UNIVERSIDADE ESTADUAL DE CAMPINAS
FACULDADE DE ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO
DEPARTAMENTO DE ENGENHARIA DE COMPUTAÇÃO E AUTOMAÇÃO INDUSTRIAL

UMA ABORDAGEM MULTI-OBJETIVO E MULTIMODAL PARA RECONSTRUÇÃO DE ÁRVORES FILOGENÉTICAS

Ana Estela Antunes da Silva

Orientador: Prof. Dr. Fernando José Von Zuben
DCA/FEEC/Unicamp

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos exigidos para a obtenção do título de Doutor em Engenharia Elétrica.

Area de Concentração: Engenharia de Computação

Campinas – São Paulo – Brasil
Dezembro de 2007

Este exemplar corresponde à redação final da tese defendida por: Ana Estela Antunes da Silva
e aprovada pela Comissão
Julgada em 12 / 12 / 2007 Fernando José Von Zuben
Orientador

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

Silva, Ana Estela Antunes da
SI38a Uma abordagem multi-objetivo e multimodal para reconstrução de árvores filogenéticas / Ana Estela Antunes da Silva. --Campinas, SP: [s.n.], 2007.

Orientador: Fernando José VonZuben.
Tese de Doutorado - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Árvores. 2. Filogenia. 3. Otimização matemática. 4. Algoritmos geneticos. 5. Biologia - Processamento de dados. I. VonZuben, Fernando José. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: A multimodal and multiobjective approach for phylogenetic trees reconstruction
Palavras-chave em Inglês: Phylogenetic trees, Multiobjective otimization, Phylogeny reconstruction algorithms, Minimum evolution
Área de concentração: Engenharia de Computação
Titulação: Doutor em Engenharia Elétrica
Banca examinadora: Rafael Santos Mendes , Gilmar Barreto, Maria Emilia Machado Telles Walter, Ricardo José Gabrielli Barreto Campello
Data da defesa: 12/12/2007
Programa de Pós Graduação: Engenharia Elétrica

Banca Examinadora

Orientador: Prof. Dr. Fernando José Von Zuben
(DCA/FEEC/Unicamp)

1º Membro Interno: Prof. Dr. Rafael Santos Mendes
(DCA/FEEC/Unicamp)

2º Membro Interno: Prof. Dr. Gilmar Barreto
(DCA/FEEC/Unicamp)

1º Membro Externo: Profa. Dra. Maria Emilia Machado Telles Walter
(CIC/UnB)

2º Membro Externo: Prof. Dr. Ricardo José Gabrielli Barreto Campello
(ICMC/USP-São Carlos)

Prof. Dr. Fernando José Von Zuben (Presidente): _____

Profa. Dra. Maria Emilia Machado Telles Walter: _____

Prof. Dr. Ricardo José Gabrielli Barreto Campello: _____

Prof. Dr. Rafael Santos Mendes: _____

Prof. Dr. Gilmar Barreto: _____

Resumo

A reconstrução de árvores filogenéticas pode ser interpretada como um processo sistemático de proposição de uma descrição arbórea para as diferenças relativas que se observam em conjuntos de atributos genéticos homólogos de espécies sob comparação. A árvore filogenética resultante apresenta uma certa topologia, ou padrão de ancestralidade, e os comprimentos dos ramos desta árvore são indicativos do número de mudanças evolutivas desde a divergência do ancestral comum. Tanto a topologia quanto os comprimentos de ramos são hipóteses descritivas de eventos não-observáveis e condicionais, razão pela qual tendem a existir diversas hipóteses de alta qualidade para a reconstrução, assim como múltiplos critérios de desempenho. Esta tese (*i*) aborda árvores sem raiz; (*ii*) enfatiza os critérios de quadrados mínimos, evolução mínima e máxima verossimilhança; (*iii*) propõe uma extensão ao algoritmo *Neighbor Joining* que oferece múltiplas hipóteses de alta qualidade para a reconstrução; e (*iv*) descreve e utiliza uma nova ferramenta para otimização multiobjetivo no contexto de reconstrução filogenética. São considerados dados artificiais e dados reais na apresentação de resultados, os quais apontam vantagens e aspectos diferenciais das metodologias propostas.

Abstract

The reconstruction of phylogenetic trees can be interpreted as a systematic process of proposing an arborean description to the relative dissimilarities observed among sets of homologous genetic attributes of species being compared. The resulting phylogenetic tree presents a certain topology, or ancestrality pattern, and the length of the edges of the tree will indicate the number of evolutionary changes since the divergence from the common ancestor. Both topology and edge lengths are descriptive hypotheses of non-observable and conditional events, which implies the existence of diverse high-quality hypotheses for the reconstruction, as long as multiple performance criteria. This thesis (*i*) deals with unrooted trees; (*ii*) emphasizes the least squares, minimum evolution, and maximum likelihood criteria; (*iii*) proposes an extension to the Neighbor Joining algorithm which offers multiple high-quality reconstruction hypotheses; and (*iv*) describes and uses a new tool for multiobjective optimization in the context of phylogenetic reconstruction. Artificial and real datasets are considered in the presentation of results, which points to some advantages and distinctive aspects of the proposed methodologies.

Dedicatória

Ao Bruno, Terezinha e João

Agradecimentos

Agradeço, primeiramente, ao meu orientador Fernando José Von Zuben pela orientação, ensinamentos e amizade. Sinto que meus horizontes de pesquisa foram em muito expandidos pelas nossas conversas. Espero levar em meu trabalho o respeito à ciência que tive a chance de presenciar durante todo o tempo dessa orientação.

Agradeço a todos do LBiC, em especial ao Guilherme, pela generosidade e extraordinária ajuda.

Agradeço à Universidade Estadual de Campinas pela oportunidade de desenvolvimento desse trabalho.

Agradeço à professora Cláudia Maria Bauzer Medeiros pelo apoio oferecido no início de meu doutorado.

Agradeço à Universidade Metodista de Piracicaba e à Fundação Municipal de Ensino pelas horas cedidas de meu trabalho ao desenvolvimento dessa tese.

Agradeço à minha amiga Fran pelo incentivo nas horas de desânimo e também à minha amiga Nilza por ter me ajudado a descobrir forças que eu não imaginava que tinha e por ser fonte de inspiração, de força e solidariedade.

Agradeço à minha família pelo apoio de todas as horas.

Finalmente, agradeço ao meu filho, por aceitar a minha ausência em tantos momentos de sua infância.

ÍNDICE

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução e Motivação

Resumo – Este capítulo introduz os principais conceitos tratados nesse trabalho, voltados para a reconstrução de árvores filogenéticas empregando abordagens multimodais e de otimização multi-objetivo. Apresenta também o escopo da pesquisa e a organização do texto.

1.1 Introdução

As áreas de pesquisa em bioinformática e biologia computacional empregam os recursos de processamento e memória do computador na busca de solução para problemas de biologia, desde o nível molecular até o estudo de nichos ecológicos. São adotadas técnicas e metodologias de matemática aplicada, de estatística, de ciência da computação e de bioquímica, dentre outras.

Embora os termos bioinformática e biologia computacional possam ser usados como sinônimos, existem algumas distinções que geralmente são feitas. A bioinformática é mais propriamente associada à criação de algoritmos e procedimentos estatísticos para análise de dados, como no caso do tratamento de dados de expressão gênica e no processamento de imagens biomédicas. Já a biologia computacional está diretamente vinculada ao uso do computador na validação de hipóteses e na modelagem e simulação de fenômenos biológicos, também a partir de dados experimentais. Como exemplos de biologia computacional pode-se mencionar a modelagem de dinâmica populacional e a obtenção de índices morfométricos em botânica.

O *National Institute of Health* dos Estados Unidos propôs a seguinte distinção:

> Bioinformática: pesquisa, desenvolvimento ou aplicação de ferramentas computacionais e metodologias para a expansão do uso de dados biológicos, médicos, comportamentais e de saúde, incluindo ferramentas computacionais para adquirir, armazenar, organizar, analisar ou visualizar tais dados.
>
> Biologia computacional: desenvolvimento e aplicação de métodos teóricos, técnicas analíticas baseadas em hipóteses ou em dados experimentais, procedimentos de modelagem matemática e simulação computacional para o estudo de sistemas biológicos, comportamentais e sociais.

Com base nas distinções acima enunciadas, esta tese propõe contribuições junto a um dos mais desafiadores problemas na área de bioinformática: a reconstrução de árvores filogenéticas. Trata-se de um processo de inferência que visa representar as relações evolutivas existentes entre as espécies vivas, a partir de atributos comparáveis que caracterizam os indivíduos em análise. Essa inferência se dá pelo uso de algoritmos computacionais que agrupam as espécies sob a forma de uma árvore filogenética, partindo da hipótese de que todos os seres vivos descendem de um ancestral comum e que cada espécie com atributos conhecidos ocupará uma folha na árvore filogenética. Embora não seja a intenção deste trabalho de tese, quando todas as espécies vivas são consideradas, obtém-se uma proposta para a árvore da vida (ALURU, 2006).

1.2 O problema de reconstrução de árvores filogenéticas

A árvore filogenética pode ser com ou sem raiz. A existência de raiz implica na definição da seqüência em que se dá o processo de diferenciação evolutiva (FELSENSTEIN, 2004). Os procedimentos de reconstrução geralmente são específicos para o caso com raiz ou para o caso sem raiz. Serão abordados neste trabalho procedimentos que operam com árvores sem raiz.

A reconstrução de árvores filogenéticas confiáveis envolve um esforço significativo em razão das dificuldades inerentes ao processo de aquisição de dados biológicos e, posteriormente, à complexidade computacional associada ao problema de reconstrução (HOLMES, 1999). Ao longo de todo o trabalho, será considerado que o processo de aquisição de dados biológicos foi realizado a priori, permitindo abordar diretamente o problema de reconstrução da árvore a partir das diferenças verificadas nas folhas. As diferenças nas folhas podem estar expressas apenas na forma de uma matriz de distâncias par-a-par entre as espécies, ou então na forma de seqüências de atributos homólogos associadas a cada espécie.

A complexidade computacional do procedimento de reconstrução filogenética e a dificuldade de se chegar a um resultado inquestionável se devem a quatro fatores principais, os quais estão interligados:

- ✓ Ausência de informação referente aos ancestrais comuns (estão disponíveis apenas informações referentes às folhas da árvore) e limitação de informação referente às folhas;
- ✓ Existência de múltiplos objetivos que devem ser atendidos simultaneamente;
- ✓ Mesmo quando se considera apenas um objetivo a ser otimizado, podem existir múltiplas soluções e ausência de definição em certas regiões da árvore (ausência de informação indicativa da topologia preferencial, segundo o critério de otimização e as hipóteses evolutivas);
- ✓ Crescimento fatorial do número de topologias de árvores candidatas com o aumento do número de folhas.

A necessidade da proposição de novos algoritmos e aperfeiçoamento dos algoritmos já existentes para reconstrução filogenética têm proporcionado desafios expressivos à comunidade de computação. O quarto fator mencionado acima é, sem dúvida, o maior desafio computacional de um procedimento de reconstrução. Para um número n de espécies (folhas da árvore), existe o seguinte número de árvores com raiz (topologias candidatas a explicar a diferença observada nas folhas):

$$\frac{(2n-3)!}{2^{n-2} * (n-2)!},$$

e o seguinte número de árvores sem raiz:

$$\frac{(2n-5)!}{2^{n-3} * (n-3)!}.$$

Essa é a razão pela qual os problemas de inferência filogenética foram provados serem NP-completos (ROCH, 2006), não sendo conhecido um algoritmo capaz de obter a árvore ótima, segundo algum critério, em tempo polinomial.

1.3 Procedimentos para reconstrução de árvores filogenéticas

Dada a ausência de um algoritmo de reconstrução capaz de operar em tempo polinomial, conforme o número *n* de folhas cresce, as únicas alternativas possíveis são os procedimentos de reconstrução construtivos, como o *Neighbor-Joining* (SAITOU & NEI,1987; ATTESON, 1996), e os procedimentos de reconstrução empíricos, que adotam heurísticas capazes de explorar um espaço de busca, como os algoritmos baseados em máxima parcimônia e verossimilhança máxima (SAITOU & IMANISHI,1989; SOURDIS & NEI, 1988; TAKAHASHI & NEI, 2000).

O espaço de busca é uma concepção matemática que associa a cada ponto (elemento do espaço) uma proposta de topologia para a árvore filogenética. Quando se trabalha apenas com as relações de ancestralidade (que definem a conformação estrutural da árvore), sem se preocupar com o comprimento dos ramos, este espaço de busca é discreto. Por outro lado, quando, além da conformação estrutural da árvore, deve-se definir o comprimento dos ramos, o espaço de busca é contínuo. Um tratamento formal para o espaço das árvores filogenéticas foi realizado em BILLERA *et al.* (2001).

Os procedimentos empíricos geralmente adotam passos aleatórios em sua execução, o que permite obter soluções diferentes a cada execução, e não há como definir a priori o tempo até a convergência e o nível de qualidade da melhor solução a ser encontrada numa dada

execução do algoritmo correspondente. Mesmo com essas limitações, na ausência de algoritmos tratáveis que sejam exatos ou que garantam atingir um nível previamente especificado de qualidade para a solução, os procedimentos construtivos e empíricos se apresentam como as únicas possibilidades para tratar o problema, o que tem sido realizado com expressivo sucesso na literatura (HOLDER & LEWIS, 2003).

O Capítulo 3 da tese irá tratar em detalhes esses dois procedimentos de reconstrução, os quais podem ser sucintamente definidos como segue:

- ✓ Procedimento construtivo: é aquele que parte de uma árvore-estrelae gradualmente promove junções de nós-folha, implicando na inserção de nós-hipotéticos que vão indicar a existência de um ancestral comum. Com isso, a conformação estrutural da árvore é definida incrementalmente, passo a passo. O procedimento pode ou não envolver a definição conjunta dos comprimentos de ramos que já foram estabelecidos.
- ✓ Procedimento empírico: é aquele que possui um índice de avaliação que é capaz de "atribuir uma nota" a qualquer proposta de árvore, ou seja, a qualquer ponto do espaço de busca. Sendo assim, deve-se explorar o espaço de busca visando localizar regiões promissoras, as quais são aquelas que contêm propostas de árvores que recebem notas altas segundo o índice de avaliação. Também aqui pode-se trabalhar apenas com a conformação estrutural da árvore, ou envolver a definição de todos os comprimentos de ramos.

1.4 Múltiplos objetivos para árvores filogenéticas

Quando se busca uma árvore filogenética, há uma diversidade de critérios que podem ser adotados visando selecionar as árvores que melhor explicam as diferenças presentes nas folhas. Basicamente, existem os critérios não baseados em modelo evolutivo e aqueles baseados em modelo evolutivo, conforme discriminados a seguir:

- ✓ Critérios não baseados em modelo evolutivo:
    - o Evolução mínima: visa minimizar o somatório do comprimento dos ramos da árvore filogenética;

- Quadrados mínimos: visa minimizar o somatório do quadrado da diferença elemento-a-elemento entre a matriz de distâncias original (extraída das comparações par-a-par entre as folhas) e aquela obtida a partir da árvore proposta (pelo somatório dos comprimentos de ramos para se deslocar de cada folha a todas as demais folhas);
- Máxima parcimônia: visa minimizar a quantidade de eventos de diferenciação necessários para produzir a diferença observada nas folhas com base em uma dada topologia de árvore.

✓ Critérios baseados em modelo evolutivo:

- Verossimilhança máxima: visa maximizar a probabilidade de se chegar àquela diferença nas folhas a partir de uma dada proposta de árvore filogenética. Para tanto, deve-se partir de hipóteses que indiquem com que probabilidade os eventos evolutivos podem ocorrer, ou seja, deve-se partir de um modelo evolutivo.

Dependendo do critério adotado, percebe-se, então, que a árvore filogenética resultante pode ser diferente. Na verdade, o problema de reconstrução de uma árvore filogenética é apenas mais um dentre os inúmeros problemas que a ciência e o progresso tecnológico enfrentam e que envolvem múltiplos objetivos, muitas vezes conflitantes entre si. Há três abordagens possíveis frente a problemas multi-objetivo:

✓ Concentrar-se em um único objetivo, negligenciando os demais, e obter a solução com base em uma formulação mono-objetivo para o problema.

✓ Levar em consideração um subconjunto de objetivos simultaneamente, ponderados em uma única função-objetivo. Tem-se, assim, um problema mono-objetivo que procura compor a importância relativa de cada objetivo em uma função matemática. Como será visto no Capítulo 6, mesmo que se tenha uma noção prévia da importância relativa de cada objetivo, ela não pode ser expressa em coeficientes de uma combinação linear, ou seja, se um objetivo é duas vezes mais relevante que um outro objetivo, multiplicar por 2 a participação do primeiro num critério de custo não implica que ele será atendido com o dobro de intensidade quando comparado ao

outro. Tudo depende de fatores de escala no espaço dos objetivos, os quais não são conhecidos previamente.

- ✓ Tratar o problema em sua forma original, ou seja, como um problema multi-objetivo em que se buscam soluções não-dominadas, no sentido de não ser possível melhorar o atendimento de qualquer objetivo sem piorar algum outro. Resolver o problema então equivale a amostrar o lugar geométrico das soluções não-dominadas. Caberá ao usuário escolher a posteriori uma dessas soluções, por análise comparativa e levando-se em conta critérios adicionais que não foram empregados ao longo da busca.

Nesta tese, iremos tratar os três tipos de abordagem, sendo que as principais contribuições encontram-se presentes junto à última abordagem.

1.5 Múltiplas soluções para árvores filogenéticas

Assim como os problemas multi-objetivos em reconstrução de árvores filogenéticas requerem a obtenção de múltiplas propostas de árvore como solução, existem formulações mono-objetivo que também podem se beneficiar da obtenção de múltiplas soluções. Particularmente, essas múltiplas soluções são altamente desejáveis em problemas multi-modais, ou seja, problemas que apresentam múltiplos ótimos locais.

Um caso a ser explorado nesta tese está associado à aplicação de uma metodologia construtiva, mais especificamente o algoritmo *Neighbor-Joining* (NJ) (SAITOU & NEI, 1987). É sabido que o NJ leva em consideração dois objetivos em uma formulação mono-objetivo (segunda abordagem apresentada na seção 1.4) e que sua estratégia gulosa de reconstrução, a ser melhor detalhada no Capítulo 3, não garante que o ótimo global seja encontrado, pois o processo de reconstrução pode convergir para um ótimo local.

Nessas circunstâncias, atuar durante o processo de reconstrução visando convergir para múltiplos ótimos locais, em paralelo, se mostra como um procedimento válido, por três motivações principais:

- ✓ O custo computacional do procedimento construtivo é baixo;
- ✓ Há um aumento das chances de se encontrar um ótimo global;

- ✓ Múltiplos ótimos locais de qualidade próxima ou igual àquela do ótimo global podem ser encontrados.

O Capítulo 5 irá tratar diretamente das propostas e resultados obtidos junto a este tópico da pesquisa.

1.6 Múltiplas soluções e técnicas de consenso

São dois os cenários em que a solução para o procedimento de reconstrução de árvores filogenéticas está associada à proposição de múltiplas topologias de árvores: abordagem multi-modal e abordagem multi-objetivo.

Frente à existência de múltiplas propostas de árvores, geralmente adota-se um de dois procedimentos:

- ✓ Obtém-se uma única árvore que atenda a algum critério de consenso entre as diversas propostas de árvore;
- ✓ Apresenta-se ao usuário as múltiplas propostas de solução, permitindo que este realize uma análise comparativa, por exemplo, visual.

Ambos serão considerados nesta tese, sendo que as principais técnicas de consenso são tratadas no Capítulo 4.

1.7 Organização do Texto

A metodologia utilizada para a realização desse trabalho está refletida na divisão e seqüência de apresentação dos capítulos da tese. O capítulo corrente, Introdução, trata da motivação para o desenvolvimento do trabalho, seus objetivos e principais conceitos envolvidos.

O segundo capítulo, Filogenia, apresenta os principais conceitos relacionados à reconstrução de árvores filogenéticas e apresenta os principais modelos evolutivos utilizados pelos algoritmos de filogenia.

O terceiro capítulo, Métodos de reconstrução de árvores filogenéticas, apresenta dois algoritmos de reconstrução. O primeiro deles, o de verossimilhança máxima, representa o algoritmo que utiliza como dados de entrada bases nucleotídicas, e, dada uma árvore candidata, realiza uma busca no espaço de topologias candidatas pela árvore que mais se aproxima dos critérios de otimização. O segundo algoritmo é o *Neighbor-Joining* (NJ), o qual tem como característica a apresentação de uma única árvore filogenética como solução. Essa árvore, a qual é construída de acordo com os critérios de evolução mínima e quadrados mínimos, é gerada a partir de uma matriz de distâncias entre as espécies.

O quarto capítulo, Consenso e medidas de distância, apresenta alguns métodos que podem ser levados em consideração para a escolha de uma solução quando muitas soluções são apresentadas. Esses métodos são: técnicas de consenso e medidas de distância.

O *MultiNeighbor-Joining*, uma das principais contribuições dessa tese, é apresentado no quinto capítulo. Esse capítulo apresenta: a descrição do algoritmo, exemplos de sua execução e um estudo de caso que comprova a adequação das soluções ao problema apresentado.

No sexto capítulo, os principais conceitos de otimização multi-objetivo são apresentados, seguidos da descrição detalhada da abordagem multi-objetivo, baseada em uma meta-heurística denominada *omni-aiNet*, a qual foi desenvolvida junto ao grupo de pesquisa em que se insere esta tese. A *omni-aiNet* sofreu, neste trabalho, uma série de extensões necessárias para permitir sua aplicação na reconstrução de árvores filogenéticas.

No sétimo capítulo são apresentados e comparados os resultados obtidos com a utilização dos algoritmos *Neighbor-Joining*, *Multi-Neighbor Joining* e *omni-aiNet*. Há também uma comparação com a solução obtida via verossimilhança máxima.

Por fim, o oitavo capítulo apresenta os comentários conclusivos e aponta as perspectivas futuras da pesquisa.

Capítulo 2

Filogenia

Resumo – Este capítulo apresenta o conceito de filogenia, exemplifica os principais tipos de árvores filogenéticas, define a representação de árvores filogenéticas utilizada nesse trabalho, a notação Newick, e cita os principais modelos evolutivos utilizados em algoritmos de reconstrução de árvores filogenéticas.

2.1 Introdução

A filogenia ou história evolutiva das espécies está fundamentada em um conceito da Teoria da Evolução que afirma que grupos com organismos que apresentam atributos similares descendem de um ancestral comum.

A Teoria da Evolução foi proposta por Darwin em seu livro Origem das Espécies (DARWIN, 1859). A idéia principal da Teoria é a de que todos os seres vivos têm um determinado grau de parentesco. Há evidências de que toda vida existente na Terra é descendente de um único ancestral comum. Durante um período de 3,8 bilhões de anos, esse ancestral comum diferenciou-se, de forma cumulativa, em novas e independentes espécies (LINDER & WARNOW, 2006). Essa divisão deu origem à diversidade observada atualmente, sendo que há muito poucos vestígios históricos desse processo de diferenciação, particularmente no caso de formas de vida já extintas.

A representação dessa diversidade por meio de uma árvore surgiu no próprio livro Origem das Espécies, no qual a única ilustração é a primeira representação de relações evolutivas entre espécies sob a forma de uma árvore filogenética (DELSUC *et al.*, 2005), conforme Figura 2.1.

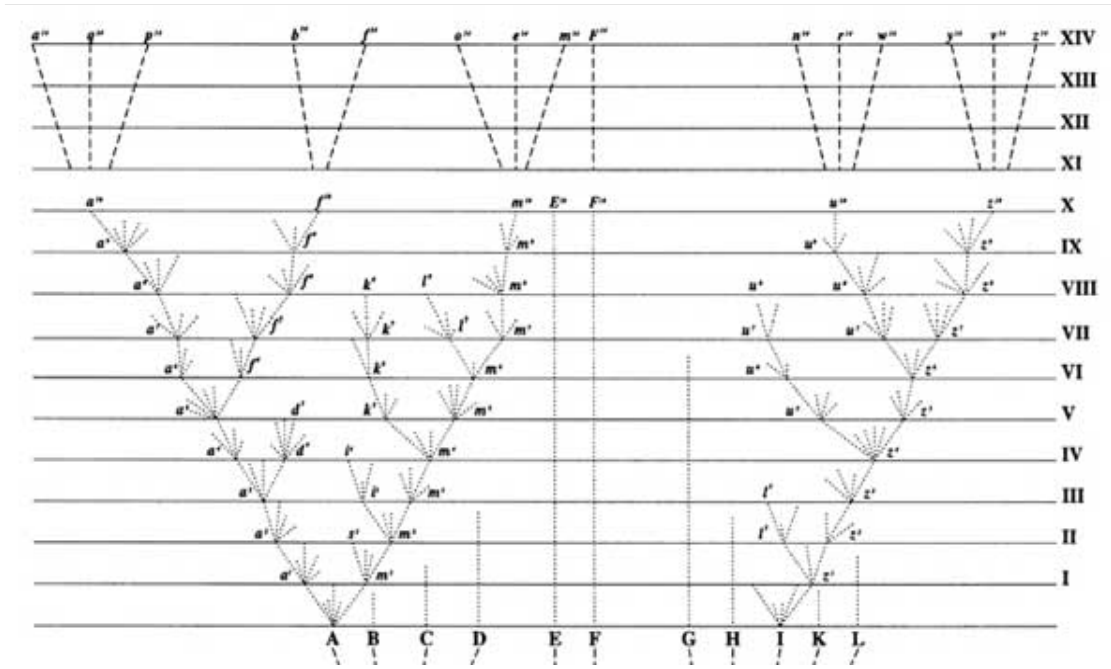


Figura 2.1 - Diagrama apresentado por Charles Darwin em Origem das Espécies (VAN WYHE, 2006)

2.2 Árvore Filogenética

As relações evolutivas são representadas na forma de árvores filogenéticas, que descrevem os relacionamentos entre as espécies. São construídas com base em dados de caracteres, sendo um caractere algum atributo relevante de um organismo, que pode assumir diferentes estados. Um exemplo biológico típico de um caractere é uma posição de um nucleotídeo em uma seqüência de DNA. O estado do caractere é o próprio nucleotídeo (A,C,G,T). A Figura 2.2 ilustra uma árvore filogenética de seqüências de DNA.

No exemplo hipotético da Figura 2.2, a primeira espécie representada (raiz) teria surgido há três intervalos de tempo. Em um período de um intervalo de tempo (possivelmente milhões de anos), essa espécie evoluiu para as duas espécies descendentes. Em um outro período de um intervalo de tempo, essas duas espécies teriam evoluído para duas espécies cada uma, surgindo quatro novas espécies. Novamente, em outro período de um intervalo de tempo teríamos, atualmente, oito diferentes espécies.

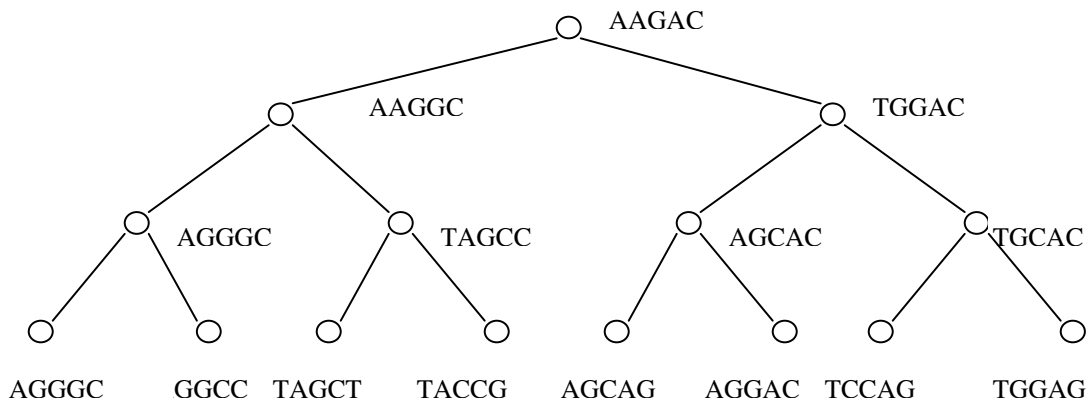


Figura 2.2 - Evolução hipotética de seqüências de DNA

2.3 Tipos de Árvores

Uma árvore filogenética pode ser classificada de várias formas (FELSENSTEIN, 2004). As duas maneiras mais comuns de classificação são:

- com e sem raiz; e
- nós internos com dois descendentes (*bifurcating*) e nós internos com vários descendentes (*multifurcating*).

Uma árvore com raiz é aquela que possui um único ancestral comum para todas as espécies. Esse ancestral comum é chamado de raiz da árvore. Conseqüentemente, uma árvore sem raiz é aquela que não possui um ancestral comum, pois não é indicada a direção em que a diferenciação ocorre.

Uma árvore com nós internos com dois descendentes (*bifurcating*) é aquela em que cada nó interior tem grau 3 (conecta-se a três outros nós) e cada folha tem grau um. Uma árvore com nós internos com número variado de descendentes (*multifurcating*) é aquela que pode ter alguns nós internos com graus maiores do que 3. A Figura 2.3 apresenta uma árvore com raiz e nós internos de grau 3 e a Figura 2.4 apresenta a mesma árvore, porém, sem o nó raiz.

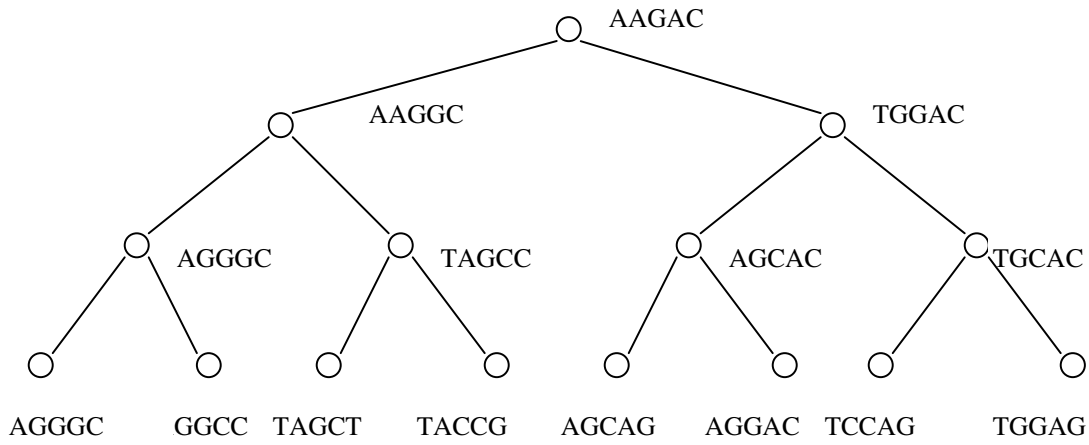


Figura 2.3 - Árvore com raiz e com nós interiores de ordem 3

A filogenia é geralmente representada por árvores com nós internos de ordem 3, uma vez que os eventos de divisão de uma espécie são tomados como eventos de bifurcação, ou seja, uma divisão usualmente ocorre quando uma espécie ancestral transforma-se em duas novas espécies independentes, conforme Figuras 2.3 e 2.4.

As Figuras 2.3 e 2.4 apresentam a árvore filogenética representada como um grafo cujos nós são as espécies e as arestas representam o tempo de evolução entre as espécies, caso tenham valores atribuídos a elas.

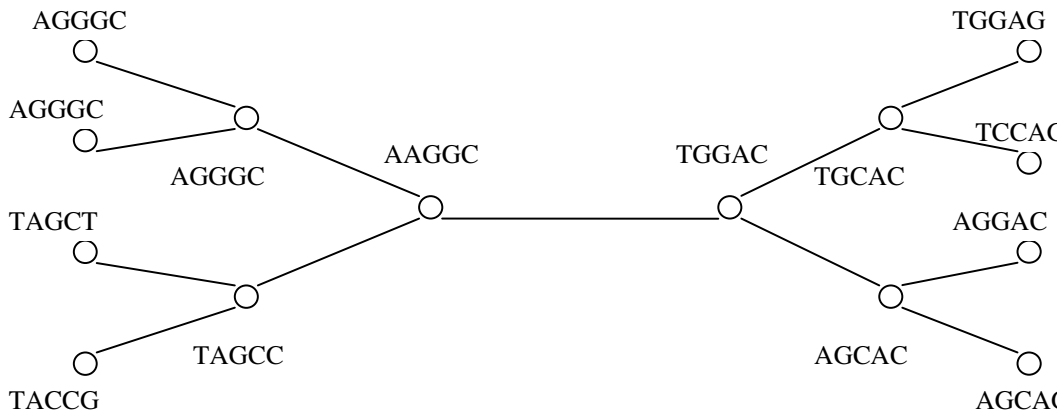


Figura 2.4 - Árvore sem raiz e com nós interiores de grau 3

Árvores com arestas sem valores atribuídos indicam apenas a relação de descendência entre as espécies. Uma outra maneira de representação de uma árvore filogenética é dada pela notação Newick, a qual se aplica tanto para o caso de arestas com valores como sem valores atribuídos. A notação Newick é apresentada na seção 2.2.3.

2.4 Metodologias para reconstrução de árvores filogenéticas

Existe uma classificação para os métodos de reconstrução de árvores filogenéticas: métodos fenéticos (ou não-baseados em modelo evolutivo) e métodos cladísticos. Os métodos fenéticos são aqueles que consideram o estado corrente das seqüências de atributos, não importando a história evolutiva, ou seja, a dinâmica dos passos intermediários. A árvore que melhor explica os relacionamentos entre as seqüências de atributos é denominada fenograma. A Figura 2.5 apresenta um fenograma construído a partir das distâncias entre as espécies A, B e C, extraídas diretamente de suas seqüências de atributos. O fenograma pode ser com ou sem raiz e a árvore resultante pode ser ultramétrica ou não. Uma árvore ultramétrica é aquela em que todas as folhas apresentam a mesma distância à raiz.

Na Figura 2.5, se as distâncias entre as espécies forem consideradas, a árvore retrata apenas aproximadamente as relações de similaridade presentes na matriz de distâncias, pois ela foi tomada como sendo ultramétrica.

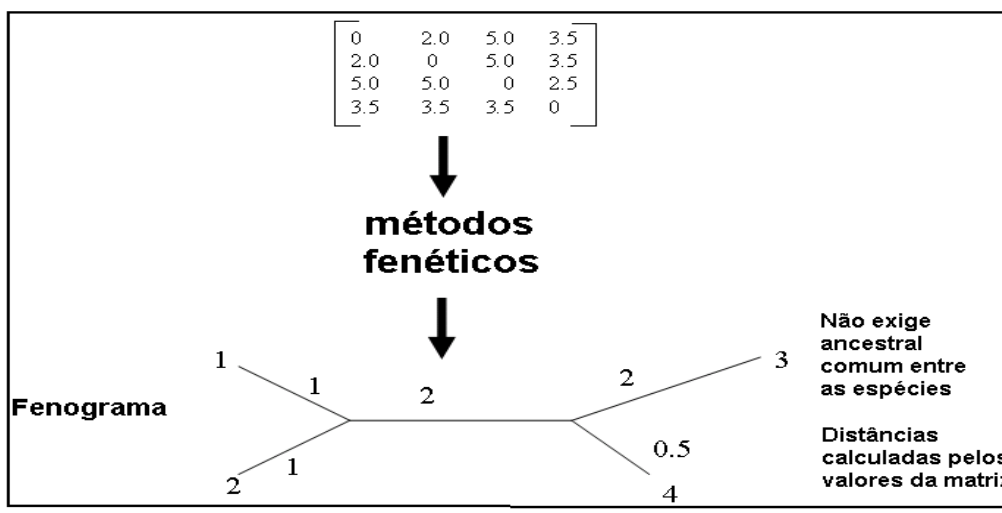


Figura 2.5 - Fenograma: árvore filogenética baseada em relações de similaridade

Já os métodos cladísticos ou baseados em modelo evolutivo são aqueles que levam em conta as possibilidades de resultado de um processo evolutivo, considerando a dinâmica dos passos intermediários. Tais métodos adotam a árvore que melhor explica os relacionamentos entre as seqüências de atributos resultantes, sempre com base em uma hipótese evolutiva. Esta hipótese evolutiva pode estar baseada em algum modelo evolutivo ou em algum critério de otimalidade. A árvore que melhor explica os relacionamentos entre as seqüências de atributos é denominada cladograma. Um exemplo de cladograma é apresentado na Figura 2.6.

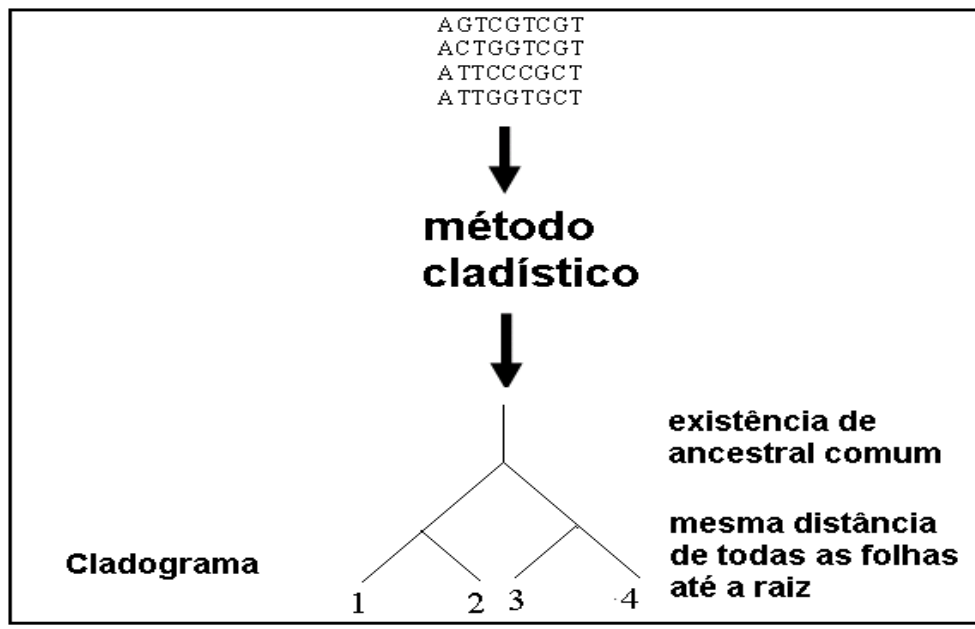


Figura 2.6 - Cladograma: árvore filogenética baseada em relações de ancestralidade comum

Pode-se distinguir fenogramas de cladogramas pelos métodos que os originam. Métodos não baseados em modelos evolutivos geram fenogramas e métodos baseados em modelos evolutivos geram cladogramas.

Quando relações de ancestralidade são consideradas e os comprimentos de ramos são levados em conta, ou seja, os comprimentos de ramos são informativos, resulta um filograma, conforme enfatizado na Figura 2.7.

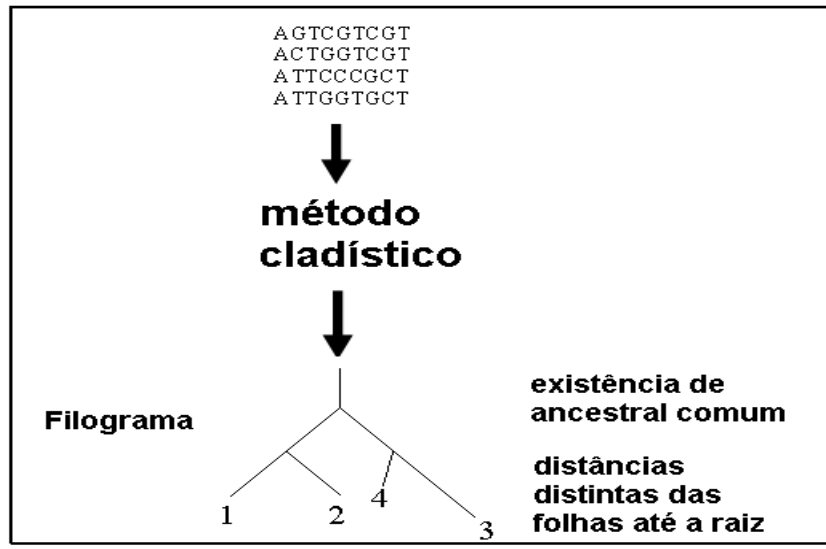


Figura 2.7 - Filograma

2.4.1 Modelos Evolutivos de Substituição de Bases em Seqüências de Nucleotídeos

A hipótese evolutiva na qual se baseiam os métodos cladísticos pode ser representada por meio de algum modelo evolutivo os quais são modelos matemáticos que tentam representar o processo de substituição de bases nucleotídicas, ou o processo de substituição de aminoácidos, entre diferentes espécies (LIÒ & GOLDMAN, 1998).

O primeiro modelo surgiu em 1965 com ZUCKERKANDL & PAULING (1965). A teoria chamada de relógio molecular afirma que a taxa de evolução molecular é constante ao longo da linha do tempo. De acordo com esta teoria, qualquer diferença entre espécies durante certo intervalo de tempo pode ser calculada como sendo proporcional à soma do número de mudanças entre as seqüências de DNA ou proteínas.

Outro modelo que serve de base para a explicação do processo evolutivo é o modelo de Markov (KOSIOL & GOLDMAN, 2005; LIÒ & GOLDMAN, 1998). O modelo de Markov pode ser utilizado para seqüências evolutivas de DNA ou aminoácidos. Para a utilização do modelo, a evolução de cada sítio é considerada ser independente da evolução de todos os demais na cadeia.

O modelo de Markov pode conter três propriedades importantes: homogeneidade, estacionariedade e reversibilidade. Homogeneidade significa que a matriz de taxas de substituição é independente do tempo, ou seja, os padrões de substituição de nucleotídeos (ou aminoácidos) permanecem os mesmos durante todo o processo evolutivo. A matriz **Q** da Figura 2.8 representa a matriz de substituição de bases de nucleotídeos. Os elementos da matriz representam o valor da taxa de substituição de uma base por outra. Por exemplo , r_{AC} indica a taxa de substituição da base A pela base C em uma cadeia. Para o caso da utilização desses modelos para a reconstrução de árvores filogenéticas, pode-se dizer que os padrões de substituição de nucleotídeos permanecem os mesmos em diferentes partes da árvore.

$$\mathbf{Q} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} r_{AA} & r_{AC} & r_{AG} & r_{AT} \\ r_{CA} & r_{CC} & r_{CG} & r_{CT} \\ r_{GA} & r_{GC} & r_{GG} & r_{GT} \\ r_{TA} & r_{TC} & r_{TG} & r_{TT} \end{bmatrix} \end{array}$$

Figura 2.8 - Matriz Q de substituição de bases de nucleotídeos

Estacionariedade significa que o processo evolutivo está em equilíbrio, ou seja, as freqüências de bases de nucleotídeos permanecem as mesmas durante o processo de evolução. As freqüências de bases de nucleotídeos são representadas por π_A, π_C, π_G e π_T, para as bases A, C, G e T, respectivamente.

Reversibilidade significa que, para a matriz da Figura 2.8, $\mathbf{Q}_{ij} = \mathbf{Q}_{ji}$, para todo i,j. A conseqüência da reversibilidade é que o processo de evolução de seqüências é, teoricamente, indistinguível do mesmo processo observado de forma reversa.

A fim de incorporar as propriedades acima descritas, a matriz de taxas de substituição pode ser apresentada conforme mostrado na Figura 2.9. Na matriz F, as freqüências são representadas por π_A, π_C, π_G e π_T . As variáveis v_1, v_2, v_3 e v_4 representam as taxas das quatro transversões e s_1 e s_2 as taxas das duas transições. As transições são as substituições

entre as bases A e G, e C e T, as quais são representadas pelas letras *s*. As tranversões são as substituições ocorridas entre as bases: A e C, A e T, C E G e G e T, representadas pela letra *v*.

$$\mathbf{F} = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{bmatrix} r_{AA} & \pi_A.v_1 & \pi_A.s_1 & \pi_A.v_2 \\ \pi_C.v_1 & r_{CC} & \pi_C.v_3 & \pi_C.s_2 \\ \pi_G.s_1 & \pi_G.v_3 & r_{GG} & \pi_G.v_4 \\ \pi_T.v_2 & \pi_T.s_2 & \pi_T.v_4 & r_{TT} \end{bmatrix} \end{array}$$

Figura 2.9 - Matriz F de taxas com freqüências de equilíbrio e taxas de transversão e transição

Um esquema representando as transições e transversões é mostrado na Figura 2.10. Em casos reais de evolução de seqüências, transições são observadas mais freqüentemente do que transversões.

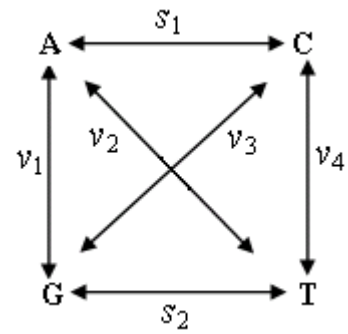


Figura 2.10 - Esquema de transições e transversões

Há vários modelos que tentam representar a evolução de seqüências de nucleotídeos. Todos eles são baseados no modelo de Markov e, portanto, consideram uma ou mais de suas propriedades.

2.4.1.1 Jukes-Cantor

O mais simples dos modelos é o Jukes-Cantor (FELSENSTEIN, 2004; HUELSENBECK & CRANDALL, 1997). O modelo, conhecido como JC69 considera que todas as substituições são igualmente prováveis e que todos os nucleotídeos ocorrem com igual freqüência em

uma seqüência. Para o modelo, $\pi_A = \pi_C = \pi_G = \pi_T$ e $v_1 = v_2 = v_3 = v_4 = s_1 = s_2$. Como a soma de todas as freqüências deve ter valor 1, o valor da freqüência é 0,25 para qualquer uma das quatro bases. Como todos os outros parâmetros são iguais, o modelo possui apenas um parâmetro.

2.4.1.2 Kimura

Outro modelo é o Kimura-dois-parâmetros conhecido como K2P. Para esse modelo, $\pi_A = \pi_C = \pi_G = \pi_T$, $v_1 = v_2 = v_3 = v_4$ e $s_1 = s_2$. Novamente, a freqüência de cada base tem valor 0,25. O modelo, conforme o nome indica, tem dois parâmetros: um para transições e outro para transversões.

2.4.1.3 Felsenstein

Um último modelo a ser mencionado aqui é o de Felsenstein, conhecido como F84. Diferentemente dos dois modelos acima, esse modelo não considera a igualdade de freqüências de bases. As quatro freqüências são calculadas de acordo com a porcentagem com que cada freqüência aparece na seqüência de nucleotídeos. Dessa forma, cada exemplo real tem suas próprias freqüências. As taxas de transição e transversão são consideradas equivalentes, fazendo com que o modelo tenha cinco parâmetros.

Uma descrição do número de parâmetros e quais são os parâmetros utilizados pela maioria dos modelos evolutivos pode ser encontrada em SASB (2006).

2.5 Representação de árvores filogenéticas pela notação Newick

A notação Newick representa árvores filogenéticas na forma de uma lista de atributos, facilitando a sua manipulação computacional. Essa notação é baseada na correspondência existente entre uma árvore e uma lista de parênteses aninhados. Seja a árvore com raiz apresentada pela Figura 2.11. A notação Newick correspondente a essa árvore seria a seguinte seqüência de caracteres entre parênteses: (B,(A,C,E),D); (PHYLIP, 2006).

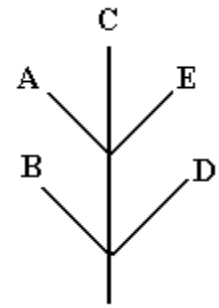


Figura 2.11 - Árvore filogenética com raiz e cinco espécies

A árvore sempre termina com um ponto-e-vírgula. Nós interiores são representados por um par de parênteses. Dentro desse par de parênteses, ficam os nós que são imediatamente descendentes desse nó, separados por vírgulas. No exemplo da Figura 2.11, os descendentes imediatos são: B, um outro nó interior e D. O nó interior é novamente representado por um par de parênteses, incluindo as representações de seus descendentes imediatos: A, C e E.

Folhas são representadas pelos seus próprios nomes. Um nome pode ser, por exemplo, uma seqüência de bases nucleotídicas.

A notação Newick também permite a inclusão dos comprimentos dos ramos da árvore. Estes podem ser incorporados à árvore por meio da inclusão de um número real colocado depois de um nó e precedido pelo símbolo de dois-pontos. Esse número representa o comprimento do ramo imediatamente abaixo desse nó. A árvore da Figura 2.11 poderia ter a notação Newick alterada para: (B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0); , o que implica na inclusão de valores para as arestas, também denominados comprimentos dos ramos da árvore filogenética.

A notação Newick não corresponde a uma única representação para uma árvore. As duas representações: (A,(B,C),D); e (A,(C,B),D); representam a mesma árvore. Além disso, a notação está representando uma árvore com raiz. Considerando propósitos biológicos, pode não ser possível ou necessário inferir a posição da raiz na árvore. Nesses casos, há a intenção de buscar uma representação de uma árvore sem raiz. Seja a seguinte árvore com raiz: (B,(A,D),C);. Essa árvore é a mesma árvore sem raiz dada por: ((A,D),(C,B));.

2.5.1 Representação formal da notação Newick

A notação Newick é baseada em uma representação na forma de uma lista composta por sub-árvores, nós e comprimentos de arestas. A sua sintaxe é composta por: parênteses, vírgulas, dois-pontos e ponto-e-vírgula. Esses símbolos são agrupados de maneira a refletir a topologia e o comprimento dos ramos da árvore.

Convenções:

●Itens em { } podem aparecer nenhuma ou várias vezes.

●Itens em [ ] são opcionais, eles podem aparecer uma vez ou nenhuma vez.

●Todos os símbolos de pontuação: dois-pontos, ponto-e-vírgula, parênteses, vírgula e apóstrofe são partes da notação.

A sintaxe da notação é apresentada a seguir por meio de regras de produção que descrevem a gramática de representação de uma árvore sob a forma Newick. Uma regra é definida sob a forma LE -> LD, sendo LE um não-terminal e LD um terminal e/ou um não-terminal. Um símbolo não-terminal pode ser substituído por um ou mais símbolos.

Sintaxe:

árvore →lista_de_descendentes [rótulo_da_raiz] [:comprimento_do_ramo] ;
lista_de_descendentes → (sub-árvore { , sub-árvore})
sub-árvore → lista_de_descendentes [rótulo_nó_interno] [:comprimento_do_ramo]
→ rótulo_da_folha [: comprimento_do_ramo]
rótulo_da_raiz → rótulo
rótulo_do_nó_interno → rótulo
rótulo_da_folha → rótulo
rótulo → rótulo_sem_apóstrofes
→ rótulo_com_apóstrofes
rótulo_sem_apóstrofes → cadeia_de_caracteres
rótulo_com_apóstrofes → ´cadeia_de_caracteres´
comprimento_de_ramos → número_com_sinal
→ número_sem_sinal

A Figura 2.12 mostra um exemplo de uma topologia referente à representação Newick para uma árvore sem raiz, com cinco espécies, três sub-árvores e seus respectivos comprimentos de ramos.

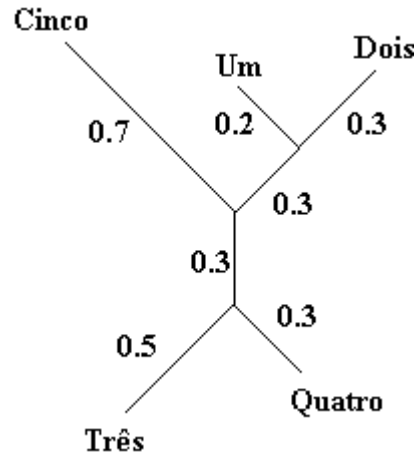


Figura 2.12 - Topologia correspondente à notação ((Um:0.2,Dois:0.3):0.3, ((Três:0.5, Quatro:0.3):0.3,Cinco:0.7):0.0)

Capítulo 3

Métodos de Reconstrução de Árvores Filogenéticas

Resumo – Este capítulo descreve dois tipos de métodos relacionados à reconstrução de árvores filogenéticas. O primeiro deles, representado pelo método da verossimilhança máxima, consiste em um algoritmo de busca. A busca pela árvore mais verossímil se dá em um espaço de topologias de árvores filogenéticas, sendo que a árvore mais verossímil é aquela que melhor explica as diferenças observadas nas folhas, dado um modelo evolutivo. O segundo tipo de método é caracterizado por algoritmos para reconstrução iterativa de árvores filogenéticas. Esses algoritmos possuem como entrada dados observados e estruturados em matrizes de distâncias. Esse trabalho descreve três métodos que utilizam uma matriz de distâncias como entrada: Método dos Quadrados Mínimos, Método da Evolução Mínima e o *Neighbor Joining*. Os passos do *Neighbor-Joining* são devidamente detalhados.

3.1 Introdução

No capítulo anterior, a classificação tradicional de métodos de reconstrução de árvores em métodos fenéticos e cladísticos foi apresentada. Nesse capítulo, uma outra classificação dos métodos de reconstrução de árvores filogenéticas é apresentada: métodos de busca em espaços de topologias de árvores e métodos de reconstrução de árvores.

3.2 Métodos de busca em espaços de topologias: Método da Verossimilhança Máxima

Alguns dos métodos de reconstrução de árvores filogenéticas requerem a implementação de uma busca no espaço de topologias, visando localizar a árvore que maximiza um dado critério de desempenho. Segundo o critério, a árvore buscada é aquela que melhor explica a diferença observada nas folhas. Esse é o caso do algoritmo da verossimilhança máxima,

sendo sempre possível obter a verossimilhança de cada proposta de topologia que se apresente.

3.2.1 O conceito de verossimilhança

A verossimilhança é uma medida proporcional à probabilidade dos dados observados uma vez definidos:

- a topologia da árvore;
- os comprimentos de ramos para esta topologia;
- o modelo evolutivo (também denominado de modelo de substituição).

A verossimilhança, portanto, não se refere à probabilidade de que a árvore seja correta. A menos de um fator de escala α, pode-se afirmar então que a verossimilhança de uma árvore filogenética (também chamada de modelo), a partir dos dados observados, $L(M/D)$, é igual à probabilidade dos dados, a partir da árvore, $\Pr(D/M)$, o que leva à seguinte expressão:

$$L(M/D) = \alpha \Pr(D/M) \tag{3.1}$$

sendo M o modelo evolutivo utilizado e D os dados observados.

O fator de escala α é irrelevante, pois as árvores filogenéticas que levam aos valores máximos de $L(M/D)$ também levam aos valores máximos de $\Pr(D/M)$, independentemente da variação em α.

3.2.2 Cálculo da Verossimilhança Máxima

Alguns pesquisadores apresentam o processo de reconstrução de árvores filogenéticas como um problema de inferência estatística (HUELSENBECK & CRANDALL, 1997). O método de verossimilhança máxima se enquadra nesta descrição, uma vez que usa modelos de inferência estatística para avaliar a capacidade explicativa da árvore filogenética. Esse método foi descrito pelo estatístico inglês R.A. Fisher (FISHER, 1922). Mais tarde, a idéia de utilizá-lo para inferência filogenética foi apresentada por CAVALLI-SFORZA & EDWARDS

(1967). Nessa época, porém, um algoritmo computacional para executar o modelo era extremamente custoso, em razão da grande demanda computacional do método e do desempenho dos processadores da época.

Uma mesma topologia de árvore filogenética pode apresentar verossimilhanças distintas para conjuntos distintos de comprimentos de ramos e também para modelos evolutivos distintos. Mesmo fixando-se o modelo evolutivo, o problema de estimação da árvore de verossimilhança máxima não é apenas um problema de estimação de parâmetros (comprimento dos ramos), mas também um problema de estimação de topologia. Fixar o modelo evolutivo implica definir um modelo de probabilidade para a seqüência de eventos evolutivos que supostamente levaram à diferenciação observada nas folhas.

O método da verossimilhança máxima aceita como parâmetros de entrada seqüências de bases nucleotídicas, ou seja, uma especificação completa dos atributos das folhas. Diferentemente dos métodos que têm como parâmetros de entrada uma matriz de distâncias (como o *Neighbor-Joining*, por exemplo), esse método trabalha diretamente com os caracteres biológicos observados e não com uma estimativa de distância entre eles.

A seqüência de passos a serem seguidos por um algoritmo de verossimilhança máxima é apresentada a seguir:

1. Parte-se das seqüências de bases nucleotídicas associadas às folhas da árvore;
2. Define-se um modelo evolutivo (modelo de substituição);
3. Busca-se uma proposta de topologia no espaço das topologias possíveis;
4. Para cada proposta de topologia, define-se o melhor conjunto de comprimentos de ramos de modo a obter a verossimilhança máxima dessa topologia específica;
5. Caso o critério de parada não tenha sido atendido, retorna-se ao passo 3.

O critério de parada pode ser um número máximo de iterações ou então a conclusão da busca em uma vizinhança da proposta de topologia inicial (FELSENSTEIN, 2004). A única forma de garantir a obtenção da árvore de verossimilhança máxima é testar todas as

topologias existentes no espaço de busca e ficar com aquela que conduz à maior verossimilhança no passo 4. Isso implica que os passos 3 e 4 acima deveriam ser executados um número de vezes igual à cardinalidade do espaço de busca. O problema é que esta tarefa é inviável na prática, pois já foi visto que o número de diferentes topologias cresce fatorialmente com o número de folhas da árvore. Conclui-se então que deve-se optar por um procedimento de busca que não garante a obtenção da árvore mais verossímil, mas que explore convenientemente o espaço de busca com recursos computacionais razoáveis. Existem meta-heurísticas para se realizar esta tarefa com base em busca exploratória (REEVES, 1993), assim como métodos contrutivos gulosos (LEVITIN, 2003). O passo 4 acima também é muito custoso computacionalmente, mas FELSENSTEIN (1981) apresentou uma formulação matemática que reduziu bastante o número de operações algébricas necessárias para se chegar à medida de verossimilhança de uma árvore com topologia, comprimento de ramos e modelo evolutivo já definidos.

3.2.3 Um exemplo de lançamento de moeda

O modelo no qual se baseia o método da verossimilhança máxima pode ser exemplificado por meio de um estudo de caso probabilístico simples (HUELSENBECK & CRANDALL, 1997). O exemplo consiste do experimento de se lançar uma moeda para cima e observar o resultado. Dois resultados são possíveis: Cara(C) ou Coroa (K).

Considerando-se as premissas (*i*) independência entre os experimentos e (*ii*) obtenção de probabilidades constantes, pode-se atribuir à probabilidade desconhecida de Cara a quantidade p, e à probabilidade desconhecida de Coroa a quantidade $1-p$. O que se sabe, é que o valor de p varia no intervalo [0,1].

Suponha que a moeda seja lançada onze vezes e que o seguinte resultado CCKKCKCCKKK (cinco caras e seis coroas) seja observado. A pergunta base do método da verossimilhança máxima é a seguinte: Qual é o valor de p que maximiza a probabilidade de ocorrência do resultado CCKKCKCCKKK? O que se tem nesse modelo é um resultado conhecido e uma probabilidade desconhecida.

Levando-se em conta as considerações feitas acima, o modelo probabilístico desse experimento pode ser dado por:

$$p \times p \times (1-p) \times (1-p) \times p \times (1-p) \times p \times p \times (1-p) \times (1-p) \times (1-p)$$

ou seja, $p^5 \times (1-p)^6$. Dessa forma, existe um resultado conhecido CCKKCKCCKKK e um modelo probabilístico que o descreve: $p^5 \times (1-p)^6$.

Para se poder responder à pergunta base do método da verossimilhança máxima, calcula-se a chance de se observar esse resultado para diferentes valores da quantidade p no intervalo [0,1], conforme mostra a Tabela 3.1. O valor de p no intervalo [0,1] para o qual se tem a maior chance de se observar o resultado CCKKCKCCKKK é ~0,4545. Esse mesmo valor pode ser obtido analiticamente, obtendo-se o ponto de máximo da superfície apresentada pelo gráfico da Figura 3.1. Para isso, faz-se a derivada primeira da função igual a zero, ou seja, $df(p)/dp = 0$. Seja então, $f(p) = p^5 \times (1-p)^6$. Para a condição de máximo, resulta $p \cong 0{,}4545$. Este valor da quantidade p é a que proporciona maior chance de se observar a seqüência CCKKCKCCKKK no experimento.

Tabela 3.1 - Valores de *p* e respectivos valores assumidos pelo modelo probabilístico

Valor de p	$p^5 \times (1-p)^6$
p=0,0	0,0000000
p=0,1	0,0000053
p=0,2	0,0000839
p=0,3	0,0002859
p=0,4	0,0004778
p=0,5	0,0004883
p=0,6	0,0003185
p=0,7	0,0001225
p=0,8	0,0000210
p=0,9	0,0000006
p=1,0	0,0000000

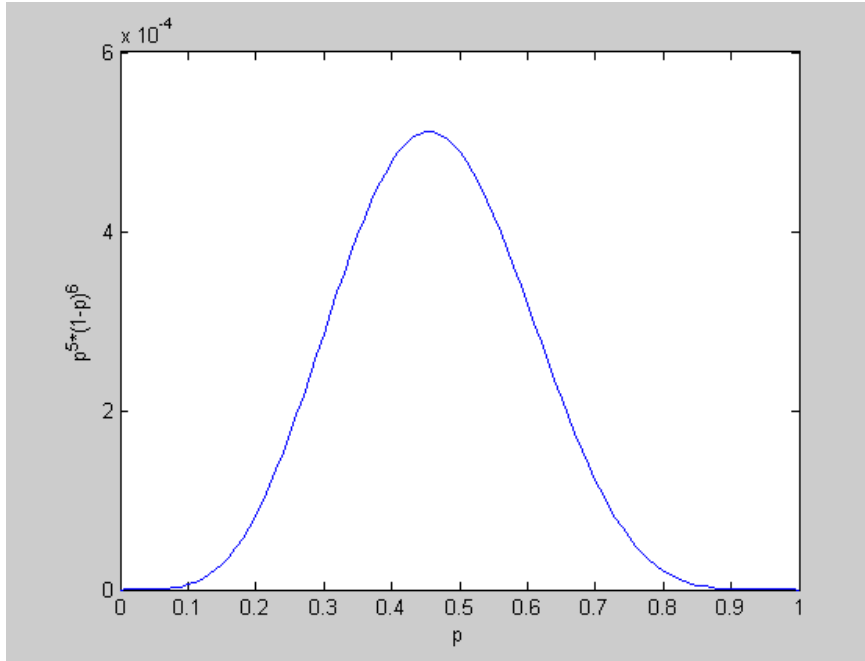


Figura 3.1 - Valor da verossimilhança para o experimento de lançamento da moeda 11 vezes

Conforme mostrado pelo gráfico da Figura 3.1, existe um valor de *p* no intervalo [0,1] para o qual a chance de se observar o resultado CCKKCKCCKKK é máxima.

Para o modelo probabilístico, a quantidade *p* é um parâmetro de entrada e a seqüência CCKKCKCCKKK é um dado (resultado) observado. Este modelo baseia-se num problema de estimação. Para o exemplo dado, estimou-se um valor de *p* adequado. O problema de estimação consiste em definir maneiras de utilizar observações amostrais para construir boas estimativas de um ou mais parâmetros.

3.2.4 Componentes do método

O método da verossimilhança máxima possui dois elementos: dados e modelo probabilístico Os dados são os resultados observados para um determinado experimento e o modelo probabilístico é sempre fornecido.

Os dados de entrada para o método consistem de observações $x_1,...,x_n$ cuja distribuição tem função de probabilidade dada por $f(X, \Theta_1,...,\Theta_k)$. A função f tem forma conhecida e

depende de parâmetros desconhecidos $\Theta_1,...,\Theta_k$. O conjunto de valores admissíveis do parâmetro é denominado Espaço Paramétrico. Para obter-se estimativas de um parâmetro são utilizados os dados disponíveis e uma função que opera sobre esses dados. Essa função é um mapeamento do espaço de dados para o espaço de valores admissíveis do parâmetro. Dessa forma, um estimador é uma função no espaço das observações $(x_1,...,x_n)$ com valores no espaço paramétrico $(\Theta_1,...,\Theta_k)$ e pode-se classificar o método da verossimilhança máxima como um método de estimação. A verossimilhança L (do inglês, *Likelihood*) é proporcional à probabilidade dos dados, D, dada uma hipótese H: $L \propto P(D \mid H)$.

Voltando ao exemplo do lançamento da moeda, podemos concluir que os dados são representados pelo resultado CCKKCKCCKKK, o modelo probabilístico é dado por $f(p) = p^5 \times (1-p)^6$ e as hipóteses atribuíram valores no intervalo [0,1] para o parâmetro p de maneira a encontrar o valor de p que maximizasse a verossimilhança do modelo.

Essa estimativa de verossimilhança máxima do parâmetro Θ (ou p, no caso do exemplo) à uma dada amostra observada é obtida encontrando o valor de Θ que maximiza a função de verossimilhança. Seja a função $L(\Theta)$ essa função de verossimilhança. A função pode ser definida como:

$$L(\Theta) = f(x_1,\Theta) \times f(x_2,\Theta) \times ... \times f(x_n,\Theta)$$

sendo x_1, x_2,..., x_n uma amostra aleatória de uma família de distribuições dada por $f(X, \Theta)$, e Θ um parâmetro desconhecido.

Para obter o valor de Θ que maximiza a função $L(\Theta)$, é necessário obter a derivada primeira da função e igualar a zero, produzindo:

$$\frac{\partial L(\Theta)}{\partial \Theta} = 0$$

No caso de lançamento de moedas, a função $L(\Theta)$ pode ser representada como uma distribuição binomial na forma:

$$L(\Theta) = \Theta^{C} \times (1-\Theta)^{N-C}$$

sendo:

- Θ: parâmetro de otimização para a função L;
- C: valor observado de caras;
- N: número de vezes que a moeda é lançada.

3.2.5 Representação em bases de nucleotídeos

Os dados geralmente considerados para problemas de filogenia são os padrões individuais de seqüências homólogas de nucleotídeos. Por exemplo, considere as seqüências de DNA de quatro indivíduos ou espécies ou unidades taxonômicas (taxa) apresentadas na Figura 3.2. As linhas dessa matriz são preenchidas pelos taxa 1, 2, 3 e 4, que representam seqüências de DNA para as espécies 1, 2, 3 e 4. As colunas representam os sítios, ou seja, a base de nucleotídeo para determinada posição homóloga entre as espécies.

O método da verossimilhança máxima possui três elementos: dados, modelo probabilístico e hipóteses estatísticas. Os dados são os resultados observados para um determinado experimento e o modelo probabilístico é sempre fornecido.

	1	2	3	4	5	6
Taxon 1	A	C	C	A	G	C
Taxon 2	A	A	C	A	G	C
Taxon 3	A	A	C	A	T	T
Taxon 4	A	A	C	A	T	C

Figura 3.2 - Seqüências de DNA para quatro espécies e seis sítios (adaptada de HUELSENBECK & CRANDALL, 1997)

Para essas seqüências, os sítios podem ser descritos pelos vetores $\mathbf{x}_1$ = {A,A,A,A}, $\mathbf{x}_2$ = {C,A,A,A}, $\mathbf{x}_3$ = {C,C,C,C}, $\mathbf{x}_4$ = {A,A,A,A}, $\mathbf{x}_5$ = {G,G,T,T} e $\mathbf{x}_6$ = {C,C,T,C}. Há um total de 4^e sítios possíveis, tomando e taxa.

O número de vezes que cada padrão distinto de sítio (de acordo com sua seqüência de nucleotídeos) aparece em um conjunto de taxa também é considerado relevante em análise de filogenia. Por exemplo, considere a matriz da Figura 3.3. Ela pode ser descrita como

uma matriz de 4 linhas por 256 colunas, totalizando o número de sítios da matriz ($4^4 = 256$). Os números que aparecem na última linha indicam quantas vezes o respectivo padrão de seqüência de bases aparece nos sítios da Figura 3.2.

	1	2	3	4	5	6	7	8	65	86	94	176	256
	A	A	A	A	A	A	A	A	...C...	C...	C	...G ...	T
	A	A	A	A	A	A	A	A	...A...	C...	C	...G ...	T
	A	A	A	A	C	C	C	C	...A...	C...	T	... T ...	T
	A	C	G	T	A	C	G	T	...A...	C...	C	... T ...	T
número de sítios	2	0	0	0	0	0	0	0	0... 1...	1...	1	... 1 ...	0

Figura 3.3 - Matriz completa com todos os sítios possíveis e o número de vezes que cada sítio é observado na Figura 3.2

Como é possível observar pelas Figuras 3.2 e 3.3, a maioria dos sítios não foi observada, com exceção dos sítios de números:1, 65, 86, 94 e 176. O número de vezes que cada padrão de sítio é observado também pode ser considerado como dado de entrada e pode ser representado por um vetor **n** que corresponde à última linha na figura 3.3, produzindo $n_1 = 2$, $n_{65} = 1$, $n_{86} = 1$, $n_{94} = 1$, $n_{176} = 1$ e todos os outros n_i iguais a zero($i = 1, ..., 256$).

É possível representar sítios e espécies utilizando uma matriz *mxn* sendo *n* o número de sítios e *m* o número de espécies. Essa matriz **X** é mostrada na Figura 3.4.

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} \ldots & X_{1h} \ldots & X_{1n} \\ X_{21} & X_{22} \ldots & X_{2h} \ldots & X_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ X_{s1} & X_{s2} \ldots & X_{sh} \ldots & X_{sn} \\ \ldots & \ldots & \ldots & \ldots \\ X_{m1} & X_{m2} \ldots & X_{mh} \ldots & X_{mn} \end{bmatrix} \begin{matrix} \\ \rightarrow \text{espécie } 1 \\ \rightarrow \text{espécie } 2 \\ \\ \rightarrow \text{espécie } s \\ \\ \rightarrow \text{espécie } m \end{matrix}$$

Figura 3.4 - Matriz de seqüências moleculares (aminoácidos ou bases) de acordo com o número de sítios e espécies

3.2.6 Função verossimilhança em filogenia

Conceitualmente, a função verossimilhança em filogenia segue os mesmos passos do exemplo do lançamento de uma moeda (FELSENSTEIN, 2004; HUELSENBECK & CRANDALL, 1997). Os dados são, novamente, interpretados como variáveis randômicas. A diferença é que, ao invés de duas possibilidades de resultados para cada sítio, agora existem 4^e possibilidades, sendo e o número de espécies. Para o exemplo das seqüências de DNA, a função pode ter uma distribuição multinomial, ao invés de uma função binomial. A distribuição multinomial é uma generalização da binomial e tem a forma:

$$L(n_1, n_2, ..., n_s \mid p_1, p_2, ..., p_s) = \prod_{i=1}^{s} p_i^{n_i} \qquad (3.2)$$

onde

- n_i: número de observações do i-ésimo padrão de sítio;
- p_i: probabilidade do sítio i ocorrer; e
- s: número de padrões de sítios possíveis, dado por 4^e;
- e: número de espécies ou taxa.

Uma estimativa para p_i, i=1,...,s, é dada por $p_i = {}^{n_i}\!/_{n}$, onde n é o número de sítios.

A topologia da árvore e os comprimentos dos ramos devem fazer parte da hipótese a ser verificada em um modelo de verossimilhança para filogenia. A Figura 3.5 abaixo mostra os componentes da função verossimilhança, também denominada de L, para filogenia.

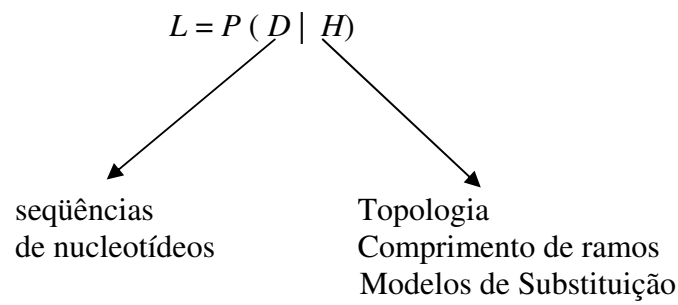


Figura 3.5 - Componentes da função *L* para filogenia

3.2.6.1 Verossimilhança da árvore mais simples

A árvore mais simples de ser construída é aquela com um único nó e sem nenhum ramo. A verossimilhança para tal árvore seria a probabilidade dos dados observados para esse nó. Para ilustrar esse caso, podemos considerar o exemplo da Figura 3.6 (LEWIS, 2002).

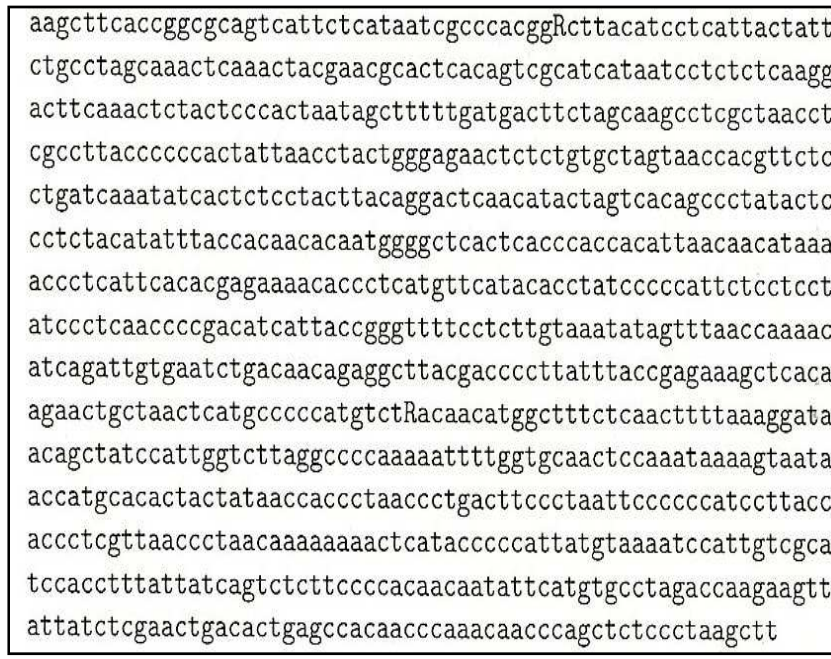

aagcttcaccggcgcagtcattctcataatcgcccacggRcttacatcctcattactatt
ctgcctagcaaactcaaactacgaacgcactcacagtcgcatcataatcctctctcaagg
acttcaaactctactcccactaatagctttttgatgacttctagcaagcctcgctaacct
cgccttacccccccactattaacctactgggagaactctctgtgctagtaaccacgttctc
ctgatcaaatatcactctcctacttacaggactcaacatactagtcacagccctatactc
cctctacatatttaccacaacacaatggggctcactcacccaccacattaacaacataaa
accctcattcacacgagaaaacaccctcatgttcatacacctatcccccattctcctcct
atccctcaaccccgacatcattaccgggttttcctcttgtaaatatagtttaaccaaaac
atcagattgtgaatctgacaacagaggcttacgacccctttatttaccgagaaagctcaca
agaactgctaactcatgcccccatgtctRacaacatggctttctcaacttttaaaggata
acagctatccattggtcttaggccccaaaaattttggtgcaactccaaataaaagtaata
accatgcacactactataaccaccctaaccctgacttccctaattccccccatccttacc
accctcgttaaccctaacaaaaaaaactcataccccccattatgtaaaatccattgtcgca
tccacctttattatcagtctcttccccacaacaatattcatgtgcctagaccaagaagtt
attatctcgaactgacactgagccacaacccaaacaacccagctctccctaagctt

Figura 3.6 - Seqüência de ψη-globina de seres humanos (LEWIS, 2002)

A Figura 3.6 mostra 894 sítios sendo: dois sítios ambíguos (R), 272 bases A, 297 bases C, 95 bases G e 230 bases T. Considerando o modelo de substituição de Jukes-Cantor, todas as bases teriam a mesma freqüência e a verossimilhança seria simplesmente: $(0,25)^{894}$. A probabilidade de se observar qualquer base em qualquer sítio é 0,25 e, assim, a verossimilhança total é o produto da mesma probabilidade multiplicada 894 vezes, visto que a verossimilhança da árvore é dada pelo produto das verossilhanças de cada sítio

Usando outros modelos de substituição, como F81, HKY ou F84, diferentes freqüências poderiam ser utilizadas, conforme indicado abaixo:

$$L = (\pi_A)^{nA}(\pi_C)^{nC}(\pi_G)^{nG}(\pi_T)^{nT} = 0{,}30425^{272} \times 0{,}33221^{297} \times 0{,}10626^{95} \times 0{,}25727^{230}$$

Em razão do número resultante ser muito pequeno, utiliza-se o logaritmo natural de *L*. Com isso, no caso acima a função verossimilhança assume o valor:

$$\ln L = -11{,}7617675.$$

Aplicando o logaritmo natural à verossimilhança calculada com o modelo JC, resulta:

$$\ln L = 894 \times \ln(0{,}25) = -1239{,}34715,$$

a qual é muito mais baixa do que a verossimilhança calculada utilizando-se modelos com freqüências diferentes. Para esse exemplo, modelos com freqüências diferentes seriam mais adequados do que o modelo JC.

3.2.6.2 Verossimilhança de árvores com dois nós e um ramo

Sejam consideradas agora árvores com dois nós e um ramo. Sejam as duas seqüências em cada nó pertencentes às espécies de humanos e de chimpanzés de acordo com a Figura 3.7.

Humano
GTAAATATAGTTTAACCAAAACATCAGATTGTGAATCTGACAACAGAGGCTTACGACCCCTTATTTACC
Chimpanzé
GTAAATATAGTTTAACCAAAACATCAGATTGTGAATCTGACAACAGAGGCTCACGACCCCTTATTTACC

Figura 3.7 - Seqüências homólogas de humanos e chimpanzés, com 69 sítios

Para o modelo JC, as probabilidades de substituição junto a cada sítio são dadas por (LEWIS, 2002; NEI & KUMAR, 2000):

$$P_{ii}(\alpha t) = \Pr(i \text{ em chimpanzé} \mid i \text{ em humano}) = \tfrac{1}{4}\,(1 + 3e^{-4\alpha t}) \text{ e}$$

$$P_{ij}(\alpha t) = \Pr(j \text{ em chimpanzé} \mid i \text{ em humano}) = \tfrac{1}{4}\,(1 - e^{-4\alpha t}) \text{ , sendo } i \neq j.$$

Essas probabilidades são obtidas a partir das seguintes hipóteses:

- A probabilidade de um nucleotídeo mudar para um dos três outros nucleotídeos num tempo *t* é dado por αt, o qual é um parâmetro a ser determinado;
- Quando o tempo *t* cresce, o número de mudanças verificadas entre duas seqüências de bases é uma sub-estimativa do número de mudanças que efetivamente ocorreram, pois pode haver retorno a uma base já existente após duas ou mais mudanças e cada diferença entre bases pode ser resultado de uma ou mais mudanças de bases.

Considerando o exemplo da Figura 3.7, há apenas um sítio para o qual a base no DNA humano é T e para o DNA do chimpanzé é C. Os outros 68 sítios não possuem variação de bases. Dessa forma, o cálculo da verossimilhança seria, considerando, por exemplo, o primeiro sítio:

L = Pr (começar com A em humano)×Pr(terminar com A em chimpanzé | começou com A em humano) = (¼)×(¼ (1 + $3e^{-4\alpha t}$)) = (1/16)×(1 + $3e^{-4\alpha t}$).

A verossimilhança para os 68 sítios sem substituição de bases é dada por:

$$L = [(1/16)\times (1 + 3e^{-4\alpha t})]^{68}.$$

Para o único sítio em que houve mudança de base, a verossimilhança, segundo o modelo JC, é dada por:

L_{52} = Pr(começar com T em humano)×Pr(terminar com um C em chimpanzé | começou com um T em humano) = $(1/16) \times (1 - e^{-4\alpha t})$.

A verossimilhança total produz: $L = [(1/16) \times (1 + 3e^{-4\alpha t})]^{68} \times (1/16) \times (1 - e^{-4\alpha t})$. Aplicando o logaritmo natural o resultado é dado por:

$$\ln L = 68 \ln [(1/16) \times (1 + 3e^{-4\alpha t})] + \ln[(1/16) \times (1 - e^{-4\alpha t})].$$

O valor αt que maximiza L é dado por 0,0309167 (LEWIS, 2002). Logo, $\ln L = -105{,}3995$.

3.2.6.3 Verossimilhança de árvores com quatro ou mais espécies

A dificuldade adicional para árvores com quatro ou mais espécies são os nós internos, os quais não existiam nos casos anteriores. O problema é que não há observações para esses nós e, assim, todas as hipóteses de substituição de bases devem ser consideradas.

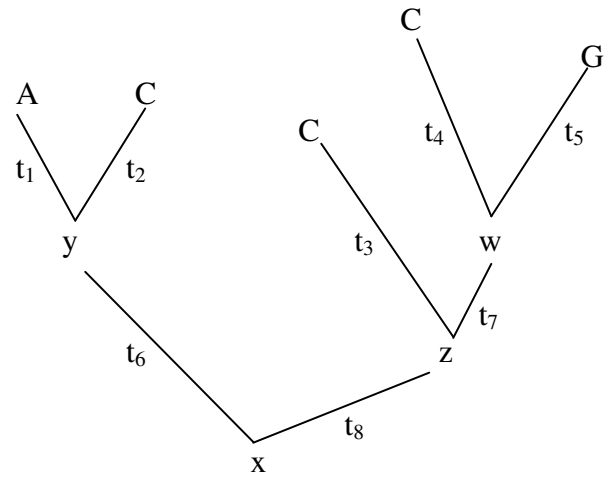


Figura 3.8 - Uma topologia de árvore, com comprimento de ramos e dados de um sítio individual (FELSENSTEIN, 2004)

Para o exemplo da Figura 3.8, os nós y, w, z e x são desconhecidos. Para calcular a verossimilhança dessa árvore, considere um conjunto de seqüências de DNA com *n* sítios. As informações fornecidas são: topologia, comprimento dos ramos e um modelo evolutivo que permite o cálculo de probabilidades de trocas de bases nessa árvore.

Esse modelo permite o cálculo de probabilidades de transição $P_{ij}(t)$, a qual descreve a probabilidade de que um estado *j* existirá no final do ramo *t*, se o estado do início do ramo for *i*. Para essa formalização da função de verossimilhança duas condições são consideradas:

(1) A evolução em diferentes sítios para a topologia apresentada é independente; e

(2) A evolução em diferentes linhas de descendência é independente.

Considerando a condição (1) e a distribuição multinomial da equação (3.2), é possível apresentar a função verossimilhança como (FELSENSTEIN, 2004):

$$L(D \mid T) = \prod_{i=1}^{m} \Pr(D^i \mid T), \tag{3.3}$$

sendo D^i dados do *i*-ésimo sítio e *m* o número de sítios. Com isso, deduz-se que é possível calcular a verossimilhança para cada sítio separadamente.

Considere agora a equação (3.3) para o exemplo da Figura 3.8. A função verossimilhança da árvore para o sítio do exemplo aqui arbitrariamente denominado i, é o somatório das probabilidades de cada cenário possível de eventos:

$$\Pr(D^i \mid T) = \sum_x \sum_y \sum_z \sum_w \Pr(\mathrm{A,C,C,C,G}, x, y, z, w \mid T), \tag{3.4}$$

sendo cada somatório executado para os quatro nucleotídios (A,C,G e T).

Considerando agora a condição (2), ou seja, que a evolução em diferentes linhas de descendência é independente, é possível decompor a probabilidade no lado direito da equação (3.4) em um produto de termos:

$$\begin{aligned}\Pr(\mathrm{A,C,C,C,G}, x, y, z, w \mid T) = \Pr(x) \times \Pr(y \mid x, t_6) \times \Pr(\mathrm{A} \mid y, t_1) \times \Pr(\mathrm{C} \mid y, t_2) \times \\ \times \Pr(z \mid x, t_8) \times \Pr(\mathrm{C} \mid z, t_3) \times \Pr(w \mid z, t_7) \times \Pr(\mathrm{C} \mid w, t_4) \times \Pr(\mathrm{G} \mid w, t_5)\end{aligned} \tag{3.5}$$

3.2.7 Verossimilhança com poda de ramos

Em termos computacionais, a equação (3.5) ainda apresenta dificuldades em razão do grande número de termos a serem calculados, visto que há um crescimento exponencial com o aumento do número de taxa. Numa árvore com *e* espécies, há $e-1$ nós interiores, e cada um pode ter quatro estados, resultando 4^{e-1} termos. Para $e = 10$, existem 262.144 termos; para $e = 20$, existem 274.877.906.944 termos.

Para proporcionar computabilidade ao método, FELSENSTEIN (1981) propôs um algoritmo de poda. O algoritmo calcula apenas uma vez a verossimilhança de determinado nó e dos seus descendentes. Uma vez que esse resultado seja armazenado, não há necessidade de se efetuar esses cálculos novamente. Esse ramo poderia ser então podado da árvore. O método é baseado na movimentação dos somatórios da equação (3.3) de maneira que:

- eles se situem o mais possivelmente para a direita;
- sempre que possível eles sejam colocados entre parênteses.

Considerando que a equação (3.4) pode ser apresentada na forma:

$$\Pr(D^i \mid T) = \sum_x \sum_y \sum_z \sum_w \Pr(x) \times \Pr(y \mid x, t_6) \times \Pr(\mathrm{A} \mid y, t_1) \times \Pr(\mathrm{C} \mid y, t_2) \times \\ \times \Pr(z \mid x, t_8) \times \Pr(\mathrm{C} \mid z, t_3) \times \Pr(w \mid z, t_7) \times \Pr(\mathrm{C} \mid w, t_4) \times \Pr(\mathrm{G} \mid w, t_5) \quad (3.6)$$

então a movimentação dos somatórios produz:

$$\Pr\left(D^i \mid T\right) = \sum_x \Pr(x) \times \left(\sum_y \Pr(y \mid x, t_6) \times \Pr(\mathrm{A} \mid y, t_1) \times \Pr(\mathrm{C} \mid y, t_2) \times \mathrm{termo} \right) \quad (3.7)$$

onde $\mathrm{termo} = \left(\sum_z \Pr(z \mid x, t_8) \times \Pr(\mathrm{C} \mid z, t_3) \times \left(\sum_w \Pr(w \mid z, t_7) \times \Pr(\mathrm{C} \mid w, t_4) \times \Pr(\mathrm{G} \mid w, t_5) \right) \right)$.

É possível notar que os padrões de parênteses e termos das pontas da equação (3.7) acompanham a estrutura da árvore, a qual pode ser representada por: (A,C),(C,(C,G)). O fluxo dos cálculos da equação (3.7) acontece de dentro dos parênteses que estão mais interiorizados para fora. Isto sugere um fluxo de informação das folhas para a raiz da árvore. Esta quantidade de informação é chamada de verossimilhança condicional de uma sub-árvore e pode ser representada por $L_k^i(s)$ que é a probabilidade de tudo que pode ser observado no *i*-ésimo sítio, do nó *k* para cima na árvore (do interior para as folhas). A condição é que o nó *k* tenha estado *s*. Na equação (3.7), o termo $\Pr(\mathrm{C} \mid w, t_4) \times \Pr(\mathrm{G} \mid w, t_5)$ é uma dessas quantidades e representa a probabilidade de tudo que pode ser observado acima ou no nó *w*, lembrando que há quatro possibilidades para o nó *w* (A,C,G e T). A chave do algoritmo de poda é que, uma vez que esses quatro valores são calculados, não há mais necessidade de serem calculados novamente.

O algoritmo pode ser expresso como uma função recursiva que calcula o $L_k^i(s)$ em cada nó da árvore. Isto significa calcular a verossimilhança daquele nó e dos nós imediatamente descendentes. Suponha que o nó *k* tenha como descendentes imediatos os nós *l* e *m*, os quais estão nas pontas dos ramos de tamanhos t_l e t_m. A Figura 3.9 mostra a ligação entre os três nós.

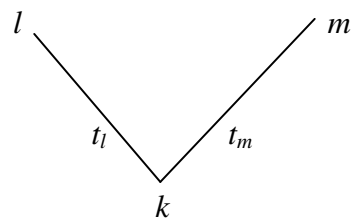


Figura 3.9 - Sub-árvore com três nós *k*, *l* e *m*; e ramos t_l e t_m.

Note que as letras *k*, *l* e *m* identificam a posição do nó na árvore. As letras *s*, *x* e *y* que aparecem na equação (3.8) abaixo denotam os estados que os nós podem assumir para os nós *k*, *l* e *m*, respectivamente. Para *s*, *x* e *y*, os estados podem variar entre as quatro bases de nucleotídeos: A, C, G e T.

A probabilidade dessa sub-árvore pode ser calculada por:

$$L^i(s) = \left(\sum_x \Pr(x \mid s, t_1) \times L^i(x) \right) \times \left(\sum_y \Pr(y \mid s, t_m) \times L^i(y) \right) \qquad (3.8)$$

A equação (3.8) representa a probabilidade de tudo que pode ser observado no nó *k* e acima do nó *k*, dado que o nó *k* tem estado *s*. Ela é dada pelo produto dos eventos que acontecem nas duas linhas descendentes. Na descendência da esquerda, todos os estados para os quais *s* poderia ter sido mudado são somados, e, para cada um desses estados, é calculada a probabilidade de mudar para aquele estado multiplicada pela probabilidade de tudo no nó *l* e acima do nó, considerando que o nó *l* tem estado *x*. A descendência da direita é calculada da mesma forma, considerando-se agora o ramo que vai de *k* a *m*, com distância t_m. A extensão para árvores com mais descendentes é feita com a adição de mais fatores do lado direito da equação (3.8).

A idéia principal do algoritmo é calcular a função verossimilhança para determinado nó e todos os seus descendentes até as folhas, não aplicando o algoritmo a outro nó sem que o nó atual tenha tido todos os seus valores calculados. Quando o cálculo chega a uma folha, um valor observado é atribuído ao nó, pois, qualquer que seja a base encontrada na extremidade da árvore, a mesma terá seu valor definido como 1, e todas as outras bases terão seus valores definidos como zero. A partir daí, a recursividade devolve os valores para o cálculo da verossimilhança do nó anterior. A notação de cálculo de verossimilhança adotada por FELSENSTEIN (2004) é também interessante pelo fato de levar em conta a aplicação do algoritmo de poda.

3.2.8 O princípio *Pulley*

As afirmações feitas na subseção 3.1.6 referem-se a árvores com raiz, como, por exemplo, a árvore apresentada na Figura 3.8. Considere a região próxima à raiz (nós *x*, *y* e *z*). Usando a verossimilhança condicional, a função verossimilhança pode ser descrita como:

$$L^i = \sum_y \sum_z \sum_x \Pr(x) \times \Pr(y \mid x, t_6) \times \Pr(z \mid x, t_8) . \quad (3.9)$$

Porém, o processo de substituição de bases permite que (FELSENSTEIN,2004):

$$\Pr(x) \times \Pr(y \mid x, t_6) = \Pr(y) \times \Pr(x \mid y, t_6) . \quad (3.10)$$

Substituindo a equação (3.10) na equação (3.9), resulta:

$$L^i = \sum_y \sum_z \sum_x \Pr(y) \times \Pr(x \mid y, t_6) \times \Pr(z \mid x, t_8) . \quad (3.11)$$

Isso permite que a raiz da árvore possa ser colocada no nó *y* sem quaisquer mudanças na função verossimilhança daquele sítio na árvore. Usando a mesma explicação, a raiz pode ser movida à direita ou à esquerda da árvore sem afetar a função verossimilhança. Dessa forma, para efeito do cálculo da função verossimilhança, a árvore se comporta como uma árvore sem raiz. Essa propriedade recebe o nome de princípio *Pulley*.

3.2.9 Verossimilhança a partir da notação Newick

Uma das dificuldades do cálculo da verossimilhança em filogenia é o aumento exponencial de operações com o aumento do número de folhas da árvore. Outra questão é a maneira como os atributos associados a cada folha devem ser apresentados como parâmetro de entrada. Uma possibilidade de representação dos dados é aquela mostrada pela matriz da Figura 3.2, a qual possui as linhas como espécies e as colunas como sítios.

Após a definição dos dados de entrada, há a necessidade de se ter uma representação para as verossimilhanças de cada sítio, uma vez que a verossimilhança da árvore é dada pelo produto das verossimilhanças de cada sítio. Essa representação deve conter também as verossimilhanças de cada sub-árvore, ou verossimilhanças parciais, uma vez que a verossimilhança de cada sítio é calculada pelo produto das verossimilhanças de cada sub-árvore. Em ADACHI & HASEGAWA (1996), é possível encontrar uma representação para as verossimilhanças parciais, considerando alguns conceitos como:

1. Princípio *Pulley*: significa que uma árvore com raiz pode ser representada como uma árvore sem raiz sem danos para o cálculo da verossimilhança, conforme brevemente apontado na sub-seção 3.2.7;
2. Reversibilidade: significa que o resultado de uma substituição de base será o mesmo se olhado para frente ou para trás no processo evolutivo (FELSENSTEIN, 1981);
3. Estrutura do cálculo da verossimilhança: indica a maneira como a função verossimilhança pode ser representada e calculada, seguindo a estrutura da própria árvore, conforme equação (3.7);
4. Estruturas de dados para representação de espécies, sítios e sub-árvores (ADACHI & HASEGAWA, 1996);
5. Notação Newick (PHYLIP, 2005): notação utilizada para representar a estrutura de árvores filogenéticas.

É possível conseguir um algoritmo que calcule a verossimilhança de uma árvore filogenética tendo como entrada a árvore em sua notação Newick. Os passos para o cálculo da verossimilhança se baseiam nos símbolos da notação. A cada abertura de parênteses todas as bases dessa sub-árvore têm sua verossimilhança calculada de acordo com o modelo evolutivo escolhido. O modelo utiliza: o valor do ramo que liga a base até seu ancestral (apresentado logo após os dois pontos que seguem a base) e as probabilidades do nó ancestral à base assumir os valores: A, C, G ou T. Dessa forma, o algoritmo calculará as verossimilhanças de cada sítio e apresentará o resultado de acordo com a multiplicação de todas as verossimilhanças. O final do algoritmo é baseado no fechamento do último parênteses. Considere como exemplo a árvore da Figura 3.8. Ela poderia ser representada na notação Newick como:

1. (A:t_1 , C:t_2 , ((C:t_3 , (C:t_4 , G:t_5):t_7):t_8):t_6), com raiz ou
2. ((A:t_1,C:t_2):t_6):t_8,C:t_3):t_7,(C:t_4,G:t_5)), sem raiz e com todos os tamanhos dos ramos sendo colocados imediatamente à direita da formação do nó.

A notação 2 pode ser utilizada em razão dos princípios *Pulley* e reversibilidade.

3.3 Métodos de reconstrução de topologias

Alguns métodos de geração de árvores filogenéticas caracterizam-se pela sua construção de acordo com características observadas nos dados de entrada. Esses dados observados são estruturados em matrizes de distância. O conceito fundamental de métodos que utilizam matrizes de distâncias como entrada pode ser resumido pelos seguintes aspectos:

-existe uma matriz de distâncias que possui distâncias observadas entre as espécies (ou OTU´s, do inglês, *Operational Taxonomic Units*) e,

-qualquer árvore que determina o comprimento de seus ramos acaba por predizer um conjunto de distâncias d_{ij} o qual pode ser calculado pela soma dos comprimentos dos ramos entre as espécies i e j.

A matriz de distâncias pode ser gerada pela observação das diferenças existentes entre seqüências de bases nucleotídicas, como as seqüências mostradas pela Figura 3.10.

Humano . . . T G T A T C G C T C ...
Coelho . . . T G T G T C G C T C ...

Humano . . . T G T A T C G C T C . . .
Galinha . . . A G T C T C G T T C . . .

Coelho . . . T G T G T C G C T C . . .
Galinha . . . A G T C T C G T T C . . .

Figura 3.10 - Partes de seqüências de código genético de três diferentes espécies (HUSMEIER, 2006)

A matriz de distâncias gerada pela observação das diferenças existentes entre as bases de cada par de espécies da Figura 3.10 é mostrada pela Figura 3.11. Nesse exemplo, a distância é dada pelo número de vezes que uma base nucleotídica difere da sua base correspondente em cada sítio de cada par de espécies comparadas. Existem formas mais sofisticadas de se obter essas distâncias par-a-par (FELSENSTEIN, 2004).

	Humano	Coelho	Galinha
Humano	0	1	3
Coelho	1	0	3
Galinha	3	3	0

Figura 3.11 - Matriz de distâncias observadas pelas diferenças entre as espécies da Figura 3.10.

As linhas e colunas contêm as espécies sendo observadas. Os elementos da matriz são valores que representam a distância entre cada par de espécies, dadas a linha e a coluna, conforme Figura 3.11. A matriz de distâncias é sempre uma matriz quadrada $e \times e$, sendo e o número de espécies analisadas (para o exemplo da Figura 3.11, e=3). A diagonal principal contém apenas valores nulos, dado que a distância ou diferença entre uma espécie e ela mesma é nula.

A matriz é também simétrica, pois a direção da evolução não é considerada para o cálculo da distância. Essa é uma das propriedades de alguns modelos de substituição de bases nucleotídicas (ou contendo aminoácidos), que consideram a mesma distância evolutiva independentemente da direção evolutiva no tempo. Essa propriedade, chamada reversibilidade (FELSENSTEIN, 1981), é ilustrada pela Figura 3.12. O cálculo da distância entre as espécies *A* e *B*, por exemplo, independe da direção evolutiva ser de *A* para *B* ou de *B* para *A*, conforme visto na sub-seção 2.7.

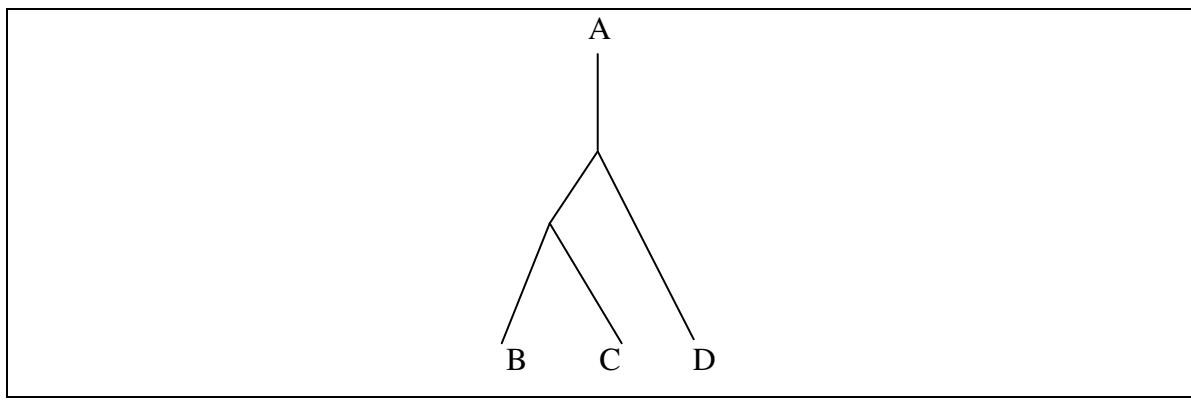


Figura 3.12 - Árvore com raiz e três espécies descendentes

Uma característica desejável da matriz de distâncias observadas é a aditividade. Um conjunto de espécies (taxa) constitui um espaço de distribuição de dados. Esse espaço é um espaço métrico se, para a distância $d(i,j)$ entre espécies i e j forem satisfeitas as condições:

- $d(i,j) > 0$, se $i \neq j$;
- $d(i,j) = 0$, se $i = j$;
- $d(i,j) = d(j,i)$; e
- $d(i,j) \leq d(i,k) + d(k,j)$.

Se uma matriz de distâncias observadas obedece às quatro condições apresentadas, essa matriz é chamada de matriz aditiva. Outra característica presente em métodos de reconstrução filogenética a partir de matrizes de distâncias é o fato de que cada árvore gerada leva à criação de outra matriz de distâncias, baseada na soma dos comprimentos dos ramos entre cada par de espécies na árvore. A Figura 3.13 mostra a topologia de uma árvore e seus comprimentos de ramos. A Figura 3.14 mostra a matriz de distâncias, também denominada de matriz de distâncias patrísticas, gerada a partir da árvore da Figura 3.13.

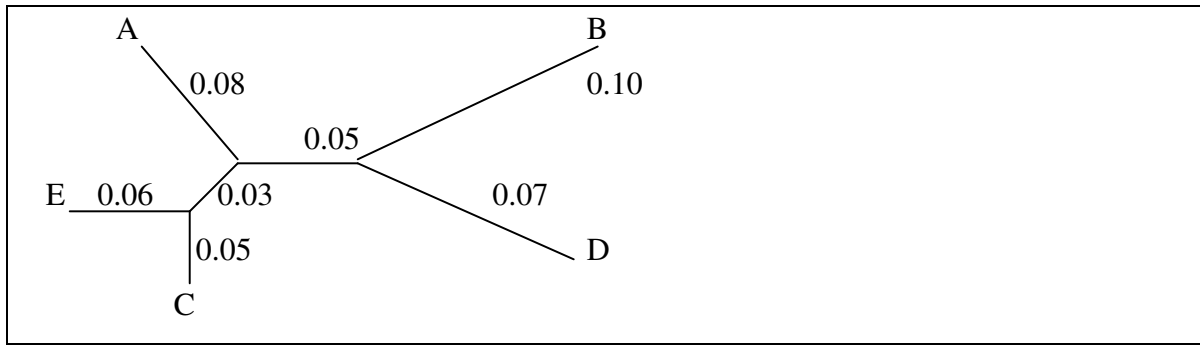


Figura 3.13 - Árvore com cinco espécies e seus comprimentos de ramos

	A	B	C	D	E
A	0	0.23	0.16	0.20	0.17
B	0.23	0	0.23	0.17	0.24
C	0.16	0.23	0	0.20	0.11
D	0.20	0.17	0.20	0	0.21
E	0.17	0.24	0.11	0.21	0

Figura 3.14 - Matriz de distâncias gerada pela árvore da Figura 3.13

A diferença existente entre a matriz de distâncias que foram observadas e a matriz de distâncias geradas pela topologia e comprimento de ramos da árvore filogenética (matriz patrística) é utilizada por alguns métodos de inferência filogenética, como o método dos quadrados mínimos.

3.3.1 Método dos quadrados mínimos

A versão proposta por FITCH & MARGOLIASH (1967) é chamada de *Weighted least squares* e a versão proposta por CAVALLI & EDWARDS (1967) é conhecida como *Unweighted least squares.* Os dois métodos, claramente relacionados, utilizam um critério de otimalidade para analisar dados de distância entre espécies.

Um critério de otimalidade é definido quando:

- diferentes filogenias devem ser comparadas umas às outras para determinar qual é a melhor;
- existe uma função-objetivo e um algoritmo que calcula essa função-objetivo para as diferentes filogenias.

O uso de um critério de otimalidade pode tornar mais lenta a busca de uma filogenia ótima, uma vez que todas as filogenias possíveis devem ser avaliadas pela função-objetivo.

Para o método dos quadrados mínimos, o critério consiste em escolher a árvore que minimiza o erro, ou seja, minimiza uma medida de discrepância entre as distâncias observadas entre as espécies (aquelas contidas na matriz original) e as distâncias patrísticas, ou seja, os comprimentos dos ramos que conectam cada par de espécies na árvore gerada a partir das distâncias observadas.

Para cada árvore possível, o método compara a soma de todas as diferenças quadráticas entre as distâncias na matriz original e as distâncias de cada par de espécies na árvore e, então, o método escolhe a árvore que minimiza essa soma.
A função-objetivo utilizada pelo método é dada por:

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (D_{ij} - d_{ij})^2 \qquad (3.12)$$

A variável w_{ij} diferencia os dois métodos mencionados anteriormente, *Unweighted least squares* e *Weighted least squares* . Para Cavalli-Sforza e Edwards a variável tem valor 1, e para Fitch e Margoliash, $w_{ij} = 1/(D_{ij})^2$ (FELSENSTEIN, 2004).

O critério consiste em encontrar o valor ótimo de Q. No caso dos quadrados mínimos, o valor ótimo de Q é o menor valor encontrado. Um exemplo da execução do método dos quadrados mínimos é apresentado abaixo. Seja a matriz de distâncias observadas dada na Figura 3.15.

	A	B	C	D
A	0	17	21	27
B	17	0	12	18
C	21	12	0	14
D	27	18	14	0

Figura 3.15 - Matriz de distâncias observadas

A Figura 3.16 apresenta a topologia e os comprimentos de ramos da árvore gerada a partir da matriz de distâncias da Figura 3.15. Aplicando-se a função-objetivo da equação 3.12 para as espécies B e D, os resultados são:

D_{BD}= 18 (distância observada);

d_{BD}= 6 + 2 + 8 = 16 (distância patrística);

Erro MínimoQuadrático$_{BD}$= $(18 - 16)^2 = 4$.

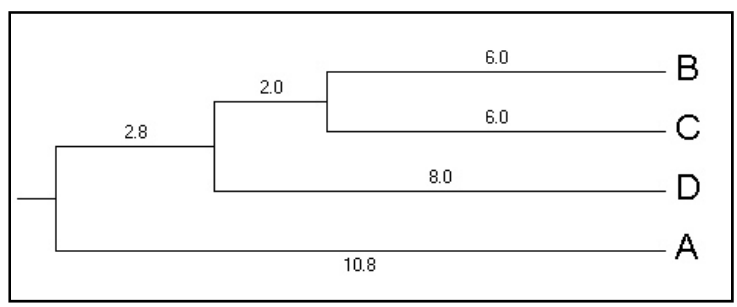


Figura 3.16 - Árvore gerada pelas distâncias observadas na matriz da Figura 3.15

Para o exemplo acima, o erro foi calculado para apenas uma árvore e para apenas as espécies B e D. Na prática, esse cálculo deve ser feito para todos os pares de espécies de cada possível topologia. O problema é o aumento do número de árvores possíveis com o aumento do número de espécies em uma árvore. Uma filogenia com quatro espécies tem a possibilidade de três diferentes topologias de árvores sem raiz. Uma filogenia com cinco espécies tem a possibilidade de quinze diferentes topologias de árvores sem raiz. Uma filogenia com dez espécies tem a possibilidade de 2.027.025 diferentes topologias de árvores sem raiz, o que torna o método inviável para um número grande de espécies em uma árvore.

3.3.2 Método da evolução mínima

O método da evolução mínima estima o comprimento total dos ramos de cada possível topologia. Depois de avaliar todas as possíveis topologias, o método escolhe a topologia com o menor comprimento total de ramos.

Este método é computacionalmente intensivo e, portanto, lento, principalmente com o aumento do número de espécies. Assim como o método dos quadrados mínimos, o método de evolução mínima utiliza um critério de otimalidade para analisar dados de distâncias entre espécies.

A função-objetivo utilizada pelo método calcula o comprimento total dos ramos de cada árvore. O valor ótimo da função é encontrado quando o menor valor de comprimento total de ramos é encontrado (MOLECULAR, 2006). O menor valor é o escolhido uma vez que a árvore que apresentar a menor variação entre as espécies, ou seja, os menores comprimentos de ramos, é considerada como a que melhor explica a evolução entre as espécies. O método é eficiente, pois, certamente encontrará a árvore, uma vez que todas as árvores serão consideradas.

Seja S o comprimento total de uma árvore o qual é dado pela soma de todos os comprimentos dos ramos da árvore. Seja a função-objetivo, S, dada por:

$$S = \sum_{i=1}^{T} b_i$$

sendo T o número de ramos e b_i a estimativa do comprimento do i-ésimo ramo.

Considerando-se que as árvores das Figuras 3.17 e 3.18 estivessem sendo analisadas pelo método de evolução mínima, a árvore da Figura 3.18 seria escolhida como a melhor árvore, uma vez que a soma total dos comprimentos de seus ramos é menor do que a soma total dos comprimentos dos ramos da árvore da Figura 3.17.

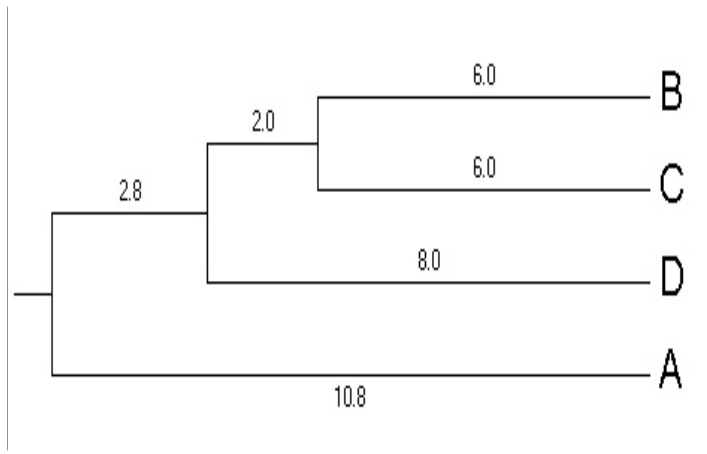


Figura 3.17 - Árvore filogenética com comprimento total S=35,6

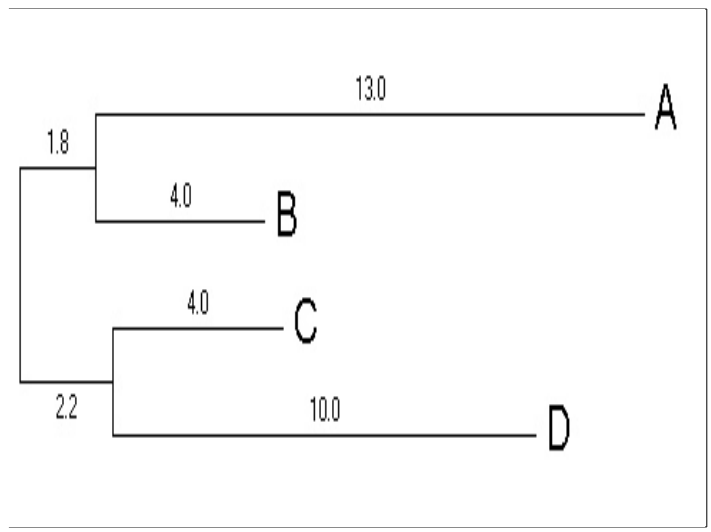


Figura 3.18 - Árvore filogenética com comprimento total S=35

O método da evolução mínima foi primeiramente empregado por KIDD & SGARAMELLA-ZONTA (1971) que usaram a soma dos valores absolutos dos ramos para calcular o

comprimento total da árvore. O método tem início com o cálculo das distâncias entre as espécies gerando uma matriz de distâncias entre pares de espécies. Os autores apresentam três possíveis métricas de distâncias entre as espécies. A aplicação de quaisquer das três métricas resulta em uma matriz de distâncias simétrica com zeros na diagonal principal. As árvores geradas pelo método:

- contém espécies que, uma vez separadas, não podem ser agrupadas a outras espécies.
- geram apenas duas outras espécies a partir de uma espécie ancestral e
- não possuem raiz.

Para um problema com N espécies um número de $\prod_{K=3}^{N}(2K-5)$ árvores sem raiz pode ser gerado pelo agrupamento de diferentes espécies. As árvores são consideradas iguais se suas topologias e folhas coincidem. Para uma determinada topologia várias outras podem ser produzidas por meio da permutação de suas folhas. Se árvores iguais são produzidas pelas permutações, o número de árvores diferentes é reduzido.

O comprimento dos ramos das árvores é calculado utilizando o método dos quadrados mínimos. O critério da evolução mínima é aplicado às árvores escolhendo a árvore que apresentar o menor comprimento. Os valores dos ramos das árvores é adicionado ao comprimento total considerando seus valores absolutos. Dessa forma, o método da evolução mínima utiliza dois critérios para a escolha de uma árvore a partir de uma matriz de distâncias: um para a escolha dos ramos e outro para a escolha da topologia.

Mais recentemente, o método da evolução mínima tem sido utilizado conforme proposto por RZHESKY & NEI (1992). Nessa versão, os comprimentos dos ramos das árvores são calculados utilizando-se o método dos quadrados mínimos respeitando o valor negativo dos ramos, caso existam. O algoritmo utilizado pelo método consiste em, primeiramente, construir uma árvore usando o algoritmo *Neighbor-Joining* (SAITOU & NEI, 1987) e então calcular a soma total dos ramos dessa árvore. Depois disso, algum critério de proximidade entre árvores é utilizado, e todas as topologias próximas à topologia gerada pelo neighbor-joining têm seu comprimento S calculado. Todos os valores de S são então comparados uns aos outros e, a árvore com menor valor de S é escolhida como a árvore final. Segundo

RZHESKY & NEI (1992), essa árvore final é, usualmente, a árvore gerada pelo *Neighbor-Joining*. Variações do NJ que utilizam topologias próximas à topologia gerada pelo algoritmo original são apresentadas na seção 3.3.3.3.

3.3.3 *Neighbor-Joining*

O algoritmo *Neighbor-Joining* (NJ) foi proposto por Saitou e Nei (SAITOU & NEI, 1987) e baseia-se na reconstrução de árvores filogenéticas sem raiz. Ao final de sua execução, o método fornece a topologia da árvore filogenética e também os comprimentos dos seus ramos.

O algoritmo tem como entrada uma matriz com *N* espécies. O número de pares de espécies cujas distâncias devem ser calculadas para se aplicar o algoritmo é dado pela fórmula $N(N\text{-}1)/2$, os quais representam o número de elementos do triângulo superior (ou inferior) da matriz. O método tem início com uma árvore-estrelasem raiz com N espécies, conforme a Figura 3.19, onde *N* foi tomado como 8.

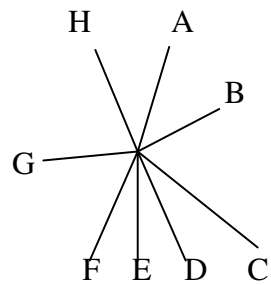


Figura 3.19 - **Árvore-estrela que dá início à execução do método NJ**

Para criar a topologia da árvore filogenética, o algoritmo segue o critério de evolução mínima (EM). O NJ utiliza o método EM para calcular a topologia da árvore escolhendo, a cada passo do algoritmo, o par de espécies que irá minimizar o comprimento total da árvore. A árvore resultante não é, necessariamente, a árvore de comprimento total mínimo, pois a construção da árvore escolhendo os ramos de menor comprimento a cada passo não garante a construção do ótimo global, segundo o critério de evolução mínima.

De qualquer modo, um par de espécies é escolhido para ser agrupado a cada passo. O par escolhido é aquele cuja junção leva à criação de uma nova árvore com o valor mínimo de

comprimento total dos ramos, considerando-se todas as junções possíveis entre as linhas e as colunas da matriz de entrada. Esse cálculo resulta na construção de uma nova matriz *M* cujos valores *m*(*i*,*j*) representam o comprimento total dos ramos, como resultado da junção da linha *i* e coluna *j*. A matriz da Figura 3.20 representa as somas das distâncias dos ramos para a junção de cada par de espécies.

	A	B	C	D	E	F	G	H
A	0							
B	36.67	0						
C	38.33	38.33	0					
D	39	39	38.67	0				
E	40.33	40.33	40	39.67	0			
F	40.33	40.33	40	39.67	37	0		
G	40.17	40.17	39.83	39.50	38.83	38.83	0	
H	40.17	40.17	39.83	39.50	38.83	38.83	37.67	0

Figura 3.20 - Matriz das somas dos ramos resultantes da junção de cada par de espécies

Para a matriz da Figura 3.20, o par que minimiza a soma dos comprimentos dos ramos da árvore é o par A-B. Esse par é substituído na árvore original (Figura 3.19) por um ancestral comum, como mostrado na Figura 3.21, que representa a nova topologia da árvore.

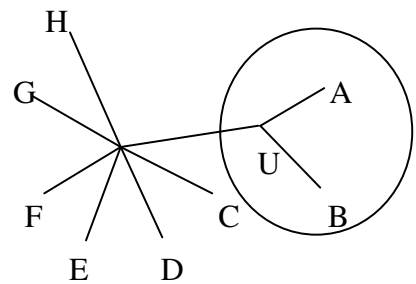


Figura 3.21 - Topologia resultante da junção das espécies A e B

Para cada junção, a árvore original perde dois nós (espécies) e ganha um novo nó (nó U na Figura 3.21). A junção dos dois nós provoca a formação de novos ramos. Para calcular o comprimento dos novos ramos que surgiram com a junção, o NJ utiliza o método de Fitch-

Margoliash (FITCH & MARGOLIASH, 1967). Dessa forma, uma nova matriz é formada por meio da eliminação da linha e da coluna relativas às espécies que foram agrupadas e acréscimo de uma linha e uma coluna para o ancestral comum recém criado, conforme Figura 3.21. O algoritmo é então novamente executado para essa nova matriz e assim sucessivamente, até que a matriz tenha dimensão 3×3 e a árvore-estrelaoriginal tenha duas espécies. O número de junções realizadas é dado por $N-2$.

É possível resumir o método NJ em quatro procedimentos principais:

1. A escolha de duas espécies da árvore atual que devam ser agrupadas;
2. O cálculo do comprimento do novo ramo resultante da junção (para o exemplo da Figura 3.21 o comprimento do nó U até o nó da árvore estrela);
3. O cálculo dos comprimentos dos ramos que começam nas duas espécies as quais foram agrupadas até o novo nó na árvore-estrela(para o exemplo da Figura 3.21, a distância do nó A até o nó U e a distância do nó B até o nó U) ;
4. O cálculo da nova matriz de distância depois da junção das OTUs com uma linha e uma coluna a menos (como a matriz da Figura 3.22).

	U	C	D	E	F	G	H
U	0						
C	6.5	0					
D	9.5	5	0				
E	11.5	7	8	0			
F	14.5	10	11	5	0		
G	11.5	7	8	6	9	0	
H	15.5	11	12	10	13	8	0

Figura 3.22 - Nova matriz depois da junção das espécies A e B

3.3.3.1 Detalhamento dos passos do algoritmo

Os principais passos do NJ são detalhados a seguir.

3.3.3.1.1 Cálculo do comprimento total da árvore-estrela inicial

Seja S a árvore-estrelainicial e seu comprimento total dado por S_{total}. Para se iniciar o algoritmo *Neighbor-Joining*, a única informação disponível é a matriz de distâncias, ou seja, a distância entre todos os pares de nós.

Pode-se iniciar a demonstração do cálculo com a menor das matrizes, ou seja, uma matriz de duas linhas e duas colunas. Essa matriz representa as distâncias entre os nós 1 e 2 de uma árvore-estrela conforme apresentado na Figura 3.23.

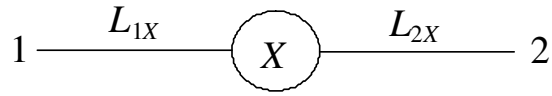


Figura 3.23 - Árvore-estrela com dois nós

A matriz de distâncias que fornece informações dessa árvore é dada por:

$$M = \begin{bmatrix} 0 & D_{12} \\ D_{12} & 0 \end{bmatrix}$$

Conforme apresentado na Figura 3.23, o comprimento da árvore-estrela inicial (S_{total}) é dado na forma: $S_{total} = L_{1X} + L_{2X}$. Embora L_{1X} e L_{2X} não sejam conhecidos, da Figura 3.23 e da matriz M de distâncias é possível deduzir diretamente que $L_{1X} + L_{2X} = D_{12}$, levando a $S_{total} = D_{12}$.

Considerando-se uma árvore-estrela com três nós, as distâncias entre os três seriam apresentadas conforme Figura 3.24.

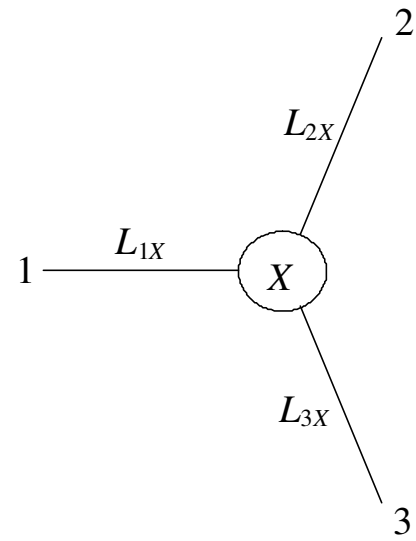


Figura 3.24 - Árvore-estrela com três nós.

As informações dessa árvore viriam de uma matriz *M* de distâncias que seria dada por:

$$M = \begin{bmatrix} 0 & D_{12} & D_{13} \\ D_{12} & 0 & D_{23} \\ D_{13} & D_{23} & 0 \end{bmatrix}$$

Conforme apresentado na Figura 3.24, o comprimento da árvore-estrela inicial (S_{total}) é dado na forma:. $S_{total} = L_{1X} + L_{2X} + L_{3X}$.

Embora L_{1X} , L_{2X} e L_{3X} não sejam conhecidos, da Figura 3.24 é possível deduzir que:

$$S_{total} = L_{1X} + L_{2X} + L_{3X} = \frac{1}{2}\left[\left(L_{1X} + L_{2X}\right) + \left(L_{1X} + L_{3X}\right) + \left(L_{2X} + L_{3X}\right)\right] = \frac{D_{12} + D_{13} + D_{23}}{2}.$$

Seguindo a mesma seqüência de passos, para uma árvore-estrela com um número arbitrário $N \geq 2$ de nós, o comprimento total da árvore é dado por: $S_{total} = \frac{1}{N-1}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} D_{ij}$, pois os comprimentos L_{iX} , i=1,...,N, serão somados N–1 vezes para que S_{total} possa ser apresentada em função apenas dos $\frac{N(N-1)}{2}$ elementos do triângulo superior da matriz de distâncias.

3.3.3.1.2 Cálculo do comprimento do novo ramo

Na subseção anterior, foi mostrada a forma de se efetuar o cálculo da soma dos ramos de uma árvore estrela. Na Figura 3.25, está representada uma árvore com *N* nós que acabou de sofrer a junção entre os nós *i* e *j*. Os passos a seguir mostram como obter o comprimento total desta árvore.

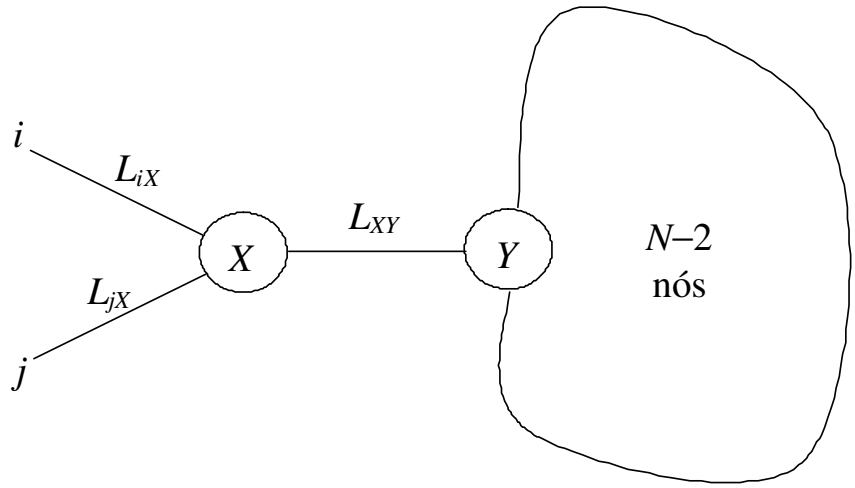


Figura 3.25 - Árvore-estrela com *N* nós após sofrer uma junção entre os nós *i* e *j*

Conforme mostrado na subseção anterior, o comprimento total dos ramos da parte remanescente da árvore-estrela (aquela com *N*–2 nós), mesmo sem conhecer L_{pX}, com $p \in \{1,..,N\}$ e $p \neq i, p \neq j$, é dado na forma:

$$S_{[-i][-j]} = \frac{1}{N-3} \sum_{\substack{l=1 \\ l \neq i \\ l \neq j}}^{N-1} \sum_{\substack{k=l+1 \\ k \neq i \\ k \neq j}}^{N} D_{lk}$$

Mesmo sem conhecer L_{iX} e L_{jX}, também é possível calcular o comprimento total relativo à parte da junção, como segue:

$$S_{ij} = L_{iX} + L_{jX} = D_{ij}$$

Logo, para se chegar a L_{XY}, é necessário:

- somar a distância entre os nós *i* e *j* e todos os demais *N*–2 nós da Figura 3.25, produzindo: $\sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{N}\left(D_{ik}+D_{jk}\right)$;

- subtrair desta soma o comprimento de todos os ramos que não forem *XY*, comprimento este que totaliza: $(N-2)D_{ij}+2\sum_{\substack{l=1 \\ l\neq i \\ l\neq j}}^{N-1}\sum_{\substack{k=l+1 \\ k\neq i \\ k\neq j}}^{N}D_{lk}$;

- dividir o resultado desta subtração pelo número de vezes que L_{XY} foi somado, dado por: $2(N-2)$.

Logo, obtém-se:

$$L_{XY}=\frac{1}{2(N-2)}\left[\sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{N}\left(D_{ik}+D_{jk}\right)-(N-2)D_{ij}-2\sum_{\substack{l=1 \\ l\neq i \\ l\neq j}}^{N-1}\sum_{\substack{k=l+1 \\ k\neq i \\ k\neq j}}^{N}D_{lk}\right].$$

3.3.3.1.3 Cálculo do comprimento total para todos os possíveis pares de nós candidatos à junção

Considerando que a árvore-estrela ainda possui *N* nós candidatos a sofrerem junção, o par (*i*,*j*) a ser escolhido é aquele que produz a menor distância total, a qual será calculada a seguir, ainda tomando como ponto de referência a Figura 3.25.

Todas as informações constam das subseções anteriores, pois a distância total, para um par arbitrário (*i*,*j*), é dada por:

$$S_{total}^{ij}=\frac{1}{N-3}\sum_{\substack{l=1 \\ l\neq i \\ l\neq j}}^{N-1}\sum_{\substack{k=l+1 \\ k\neq i \\ k\neq j}}^{N}D_{lk}+D_{ij}+\frac{1}{2(N-2)}\left[\sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{N}\left(D_{ik}+D_{jk}\right)-(N-2)D_{ij}-2\sum_{\substack{l=1 \\ l\neq i \\ l\neq j}}^{N-1}\sum_{\substack{k=l+1 \\ k\neq i \\ k\neq j}}^{N}D_{lk}\right]$$

que ainda pode ser simplificada, produzindo:

$$S_{total}^{ij} = \frac{1}{2(N-2)} \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{N} \left(D_{ik} + D_{jk} \right) + \frac{1}{2} D_{ij} + \frac{1}{N-2} \sum_{\substack{l=1 \\ l \neq i \\ l \neq j}}^{N-1} \sum_{\substack{k=l+1 \\ k \neq i \\ k \neq j}}^{N} D_{lk} \quad .$$

Para todos os $\frac{N(N-1)}{2}$ pares (*i*,*j*) candidatos, a distância S_{total}^{ij} deve ser calculada. O par (*i*,*j*) a ser escolhido é aquele que produz o menor valor para S_{total}^{ij}.

3.3.3.1.4 Cálculo do comprimento dos ramos dos nós que sofreram junção (Fitch-Margoliash)

Considere que o par (*i*,*j*) seja aquele que sofreu a junção. O conhecimento das distâncias dos nós *i* e *j* ao nó interno *X* (veja Figura 3.25) não foi necessário para se obter S_{total}^{ij}. No entanto, o método NJ deve fornecer também o comprimento dos ramos, e não apenas a topologia da árvore. Sendo assim, é necessário obter L_{iX} e L_{jX}.

Para tanto, considere a representação apresentada na Figura 3.26. *Z* é uma folha auxiliar que representa o agrupamento de todas as folhas da árvore, com exceção de *i* e *j*.

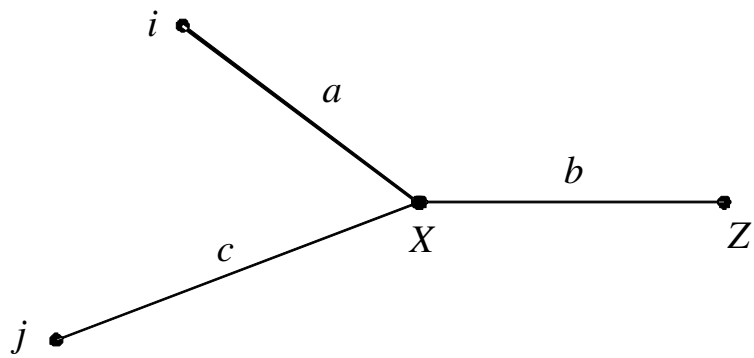


Figura 3.26 - Árvore da Figura 3.25 com um agrupamento dos *N*–2 nós que não sofreram junção, representado pelo nó *Z*.

O que se quer determinar aqui são os valores de *a* e *c*, ou seja, de L_{iX} e L_{jX}, respectivamente. Embora a distância $D_{ij} = a + c$ seja conhecida, *a* e *c* são desconhecidas.

Outras distâncias que podem ser conhecidas são: $D_{iZ} = a + b$ e $D_{jZ} = c + b$. Como Z é uma folha auxiliar que representa o agrupamento de todas as folhas da árvore, com exceção de i e j, então D_{iZ} e D_{jZ} são dadas, respectivamente, pela média das distâncias das folhas i e j a todas as demais folhas da árvore, na forma:

$$D_{iZ} = a + b = \frac{1}{N-2} \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{N} D_{ik} \qquad D_{jZ} = c + b = \frac{1}{N-2} \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{N} D_{jk}$$

Com o conhecimento de D_{ij}, D_{iZ} e D_{jZ}, temos:

- $L_{iX} = a = \frac{2a}{2} = \frac{(a+c)+(a+b)-(c+b)}{2} = \frac{D_{ij} + D_{iZ} - D_{jZ}}{2}$

- $L_{jX} = c = \frac{2c}{2} = \frac{(a+c)+(c+b)-(a+b)}{2} = \frac{D_{ij} + D_{jZ} - D_{iZ}}{2}$.

Fazendo as substituições para D_{iZ} e D_{jZ}, resulta:

$$L_{iX} = \frac{D_{ij} + \frac{1}{N-2} \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{N} (D_{ik} - D_{jk})}{2}$$

$$L_{jX} = \frac{D_{ij} + \frac{1}{N-2} \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{N} (D_{jk} - D_{ik})}{2}$$

3.3.3.1.5 Cálculo da nova matriz de distâncias

A nova matriz de distâncias terá dimensões $(N-1) \times (N-1)$, pois irá perder as linhas e colunas i e j e ganhar uma nova linha e uma nova coluna, as quais irão ocupar a posição $\min(i,j)$, tanto para a nova linha como para a nova coluna.

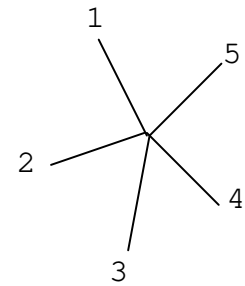
- antes da junção, para cada nó k, com $k \neq i$ e $k \neq j$, havia duas distâncias: D_{ik} e D_{jk}. Com a junção, fica apenas uma distância para cada nó k: $\overline{D} = \frac{D_{ik} + D_{jk}}{2}$.

3.3.3.2 Um exemplo de execução do NJ

O algoritmo NJ será apresentado a seguir por meio da execução de um exemplo. Seja a árvore inicial composta por cinco espécies, com as distâncias observadas dadas pela matriz *M* conforme Figura 3.27.

	1	2	3	4	5
1	0	5	4	7	6
2	5	0	7	10	9
M= 3	4	7	0	7	6
4	7	10	7	0	5
5	6	9	6	5	0

a



b

Figura 3.27(a) - Matriz de distâncias observadas. (b) - Árvore estrela

Primeira Iteração (N=5)

Passo 1. Cálculo da Matriz Soma

Para todos os $\frac{N(N-1)}{2}$ pares (*i*,*j*) candidatos, a distância S_{total}^{ij} deve ser calculada. O par (*i*,*j*) a ser escolhido é aquele que minimiza S_{total}^{ij}. O cálculo é dado pela fórmula:

$$S_{total}^{ij} = \frac{1}{2(N-2)} \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{N} \left(D_{ik} + D_{jk} \right) + \frac{1}{2} D_{ij} + \frac{1}{N-2} \sum_{\substack{l=1 \\ l \neq i \\ l \neq j}}^{N-1} \sum_{\substack{k=l+1 \\ k \neq i \\ k \neq j}}^{N} D_{lk}$$

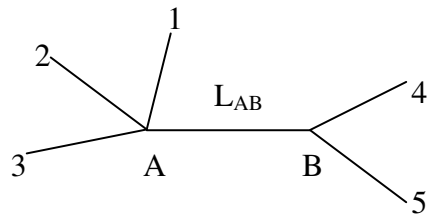
Para cada par candidato, a fórmula acima deve ser executada, como, por exemplo, para o par de espécies 1-2:

$$S_{total}^{12} = \frac{1}{2(3-2)} \sum_{\substack{k=1 \\ k \neq i \\ k \neq j}}^{5} \left(D_{1k} + D_{2k} \right) + \frac{1}{2} D_{12} + \frac{1}{5-2} \sum_{\substack{l=1 \\ l \neq i \\ l \neq j}}^{4} \sum_{\substack{k=l+1 \\ k \neq i \\ k \neq j}}^{5} D_{lk} = 15.6667$$

Os cálculos para cada par possível resultam em:

$$
S = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}
\begin{array}{c}
\begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\
\left[\begin{array}{ccccc}
0 & 15.6667 & 16.3333 & 17.0000 & 17.0000 \\
15.6667 & 0 & 16.3333 & 17.0000 & 17.0000 \\
16.3333 & 16.3333 & 0 & 16.6667 & 16.6667 \\
17.0000 & 17.0000 & 16.6667 & 0 & 15.3333 \\
17.0000 & 17.0000 & 16.6667 & 15.3333 & 0
\end{array}\right]
\end{array}
$$

Passo 2. Escolha do par de espécies que minimiza o comprimento total da árvore
Par = [4 5] , correspondente ao valor 15.3333.



Passo 3. Cálculo da nova matriz de distâncias

As novas distâncias a serem calculadas são as distâncias entre cada espécie restante na árvore-estrelae o novo nó B. Essas distâncias são calculadas pela média aritmética entre as distâncias de cada nó na árvore-estrelae os nós 4 e 5 que sofreram a junção. As distâncias consideradas são aquelas anteriores à junção. Antes da junção, para cada nó k, com $k \neq i$ e $k \neq j$, havia duas distâncias: D_{ik} e D_{jk}. Com a junção, fica apenas uma distância para cada nó k: $\overline{D} = \dfrac{D_{ik} + D_{jk}}{2}$.

Para essa iteração do algoritmo, haverá três cálculos de novas distâncias : $D_{1\text{-}B}$, $D_{2\text{-}B}$ e $D_{3\text{-}B,}$ dados por:

$$\overline{D}_{1B} = \frac{D_{41} + D_{51}}{2} = \frac{7+6}{2} = 6.5$$

$$\overline{D}_{2B} = 9.5$$

$$\overline{D}_{3B} = 6.5$$

Resultando na nova matriz M :

M =	1	2	3	B
1	0	5.0000	4.0000	6.5000
2	5.0000	0	7.0000	9.5000
3	4.0000	7.0000	0	6.5000
B	6.5000	9.5000	6.5000	0

Passo 4. Cálculo dos comprimentos dos ramos criados

O cálculo dos comprimentos dos ramos que sofreram a junção é dado por:

$$L_{iX} = \frac{D_{ij} + \frac{1}{N-2}\sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{N}\left(D_{ik} - D_{jk}\right)}{2}$$

$$L_{jX} = \frac{D_{ij} + \frac{1}{N-2}\sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{N}\left(D_{jk} - D_{ik}\right)}{2}$$

Nessa iteração do algoritmo, os ramos criados pela junção são:

•Ramo 4-B ou, pela fórmula acima:

$$L_{4B} = \frac{D_{45} + \frac{1}{5-2}\sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{5}\left(D_{4k} - D_{5k}\right)}{2}$$

$$L_{4B} = \frac{5 + \frac{1}{3}\sum_{\substack{k=1 \\ k\neq i \\ k\neq j}}^{5}\left(D_{4k} - D_{5k}\right)}{2} = 3$$

•Ramo 5-B ou, pela fórmula acima: $L_{5B} = 2$.

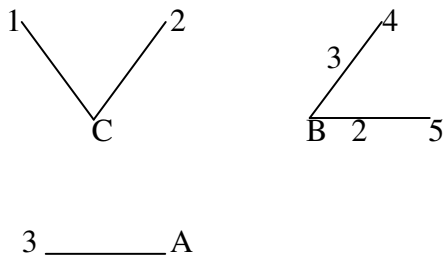
Segunda Iteração (N = 4)

Passo 1. Cálculo da matriz Soma

S=	1	2	3	B
1	0	12.5000	13.0000	13.0000
2	12.5000	0	13.0000	13.0000
3	13.0000	13.0000	0	12.5000
B	13.0000	13.0000	12.5000	0

Passo 2. Escolha do novo par

Par = [1 2], correspondente ao valor 12.5000.



Passo 3. Cálculo da nova matriz de distâncias

M=	C	3	B
C	0	5.5000	8.0000
3	5.5000	0	6.5000
B	8.0000	6.5000	0

Passo 4. Cálculo dos comprimentos dos ramos criados

C-1: 1
C-2: 4

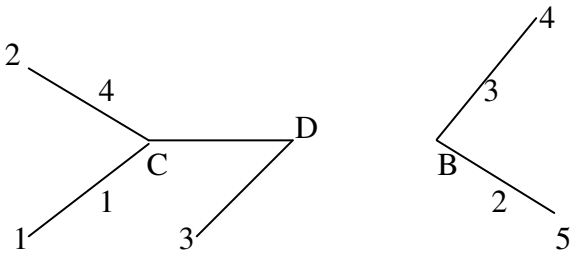
Terceira iteração (N = 3)

Passo 1 . Cálculo da matriz soma

		C	3	B
	C	0	10	10
S=	3	10	0	10
	B	10	10	0

Passo 2. Escolha do novo par

Par=[C 3]



Passo 3. Cálculo dos comprimentos dos ramos criados

D-C: 1
D-3: 2

Final das iterações, pois *N*=2.

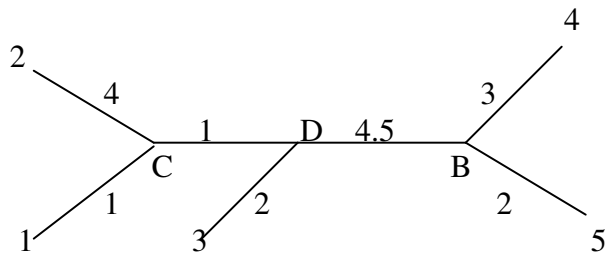
Cálculo do comprimento do ramo que une as duas partições:

$$L_{DB} = \frac{1}{2(N-2)}\left[\sum_{\substack{k=1\\k\neq i\\k\neq j}}^{N}\left(D_{ik}+D_{jk}\right)-(N-2)D_{ij}-2\sum_{\substack{l=1\\l\neq i\\l\neq j}}^{N-1}\sum_{\substack{k=l+1\\k\neq i\\k\neq j}}^{N}D_{lk}\right]$$

$$L_{DB} = \frac{1}{2(3-2)}\left[14.5-5.5\right]$$

$L_{DB} = 4.5$

Assim, a árvore final é dada por:



3.3.3.3 Variações do *Neighbor-Joining*

Sendo o *Neighbor-Joining* um algoritmo de baixo custo e muito empregado na reconstrução de árvores filogenéticas, alguns trabalhos adicionam-lhe extensões ou modificações (BRUNO, 2000; KUMAR, 1996; PEARSON *et al.*, 1999; GASCUEL, 1997).

O trabalho desenvolvido por KUMAR (1996) mantém os mesmos cálculos efetuados pelo NJ original. O objetivo do trabalho é avaliar outras possibilidades de soluções além da árvore de evolução mínima gerada pelo NJ. O algoritmo busca árvores alternativas por meio de um conceito chamado taxon líder. A partir da matriz soma – matriz que apresenta a soma dos ramos da árvore para cada possível agrupamento de pares de taxa - escolhe-se o par que apresentar a soma mínima de comprimento de ramos. Esses dois taxa são então eleitos como possíveis taxa líderes. Para se escolher um dos dois, calcula-se a soma dos comprimentos dos ramos de cada um dos dois taxa aos outros taxa restantes. Dessa forma, haverá dois conjuntos de somas referentes às somas dos dois taxa aos outros. O taxon líder será aquele que possuir em seu conjunto de somas o menor valor dos dois conjuntos.

Depois da escolha do taxon líder, há a necessidade de se escolher o taxon que será a ele agrupado. Para isso, utiliza-se o seu conjunto de somas. O conjunto de somas do taxon líder deve ser colocado em ordem ascendente de valores. O taxon que apresentar o menor valor, ou seja, o primeiro valor do conjunto, quando agrupado ao taxon líder, será o taxon escolhido. Os outros serão considerados como vizinhos potenciais. Após a primeira bifurcação, as outras serão feitas de acordo com os vizinhos potenciais do taxon líder contidos no seu conjunto de somas.

Para verificar se a árvore formada a partir de determinado agrupamento do taxon líder com um vizinho potencial é uma árvore adequada, verifica-se se o valor da sua soma é consideravelmente pequeno quando comparado ao menor valor do conjunto de somas do taxon líder. Se esse valor for maior que determinado limite, é improvável que a árvore gerada por esse agrupamento tenha um valor de soma menor do que o valor da árvore gerada pelo agrupamento do taxon líder ao vizinho que gerou a soma mínima. Nesse caso, essa árvore é descartada. Dessa forma, o espaço de árvores a serem analisadas é reduzido.

A cada possível agrupamento do taxon líder a um vizinho potencial, uma nova árvore é gerada. Essa árvore é submetida ao teste de heurística e, se o valor da soma dos seus ramos for menor que o limite estabelecido, a árvore é considerada como possível solução para a próxima iteração do algoritmo. Ao final de todas as iterações, a árvore escolhida será a que apresentar o menor valor de soma dos ramos.

PEARSON *et al.* (1999) propõem uma generalização do *Neighbor-Joining* clássico. Esse trabalho aponta como ponto vulnerável do NJ a apresentação de uma única solução. O objetivo do trabalho é o de apresentar várias árvores alternativas como possíveis soluções. Para isso, o algoritmo explora o espaço de topologias apresentado a cada iteração do algoritmo, por meio da seleção das soluções mais promissoras e de sua apresentação para a próxima iteração do algoritmo.

Dessa forma, o algoritmo mantém um conjunto de soluções alternativas a cada iteração. As soluções alternativas não necessariamente representam ótimos locais, mas possuem um potencial de minimização das funções dos quadrados mínimos e evolução mínima para as próximas iterações. O número de soluções K é um parâmetro de entrada do algoritmo.

O algoritmo pode selecionar de várias formas as K soluções a cada iteração. Uma estratégia simples é a de escolher as K soluções de menor custo, ou seja, as K soluções que apresentam os menores valores para a função evolução mínima. Dessa forma, as K árvores que apresentarem os menores valores para a soma de seus ramos são escolhidas. Outra estratégia apontada pelo trabalho é a escolha aleatória de soluções. A fim de manter soluções com qualidade e diversidade, o algoritmo implementa um esquema híbrido que

equilibra os dois extremos: as melhores soluções em termos de custo e soluções que apresentem diversidade topológica.

O algoritmo é chamado *Generalized Neighbor-Joining* e pode ser dividido em duas partes principais: um componente que gera várias árvores alternativas utilizando o mesmo procedimento do NJ original e um segundo componente composto por uma função de avaliação que escolhe determinadas soluções de acordo com o critério que equilibra soluções de baixo custo e soluções que apresentam diversidade topológica.

Uma outra extensão do NJ é feita por GASCUEL (1997). Ao invés de tentar encontrar soluções alternativas àquela fornecida pelo NJ original, o princípio básico do algoritmo é reconsiderado. Uma modificação das fórmulas originais é feita levando em consideração as características dos dados biológicos de entrada.

O algoritmo é denominado BIONJ e trata as distâncias evolutivas obtidas do alinhamento de seqüências. Essas distâncias, contidas na matriz de distâncias de entrada, são multiplicadas por um fator λ a cada agrupamento de pares executado pelo algoritmo. O fator tem o objetivo de minimizar a variância dos valores da matriz de distâncias de entrada. Dessa forma, o algoritmo tenta melhorar os valores disponíveis na matriz de maneira que a escolha de pares a serem agrupados seja executada mais adequadamente. Como o algoritmo é iterativo, essa minimização é cumulativa para as próximas iterações.

Nesse trabalho, a tentativa de melhoria do NJ é feita diretamente a partir dos dados de entrada, diferentemente da abordagem anterior, que aponta para diferentes soluções como uma melhor aplicação do NJ.

Capítulo 4

Consenso e Medidas de Distância

Resumo – Este capítulo pretende discutir procedimentos comumente adotadas na reconstrução de árvores filogenéticas, mais especificamente a busca de uma árvore de consenso e medidas de distância entre árvores filogenéticas. A clara relação entre os assuntos acontece pela necessidade de se medir a distância entre árvores filogenéticas para permitir realizar análises comparativas entre elas e extrair uma única topologia a partir de múltiplas propostas de topologia. Os principais métodos de consenso e medidas de distância são brevemente descritos.

4.1 Consenso

Um dos problemas dos algoritmos que constroem árvores filogenéticas é o fato de poderem existir várias árvores com diferentes topologias como resultado da execução desses algoritmos (HEIN *et al.*, 1996). Reconstruir a árvore filogenética correta para um conjunto de espécies é ainda um dos problemas fundamentais da Genética Evolutiva (DASGUPTA *et al.*, 1997; DASGUPTA *et al.*, 1998). Os algoritmos podem levar a diferentes árvores para um mesmo conjunto de atributos das espécies. Os métodos de verossimilhança máxima e *Neighbor-Joining* são alguns desses algoritmos, os quais adotam critérios e sistemáticas distintas junto ao processo de reconstrução.

Além disso, para os métodos que recorrm a uma busca exploratória no espaço de todas as topologias possíveis, obter a melhor árvore para um dado conjunto de dados é um problema NP-completo (BILLERA *et al.*, 2001). O método de verossimilhança máxima, por exemplo, requer que se busque a melhor topologia entre todas as existentes. O problema está no crescimento fatorial do espaço de árvores a serem verificadas com o aumento do número de folhas. Esse aumento faz com que existam, por exemplo, para um número de 25 espécies, $1{,}19 \times 10^{30}$ árvores a serem verificadas.

Para contornar esse problema, o *Neighbor-Joining* utiliza um algoritmo de busca gulosa apresentado no Capítulo 3. No caso do *Neighbor-Joining* clássico, apenas uma árvore é apresentada ao final da execução do algoritmo e não há garantias de que essa árvore seja a que melhor representa o compromisso entre os critérios de evolução mínima e de quadrados mínimos. Em KUHNER & FELSENSTEIN (1994), um estudo comparativo do desempenho desses métodos para um mesmo conjunto de dados pode ser encontrado.

Diferentes árvores podem também ser geradas pelo uso de um mesmo método, apenas variando-se os atributos que caracterizam as espécies em análise. Em HEIN (1990), um estudo sobre diferentes árvores filogenéticas resultantes da recombinação de seqüências homólogas pode ser encontrado.

Dado que diferentes algoritmos para um mesmo conjunto de dados ou diferentes conjuntos de dados executados por um mesmo algoritmo podem produzir diferentes árvores filogenéticas, existe a necessidade de se decidir sobre a escolha da melhor árvore entre o conjunto de árvores resultantes.

Essa árvore, chamada de árvore de consenso, não é propriamente escolhida entre um conjunto de árvores candidatas, mas construída a partir da análise de todas as diferentes árvores resultantes. A árvore de consenso pode ser definida como aquela que contém a maior quantidade possível de informações que são comuns a todas as árvores que foram geradas a partir do mesmo conjunto de dados e que contêm o mesmo conjunto de *n* folhas (FELSENSTEIN, 2004). Do ponto de vista matemático, uma árvore de consenso seria um mapeamento ou uma função que transforma várias árvores de entrada com o mesmo conjunto de *n* folhas em uma única árvore de saída com *n* folhas (BRYANT, 2003).

Existem várias técnicas que podem ser escolhidas para a obtenção da árvore de consenso. Algumas delas serão discutidas neste capítulo.

Antes da apresentação das técnicas, serão definidos alguns termos que serão utilizados ao longo do capítulo. A denominação **grupos** será utilizada para árvores com raiz. Para

árvores sem raiz, as unidades de divisão e comparação das árvores serão chamadas de **partições.**

As diferentes árvores que podem resultar da aplicação de determinada técnica de reconstrução de árvores filogenéticas (ou de diferentes técnicas) tanto para um mesmo banco de dados, ou para bancos de dados diferentes, serão chamadas de **árvores de contexto.** A idéia é que árvores de contexto necessitam da aplicação de alguma técnica de consenso para que estudos filogenéticos posteriores possam ser efetuados com maior grau de confiabilidade junto às informações evolutivas.

4.1.1 Consenso estrito

Essa técnica constrói a árvore filogenética de consenso por meio da reunião de todos os grupos ou partições que são comuns a todas as árvores de contexto. É considerado o mais simples de todos os métodos de consenso (FELSENSTEIN, 2004) e não leva em conta o comprimento dos ramos. Considere as seguintes árvores com raiz apresentadas na Figura 4.1.

Figura 4.1 - Duas topologias distintas de árvore com raiz e quatro folhas

Os grupos {a,b,c,d} e {a,b,c} aparecem em ambas as árvores. Assim, a árvore de consenso é dada pela árvore da Figura 4.2.

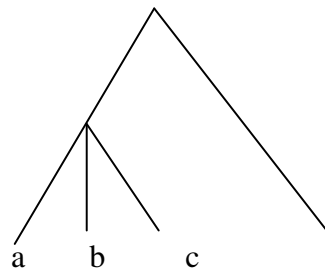


Figura 4.2 - Árvore resultante do método de consenso estrito entre as árvores da Figura 4.1

4.1.2 Consenso estrito utilizando comprimento dos ramos

Uma generalização natural do consenso estrito é considerar os pesos ou comprimentos dos ramos das árvores de contexto. O comprimento considerado na árvore de consenso é o comprimento mínimo de cada ramo pertencente ao grupo ou partição comum.

Figura 4.3 - Duas topologias distintas de árvore com raiz e quatro folhas, incluindo o comprimento dos ramos

Para as árvores da Figura 4.3, aplicando a técnica de consenso estrito e considerando o comprimento dos ramos, a árvore de consenso seria aquela mostrada na Figura 4.4.

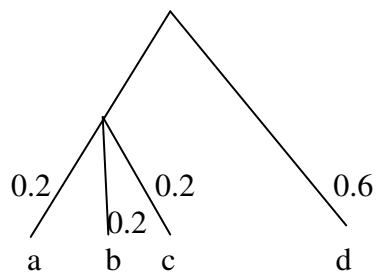


Figura 4.4 - Árvore resultante do método de consenso estrito entre as árvores da Figura 4.3, considerando o comprimento dos ramos

4.1.3 Consenso da regra majoritária

A árvore de consenso é construída considerando grupos ou partições que são comuns a uma determinada porcentagem das árvores de contexto. Essa porcentagem recebe o nome de regra majoritária (FELSENSTEIN, 2004). Seja uma árvore de consenso com, por exemplo, 60% de regra majoritária. Para reconstruir essa árvore, há a necessidade de se obter grupos ou partições comuns a mais de 60% do total das árvores. Considere, como exemplo, o seguinte conjunto de três árvores com raiz, conforme apresentado na Figura 4.5.

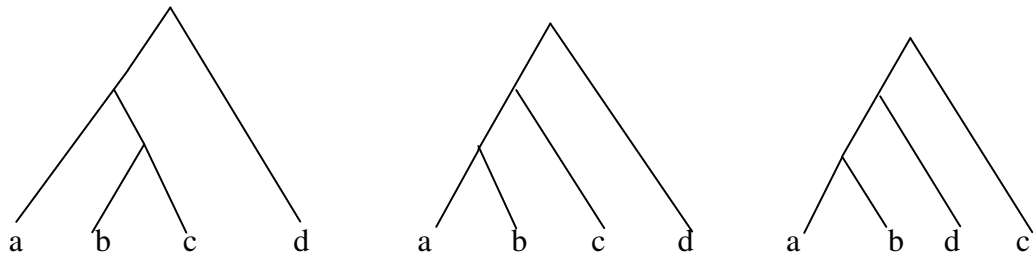


Figura 4.5 - Três topologias distintas de árvores com raiz e quatro folhas

Os grupos {a,b}, {a,b,c} e {a,b,c,d} aparecem em duas das três árvores da Figura 4.5. Dessa forma, a árvore com 66% de regra majoritária é dada pela árvore (((a,b),c),d) apresentada na Figura 4.6.

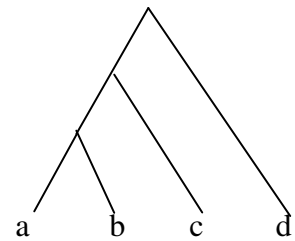


Figura 4.6 - Árvore resultante da técnica de consenso com 66% de regra majoritária para as árvores da Figura 4.5

Uma conclusão direta é a de que árvores de consenso com 100% de regra majoritária são árvores de consenso estrito.

4.1.4 Consenso de regras majoritárias considerando comprimento de ramos

Uma generalização da técnica de consenso de regras-majoritárias seria o uso dos comprimentos dos ramos das árvores de contexto para o cálculo da árvore de consenso. A técnica consiste em calcular o comprimento de determinado ramo utilizando a média de todos os comprimentos daquele ramo nas árvores de contexto. FELSENSTEIN (2004) sugere que, quando não houver o ramo para determinada árvore, seu comprimento deve ser considerado como zero para efeito do cálculo da média.

4.1.5 Consenso de ADAMS

A técnica de consenso de Adams é a mais antiga das técnicas de consenso para árvores e, talvez em razão disso, uma das mais populares (BRYANT, 2003). Antes de explicar a técnica, há a necessidade de se definir dois conceitos: produto e grupos máximos.

Sejam P_1, P_2, ... ,P_k todas as partições de um conjunto de espécies. Um exemplo do conceito produto dessas partições é a partição P para a qual duas espécies *a* e *b* estão no mesmo bloco se e somente se elas estão no mesmo bloco para cada partição P_1, P_2, ... ,P_k. Por exemplo, o produto de ab|cde e ac|bde é a|b|c|de, pois os elementos d e e estão no mesmo bloco de ambas as partições ab|cde e ac|bde. A árvore de consenso construída pela técnica de Consenso de Adams é formada pelos produtos que não são contraditórios nas árvores sendo analisadas. A árvore de consenso é formada pela busca de todos os produtos que não entram em contradição com a topologia de nenhuma outra árvore, ou seja, o produto existe em todas as árvores de contexto. Esses produtos formam a árvore de consenso.

Os grupos máximos de uma árvore de entrada T_i são os maiores grupos no conjunto de árvores de contexto T. A partição de grupo máximo para T_i é a partição $P(T_i)$ do conjunto de espécies com blocos iguais aos grupos máximos de T_i . A árvore de consenso é obtida pela formação recursiva de partições P para T, considerando as restrições dos grupos máximos. Considere, por exemplo, as duas árvores apresentadas na Figura 4.7.

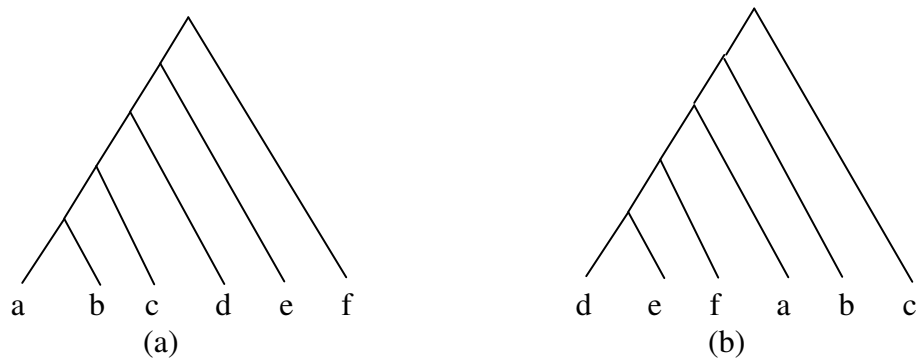


Figura 4.7 - Duas topologias distintas para uma árvore com raiz e seis folhas

A partição de grupo máximo de T1 é abcde|f e a partição de grupo máximo de T2 é defab|c e o produto dessas duas partições é abde|c|f. Dessa forma, a árvore de Adams tem três grupos máximos: {a,b,d,e} , {c} e {f}. As restrições de T1 e T2 a {a,b,d,e} são: (((a,b),d),e) e (((d,e),a),b), respectivamente. As árvores que representam as restrições são mostradas pelas Figuras 4.8a e 4.8b. As partições de grupo máximo dessas restrições são adb|e e ade|b as quais têm o produto ad|b|e. Sendo assim, a árvore de consenso de Adams para T1 e T2 é dada por: (((a,d),b,e),c,f) conforme Figura 4.9.

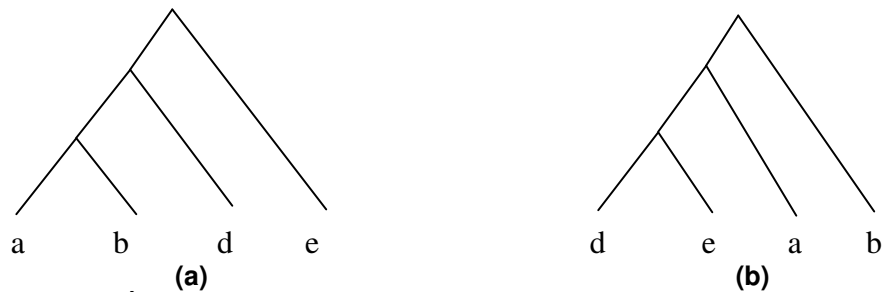


Figura 4.8 - (a) Árvore resultante das restrições da árvore da Figura 4.7a ao grupo {a,b,d,e}. (b) - Árvore resultante das restrições da árvore da Figura 4.7b ao grupo {a,b,d,e}

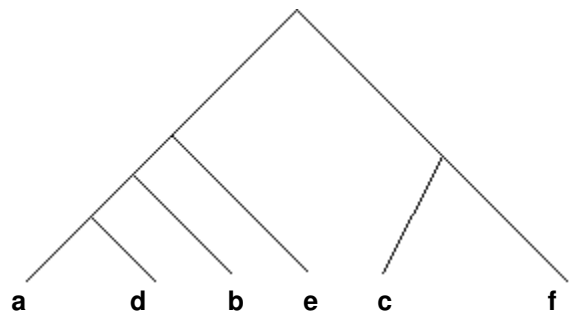


Figura 4.9 - Árvore de consenso a partir das árvores da Figura 4.7 utilizando a técnica de consenso de Adams

4.2 Considerações sobre as técnicas de consenso

Nem todas as técnicas são apropriadas para todas as situações. Um aspecto que pode dificultar a utilização de determinada técnica é a existência ou não existência de raiz em uma árvore. Algumas técnicas como, por exemplo, o Consenso de Adams não são passíveis de aplicação em árvores sem raiz. Isto, certamente, é um fator limitante à aplicação da técnica.

O problema da extensão para os casos de árvores sem raiz, a partir de técnicas adequadas a árvores com raiz, é a variação do resultado final da árvore de consenso. Dependendo do ramo ao qual a raiz for associada, podem-se obter diferentes resultados de árvores de consenso (BRYANT, 2003). Seria gerado, a partir daí, um problema adicional que seria o da análise dos diferentes resultados de árvores dependendo da colocação da raiz nas mesmas.

Outro aspecto importante a ser considerado para as técnicas de geração de árvore de consenso é a análise do comprimento dos ramos das árvores de contexto. Muitas técnicas levam em consideração apenas os aspectos topológicos das árvores, sem considerar a distância que separa a evolução dos taxa (FELSENSTEIN, 2004).

De acordo com BRYANT (2003), o problema das técnicas de consenso não está na escolha de alguma das técnicas, mas na maneira como as árvores resultantes da aplicação de técnicas de consenso têm sido interpretadas.

Em STEEL *et al.* (2000), pode-se encontrar uma análise sobre alguns aspectos limitantes das técnicas de consenso. O autor também apresenta algumas propriedades básicas que deveriam ser atendidas pelas técnicas de consenso. As três principais são:
P1- O método de consenso pode ser aplicado a qualquer conjunto não ordenado de árvores de entrada.
P2- Se todas as espécies são renomeadas e o método de consenso é aplicado às novas árvores de entrada, a árvore de consenso deve ser a mesma árvore anterior, porém, com as espécies renomeadas de acordo com a mudança efetuada.
P3- Se as árvores de entrada são compatíveis, então a árvore de saída é uma árvore que pode dar origem a todas as outras, uma vez que o método de consenso deve selecionar a árvore que atinge essa compatibilidade.

A propriedade P1 requer que, caso haja grau de confiança igual nas árvores de entrada, então a ordem de entrada na qual as árvores são submetidas ao algoritmo não importa. A propriedade P2 requer que a maneira pela qual as espécies são nomeadas ou rotuladas não afete a árvore de saída. A propriedade P3 requer que as árvores de entrada sejam

compatíveis e que o método conduza a uma árvore de consenso que seja compatível com todas as árvores de contexto.

Ainda segundo STEEL *et al.* (2000), alguns métodos como o Consenso das Regras Majoritárias e o Consenso de Adams satisfazem as duas primeiras propriedades mencionadas acima. O problema apontado é que, quando as árvores não são compatíveis, esses métodos freqüentemente resultam em árvores não-resolvidas (do tipo estrela), e até mesmo relacionamentos filogenéticos que eram compartilhados pelas árvores de contexto podem desaparecer na árvore de consenso, ou seja, P3 não é satisfeita.

4.3 Medidas de Distância

Encontrar informações comuns entre árvores filogenéticas é um assunto que pode estar relacionado a encontrar diferenças ou medidas de distância entre as árvores. Há casos em que o conhecimento da medida de distência entre árvores é mais significativo do que encontrar a árvore de consenso.

Da mesma forma que para árvores de consenso, vários métodos para se medir a distância entre árvores filogenéticas têm sido apresentados. A seguir, um dos mais adotados na literatura é apresentado: distância de Robinson-Foulds.

4.3.1 Distância de Robinson-Foulds sem comprimento de ramos

Robinson e Foulds (R-F) (ROBINSON & FOULDS, 1981) desenvolveram uma medida de distância para árvores filogenéticas motivados pelo problema de árvores diferentes serem produzidas por métodos distintos de construção quando aplicados ao mesmo conjunto de dados. Esta medida ficou conhecida como diferença simétrica ou métrica da partição.

Essa métrica consiste em dividir uma árvore filogenética sem raiz em duas partições. Para dividir a árvore, qualquer ramo ou aresta pode ser escolhido. Essa aresta divide a árvore em dois conjuntos de nós, ou novas árvores, cada uma conectada ao ponto final da aresta. Essa divisão é feita para todas as arestas existentes nessa árvore.

A diferença simétrica entre as árvores se dá pela soma dos módulos das diferenças entre os comprimentos dos ramos correspondentes em cada árvore. Quando um ramo existe em uma árvore e não existe em outra, o comprimento do ramo inexistente é considerado zero. Considere a árvore T_1 da Figura 4.10. Essa árvore pode gerar as seguintes partições: {ADF | BCEG}, {DF | ABCEG}, {BC | ADEFG}, {EG | ABCDF}, {A | BCDEFG}, {B | ACDEFG}, {C | ABDEFG}, {D | ABCEFG}, {E | ABCDFG}, {F | ABCDEG}, {G | ABCDEF}. Repare que o número de partições é igual ao número de ramos da árvore.

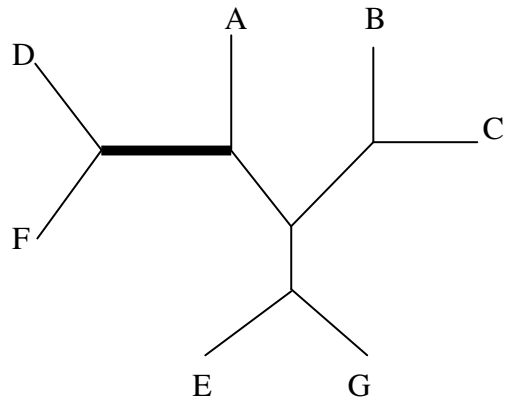


Figura 4.10 - Árvore filogenética T_1 (FELSENSTEIN, 2004)

Considere agora uma outra árvore T_2, similar à arvore T_1 conforme Figura 4.11. As partições que podem ser geradas a partir de T_2 são: {ADF | BCEG}, {AD | BCEFG}, {BC | ADEFG}, {EG | ABCDF}, {A | BCDEFG}, {B | ACDEFG}, {C | ABDEFG}, {D | ABCEFG}, {E | ABCDFG}, {F | ABCDEG}, {G | ABCDEF}.

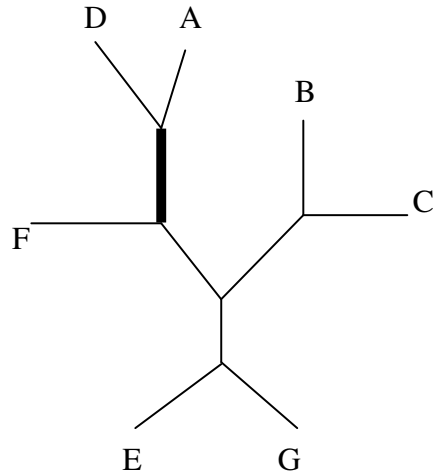


Figura 4.11 - Árvore filogenética T_2 (FELSENSTEIN, 2004)

A diferença simétrica entre essas árvores está representada pelos ramos destacados em T_1 e T_2, os quais, quando escolhidos para dividir as árvores, geram as seguintes partições: {DF | ABCEG} em T_1 e {AD | BCEFG} em T_2. Dessa forma, a distância R-F entre as duas árvores é dada por $D_{R-F} = 2$. O trabalho original da métrica R-F descreve a mesma métrica apresentada acima definindo-a como a aplicação de operações elementares em árvores filogenéticas. Essas operações podem transformar uma árvore em outra. As operações são definidas como Contração e Dilatação. Considerando novamente a árvore T_1 da Figura 4.10, poderíamos aplicar uma operação de contração à aresta que liga a partição {D,F} à partição {A} resultando na árvore T_3 conforme mostra a Figura 4.12.

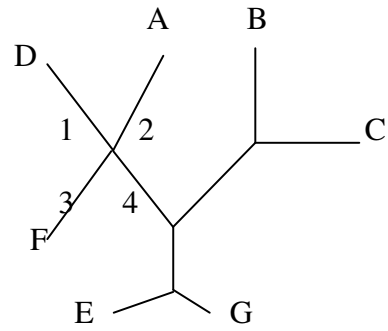


Figura 4.12 - Árvore filogenética T_3

Aplicando-se à árvore T_3 uma operação de dilatação mantendo ligadas as arestas 1 e 2 e separando-as das arestas 3 e 4 pela colocação de uma nova aresta, poderíamos obter a árvore T_4 conforme mostra a Figura 4.13. Pode-se observar que as árvores T_2 e T_4 têm a mesma tolopologia, ou seja, a árvore T_1 foi transformada na árvore T_2 pela aplicação de uma contração seguida de uma dilatação, produzindo $D_{R-F} = 2$.

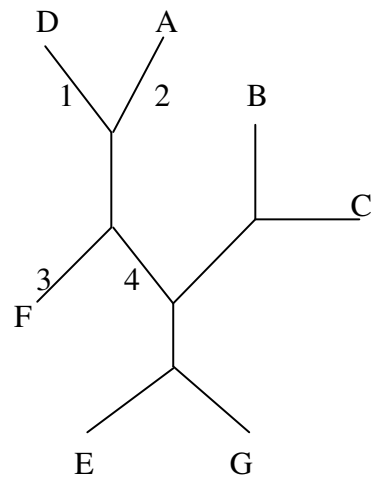


Figura 4.13 - Árvore filogenética T_4

4.3.2 Distância de Robinson-Foulds com comprimento de ramos

A técnica R-F, conforme mencionada acima, pode ser aplicada a árvores filogenéticas para medir a distância entre as árvores de acordo com a topologia das mesmas, sem considerar o comprimento dos ramos. Porém, a mesma medida de distância de R-F pode ser aplicada a árvores filogenéticas considerando o comprimento dos ramos como parte da diferença entre elas. Considere novamente as árvores T_1 e T_2, porém, agora com comprimentos atribuídos aos seus ramos, conforme Figura 4.14.

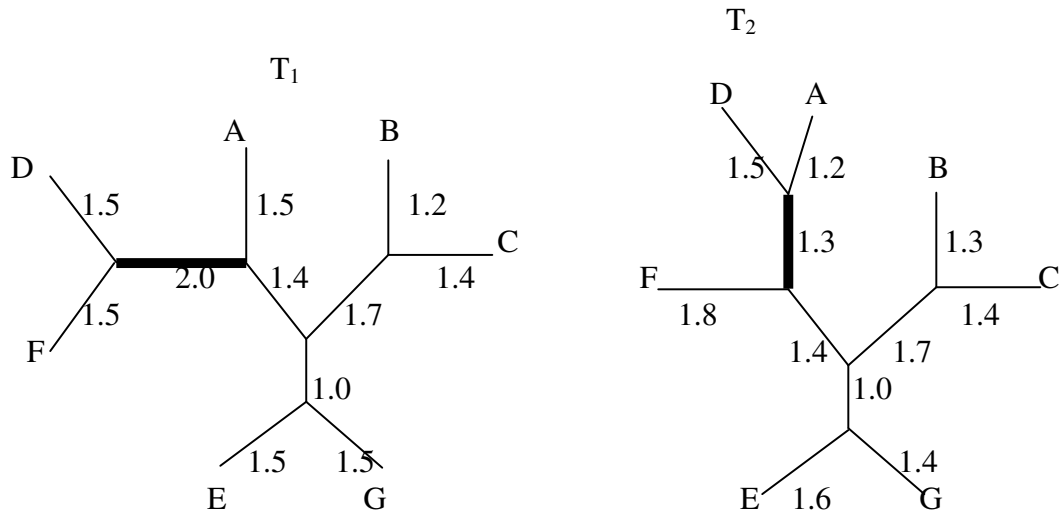


Figura 4.14 - Árvores filogenéticas com ramos valorados

As partições relativas às duas árvores e o comprimento dos ramos são apresentados na Figura 4.15. Quando a partição em uma árvore não tem correspondente na outra, o ramo é indicado como tendo valor zero. A distância de Robinson-Foulds é calculada pelas somas dos valores absolutos das diferenças entre as colunas na Figura 4.15.

O resultado da distância R-F para as árvores T_1 e T_2 da Figura 4.14, ignorando os comprimentos que são iguais nas duas árvores, é dado por:

$$d_{R\text{-}F} = |0 - 1.3| + |2 - 0| + |1.5 - 1.2| + |1.2 - 1.3| + |1.5 - 1.6| + |1.5 - 1.8| + |1.5 - 1.4| = 4.2$$

Tabela 4.1 - Comprimento dos ramos das árvores T_1 e T_2 necessários para o cálculo da distância R-F

Partições	Comprimento dos Ramos	
	T_1	T_2
{ADF | BCEG}	1.4	1.4
{AD | BCEFG}	0	1.3
{BC | ADEFG}	1.7	1.7
{DF | ABCEG}	2.0	0
{EG | ABCDF}	1.0	1.0
{A | BCDEFG}	1.5	1.2
{B | ACDEFG}	1.2	1.3
{C | ABDEFG}	1.4	1.4
{D | ABCEFG}	1.5	1.5
{E | ABCDFG}	1.5	1.6
{F | ABCDEG}	1.5	1.8
{G | ABCDEF}.	1.5	1.4

4.3.3 *Nearest Neighbor Interchange* sem comprimento de ramos

A medida de distância Nearest Neighbor Interchange (NNI) é determinada por meio do número mínimo de operações NNI efetuadas em uma dada árvore T_1 para transformá-la em uma outra árvore T_2 (WATERMAN, 1978) . Uma operação NNI pode ser definida como uma operação que troca duas sub-árvores que são separadas por uma aresta interna. As Figuras 4.15b e 4.15c mostram as duas operações NNI possíveis de serem realizadas na Figura 4.15a. Na Figura 4.15, cada letra A, B, C e D representa uma sub-árvore descendente dos nós internos u ou v.

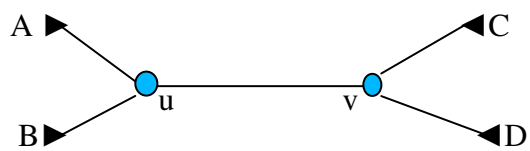


Figura 4.15(a) - Árvore com aresta interna (u,v)

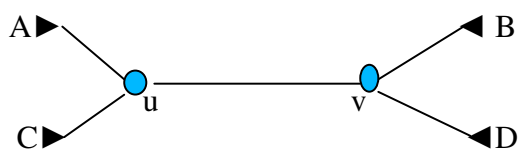


Figura 4.15(b) - Árvore com troca efetuada efetuada entre sub-árvores B e C

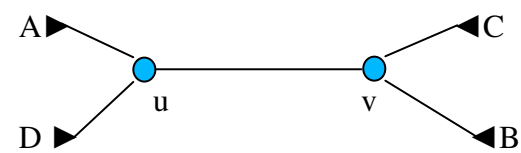


Figura 4.15(c) - Árvore com troca entre sub-árvores B e D

Nesse exemplo, a distância NNI é dada por $D_{NNI}=2$, o que significa que a distância entre as árvores da Figura 4.15a e 4.15b é dada por $D_{NNI}=2$ e é resultante da mudança entre as sub-árvores B e C; e a distância entre as árvores das Figura 4.15a e 4.15c é $D_{NNI}=2$ e é resultante da mudança entre as sub-árvores B e D.

4.3.4 *Nearest Neighbor Interchange* considerando comprimento de ramos

A extensão da distância NNI considerando os comprimentos dos ramos pode ser feita tomando o custo da operação como sendo igual ao peso da aresta ou ramo na qual a operação está sendo efetuada. Esse custo é considerado para casos como os da Figura 4.15.

Uma das dificuldades da distância NNI é que, para árvores extensas, que contêm muitas diferenças, o algoritmo é computacionalmente intratável (FELSENSTEIN, 2004). A demonstração de que o problema de se calcular a distância NNI para árvores extensas é NP-completo pode ser encontrada em DASGUPTA *et al.* (1997).

4.3.5 Distância de quartetos

A técnica conhecida como distância de quartetos (do inglês, *quartet distance*) é medida por meio da contagem de quartetos semelhantes nas árvores filogenéticas. Um quarteto é uma sub-árvore com topologia de quatro espécies. As topologias possíveis para um quarteto são mostradas na Figura 4.16. A distância entre duas árvores filogenéticas é dada pelo número de diferentes quartetos existentes entre as árvores. Essa distância pode ser obtida pela comparação dos quartetos um por um. Esse algoritmo tem complexidade $O(n^4)$, sendo n o número de folhas. Em BRYANT (2003), um algoritmo da ordem $O(n^2)$ para a técnica distância de quartetos pode ser encontrado.

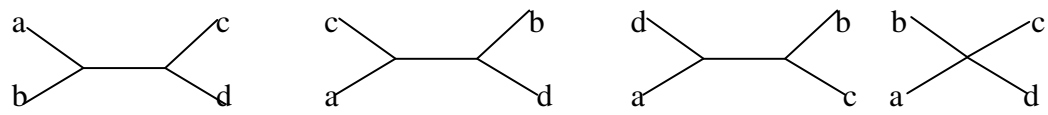


Figura 4.16 - As quatro possíveis topologias de quarteto das espécies a,b,c e d (BRODAL *et al.*, 2004)

Capítulo 5

Multi-Neighbor Joining

Resumo – A proposta principal deste capítulo é apresentar o algoritmo *Multi-Neighbor Joining*. Conforme visto no capítulo 3, o *Neighbor-Joining*, embora seja um bom algoritmo de reconstrução de árvores filogenéticas, tem a desvantagem de proporcionar apenas uma árvore como solução. Dado que essa árvore pode não ser a de evolução mínima, podem existir no conjunto de soluções várias possíveis árvores. O *Multi-Neighbor Joining* é um algoritmo capaz de executar múltiplas decisões de agrupamento a cada nível de reconstrução da árvore, mantendo várias possíveis soluções parciais ao longo da execução recursiva do *Neighbor-Joining*. Dessa forma, várias possíveis árvores são apresentadas ao final da execução do algoritmo que aumentando a chance de apresentação da árvore de evolução mínima, e mantendo o baixo custo de reconstrução do algoritmo *Neighbor-Joining*.

5.1 Introdução ao algoritmo

Baseando-se nos atributos evolutivos extraídos de espécies sob análise, a proposta do modelo de árvore filogenética para descrever os eventos evolutivos responsáveis pela observância de dissimilaridades entre espécies é geralmente precedida pela seleção de um método apropriado para a reconstrução da árvore.

Conforme já discutido no Capítulo 3, várias abordagens usadas para a inferência de topologias já foram propostas na literatura, e não há evidência clara de superioridade principalmente quando diferentes perspectivas de desempenho são consideradas.

Os algoritmos de reconstrução de árvores filogenéticas podem ser classificados como: 1) algorítmicos: um procedimento sistemático de reconstrução é empregado para determinar a árvore; 2) baseados em busca: um critério de otimização é definido para comparar as topologias alternativas, e cada árvore representa um ponto no espaço de busca a ser

explorado. Os procedimentos algorítmicos tentam simultaneamente definir a topologia e alcançar os objetivos do critério de otimização, seguindo um critério de desempenho.

Os critérios baseados em busca adotam um paradigma distinto, composto de dois passos: um para a avaliação das topologias, e outro para a determinação das topologias com a maior avaliação entre as candidatas, fazendo uso de estratégias de busca computacionais.

As principais desvantagens dos procedimentos algorítmicos são a possibilidade do algoritmo ficar preso em um mínimo local e a apresentação de uma única topologia como solução. A maior vantagem é o custo computacional reduzido, com o número de antecessores a serem determinados sendo linearmente associado ao número de objetos sob análise. Esses métodos incorporam todos os métodos baseados em distância, incluindo o UPGMA (do inglês, *Unweighted Pair Group Method withArithmetic Mean*) e o *Neighbor-Joining*.

Para o caso dos procedimentos baseados em busca, o custo computacional é a maior desvantagem, devido ao aumento fatorial do número de topologia candidatas com o aumento do número de objetos sendo analisados. Os aspectos vantajosos são: o fato de se poder evitar mínimos locais e a capacidade de se apresentar várias topologias de alta qualidade ao invés de apenas uma.

O *Neighbor-Joining* (NJ), como um exemplo de procedimentos algorítmicos para a reconstrução de árvores, encontra a única árvore que representa perfeitamente os dados (GASCUEL, 1997), somente se a matriz de distâncias dada como entrada obedecer à propriedade de aditividade (BARTHÉLEMY & GUÉNOCHE, 1991). A propriedade de aditividade, descrita na seção 3.2, confere ao algoritmo NJ a certeza de que a árvore que minimiza o método dos quadrados mínimos será encontrada. Para esses casos, a diferença entre a matriz de distâncias da árvore gerada (patrística) e a matriz observada é zero.

Mesmo quando a condição de aditividade não se verifica, soluções de alta qualidade produzidas pelo NJ são encontradas (SAITOU & IMANISHI, 1989; KUHNER, 1994;

HUELSENBECK, 1995). Além disso, resultados teóricos obtidos por ATTESON (1996) indicam que o NJ é tão eficiente quanto possível, em termos do que é possível realizar a partir de uma busca gulosa.

Dessa forma, tomando o espaço de soluções possíveis, a natureza heurística gulosa das decisões adotadas pelo NJ leva a uma exploração apenas parcial do espaço de soluções, mantendo seu custo relativamente baixo. Entretanto, a ausência de uma estratégia de busca mais exploratória impede a proposição de topologias distintas e capazes de atender bem aos critérios de quadrados mínimos e evolução mínima. As topologias alternativas podem se adequar aos dados de entrada tão bem quanto ou até mesmo de uma maneira melhor do que a única solução considerada pelo *Neighbor-Joining*.

Para aumentar a confiabilidade dos processos de reconstrução de árvores filogenéticas, há a necessidade de se aproveitar os baixos custos computacionais dos processos algorítmicos (SALEMI, 2003) e associá-los a procedimentos de produção de árvores alternativas (HOLDER & LEWIS, 2003; HOLMES, 2002) como resultado do algoritmo.

Com o objetivo de contribuir para tanto, uma extensão do NJ foi proposta neste trabalho (DA SILVA *et al.*, 2004; DA SILVA *et al.*, 2005). Se o NJ gerasse várias soluções ao invés de uma única, poder-se-ia observar a possibilidade de sub-árvores mais estáveis que outras, ou seja, em um conjunto de várias soluções possíveis, podería-se encontrar sub-árvores que se repetissem em várias soluções. Isso conferiria à essa sub-árvore maior grau de estabilidade quando comparada a outras que aparecessem em menor número de vezes nas diversas soluções. Além disso, não há garantia de que a árvore de evolução mínima é sempre obtida pelo NJ clássico, pois o fato do algoritmo escolher sempre a melhor opção (de acordo com critério de evolução mínima) a cada iteração, não garante a obtenção do ótimo global, pois seu processo de reconstrução pode convergir para um ótimo local, conforme discutido na seção 3.2.3.

Dessa forma, a nova proposta se presta a tomar decisões a cada nível de reconstrução da árvore, escolhendo vários pares a sofrerem junção e mantendo várias soluções parciais durante a execução recursiva do NJ.

As principais vantagens do novo algoritmo são:1) mantém o baixo custo de reconstrução do NJ clássico (complexidade $O(n^3)$, onde n é o número de folhas da árvore; 2) a chance de se alcançar a árvore de evolução mínima é maior; 3) topologias com desempenhos similares são mostradas ao final da execução do algoritmo.

Essa extensão do NJ clássico é chamada de *Multi-Neighbor Joining* (MNJ), uma vez que a saída do algoritmo não se restringe a uma única árvore final, dado que junções de pares distintos podem também produzir árvores de boa qualidade. A ocorrência de decisões de criação de mais pares de junções é controlada por um limiar fornecido como parâmetro de entrada do algoritmo. Esse limiar é uma porcentagem que permite que outras possibilidades de árvores, além da árvore de evolução mínima, sejam consideradas como possíveis soluções, desde que seus valores de soma dos comprimentos dos ramos não ultrapassem tal porcentagem em relação à única árvore de evolução mínima encontrada a cada iteração do NJ clássico.

Essencialmente, ao invés de selecionar um único par a ser agrupado, aquele que se supõe corresponda à reconstrução da árvore mínima, pares alternativos serão considerados respeitando o limiar dado como parâmetro. Esse limiar foi estabelecido como sendo 2% acima da soma total dos comprimentos dos ramos da árvore corrente sendo construída com base no NJ clássico.

Conseqüentemente, como o NJ tem que desempenhar n-2 junções, haverá n-3 possibilidades de geração de topologias alternativas de árvores. As árvores escolhidas podem ser tão próximas da árvore fornecida pelo NJ clássico quanto for o limiar adotado pelo algoritmo, sempre considerando a configuração corrente da árvore.

As novas alternativas de árvores terão o poder de gerar suas próprias topologias alternativas nas escolhas seguintes de junções de pares. Diferentes junções de pares podem produzir as mesmas topologias finais (mesmas junções, mas em ordens distintas), indicando que apenas um subconjunto das topologias alternativas serão efetivamente distintas umas das outras.

O já mencionado trabalho de PEARSON *et al.* (1999) pode ser considerado como a primeira tentativa na direção de se produzir múltiplas topologias de árvores sob a aplicação recursiva do algoritmo NJ. Entretanto, sua abordagem é conceitualmente incompatível com a do MNJ, uma vez que seu propósito principal foi o de obter topologias maximamente diversas. Isto não quer dizer que o critério de otimização foi negligenciado, mas um papel importante foi dado à produção de diversidade.

Ao invés de ter como objetivo principal a obtenção de topologias distintas, o MNJ tenta induzir a detecção de topologias alternativas de alta qualidade. Se as soluções múltiplas, incluindo aquela associada ao NJ clássico, serão ou não topologias similares, é dependente do problema.

5.2 Passos do algoritmo MNJ

A seqüência de passos do MNJ é bastante semelhante à do NJ clássico. A diferença está na reconstrução de várias alternativas de solução. A primeira árvore gerada pelo MNJ é sempre a árvore original do NJ clássico. A seguir, os passos do MNJ são apresentados por meio de uma descrição em pseudo-código.

```
Ler matriz de distâncias M
Inicializar bif =1                    //número de bifurcações
Inicializar n=0                       //número de árvores criadas
Calcular dim                          //dimensão de M
Criar uma matriz Maux de dimensão dim
//Maux deve conter 0´s ou 1´s em suas posições de maneira a indicar as junções sendo executadas
Enquanto dim >= 3  //a matriz de distâncias não pode ter dimensão menor do que três
    Para todas as árvores sendo reconstruídas faça
        Calcular matriz soma dos comprimentos dos ramos S
        Escolher par que minimiza a soma S
        Diminuir dim em 1
        Atualizar a estrutura Maux com a junção executada
        Calcular nova matriz de distâncias com dimensão dim //O comprimento do novo nó é calculado pela
        média aritmética das distâncias do novo nó aos dois nós que sofreram junção
        Calcular comprimentos dos ramos associados ao par que saiu da estrela
        Enquanto  houver outro par a ser testado
          Se S_outropar / S_primeiropar < 1.02  // limite para aceitar uma árvore alternativa
          Então
                    Atualizar a estrutura Maux com a junção executada
                    Calcular nova matriz de distâncias com dimensão dim
                Calcular comprimentos dos ramos associados ao par que saiu da árvore-estrela
                Incrementar bif em  1
          Fim_Se
        Fim_Enquanto
        Diminuir dim  em 1
        Incrementar n em 1
    Fim_Para
Fim_Enquanto
```

A figura 5.1 mostra algumas das possíveis combinações de agrupamento e uma possível seleção de árvore alternativa. Para o exemplo, nas árvores selecionadas, as espécies B e D foram agrupadas em diferentes sub-árvores. Assim, o MNJ teria escolhido o agrupamento de A e B como primeira alternativa (solução original) e teria escolhido como solução alternativa o agrupamento de A e D. A partir da criação de duas árvores, o MNJ seguiria com as mesmas escolhendo, a partir delas, outras possíveis alternativas de agrupamento a cada nova iteração.

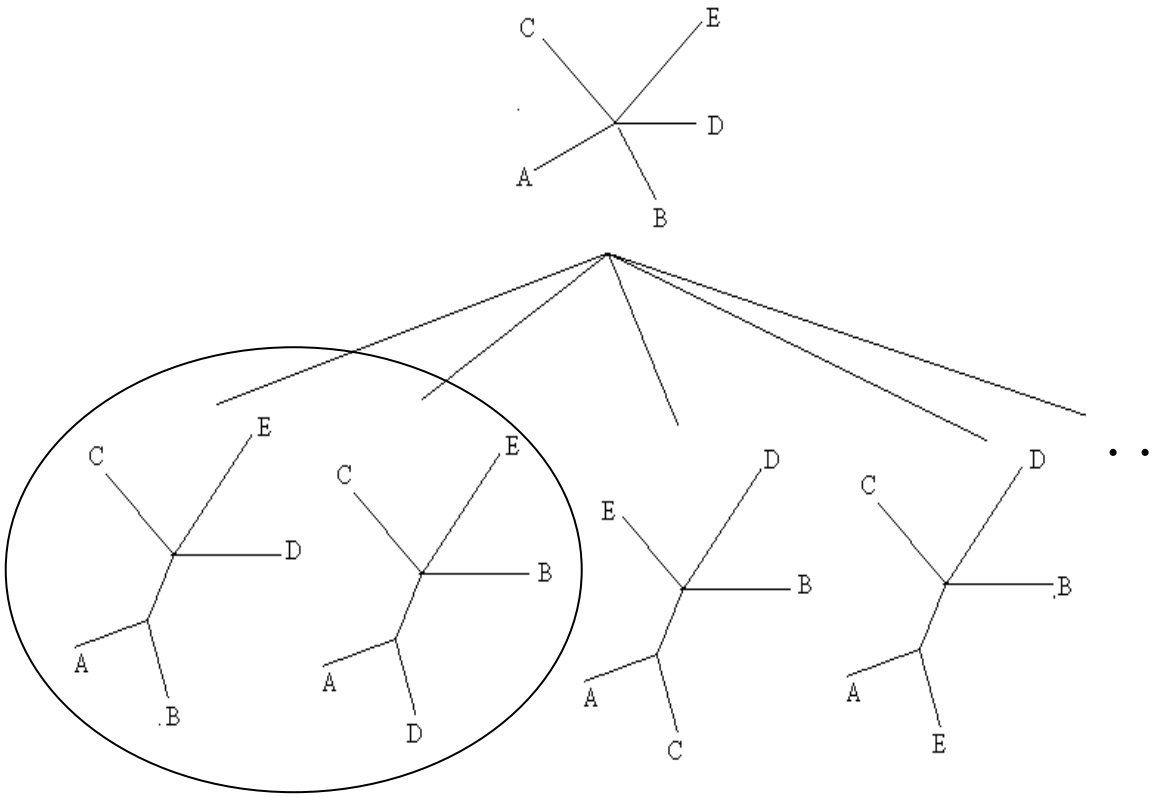


Figura 5.1 - Possíveis árvores geradas pelo MNJ

Os dois exemplos apresentados a seguir ilustram a execução do MNJ sob dois diferentes contextos. O primeiro estudo de caso é um estudo didático que apresenta uma matriz com oito espécies. O segundo é um estudo de caso real com dados de um experimento realizado com roedores. Nos dois casos, o MNJ apresentou interessantes soluções alternativas.

5.3 Geração de árvores alternativas utilizando o MNJ: Um estudo de caso didático

A partir de uma matriz D de distâncias de dimensão 8×8, conforme Figura 5.2, foram obtidas, a partir da execução do MNJ, três diferentes topologias de árvores filogenéticas, conforme Figura 5.3.

$$
D = \begin{bmatrix}
0 & 5.1680 & 5.1680 & 5.1680 & 3.9520 & 3.9520 & 5.1680 & 3.9520 \\
5.1680 & 0 & 3.0400 & 3.9520 & 4.8640 & 5.7760 & 4.8640 & 3.6480 \\
5.1680 & 3.0400 & 0 & 5.1680 & 5.4720 & 5.1680 & 5.4720 & 4.5600 \\
5.1680 & 3.9520 & 5.1680 & 0 & 4.5600 & 4.5600 & 3.3440 & 3.6480 \\
3.9520 & 4.8640 & 5.4720 & 4.5600 & 0 & 3.9520 & 4.5600 & 3.6480 \\
3.9520 & 5.7760 & 5.1680 & 4.5600 & 3.9520 & 0 & 4.5600 & 4.2560 \\
5.1680 & 4.8640 & 5.4720 & 3.3440 & 4.5600 & 4.5600 & 0 & 4.5600 \\
3.9520 & 3.6480 & 4.5600 & 3.6480 & 3.6480 & 4.2560 & 4.5600 & 0
\end{bmatrix}
$$

Figura 5.2 - Matriz de distâncias observadas com 8 espécies

A Figura 5.3 apresenta as três árvores consideradas na execução do MNJ. A primeira árvore gerada pelo MNJ é sempre a árvore gerada pelo algoritmo do NJ clássico. Essa árvore apresenta a topologia apresentada pela Figura 5.3a e tem como comprimento total o valor 17.270.

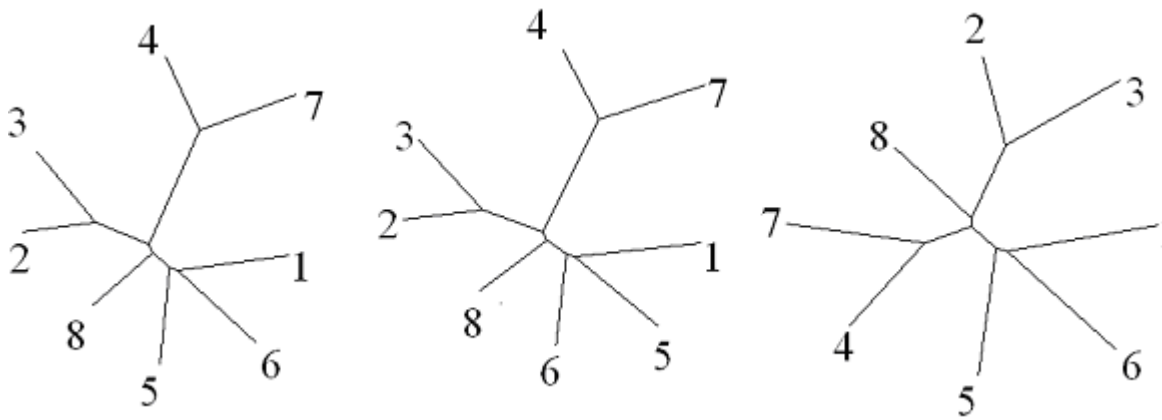


Figura 5.3 - (a) Árvore gerada pelo NJ clássico. (b) Primeira árvore alternativa gerada pelo MNJ. (c) Segunda árvore alternativa gerada pelo MNJ.

A árvore da Figura 5.3b é a primeira árvore alternativa com comprimento total 17.270. Os valores de comprimento total das árvores são iguais para as duas primeiras árvores, porém, suas topologias apresentam pequenas diferenças refletidas pelo agrupamento dos nós 1, 5 e 6.

A segunda árvore alternativa é mostrada pela Figura 5.3c. Essa árvore tem o comprimento total de ramos de 16.074. Nesse exemplo, a segunda árvore alternativa mostrou maior proximidade à árvore de evolução mínima do que a solução original. Essa diferença sugere

que a árvore da figura 5.3b poderia explicar de maneira mais adequada o agrupamento entre as oito espécies.

Além da diferença do comprimento total dos ramos, há diferenças de topologia entre as duas árvores. Na árvore da Figura 5.3a, o agrupamento das espécies 4 e 7 forma uma sub-árvore com o agrupamento das espécies 2 e 3, enquanto que na árvore da Figura 5.3c o agrupamento das espécies 4 e 7 forma uma sub-árvore com a sub-árvore formada pelo agrupamento das espécies 2 e 3 com a espécie 8. Em outras palavras, as espécies 4 e 7 estão mais próximas às espécies 2 e 3 na solução da Figura 5.3a do que estão na solução da Figura 5.3c. Essa diferença proporciona ao biólogo material de estudo para avaliar a topologia que melhor explica o agrupamento entre as espécies. A seguir, um estudo de caso real para o MNJ é apresentado.

5.4 Resultado do processamento do MNJ para dados morfológicos: Um estudo de caso real

O estudo de caso apresentado a seguir foi gerado a partir de dados fornecidos pelo grupo de pesquisa coordenado pelo prof. Dr. Sérgio Furtado dos Reis (Departamento de Parasitologia, Instituto de Biologia, Unicamp). Esse grupo coletou e estudou amostras de características morfológicas de roedores nas regiões Nordeste, Sudeste e Centro-Oeste do Brasil. O objetivo desse trabalho foi o de detectar padrões de variação em características morfológicas de roedores. Os dados coletados foram formatados sob a forma de uma matriz de distâncias a qual foi submetida ao MNJ.

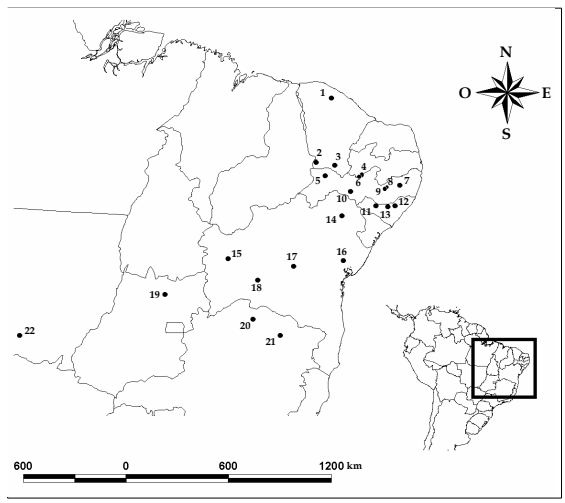
5.4.1 Caracterização do cenário de estudo

O estudo analisou caracteres morfológicos de indivíduos de uma mesma espécie de roedores (Thrichomys apereoides), localizada em diferentes regiões do Brasil (MONTEIRO *et al*., 2005). Esse estudo forneceu dados de fundamental importância para definição de fronteiras entre unidades evolucionárias na natureza.

Um passo inicial importante em reconhecer tais unidades é a identificação dos grupos ou populações que compartilham características morfológicas e continuidade geográfica ao longo de um dado espaço geográfico (CARLETON, 1988; MYERS *et al.* , 1989).

Em biologia sistêmica, informações que permitem o reconhecimento de tais unidades têm sido classicamente derivadas da análise da variação na forma das estruturas morfológicas.

Em BONATO (2004) foi analisada a variação geográfica da forma do crânio em 22 populações de roedores da espécie *Thrichomys apereoides*, amostradas das regiões Nordeste, Sudeste e Centro-Oeste do Brasil (vide Figura 5.4). A variação na forma foi descrita usando *partial warps*, as quais são variáveis derivadas do formalismo de morfometria geométrica.



1 – Itapagé;
2 – Campos Sales;
3 – Crato;
4 – Princesa Isabel;
5 – Bodocó;
6 – Triunfo;
7 – Caruaru;
8 – Pesqueira;
9 – Alagoinha;
10 – Floresta;
11 – Santana do Ipanema;
12 – Viçosa;
13 – Palmeira dos Índios;
14 – Canudos;
15 – Barreiras;
16 – Feira de Santana;
17 – Palmeiras;
18 – Bom Jesus da Lapa;
19 – Serra da Mesa;
20 – Jaíba;
21 – Salinas;
22 – Barão de Melgaço.

Figure 5.4 - Locais das amostras coletadas

A partir dos dados coletados acerca da forma do crânio das 22 populações, foi gerada a matriz de distâncias a ser considerada como dado de entrada para o MNJ. Dessa forma, essa matriz reflete as diferenças entre pares de populações de acordo com as diferenças nas formas do crânio dos roedores.

Os dados usados para obter as variáveis de forma são coordenadas tridimensionais de marcos, os quais representam demarcações previamente definidas nos crânios de *Thrichomys apereoides.*

Começando pela matriz de distâncias, o propósito é determinar alternativas múltiplas para explicar os dados de entrada, sempre respeitando os critérios de evolução mínima e quadrados mínimos. A matriz de distâncias utilizada para o estudo encontra-se no Apêndice B.

5.4.2 Resultados do estudo de caso

A Figura 5.5 apresenta um exemplo da execução do NJ clássico de uma árvore sem raiz e duas sub-árvores extraídas das topologias alternativas produzidas pelo MNJ (DA SILVA *et al.*, 2005) para os dados dos roedores.

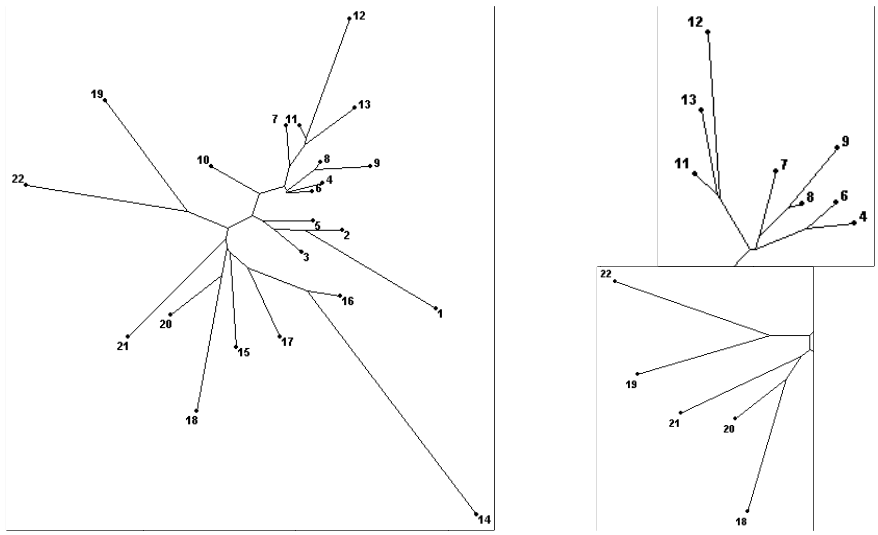


Figura 5.5 - A árvore produzida pelo NJ clássico (à esquerda), e duas propostas alternativas de sub-árvores (à direita), ambas extraídas de topologias produzidas pelo MNJ.

Considerando um limite empiricamente definido de 2% de aumento admissível no tamanho total da árvore, levando-se em conta a configuração corrente da árvore e a melhor decisão no momento, escolhas de pares múltiplos foram efetuadas, resultando em cinco topologias finais distintas. A saída simples do NJ clássico (vide Figura 5.5, à esquerda) é com certeza uma das topologias produzidas pelo MNJ, o que significa que o MNJ é uma generalização do NJ.

Uma análise de topologias alternativas nos leva à conclusão de que as sub-árvores: 4-6-7-8-9-11-12-13 e 18-20-21 são menos estáveis, no sentido de que as topologias alternativas propõem configurações distintas para essas sub-árvores (vide Figura 5.5 , à direita).

Basicamente, mais duas distinções significativas podem ser extraídas da comparação visual envolvendo a topologia original e as alternativas:

- a OTU 7 é agrupada com as OTUs 8 e 9, sendo que as demais OTUs na sub-árvore são realocadas em várias configurações;

- a OTU 21 é primeiramente agrupada com as OTUs 18 e 20 antes de ser agrupada com as OTUs restantes.

Dado que nosso interesse é identificar grupos que compartilham características morfológicas e continuidade geográfica ao longo de um espaço geográfico, as topologias alternativas são muito sugestivas. Baseando-se na Figura 5.5, colocando-se a OTU 7 mais próxima às OTUs 8 e 9, e a OTU 21 mais próxima às OTUs 18 e 20, fica-se mais próximo da hipótese de continuidade geográfica.

Note que não estão sendo procuradas explicações distintas que melhor apóiem um conjunto de hipóteses, simplesmente porque as topologias alternativas são também soluções de alta qualidade em termos do critério de evolução mínima e do critério dos quadrados mínimos. Dessa forma, elas representam perpectivas confiáveis e opcionais, extraídas do mesmo conjunto de dados.

A análise das sub-árvores da Figura 5.5 mostrou que as mesmas são soluções interessantes dado que o foco do problema dos roedores é o de identificar grupos de populações que compartilham características morfológicas e continuidade geográfica. Essas sub-árvores explicam de maneira adequada o agrupamento de espécies com características semelhantes em determinado espaço geográfico.

Capítulo 6

Otimização Multi-objetivo e Soluções Alternativas para Árvores Filogenéticas

Resumo – A proposta principal desse capítulo é apresentar os principais conceitos envolvidos na reconstrução baseada em princípios de otimização multi-objetivo. O algoritmo omni-aiNet para otimização multi-objetivo é apresentado e sua extensão junto a problemas de reconstrução de árvores filogenéticas é então proposta como uma contribuição desse trabalho. O conceito de fronteira de Pareto é empregado na obtenção de árvores filogenéticas (soluções) não-dominadas.

6.1 Introdução

Os problemas tratados na área de otimização normalmente podem apresentar diferentes tipos e formas (COELHO & VONZUBEN, 2006), que acabam exigindo um tratamento distinto de acordo com suas características. De maneira geral, tais problemas de otimização podem ser classificados de acordo com dois critérios:

i. Número de Objetivos:

   Um único objetivo: conhecidos como problemas de otimização de objetivo simples ou mono-objetivo;

   Vários objetivos conflitantes entre si: conhecidos como problemas de otimização multi-objetivo;

ii. Número de Soluções Ótimas:

   Uma única solução ótima global: conhecidos como problemas de otimização uni-globais ou problemas unimodais;

   Várias soluções ótimas globais: conhecidos como problemas de otimização multi-globais ou multimodais.

O problema de reconstrução de árvores filogenéticas, descrito no Capítulo 3, também pode ser visto como um problema de otimização, em que se busca minimizar ou maximizar um ou mais critérios, tais como evolução mínima (KIDD & SGARAMELLA-ZONTA, 1971), erro quadrático (BULMER, 1991) ou alguma métrica de verossimilhança (FELSENSTEIN, 2004) , por exemplo. Na literatura desta área, os critérios sendo otimizados na reconstrução de árvores filogenéticas são geralmente aplicados isoladamente ou de maneira iterativa, como no algoritmo *Neighbor-Joining*, descrito no Capítulo 3.

Como pôde ser visto no Capítulo 3, existe uma grande quantidade de métricas para avaliar a qualidade de uma dada árvore filogenética reconstruída, as quais poderiam ser utilizadas em conjunto de forma a levar a resultados mais abrangentes. Desta forma, o problema de reconstrução de árvores filogenéticas passaria a ser caracterizado como um problema de otimização multi-objetivo, que é onde se encaixa a proposta que será discutida neste capítulo. Neste trabalho, os critérios de Evolução Mínima (KIDD & SGARAMELLA-ZONTA, 1971) e Erro Quadrático Médio (BULMER, 1991), descritos no Capítulo 3, foram considerados em conjunto, e um algoritmo imuno-inspirado denominado *omni-aiNet* (COELHO & VONZUBEN, 2006) foi devidamente adaptado para resolução deste problema multi-objetivo de reconstrução de árvores filogenéticas, a partir da matriz de distâncias original do problema. Foram empregadas aqui árvores filogenéticas sem raiz.

Apesar desta abordagem multi-objetivo de reconstrução de árvores se mostrar uma iniciativa natural, dada a enorme quantidade de métricas para avaliação de propostas de árvores, isto ainda não foi bem explorado na literatura. A única proposta de aplicação de um algoritmo evolutivo para otimização multi-objetivo em filogenia foi feita por POLADIAN & JERMIIN (2004). No entanto, sua proposta é diferente da adotada aqui, uma vez que em POLADIAN & JERMIIN (2004) os autores evoluem um conjunto de árvores filogenéticas considerando-se apenas um único objetivo (critério de verossimilhança máxima) e dois conjuntos de dados distintos. Dessa forma, as diversas soluções conflitantes presentes na fronteira de Pareto (este conceito será tratado na Seção 6.2.2) são motivadas pelas informações conflitantes presentes nos dois conjuntos de dados distintos.

Este capítulo está estruturado da seguinte forma: na Seção 6.2 os principais conceitos associados à otimização multi-objetivo serão apresentados, bem como algumas abordagens clássicas de resolução deste tipo de problema, que servirão de motivação para o uso de abordagens evolutivas e imuno-inspiradas, como o algoritmo *omni-aiNet*, que também será tratado nesta seção; na Seção 6.3, a metodologia adotada aqui será formalmente apresentada, bem como as principais modificações necessárias para permitir a evolução de árvores filogenéticas através da aplicação da *omni-aiNet*. Os resultados experimentais da aplicação deste algoritmo para o problema multi-objetivo de reconstrução de árvores filogenéticas serão apresentados no próximo capítulo.

6.2 Conceitos de Otimização Multi-objetivo

Nesta seção, os problemas de otimização multi-objetivo serão formalmente definidos e dois dentre os principais conceitos associados a este tipo de problemas serão apresentados: *dominância* e *fronteira de Pareto*. Além disso, será feita uma breve revisão sobre algumas das principais abordagens clássicas e evolutivas para solução deste tipo de problema e, em seguida, o algoritmo *omni-aiNet* será apresentado.

6.2.1 Formalismo Matemático

Sem perda de generalidade, os conceitos e definições associados aos problemas multi-objetivos tratados neste trabalho serão formalizados considerando-se um problema de *minimização* de todos os objetivos envolvidos. Dessa forma, um problema de otimização multi-objetivo pode ser definido como:

$$\begin{array}{lll} \text{Minimizar} & \left(f_1(\mathbf{x}) \quad f_2(\mathbf{x}) \quad \cdots \quad f_M(\mathbf{x})\right) & \\ \text{s.a.} & g_j(\mathbf{x}) \geq 0 & j = 1,2,\cdots,J \\ & h_k(\mathbf{x}) = 0 & k = 1,2,\cdots,K \\ & \mathbf{x} \in \Omega \subset R^n & \end{array} \tag{6.1}$$

onde $f_i(\mathbf{x})$, i=1,...,M, são as M funções-objetivo do problema, $g_j(\mathbf{x})$ é o conjunto de J restrições de desigualdade, $h_k(\mathbf{x})$ é o conjunto de K restrições de igualdade, Ω é o domínio do problema e n é a dimensão de $\mathbf{x}$.

Em problemas de otimização com um único objetivo, a solução corresponde a um ou mais pontos factíveis, cujos valores levam a um extremo da função-objetivo, ou seja, se o problema tratado for de minimização, sua solução será o conjunto de pontos factíveis que apresentam o menor valor possível da função-objetivo.

Já para o caso de problemas multi-objetivo, este conceito de soluções correspondendo a valores extremos das funções-objetivo não pode ser diretamente estendido, uma vez que os objetivos envolvidos podem ser (e geralmente são) conflitantes.

Para ilustrar este conceito de objetivos conflitantes, tomemos um exemplo didático: suponha que uma dada montadora de veículos deseja construir um veículo que deve ter a maior resistência a colisões possível, gastar o mínimo de combustível e ter o menor custo. Suponha também que só é possível atuar no material de fabricação da carroceria. Dessa forma, supondo que seja escolhido o aço como material, seriam necessárias chapas mais grossas para aumentar a resistência a colisões, mas isso acabaria deixando o veículo mais pesado e, conseqüentemente, levaria a um maior consumo de combustível. Por outro lado, o fabricante poderia decidir construir a carroceria do veículo com algum material como fibra de carbono, o que é altamente resistente e, ao mesmo tempo, leve, o que reduziria o consumo de combustível. Só que, nesse caso, o custo do veículo seria muito mais alto que no caso do uso de aço como matéria-prima.

Este exemplo deixa claro que os três objetivos envolvidos no problema são conflitantes, uma vez que não podemos tomar a solução extrema para um deles (por exemplo, a resistência com o uso de fibra de carbono) e esperar que o resultado obtido seja ótimo também para os demais objetivos (novamente considerando o caso da fibra de carbono, o custo do veículo acaba sendo muito alto). O que ocorre geralmente em problemas multi-objetivo, com objetivos conflitantes, é que existem diversas soluções que correspondem a

compromissos diferentes entre os objetivos, ou seja, cada solução de um problema multi-objetivo indica, para um dado valor de um dos objetivos, qual é o melhor valor que pode ser obtido para os demais.

Dessa forma, podem existir infinitas soluções para um dado problema, que correspondem aos infinitos compromissos existentes entre todos os objetivos. A este conjunto de soluções ótimas para um problema multi-objetivo dá-se o nome de *Fronteira de Pareto*, que será formalmente descrita na próxima subseção.

6.2.2 Fronteira de Pareto e Dominância

Um conceito que deve ser definido antes que seja formalizado o que é exatamente uma fronteira de Pareto é o conceito de *dominância*. Diz-se que uma dada solução **u** domina outra solução **v** se **u** é melhor ou igual a **v** em todos os objetivos do problema e, para pelo menos um dos objetivos, **u** é estritamente melhor que **v**. De uma maneira mais formal, considerando-se a definição adotada na Expressão 6.1, diz-se que **u** domina **v** (adotaremos a notação $\mathbf{u} \prec \mathbf{v}$) se e somente se:

$$i) \qquad f_i(\mathbf{u}) \leq f_i(\mathbf{v}) \quad \forall i \in \{1,2,\cdots,M\}$$

$$ii) \qquad \exists j \in \{1,2,\cdots,M\} : f_j(\mathbf{u}) < f_j(\mathbf{v})$$

A Figura 6.1 ilustra o conceito de dominância entre pontos de um problema de minimização de objetivos. Nesta Figura, $\mathbf{B} \prec \mathbf{C}$**.** Repare que B e C são soluções associadas a valores distintos de **x**, embora estes pontos estejam sendo plotados no espaço das funções-objetivo e não no espaço das variáveis.

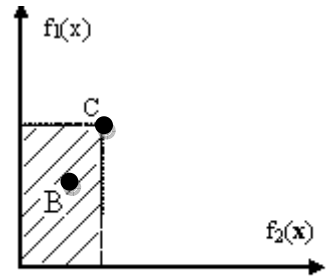


Figura 6.1 - Qualquer ponto na região hachurada, desde que corresponda a uma solução factível, domina a solução representada pelo ponto C

Dessa forma, diz-se então que uma dada solução **x** de um problema multi-objetivo pertence à *Fronteira de Pareto* se e somente se não existe nenhuma outra solução **x'**, factível para o problema em questão, que domine **x**. A todo este conjunto de soluções não-dominadas dá-se o nome de *conjunto ótimo de Pareto*. Formalmente temos então:

→Conjunto ótimo de Pareto $P^* = \{x \in \Omega \mid \neg\exists x' \in \Omega, x' \prec x\}$

onde Ω é o conjunto de soluções factíveis do problema.

→Fronteira de Pareto (P_F):

$$P_F = \{\mathbf{f} = (f_1(\mathbf{x}) \quad f_2(\mathbf{x}) \quad \cdots \quad f_M(\mathbf{x})) \mid \mathbf{x} \in P^*\}$$

Na Figura 6.2, estão ilustradas três fronteiras para um dado problema de minimização de dois objetivos (três conjuntos de soluções, cada um contendo soluções não dominadas pelos demais daquele mesmo conjunto), sendo que a linha pontilhada corresponde à *Fronteira de Pareto* propriamente dita, ou seja, à solução do problema.

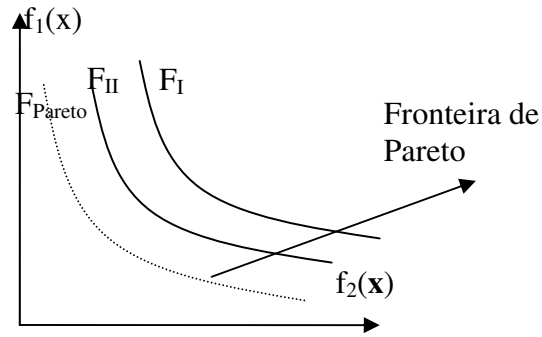


Figura 6.2 - Fronteiras no espaço de objetivos tal que $F_{II} \prec F_I$ e $F_{Pareto} \prec F_{II}$.

Definidos estes dois conceitos, podemos perceber que, diferentemente de um problema de otimização com um único objetivo, em que se deseja obter o ótimo global (ou os múltiplos ótimos, no caso de problemas multimodais), a resolução de problemas multi-objetivo engloba duas metas: obter um conjunto de soluções na fronteira de Pareto e, ao mesmo tempo, garantir que tais soluções sejam as mais diversas possíveis, de forma a promover uma cobertura uniforme da fronteira de Pareto. Tal cobertura uniforme é desejável, pois todos os objetivos do problema são considerados, *a priori,* de igual importância.

Pode-se dizer que estas duas metas presentes em otimização multi-objetivo são de certa forma ortogonais, uma vez que a obtenção de uma delas não necessariamente implica na obtenção da outra, sendo necessários mecanismos explícitos ou implícitos, nos algoritmos de otimização multi-objetivo, tanto para enfatizar a convergência para a fronteira de Pareto quanto para manter um conjunto diverso de soluções (DEB, 2001).

6.2.3 Espaço de Variáveis e Espaço de Objetivos

Outro ponto que distingue os problemas de otimização multi-objetivo dos problemas de otimização mono-objetivo tradicionais é a presença de dois espaços de busca distintos do problema: o *espaço de variáveis* e o *espaço de objetivos.* O *espaço de variáveis* é onde se faz a busca pelas soluções do problema, ou seja, é o domínio das variáveis do problema. Já o *espaço de objetivos* é o espaço formado pelas funções-objetivo do problema. Dessa forma, se tivermos as variáveis do problema no R^n e um conjunto de M objetivos, a aplicação das funções-objetivo em uma dada variável corresponderá a um mapeamento. Este mapeamento está ilustrado na Figura 6.3 (no caso, g: $R^2 \rightarrow R^2$).

Apesar de estes dois espaços estarem relacionados pelo mapeamento único descrito acima, geralmente suas propriedades são distintas e o mapeamento em questão é não-linear, o que introduz mais algumas dificuldades à tarefa de encontrar as soluções para este tipo de problema (DEB, 2001).

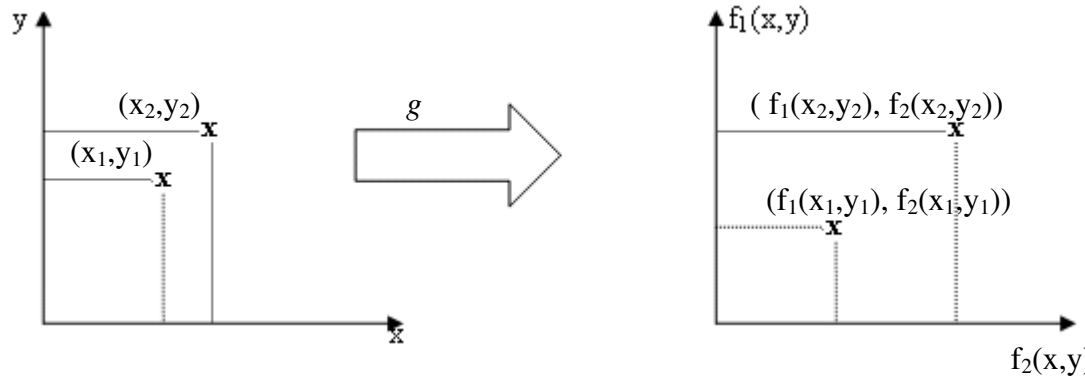


Figura 6.3 - Mapeamento do espaço de variáveis (à esquerda) para o espaço de objetivos (à direita)

Por exemplo, a proximidade de duas soluções em um dos espaços não necessariamente implica em uma proximidade no outro espaço. Dessa forma, ao se tentar atingir a meta de manutenção de diversidade das soluções encontradas, deve-se decidir previamente em qual espaço tal diversidade será considerada.

Em qualquer algoritmo de otimização, a busca sempre é feita no espaço de variáveis. No entanto, a movimentação das soluções encontradas no espaço de variáveis pode ser rastreada no espaço de objetivos, e esta representação das soluções no espaço de objetivos pode ser utilizada como um mecanismo auxiliar para a busca no espaço de variáveis, de forma a permitir o atendimento das duas metas envolvidas na otimização multi-objetivo. Como será visto mais adiante, o algoritmo *omni-aiNet*, a ser devidamente adaptado ao contexto deste trabalho, atua em ambos os espaços envolvidos ao longo de sua busca por soluções.

6.2.4 Abordagens Clássicas

Nesta subseção, serão apresentados alguns algoritmos clássicos para solução de problemas de otimização multi-objetivo. Esses algoritmos foram chamados de *clássicos* como uma forma de distinção para com as abordagens *evolutivas*, que serão tratadas posteriormente neste capítulo.

O método mais simples de solução de problemas multi-objetivos é conhecido por *Método da Soma Ponderada* que, como o próprio nome sugere, escalona um conjunto de objetivos em um único objetivo, através da pré-multiplicação de cada um dos objetivos originais por um valor previamente definido pelo usuário. Dessa forma, o problema original dado pela Expressão 6.1, se torna:

$$\begin{aligned} \text{Minimizar} \quad & F(\mathbf{x}) = \sum_{m=1}^{M} w_m . f_m(\mathbf{x}) \\ \text{s.a.} \quad & g_j(\mathbf{x}) \geq 0 \qquad j = 1,2,\cdots,J \\ & h_k(\mathbf{x}) = 0 \qquad k = 1,2,\cdots,K \\ & \mathbf{x} \in \Omega \subset R^n \end{aligned} \tag{6.2}$$

onde M é o número de objetivos originais e w_m, m = 1,2,..., M, são os pesos definidos *a priori*. Aqui, os pesos normalmente são adotados no intervalo [0, 1] e, como o mínimo do problema 6.2 não muda se multiplicarmos todos os pesos por uma constante, também se costuma adotar um conjunto de pesos cuja soma seja 1 ($\sum_{m=1}^{M} w_m = 1$).

Dessa forma, pode-se aplicar qualquer algoritmo tradicional de otimização ao problema 6.2 que se obterá uma das possíveis soluções para o problema multi-objetivo original. No entanto, este método exige que se defina o conjunto de pesos, o que pode ser uma tarefa não trivial quando não se tem um conhecimento mais aprofundado do problema que permita determinar as importâncias relativas de cada um dos objetivos. Além disso, um problema introduzido pela ponderação de objetivos está relacionado à escala dos objetivos: é comum que dois objetivos de um mesmo problema tenham ordens de magnitude diferentes, o que pode levar a resultados não satisfatórios caso a ponderação seja aplicada sem que haja uma normalização prévia dos valores de cada objetivo.

Outra questão associada a esta técnica clássica de resolução de problemas de otimização multi-objetivo está na incapacidade de se encontrar soluções para o problema multi-objetivo em regiões não-convexas da fronteira de Pareto. Para ilustrar este problema, vamos considerar aqui apenas a minimização de dois objetivos. Neste caso, a resolução do problema dado pela Expressão 6.2 consiste em encontrar uma solução factível para o problema que apresente o menor valor possível de F. O efeito das ponderações das funções-

objetivo originais será indicado por retas cuja inclinação depende dos pesos escolhidos para cada objetivo (vide Figura 6.4, retas ‘a’, ‘b’, ‘c’ e ‘d’).

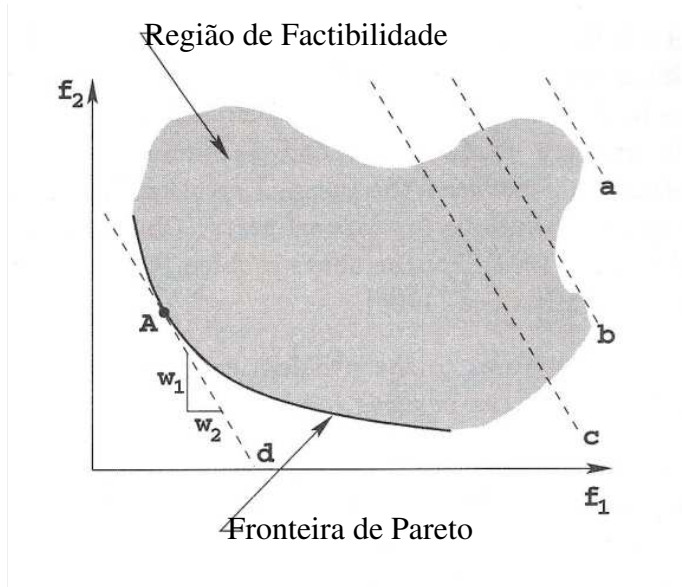


Figura 6.4 - Região de factibilidade e retas indicando a ponderação dos pesos junto a um problema com dois objetivos

Pela Figura 6.4, podemos perceber que, para o conjunto de pesos escolhido, a solução do problema será dada pelo ponto A, que corresponde ao ponto onde a seqüência de retas com a mesma inclinação tangencia a região de factibilidade do problema. Podemos perceber também que, ao alterarmos os pesos associados aos dois objetivos, as inclinações das retas também serão alteradas e, conseqüentemente, o ponto ótimo obtido (ponto em que se tangencia a região de factibilidade) também será diferente. No entanto, se a região de factibilidade do problema resultar em uma fronteira de Pareto não-convexa, como a representada na Figura 6.5, não será possível encontrar soluções na região de não-convexidade (região entre os pontos B e C na Figura 6.5), uma vez que, para isto, esta curva sempre tangenciará outros pontos que acabam sendo interceptados primeiramente pelas retas, não importando a ponderação (esses pontos estarão localizados nas regiões AB ou CD na Figura 6.5) (DEB, 2001).

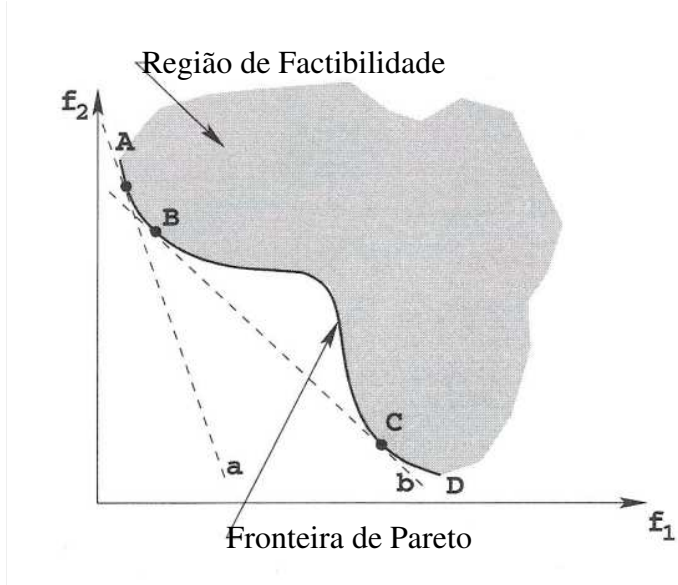


Figura 6.5 - Falha do método de soma ponderada para obtenção de soluções em regiões não-convexas da fronteira de Pareto

Outro método clássico bem conhecido de resolução de problemas de otimização multi-objetivo, e que não é susceptível a regiões não-convexas da fronteira de Pareto, é o método conhecido como *ε-restrição* (do inglês, *ε-constraint*) (HAIMES *et al.,* 1971). Nesta estratégia, um dos objetivos é escolhido como o único objetivo a ser otimizado, com os demais sendo tratados como restrição de desigualdade do problema, o que também permite que os métodos clássicos de resolução para otimização mono-objetivo possam ser aplicados. Dessa forma, o problema dado pela Expressão 6.1 se torna:

$$\begin{aligned} \text{Minimizar} \quad & f_\mu(\mathbf{x}) \\ \text{s.a.} \quad & \mathrm{f_m}(\mathbf{x}) \le \varepsilon_m \qquad m = 1,2,\cdots,M \text{ e } m \ne \mu \\ & \mathrm{g}_j(\mathbf{x}) \ge 0 \qquad j = 1,2,\cdots,J \\ & h_k(\mathbf{x}) = 0 \qquad k = 1,2,\cdots,K \\ & \mathbf{x} \in \Omega \subset R^n \end{aligned} \tag{6.3}$$

Nas expressões dadas em 6.3, o parâmetro ε_m representa um limite superior para o valor do objetivo f_m e não necessariamente corresponde a um valor próximo de zero. Para ilustrar o funcionamento do método da *ε-restrição*, considere novamente o problema de minimização de dois objetivos, agora representado na Figura 6.6, onde o objetivo f_1 é tratado como a restrição de desigualdade. Em um primeiro cenário, tomemos $f_1(\mathbf{x}) \le \varepsilon_1^{\,b}$. Neste caso, ao aplicarmos um algoritmo de otimização mono-objetivo tradicional, a solução para nosso

problema corresponderá ao ponto **B** na fronteira de Pareto dada na Figura 6.6. Por outro lado, se considerarmos $f_1(\mathbf{x}) \leq \varepsilon_1^{\,c}$, o ótimo encontrado corresponderá ao ponto **C**, que está localizado na região de não-convexidade da fronteira de Pareto. Isto ilustra bem que tal técnica não é suscetível a regiões não-convexas como o método da soma ponderada.

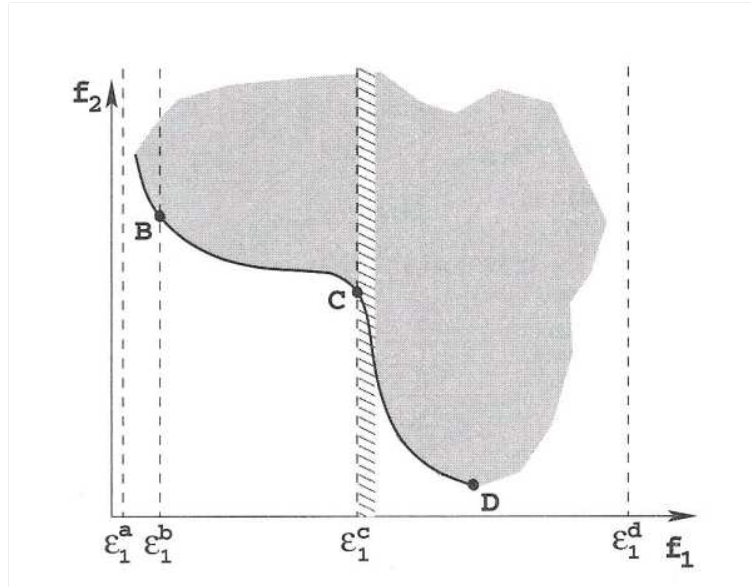


Figura 6.6 - O método da ε-restrição

Por outro lado, a determinação dos valores de ε_m também exige um conhecimento prévio do problema, para evitar situações como a ilustrada na Figura 6.6 considerando-se $f_1(\mathbf{x}) \leq \varepsilon_1^{\,a}$. Neste caso, esta escolha de limitante superior para o primeiro objetivo acaba por tornar o problema infactível, conseqüentemente impedindo a obtenção de qualquer solução sobre a fronteira de Pareto.

Outros métodos clássicos bem populares, mas que não serão tratados a fundo aqui são: métodos de *Métrica Ponderada*, que buscam minimizar a soma ponderada do resultado da aplicação de uma métrica de distância entre o valor de cada objetivo e um determinado valor ideal para este objetivo; métodos de *Função Valor* (ou *Função Utilidade*), que buscam otimizar uma função que relacione os *M* objetivos do problema, fazendo assim uma transformação $g : \Re^M \rightarrow \Re^1$; métodos de *Programação de Metas* (do inglês *Goal Programming*), que definem metas a um sub-conjunto de objetivos e buscam soluções que atendam a estas metas pré-estabelecidas; e, por fim, os chamados *Métodos Interativos*, que

se baseiam no conhecimento do usuário sobre o problema para direcionar a busca por soluções em regiões de interesse do usuário. Todos estes métodos estão detalhados em (DEB, 2001), onde uma discussão sobre suas vantagens e desvantagens também é apresentada.

De maneira geral, os algoritmos clássicos para solução de problemas multi-objetivo apresentam algumas dificuldades, principalmente quando se tem o interesse em encontrar múltiplas soluções Pareto-ótimas:

-Normalmente apenas uma solução Pareto-ótima é encontrada a cada execução de um algoritmo clássico;

-Nem todas as soluções da fronteira de Pareto podem ser encontradas por alguns algoritmos, particularmente em problemas multi-objetivo com fronteira de Pareto não-convexa;

-Todos os algoritmos exigem alguma forma de conhecimento sobre o problema, tais como pesos adequados, valores de ε-restrição e metas para os objetivos.

Diante disso, uma abordagem que tem recebido grande atenção e vem sendo aplicada com sucesso nos últimos anos são os algoritmos evolutivos para otimização multi-objetivo (do inglês ***M**ulti-**O**bjective **E**volutionary **A**lgorithms - MOEAs*) (COELLO COELLO,1998; DEB, 2001; COELLO COELLO, 2006), que serão apresentados na próxima subseção.

6.2.5 Algoritmos Evolutivos para Otimização Multi-objetivo

Atualmente, o uso de meta-heurísticas para solução dos mais diversos tipos de problemas está amplamente difundido entre pesquisadores de diversas disciplinas, sendo que tal aceitação cresceu rapidamente nos últimos anos, dadas as vantagens que tais abordagens apresentam frente às demais. Especificamente para o caso de otimização multi-objetivo, dentre as diversas meta-heurísticas existentes, os *algoritmos evolutivos* (que se baseiam na emulação do mecanismo de seleção natural) estão entre os mais populares (GOLDBERG, 1989; FOGEL, 1999), uma vez que trabalham com um conjunto de possíveis soluções para o problema (chamado de *população*), o que permite encontrar diversos membros do conjunto

ótimo de Pareto em uma única execução do algoritmo. Além disso, os algoritmos evolutivos são menos susceptíveis à forma da fronteira de Pareto (sendo capazes de tratar também problemas com superfície de Pareto não-convexas ou não-contínuas) e geralmente não exigem um conhecimento muito amplo sobre o problema sendo tratado.

Existem na literatura diversas propostas de algoritmos evolutivos para otimização multi-objetivo (ZITZLER & THIELE,1999; CORNE *et al.*, 2000; KNOWLES & CORNE, 2000; COELLO COELLO & TOSCANO PULIDO, 2001; CORNE *et al*., 2001; ZITZLER *et al.*, 2001; DEB *et al.*, 2002), mas dois destes algoritmos são considerados atualmente o estado-da-arte: NSGA-II (*Nondominated Sorting Genetic Algorithm II*), proposto por Deb *et al.* (2002) e SPEA2 (*Strength Pareto Evolutionary Algorithm 2*), proposto por Zitzler *et al.* (2001).

O algoritmo NSGA-II (DEB *et al.*, 2002) é uma evolução do algoritmo NSGA original (*Nondominated Sorting Genetic Algorithm*) proposto por SRINIVAS & DEB (1994), e que tem como principais características *elitismo* (manutenção dos melhores indivíduos da população para a próxima iteração), ordenação dos indivíduos pelo critério de não-dominância e manutenção de diversidade por *crowding distance* (que consiste em analisar a distância, no espaço de objetivos, de cada indivíduo em relação ao seu vizinho mais próximo). Já o algoritmo SPEA2 (ZITZLER *et al*., 2001) é uma evolução do algoritmo SPEA original, proposto por Zitzler & Thiele (1999), e se baseia em um mecanismo de atribuição de *fitness* aos indivíduos que leva em conta a relação de dominância do indivíduo em questão com os demais na população e a distância aos k vizinhos mais próximos. Além disso, o algoritmo SPEA2 trabalha com duas populações de tamanho fixo, sendo que uma corresponde a um arquivo onde são armazenadas as soluções não-dominadas obtidas até o momento, e a outra à população da busca na geração atual.

Nesta tese, foi adotado um algoritmo que se baseia num paradigma que apresenta algumas semelhanças com algoritmos evolutivos: o algoritmo imuno-inspirado (que busca emular alguns mecanismos presentes no sistema imunológico natural dos seres humanos (DE CASTRO & TIMMIS, 2002) conhecido como *omni-aiNet* (COELHO & VONZUBEN, 2006), que tem como principal vantagem a capacidade de se adequar ao problema sendo tratado e

automaticamente ajustar o tamanho de sua população às condições encontradas. A versão original deste algoritmo será descrita a seguir, enquanto que as modificações necessárias para reconstrução de árvores filogenéticas serão apresentadas na Seção 6.3.

6.2.6 O algoritmo *omni-aiNet*

O algoritmo *omni-aiNet* (do inglês *Artificial Immune Network for Omni Optimization*) é um algoritmo populacional e imuno-inspirado, proposto por COELHO & VON ZUBEN (2006), e desenvolvido para otimização de qualquer tipo de função no espaço contínuo (o que os autores chamaram de *omni-otimização*). Este algoritmo é capaz de automaticamente degenerar seus mecanismos internos de forma a tratar eficientemente problemas mono-objetivo unimodais, mono-objetivo multimodais, multi-objetivo unimodais e multi-objetivo multimodais. Neste trabalho, como se está interessado em resolver problemas multi-objetivo, apenas esta característica do algoritmo será tratada aqui.

Além da capacidade de otimizar qualquer tipo de função no espaço contínuo, o algoritmo *omni-aiNet* se distingue da maioria dos demais algoritmos populacionais para otimização multi-objetivo por incluir as principais características inerentes aos sistemas imunológicos artificiais (do inglês *AIS – Artificial Immune Systems*), que são (DE CASTRO & TIMMIS, 2002):

- Capacidade intrínseca de manter a diversidade da população, o que favorece a resolução de problemas multimodais e contribui para uma boa cobertura das fronteiras de Pareto;
- Ajuste dinâmico do tamanho da população, através da eliminação das piores soluções presentes em regiões muito povoadas e do favorecimento da manutenção de soluções em regiões pouco povoadas;
- Tendência a localizar e preservar soluções ótimas locais.

Os principais passos do algoritmo *omni-aiNet* podem ser vistos no fluxograma representado na Figura 6.7. O algoritmo começa com a geração aleatória dos N_{inic} indivíduos que constituirão a população inicial do algoritmo (N_{inic} é definido pelo usuário) e, em seguida, verifica se a condição de parada do algoritmo foi atingida (no algoritmo original, tal

condição é dada por um número máximo de iterações definido pelo usuário). Caso tal condição de parada esteja satisfeita, o algoritmo encerra sua execução e retorna as soluções não-dominadas presentes na população final. Caso contrário, o algoritmo entra em seu ciclo principal, que consiste nos seguintes passos:

- Clonagem: para cada indivíduo na população, N_c clones (cópias) são gerados.

- Hipermutação: os clones gerados na etapa anterior passam então por um processo de mutação polinomial com taxa de variabilidade genética inversamente proporcional à sua *afinidade,* ou seja, ao seu *fitness*. Nesta etapa, os melhores indivíduos sofrem uma mutação com pequena variação genética (ajuste fino), enquanto que os piores indivíduos sofrem uma variação genética maior, uma vez que estão mais distante de regiões promissoras do espaço de busca.

- Seleção: após as etapas de clonagem e hipermutação, a população do algoritmo que anteriormente continha N indivíduos, terá agora $(N_c + 1).N$ elementos. Dessa forma, a etapa de seleção consiste, como o próprio nome diz, em selecionar N indivíduos, entre os clones gerados e os indivíduos iniciais, para permanecerem na população. Esta etapa utiliza um mecanismo de *ranking* baseado no critério de não-dominância para escolher os melhores indivíduos na população (indivíduos "menos dominados", ou seja, dominados pelo menor número de outros indivíduos) e um mecanismo de *grid* para também selecionar os indivíduos mais espaçados entre si, no espaço de objetivos. Estes dois mecanismos (*ranking* e *grid*) não serão detalhados aqui (ambos são devidamente apresentados e descritos em COELHO & VONZUBEN (2006)), mas é importante ressaltar que esta etapa é caracterizada pela presença de *elitismo* (uma vez que os melhores indivíduos sempre são mantidos na nova população) e manutenção de diversidade no espaço de objetivos.

- Mutação por Duplicação Gênica: o algoritmo *omni-aiNet* incorpora também um segundo mecanismo de mutação, conhecido como duplicação gênica, que emula o fenômeno de duplicação gênica natural, o qual consiste na repetição de partes da cadeia de DNA durante a leitura de um cromossomo. Este mecanismo foi considerado um operador relevante em outro algoritmo da família *aiNet* (no

caso, o algoritmo *dopt-aiNet* (DE FRANÇA *et al.*, 2005), proposto para otimização dinâmica), e consiste basicamente na seleção aleatória de uma coordenada *i* do indivíduo e replicação do valor presente nesta coordenada em todas as demais coordenadas do indivíduo, sempre que isso resultar em uma melhora de *fitness*.

- Supressão e inserção de novos indivíduos: estes dois últimos mecanismos da *omni-aiNet* são os principais responsáveis pela boa exploração do espaço de busca promovida pelo algoritmo e pela tendência de localização e manutenção de soluções ótimas locais (ou seja, pela manutenção de diversidade no espaço de variáveis). A cada N_s iterações, a distância euclidiana entre cada indivíduo e os demais é calculada (normalmente chamada de *afinidade entre anticorpos*) e, caso esta distância entre dois indivíduos seja menor que um determinado limiar (definido pelo usuário), o indivíduo de pior *fitness* é suprimido, eliminando assim a redundância de indivíduos em uma determinada região do espaço de busca. Após esta supressão, uma porcentagem de novos indivíduos (também definida pelo usuário) gerados aleatoriamente é inserida na população.

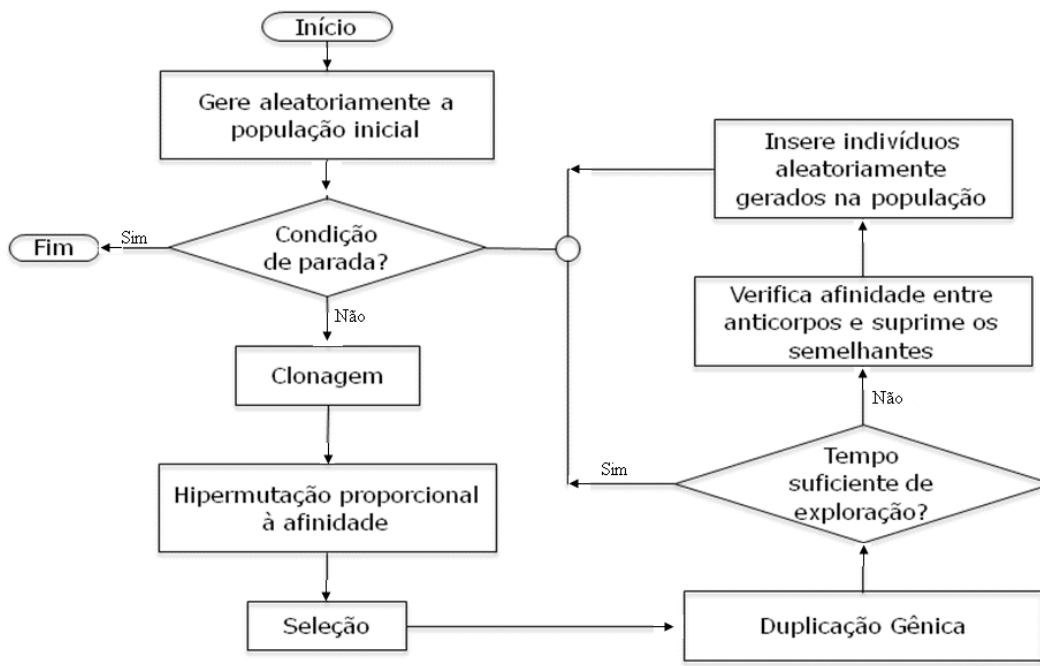


Figura 6.7 - Fluxograma do algoritmo omni-aiNet

Por fim, após a execução destes cinco passos principais, o algoritmo volta a verificar se a condição de parada foi atendida e, caso ainda não tenha sido, reinicia seu ciclo principal. No trabalho original (COELHO & VONZUBEN, 2006), o algoritmo foi aplicado a um conjunto de problemas tanto mono-objetivos quando multi-objetivos e se mostrou competitivo com outras abordagens presentes na literatura, inclusive superando-as em alguns casos. Para o problema de reconstrução de árvores filogenéticas, várias adaptações e extensões ao algoritmo foram realizadas nesta tese. São elas: 1) a utilização da medida de distância Robinson-Foulds em substituição à distância Euclidiana; 2) o emprego das funções-objetivo Evolução Mínima e Quadrados Mínimos; 3) mecanismo específico para geração da população inicial e, principalmente, 4) a utilização do algoritmo Neighbor-Joining como construtor do fenótipo a partir do genótipo. Essas adaptações e extensões são explicadas a seguir.

6.3 Abordagem Multi-objetivo para Reconstrução de Árvores Filogenéticas

Como mencionado anteriormente, neste trabalho o algoritmo *omni-aiNet* foi estendido para evoluir uma população de árvores filogenéticas de topologias possivelmente distintas (COELHO *et al.*, 2007). Para isso, três módulos do algoritmo original foram modificados:

i) Codificação adequada dos indivíduos através de um mapeamento genótipo-fenótipo;

ii) Um mecanismo diferenciado de geração da população inicial;

iii) Um mecanismo mais apropriado para comparar indivíduos similares na etapa de supressão.

Nesta subseção, estas modificações feitas no algoritmo original serão descritas e as condições de factibilidade de soluções adotadas serão apresentadas.

6.3.1 Codificação de Indivíduos

Como foi apresentado no Capítulo 3, o algoritmo *Neighbor-Joining* (NJ) é um método eficiente de se obter uma árvore filogenética a partir de uma matriz de distâncias, mesmo levando geralmente apenas a soluções sub-ótimas. Diante disso, neste trabalho tal característica foi explorada, permitindo que o algoritmo *omni-aiNet* trabalhasse apenas com matrizes de distâncias como indivíduos (*genótipo*) e o NJ fosse utilizado para converter tais matrizes de distâncias em árvores filogenéticas propriamente ditas (*fenótipo* dos indivíduos).

Esta abordagem permitiu manter a codificação real usada pela *omni-aiNet* original e, ao mesmo tempo, permitiu indiretamente a evolução de árvores com topologias distintas, sem a necessidade do uso de estruturas de dados refinadas para manipulação de árvores.

Resumindo, cada indivíduo do algoritmo corresponde então a uma dada matriz de distâncias que, ao ser submetida ao algoritmo *Neighbor-Joining*, resulta em uma dada árvore filogenética com *evolução mínima* e *erro quadrático mínimo* próprios. Uma ilustração deste mecanismo de codificação e o mecanismo de conversão do indivíduo codificado (genótipo) para a respectiva árvore filogenética (fenótipo) é dada na Figura 6.8.

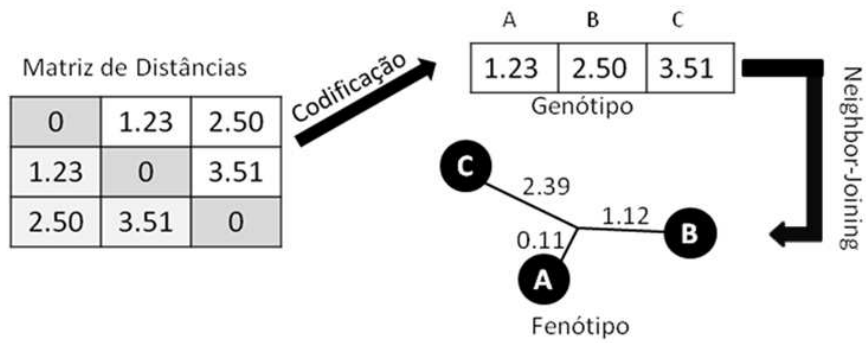


Figura 6.8 - Exemplo do mecanismo de codificação de indivíduos na omni-aiNet

6.3.2 Demonstração da Universalidade do Construtor

A codificação de indivíduos apresentada na seção 6.3.1 pode ser resumida de acordo com os seguintes passos:

→ apresentação de uma matriz de distâncias ao algoritmo omni-aiNet;

→ essa matriz torna-se o primeiro indivíduo da população, sendo que os demais indivíduos são produzidos a partir da introdução de perturbações aleatórias junto aos elementos da matriz de distâncias apresentada ao algoritmo, visando assim construir árvores que se encontrem predominantemente em uma vizinhança topológica, mas que sejam arbitrariamente distintas, em termos de topologia, da árvore que resulta da matriz de distâncias apresentada ao algoritmo;

→ cada genótipo, ou seja, cada vetor correspondente a um indivíduo da população (matriz de distâncias), é submetido ao NJ, o qual serve como construtor para o fenótipo do indivíduo, no caso, a árvore.

A execução dos passos mencionados acima leva à uma forma completa de representação de indivíduos. Com essa representação, pode-se gerar qualquer árvore, ou seja, uma árvore com qualquer topologia e quaisquer comprimentos de ramos, utilizando o NJ como construtor a partir de uma matriz de distâncias. Considere o exemplo da Figura 6.9.

(a) (b)

$$D = \begin{bmatrix} 0 & & & & & & & \\ 7 & 0 & & & & & & \\ 8 & 5 & 0 & & & & & \\ 11 & 8 & 5 & 0 & & & & \\ 13 & 10 & 7 & 8 & 0 & & & \\ 16 & 13 & 10 & 11 & 5 & 0 & & \\ 13 & 10 & 7 & 8 & 6 & 9 & 0 & \\ \underline{17} & 14 & 11 & 12 & 10 & 13 & 8 & \underline{0} \end{bmatrix}$$

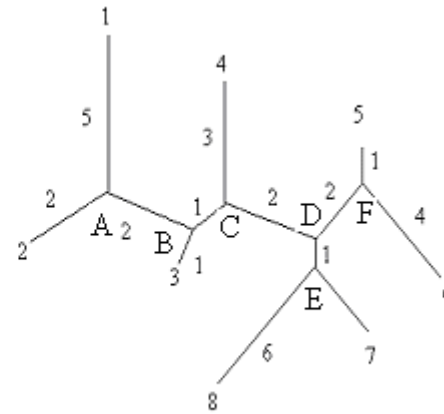


Figura 6.9 - (a) Matriz de distâncias para oito espécies. (b) árvore correspondente à matriz da Figura 6.9(a).

A partir de uma matriz de distâncias, o NJ sempre será capaz de gerar uma árvore com comprimento de ramos, conforme exemplo da Figura 6.9. Se aplicarmos uma pequena perturbação na árvore da Figura 6.9b, por exemplo, um δ=0.2 para o ramo 2-A (Figura 6.10b), haverá um efeito na matriz de distâncias refletindo essa perturbação, conforme

mostrado na Figura 6.10a. Se o mesmo procedimento for feito para todos os ramos terminais, haverá também um efeito na matriz de distâncias, embora para elementos distintos da matriz, dependendo do ramo alterado na árvore.

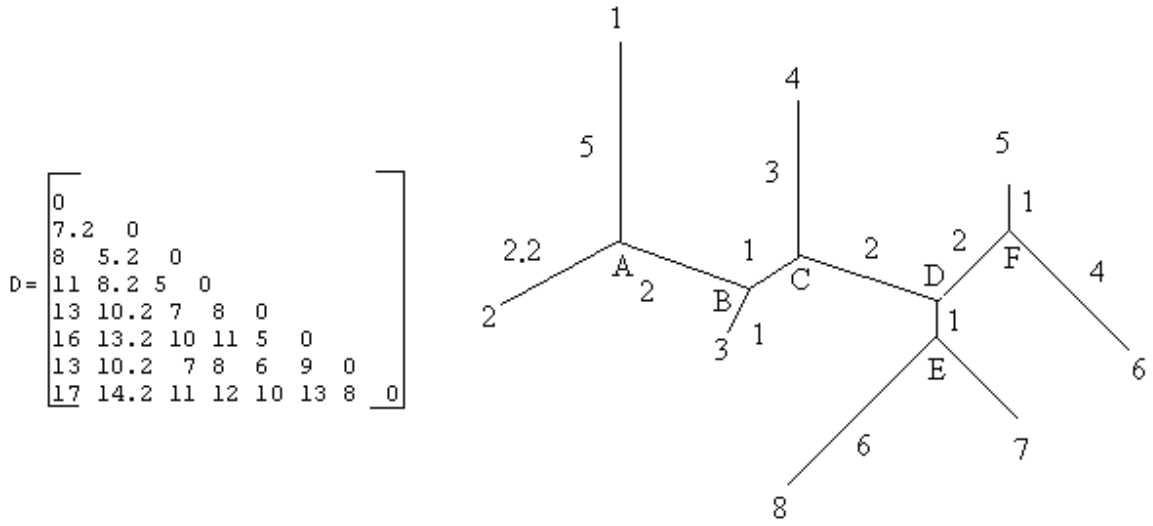


Figura 6.10 - (a) Matriz de distâncias refletindo perturbação aplicada à árvore da Figura 6.9b. (b) Árvore resultante da aplicação de perturbação no ramo 2-A da árvore da Figura 6.9b.

Novamente, se aplicarmos uma pequena perturbação na árvore da Figura 6.9b, por exemplo, um δ=0.2 para o ramo B-C, haverá um efeito na matriz de distâncias refletindo essa perturbação, conforme mostrado na Figura 6.11a. Se o mesmo procedimento for feito para todos os ramos internos, haverá também um efeito na matriz de distâncias, embora para elementos distintos da matriz, dependendo do ramo alterado na árvore.

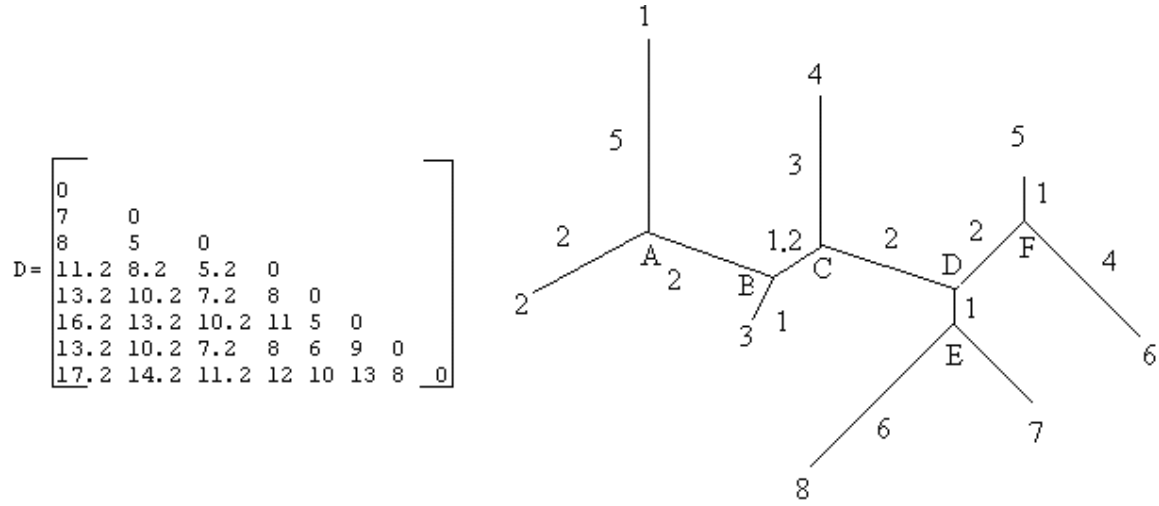


Figura 6.11 - (a) Matriz de distâncias refletindo perturbação aplicada à árvore da Figura 6.11b.(b) Árvore resultante da aplicação de perturbação no ramo B-C da árvore da Figura 6.9b.

Dessa forma, toda alteração realizada em uma árvore resulta em uma única e válida alteração na matriz de distâncias. Assim, se existe uma matriz de distâncias que leva o NJ a produzir uma certa topologia de árvore T_1, então existem matrizes de distância para qualquer outra topologia possível $T_{possível}$ (com o mesmo número de nós terminais), pois $T_{possível}$ pode ser obtida introduzindo um número finito de alterações controladas sobre T_1.

6.3.3 Geração da População Inicial

No algoritmo omni-ainet, a população inicial de N_{inic} indivíduos é gerada aleatoriamente dentro do domínio do problema. Tal abordagem poderia ser utilizada aqui também, mas como temos uma matriz de distâncias inicial fornecida com o problema, decidiu-se inicializar os indivíduos do algoritmo através da aplicação de pequenas perturbações aleatórias a esta matriz inicial. Dessa forma, os indivíduos tendem a ser inicializados em uma região mais próxima à árvore gerada para a matriz original do que no caso de indivíduos inicializados de forma totalmente aleatória.

É importante ressaltar que tais matrizes de distância nos indivíduos da população não têm um significado relevante e servem apenas como entradas adequadas ao algoritmo *Neighbor-Joining*, que gerará uma árvore com significado para o problema. O objetivo aqui é encontrar matrizes de distância que, quando submetidas ao *Neighbor-Joining*, produzam árvores filogenéticas de alta qualidade. Dessa forma, a aplicação da *omni-aiNet* para geração de um conjunto de matrizes de distância que levem a árvores de qualidade pode ser vista como o resultado da aplicação de perturbações (proporcionais à qualidade individual de cada elemento na população, como foi visto na Seção 6.2.6) à matriz de distâncias original, que possam compensar os resultados não satisfatórios gerados pelo *Neighbor-Joining,* quando aplicado diretamente sobre a matriz original.

6.3.4 Afinidade entre Anticorpos e Supressão

A métrica de afinidade entre anticorpos (indivíduos da população) do algoritmo *omni-aiNet* em conjunto com sua etapa de supressão é um dos mecanismos-chave para manutenção de

uma boa diversidade entre os elementos da população, que leva o algoritmo a executar uma exploração mais ampla do espaço de buscas, permitindo assim a obtenção de melhores soluções e o melhor espalhamento das soluções sobre a fronteira de Pareto.

No algoritmo original, a métrica de afinidade entre anticorpos é dada pela distância euclidiana entre dois anticorpos (vide Seção 6.2.6). No entanto, se tomarmos aqui esta métrica como indicador do grau de dissimilaridades entre dois indivíduos, não estaremos medindo diretamente um grau de diferença entre duas árvores filogenéticas, uma vez que a distância euclidiana entre as matrizes de distância não indica diretamente a distância entre as árvores correspondentes no espaço de topologias, pois a mesma não leva em conta a diferença entre as topologias, mas apenas os comprimentos dos ramos. Como isso poderia levar a resultados equivocados, a distância euclidiana do algoritmo original foi substituída pela métrica de *Robinson-Foulds* (vide seção 4.3.2), que é capaz de avaliar o quão diferente uma dada árvore é de outra. Como esta métrica também é simétrica e retorna valores reais, tal substituição pôde ser feita sem que fossem necessárias modificações mais profundas no algoritmo *omni-aiNet*.

6.3.5 Factibilidade de Soluções

Um último item que deve ser ressaltado são as restrições de factibilidade das soluções geradas pela *omni-aiNet*. No algoritmo original, desenvolvido para otimização de funções, uma dada solução é dita factível se atender às restrições de igualdade, desigualdade e de domínio. Neste trabalho o procedimento é análogo, exceto que não existe nenhuma restrição de igualdade e devem ser analisados não só os genótipos dos indivíduos (suas matrizes de distância), mas também seus fenótipos (as árvores filogenéticas resultantes), uma vez que a simples verificação de não-negatividade dos elementos das matrizes de distância (não faz sentido a presença de distâncias negativas) não garante que serão geradas árvores com comprimento de ramos não-negativos.

Comprimentos de ramos negativos podem ser gerados pelo *Neighbor-Joining* graças à presença de valores com ruído na matriz de distâncias, que podem potencialmente levar a

distâncias não-aditivas nas matrizes (ATTESON, 1999). Dessa forma, além de analisar a não-negatividade das matrizes de distâncias (genótipos) para determinar se uma dada solução é factível, devemos analisar também se o comprimento de cada ramo da árvore gerada pelo *Neighbor-Joining* também é positivo.

Capítulo 7

Resultados Experimentais em Filogenia via uma abordagem multi-objetivo

Resumo – O objetivo desse capítulo é apresentar os principais resultados obtidos a partir da utilização do algoritmo omni-aiNet para geração de árvores filogenéticas alternativas. Três casos de estudo foram adotados:o primeiro deles compara as árvores geradas pela execução da *omni-aiNet* (adaptada para reconstrução de árvores filogenéticas) com a árvore gerada pelo *Neighbor-Joining*. No segundo, o algoritmo *Multi-Neighbor Joining* foi utilizado e as árvores alternativas propostas por este algoritmo foram então comparadas à árvore original gerada pelo *Neighbor-Joining* e às árvores geradas pela *omni-aiNet*. No último, um exemplo de execução da omni-aiNet foi executado e comparado aos resultados da árvore NJ e da árvore gerada pelo algoritmo de verossimilhança máxima.

7.1 *Omni-aiNet* e a solução NJ

Para ilustrar a diferença entre os resultados gerados pela *omni-aiNet* e pelo *Neighbor-Joining ,* dois exemplos didáticos de matrizes de distância foram utilizados.

Para o primeiro exemplo, a matriz aditiva representada na Figura 7.1(a) foi utilizada (SAITOU & NEI, 1987) . Esta matriz D1 mostra os valores para as distâncias observadas entre oito espécies. A Figura 7.1(b) mostra a árvore gerada pelo *Neighbor-Joining* tendo como entrada a matriz D1.

Os resultados obtidos a partir da matriz de distâncias D1, representada na Figura 7.1(a), tanto pela *omni-aiNet* (círculos em branco) quanto pelo *Neighbor-Joining* (círculo em cinza) podem ser observados na Figura 7.2. Nesse exemplo, podemos perceber que a árvore

original faz parte da superfície de Pareto, ou seja, é uma solução não-dominada e possui EQM de valor zero.

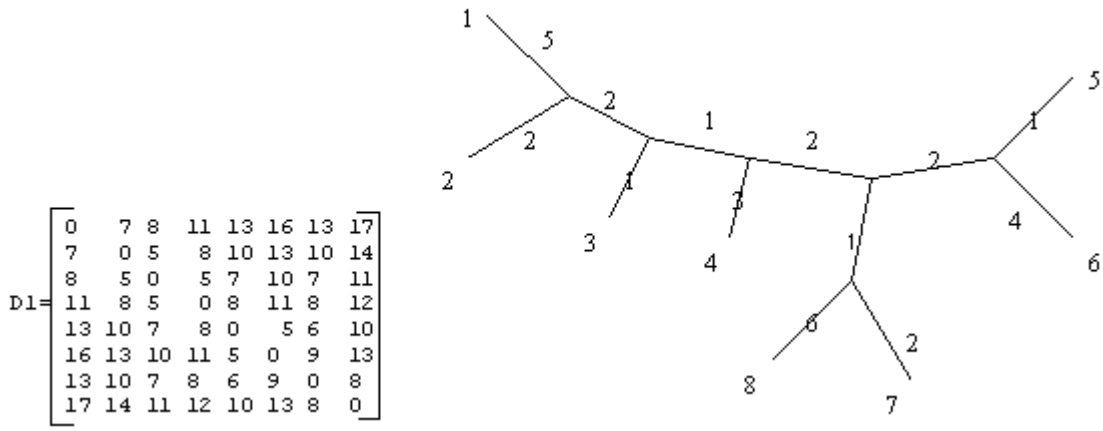


Figura 7.1(a) - Matriz aditiva com oito espécies. (b) Topologia e comprimento de ramos obtidos pelo *Neighbor-Joining*, tendo como entrada a matriz D1.

Isto acontece em razão da matriz de distâncias observadas ser uma matriz aditiva, ou seja, a diferença entre a matriz de distâncias da árvore gerada (patrística) e a matriz observada é zero, o que indica que o *Neighbor-Joining* foi capaz de encontrar a solução ótima para o problema. É importante notar que, mesmo para este problema, o algoritmo *omni-aiNet* foi capaz de encontrar várias soluções, diferentes da encontrada pelo *Neighbor-Joining*, e que, apesar de apresentarem EQM maior, possuem evolução mínima menor, ou seja, soluções que não são dominadas pela solução encontrada pelo *Neighbor-Joining*.

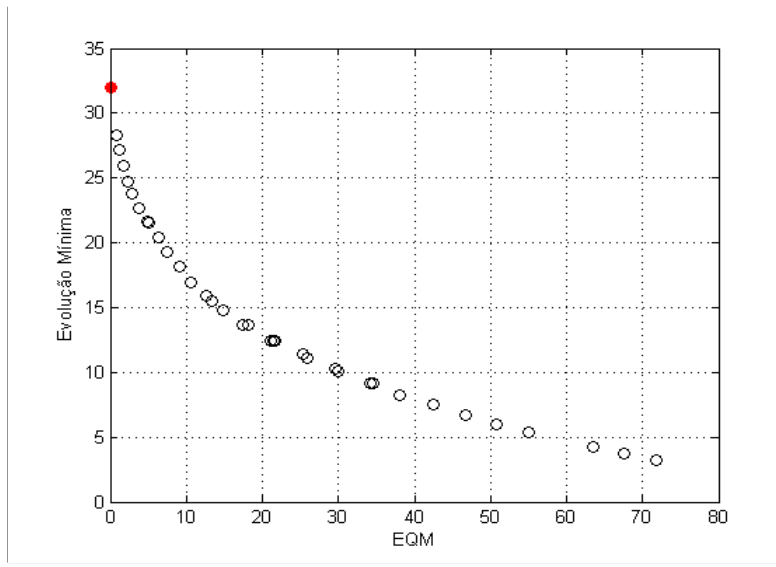


Figura 7.2 - Fronteira de Pareto obtida pela *omni-aiNet* (círculos brancos) e pelo *Neighbor-Joining* (círculo cinza) para a matriz de distâncias D1, representada na Figura 7.1(a).

Para o segundo exemplo desta primeira etapa, foi utilizada a mesma matriz de distâncias D1, mas acrescida de pequenas perturbações que têm a finalidade de transformar a matriz originalmente aditiva em uma matriz não-aditiva, representada na Figura7.3(a). A quebra de aditividade é dada pela não observância da desigualdade triangular entre as espécies 6 e 8, considerando-se o terceiro ponto como a espécie 5. Esta desigualdade é dada por: $d_{6,8} \geq d_{6,5} + d_{5,8}$, o que infringe uma das condições de aditividade.

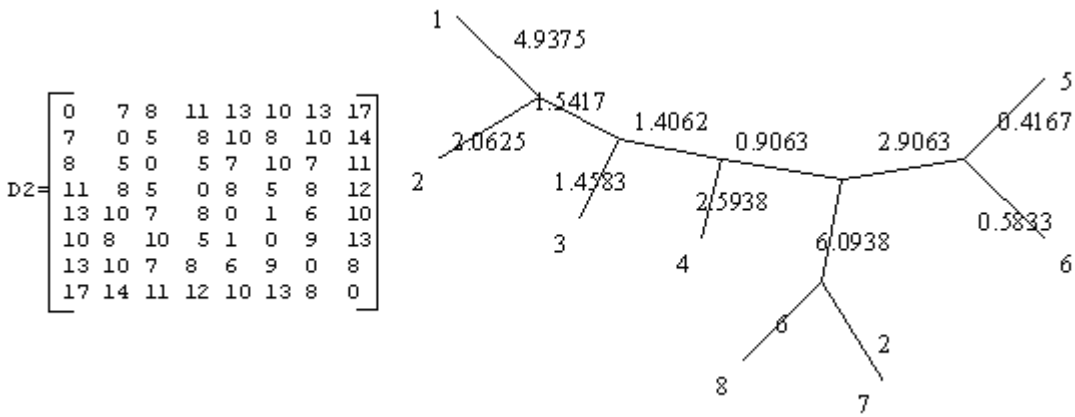


Figura 7.3 - (a) Matriz não aditiva com oito espécies. (b) Topologia e comprimento de ramos obtidos pelo *Neighbor-Joining* tendo como entrada a matriz D2.

A partir da matriz D2, o algoritmo *Neighbor-Joining* deu origem à árvore apresentada na Figura 7.3(b). O resultado das árvores evoluídas pela *omni-aiNet* a partir da matriz de distâncias D2 dada pela Figura 7.3(a) pode ser observado na Figura 7.4 (círculos pretos). O ponto representado por um losango corresponde à árvore gerada pelo *Neighbor-Joining* (Figura7.3(b)). Além disso, estão representadas graficamente três árvores geradas pela *omni-aiNet*, além da árvore gerada pelo NJ.

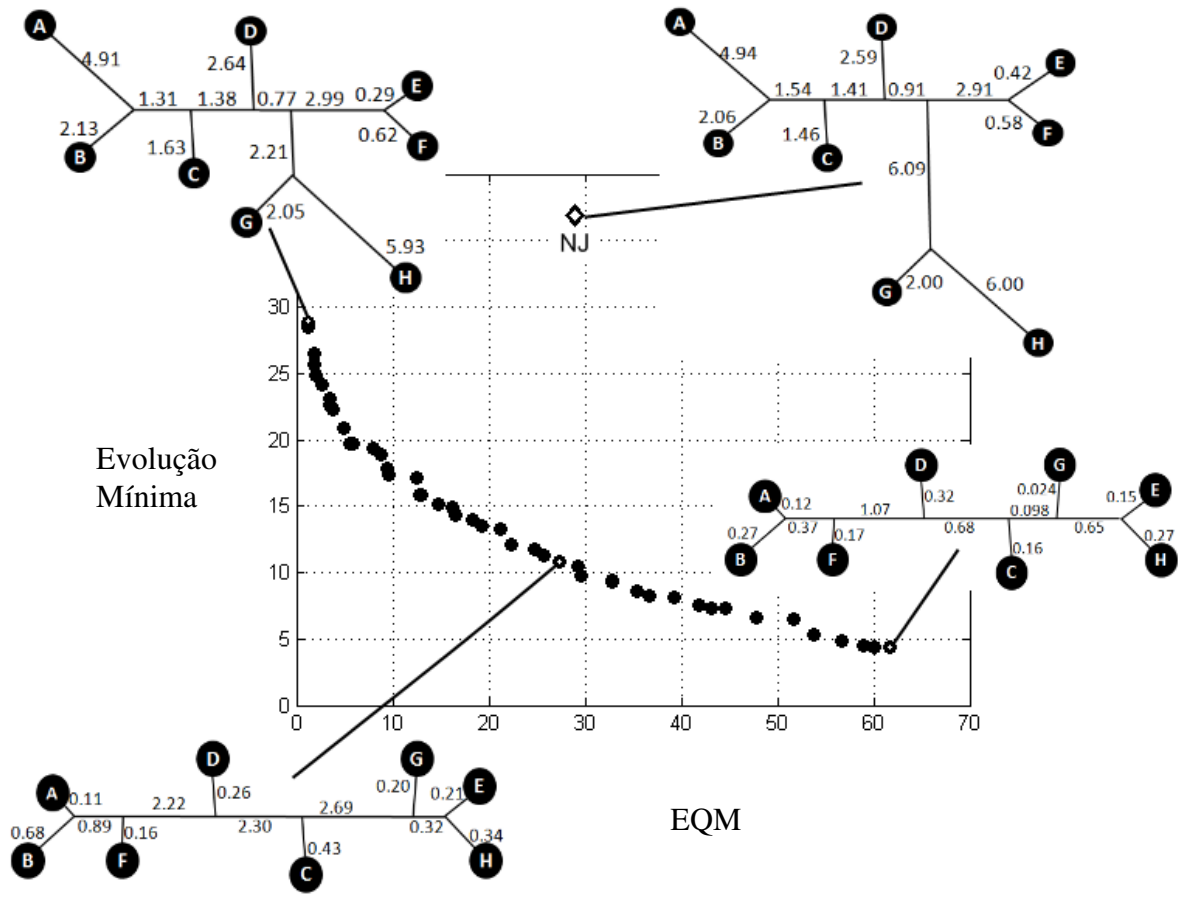


Figura 7.4 - Árvores obtidas pela *omni-aiNet* (círculos pretos) e pelo *Neighbor-Joining* (losango) para a matriz de distâncias D2, apresentada na Figura 7.3(a).

Nesse exemplo, é possível perceber que a árvore gerada pelo NJ não faz parte da fronteira de Pareto, ou seja, é uma solução dominada pelas outras soluções geradas pela *omni-aiNet*. Isto acontece em razão da matriz de distâncias utilizada não ser uma matriz aditiva, ou seja, a diferença entre a matriz de distâncias observadas e a matriz patrística é diferente de zero. Nesses casos, para os quais a matriz de distâncias não é uma matriz aditiva, a solução NJ será sub-ótima e, portanto, não estará sobre a superfície de Pareto. É importante notar aqui que as soluções geradas pela *omni-aiNet*, mesmo partindo da mesma matriz que o algoritmo *Neighbor-Joining*, são superiores (boa parte destas soluções domina a solução NJ), o que mostra bem a capacidade de geração de árvores de boa qualidade pela *omni-aiNet*.

Podemos observar também na Figura 7.4 que o algoritmo *omni-aiNet* é capaz de gerar árvores de topologias distintas (veja, por exemplo, as árvores de menor EQM e menor EM, respectivamente), além de árvores de mesma topologia e comprimento de ramos diferentes (por exemplo, as duas árvores representadas na parte inferior da Figura 7.4). Outro ponto a ser ressaltado é que a árvore de menor EQM possui a mesma topologia que a árvore gerada pelo NJ, mas apresenta comprimentos diferentes para os ramos, o que indica que a deficiência do algoritmo *Neighbor-Joining* para este problema se deu na determinação do comprimento dos ramos da árvore.

7.1.1 Soluções de Consenso

As soluções apresentadas na Figura 7.4 foram escolhidas, de maneira a representar o espaço de soluções da fronteira de Pareto estimada. O usuário final que fará a análise das árvores poderá escolher a que melhor explique os aspectos evolutivos dos dados de entrada. Uma outra forma de proporcionar alternativas ao usuário final seria apresentar, juntamente com essas soluções, uma solução de consenso. Representando as soluções da Figura 7.4 conforme a notação Newick, tem-se as seguintes topologias:

1. Solução com menor EQM: (((((A,B),C),D),(G,H)),(E,F));
2. Solução NJ: (((((A,B),C),D),(G,H)),(E,F));
3. Solução com menor EM: ((((((A,B),F),D),C),G),(E,H));
4. Solução com EQM e EM médios: ((((((A,B),F),D),C),G),(E,H)).

Utilizando a técnica de consenso estrito para as soluções 1, 2, 3 e 4, resulta a topologia: ((A,B),C,D,E,F,G,H). Essa solução é apresentada pela Figura 7.5.

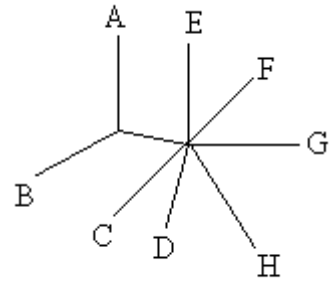


Figura 7.5 - Solução de consenso estrito entre as árvores da Figura 7.4

7.2 *Omni-aiNet* e *Multi-Neighbor Joining*

Nesta seção, novamente dois exemplos são apresentados, sendo que aqui, além da comparação entre as soluções geradas pela *omni-aiNet* e pelo *Neighbor-Joining* (NJ), também serão analisadas as soluções geradas pelo *Multi-Neighbor Joining* (MNJ).

No primeiro exemplo, a matriz de distâncias D3, apresentada na Figura 7.6, foi utilizada como entrada dos três algoritmos. A matriz representa dados de um exemplo didático. O MNJ apresentou, além da árvore do *Neighbor-Joining* clássico (chamada aqui de árvore NJ), outras sete árvores como soluções alternativas. Tais árvores (em notação *Newick*) estão representadas a seguir, em conjunto com seus respectivos comprimentos.

$$D3=\begin{bmatrix}
0 & 5.1680 & 5.1680 & 5.1680 & 3.9520 & 3.9520 & 5.1680 & 3.9520 \\
5.1680 & 0 & 3.0400 & 3.9520 & 4.8640 & 5.7760 & 4.8640 & 3.6480 \\
5.1680 & 3.0400 & 0 & 5.1680 & 5.4720 & 5.1680 & 5.4720 & 4.5600 \\
5.1680 & 3.9520 & 5.1680 & 0 & 4.5600 & 4.5600 & 3.3440 & 3.6480 \\
3.9520 & 4.8640 & 5.4720 & 4.5600 & 0 & 3.9520 & 4.5600 & 3.6480 \\
3.9520 & 5.7760 & 5.1680 & 4.5600 & 3.9520 & 0 & 4.5600 & 4.2560 \\
5.1680 & 4.8640 & 5.4720 & 3.3440 & 4.5600 & 4.5600 & 0 & 4.5600 \\
3.9520 & 3.6480 & 4.5600 & 3.6480 & 3.6480 & 4.2560 & 4.5600 & 0
\end{bmatrix}$$

Figura 7.6 - Matriz D3, correspondente às distâncias entre oito espécies

- **Árvore NJ**

(((((N1:1.976,N6:1.976):0.152,N5:1.824):0.399,N8:1.501):0.171,(N4:1.52,N7:1.824):2.299):1.045,(N2:1.292,N3:1.748))

Comprimento: 17.727

- **Árvore 1**

(((((N2:1.292,N3:1.748):1.121,N8:1.463):0.114,(N4:1.52,N7:1.824):0.646):0.418,N5:1.824):0.152,(N6:1.938,N1:2.014))

Comprimento: 16.074

- **Árvore 2**

(((((N2:1.292,N3:1.748):1.045,(N4:1.4187,N7:1.9253):2.299):0.171,N8:1.501):0.399,N5:1.824):0.152,(N1:2.0368,N6:1.9152))

Comprimento: 17.727

- **Árvore 3**

(((((N2:1.292,N3:1.748):1.0893,(N4:1.52,N7:1.824):0.5827):0.171,N8:1.501):2.375,N5:1.805):0.171,(N1:1.976,N6:1.976))

Comprimento: 18.031

- **Árvore 4**

(((((N2:1.292,N3:1.748):1.121,N8:1.463):0.114,(N4:1.52,N7:1.824):2.318):0.418,N5:1.824):0.152,(N6:1.9253,N1:2.0267))

Comprimento: 17.746

- **Árvore 5**

(((((N2:1.292,N3:1.748):1.045,(N4:1.406,N7:1.938):2.964):0.171,N8:1.5453):0.3547,N5:1.824):0.152,(N1:2.0368,N6:1.9152))

Comprimento: 18.392

- **Árvore 6**

(((((N2:1.292,N3:1.748):1.0893,(N4:1.52,N7:1.824):0.5827):0.171,N8:1.501):0.399,N5:1.805):0.399,(N1:1.979,N6:1.976))

Comprimento: 16.055

- **Árvore 7**

(((((N2:1.292,N3:1.748):1.121,N8:1.463):0.114,(N4:1.52,N7:1.824):0.646):0.475,N6:1.995):0.019,(N5:1.881,N1:2.071))

Comprimento: 16.131

Os resultados obtidos pela *omni-aiNet* a partir da matriz de distâncias D3 podem ser observados na Figura 7.7 (círculos brancos). Os pontos coloridos representam os valores das funções-objetivo para as soluções alternativas apresentadas pelo MNJ. Nesse exemplo, os valores relativos à árvore do NJ original não fazem parte da fronteira de Pareto, mas os valores relativos a uma das árvores alternativas (de número 1) estão localizados na fronteira de Pareto estimada. A Figura 7.8 apresenta um gráfico ampliado mostrando apenas os valores das funções-objetivo para as oito árvores retornadas pelo MNJ.

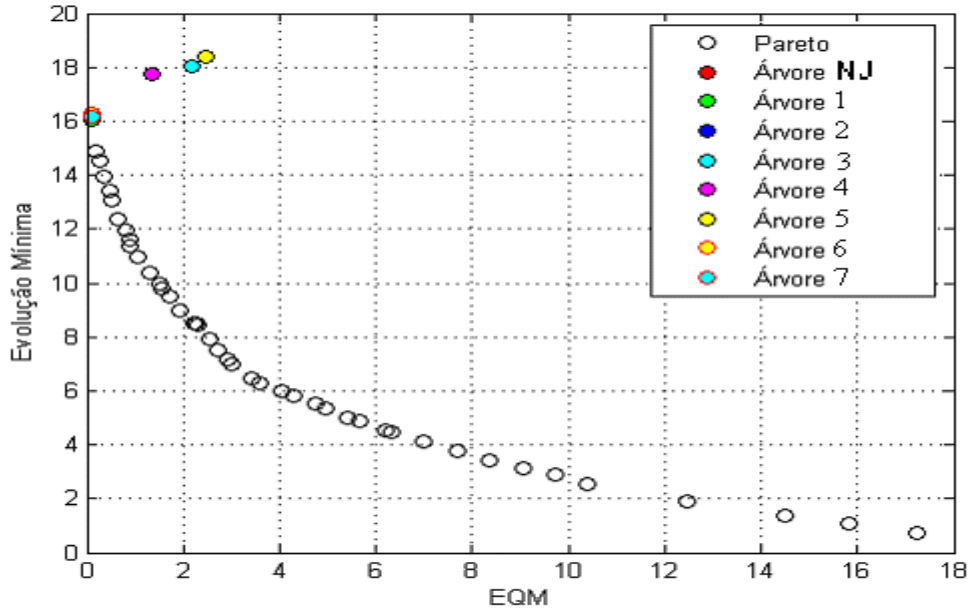


Figura 7.7 - Resultado da evolução de árvores (a partir da matriz D3) pela *omni-aiNet* (círculos brancos), árvore gerada pelo NJ clássico (círculo vermelho) e valores das funções-objetivo para as árvores alternativas propostas pelo MNJ (círculos coloridos).

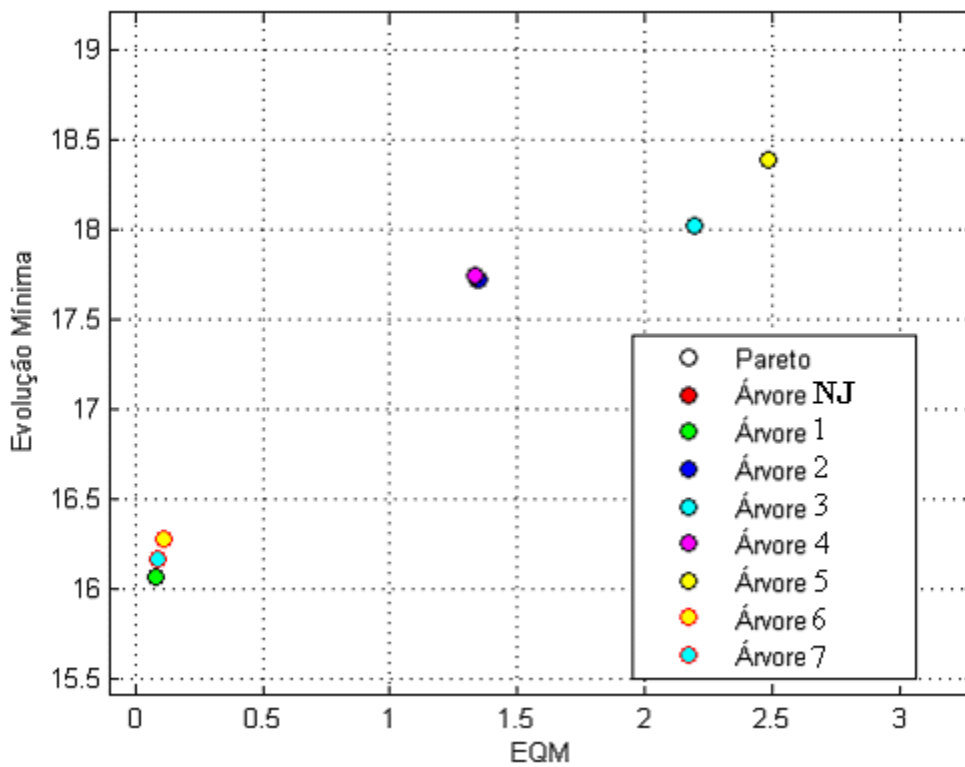


Figura 7.8 - Distribuição dos valores das funções-objetivo relativos às oito árvores retornadas pelo MNJ.

Para esse exemplo, pôde-se observar que as oito árvores mantiveram em suas topologias as sub-árvores (N2, N3) e (N4, N7). Para sete das árvores, além das sub-árvores (N2, N3) e (N4, N7), a sub-árvore (N1, N6) foi incluída em suas topologias. A exceção diz respeito à árvore 7, a qual é a única a incluir a sub-árvore (N1, N5) em sua topologia. Além disso, a árvore 1 é uma solução que está localizada sobre a fronteira de Pareto encontrada pela *omni-aiNet*, o que a inclui no conjunto de possíveis soluções para este problema.

No segundo exemplo tratado aqui, a matriz de distâncias adotada foi a mesma utilizada no exemplo dos roedores apresentado no Capítulo 5 (esta matriz de distâncias é dada no Apêndice B). A árvore original gerada pelo NJ para este problema está representada na Figura 7.9. O resultado da execução da *omni-aiNet* para este exemplo é apresentado na Figura 7.10.

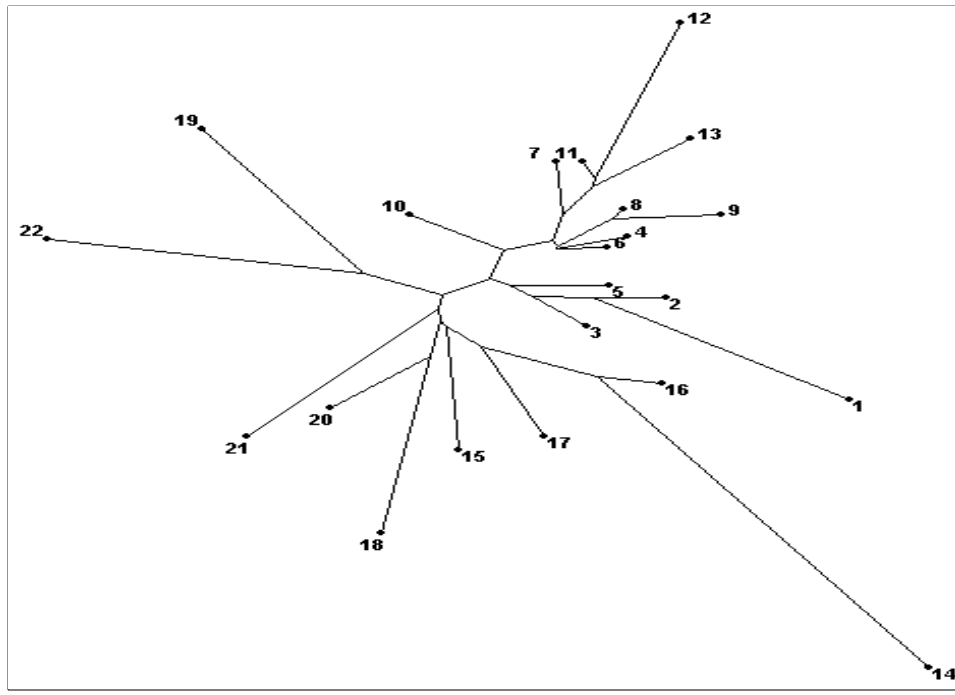


Figura 7.9 - Árvore original gerada pelo NJ.

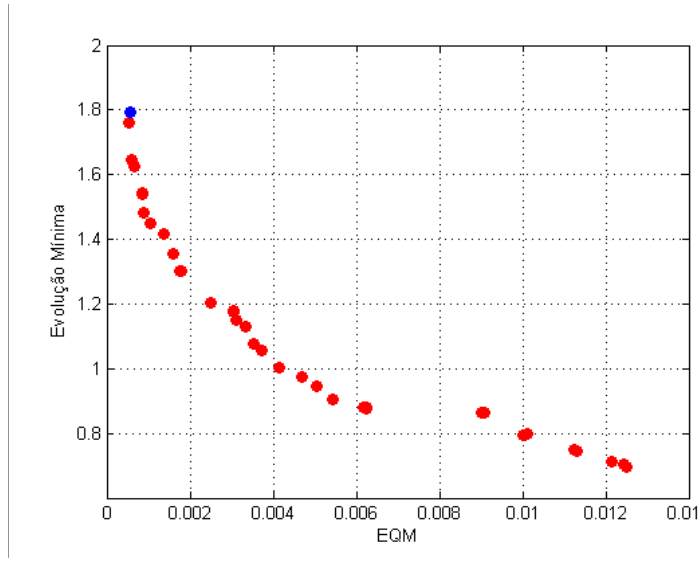


Figura 7.10 - Resultado da execução da *omni-aiNet* para o problema dos roedores (em vermelho) e a árvore NJ (em azul).

A Figura 7.11 apresenta os valores das funções-objetivo para a árvore original (em azul), para as soluções alternativas geradas pelo MNJ (em preto) e a fronteira de Pareto encontrada pela *omni-aiNet* (pontos em vermelho).

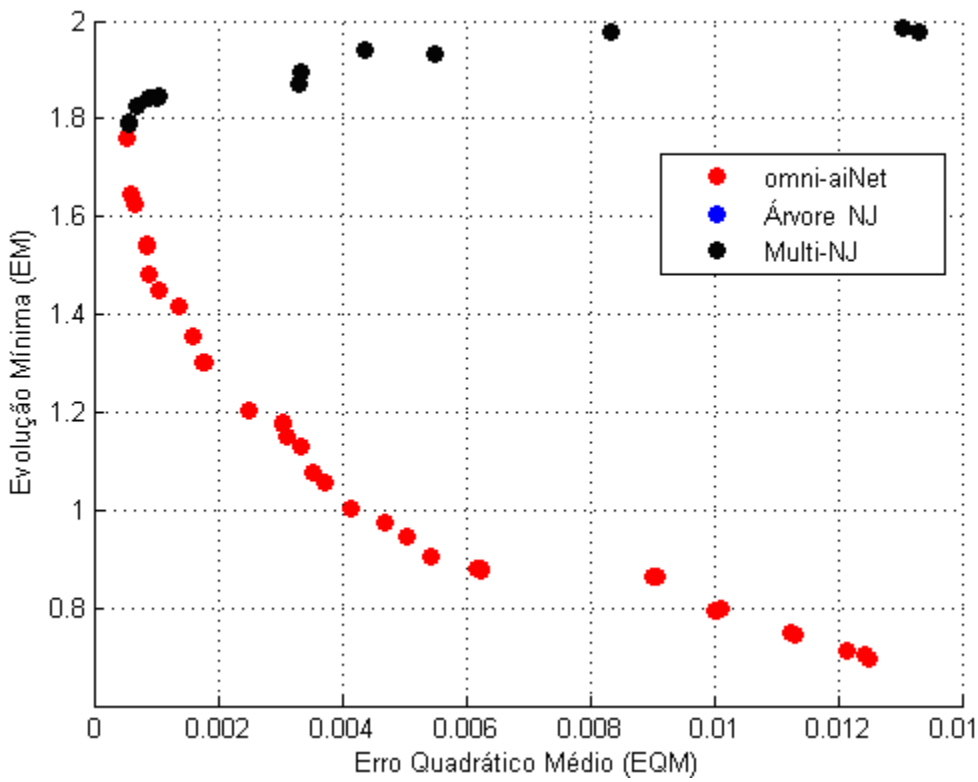


Figura 7.11 - Fronteira de Pareto encontrada pela *omni-aiNet* (em vermelho), solução NJ (em azul) e as soluções alternativas geradas pelo MNJ (em preto).

Como pode ser observado na Figura 7.11, a fronteira de Pareto estimada encontrada pela *omni-aiNet* para este problema não é tão suave quando à obtida para os demais problemas, o que indica que ainda há espaço para melhorias nas soluções encontradas, ou seja, tais soluções provavelmente não correspondem ao conjunto de Pareto para o problema. No entanto, mesmo assim estas soluções não são dominadas por nenhuma das outras soluções propostas pelo MNJ nem pela solução NJ do problema, o que novamente evidencia a capacidade da metodologia proposta em obter um conjunto de possíveis soluções de bom desempenho para o problema de reconstrução de árvores filogenéticas.

Algumas das soluções apresentadas pelo MNJ ficaram muito distantes da fronteira de Pareto estimada. Essas soluções provavelmente priorizaram agrupamentos de pares de espécies que as levaram a valores ruins de EMQ e Evolução Mínima, distanciando-as da árvore NJ. Outro fator que pode ter influenciado o aparecimento de soluções do MNJ distantes da fronteira estimada é o limite de escolha de árvores alternativas. Como essa heurística não cobre todo o espaço de árvores, algumas boas soluções podem ter sido desprezadas.

7.3 *Omni-aiNet*, a solução NJ e a solução de verossimilhança máxima

Nesse exemplo, os seguintes passos foram seguidos. Primeiramente, um conjunto de seqüências de DNA (WELLS *et al.*, 2001) de espécies de Sarcophagidae foi utilizado como entrada para o algoritmo de cálculo da árvore de verossimilhança máxima (PHYLIP, 2007). As seqüências utilizadas encontram-se no Apêndice C. O resultado da execução do algoritmo de verossimilhança máxima é apresentado na Figura 7.12.

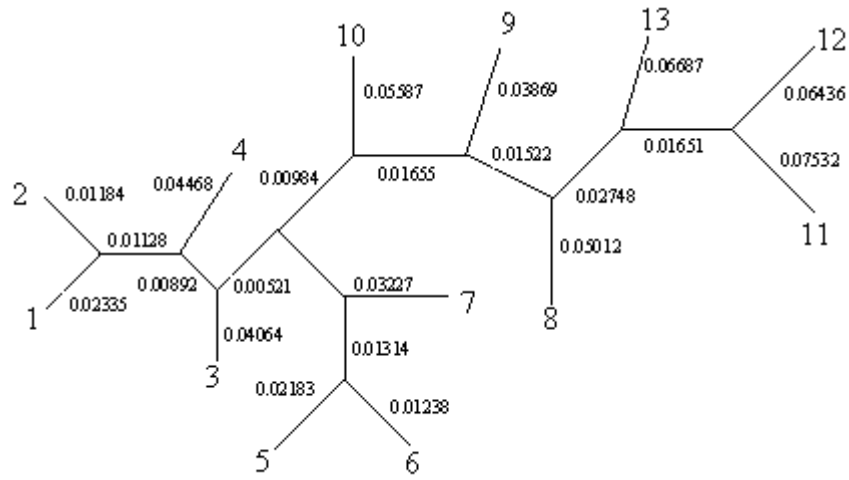


Figura 7.12 - Árvore resultante da execução do algoritmo de verossimilhança máxima.

A Figura 7.13 apresenta a matriz de distâncias utilizada como indivíduo inicial para a execução do omni-aiNet. Essa matriz corresponde às distâncias entre as seqüências utilizadas para a execução do algoritmo que resultou na árvore da Figura 7.12.

	1	2	3	4	5	6	7	8	9	10	11	12	13
1													
2	0.033												
3	0.061	0.054											
4	0.075	0.066	0.068										
5	0.068	0.056	0.060	0.034									
6	0.084	0.072	0.080	0.070	0.061								
7	0.063	0.072	0.074	0.066	0.063	0.077							
8	0.084	0.079	0.092	0.084	0.087	0.092	0.092						
9	0.095	0.089	0.086	0.075	0.077	0.093	0.091	0.091					
10	0.096	0.088	0.100	0.101	0.097	0.109	0.107	0.112	0.087				
11	0.128	0.117	0.120	0.121	0.112	0.128	0.120	0.120	0.110	0.121			
12	0.138	0.130	0.129	0.128	0.128	0.147	0.132	0.135	0.124	0.121	0.129		
13	0.125	0.123	0.119	0.123	0.126	0.137	0.138	0.125	0.129	0.121	0.123	0.124	

Figura 7.13 - Matriz de distâncias utilizada como entrada da omni-aiNet.

A Figura 7.14 apresenta o resultado da execução do NJ para a matriz da Figura 7.13.

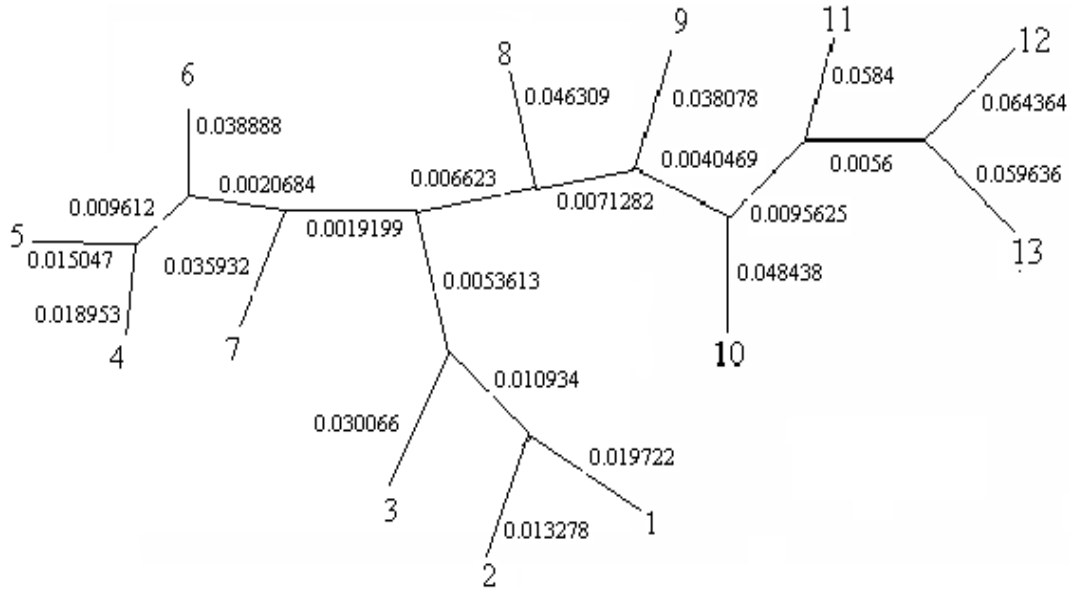


Figura 7.14 - Árvore resultante da execução do NJ para a matriz da Figura 7.13.

A Figura 7.15 apresenta o resultado da execução da omni-aiNet para a matriz da Figura 7.13.

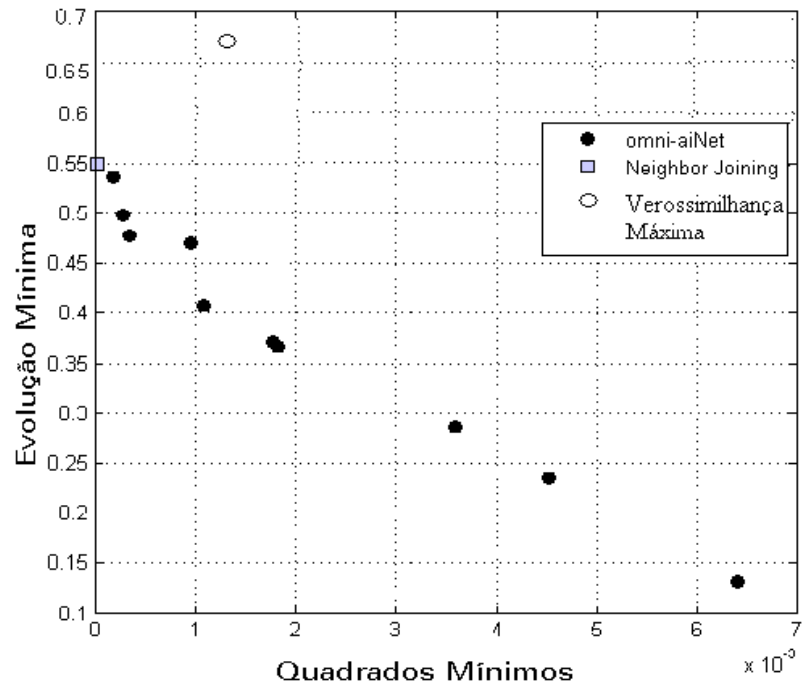


Figura 7.15 - Resultado da execução do omni-aiNet para a matriz da Figura 7.13.

Para esse exemplo, a árvore NJ faz parte da superfície de Pareto com valor de evolução mínima 0,55 e erro quadrático médio 0 (zero). Para a árvore gerada pelo algoritmo de

verossimilhança máxima, o valor de evolução mínima foi 0,675 e o valor do erro quadrático médio foi 0,00136. A solução NJ mostrou-se mais apropriada que a solução do algoritmo de verossimilhança máxima, sempre lembrando que estão sendo levados em conta apenas os critérios de evolução mínima e quadrados mínimos.

7.4 Comentários Finais

Neste capítulo, foi apresentada uma metodologia imuno-inspirada e multi-objetivo para reconstrução de árvores filogenéticas a partir de matrizes de distâncias originais, baseada no algoritmo *omni-aiNet*, que leva em consideração simultaneamente os critérios de evolução mínima e erro quadrático médio, e se baseia no algoritmo *Neighbor-Joining* para fazer a conversão entre genótipo e fenótipo dos indivíduos da população. O principal objetivo desta proposta é fornecer ao usuário um conjunto de soluções que representem compromissos entre os objetivos em questão, de forma a permitir que o usuário tenha uma visão mais ampla das relações filogenéticas entre as espécies em estudo.

Foram feitos experimentos utilizando-se a metodologia proposta e comparando-a com as soluções geradas pelos algoritmos *Neighbor-Joining* clássico e *Multi-Neighbor Joining,* onde foi possível observar que para nenhuma situação tratada os algoritmos NJ e MNJ foram capazes de gerar soluções que dominassem as soluções geradas pela extensão proposta para a *omni-aiNet*. Isto evidencia que, por tratarem os critérios de evolução mínima e erro quadrático médio de forma iterativa, o NJ e o MNJ geralmente levam a soluções sub-ótimas. Além disso, esta não dominância das soluções geradas pelo algoritmo imuno-inspirado leva a crer que, apesar da fronteira de Pareto dos problemas tratados ser desconhecida, os resultados obtidos pela *omni-aiNet* provavelmente estão muito próximos desta fronteira.

O exemplo utilizando as seqüências de DNAs e o algoritmo de verossimilhança máxima mostrou-se também uma solução que foi dominada pelas soluções apresentadas pelo omni-aiNet e pela solução NJ, o que evidencia o fato de que diferentes critérios de otimalidade conduzem a árvores filogenéticas distintas.

Capítulo 8

Conclusão

O objetivo principal desse trabalho foi o de proporcionar a oportunidade de escolha a pesquisadores que se encontram frente a um problema de reconstrução de uma árvore filogenética. Foi enfatizado ao longo do texto que a inferência filogenética é um problema multimodal, ao apresentar ótimos locais quando se emprega busca gulosa, e ganha novas perspectivas ao ser interpretado como um problema multi-objetivo. Nos dois casos, a proposição de múltiplas soluções candidatas de alta qualidade, seguida ou não da aplicação de técnicas de consenso, torna-se um procedimento inerente às técnicas de reconstrução.

Como foi enfatizado ao longo do texto, os principais desafios do problema de reconstrução filogenética estão vinculados: (*i*) à ausência de informação acerca do processo evolutivo de diferenciação, pois tem-se apenas um 'retrato parcial' das diferenças atuais entre espécies; (*ii*) à possibilidade de emprego de vários critérios de otimização e modelos evolutivos; (*iii*) aos desafios dos problemas de otimização resultantes; e (*iv*) ao elevado custo computacional vinculado ao processo de reconstrução, o que também leva à existência de diversos processos de reconstrução já propostos na literatura.

A partir do cenário desafiador caracterizado pelos seguintes fatores:

- os dados disponíveis não são totalmente informativos;
- os problemas de otimização associados são multimodais;
- existem múltiplos critérios e modelos evolutivos que podem ser considerados para guiar o processo de reconstrução;
- existem múltiplas propostas para o processo de reconstrução;

esta tese procura oferecer ferramentas de reconstrução capazes de produzir múltiplas alternativas de árvores filogenéticas, além de mostrar que boa parte dessas árvore obtidas têm qualidade superior àquelas fornecidas por ferramentas de reconstrução bem difundidas na literatura.

Embora as abordagens propostas estejam restritas a árvores sem raiz e partam apenas de matrizes de distâncias como dados de entrada, elas admitem várias extensões importantes. Por exemplo, foram considerados como funções-objetivo na abordagem multi-objetivo os critérios de evolução mínima e de quadrados mínimos. Nada impede, no entanto, que outras funções-objetivo sejam empregadas e em número maior ou igual a dois.

Um dos focos desse trabalho foi o de aliar vantangens do algoritmo *Neighbor-Joining* (NJ) à apresentação de várias alternativas de soluções para uma única matriz de distâncias apresentada como entrada. O *Multi-Neighbor Joining* (MNJ) caracteriza-se como uma extensão do NJ, uma vez que preserva suas características heurísticas e de complexidade computacional, além de sempre incluir entre as soluções propostas a solução única do NJ. A partir dos exemplos executados para o MNJ, pôde-se concluir que, para matrizes não-aditivas, o algoritmo é indicado, pois pode produzir árvores com valores mais adequados para o critério de evolução mínima, isso quando comparado ao NJ.

Além da utilização do NJ como base para a criação do MNJ, o NJ foi utilizado como construtor na abordagem multi-objetivo. Como construtor, o NJ produz o fenótipo de qualquer matriz de distâncias apresentada ao algoritmo de busca no espaço de árvores filogenéticas, no caso, um algoritmo imuno-inspirado denominado *omni-aiNet*. Este algoritmo de busca sofreu adaptações e extensões para permitir sua aplicação no contexto de árvores filogenéticas. As múltiplas soluções não-dominadas encontradas são então apresentadas, representando uma aproximação da fronteira de Pareto. A partir daí, pode-se, por exemplo, escolher soluções que expliquem os fatores evolutivos de maneira mais adequada, ou então empregar alguma técnica de consenso.

Como trabalhos futuros, pretende-se incorporar outras funções-objetivo ao processo de reconstrução, assim como aplicar as ferramentas de reconstrução propostas a outros problemas práticos. Pretende-se também investir mais em ferramentas de visualização gráfica das árvores resultantes. Como uma proposta mais ambiciosa, vislumbra-se o tratamento multi-objetivo para reticulados filogenéticos ou redes filogenéticas.

Apêndice A

Definição dos principais conceitos envolvidos na reconstrução de árvores filogenéticas

Resumo – Este apêndice define e apresenta brevemente, em ordem alfabética, os principais conceitos de reconstrução de árvores filogenéticas utilizados nos capítulos desta tese.

Árvore filogenética: Representa graficamente as relações de parentesco entre grupos de taxa ou genes. Pode ser com raiz ou sem raiz e é um caso particular de um grafo, constituído por nós e arestas ou ramos. Um árvore apresenta um caminho único entre qualquer par de nós, seguindo pelas arestas. Os nós podem ser terminais, quando são chamados de folhas, e não-terminais. Um caso particular muito utilizado é a árvore binária, em que todos os nós não-terminais se ligam por aresta ou ramo a outros três nós.

Bases Nitrogenadas: Podem ser de 5 tipos: A (Adenina), C (Citosina), G (Guanina), T (Timina), ou U (Uracila) e fazem parte do DNA e RNA. As bases A-T ou A-U e C-G são complementares. Cada seqüência de 3 bases, chamada códon, irá determinar um aminoácido (unidade de uma proteína). O arranjo dos aminoácidos, por sua vez, irá determinar a estrutura e a função de uma proteína.

Bioinformática: Se ocupa da análise e organização de dados biológicos usando técnicas avançadas de computação. Tem amplo uso na pesquisa genômica.

DNA (ADN): Sigla em inglês para ácido desoxirribonucléico, um complexo filamento de substâncias químicas, em forma de hélice dupla, que se encontra principalmente no núcleo das células. Todos os cromossomos e genes compõem-se de DNA, o qual controla a atividade celular e é responsável pela informação hereditária a ser transmitida aos descendentes.

Espécie: Representa uma das unidades básicas de classificação biológica. Um grupo de organismos constitui uma espécie quando possuem características (genéticas e/ou morfológicas) presentes em todos os seus membros, e ausentes nos grupos relacionados a outras espécies, de modo que seus membros podem gerar descendentes férteis.

Espaço de árvores: Entidade matemática que define um espaço contínuo ou discreto em que cada topologia de árvore corresponde a um ponto do espaço. Quando o comprimento dos ramos não é determinado, o espaço é discreto, enquanto que para árvores com comprimento de ramos, o espaço é contínuo.

Folhas: Nós terminais de uma árvore.

Formato Newick: Método de representação de árvores filogenéticas que utiliza parênteses aninhados, sendo adequado para tratamento computacional. Uma sub-árvore é fechada em um par de parênteses, separados por uma vírgula. Essa sub-árvore será fechada em uma par de parênteses com a sub-árvore mais próxima, e assim sucessivamente.

Gene: Foi originalmente definido como a unidade básica de hereditariedade em organismos vivos, mas esta definição se mostrou inadequada. Um gene corresponde a um segmento de DNA que ocupa um lugar específico no cromossomo e determina um subconjunto de características do indivíduo, sendo transmitido de geração em geração. O ser humano tem cerca de 30 mil genes. Sabe-se que mais de mil genes, quando alterados, estão associados a doenças.

HTU (Unidade taxonômica hipotética, do inglês, Hypothetical Taxonomic Unit): Corresponde a grupos de taxa hipotéticos, ou seja, não observados na realidade, normalmente associados aos nós não-terminais de árvores filogenéticas. Alguns métodos de inferência filogenética, como o de Verossimilhança Máxima e o do Parcimônia Máxima, procuram operar com as configurações possíveis para as HTUs.

Matriz de distâncias: É uma matriz de pares de distâncias entre taxa, baseadas em algum subconjunto de atributos homólogos. Para dados moleculares, pode corresponder ao número de diferentes nucleotídeos observados entre cada par de taxa, ou alternativamente, pode ser a distância entre as seqüências, tomando algum modelo evolutivo. Suas distâncias poderiam igualmente ser baseadas em dados morfológicos, ecológicos e/ou comportamentais. A matriz é simétrica e, sendo n o número de folhas da árvore, tem dimensão $n \times n$ e contém $n(n-1)/2$ pares de distâncias.

Matriz patrística: É a matriz de pares de distâncias obtidas diretamente da árvore filogenética inferida. Evidentemente, a árvore filogenética deve apresentar o comprimento de todos os ramos.

Modelo Evolutivo: São modelos matemáticos que tentam representar o processo de substituição de bases nucleotídicas ao longo da evolução das espécies. Geralmente envolvem simplificações e aproximações, além de apresentarem parâmetros a serem determinados.

Nucleotídeo: Sub-unidade de DNA ou RNA que consiste numa base nitrogenada, uma molécula de açúcar e ácido fosfórico. Existem 4 tipos diferentes de bases nitrogenadas no DNA (Adenina, Timina, Citosina e Guanina), que constituem as 4 "letras" com as quais é escrito o código genético. Milhares de nucleotídeos ligam-se para formar uma molécula de DNA ou RNA.

OTU (Unidade taxonômica operacional, do inglês, Operational Taxonomic Unit): Ver táxon.

Parcimônia máxima: Método que assume o princípio de que soluções mais simples ou econômicas são preferíveis a soluções mais complexas. Isto significa que árvores filogenéticas que podem explicar os dados observados por meio de um menor número de eventos evolutivos são preferíveis àquelas que requerem maior número de eventos para explicar os mesmos dados.

Pirimidina: Um nucleotídeo Citosina ou Timina.

Purina: Um nucleotídeo Adenina ou Guanina.

Relógio molecular: É um índice utilizado em evolução molecular que relaciona o tempo de divergência entre duas espécies e o número de diferenças moleculares medidas entre as seqüências homólogas de DNA.

Sítio: Uma posição específica do DNA, ocupada por uma base nucleotídica.

Sub-árvore: É também uma árvore, porém foi separada da árvore principal por meio da extinção de um determinado ramo.

Táxon (plural: Taxa): Também denominada de unidade taxonômica, designa um organismo ou grupo de organismos e pode ser posicionado em um nível particular de uma hierarquia que reflete relacionamentos evolutivos. Normalmente, o conjunto de atributos em cada folha de uma árvore filogenética está associado a um táxon.

Topologia em estrela: Topologia de árvore com uma única origem de onde todos os ramos terminais derivam.

Transição: Substituição de uma purina por uma purina ou de uma pirimidina por uma pirimidina.

Transversão: Substituição de uma purina por uma pirimidina ou vice-versa.

Verossimilhança Máxima: método de inferência de relações de árvores filogenéticas usando algum modelo evolutivo. Dada uma árvore (topologia e comprimento de ramos) o método faz a pergunta: “qual a verossimilhança dos dados observados considerando determinado modelo evolutivo?”.

Apêndice B

Matriz de distâncias do estudo de caso dos roedores

Observação: As distâncias que compõem a matriz a seguir não se originaram de comparações genéticas, mas sim morfológicas, a partir de marcos homólogos de crânio (BONATO, 2004).

Colunas 1 a 7

0	0.1517	0.1979	0.1870	0.2174	0.2100	0.2701
0.1517	0	0.0608	0.1378	0.0969	0.1419	0.1953
0.1979	0.0608	0	0.1138	0.0761	0.1070	0.1506
0.1870	0.1378	0.1138	0	0.1178	0.0505	0.0913
0.2174	0.0969	0.0761	0.1178	0	0.1290	0.1485
0.2100	0.1419	0.1070	0.0505	0.1290	0	0.0822
0.2701	0.1953	0.1506	0.0913	0.1485	0.0822	0
0.2107	0.1550	0.1347	0.0769	0.1662	0.0570	0.0801
0.2625	0.1891	0.1732	0.0955	0.1998	0.0939	0.0985
0.2369	0.1139	0.0830	0.1122	0.1058	0.0875	0.1140
0.2800	0.1824	0.1339	0.0898	0.1219	0.0881	0.0677
0.3403	0.2660	0.2289	0.1696	0.1835	0.1797	0.1574
0.2861	0.2036	0.1787	0.1092	0.1681	0.1113	0.1053
0.4931	0.3841	0.3616	0.3965	0.3437	0.4403	0.4570
0.2776	0.1909	0.1791	0.2116	0.1496	0.1988	0.2199
0.2888	0.1986	0.1820	0.1864	0.1846	0.1905	0.2446
0.3217	0.2122	0.1831	0.1967	0.1704	0.1799	0.2144
0.3242	0.2608	0.2741	0.2257	0.2564	0.2298	0.2484
0.3313	0.2492	0.2095	0.2644	0.2285	0.2675	0.2759
0.3073	0.1692	0.1823	0.2024	0.1519	0.1784	0.2021
0.2857	0.2151	0.2221	0.1882	0.1929	0.2178	0.2292
0.3494	0.2331	0.2169	0.2376	0.2240	0.2128	0.2795

Colunas 8 a 14

0.2107	0.2625	0.2369	0.2800	0.3403	0.2861	0.4931
0.1550	0.1891	0.1139	0.1824	0.2660	0.2036	0.3841
0.1347	0.1732	0.0830	0.1339	0.2289	0.1787	0.3616
0.0769	0.0955	0.1122	0.0898	0.1696	0.1092	0.3965

0.1662	0.1998	0.1058	0.1219	0.1835	0.1681	0.3437
0.0570	0.0939	0.0875	0.0881	0.1797	0.1113	0.4403
0.0801	0.0985	0.1140	0.0677	0.1574	0.1053	0.4570
0	0.0529	0.0973	0.1026	0.2125	0.1376	0.4593
0.0529	0	0.1234	0.1276	0.2435	0.1779	0.5114
0.0973	0.1234	0	0.1095	0.2256	0.1580	0.4161
0.1026	0.1276	0.1095	0	0.1123	0.0677	0.4095
0.2125	0.2435	0.2256	0.1123	0	0.1530	0.5007
0.1376	0.1779	0.1580	0.0677	0.1530	0	0.4516
0.4593	0.5114	0.4161	0.4095	0.5007	0.4516	0
0.2279	0.2697	0.1872	0.1925	0.2232	0.2193	0.3579
0.2128	0.2459	0.2236	0.2161	0.2780	0.2270	0.2508
0.1976	0.2463	0.1804	0.1846	0.2671	0.2357	0.3230
0.1951	0.2376	0.2601	0.2470	0.3418	0.2472	0.4176
0.2680	0.3344	0.2252	0.2571	0.3205	0.3186	0.3712
0.1989	0.2371	0.1856	0.1836	0.2532	0.1989	0.3703
0.2488	0.2711	0.2525	0.2249	0.2536	0.2303	0.4846
0.2696	0.3109	0.2429	0.2737	0.2974	0.3185	0.4413

Colunas 15 a 22

0.2776	0.2888	0.3217	0.3242	0.3313	0.3073	0.2857	0.3494
0.1909	0.1986	0.2122	0.2608	0.2492	0.1692	0.2151	0.2331
0.1791	0.1820	0.1831	0.2741	0.2095	0.1823	0.2221	0.2169
0.2116	0.1864	0.1967	0.2257	0.2644	0.2024	0.1882	0.2376
0.1496	0.1846	0.1704	0.2564	0.2285	0.1519	0.1929	0.2240
0.1988	0.1905	0.1799	0.2298	0.2675	0.1784	0.2178	0.2128
0.2199	0.2446	0.2144	0.2484	0.2759	0.2021	0.2292	0.2795
0.2279	0.2128	0.1976	0.1951	0.2680	0.1989	0.2488	0.2696
0.2697	0.2459	0.2463	0.2376	0.3344	0.2371	0.2711	0.3109
0.1872	0.2236	0.1804	0.2601	0.2252	0.1856	0.2525	0.2429
0.1925	0.2161	0.1846	0.2470	0.2571	0.1836	0.2249	0.2737
0.2232	0.2780	0.2671	0.3418	0.3205	0.2532	0.2536	0.2974
0.2193	0.2270	0.2357	0.2472	0.3186	0.1989	0.2303	0.3185
0.3579	0.2508	0.3230	0.4176	0.3712	0.3703	0.4846	0.4413
0	0.1777	0.1584	0.2260	0.1946	0.1368	0.2080	0.2859
0.1777	0	0.1513	0.2344	0.2599	0.1931	0.2377	0.3167
0.1584	0.1513	0	0.1974	0.2103	0.1752	0.1933	0.3188
0.2260	0.2344	0.1974	0	0.2803	0.1612	0.2236	0.3612
0.1946	0.2599	0.2103	0.2803	0	0.2244	0.2553	0.2408
0.1368	0.1931	0.1752	0.1612	0.2244	0	0.1913	0.2748
0.2080	0.2377	0.1933	0.2236	0.2553	0.1913	0	0.2949
0.2859	0.3167	0.3188	0.3612	0.2408	0.2748	0.2949	0

Apêndice C

Seqüências de DNAs para as espécies de Sarcophagidae

13 espécies e 783 sítios

AF259506
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATATTTTATTTTCGGAGCTT
GAGCAGGTATAGTAGGAACTTCATTAAGAATTCTTATTCGAGCAGAACTAGGCCATCCGGGTGCATTAATTGG
AGATGACCAAATTTATAATGTAATTGTTACAGCCCATGCTTTTATTATAATTTTTTTTATAGTAATACCAATT
ATAATTGGAGGATTTGGAAATTGATTAGTACCAATTATACTAGGAGCCCCAGACATAGCTTTCCCTCGAATAA
ATAATATAAGTTTTTGACTTTTACCCCCAGCATTAACATTACTTCTAGTAAGTAGTATAGTAGAAAACGGAGC
TGGAACAGGATGAACTGTTTACCCTCCTTTATCTTCTAACATCGCCCATGGAGGAGCTTCTGTTGATTTAGCC
ATTTTTTCCCTACATTTAGCCGGAATTTCTTCAATTTTAGGAGCAGTAAATTTTATTACTACAGTTATTAATA
TACGATCTACAGGTATTACATTTGATCGAATACCTTTATTTGTTTGATCTGTAGTAATTACAGCTTTACTTTT
ACTTCTTTCCCTACCTGTACTTGCTGGAGCAATTACTATACTATTAACTGATCGAAATATTAATACTTCATTC
TTTGACCCTGCAGGAGGGGGAGATCCAATTCTTTATCAACATTTATTTTGATTCTTTGGACATCCTGAAGTTT
ATATTTTAATTTTACCAGGATTTGGAATAATTTCCCATATTATTAGTCAAGAA

AF259507
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATATTTTATTTTCGGAGCTT
GAGCAGGTATAGTAGGAACTTCATTAAGAATTCTTATTCGAGCAGAACTGGGTCACCCTGGTGCATTAATTGG
AGATGATCAAATTTATAACGTAATTGTTACAGCTCATGCTTTTATTATAATTTTTTTTATAGTAATGCCAATT
ATAATTGGAGGGTTTGGAAATTGATTAGTACCAATTATACTAGGAGCTCCAGACATAGCTTTCCCTCGAATAA
ATAATATAAGTTTTTGACTTTTACCTCCAGCATTAACATTACTTCTAGTAAGTAGTATAGTAGAAAACGGAGC
TGGAACAGGATGAACTGTTTACCCTCCTTTATCATCTAATATTGCTCATGGAGGAGCTTCTGTTGATTTAGCT
ATTTTTTCCCTACACTTAGCTGGAATTTCTTCAATTTTAGGAGCAGTAAATTTTATTACTACAGTTATTAATA
TACGATCTACAGGTATTACATTTGACCGAATACCTTTATTTGTTTGATCTGTAGTAATTACAGCTTTACTTTT
ACTTCTTTCTCTACCTGTACTTGCTGGAGCAATTACTATACTATTAACTGATCGAAATATTAATACTTCATTC
TTTGACCCTGCAGGAGGAGGAGACCCAATTTTATACCAACATTTATTTTGATTCTTTGGGCACCCTGAAGTTT
ATATTTTAATTTTACCAGGATTTGGAATAATTTCCCATATTATTAGTCAAGAA

AF259508
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATATTTTATTTTCGGAGCTT
GAGCAGGTATAGTAGGAACTTCATTAAGAATTCTTATTCGAGCAGAATTAGGTCACCCTGGTGCATTAATTGG
TGATGATCAAATTTATAATGTAATTGTTACAGCCCATGCTTTCATTATAATTTTTTTTATAGTAATACCAATT
ATAATTGGAGGATTTGGAAATTGATTAGTGCCAATTATACTAGGAGCTCCAGATATAGCCTTCCCTCGGATAA
ACAATATAAGTTTTTGACTTTTACCTCCTGCATTAACATTGCTTCTAGTAAGTAGTATAGTAGAAAATGGAGC
TGGAACAGGTTGAACTGTATACCCTCCTTTATCTTCTAATATTGCTCATGGAGGAGCTTCTGTTGATTTAGCT
ATTTTTTCTCTCCATTTAGCTGGAATTTCTTCAATTCTAGGAGCAGTAAATTTTATTACTACAGTTATTAATA
TACGATCAACAGGAATCACTTTGGATCGAATACCTTTATTTGTATGATCTGTAGTAATCACAGCCCTACTTTT
ATTACTTTCTTTACCTGTACTTGCCGGAGCTATTACTATATTATTAACTGATCGAAATATTAATACTTCATTT
TTTGACCCGGCAGGAGGAGGAGATCCTATTCTATATCAACATTTATTTTGATTCTTTGGGCATCCTGAAGTTT
ACATTTTAATTTTACCAGGATTTGGAATAATTTCTCACATTATTAGTCAAGAA

AF259509
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATACTTTATTTTCGGAGCTT
GAGCAGGTATAGTAGGAACTTCATTAAGAATTCTTATTCGAGCAGAATTAGGTCACCCTGGTGCATTAATTGG
AGATGACCAAATTTATAACGTAATTGTTACAGCTCATGCCTTTATTATAATTTTTTTTATAGTAATGCCAATT
ATAATTGGAGGATTTGGAAATTGACTGGTACCAATTATATTAGGAGCCCCAGATATAGCTTTTCCTCGAATAA
ATAATATAAGTTTTTGACTTTTACCTCCAGCATTAACACTACTTCTAGTAAGCAGCATAGTAGAAAATGGAGC
TGGAACAGGATGAACTGTTTACCCTCCTTTATCTTCTAATATTGCCCATGGAGGTGCTTCTGTTGATTTAGCT
ATCTTCTCCCTTCATTTAGCTGGAATTTCATCAATTTTAGGAGCAGTAAATTTTATTACTACAGTTATTAATA
TACGATCTTCTGGTATTACATTTGATCGAATGCCTTTATTTGTATGATCAGTAGTAATTACAGCTTTACTTTT
ATTACTTTCTTTACCCGTTCTTGCCGGAGCAATTACAATATTATTAACTGATCGAAATATTAATACTTCATTT
TTTGATCCTGCAGGAGGAGGAGACCCAATTCTATACCAACATCTATTTTGATTTTTTGGACACCCTGAAGTAT
ACATTTTAATTTTACCTGGATTTGGAATAATTTCTCATATTATTAGTCAAGAA

AF259510
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATATTTTATCTTCGGAGCTT
GAGCAGGAATAGTAGGAACTTCACTAAGAATTCTTATTCGAGCAGAATTAGGTCATCCTGGTGCATTAATTGG
AGATGATCAAATTTATAATGTAATTGTTACAGCTCATGCTTTTATTATAATTTTTTTTATAGTAATACCAATT
ATGATTGGAGGATTTGGAAACTGATTAGTTCCAATTATACTAGGAGCTCCAGATATAGCCTTTCCTCGAATAA
ATAATATAAGTTTTTGACTTTTACCCCCAGCATTAACATTGCTTCTAGTAAGTAGTATAGTAGAAAATGGAGC
TGGAACGGGGTGAACTGTTTACCCTCCTTTATCTTCTAATATTGCTCATGGAGGAGCTTCTGTTGATTTAGCT
ATTTTTTCTCTACATTTAGCTGGAATTTCTTCAATTTTAGGAGCAGTAAATTTTATTACTACAGTAATTAATA
TACGATCTACAGGAATTACCTTTGATCGAATACCTTTATTTGTTTGATCAGTAGTAATTACAGCCCTACTTTT
ACTTTTATCTTTACCGGTACTTGCAGGAGCTATTACTATATTATTAACTGACCGAAATATTAATACCTCTTTT
TTCGACCCAGCAGGAGGAGGAGATCCTATTTTATACCAACACCTATTTTGATTTTTCGGTCACCCTGAAGTTT
ATATTTTAATTTTACCAGGATTCGGAATAATTTCTCACATTATTAGTCAAGAA

AF259511
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATATTTTATCTTCGGAGCTT
GAGCAGGAATAGTAGGAACTTCACTAAGAATTCTTATTCGAGCAGAATTAGGTCATCCTGGTGCATTAATTGG
AGATGACCAAATTTATAATGTAATTGTTACAGCTCATGCTTTTATTATAATTTTTTTTATAGTAATACCAATT
ATAATTGGAGGATTTGGAAACTGACTAGTTCCAATTATATTAGGAGCTCCAGATATAGCTTTTCCTCGAATAA
ATAATATAAGTTTTTGACTTTTACCTCCAGCATTAACATTACTTCTAGTAAGTAGCATAGTAGAAAACGGAGC
TGGAACAGGATGAACTGTTTACCCTCCTTTATCATCTAATATTGCTCATGGAGGAGCTTCTGTTGATTTAGCT
ATTTTTTCTCTTCATTTAGCCGGAATTTCTTCAATTTTAGGAGCAGTAAATTTTATTACTACAGTAATTAATA
TACGATCTACAGGAATTACCTTTGATCGAATACCTTTATTTGTTTGATCAGTAGTAATTACAGCTCTACTTTT
ACTTTTATCTTTACCTGTACTTGCAGGAGCTATTACTATATTATTAACTGACCGAAATATTAACACTTCCTTC
TTTGACCCAGCAGGAGGAGGAGACCCTATTTTATACCAACACTTATTTTGATTTTTTGGTCACCCTGAAGTTT
ATATTTTAATTTTACCAGGATTCGGGATAATTTCTCATATTATTAGTCAAGAA

AF259512
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATATTTTATTTTCGGAGCTT
GAGCAGGAATAGTAGGAACTTCACTAAGAATCCTAATTCGAGCAGAACTAGGTCACCCTGGTGCATTAATTGG
AGATGATCAAATTTATAATGTAATTGTTACAGCTCATGCTTTTATTATAATTTTTTTTATAGTAATACCAATC
ATAATTGGAGGATTTGGAAACTGACTAGTTCCAATTATACTAGGAGCTCCAGATATAGCTTTCCCTCGAATAA
ATAATATAAGATTTTGACTTTTACCTCCTGCATTAACATTACTACTAGTAAGTAGTATAGTAGAAAATGGAGC
TGGAACAGGATGAACTGTTTACCCTCCTTTATCATCTAATATTGCTCATGGAGGAGCTTCTGTTGATCTAGCT
ATTTTTTCTCTTCACTTAGCTGGAATTTCTTCAATTTTAGGAGCAGTAAATTTTATTACTACAGTAATTAATA
TACGATCTACAGGTATTACTTTTGATCGAATACCCCTTTTTGTTTGATCAGTAGTAATTACCGCTTTACTTCT
CCTTCTATCCCTACCCGTACTTGCAGGAGCAATTACTATATTATTAACTGACCGAAATATTAATACTTCATTT
TTTGATCCAGCAGGAGGAGGAGATCCAATTCTATATCAACACTTATTTTGATTTTTTGGTCATCCTGAAGTTT
ATATTTTAATTTTACCAGGATTTGGAATAATTTCTCATATTATTAGTCAAGAA

AF259513
ATTTAATCGCGACAATGGTTATTCTCTACTAATCATAAAGATATTGGGACTTTATATTTTATTTTTGGTGCTT
GATCAGGAATAGTAGGAACTTCTTTAAGAATTCTTATTCGAGCAGAATTAGGACATCCAGGAGCATTAATTGG
AGATGACCAAATTTATAATGTTATTGTTACAGCTCATGCTTTTATTATAATTTTCTTTATAGTAATACCTATT

ATAATTGGAGGATTTGGAAATTGATTGGTTCCAATTATACTTGGTGCTCCAGATATAGCTTTCCCTCGAATAA
ATAATATAAGTTTTTGATTACTTCCTCCAGCTCTTACATTATTACTAGTAAGTAGTATAGTAGAAAACGGAGC
TGGAACTGGATGAACTGTTTACCCACCATTATCTTCTAATATTGCTCATGGAGGAGCCTCTGTTGATCTAGCT
ATCTTCTCTCTACATTTAGCAGGAATTTCATCAATTTTAGGTGCTGTAAATTTTATTACTACAGTTATTAATA
TACGATCAACAGGAATTACTTTCGATCGAATACCTTTATTTGTTTGATCAGTAATAATCACTGCTTTATTACT
TCTTTTATCATTACCAGTTCTTGCTGGAGCTATTACTATATTATTAACTGACCGAAATATTAATACTTCATTT
TTTGACCCAGCAGGAGGAGGAGACCCAATTTTATACCAACATTTATTTTGATTCTTTGGACACCCTGAAGTTT
ATATTTTAATTTTACCAGGATTCGGAATAATCTCTCATATTATTAGTCAAGAA

AF259514
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATACTTCATTTTTGGAGCTT
GATCCGGAATAGTAGGAACTTCGTTAAGAATTCTTATTCGAGCTGAATTAGGACATCCAGGTGCACTTATTGG
TGACGATCAAATTTATAATGTAATCGTTACAGCTCATGCTTTTATTATAATTTTCTTCATAGTAATACCTATT
ATAATTGGAGGATTTGGAAATTGATTAGTTCCAATTATACTTGGAGCACCAGATATAGCTTTCCCTCGAATAA
ATAATATAAGTTTTTGACTTCTTCCTCCAGCTTTAACATTATTACTAGTAAGTAGTATAGTAGAAAATGGAGC
TGGAACAGGTTGAACTGTTTACCCTCCTTTATCTTCTAATATTGCCCATGGAGGAGCATCTGTTGATTTAGCA
ATTTTCTCTCTTCACTTAGCTGGAATTTCATCTATTTTAGGAGCAGTAAATTTTATTACTACAGTAATTAATA
TACGATCTACAGGTATTACTTTTGATCGAATACCTTTATTTGTTTGATCTGTAGTAATTACTGCTTTATTATT
ACTTCTTTCTTTACCTGTACTTGCTGGTGCAATTACTATATTATTAACTGATCGAAATATTAATACTTCATTC
TTTGACCCTGCAGGAGGAGGAGATCCAATTCTATACCAACACTTATTCTGATTCTTTGGACATCCTGAAGTTT
ATATTTTAATTTTACCTGGATTTGGAATAATTTCCCATATTATTAGTCAAGAA

AF259515
ATTTAATCGCAACAATGGTTATTCTCTACTAATCATAAAGATATTGGAACATTATATTTCATTTTTGGAGCTT
GAGCAGGTATAGTAGGAACATCTCTAAGAATTCTTATTCGAGCCGAATTAGGTCATCCAGGAGCTCTAATTGG
AGATGATCAAATTTATAATGTAATTGTCACAGCTCATGCTTTTATTATAATTTTCTTTATAGTAATGCCAATT
ATAATTGGTGGATTTGGAAATTGACTAGTACCAATTATATTAGGAGCCCCAGATATAGCTTTCCCTCGAATAA
ATAATATAAGTTTCTGACTTTTACCTCCAGCATTAACATTACTTTTAGTAAGTAGTATAGTAGAAAATGGAGC
TGGAACAGGATGAACTGTTTATCCACCATTATCTTCTAATATTGCTCATGGAGGGGCTTCTGTTGATTTAGCA
ATTTTTTCTCTTCATTTAGCAGGAATTTCTTCAATTTTAGGAGCAGTAAATTTTATTACAACAGTAATTAATA
TACGATCGACAGGAATTACCTTTGATCGAATACCTTTATTTGTTTGATCTGTAGTTATTACAGCCCTATTATT
ACTTCTTTCTTTACCAGTACTTGCAGGAGCAATTACAATATTATTAACAGATCGAAATATTAATACATCATTC
TTTGATCCAGCTGGAGGAGGAGATCCTATTCTTTATCAACATTTATTCTGATTTTTCGGACACCCTGAAGTTT
ATATTTTAATTTTACCGGGATTTGGAATAATTTCTCATATTATTAGTCAAGAA

AF259516
ATTTAATCGCGACAGTGGTTATTCTCTACTAATCATAAAGATATTGGTACTTTATATTTTCTATTTGGAGCTT
GATCAGGAATAGTAGGAACTTCATTAAGAATTTTAATTCGAGCAGAATTAGGACATCCTGGAGCTTTAATTGG
TAATGATCAAATTTATAACGTAATTGTTACAGCCCATGCTTTTATTATGATTTTTTTCATAGTAATACCTATT
ATAATCGGAGGTTTCGGAAATTGATTAGTTCCTTTAATGTTAGGGGCCCCAGATATAGCATTCCCTCGAATAA
ATAATATAAGTTTTTGATTACTCCCTCCTGCATTAACATTATTATTAGTAAGTAGTATAGTAGAAAACGGAGC
TGGAACTGGTTGAACTGTTTACCCTCCACTTTCAGCTAATATTGCTCATAGAGGAGCTTCTGTGGATTTAGCA
ATCTTCTCTCTTCATTTGGCTGGAATTTCTTCTATTTTAGGGGCTGTAAATTTTATTACAACTGTTATTAATA
TACGATCAACAGGAATTACATTTGATCGAATACCTCTATTTGTTTGATCCGTAGTGATTACTGCTTTATTACT
TCTTCTATCCTTACCTGTATTAGCTGGAGCAATCACTATACTTTTAACAGATCGAAATCTTAATACTTCCTTT
TTTGACCCCGCAGGTGGAGGAGATCCTATTCTTTATCAACATTTATTTTGATTTTTTGGGCACCCAGAAGTTT
ATATTTTAATTCTACCTGGATTTGGTATAATTTCACACATTATTAGTCAAGAA

AF259517
ATTTAATCGCAACAGTGGTTATTCTCTACTAATCATAAAGATATTGGAACTTTATATTTTATCTTCGGAATTT
GATCAGGAATAATTGGAACTTCTTTAAGTATCTTAATTCGAACTGAATTAGGACATCCAGGAGCATTAATTGG
AGATGATCAAATTTATAATGTAATTGTAACAGCTCATGCTTTCATTATAATTTTCTTTATAATTATACCAATT
ATAATTGGAGGATTTGGAAATTGATTAGTACCTTTAATATTAGGAGCTCCAGACATAGCATTTCCTCGAATAA
ATAATATAAGTTTTTGATTACTACCCCCTGCATTAACTTTATTATTAGTAAGTAGTATAGTAGAAAACGGAGC
TGGGACAGGATGAACTGTTTACCCTCCCCTATCTTCTAATATTGCTCATGGAGGAGCCTCTGTAGATTTAGCT
ATTTTCTCTTTACATTTAGCAGGAATTTCCTCTATTTTAGGGGCTGTTAATTTTATTACAACAGTAATTAATA
TACGTGCAACAGGAATTTCATTTGATCGAATACCCCTATTTGTTTGATCAGTAGTAATTACAGCTTTATTATT

ACTTTTATCTCTTCCAGTTTTAGCAGGAGCAATTACAATATTATTAACAGATCGAAATCTTAATACTTCATTT
TTTGATCCTGCAGGAGGAGGGGATCCAATTCTTTACCAACATTTATTTTGATTTTTTGGTCATCCAGAAGTTT
ATATTTTAATTTTACCAGGATTTGGATTAGTTTCTCATGTTATTAGTCAAGAA

AF259518
ATTTAATCGCAACAATGGTTATTTTCTACTAATCATAAAGATATTGGTACTTTATATTTTATCTTCGGAGCAT
GATCTGGTATAGTAGGAACATCATTAAGAATTTTAATTCGAGCTGAATTAGGACACCCTGGTGCTCTAATTGG
AGACGATCAAATTTATAATGTTATTGTAACAGCTCATGCTTTTATTATAATTTTCTTTATAGTAATACCTATT
ATAATTGGAGGGTTTGGAAATTGATTAGTTCCTTTAATATTAGGAGCTCCAGATATAGCATTCCCTCGAATGA
ATAATATAAGTTTTTGATTATTACCTCCTGCATTAACTCTATTATTAGTAAGAAGTACAGTAGAAAAGGGAGC
TGGAACAGGTTGAACTGTTTATCCACCTTTATCATCAATTATTGCTCATGGTGGAGCTTCAGTTGATTTAGCT
ATTTTCTCTCTTCACTTAGCAGGAATTTCTTCAATTTTAGGAGCAGTAAATTTTATTACAACTGTTATTAACA
TACGATCAACAGGAATTACATTCGATCGAATGCCTTTATTTGTTTGATCAGTTGTAATTACTGCATTATTATT
ATTATTATCTCTTCCTGTTCTTGCTGGAGCTATTACTATACTATTAACTGATCGAAATTTAAATACTTCATTC
TTTGACCCAGCTGGAGGAGGTGATCCAATTCTTTACCAACACTTATTCTGATTCTTTGGACATCCAGAAGTTT
ATATTTTAATTTTACCTGGATTTGGAATAATTTCTCATATTATTAGTCAAGAA

Referências

ALURU, S. “Handbook of Computational Molecular Biology”. Taylor and Francis Group. 2006.

ADACHI, J.; HASEGAWA, M. “ MOLPHY: Programs for Molecular Phylogenetics Based on Maximum Likelihood”. <http://www.is.titech.ac.jp/~shimo/class/doc/csm96.pdf> <acesso em abril2007>.1996.

ATTESON, K. “An Analysis of the Performance of the Neighbor-Joining Method of Phylogeny Reconstruction”. DIMACS Workshop on Mathematical Hierarchies and Biology, Rutgers University, N.J. 1996.

ATTESON, K. “The Performance of Neighbor-Joining Methods of Phylogenetic Reconstruction. Algorithmica”. vol 25, pp. 251-278, 1999.

BARTHÉLEMY, J.P.; GUÉNOCHE, A. “Trees and Proximity Representations. Wiley. Chichester. England. 1991.

BILLERA, L.J.; HOLMES, S.P.; VOGTMANN, K. “Geometry of the Space of Phylgenetic Trees”. Advances in Applied Math. vol. 27, no. 4, pp. 733-767, 2001.

BONATO, V. “Padrões de Variação Geográfica em Thrichomys apereoides (Rodentia: Echimyidae)”. Tese de Doutorado, Instituto de Biologia, Unicamp, 2004.

BRODAL, G.S.; FAGERBER G R.; PEDERSEN C.N.S. “Computing the quartet distance between evolutionary trees in time O(n log n)”. Algoritmica 38, pp. 377-395, 2004.

BRUNO, W.J.; SOCCI, N.D.; HALPERN, A.L. “Weighted *Neighbor-Joining*: A Likelihood-Based Approach to Distance-Based Phyloneny Reconstruction. Society of Molecular Biology and Evolution. vol 17, no.1, pp.189-197, 2000.

BRYANT, D. “A Classification of Consensus Methods for Phylogenetics”. Bioconsensus. Ed. Janowitz, M.F.; Lapointe, F-J; Morris, F. R.; Mirkin, B.; Roberts, F. S . Dimacs Series in Discrete Mathematics and Theoretical Computer Science. American Mathematical Society. Providence, Rhode Island. vol. 61, pp. 163-183, 2003.

BULMER, M. “Use of the method of generalized least squares in reconstructing phylogenies from sequence data”. Molecular Biology and Evolution. vol 8, pp. 868–883, 1991.

CARLETON, M.D. “Systematics and Evolution. in Advances in the study of Peromyscus (Rodentia)”. Eds. Kirkland Jr., G.L.; Layne, J.N.. Texas Tech University Press, Lubbock 367. pp. 7-140, 1988.

CAVALLI-SFORZA, L.L.; EDWARDS, A.W.F. “Phylogenetic analysis: models and estimation procedures”. American Journal of Human Genetics. vol 19, pp. 233-257, 1967.

COELHO, G.P.; VON ZUBEN, F.J. “omni-aiNet: An Immune-Inspired Approach for Omni Optimization”. Lecture Notes in Computer Science, Springer, vol. 4163, pp. 294-308, 2006.

COELHO, G.P.; DA SILVA, A.E.A.; VON ZUBEN, F.J. “Evolving Phylogenetic Trees: A Multiobjective Approach”. *in* Advances in Bioinformatics and Computational Biology, Lecture Notes in Computer Science, Springer, vol. 4643, pp 113-125, 2007.

COELLO COELLO, C.A. “A Comprehensive Survey of Evolutionary-Based Multiobjective Optimization Techniques”. Knowledge and Information Systems. vol. 1, no.3, pp. 129-156, 1999.

COELLO COELLO, C.A. “Evolutionary Multi-Objective Optimization: A Historical View of the Field”. IEEE Computational Intelligence Magazine. vol 1, no.1, pp. 28-36, 2006.

COELLO COELLO, C.A.; TOSCANO P. G., "Multiobjective optimization using amicro-genetic algorithm". Ed. Spector, L. Proceedings of the Genetic and Evolutionary Computation Conference. Morgan Kaufmann Publishers. San Francisco, California. pp. 274–282, 2001.

CORNE, D.W.; KNOWLES, J.D.; OATES, M.J. "The pareto envelope-based selection algorithm for multiobjective optimization," Ed. Schoenauer, M. Proceedings of the Parallel Problem Solving from Nature VI Conference. Lecture Notes in Computer Science. vol. 1917, pp. 839–848, 2000. Paris, France.

CORNE, D.W.; JERRAM, N.R.; KNOWLES, J.D.; OATES, M.J. "PESA-II: Region-based Selection in Evolutionary Multiobjective Optimization". Ed. Spector, L. Proceedings of the Genetic and Evolutionary Computation Conference. Morgan Kaufmann Publishers,.San Francisco, California. pp. 283–290, 2001..

DARWIN, C. "On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life". John Murray. Londres. 1859.

DASGUPTA, B.; HE X.; JIANG, T.; LI, M.; TROMP, J.; ZHANG, L. "On distances between phylogenetic trees". Proceedings of the 8th Annual ACM-SIAM Symposium on Discrete Algorithms. pp. 427-436, 1997.

DASGUPTA, B.; HE X.; JIANG, T.; LI, M.; TROMP, J.; WANG, L.; ZHANG, L. "Computing distances between evolutionary trees". Handbook of Combinatorial Optimization D-Z. Ed. Pardalos, P.M. Kluwer Academic Publishers. 1998.

DA SILVA, A.E.A., VILLANUEVA, W.J.P., KNIDEL, H., BONATO, V., DOS REIS, S.F., VON ZUBEN, F.J. "A Multi-Neighbor Joining approach for phylogenetic tree reconstruction and visualization, Genetics and Molecular Research, vol. 4, no. 3, pp. 525-534, 2005.

DEB. K. “Multi-objective Optimization using Evolutionary Algorithms”. John Wiley & Sons. UK. 2001.

DEB, K.; AGRAWAL, S.; PRATAB, A.; MEYARIVAN, T. “A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II”. IEEE Transactions on Evolutionary Computation, vol 6, no. 2, pp. 182-197, 2002.

DE CASTRO, L.N.; TIMMIS, J. “An Introduction to Artificial Immune Systems: A New Computational Intelligence Paradigm”. Springer-Verlag. 2002.

DE FRANÇA, F.O.; VON ZUBEN, F.J.; DE CASTRO, L.N.: “An Artificial Immune Network for Multimodal Function Optimization on Dynamic Environments.”. Proceedings of the Genetic and Evolutionary Computation Conference. Washington, DC, USA. pp. 289–296, 2005.

DELSUC, F.; BRINKMANN, H.; PHILIPPE, H. “Phylogenomics and the Reconstruction of the Tree of Life”. Nature Reviews|Genetics. Nature Publishing Group. vol 6, pp. 361-375, 2005.

FELSENSTEIN, J. “Evolutionary Trees from DNA Sequences: a Maximum Likelihood Approach”. Journal of Molecular Evolution. vol 17, pp. 368-376, 1981.

FELSENSTEIN, J. “Inferring Phylogenies”. Sinauer Associates. Sunderland. Massachusetts. 2004.

FISHER, R.A. “On the Mathematical Foundations of Theoretical Statistics”. Philosophical Transactions of the Royal Society of London. vol. 222, pp. 309-368, 1922.

FITCH, W.M.; MARGOLIASCH, E. “Construction of Phylogenetic Trees”. Science 155, pp. 279-284, 1967.

FOGEL, L.J. “Artificial Intelligence through Simulated Evolution”. Forty Years of Evolutionary Programming. John Wiley & Sons. Inc. New York. 1999.

GASCUEL, O. “BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data”. Molecular Biology and Evolution. Vol. 14, no. 7, pp. 685-695, 1997.

GOLDBERG, D.E. “Genetic Algorithms in Search, Optimization and Machine Learning”. Addison-Wesley Publishing Company. Massachusetts. 1989.

HAIMES, Y.Y.; LASDON, L.S.; WISMER, D.A. “On a bicriterion formulation of the problems of integrated system identification and system optimization”. IEEE Transactions on Systems Man and Cybernetics. vol 1, no. 3, pp. 296-297, 1971.

HEIN, J. “Reconstructing evolution of sequences subject to recombination using parsimony”. Math. Biosci.Mar. vol. 98, no.2, pp. 185-200, 1990.

HEIN, J.; JIANG, T; WANG,L.; ZHANG, K. “On the complexity of comparing evolutionary trees”. Discrete Applied Mathematics. vol 71, pp. 153-169; 1996.

HOLDER, M.T.; LEWIS, P.O. “Phylogeny Estimation: Traditional and Bayesian Approaches”. Nature Reviews Genetics. vol. 4, pp. 275-284, 2003.

HOLMES, S.P. “Phylogenies: An Overview”. Springer Verlag. NY. Statistics and Genetics. Vol. 112, pp. 81-119, 1999.

HOLMES, S.P. “Statistics for Phylogenetic Trees. Theoretical Population Biology”,vol. 63, pp. 17-32, 2002.

HUELSENBECK, P.J. “Performance of Phylogenetic Methods in Simulation”. Systatic Biology. vol 44, pp. 17-48, 1995.

HUELSENBECK, J.P.; CRANDALL, K.A. “Phylogeny Estimation and Hypothesis Testing using Maximum Likelihood”. Annual Review of Ecological Systems. vol 28, pp. 437-466, 1997.

HUSMEIER, D. ”Introduction to Phylogenetics”. Biomathematics and Statistics Scotland at the Scottish Crop Research Institute Invergowrie. UK. <http://www.bioss.ac.uk/»dirk>. <acesso em 03/2006>.

KIDD, K.K.; SGARAMELLA-ZONTA, L.A.”Phylogenetic analysis: Concepts and Methods”. American Journal of Human Genetics. vol 23, pp. 235-252, 1971.

KNOWLES, J.D.; CORNE, D.W. “Approximating the nondominated front using the pareto archived evolution strategy”. Evolutionary Computation. vol. 8, no. 2, pp. 149–172, 2000.

KOSIOL, C.; GOLDMAN, N. “Different versions of the Dayhoff Rate Matrix”. Molecular Biology and Evolution. vol 22, no. 2, pp. 193-199, 2005.

KUHNER, M.K.; FELSENSTEIN, J. “A Simulation Comparison of Phylogeny Algorithms Under Equal and Unequal Evolutionary Rates”. Molecular Biology and Evolution. vol 11, pp. 459-468, 1994.

KUMAR, S. “ A Stepwise Algorithm for Finding Minimum Evolution Trees”. Society for Molecular Biology and Evolution. vol 13(4), pp. 584-593, 1996.

LEVITIN, A. “Introduction to the design and analysis of algorithms”. Addison-Wesley. 2003.

LEWIS, P.O. “Substitution Models and Evolutionary distances”. EEB 372 Lectures Notes. 2002.

LINDER, C.R.; WARNOW, T. “An Overview of Phylogeny Reconstruction”. “Handbook of Computational Molecular Biology”. Ed. Avaru, S. Chapman & Hall/CRC. Cap. 19, pp. 1-34, 2006.

LIÒ, P.; GOLDMAN, N. “Models of Molecular Evolution and Phylogeny”. Genome Research. Cold Spring Harbor Laboratory Press. vol 8, pp. 1233-1244. 1998.

MONTEIRO, L.R. ; BONATO, V; DOS REIS S.F. "Evolutionary integration and morphological diversification in complex morphological structures: mandible shape divergence in spiny rats (Rodentia, Echimyidae)". Evolution & Development. vol 7, no.5, pp. 429–439, 2005.

MYERS, P.; PATTON, J.L.; SMITH, M.F. “Revision of the Boliviensis Group of Akodon (Muridae: Sigmodontinae), with Emphasis on Perú and Bolivia”. Miscellaneous Publications of the Museum of Zoology, University of Michigan. vol 177, pp. 1-105, 1989.

NEI, M.; KUMAR, S. Molecular Evolution and Phylogenetics. Oxford University Press. 2000.

PEARSON, W.R.; ROBINS, G.; ZHANG, T. “Generalized Neighbor-Joining: More Reliable Phylogenetic Tree Reconstruction”. Molecular Biology and Evolution. vol 16, no.1, pp. 806-816, 1999.

PHYLIP. <http://evolution.genetics.washington.edu/phylip/> <acesso em Out/2007>. 2007.

PHYLIP. <http://evolution.genetics.washington.edu/phylip/newick_doc.html> <acesso em Jun/2006>. 2006.

PHYLIP. <http://evolution.gs.washington.edu/phylip/newicktree.html><acesso em Jun/2005>. 2005.

POLADIAN, L.; JERMIIN, L.S. “What might evolutionary algorithms (EA) and multi-objective optimisation (MOO) contribute to phylogenetics and the total evidence debate”. Proceedings of the Genetic and Evolutionary Computing Conference. Seattle. USA. June 2004.

POLADIAN, L.; JERMIIN, L.S. “Multi-objective evolutionary algorithms and phylogenetic inference with multiple data sets”. Soft Computing. vol 4, no. 10, pp. 359–368, 2006.

REEVES, C.R. “Modern Heuristic Techniques for Combinatorial Problems”. Blackwell. 1993.

ROBINSON, D.F.; FOULDS, L.R. “Comparison of Phylogenetic Trees”. Mathematical Biosciences. vol 53, pp. 131-147, 1981.

ROCH, S. “A Short Proof that Phylogenetic Tree Reconstruction by Maximum Likelihood is Hard”. IEEE/ACM Transactions on Computational Biology and Bioinformatics. vol 3, no.1, January-March 2006.

RZHESKY, A.; NEI, M. “Statistical properties of the ordinary least-squares, generalizaed least-squares, and minimum evolution method os phylogenetic inference”. Molecular Biology and Evolution. vol 10, pp. 1073-1095, 1992.

SAITOU, N.; NEI, N. “The *Neighbor-Joining* Method: A New Method for Reconstructing Plylogenetic Trees”. Molecular Biology and Evolution. vol 4, no.4, pp. 406-425, 1987.

SAITOU, N.; IMANISHI, T. Relative Efficiencies of the Fitch-Margoliash, Maximum-Parsimony, Maximum-Likelihood, Minimum-Evolution and Neighbor-Joining Methods of Phylogenetic Tree Construction in Obtaining the Correct Tree. Molecular Biology and Evolution. vol 6, no.5, pp. 514-525, 1989.

SALEMI, M.; VANDAMME, A. “The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny”. Cambridge. University PressScience. 2003.

SASB. “Phylogenetic Tree-Building”. Society of Australian Systematic Biologists. <http://www.sasb.org.au/TreeBuild/TBTable3.html>.<acesso em 02/2006>. 2006.

SOURDIS, J.; NEI, M. “Relative Efficiencies of the Maximum Parsimony and Distance-Matrix Methods in Obtaining the Correct Phylogenetic Tree”. Molecular Biology and Evolution. vol 5, no. 3, pp. 298-311, 1988.

SRINIVAS, N.; DEB, K. “Multiobjective optimization using nondominated sorting in genetic algorithms,” Evolutionary Computation. vol. 2, no. 3, pp. 221–248, 1994.

STEEL, M.; Dress, A.W.; Böcker S. “Simple but fundamental limitations on supertree and consensus tree methods”. Systematic Biology. Publisher: Taylor & Francis. vol 49, no. 2, pp. 363 – 368, April 1, 2000.

TAKAHASHI, K.; NEI, N. “Efficiencies of Fast Algorithms of Phylogenetic Inference Under the Criteria of Maximum Parsimony, Minimum Evolution, and Maximum Likelihood When a Large Number of Sequences Are Used”. Molecular Biology and Evolution. vol 17, no. 8, pp. 1251-1258, 2000.

UNROOTED. “Tree drawing program”.< http://pbil.univ-lyon1.fr/software/unrooted.html>. <Acesso em Fev/2007>. 2007.

VAN WYHE, J. “The writings of Charles Darwin on the Web”. http://pages.britishlibrary.net/charles.darwin/texts/origin1859/origin04.html. <acesso em Jun/2006>. 2006.

WATERMAN, M. S. “On the similarity of dendrograms”. Journal of Theoretical Biology. vol 73, pp. 789-800, 1978.

WELLS, J.D.; PAPE, T.; SPERLING, F.A.H. “DNA-based identification and molecular systematics of forensically important sarcophagidae (diptera)”. Journal of Forensic Sciences. vol 46, no. 5, pp. 1098–1102, 2001.

ZITZLER, E.; THIELE, L. “Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach”. IEEE Transactions on Evolutionary Computation. vol 3, no. 4, pp. 257–271, 1999.

ZITZLER, E.; LAUMANNS, M.; THIELE, L. “SPEA2: Improving the Strength Pareto Evolutionary Algorithm”. Ed. Giannakoglou, K. EUROGEN 2001.

ZUCKERKANDL, E.; PAULING, L. “Evolutionary divergence and convergence in proteins”. Evolving genes and proteins. Academic Press. pp. 97-166, 1965.

Índice Remissivo de Autores

FISHER, R.A., 26.

FITCH, W.M.; MARGOLIASCH, E., 47, 48, 54.

FOGEL, L.J., 107.

GASCUEL, O., 66, 68, 84.

GASIENIEC, L; JANSSON, J.; LINGAS, A.; ÖSTLIN, A..

GOLDBERG, D.E., 107.

HAIMES, Y.Y.; LASDON, L.S.; WISMER, D.A., 105.

HEIN, J., 70.

HEIN, J.; JIANG, T; WANG,L.; ZHANG, K., 69.

HOLDER, M.T.; LEWIS, P.O., 5, 85.

HOLMES, S.P., 3, 85.

HUELSENBECK, P.J., 85.

HUELSENBECK, J.P.; CRANDALL, K.A., 19, 26, 28, 32, 34.

HUSMEIER, D., 44.

KIDD, K.K.; SGARAMELLA-ZONTA, L.A., 50, 96.

KNOWLES, J.D.; CORNE, D.W., 108.

KOSIOL, C.; GOLDMAN, N., 17.

KUHNER, M.K.; FELSENSTEIN, J., 70, 84.

KUMAR, S., 66.

LEVITIN, A., 28.

LEWIS, P.O., 35, 36, 37.

LINDER, C.R.; WARNOW, T., 11.

LIÒ, P.; GOLDMAN, N., 17.

MONTEIRO, L.R. ; BONATO, V; DOS REIS S.F., 91.

MYERS, P.; PATTON, J.L.; SMITH, M.F., 92.

NEI, M.; KUMAR, S., 36.

PEARSON, W.R.; ROBINS, G.; ZHANG, T., 66, 67, 87.

POLADIAN, L.; JERMIIN, L.S., 96.

REEVES, C.R., 28.

ROBINSON, D.F.; FOULDS, L.R., 77.

ROCH, S., 4.

RZHESKY, A.; NEI, M., 51, 52.

SAITOU, N.; NEI, N., 4, 7, 51, 52, 119.

SAITOU, N.; IMANISHI, T., 4, 84.

SALEMI, M.; VANDAMME, A., 85.

SOURDIS, J.; NEI, M., 4.

SRINIVAS, N.; DEB, K., 108.

STEEL, M.; Dress, A.W.; Böcker S., 76, 77.

TAKAHASHI, K.; NEI, N., 4.

VAN WYHE, J., 12.

WATERMAN, M. S., 81.

WELLS, J.D.; PAPE, T.; SPERLING, F.A.H., 129.

ZITZLER, E.; THIELE, L., 108.

ZITZLER, E.; LAUMANNS, M.; THIELE, L., 108.

ZUCKERKANDL, E.; PAULING, L., 17.