tfrac{1}{2}d_{45} + \tfrac{1}{6}\left(d_{14} - d_{15} + d_{24} - d_{25} + d_{34} - d_{35}\right)\\
\hat{b_5} &= \tfrac{1}{2}d_{45} + \tfrac{1}{6}\left(d_{14} - d_{15} + d_{24} - d_{25} + d_{34} - d_{35}\right)\\
\hat{b_6} &= -\tfrac{1}{2}d_{12} + \tfrac{1}{4}\left(d_{13} + d_{23} - d_{34} - d_{35}\right) + \tfrac{1}{8}\left(d_{14} + d_{24} + d_{15} + d_{25}\right)\\
\hat{b_7} &= \tfrac{1}{4}\left(d_{34} + d_{35} - d_{13} - d_{23}\right) + \tfrac{1}{8}\left(d_{14} + d_{24} + d_{15} + d_{25}\right) - \tfrac{1}{2}d_{45}
\end{aligned}$$

A determinação das estimativas dos pesos $b$ através deste sistema torna-se bastante complicado se o número de sequências em estudo for grande. Para ultrapassar essa dificuldade Rzhetsky e Nei [27] desenvolveram uma outra fórmula para calcular os pesos, a qual exemplificaremos através de um exemplo. Considerando agora uma topologia com sete vértices exteriores (ver Figura 3.11). Escolhendo uma aresta interior, por exemplo

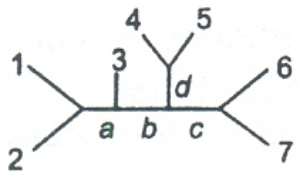


Figura 3.11: Topologia com sete vértices exteriores.

a aresta $b$ pode-se representar a topologia dada na forma da topologia apresentada na Figrua 3.12, onde $A, B, C$ e $D$ representam, respectivamente, os seguintes agrupamentos de vértices exteriores (3), (1, 2), (4, 5) e (6,7).

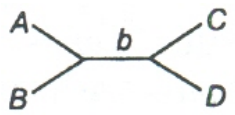


Figura 3.12: Topologia com vértices agrupados.

Neste caso o peso da aresta $b$ é determinado recorrendo à seguinte fórmula:

$$\hat{b} = \tfrac{1}{2}\left[\gamma\left(\tfrac{d_{AC}}{n_A n_C} + \tfrac{d_{BD}}{n_B n_D}\right) + (1-\gamma)\left(\tfrac{d_{BC}}{n_B n_C} + \tfrac{d_{AD}}{n_A n_D}\right) - \tfrac{d_{AB}}{n_A n_B} - \tfrac{d_{CD}}{n_C n_D}\right]$$

onde $n_A, n_B, n_C$ e $n_D$ representam o número de vértices no agrupamento $A, B, C$ e $D$, respectivamente, $d_{IJ}$ é a soma das distâncias entre todos os vértices do agrupamento $I$ e todos os vértices do agrupamento $J$ e $\gamma$ é dado pela seguinte expressão:

$$\gamma = \frac{n_B n_C + n_A n_D}{(n_A + n_B)(n_C + n_D)}.$$

Para determinar o peso de uma aresta exterior (ver Figura 3.13) a fórmula é

$$\hat{b} = \frac{\frac{d_{AB}}{n_B} + \frac{d_{AC}}{n_C} - \frac{d_{BC}}{n_A n_B}}{2}$$

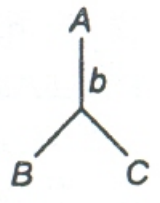


Figura 3.13: Topologia com vértices agrupados

Depois de obter os pesos de todas as arestas, para determinar as estimativas das distâncias esperadas $e_{ij}$, basta somar os pesos das arestas que ligam os vértices $i$ e $j$ [21, 26, 27].

# Capítulo 4

# Problema da evolução mínima

Um dos critérios mais usados para inferir árvores filogenéticas é o da evolução mínima, que foi introduzido pela primeira vez por Kidd e Sgaramella-Zonta [18] e mais tarde reinterpretado por Rzhetsky e Nei [26]. Este critério afirma que a melhor árvore filogenética é aquela que tem comprimento mínimo. Usando este critério como função objectivo, a topologia de árvore que optimizar esta função será a árvore filogenética óptima, para um conjunto de espécies em estudo [8]. Assim, encontrar essa árvore filogenética óptima envolve resolver um problema de optimização que se designa por problema de evolução mínima. Este problema é estatisticamente consistente mas NP-Difícil [2].

Dada a matriz de distâncias D o problema da evolução mínima na sua forma geral tem a seguinte formulação [3]:

$$\begin{aligned} & \min_{(X,w)} L(X,w) \\ \text{s.a.} \quad & f(D,X,w) = 0 \\ & X \in \chi \\ & w \in \Re_{0^+}^{(2n-3)} \end{aligned}$$

onde $X$ é uma árvore filogenética representada pela matriz de incidência aresta-caminho, $\chi$ representa o conjunto de todas as possíveis topologias de árvores , $L(X,w)$ indica o comprimento da árvore filogenética $X$ em função dos pesos das arestas $w$ e $f(D,X,w)$ é uma função que relaciona a matriz de distâncias $D$ com a árvore filogenética $X$ e os pesos das arestas $w$. Esta última função tem em consideração que o comprimento do caminho que liga dois vértices externos $i$ e $j$ na árvore não pode ser inferior à distância entre esses dois vértices, $d_{ij}$, apresentada na matriz de distâncias [5].

O problema da evolução mínima pode ser dividido em dois subproblemas [3]:

- Determinar a topologia da árvore filogenética, ou seja, determinar as entradas da matriz de incidência $X$;
- Determinar os pesos das arestas da árvore filogenética, $w$, que melhor se ajustam

à matriz de distâncias $D$.

Existem várias versões do problema da evolução mínima consoante a forma como se determinam os pesos $w$ que, por sua vez, dependem da escolha da função $f(D, X, w)$ e da função $L(X, w)$. A função $L(X, w)$ é geralmente definida como sendo a soma dos pesos das arestas.

Tradicionalmente, para resolver o problema da evolução mínima começa-se com a topologia obtida através do método neighbor-joining e faz-se uma pesquisa topológica a partir desse ponto [8].

## 4.1 Método neighbor-joining

O método neighbor-joining é eficiente em obter a topologia de árvore correcta, comparado com o método da máxima parcimónia e vários outros [26]. Ele foi inicialmente proposto por Saitou e Nei [28] e posteriormente modificado por Studier e Kepler [29]. Este método constrói a árvore sem raiz cuja soma dos todos os pesos das arestas é mínima. Inicia-se com uma topologia em forma de estrela, com um vértice interno hipotético, e os $n$ vértices externos representando cada uma das espécies em estudo. Iterativamente procura-se o vizinho mais próximo, ou seja, o par de vértices que induz uma árvore cuja soma dos pesos das arestas é a menor. O par assim escolhido é então agrupado num novo vértice interno e as distâncias entre esse novo vértice e os restantes são calculadas, obtendo-se assim uma nova matriz de distâncias que será usada na próxima iteração. O algoritmo encerra quando se tiverem inseridos na árvore $(n-2)$ vértices internos [23].

Na Figura 4.1 apresenta-se uma árvore em forma de estrela e a árvore induzida pelo agrupamento dos vértices 1 e 2.

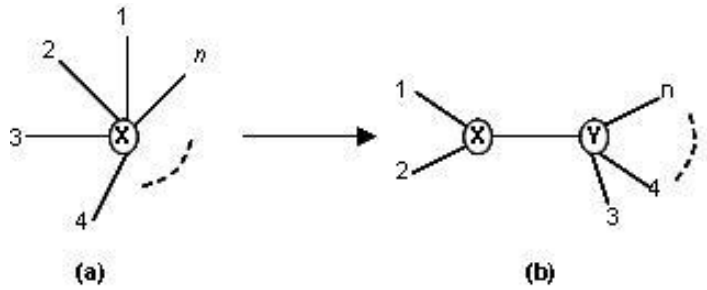


Figura 4.1: (a) Árvore em forma de estrela; (b) Árvore induzida pelo agrupamento dos vértices 1 e 2.

Em cada uma das $(n-2)$ iterações o método determina, para cada possível par de agrupamentos de vértices $i$ e $j$, a soma dos pesos das arestas $S_{ij}$, que pode ser obtida

usando as distâncias da matriz de distâncias dada. Inicialmente na árvore em forma de estrela a soma é determinada usando a seguinte expressão:

$$S_0 = \sum_{k=1}^{n} L_{kX} = \frac{1}{n-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} \tag{4.1}$$

onde $L_{kX}$ é o peso da aresta que liga o vértice externo $k$ ao hipotético vértice interno $X$ e $d_{ij}$ é a distância entre o vértice externo $i$ e o vértice externo $j$ da matriz de distâncias.

Nos outros casos $S_{ij}$ é determinado usando a seguinte fórmula:

$$S_{ij} = \frac{1}{2(n-2)} \sum_{\substack{k=1 \\ k \neq i,j}}^{n} (d_{ik} + d_{jk}) + \frac{1}{2} d_{12} + \frac{1}{n-2} \sum_{\substack{l=k+1 \\ k,l \neq i,j}}^{n} d_{kl}. \tag{4.2}$$

Escolhidos os vértices $i$ e $j$ a serem agrupados num novo vértice interno $X$ determinam-se os pesos das novas arestas formadas:

$$L_{iX} = \frac{d_{ij} + d_{iz} - d_{jz}}{2} \tag{4.3}$$

$$L_{jX} = \frac{d_{ij} + d_{jz} - d_{iz}}{2} \tag{4.4}$$

onde

$$d_{iz} = \frac{1}{n-2} \sum_{\substack{k=1 \\ k \neq i,j}} d_{ik} \tag{4.5}$$

$$d_{jz} = \frac{1}{n-2} \sum_{\substack{k=1 \\ k \neq i,j}} d_{jk} \tag{4.6}$$

e determinam-se, também, as distâncias entre esse novo vértice interno $A$ e os restantes vértices externos $(k, k \neq i, j)$ através da seguinte fórmula:

$$d_{Ak} = \frac{d_{ik} + d_{jk} - d_{ij}}{2}. \tag{4.7}$$

Desta forma obtém-se uma nova matriz de distâncias que será usada na iteração seguinte para o cálculo das novas somas $S'_{ij}$ [21, 28].

Para ilustrar o método apresentamos na Tabela 4.1 as várias iterações para a construção de uma árvore filogenética relativa a seis taxons e na Figura 4.2 a construção da árvore ao longo das várias iterações [21].

O método neighbor-joining determina uma topologia de árvore muito próxima da árvore que satisfaz o critério da evolução mínima [17] e se a matriz for aditiva Saitou e

| | Matriz de distâncias | | | | | | Soma dos pesos das arestas $S_{ij}$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Primeira Iteração | | | | | | | | | | | |
| | 1 | 2 | 3 | 4 | 5 | 6 | 1 | 2 | 3 | 4 | 5 | 6 |
| 1 | 0 | 9 | 12 | 15 | 20 | 16 | | **29.5** | 32.5 | 33 | 33.5 | 33.5 |
| 2 | | 0 | 7 | 10 | 15 | 11 | | | 32.5 | 33 | 33.5 | 33.5 |
| 3 | | | 0 | 5 | 10 | 6 | | | | 32 | 32.5 | 32.5 |
| 4 | | | | 0 | 11 | 7 | | | | | 32 | 32 |
| 5 | | | | | 0 | 8 | | | | | | 30.5 |
| | Agrupam-se os vértices 1 e 2, A=(1,2) e $d_{A1} = 7$, $d_{A2} = 2$ | | | | | | | | | | | |
| | Segunda Iteração | | | | | | | | | | | |
| | A | 3 | 4 | 5 | 6 | | A | 3 | 4 | 5 | 6 | |
| A | 0 | 5 | 8 | 13 | 9 | | | 19.7 | 20.3 | 21 | 21 | |
| 3 | | 0 | 5 | 10 | 6 | | | | 20.3 | 21 | 21 | |
| 4 | | | 0 | 11 | 7 | | | | | 20.7 | 20.7 | |
| 5 | | | | 0 | 8 | | | | | | **19.3** | |
| | Agrupam-se os vértices 5 e 6, B=(5,6) e $d_{B5} = 6$, $d_{B6} = 2$ | | | | | | | | | | | |
| | Terceira Iteração | | | | | | | | | | | |
| | A | 3 | 4 | B | | | A | 3 | 4 | B | | |
| A | 0 | 5 | 8 | 7 | | | | **11** | 11.5 | 11.5 | | |
| 3 | | 0 | 5 | 4 | | | | | 11.5 | 11.5 | | |
| 4 | | | 0 | 5 | | | | | | 11 | | |
| | Agrupam-se os vértices A e 3, C=(A,3) e $d_{CA} = 4$, $d_{C3} = 1$ | | | | | | | | | | | |

Tabela 4.1: Iterações para a construção de uma árvore filogenética usando o método neighbor-joining.

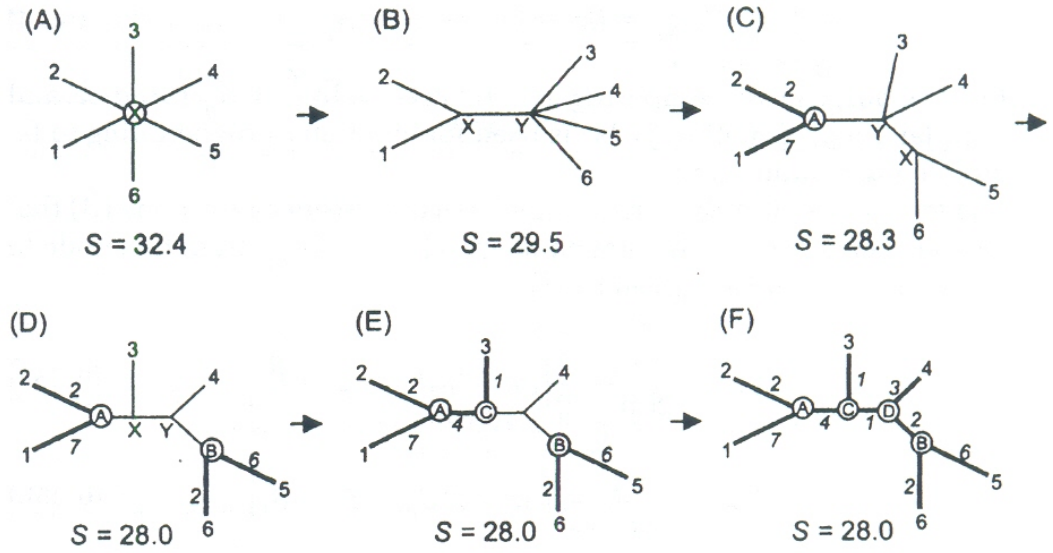


Figura 4.2: Ilustração da construção da árvore ao longo das várias iterações do método neighbor-joining.

Nei [28] provaram que o método determina o peso correcto de todas as arestas.

O método é simples e eficiente, mas a sua eficiência advêm de ele apenas explorar uma pequena parte do espaço de solução, ou seja, algumas topologias de árvores possíveis e nesse sentido pode deixar de encontrar uma solução óptima global [23].

Para resolver o problema da evolução mínima podemos, também, usar a programação linear.

## 4.2 Os modelos de programação linear

Segundo Catanzaro [3] o primeiro trabalho a considerar formulações em programação linear para o problema da evolução mínima foram de Beyer et al. [1] que observaram que as distâncias evolutivas entre dados moleculares têm de satisfazer a desigualdade triangular (3.12) uma vez que reflectem o número de mutações requeridas ao longo do tempo para transformar uma sequência noutra. Por outro lado, como os pesos das arestas numa árvore filogenética também representam distâncias evolutivas estas também deverão satisfazer a desigualdade triangular. Beyer et al. [1] propuseram um modelo de programação linear em que para uma dada topologia os $(2n-3)$ pesos devem satisfazer $\frac{n(n-1)}{2}$ desigualdades triangulares. Waterman et al. [32] modificaram o modelo de Beyer et al. ao misturar a propriedade aditiva da matriz de distâncias (3.13) com a desigualdade triangular impondo:

$$w_e \geq 0 \quad e = 1, \ldots, 2n-3 \tag{4.8}$$

$$\sum_{e \in p_{ij}} w_e \geq d_{ij} \quad i, j = 1, \ldots, n \quad i < j \tag{4.9}$$

onde $n$ é o número de vértices externos, ou seja, o número de espécies em estudo, $w_e$ representa o peso da aresta $e$ e $p_{ij}$ representa o caminho que liga o vértice $i$ ao vértice $j$.

Beyer et al . [1] e Waterman et al. [32] sugeriram as seguintes funções objectivo para a sua formulação em programação linear:

$$\text{Ł}(X, w) = \sum_{e=1}^{2n-3} w_e \tag{4.10}$$

$$\text{Ł}(X, w) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\sum_{e \in p_{ij}} w_e - d_{ij}}{d_{ij}} \tag{4.11}$$

e

$$Ł(X, w) = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{\sum_{e \in p_{ij}} w_e - d_{ij}}{d_{ij}^2} \tag{4.12}$$

Antes de descrevermos algumas das formulações para o problema de evolução mínima que envolvem a obtenção de uma árvore de suporte de custo mínimo com algumas restrições, começamos por descrever formalmente o problema.

### 4.2.1 Descrição do problema

A árvore filogenética óptima corresponde a uma árvore de suporte de custo mínimo com algumas restrições adicionais que descrevemos nesta secção.

Considere o grafo $G = (V, E)$ onde $V$ é o conjunto de vértices $V = V_{int} \cup V_{ext}$ e $E$ o conjunto das arestas. Em particular $V_{ext}$ é o conjunto das folhas ou vértices externos que representam os $n$ taxons e $V_{int}$ o conjunto dos vértices internos que representam os $(n-2)$ antepassados comuns aos taxons. O conjunto $E$ das arestas contém todas as possíveis arestas entre vértices internos (definindo assim um grafo completo no conjunto dos vértices $V_{int}$) e as arestas que ligam todos os vértices internos de $V_{int}$ a todos os vértices externos de $V_{ext}$. O conjunto das arestas $E$ tem cardinalidade $\frac{(n-2)(n-3)}{2} + (n-2)n = \frac{3(n-1)(n-2)}{2}$. A árvore filogenética é tal que todos os vértices externos são folha e tal que o grau de todos os vértices internos é 3 exceptuando nas árvores com raiz, onde o vértice interno que representa a raiz tem grau 2. As $(2n-3)$ arestas têm um peso $w \geq 0$ associado que representa a distância evolutiva entre o par de vértices ligados pela aresta. A árvore filogenética pode ser representada por uma matriz binária, a matriz de incidência aresta-comprimento.

Uma vez que uma árvore filogenética é uma árvore de suporte de custo mínimo, apresentamos, de seguida, um modelo para o problema da árvore de suporte de custo mínimo, considerando uma árvore com raiz, a qual será denotada por 0 e pertence ao conjunto $V_{int}$, com a restrição adicional de o grau de todos os vértices internos ser 3.

Consideremos as variáveis $x_{ij}$ (com $i \in V_{int}, j \in V \setminus \{0\}$ e $i \neq j$) que indicam se a aresta que liga o vértice $i$ ao vértice $j$, denotada por $e_{ij}$ pertence ou não à árvore, ou seja,

$$x_{ij} = \begin{cases} 1, & \text{se a aresta } e_{ij} \text{ está na árvore} \\ 0, & \text{caso contrário} \end{cases}$$

e as variáveis de fluxo $y_{ij}^k$ (com $i \in V_{int}, j, k \in V \setminus \{0\}$ e $i \neq j$) que indicam se a aresta $e_{ij}$ é ou não usada no caminho da raiz para o vértice $k$, ou seja,

$$y_{ij}^k = \begin{cases} 1, & \text{se a aresta } e_{ij} \text{ é usada no único caminho da raiz para o vértice } k \\ 0, & \text{caso contrário} \end{cases}$$

Uma formulação para este problema em Programação Linear Inteira como um problema de fluxos num grafo tem a seguinte forma:

Modelo MSPT

$$min \quad \sum_{e_{ij} \in E} c_{ij} x_{ij}$$

*sujeito a*

$$\sum_{\substack{i \in V_{int} \\ i<j}} x_{ij} = 1, \qquad \forall j \in V \setminus \{0\} \tag{4.13}$$

$$\sum_{\substack{j \in V \\ j>i}} x_{ij} = 2, \qquad \forall i \in V_{int} \tag{4.14}$$

$$\sum_{\substack{j \in V_{int} \cup \{k\} \\ j \neq 0}} y_{0j}^k = 1, \qquad \forall k \in V_{ext} \tag{4.15}$$

$$\sum_{i \in V_{int}} y_{ik}^k = 1, \qquad \forall k \in V_{ext} \tag{4.16}$$

$$\sum_{\substack{i \in V_{int} \\ i<j}} y_{ij}^k = \sum_{\substack{i \in V_{int} \cup \{k\} \\ i>j \\ i \neq 0}} y_{ji}^k, \qquad \forall k \in V_{ext}, \forall j \in V_{int \setminus \{0\}} \tag{4.17}$$

$$y_{ij}^k \leq x_{ij}, \qquad \forall k \in V_{ext}, \forall i \in V_{int}, \forall j \in V \setminus \{0\}, j > i \tag{4.18}$$

$$x_{ij} \in \{0,1\} \qquad \forall i \in V_{int}, \forall j \in V \setminus \{0\} \tag{4.19}$$

$$y_{ij}^k \in \{0,1\} \qquad \forall k in V_{ext}, \forall i \in V_{int}, \forall j \in V \setminus \{0\} \tag{4.20}$$

As restrições (4.13) e (4.14) impõem que cada vértice interno, à excepção da raiz, tenha grau 3. O facto da raiz ter grau 2 é imposto pela restrição (4.14) quando $i = 0$. As restrições (4.15), (4.16) e (4.17) são as restrições usuais de conservação de fluxo. As restrições (4.18) são as restrições de ligação que estabelecem uma relação entre as variáveis $x_{ij}$ e as variáveis $y_{ij}^k$ de forma que não exista nenhum fluxo na aresta $e_{ij}$ se esta não se encontra na solução. Finalmente, as restrições (4.19) e (4.20) são as restrições de integralidade das variáveis.
Denotaremos por SPT o conjunto de restrições (4.13) - (4.20) desta formulação.

Contudo agora temos de determinar os valores para os pesos $w$ de cada aresta que representam a distância evolutiva entre o par de vértices incidente na aresta. De entre as árvores de suporte cujos vértices internos têm grau três há que seleccionar aquela cujos pesos atribuídos a cada aresta representem a distância evolutiva mínima.

Catanzaro et al. [4] definiram várias formulações, para a construção de árvores filogenéticas sem raiz segundo o critério da evolução mínima, que diferem na forma

como os pesos são definidos, dos quais iremos apresentar duas, o modelo de caminhos e o modelo de fluxos.

### 4.2.2 Modelo de caminhos

O modelo de caminhos identifica a árvore filogenética óptima como sendo a árvore de suporte de custo mínimo que satisfaz mais algumas restrições que definem caminhos na árvore. Consideram-se as variáveis de decisão $x_e (e \in E)$ que indicam se a aresta $e$ pertence ou não à árvore, ou seja,

$$x_e = \begin{cases} 1, & \text{se a aresta } e \text{ pertence à árvore filogenética} \\ 0, & \text{caso contrário} \end{cases}$$

e as variáveis contínuas não negativas $w_e$ que indicam o peso de cada aresta $e$.

No modelo $p_{ij} \subseteq E$ denota o conjunto de arestas de um caminho genérico que liga os vértices externos $i$ e $j$; $P_{ij} = \{p_{ij}\}$ denota o conjunto de todos os caminhos $p_{ij}$ possíveis em $G$ e $\hat{d} = \max\{d_{ij} : i, j \in V_{ext}\}$ representa a distância máxima apresentada na matriz de distâncias. Para todo o subconjunto $S \subseteq V, E(S) \subseteq E$ denota o subconjunto das arestas induzidas pelos vértices de $S$, $\delta(i) \subseteq E$ denota o subconjunto de arestas incidentes no vértice $i \in V$ e para $\hat{E} \subseteq E, x(\hat{E}) = \sum_{e \in \hat{E}} x_e$. Assim, a formulação apresentada por Catanzano et al. [4] para o problema da evolução mínima é a seguinte:

Modelo de caminhos

$$min \quad \sum_{e \in E} w_e$$

*sujeito a*

$$x(E(S)) \leq |S| - 1 \qquad \forall S \subset V \tag{4.21}$$

$$x(E(V)) = 2n - 3 \tag{4.22}$$

$$x(\delta(i)) = 3 \qquad \forall i \in V_{int} \tag{4.23}$$

$$w_e \leq \hat{d} x_e \qquad \forall e \in E \tag{4.24}$$

$$\sum_{e \in p_{ij}} (w_e + d_{ij}(1 - x_e)) \geq d_{ij} \qquad \forall p_{ij} \in P_{ij}, \forall i, j \in V_{ext}, i < j \tag{4.25}$$

$$x_e \in \{0, 1\} \qquad \forall e \in E \tag{4.26}$$

$$w_e \geq 0 \qquad \forall e \in E \tag{4.27}$$

Através das restrições (4.21)-(4.23) (que denotaremos por SPT1) e (4.26) as variáveis $x_e$ definem uma árvore cujos vértices interiores têm grau três. As restrições (4.24) impõe que o peso $w_e$ da aresta $e$ é positivo apenas no caso em que essa aresta $e$ pertence à árvore. As restrições (4.25) obrigam a soma dos pesos ao longo do caminho $p_{ij}$ na árvore a não ser inferior a $d_{ij}$. Finalmente, as restrições (4.27) impõe que os pesos $w_e$ sejam

não negativos.

Este modelo caracteriza-se por ter um número relativamente pequeno ($O(n^2)$) de variáveis binárias e contínuas e por ter um número exponencial de restrições. Em particular o conjunto de restrições (4.25) uma vez que devem ser considerados todos os caminhos entre todos os pares de vértices externos $i < j$.

### 4.2.3 Modelo de fluxos

No modelo de fluxos os pesos não negativos são representados por potenciais desconhecidos de forma a garantir uma distância mínima entre cada par de vértices externos.

Consideram-se as variáveis $x_{ij}$ (com $i \in V_{int}, j \in V$ e $i \neq j$) que indicam se a aresta que liga o vértice $i$ ao vértice $j$, denotada por $e_{ij}$, pertence ou não à árvore filogenética, ou seja,

$$x_{ij} = \begin{cases} 1, & \text{se a aresta } e_{ij} \text{ pertence à árvore filogenética} \\ 0, & \text{caso contrário} \end{cases}$$

As variáveis $x_{ij}$ poderão também ser representadas por $x_e$ e neste caso $x_e$ toma o valor 1 se a aresta $e$ pertence à árvore filogenética e toma o valor 0 caso contrário. Consideram-se, ainda, as variáveis contínuas não negativas $w_{ij}$ que indicam o peso de cada aresta $e_{ij}$ e que também poderão ser representadas por $w_e$. Além destas variáveis, consideram-se, também, as variáveis contínuas $u_{ij}$ (com $i \in V$ e $j \in V_{ext}$) o potencial desconhecido que representa o comprimento do caminho que liga o vértice $i$ ao vértice $j (i \neq j)$.

A formulação apresentada por Catanzano et al. em [4] é a seguinte:

Modelo de fluxos

$$min \quad \sum_{e \in E} w_e$$

$sujeito\ a$

$$x \in SPT1 \tag{4.28}$$

$$u_{ij} \geq d_{ij} \qquad \forall i, j \in V_{ext}, i \neq j \tag{4.29}$$

$$w_{ij} \geq u_{jk} - u_{ik} - \widehat{d}(1 - x_{ij}) \qquad \forall i \in V_{int}, \forall j, k \in V_{ext} \tag{4.30}$$

$$x_e \in \{0, 1\} \qquad \forall e \in E \tag{4.31}$$

$$w_e \geq 0 \qquad \forall e \in E \tag{4.32}$$

Como no modelo anterior as restrições (4.28) e (4.31) obrigam a que as variáveis $x_e$ definam uma árvore e as restrições (4.32) impõem que os pesos $w_e$ sejam não negativos. As restrições (4.29) obrigam todo o comprimento do caminho que liga um vértice externo $i$ a um vértice externo $j$ a não ser inferior a $d_{ij}$. Finalmente, as restrições (4.30) asseguram que a diferença dos comprimentos dos caminhos dum vértice $i$ para um

vértice $j$ não seja maior que $w_{ij}$.

Os autores identificam como desvantagem deste modelo a existência de várias soluções equivalentes que conduzem à mesma topologia de árvore. Apesar de os vértices externos estarem associados a taxons diferentes os vértices internos são elementos idênticos. Isso implica que uma topologia de árvore está associada a várias soluções, cada uma determinada por uma permutação diferente dos vértices internos. Para evitar esta situação os autores sugerem que se fixem algumas variáveis, como por exemplo, $x_{(n-1)1} = x_{12} = x_{23} = 1$, sendo $(n-1)$ um vértice externo e 1,2 e 3 três vértices internos.

Verificamos, claramente que o modelo tem como solução óptima a solução com todas as variáveis $w_{ij}$ iguais a zero.

Consideremos $j, k \in V_{ext}$ e $i \in V_{int}$. As restrições (4.29) impõem que $u_{jk} \geq d_{jk}$, sejam, então, $u_{jk} = d_{jk}$ e $u_{ik} = \max_{\ell \in V_{ext}} d_{k\ell}$. Assim, $u_{jk} - u_{ik} \leq 0$ para todo o $j, k \in V_{ext}$ e $i \in V_{int}$ e consequentemente $u_{jk} - u_{ik} - \hat{d}(1 - x_{ij}) \leq 0$, tornando as restrições (4.30) redundantes.

Além disso, ainda fazemos as seguintes observações:

1. as variáveis $w_{ij}$ estão definidas para todas as arestas em $E$;

2. as restrições (4.29) estabelecem um limite inferior para as variáveis $u_{ij}$, mas apenas para aquelas definidas entre dois vértices externos, não sendo definido nenhum limite inferior para as variáveis $u_{ij}$ definidas entre um vértice interno e um vértice externo e entre dois vértices internos ;

3. as restrições (4.30) estabelecem um limite inferior para as variáveis $w_{ij}$, mas apenas para aquelas definidas entre um vértice interno e um vértice externo, não sendo definido nenhum limite inferior para as variáveis definidas entre dois vértices internos ;

4. o valor de $\hat{d}(1-x_{ij})$ ou é 0, se a aresta $e_{ij}$ pertence à solução, ou é $\hat{d}$ caso contrário.

Pelas observações feitas, verifica-se que o modelo apresentado está incompleto. Assim, no capítulo seguinte apresentamos um modelo para o problema da evolução mínima baseado neste e nas observações anteriormente referidas.

# Capítulo 5

# Uma Formulação Alternativa

Ao longo deste trabalho de mestrado tentou encontrar-se uma formulação que permitisse resolver o problema da evolução mínima. Tendo-se verificado que o modelo de fluxos apresentado por Catanzaro et al. [4] está incompleto vamos completá-lo. Determinamos algumas árvores filogenéticas usando a formulação. Depois tentámos usar um esquema de decomposição que nos permitisse obter soluções para o problema.

## 5.1 Formulação

Considerando as observações no final do Capítulo anterior vamos usá-las para propor uma formulação que completa a apresentada por Catanzaro et al. [4]. Consideremos as variáveis $x_{ij}$, $w_{ij}$ e $u_{ij}$ definidas no modelo de fluxos. Assim, $x_{ij}$ (com $i \in V_{int}, j \in V$ e $i \leq j$) saõ as variáveis binárias que indicam se a aresta que liga o vértice $i$ ao vértice $j$, denotada por $e_{ij}$, pertence ou não à árvore filogenética, ou seja,

$$x_{ij} = \begin{cases} 1, & \text{se a aresta } e_{ij} \text{ pertence à árvore filogenética} \\ 0, & \text{caso contrário} \end{cases}$$

As variáveis contínuas não negativas $w_{ij}$ (com $i \in V_{int}, j \in V$ e $i \leq j$) indicam o peso de cada aresta $e_{ij}$ e as variáveis continuas $u_{ij}$ (com $i \in V$, $j \in V_{ext}$ e $i \neq j$) representam o comprimento do caminho que liga o vértice $i$ ao vértice $j$.

Na nossa segunda observação verificamos que as restrições (4.29) do modelo de fluxos estabelecem um limite inferior para as variáveis $u_{ij}$, mas apenas para aquelas definidas entre dois vértices externos. Assim, começamos por acrescentar restrições de modo a definir um limite inferior para as variáveis $u_{ij}$ definidas entre um vértice interno e um vértice externo e entre dois vértices internos:

$$u_{ij} \geq w_{ij} \qquad \forall i \in V_{int}, \forall j \in V, i < j.$$

Reparamos, também, que as variáveis $u_{ij}$ deveriam tomar o mesmo valor que $w_{ij}$ quando os vértices $i$ e $j$ são adjacentes e que nenhuma das restrições apresentada obrigava a que tal sucedesse. Assim, acrescentamos restrições que em complemento com as acrescentadas para definir o limite inferior, obrigam a que $u_{ij} = w_{ij}$ quando os vértices $i$ e $j$ são adjacentes:

$$w_{ij} \geq u_{ij} - \widehat{d}(1 - x_{ij}) \qquad \forall i \in V_{int}, \forall j \in V, i < j.$$

De seguida, tendo em conta a nossa terceira observação onde verificamos que as restrições (4.30) do modelo de fluxos estabelecem um limite inferior para as variáveis $w_{ij}$, mas apenas para aquelas definidas entre um vértice externo e um vértice interno, completamos essas restrições de modo a definir, também, um limite inferior para as variáveis definidas entre dois vértices internos:

$$w_{ij} \geq u_{jk} - u_{ik} - \widehat{d}(1 - x_{ij}) \qquad \forall i \in V_{int}, \forall j \in V, i < j, \forall k \in V \backslash \{i, j\}.$$

Além disso, verificamos que as restrições (4.30) mesmo depois de serem completadas não eram suficientes, pois se $u_{jk}$ for menor que $u_{ik}$, então $u_{jk} - u_{ik}$ toma um valor negativo tornando as restrições redundantes. Assim, acrescentamos restrições de forma a contemplar a diferença simétrica, ou seja, $u_{ik} - u_{jk}$ e desta forma quando umas restrições são redundantes as outras estabelecem um limite inferior para as variáveis $w_{ij}$:

$$w_{ij} \geq u_{ik} - u_{jk} - \widehat{d}(1 - x_{ij}) \qquad \forall i \in V_{int}, \forall j \in V, i < j, \forall k \in V \backslash \{i, j\}.$$

Finalmente, tem-se pressuposto que o caminho de $i$ para $j$ tem o mesmo comprimento que o caminho de $j$ para $i$, ou seja, $u_{ij} = u_{ji}$, mas nenhuma das restrições apresentadas contemplava esse facto. Como estas variáveis $u_{ij}$ são usadas nas restrições que estabelecem um limite inferior para as variáveis $w$ tanto quando $i < j$ como quando $i > j$ acrescentamos restrições de forma a garantir que $u_{ij} = u_{ji}$.

A formulação que propomos é a seguinte:

Formulação Alternativa

$$min \quad \sum_{i \in V_{int}} \sum_{\substack{j \in V \\ j > i}} w_{ij}$$

*subject to*

$$x \in SPT \tag{5.1}$$

$$u_{ij} \geq d_{ij} \qquad \forall i, j \in V_{ext}, i \neq j \tag{5.2}$$

$$w_{ij} \geq u_{ik} - u_{jk} - \widehat{d}(1 - x_{ij}) \qquad \forall i \in V_{int}, \forall j \in V, i < j, \forall k \in V \backslash \{i, j\} \tag{5.3}$$

$$w_{ij} \geq u_{jk} - u_{ik} - \widehat{d}(1 - x_{ij}) \qquad \forall i \in V_{int}, \forall j \in V, i < j, \forall k \in V \backslash \{i, j\} \tag{5.4}$$

$$w_{ij} \geq u_{ij} - \widehat{d}(1 - x_{ij}) \qquad \forall i \in V_{int}, \forall j \in V, i < j \tag{5.5}$$

$$u_{ij} \geq w_{ij} \qquad \forall i \in V_{int}, \forall j \in V, i < j \tag{5.6}$$

$$u_{ij} = u_{ji} \qquad \forall i \in V_{int}, \forall j \in V, i < j \tag{5.7}$$

$$x_{ij} \in \{0, 1\} \qquad \forall i \in V_{int}, \forall j \in V, i < j \tag{5.8}$$

$$w_{ij} \geq 0 \qquad \forall i \in V_{int}, \forall j \in V, i < j \tag{5.9}$$

As restrições (5.1) e (5.8) obrigam a que as variáveis $x_{ij}$ definem uma árvore e as restrições (5.9) impõem que os pesos $w_{ij}$ sejam não negativos.

As restrições (5.2) obriga todo o comprimento do caminho que liga um vértice externo $i$ a um vértice externo $j$ a não ser inferior a $d_{ij}$, ou seja, estabelecem um limite inferior para as variáveis $u_{ij}$ definidas entre dois vértices externos.

As restrições (5.3) e (5.4) asseguram que a diferença dos comprimentos dos caminhos dum vértice $i$ para um vértice $j$ não seja maior que $w_{ij}$, ou seja, estabelecem um limite inferior para as variáveis $w_{ij}$. Notamos que as restrições (5.3) e (5.4) apenas diferem no cálculo da diferença entre os caminhos. Tirando o caso em que essa diferença seja zero, haverá sempre uma das duas que se torna redundante (aquela cuja diferença é negativa) estabelecendo a outra um limite superior para as variáveis $w_{ij}$.

As restrições (5.5) obrigam, por um lado, que o comprimento do caminho $u_{ij}$ que liga dois vértices $i$ e $j$, não adjacentes, não seja superior à distância máxima apresentada na matriz de distâncias e no caso dos vértices $i$ e $j$ serem adjacentes estas restrições obrigam a que o comprimento do caminho $u_{ij}$ não seja superior ao peso da aresta que liga o vértice $i$ ao vértice $j$. Estas restrições estabelecem assim um limite superior para as variáveis $u_{ij}$, definidas entre um vértice interno e um vértice externo e entre dois vértices internos.

As restrições (5.6) impõem que o comprimento do caminho $u_{ij}$ que liga um vértice interno a um vértice externo ou que liga dois vértices internos não deve ser inferior a $w_{ij}$ e estabelecem, assim, um limite inferior para as variáveis $u_{ij}$, definidas entre um vértice interno e um vértice externo e entre dois vértices internos.

Quando os vértices $i$ e $j$ são adjacentes as restrições (5.5) e (5.6) obrigam as variáveis $u_{ij}$ a tomarem o mesmo valor do peso da aresta que representam.

Finalmente, as restrições (5.7) impõem que o comprimento do caminho entre dois vértices seja o mesmo independentemente de qual o vértice inicial considerado.

Esta formulação constrói uma árvore filogenética com raiz. Esta raiz fictícia foi acrescentada relativamente ao modelo de fluxos de Catanzaro et al. [4], que determina árvores filogenéticas sem raiz, de modo a facilitar a construção da árvore.

## 5.2 Resultados

Vamos agora analisar o desempenho do modelo usando-o para resolver alguns problemas.

Usamos o software XPRESS para implementar o modelo e para resolver os problemas. Um dos problemas resolvido foi a construção de uma árvore filogenética com cinco taxons cuja matriz de distâncias retirámos de [20]. A matriz de distâncias é a seguinte:

| | $A$ | $B$ | $C$ | $D$ | $E$ |
|---|---|---|---|---|---|
| $A$ | 0 | 22 | 39 | 39 | 41 |
| $B$ | | 0 | 41 | 41 | 43 |
| $C$ | | | 0 | 18 | 20 |
| $D$ | | | | 0 | 10 |
| $E$ | | | | | 0 |

Tabela 5.1: Matriz de distâncias relativa a cinco taxons.

A solução apresentada por Mount está representada na Figura (5.1) e o valor óptimo é 66.

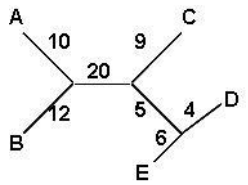


Figura 5.1: Árvore filogenética óptima apresentada por Mount partindo de uma matriz de distâncias com 5 taxons.

O XPRESS demorou 0,7 segundos a encontrar a solução óptima usando a nossa formulação alternativa. A árvore filógenética obtida com valor óptimo de 66 está representada na Figua (5.2).

Ao comparar as duas árvores verificamos que os valores correspondem e que a soma dos pesos das duas arestas que incidem na raiz fictícia da árvore obtida através da nossa

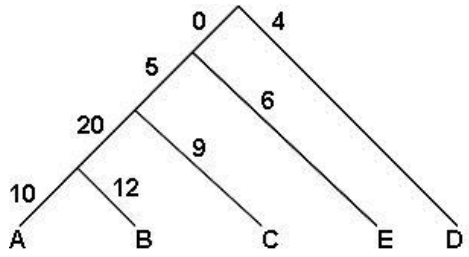


Figura 5.2: Árvore filogenética optida através da formulação alternativa para 5 taxons.

formulação alternativa, é igual ao peso da aresta correspondente da árvore sem raiz. Verificamos, ainda, que a distribuição dos pesos pelas duas arestas incidentes na raiz fictícia é feita de forma a que uma das aresta tenha peso zero. Esta situação aconteceu em todos as árvores construídas usando a nossa formulação alternativa.

A matriz de distâncias para a construção de uma árvore filogenética com sete taxons foi retirada de [24], sendo que o XPRESS demorou 184,4 segundos a encontrar a solução óptima e as conclusões retiradas ao comparar a árvore sem raiz obtida em [24] e a árvore obtida pelo XPRESS usando a formulação alternativa são idênticas às anteriores para a árvore com cinco taxons.

Usamos a matriz de distâncias do ficheiro que Juan-José Salazar-González nos enviou para construir uma árvore filogenética com nove taxons. A matriz de distâncias é a seguinte:

| | $A$ | $B$ | $C$ | $D$ | $E$ | $F$ | $G$ | $H$ | $I$ |
|---|---|---|---|---|---|---|---|---|---|
| $A$ | 0 | 1.51499 | 0.353063 | 0.273829 | 0.418288 | 1.51499 | 1.51499 | 1.51499 | 1.51499 |
| $B$ | | 0 | 0.388258 | 0.280125 | 1.51499 | 0.361509 | 0.215903 | 0.293472 | 0.390438 |
| $C$ | | | 0 | 0.130378 | 0.291302 | 0.253443 | 0.483335 | 0.289859 | 0.229979 |
| $D$ | | | | 0 | 0.258003 | 0.220236 | 0.281086 | 0.368932 | 0.318039 |
| $E$ | | | | | 0 | 0.185762 | 0.181499 | 0.240245 | 0.198209 |
| $F$ | | | | | | 0 | 0.184151 | 0.226651 | 0.198648 |
| $G$ | | | | | | | 0 | 0.0552959 | 0.0521836 |
| $H$ | | | | | | | | 0 | 0.0465828 |
| $I$ | | | | | | | | | 0 |

Tabela 5.2: Matriz de distâncias relativa a nove taxons.

Deparámo-nos com uma série de dificuldades ao tentar encontrar a árvore filogenética óptima usando o XPRESS. Não foi possível encontrar a árvore usando a formulação tal como foi apresentada devido a limitações de memória. Para contornar essa dificuldade, usamos a sugestão apresentada por Catanzaro et al. [4] fixando algumas variáveis. Assim, ao fixar $x_{1A} = x_{12} = x_{23} = 1$, representando 1, 2 e 3 vértices internos e $A$ um vértice externo, o XPRESS obteve a árvore filogenética com valor óptimo 2.01868 passados 3196.8 segundos. Ao fixar apenas $x_{12} = x_{23} = 1$ o XPRESS demorou 97272.3 segundos a obter a árvore filogenética com valor óptimo 2.01868. Verificamos que as topologias de árvore obtidas são um pouco diferente mas apenas pela mudança

de posição de um dos vértices internos, mantendo os taxons a mesma ligação, como podemos verificar nas Figuras (5.3) e (5.4).

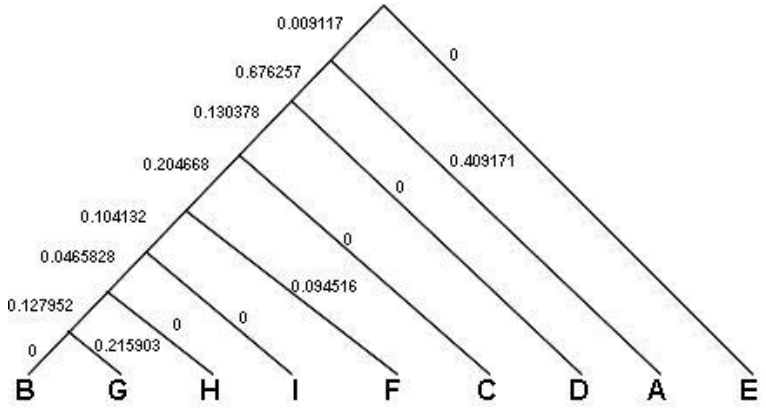


Figura 5.3: Árvore filogenética com 9 taxons obtida ao fixar $x_{1A} = x_{12} = x_{23}$.

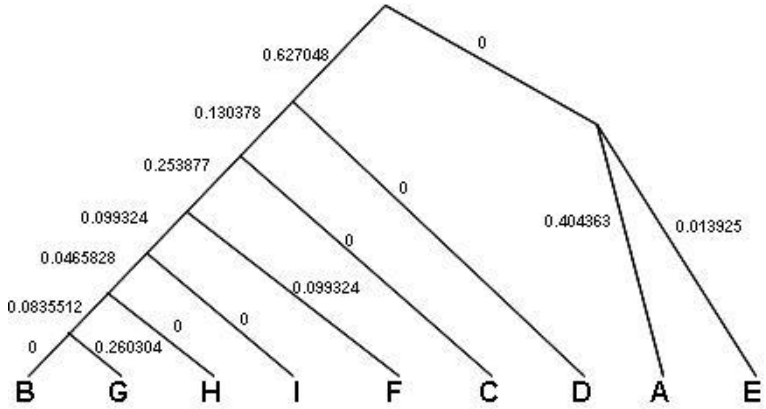


Figura 5.4: Árvore filogenética com 9 taxons obtida ao fixar $x_{12} = x_{23}$.

A diferença nas topologias é apenas de uma aresta e acontece por no primeiro caso termos fixado a variável $x_{1A}$, ou seja, a aresta $e_{1A}$ que liga um vértice interno a um vértice externo e no segundo caso já não a termos fixado. Verificamos, ainda, que além da aresta incidente na raiz ter peso zero aparecem, nas duas árvores, mais cinco arestas com peso zero e que de uma árvore para a outra os pesos não nulos das arestas também são alterados. A diferença de pesos prende-se com o facto de um dos vértices internos ter mudado de posição alterando deste modo a distribuição dos pesos. Os pesos nulos estão associados a arestas que ligam um vértice externo a um vértice interno, incidindo essas arestas com peso nulo nos mesmos vértices externos nas duas árvores.

Devido a limitações de memória e ao elevado tempo computacional não conseguimos obter árvores filogenéticas quando o número de taxons é superior a nove.

## 5.3 Decomposição de Dantzig-Wolfe

Uma vez que o uso da formulação em programação linear se torna incomputável para a obtenção de árvores filogenéticas cujo número de taxons é superior a nove, tentámos usar um outro método para a obtenção de uma solução para a formulação alternativa. Tentamos, assim, aplicar a decomposição de Dantzig-Wolfe à nossa formulação alternativa. O método de decomposição de Dantzig-Wolfe divide o conjunto de restrições em 2 subconjuntos. No nosso caso agrupamos as restrições (5.3), (5.4), (5.5) e (5.6) num subconjunto e as restantes noutro subconjunto. O método opera separadamente em 2 problemas mais pequenos, o problema mestre e um subproblema. Tendo em conta a forma como agrupamos as restrições, o nosso problema mestre é o seguinte:

Problema Mestre

$$min \quad \sum_{t=1}^{T} \left( \sum_{\substack{i \in V_{int} \\ j \in V \\ i<j}} (w_{ij})_t \right) \lambda_t$$

*sujeito a*

$$\sum_{t=1}^{T} (-(u_{jk})_t + (u_{ik})_t + (w_{ij})_t - d(x_{ij})_t)\lambda_t \geq -d \quad \forall i \in V_{int}, \forall j \in V, i < j, \forall k \in V \setminus \{i,j\} \tag{5.10}$$

$$\sum_{t=1}^{T} (-(u_{ik})_t + (u_{jk})_t + (w_{ij})_t - d(x_{ij})_t)\lambda_t \geq -d \quad \forall i \in V_{int}, \forall j \in V, i < j, \forall k \in V \setminus \{i,j\} \tag{5.11}$$

$$\sum_{t=1}^{T} (-(u_{ij})_t + (w_{ij})_t - d(x_{ij})_t)\lambda_t \geq -d \quad \forall i \in V_{int}, \forall j \in V, i < j \tag{5.12}$$

$$\sum_{t=1}^{T} ((u_{ij})_t - (w_{ij})_t)\lambda_t \geq 0 \quad \forall i \in V_{int}, \forall j \in V, i < j \tag{5.13}$$

$$\sum_{t=1}^{T} \lambda_t = 1 \tag{5.14}$$

$$x_{ij} \in \{0,1\} \quad \forall i \in V_{int}, \forall j \in V \tag{5.15}$$

$$w_{ij} \geq 0 \quad \forall i \in V_{int}, \forall j \in V \tag{5.16}$$

$$u_{ij} \geq 0 \quad \forall i \in VT, \forall j \in VT \tag{5.17}$$

$$\lambda_t \geq 0 \quad \forall t \in \{1, ...T\} \tag{5.18}$$

Para obter o subproblema temos de associar variáveis duais às restrições do problema mestre. No nosso caso associamos as variáveis $\alpha, \gamma, \delta, \eta$ e $\beta$ às restrições (5.10), (5.11), (5.12), (5.13) e (5.14), respectivamente, obtendo o seguinte subproblema:

Subproblema

$$
\begin{aligned}
max \quad & \beta \quad + \\
& + \sum_{\substack{i \in V_{int} \\ j \in V, i<j}} \Bigg( \sum_{k \in V \setminus \{i,j\}} -d(\alpha_{kji} + \gamma_{kji}) - d(\delta_{ji}) \Bigg) x_{ij} \quad + \\
& + \sum_{\substack{i \in V_{int} \\ j \in V, i<j}} \Bigg( \sum_{k \in V \setminus \{i,j\}} (\alpha_{kji} + \gamma_{kji}) + (\delta_{ji} - \eta_{ji}) - 1 \Bigg) w_{ij} \quad + \\
& + \sum_{\substack{j \in V \\ k \in V, k \neq j}} \Bigg( \sum_{\substack{i \in V_{int} \\ i<j, i \neq k}} (-\alpha_{kji} + \gamma_{kji}) \Bigg) u_{jk} \quad + \\
& + \sum_{\substack{i \in V_{int} \\ k \in V, k \neq i}} \Bigg( \sum_{\substack{j \in V \\ i<j, j \neq k}} (\alpha_{kji} - \gamma_{kji}) \Bigg) u_{ik} \quad + \\
& + \Bigg( \sum_{\substack{i \in V_{int} \\ j \in V, i<j}} (-\delta_{ji} + \eta_{ji}) \Bigg) u_{ij}
\end{aligned}
$$

*sujeito a*

$$
\begin{aligned}
& x \in SPT && && (5.19) \\
& u_{ij} \geq d_{ij} && \forall i, j \in V_{ext}, i \neq j && (5.20) \\
& u_{ij} = u_{ji} && \forall i \in V_{int}, j \in V, i < j && (5.21) \\
& x_{ij} \in \{0, 1\} && \forall i \in V_{int}, \forall j \in V && (5.22) \\
& w_{ij} \geq 0 && \forall i \in V_{int}, \forall j \in V && (5.23) \\
& u_{ij} \geq 0 && \forall i \in VT, \forall j \in VT && (5.24) \\
& \alpha_{kji} \geq 0 && \forall i \in V_{int}, \forall j \in V, i < j, \forall k \in V \setminus \{i, j\} && (5.25) \\
& \gamma_{kji} \geq 0 && \forall i \in V_{int}, \forall j \in V, i < j, \forall k \in V \setminus \{i, j\} && (5.26) \\
& \delta_{ji} \geq 0 && \forall j \in V, \forall i \in V_{int}, i < j && (5.27) \\
& \eta_{ji} \geq 0 && \forall j \in V, \forall i \in V_{int}, i < j && (5.28)
\end{aligned}
$$

A decomposição de Dantzig-Wolfe parte de uma árvore inicial e através de várias iterações melhora essa solução inicial até chegar à solução óptima. Nas várias iterações

o algoritmo resolve sucessivamente o problema mestre e o subproblema usando a solução encontrada na iteração anterior. A ideia é a seguinte. Dada uma árvore com a forma pretendida determinar valores para as variáveis $w$ e $u$ usando o problema mestre e verificar, resolvendo o subproblema, se esses valores são óptimos. Caso o não sejam, o subproblema sugere uma nova árvore para a qual o problema mestre determina, novamente, os valores para as variáveis $w$ e $u$. Este procedimento continua até serem encontrados os valores óptimos.

A árvore inicial que usamos é a árvore de suporte de custo mínimo em que todos os vértices internos têm grau três e a raiz tem grau dois. Desta forma efectuamos uma pesquisa sobre as árvores que se podem construir com a forma pretendida. Para cada uma delas obtínhamos valores para as variáveis $w$ e $u$, para depois testarmos a optimalidade desta solução. Devido a um erro, ainda, não encontrado até ao momento, não obtivemos resultados. As variáveis $\lambda_t$ tomam todas valor zero no problema mestre e as variáveis duais do problema mestre também.

# Capítulo 6

# Conclusão

Neste trabalho apresentamos o conceito de árvores filogenéticas, descrevendo-o formalmente e indicando os vários tipos de árvores existente bem como a fórmula para determinar o número de árvores filogenéticas possíveis dado o número de espécies em estudo. Estudamos alguns dos métodos usados para a inferência de árvores filogenéticas, nomeadamente, os métodos da máxima parcimónia, os métodos de verossimilhança máxima e alguns métodos de distâncias. Encontrar uma árvore de máxima parcimónia é um problema NP-Difícil e em alguns casos o critério de máxima parcimónia é estatisticamente inconsistente. O problema de determinar uma árvore de verossimilhança máxima também é NP-Difícil mas é estatisticamente consistente. Relativamente aos métodos de distância estudamos o método da média aritmética não ponderada e o método dos mínimos quadrados. Tendo numa primeira fase descrito algumas das formas de determinar as distâncias da matriz assim como algumas propriedades da matriz de distâncias. O método UPGMA pressupõe a existência de um relógio molecular e no caso de a taxa de substituição não ser constante, o que acontece na prática na maioria dos casos, este método obtém maus resultados. O método dos mínimos quadrados já não pressupõe a existência de relógio molecular e por isso também é mais usado.

Apresentamos, também, o problema da evolução mínima. Estudamos o método neighbor-joining e dois modelos de programação linear para resolver o problema. O método neighbor-joining tem a vantagem de ser eficiente mas a desvantagem de apenas explorar uma pequena parte do espaço solução podendo deixar de encontrar a solução global. Os modelos de programação linear apresentados são o modelo de caminhos que identifica a árvore filogenética óptima como sendo a árvore de suporte de custo mínimo que satisfaz mais algumas restrições e o modelo de fluxos que apresenta os pesos não negativos por potenciais desconhecidos de modo a garantir uma distância mínima entre cada par de vértices externos. Ao estudarmos o modelo de fluxos constatamos que este estava incompleto e nesse sentido elaboramos uma formulação alternativa para o problema da evolução mínima. Esta formulação alternativa foi implementada no programa XPRESS. Os resultados obtidos foram bons para um número de espécies

inferior a nove. Ao tentar encontrar a árvore filogenética óptima para um conjunto de nove espécies, deparámo-nos com uma série de dificuldades, nomeadamente, ligadas às limitações de memória e ao elevado tempo computacional. Para ultrapassar essas dificuldades tivemos de fixar algumas variáveis. Para um número de taxons superior a nove é necessário encontrar uma outra forma de construir árvores filogenéticas. Numa fase posterior efectuamos uma abordagem usando a decomposição de Dantzig-Wolfe. Mas até ao momento ainda não obtivemos resultados conclusivos.

A construção de árvores filogenéticas e o problema da evolução mínima continuam a ser problemas difíceis de resolver. Todos os métodos apresentados para a sua resolução continuam a ser estudados e melhorados. Uma vez que muitos dos métodos apresentam árvores filogenéticas diferentes, a construção de uma árvore de consenso também tem sido muito estudada. Nos últimos anos muitos investigadores têm vindo a explorar o uso da computação evolutiva na construção de árvores filogenéticas.

# Referências


[1] W. Beyer, M. Stein, T. Smith, and S. Ulam. A molecular sequence metric and evolutionary trees. *Mathematical Biosciences*, 19:9–25, 1974.

[2] D. Catanzaro. PhD thesis.

[3] D. Catanzaro. The minimum evolution problem: overview and classification. *Wiley Periodicals, Inc. NETWORKS*, 7:228:112–125, 2008.

[4] D. Catanzaro, M. Labbé, R. Pesenti, and J. J. Salazar-González. Mathematical models to reconstruct phylogenetic trees under the minimum evolution criterion. *Wiley Periodicals, Inc. NETWORKS*, 53(2):126–140, 2009.

[5] D. Catanzaro, R. Pesenti, and M. Milinkovitch. An ant colony optimization algorithm for phylogenetic estimation under the minimum evolution principle. *BMC Evolutionary Biology*, 53(2):112–125, 2008.

[6] L. Cavalli-Sforza and A. Edwards. Phylogenetic analysis: models and estimation procedures. *Am. J. Hum. Genet.*, 19:233–257, 1967.

[7] P. Darlu and P. Tassy. *La reconstruction phylogénétique*. Masson, Paris, 1993.

[8] R. Desper and O. Gascuel. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computation Biology*, 9(5):687–705, 2002.

[9] R. Eck and M. Dayhoff. *Atlas of protein sequence and structure*. National Biomedical Research Foundation, Silver Springer, Maryland, 1966.

[10] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[11] J. Felsenstein. Distance methods for inferring phylogenies: a justification. *Evolution*, 38(1):16–24, 1983.

[12] W. Fitch. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, 20(4):406–416, 1971.

[13] W. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

[14] A. Goeffon, J.-M. Richer, and J.-K. Hao. Voisinage d'arbre évolutif appliqué au problème maximum parcimonie. *Communication longue, Actes JOBIM-05*, 2005.

[15] J. Hartigan. Minimum evolution fits to a given tree. *Biometrics*, 29:53–65, 1973.

[16] D. Hillis, C. Moritz, and B. Mable. *Molecular systematics*. Sinauer Associates, Inc., Sunderland, Massachusetts U.S.A., 1996.

[17] V. Hollich, L. Milchert, L. Arvestad, and E. Sonnhammer. Assessment of protein distance measures and tree-building methods for phylogenetic tree reconstruction. *Molecular Biology and Evolution*, 22(11):2257–2264, 2005.

[18] K. Kidd and L. Sgaramella-Zonta. Phylogenetic analysis: concepts and methods. *American Jornal of Human Genetics*, 23:235–252, 1971.

[19] V. Makarenkov and F.-J. Lapointe. A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, 20(13):2113–2121, 2004.

[20] D. Mount. *Bioinformatics. Sequence and genome analysis*. Cold Spring Harbor Laboratory Press, New York, 2001.

[21] M. Nei and S. Kumar. *Molecular evolution and phylogenetics*. Oxford University Press, New York, 2000.

[22] K. Nixon. The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15:407–414, 1999.

[23] W. Pearson, G. Robins, and T. Zhang. Generalized neighbor-joining: more reliable phylogenetic tree reconstruction. *Molecular Biology and Evolution*, 16(6):806–816, 1999.

[24] L. Pinteiro, G. Viana, F. Gomes, and M. Viana. Técnicas algorítmicas para construção de árvores filogenéticas. *FLF.EDU*, 4(1):81–102, 2005.

[25] O. Prado, F. Zuben, and S. Reis. Evolving phylogenetic trees: an alternative to black-box approaches. *Biblioteca Digital Brasileira de Computação*, I Workshop Brasileiro de Bioinformática:56–63, 2002.

[26] A. Rzhetsky and M. Nei. A simple method for estimating and testing minimum-evolution tree. *Molecular Biology and Evolution*, 9(5):945–967, 1992.

[27] A. Rzhetsky and M. Nei. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10(5):1073–1095, 1993.

[28] N. Saitou and M. Nei. Neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.

[29] J. Studier and K. Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5:729–731, 1988.

[30] F. Tajima and M. Nei. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *Journal of Molecular Evolution*, 18:115–120, 1982.

[31] N. Takezaki and M. Nei. Genetic distances and reconstruction of phylogenetic tree from microsatellite DNA. *Genetics*, 144:389–399, 1996.

[32] M. Waterman, T. Smith, M. Singh, and W. Beyer. Additive evolutionary trees. *Journal of Theoretical Biology*, 64:199–213, 1977.