Eva Reda Moussa Mansour

Método automático de determinação de clados utilizando algoritmo de detecção de comunidades

(Versão Corrigida)

Dissertação de mestrado apresentada ao Programa de Pós-Graduação Interunidades Bioengenharia - Escola de Engenharia de São Carlos / Faculdade de Medicina de Ribeirão Preto / Instituto de Química de São Carlos da Universidade de São Paulo como parte dos requisitos para a obtenção do título de Mestre em Ciências.

Área de concentração: Bioengenharia

Orientador: Prof. Dr. Alexandre Cláudio Botazzo Delbem

São Carlos-SP
2013

AUTORIZO A REPRODUÇÃO TOTAL OU PARCIAL DESTE TRABALHO, POR QUALQUER MEIO CONVENCIONAL OU ELETRÔNICO, PARA FINS DE ESTUDO E PESQUISA, DESDE QUE CITADA A FONTE.

M289m

Mansour, Eva Reda Moussa
Método automático de determinação de clados utilizando algoritmo de detecção de comunidades / Eva Reda Moussa Mansour; orientador Alexandre Cláudio Botazzo Delbem. São Carlos, 2013.

Dissertação (Mestrado) - Programa de Pós-Graduação Interunidades Bioengenharia e Área de Concentração em Bioengenharia -- Escola de Engenharia de São Carlos; Faculdade de Medicina de Ribeirão Preto; Instituto de Química de São Carlos, da Universidade de São Paulo, 2013.

1. Filogenia. 2. União de FIlogenias em Rede. 3. Métodos de Distância. 4. Método de Otimalidade. 5. Identificação de Clados. I. Título.

Programa de Pós-Graduação Interunidades em Bioengenharia
EESC / FMRP / IQSC

Eva Reda Moussa Mansour
Título: "Método automático de determinação de clados utilizando algoritmo de detecção de comunidades".

DISSERTAÇÃO APRESENTADA AO PROGRAMA DE PÓS - GRADUAÇÃO INTERUNIDADES BIOENGENHARIA - EESC/FMRP/IQSC DA UNIVERSIDADE DE SÃO PAULO PARA OBTENÇÃO DO TÍTULO DE MESTRE EM CIÊNCIAS NA ÁREA DE BIOENGENHARIA.

Aprovado em: 11/10/2013

Prof. Dr. Alexandre Cláudio Botazzo Delbem (Orientador) — Instituto de Ciências Matemáticas e de Computação - USP

RESULTADO: Aprovado — ASSINATURA: Alexandre Cláudio Botazzo Delbem

Profª. Drª. Karla Vittori — Universidade Federal do ABC

RESULTADO: Aprovado — ASSINATURA: Karla Vittori

Prof. Dr. Waldo Gonzalo Cancino Ticona — Instituto Nacional de Pesquisas em Ciência da Computação e Controle

RESULTADO: Aprovado — ASSINATURA: Alexandre Cláudio Botazzo Delbem

Homologado em: __/__/__

Av. Trabalhador São-carlense, 400 - Centro - São Carlos - SP - 13566-590
Telefone/Fax: (16) 3373-9586 - E-mail: bioeng@sc.usp.br

"*Nossos conhecimentos aumentam sempre, através de nossos atos, em qualquer hora e lugar em que nos manisfestamos*"

– Waldo Vieira

Resumo

Mansour, E. M.**Método automático de determinação de clados utilizando algoritmo de detecção de comunidades**. 96pg. Dissertação (Mestrado)- Programa de Bioengenharia/Escola de Engenharia de São Carlos/Faculdade de Medicina de Ribeirão Preto/Instituto de Química, Universidade de São Paulo (USP), 2013.

Análises filogenéticas são bastante utilizadas para a compreensão das relações existentes entre objetos biológicos, beneficiando as investigações em vários campos das ciências da vida. Vários métodos computacionais para reconstruir filogenias tem sido desenvolvidos. Em geral, os métodos que fornecem filogenias mais confiáveis, requerem significativamente maior tempo computacional, restringindo a aplicação deles a conjuntos de dados relativamente pequenos. Por outro lado, a utilização de conjuntos de dados maiores é fundamental para proporcionar uma amostragem que seja suficiente, para restringir as incongruências na identificação de clados em uma filogenia. Este trabalho propõe uma abordagen (denominada *CladeNet*) de reamostragem de filogenias, obtidas por algoritmos relativamente eficientes, a fim de melhorar a identificação de clados. Experimentos com sete conjuntos de dados, que variam de dezenas a centenas de sequências de DNA mostram que, em geral, clados encontrados pela abordagem proposta tornam-se mais confiavéis, conforme os tamanhos dos conjuntos de sequências aumentam, com um moderado aumento do tempo computacional relativamente moderado. Além disso, o *CladeNet* é um método que também inova ao identificar clados de forma automáticamente por meio de um algoritmo de identificação de comunidades em redes.

Palavras–chave: Filogenia, União de Filogenias em Rede, Métodos de Distância, Método de otimalidade, Identificação de Clados, Identificação de Comunidades

Abstract

Mansour, E. M.**Method for automatic determination of clades using community detection algorithm**. 96pg. Master's Thesis. Programa de Bioengenharia/Escola de Engenharia de São Carlos/Faculdade de Medicina de Ribeirão Preto/Instituto de Química, Universidade de São Paulo (USP), 2013.

Phylogenies are useful for understanding relationships among biological objects, benefiting investigations in various fields of life sciences. Several computational methods for reconstructing phylogenies have been developed. In general, methods that provide more reliable phylogenies require significantly larger computing time, constraining their application to relatively small datasets of objects. On the other hand, the use of larger datasets is fundamental to provide enough samples in order to reduce incongruence in clade identification from a phylogeny. This work proposes an approach of resampling phylogenies (called *CladeNet*) obtained from relatively efficient algorithms, in order to improve clade identification. Experiments with seven datasets with sizes varying from dozens to hundreds of DNA sequences show that, in general, clades found by the proposed approach are more reliable as the dataset sizes augment, with relatively moderate increase of computing time. Moreover, *CladeNet* is a new method for identifying clades in an automatic way by means of community detection algorithm for networks.

Keywords: Phylogeny, Union of Phylogenies in Networks, Distance Methods, Optimality Methods, Clade Identification

Agradecimentos

Finalmente chega a hora tão esperada, a escrita dos agradecimentos. Caracterizando assim o fim deste texto e agradecendo a todos que me ajudaram neste caminho dificil. Primeiramente e especialmente agradeço aos meus pais ***Reda Mansour*** *e* ***Fatima Mansour*** *por sempre terem me apoiado, terem confiado em mim e me dado incentivos para estudar e seguir em frente, com certeza sem vocês eu não teria chegado aqui. As minhas irmãs* ***Sara*** *e* ***Noura*** *que me apoiaram, me aguentaram nas horas mais difíceis e pela paciência, sempre. Agradeço ao meu irmão* ***Moussa*** *alguém à quem me espelhei, me ajudou, tentou me convencer a partir para a computação na graduação (mas não conseguiu), e no mestrado aqui estou eu (na computação). Ao meu cunhado* ***Giovani*** *pela companhia de chimarrão e arguile, também pelas conversas aleatórias em momentos que mais precisei. A minha amiga* ***Marcilyanne*** *por sempre me aconselhar e ajudar, principalmente no inicio da minha jornada.*

Neste mestrado me deparei com algo desconhecido, eu era das Biologia de repente fui parar em um projeto na Computação, foi uma aventura e tanto, algo novo, diferente, uma bióloga de ratos de repente virou bióloga de computador, algo meio complexo. Essa jornada foi ótima, desafiadora, aprendi muito e nem me arrependo, pois cresci tanto no profissional quanto no pessoal. Por isso quero agradecer ao voto de confiança dado pelo meu orientador ***Prof.Alexandre Cláudio Botazzo Delbem*** *pela ajuda, paciência, por me guiar neste caminho, meu excelente orientador, agradeço pelos incentivos e pela motivação que tem me transmitido para seguirmos com trabalhos futuros. Também quero agradecer aos meus colegas do* ***(LCR) Márcio, Marcilyanne, Felipe, Helson, André, Daniel, Antônio, Jean, Sérgio, Waldo*** *vocês foram muito importantes nesta jornada, me auxiliaram em tudo especialmente no Linux e no Latex.*

Quero agradecer ao meu amigo ***Rodrigo Ramos*** *e a* ***Paulinha*** *que estiveram presentes nos momentos difíceis, tanto para relaxar bebendo ou saindo. Aos meus amigos* ***Eugênio, Adriano, Evandro, Manuela, Denise, Jennifer, Rodrigo, Renato*** *pelas rodas de tererê, jogos de cartas e pelas motivações sempre. Sem esquecer o grupo* ***BANDINERD*** *e ao grupo* ***Bandidoidu*** *, vocês foram meu grupo de apoio, seria estressante chegar em cada sem ler sobre jogos, animes e outros assuntos. Agradeço ao meu amado* ***Jason Jacinto****, muito obrigada por sempre aturar meu stress, por me dar força e carinho nas horas que mais precisei, por jogar comigo PSO e Diablo III para relaxar. Meu querido agradeço a você pela amizade, companheirismo, confiança e pela inspiração transmitida ao longo deste caminho. Gostaria de agradecer aos meus amigos de São Carlos,* ***Suetake, Giovani, Débora, Monaro, Chris, Everton, Daniel, Fernanda, Paula****. Em especial ao* ***Márcio Crocomo****,* ***Paulo Henrique*** *e* ***Jean Martins*** *que por livre e espontânea pressão leram meus textos e auxiliaram nas correções e em outros detalhes computacionais.*

Agradeço a ***(USP)*** *pela oportunidade de realização do mestrado de alto nível e reconhecimento internacional e a* ***CAPES*** *pelo auxilio financeiro para este projeto. Agradeço à aqueles que não foram citados mas que contribuíram neste trabalho.*

Sumário

Lista de Figuras

1 Introdução **12**

2 Evolução, Genética e Sistemática Filogenética **15**

2.1 Taxonomia 15

2.2 Teoria da evolução 17

2.3 Conceitos de Genética 19

2.4 Sistemática Filogenética 23

2.5 Considerações Finais 26

3 Métodos Computacionais Filogenéticos **28**

3.1 Métricas de Distância Biológicas 28

3.2 Métodos de agrupamento 31

3.2.1 *Neighbor Joining* 32

3.2.2 *UPGMA* 33

3.3 Métodos de Otimalidade 34

3.3.1 Máxima Parcimônia 35

3.3.2 Máxima Verossimilhança 38

3.3.3 Evolução Mínima 39

3.3.4 Inferência Bayesiana de Filogenias 39

3.4 Considerações Finais 40

4 Testes de Confiabilidade **42**

4.1 Análise de *Bootstrap* e *Jackknife* 42

4.2 Técnicas de Consenso 43

4.2.1 Consenso Estrito 43

4.2.2 Consenso estrito utilizando comprimento dos ramos 44

4.2.3 Consenso da regra majoritária 44

4.2.4 Consenso de regras majoritárias considerando comprimento de ramos 45

4.3 Considerações sobre as técnicas de consenso 45

5 Proposta do Método *CladeNet* **47**
5.1 Introdução 47
5.2 Análise de Várias Filogenias 47
5.3 Identificação de clados 48
5.4 Considerações Finais 52

6 Procedimentos Experimentais e Resultados **54**
6.1 Conjuntos de Dados 54
6.2 Procedimentos Experimentais 55
6.2.1 Reamostragem e obtenção de Clados 55
6.2.2 Avaliação de Incongruências entre Clados 56
6.3 Resultados 58
6.3.1 Clados obtidos pelo *CladeNet* usando MrBayes 58
6.3.2 Clados obtidos pelo *Consense-Phylip* 61
6.3.3 Clados obtidos pela *UFR* 65
6.3.4 Ranqueamento de Dominância 68
6.4 Considerações Finais 72

7 Considerações Finais e Trabalhos Futuros **74**
7.1 Considerações Finais 74
7.2 Trabalhos Futuros 76

Referências **78**

A Apêndice **85**
A.1 Métricas de Distância da Teoria de Informação 85

B Apêndice **88**

C Apêndice **92**

D Apêndice **94**

Lista de Figuras

2.1 Diagrama mostrando uma hierarquia de táxons juntamente com a hierarquia de categorias (AMORIM, 2005) 16

2.2 Estrutura das *purinas* e *pirimidinas*, linhas pontilhadas representam ligações de hidrogênio entre nucleotídeos (WATSON; CRICK, 1953). 21

2.3 Diagrama apresentado por Charles Darwin em *Origem das Espécies* (DARWIN, 1859). 23

2.4 Representação de uma árvore filogenética. 26

3.1 Primeira iteração do NJ: (a) Árvore-estrela e (b) Árvore após a inclusão do nó ancestral u. 32

3.2 Matriz distância e árvore filogenética correspondente obtida pelo NJ. . . . 34

3.3 Matriz distância e filogenia correspondente obtida pelo UPGMA. 35

4.1 Exemplo de consenso pelo método Consenso Estrito. 43

4.2 Exemplo de consenso pelo método Consenso Estrito utilizando comprimento de ramos. 44

4.3 Exemplo de consenso pelo método da Regra Majoritária. 45

5.1 *UFR JC+KI+F84*. 49

5.2 Rede com exemplo de cálculo de modularidade do *Fast Algorithm*. 50

5.3 Possíveis valores de modularidade Q para divisões do *Fast Algorithm*, conforme o algoritmo agrupa comunidades. 50

5.4 Diagrama contendo as principais etapas do *CladeNet* usando a *UFR*. . . . 51

5.5 Diagrama contendo as principais etapas do *CladeNet* usando o *Consense-Phylip*. 52

5.6 Diagrama contendo as principais etapas do *CladeNet* usando o *MrBayes*. . 53

6.1 Incongruência Média, usando conjunto de clados obtido pelo *CladeNet* a partir de filogenias do *MrBayes*. 60

6.2 Incongruência Média Normalizada, usando conjunto de clados obtido pelo *CladeNet* a partir de filogenias do *MrBayes*. 61

6.3 Incongruências Médias Normalizadas, usando conjunto de clados obtido pelo *CladeNet* a partir de filogenias de consenso obtidas pelo *Consense-Phylip*. 63

6.4 Incongruências Médias Normalizadas e Incongruências Líquidas, usando conjunto de clados obtido pelo *CladeNet* a partir de filogenias de consenso obtidas pelo *Consense-Phylip*. 64

6.5 Curva de tendência de queda de incongruência usando conjunto de clados obtido pelo *CladeNet* a partir do *Consense-Phylip* para a combinação *KI+F84*. 66

6.6 Curva de tendência de queda de incongruência usando conjunto de clados obtido pelo *CladeNet* a partir do *Consense-Phylip* para todas as combinações de métricas. 67

6.7 Incongruências Médias, usando conjunto de clados obtido pelo *CladeNet* a partir do *UFR*. 68

6.8 Incongruências Médias Normalizadas e Incongruências Líquidas, usando conjunto de clados obtido pelo *CladeNet* a partir do *UFR*. 69

6.9 Reta de tendência de queda de incongruência usando conjunto de clados obtido pelo *CladeNet* a partir do *UFR* para a combinação *KI+F84*. 70

6.10 Curva de tendência de queda de incongruência usando conjunto de clados obtido pelo *CladeNet* a partir da *UFR* para todas as combinações de métricas. 71

D.1 Exemplo de topologia de árvore com comprimentos de ramo (CANCINO, 2008). 95

1
Introdução

A biologia molecular é uma área que tem demonstrado avanços significativos nas últimas décadas. Grandes quantidades de informações extraídas em experimentos laboratoriais têm sido a fonte principal para o desenvolvimento dessa área. A necessidade de interpretar essas informações demandou as pesquisas em bioinformática, que aplica técnicas computacionais, matemáticas e estatísticas para tratar problemas que em geral envolvem dados moleculares.

Um dos principais problemas nesta área é a inferência filogenética. Esse problema consiste em determinar as relações evolutivas de um conjunto de espécies que se encontram representadas por um conjunto de dados moleculares ou morfológicos. Tais relações evolutivas são apresentadas usualmente na forma de uma árvore, conhecida como árvore filogenética. Assim, o objetivo principal da inferência filogenética é determinar a árvore que reflete satisfatoriamente a história evolutiva de um conjunto de espécies (CANCINO, 2008).

Este problema é importante tanto do ponto de vista biológico quanto do ponto de vista computacional. Considerando que não se possui informações suficientes das espécies extintas, deve-se considerar que cada árvore filogenética é apenas uma hipótese possível da relação evolutiva entre as espécies estudadas. Encontrar a melhor árvore filogenética é um problema computacionalmente difícil devido ao número das árvores possíveis crescerem exponencialmente conforme o número de espécies estudadas aumenta.

Diversos métodos de inferência filogenética têm sido propostos na literatura. Alguns métodos definem um critério de otimalidade (FELSENSTEIN, 1981a), que é utilizado para se escolher a melhor filogenia dentro de um conjunto de árvores candidatas. Cada árvore recebe um *score*, que é uma função do relacionamento entre a árvore e os dados. Esta função pode ser, por exemplo, baseada em um modelo de como as sequências evoluem. Por meio do critério, pode-se avaliar a qualidade das árvores e, então, escolher aquela que, dentre várias hipóteses, melhor explica os dados.

Há outros métodos que se baseiam em uma matriz de distâncias entre espécies, que

é previamente calculada segundo uma medida de distância entre pares de espécies. Este método realiza agrupamentos sucessivos das espécies menos distantes em níveis hierárquicos por meio da associação de um nó ancestral, até que não haja mais nós para agrupar, formando assim uma árvore final. Essa abordagem possui algumas vantagens como: facilidade de implementação e eficiência computacional dos programas. Tais métodos possuem a limitação de sempre produzir uma única árvore, não possibilitando a avaliação de outras candidatas para comparações (FELSENSTEIN, 1988).

Várias pesquisas mostram que a escolha do método de inferência influencia significativamente na árvore filogenética obtida. Em outras palavras, a aplicação de diferentes métodos de reconstrução resulta em recuperação de árvores que em geral possuem partes conflitantes entre si. Atualmente a Inferência Bayesiana (RONQUIST; HUELSENBECK, 2003) é considerada o método que produz a filogenia mais verossímil. Porém, esse método é computacionalmente inviável para instâncias com grande número de sequências. Por outro lado, lidar com grande número de sequências é fundamental para se evitar incongruências na determinação de clados (ROKAS et al., 2003), que são os agrupamentos principais, envolvendo espécies fortemente relacionadas com ancestral comum.

Em geral, a obtenção dos clados, é uma etapa posterior à obtenção de uma árvore filogenética ou de um conjunto de filogenias a partir do qual são inferidos os clados. Porém, se o objetivo for obter somente clados e não toda a hipótese filogenética (a árvore completa), o processo computacional pode ser menos intenso. Com isso, poder-se-ia obter clados para grande conjuntos de dados (diminuindo problemas de incongruências) por meio de procedimentos computacionalmente viáveis.

Esta pesquisa de mestrado propõe um método computacional capaz de detectar automaticamente clados a partir de filogenias e também de uma União de Filogenias *(UFR)*, denominado *CladeNet*. Para detectar clados a partir de varias filogenias distintas o *CladeNet* interliga essas filogenias construídas a partir de um mesmo conjunto de dados através dos seus nós folha formando uma *UFR*. Filogenias distintas podem ser obtidas por diferentes métodos de reconstrução (Evolução Mínima, Máxima Parcimônia, entre outras) ou por métricas de distâncias diferentes. O *CladeNet* também pode utilizar uma árvore de consenso obtida de diversas filogenias. Tanto uma árvore de consenso quanto uma *UFR* são interpretadas pelo *CladeNet* com uma Rede Complexa, à qual podem ser aplicados algoritmo de detecção de comunidades (NEWMAN, 2004). Tais algoritmos são capazes de identificar nós da rede com correlação relativamente alta, isto é, que podem corresponder a clados do ponto de vista biológico.

Apesar do *CladeNet* utilizar várias filogenias para determinar clados, ele é computacionalmente eficiente, uma vez que se baseia principalmente em métodos de distância. Além disso, *CladeNet* mostrou-se capaz de obter clados comparáveis ao clados obtidos empregando o *CladeNet* a uma árvore gerada por Inferência Bayesiana. Isso ocorreu ba-

sicamente porque o uso de várias filogenias obtidas por métodos diferentes pode resolver o conflito das informações presentes nessas filogenias, possibilitando determinar clados significativamente confiáveis.

Este texto está organizado em 7 capítulos. O Capítulo 2 apresenta os principais conceitos da Evolução, Genética e a importância da Sistemática Filogenética. O Capítulo 3 apresenta os principais Métodos de Reconstrução Filogenética. O Capítulo 4 apresenta algumas técnicas de reamostragem, importantes para o trabalho realizado. O Capítulo 5 propõe o *CladeNet*. O Capítulo 6 mostra os testes e resultados da aplicação do *CladeNet*. Finalmente, o Capítulo 7 apresenta as principais conclusões desta pesquisa e trabalhos futuros.

2

Evolução, Genética e Sistemática Filogenética

O propósito deste capítulo é apresentar os conceitos básicos necessários para o entendimento desta pesquisa em termos de sua motivação da contribuição que um método de identificação automática de clados, como o *CladeNet*, possa ter para a área. A Seção 2.1 apresenta as nomenclaturas para classificação das espécies que contribuíram para o desenvolvimento da Teoria da Evolução. A Seção 2.3 apresenta conceitos básicos da Genética. A Seção 2.4 sintetiza aspectos de Sistemática Filogenética relevantes neste trabalho. Por fim, a Seção 2.5 apresenta as considerações finais deste capítulo.

2.1 Taxonomia

Um aspecto importante para o entendimento da natureza dos seres vivos é a identificação, denominação e classificação das espécies. O naturalista sueco do século XVIII Carolus Linnaeus iniciou o trabalho de descrição dos tipos de plantas, animais e minerais. Em 1753, Linnaeus publicou o livro intitulado *Species Plantarum* no qual descrevia em latim nomes adequados para cada espécie. Ao utilizar o sistema binomial proposto por Caspar Bauhin (1560-1624) (BAUHIN,), Linnaeus tornou permanente o uso dessa nomenclatura. A finalidade desse método foi a de substituir os nomes desorganizados por nomes binomiais. Assim, os nomes das espécies consistem-se de duas partes: a primeira é o gênero e a segunda funciona como um adjetivo que especifica a primeira. Por exemplo, para a erva-dos-gatos, o nome genérico é *Nepeta*, o específico é *cataria* e o nome da espécie é *Nepeta cataria*.

Hierarquia dos organismos

A nomenclatura de Linnaeus incluía também os reinos que foram denominados: vegetal, animal e mineral. Assim, o reino era a unidade mais inclusiva usada na classificação biológica. Outras categorias taxonômicas hierárquicas também foram adicionadas, como

ordem e classe. O botânico franco-suíço Augustin-Pyramus de Candolle (1778-1841) criou a palavra taxonomia e acrescentou outra categoria para identificar grupos de classes no reino vegetal.

Em 1993, o *International Code of Botanical Nomenclature* (GREUTER; MCNEILL, 1993) definiu que o termo ***filo*** significa uma divisão em grupos taxonômicos. Esse sistema é hierárquico, isto é, cada subgrupo pertence a um grupo, formando níveis hierárquicos. O subgrupo ou grupo taxonômico situados a qualquer nível foram chamados de ***táxon***. Este termo é utilizado para determinar qualquer sistema cujos elementos são populações biológicas. Por exemplo, o conjunto de todas as espécies de felinos (gatos domésticos, gato-do-mato, onça) poderia compor um táxon; as cobras e as serpentes, formam um táxon e assim por diante. Dessa forma, o termo a Taxonomia passou a denominar *ciência da identificação* (RAVEN; EVERT; EICHORN, 2001).

Nos estudos das classes e classes de classes, Linnaeus (1800) utilizou as seguintes classificações, *Species, Genus, Ordo, Classis e Regnum* - Espécie, Gênero, Ordem, Classe e Reino, respectivamente, conforme apresentado na Figura 2.1. Além das categorias taxonômicas criadas por Linnaeus, as classes ou conjunto de ordens semelhantes estão reunidas em ***filos*** e estes em ***reinos***. Por exemplo, o gato-do-mato (espécie *Leopardus tigrinus*) pertence ao reino *Animalia*, filo *Chordata*, classe *Mammalia*, ordem *Carnivora*, família *Felidae* e gênero *Leopardus* e espécie *tigrinus*.

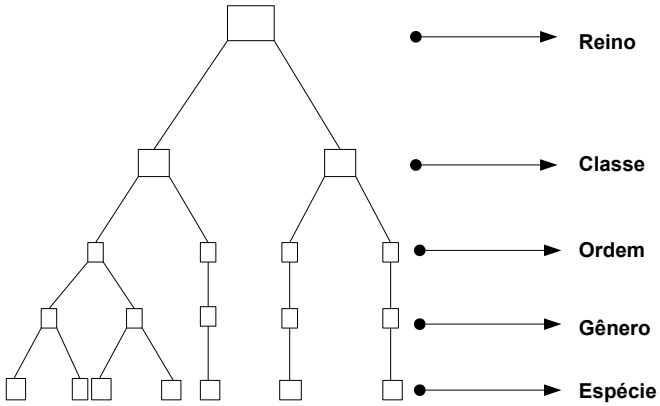


Figura 2.1: Diagrama mostrando uma hierarquia de táxons juntamente com a hierarquia de categorias (AMORIM, 2005)

No início do século XX, os cientistas sentiram a necessidade de separar os seres vivos em novos reinos. Com isso, as pesquisas do biólogo francês Edouard Chatton (1883-1947) (CHATTON, 1975) promoveram a separação das bactérias em um reino exclusivo, chamado *Monera*, visto que elas apresentam células procarióticas, ou seja, sem núcleo nem organelas membranosas, diferindo-as dos demais seres vivos que apresentam células eucarióticas (RAVEN; EVERT; EICHORN, 2001).

Em 1969, o biólogo norte-americano Robert H. Whittaker (1924-1980) (WHITTAKER, 1972) popularizou o sistema de classificação em cinco reinos quando, além dos reinos já citados *Animalia*, *Plantae* e *Monera*, foi criado reino *Fungi*, que agrupa os *fungos*, o *Protista*, que agrupa os protozoários e as algas e retirado o reino *Mineral*. A partir de 1977, Woese e Fox (1977) sugeriu a divisão do reino *Monera* em *Eubacteria* (compostos por organismos unicelulares procariotos, que possuem organelas e podem ser patogênicas) e *Archaebacterias* (compostas por organismos unicelulares, não possuem todas as organelas e não são patogênicas) passando para seis reinos. Em 1990 uma nova classificação foi proposa a qual, por enquanto, a mais difundida e aceita, ela considera três domínios (*Bacteria, Archae e Eukarya*) (RAVEN; EVERT; EICHORN, 2001).

Além da Taxonomia, a sistemática de classificação biológica estuda a biodiversidade do planeta. Uma divisão importante da sistemática é a Paleontologia, que estuda os organismos fósseis (espécies que viveram em um tempo anterior ao recente e que têm partes preservadas). Os fósseis são classificados da mesma forma que as espécies atuais, em classe, ordem, gênero e demais grupos. Uma das dificuldades de lidar com tais dados é a dificuldade em determinar se o fóssil pertence a uma espécie ancestral, extinta ou derivada (MAZZAROLO, 2005).

2.2 Teoria da evolução

A Teoria da Evolução busca explicar as mudanças hereditárias que ocorrem em todos os organismos. Para explicar os mecanismos de evolução das espécies, surgiram várias hipóteses, nas quais se destacaram as teorias de Lamarck (1839), Darwin (1859) e a Teoria Sintética da Evolução, conhecida também como Neodarwinismo (DENNETT; RODRIGUES, 1998).

Teoria de Lamarck

Jean Baptiste de Lamarck (1744-1829) realizou estudos de paleontologia dos invertebrados e formulou uma teoria que dizia que os organismos evoluem e se transformam. Com o tempo, os organismos iriam tornando-se seres mais complexos, até obterem uma condição de vida perfeita e ideal. Lamarck (1839) sustentava que as semelhanças dos organismos ocorriam por modificações evolutivas, ou seja, que as espécies tinham ancestralidade em comum. Os indivíduos teriam necessidades de modificações quando expostos as alterações ambientais. Tais necessidades podiam promover adaptações para novas condições do meio. As espécies que adquirissem os novos hábitos utilizariam com frequência certas partes do corpo, fazendo com que estas partes se desenvolvessem e as partes pouco

utilizadas se atrofiassem. As características adquiridas poderiam ser herdadas por descendentes.

Acredita-se que a força de Lamarck em defender as suas idéias evolucionistas contra os conservadores resultou em sua morte, em 1829. Por outro lado, sua teoria foi considerada viável até 1893, quando o biólogo alemão Weismann (1893) publicou um trabalho que verificou que todos os organismos superiores apresentam dois tipos de células *i*) células germinativas, que passam informações genéticas aos descendentes; *ii*) células somáticas, que compõem o organismo em suas partes que não são associadas à reprodução.

Weismann conseguiu indicar que as células somáticas não transmitiam aos seus descendentes as características adquiridas, através de um experimento com camundongos. O pesquisador cortava os rabos dos camundongos e os cruzavam. Os descendentes sempre apresentavam rabos. A partir desse experimento, Weismann evidenciou que os descendentes não herdavam as características adquiridas por seus progenitores ao longo da vida (WEISMANN, 1893).

Teoria de Darwin

Charles Darwin (1809-1882), naturalista inglês, desenvolveu teoria da Seleção Natural. Ele se interessou pela evolução durante uma viagem realizada ao redor do mundo a bordo do navio inglês *H.M.S Beagle*. Nesta viagem, Darwin coletou diversos exemplares de animais, plantas e fósseis e fez observações sobre as diferenças encontradas entre os indivíduos da mesma espécie. Em 1837, Darwin iniciou os estudos mais detalhadamente com os animais domésticos (galinhas, pombos, gatos, etc). Ele concluiu que as raças de organismos domésticos foram criadas pelo homem, que escolhia os indivíduos para os cruzamentos. Ao passar das gerações, eram reproduzidos indivíduos que possuíam uma determinada característica que fosse de interesse ao homem. Ao mesmo tempo, outros indivíduos eram impedidos de reproduzir. Esse processo foi chamado de seleção artificial.

Após comprovar que as espécies se modificavam no decorrer do tempo, Darwin começou a questionar como as espécies conseguiam se modificar, de forma a se adaptarem ao meio em que vivem. Darwin focou, então, no estudo de fósseis. Ao comparar os fósseis, ele concluiu que os seres vivos modificavam-se com o tempo e que algumas características de animais extintos continuavam em animais atuais.

Em 1838, ele conheceu a Teoria de Malthus (PEREZ, 2009) sobre o crescimento populacional. Malthus dizia que o potencial de reprodução da população humana é maior do que a capacidade da terra de produzir os meios de subsistência. Se uma população não encontra obstáculos ao crescimento, ocorre um aumento no número de indivíduos de acordo com uma progressão geométrica. Assim, se a humanidade fosse submetida às mesmas leis que regem as populações de outras espécies, as populações poderiam crescer

muito rapidamente, mas isso não ocorria na prática. Isso foi o que mais chamou a atenção de Darwin. Para explicar tal fenômeno, Darwin achava que existia uma certa *luta pela vida*. Em uma população com indivíduos diferentes, aqueles com características mais adaptadas ao meio em que vivem teriam mais oportunidades de sobreviver e produzir descendentes. Com o passar das gerações, as diferenças entre os novos indivíduos e os da população original iriam se acentuando até constituírem espécies novas.

Darwin (1859) organizou esses resultados no livro *A origem das espécies*. Os princípios básicos das ideias de Darwin podem ser resumidos da seguinte maneira:

- Os indivíduos de uma mesma espécie em geral apresentam variações em todos os caracteres, de forma que não há indivíduos idênticos;
- Todo organismo tem grande capacidade de reprodução, gerando muitos descendentes. Mas, apenas alguns dos descendentes chegam a idade adulta;
- O número de indivíduos de uma espécie é mantido mais ou menos constante ao longo das gerações;
- Existe a chamada *luta pela vida* entre os descendentes, pois apesar de nascerem muitos indivíduos, poucos atingem a maturidade;
- Na *luta pela vida*, os organismos com variações favoráveis às condições do ambiente onde vivem têm maiores chances de sobreviver;
- Os organismos com as variações vantajosas têm maiores chances de deixar descendentes;
- Dessa forma, ao longo das gerações ocorre naturalmente um processo de seleção por indivíduos que em geral são mais aptos a suportarem mudanças ocorridas no ambiente.

Neodarwinismo

O Neodarwinismo (HO; SAUNDERS, 1984) ou teoria sintética da evolução constitui uma ampliação das ideias de Darwin, ao incluir as causas das variações dos seres vivos. Basicamente, as mutações e as recombinações gênicas são as principais causas da variabilidade genética entre os seres vivos.

2.3 Conceitos de Genética

Mendel (1866) estudou as ervilhas-de-cheiro *(Pisum sativum)* durante oito anos, observando características que eram evidentes em cada planta da espécie: cor da flor, posição

da flor no caule, cor da semente, aspecto externo da semente, forma da vagem, cor da vagem e altura da planta. Ele constatou que dessas plantas (parentais), uma característica (por exemplo, a cor) de uma delas poderia não ocorrer nos descendentes diretos (todos amarelos), mas poderiam aparecer em descendentes de gerações futuras (parte amarelos e parte verdes). Isso mostrava que um indivíduo poderia ter fatores que não se manifestavam (eram recessivos), mas poderiam ser transmitidos a seus descendentes. Verificou-se também que a composição dos fatores de uma planta ocorria na composição das células reprodutivas.

Posteriormente, nos trabalhos de Mendel descobriu-se que esses fatores, que passaram a ser chamados genes, são armazenados nas células em uma sequência de nucleotídeos. O ácido *desoxirribonucleico* (DNA, do inglês *Deoxyribonucleic Acid*) é a estrutura molecular que contém os genes.

DNA e RNA

O DNA (WATSON; CRICK, 1953) é composto por duas fitas de nucleotídeos enrolados, uma ao redor da outra, na forma de uma dupla hélice. Os nucleotídeos são as unidades químicas do DNA. Cada unidade é formada por um açúcar (desoxirribose), um grupo fosfato e uma base nitrogenada (purina ou pirimidina). Os nucleotídeos são diferenciados pelas bases nitrogenadas, sendo elas: *adenina* (A), *guanina* (G), *citosina* (C) e *timina* (T).

Uma das características do DNA é a sua autoduplicação; outra é a de ser base para produção das moléculas de ácido ribonucleico (RNA, do inglês *RiboNucleic Acid*). Quando uma proteína (molécula formada por sequência de aminoácidos (NELSON; COX; CUCHILLO, 2001) cujo enovelamento gera uma estrutura tridimensional que determina sua função na célula) em particular é necessitada por uma célula, a sequência de nucleotídeos é primeiramente copiada (processo de *transcrição*) para RNA. Um outro processo, chamado de *tradução*, usa as informações contidas no RNA para produzir então a proteína no citoplasma da célula (NUSSBAUM, 2008).

O RNA é um ácido nucleico que também pode carregar as informações genéticas, e possui propriedades similares ao DNA. Porém, ao invés de ser uma dupla hélice, o RNA é composto por apenas uma fita de nucleotídeos, com a substituição da desoxirribose pela ribose e da base *timina* (T), que está presente no DNA, pela *uracila* (U).

O DNA e o RNA, de certa forma, possuem semelhanças com as proteínas, por serem constituídos por uma sequência de blocos construtivos. Por outro lado, os *nucleotídeos*, blocos construtivos dos ácidos nucleicos, podem ser considerados mais complexos do que os aminoácidos. As duas purinas, adenina e guanina, são encontradas no DNA e RNA. Quanto às pirimidinas, a citosina está presente no DNA e no RNA, mas, a *timina* ocorre

apenas na molécula de DNA, e a *uracila* aparece apenas no RNA. A Figura 2.2 mostra a ligação dessas bases nas fitas de DNA.

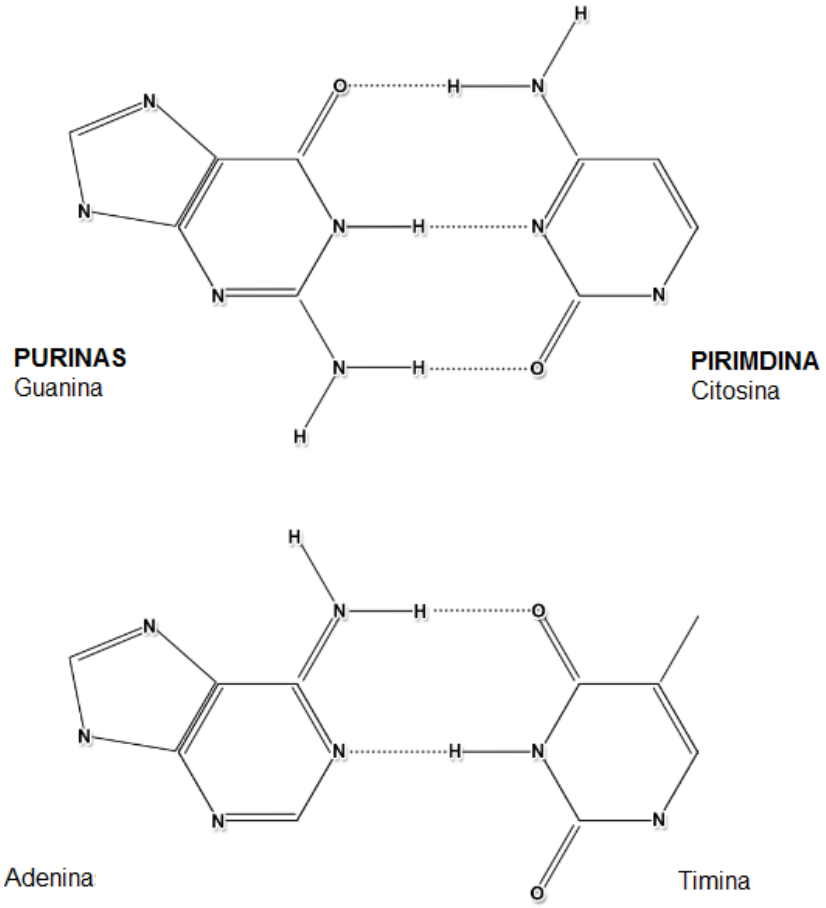


Figura 2.2: Estrutura das *purinas* e *pirimidinas*, linhas pontilhadas representam ligações de hidrogênio entre nucleotídeos (WATSON; CRICK, 1953).

A forma pela qual as cadeias são mantidas juntas na molécula de DNA foi determinada por Watson e Crick (1953) que propuseram um modelo para essa molécula, verificando a configuração mais favorável e compatível com os parâmetros helicoidais fornecidos por dados de raio X. Nesse modelo, a estrutura composta por açúcar e fosfato está do lado externo da molécula de DNA, as bases púricas e pirimídicas estão no interior, formando ligações de hidrogênio com bases da cadeia oposta.

O padrão das ligações de hidrogênio surgiu com base no fato de que se uma *purina* de uma cadeia faz sempre uma ligação de hidrogênio com uma *pirimidina* da outra cadeia, a área da cadeia ocupada pelos pares de bases deveria ser sempre a mesma, por todo o comprimento da molécula de DNA. Isso de fato ocorre, vista de fora, a molécula de DNA tem uma aparência regular. As duas *purinas*, *adenina* e *guanina*, fazem uma ligação seletiva às duas *pirimidinas*, *timina* e *citosina*. A *adenina* (A) pode parear-se somente com a *timina* (T), enquanto a *guanina* (G) pode ligar-se somente a *citosina* (C). Cada um

dos pares possui uma simetria que permite a eles estarem inseridos na dupla hélice nas seguintes formas: A=T e T=A; G=C e C=G (o símbolo = indica ligações de hidrogênio) (NUSSBAUM, 2008).

Os pares de bases são separados por ligações de hidrogênio, que também os diferencia. Três ligações de hidrogênio unem a *guanina* à *citosina*, enquanto duas mantêm os pares *adenina-timina* ligados. Duplas hélices contendo pares de bases GC são mais estáveis (se desnaturam a temperaturas mais altas) que hélices onde predominam pares de bases A-T (GENES, 1968). Amostras de pares de AT e GC em uma molécula de DNA, podem ser determinadas diretamente, medindo-se a temperatura à qual as moléculas de DNA separam-se em fitas simples. Organismos podem ser identificados através da análise da quantidade de pares AT e GC, uma vez que organismos adaptados a ambientes quentes podem ter mais prevalência de conteúdo GC do que AT (WATSON; BERRY, 2005).

Replicação do DNA e Mutações

Para as informações genéticas serem propagadas para as próximas gerações é necessário que ocorra replicação do DNA. O primeiro passo nesta replicação é a abertura da dupla hélice por proteínas conhecidas como helicases. Após a separação das duas fitas de DNA, novas cópias são originadas pelas enzimas chamadas de *polimerases*. Cada fita original age como um molde para a produção de uma nova fita. Finalmente, a fita molde é unida à fita recém-sintetizada por enzimas ligases (NELSON; COX; CUCHILLO, 2001).

Durante a replicação podem ocorrer mutações que são modificações súbitas e hereditárias que ocorrem no material genético. Usa-se o termo mutação de ponto para se referir à alteração de um único par de bases do DNA ou a um pequeno número de pares de bases adjacentes. Os dois principais tipos de mutação de ponto no DNA são as *substituições de bases* e as *inserções* ou *deleções*. Tais substituições de bases são mutações nas quais um par de bases é substituído por outro. As substituições de bases podem ser divididas em dois subtipos: transições ou transversões (GRIFFITHS, 1998). As primeiras correspondem à substituição de uma base por outra base da mesma categoria química. Tanto uma purina é substituída por outra purina (de A para G ou de G para A) quanto uma pirimidina pode ser substituída por outra pirimidina (de C para T ou T para C). A mutação por transversão corresponde à substituição de uma base de categoria química por uma base de outra. Tanto uma pirimidina é substituída por uma purina (de C por A, C por G, T por A ou T por G) quanto uma purina é substituída por uma pirimidina (de A por C, A por T, G por C ou G por T).

A maioria das mutações que ocorrem são deletérias. Por outro lado as mutações são consideradas a base da variabilidade, que permite a seleção de indivíduos mais adaptados a determinado ambiente, possibilitando ao processo de seleção natural (SCHOR; BOIM;

SANTOS, 2003).

2.4 Sistemática Filogenética

Para inferir as relações evolutivas de um determinado grupo de espécies, considerando por exemplo sequências de nucleotídeos, é necessário a escolha de um método filogenético com sequências que apresentem ancestralidade em comum, ou seja, que sejam homólogas.

O conceito principal da filogenia é a ancestralidade que ocorre entre as espécies. Uma filogenia pode ser representada por um diagrama ramificado, em que os seus elementos são espécies e as suas ramificações demonstram a proximidade genealógica existente entre uma espécie e outra, conhecida como ancestralidade exclusiva.

O termo filogenia tem sido utilizado para poder explicar as definições e a dimensão que a biodiversidade adquiriu. A representação dessa diversidade por meio de uma árvore surgiu no próprio livro *Origem das Espécies* (DARWIN, 1859), o qual representa relações evolutivas entre espécies sob a forma de uma árvore (DELSUC, 2005), conforme Figura 2.3.

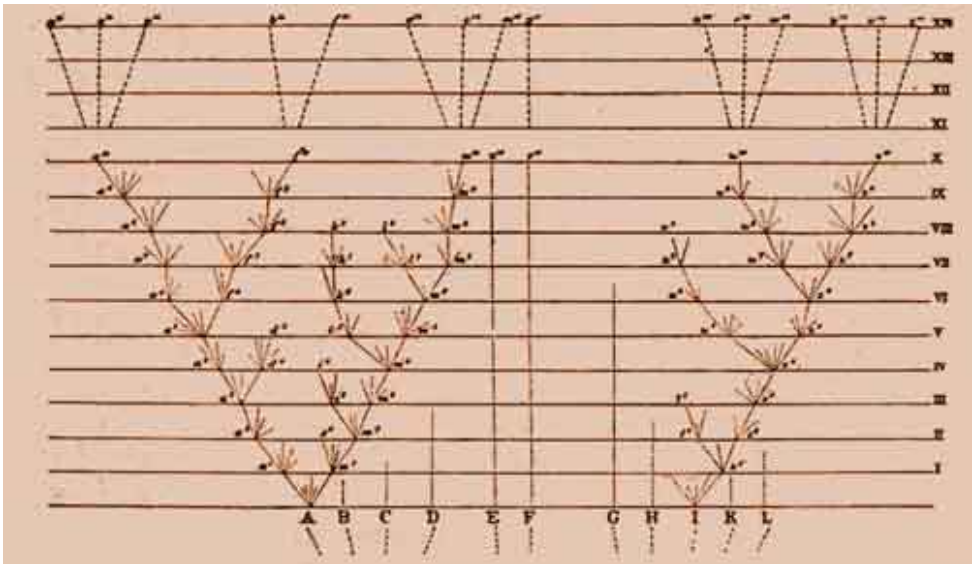


Figura 2.3: Diagrama apresentado por Charles Darwin em *Origem das Espécies* (DARWIN, 1859).

A filogenia ou história evolutiva das espécies está fundamentada na Teoria da Evolução (ver Seção 2.2). Assim, as filogenias nada mais são que a indicação das relações de ancestralidade supostas para um conjunto de espécies. Na prática, o problema fundamental é que as espécies ancestrais que existiram no passado, não podem ser observadas diretamente. Então, é necessário buscar mecanismos para recuperar a informação a res-

peito das relações da similaridade genética de ancestralidade entre os grupos, analisando os organismos atuais (ou fósseis disponíveis) (MATIOLI, 2001).

A ideia da evolução aceita pela comunidade científica é que quaisquer duas espécies devem ter pelo menos uma espécie ancestral comum. Ao se aplicar esse raciocínio a todas as espécies, obtém-se a imagem de uma enorme sequência de divisões que se fragmentaram desde a primeira espécie ancestral a todos os seres vivos até as espécies existentes hoje em dia (DARLINGTON, 1970).

A própria filogenia não segue uma única linha investigativa. Por exemplo, para a Escola Sistemática Tradicional (SANTOS; CALOR, 2008) a ordenação dos organismos é uma atividade apenas catalogatória que separa ou agrupa coisas considerando suas semelhanças ou diferenças. A Escola Fenética também utiliza critérios de similaridade, embora utilize o maior número possível de semelhanças observáveis para agrupar os organismos.

A Escola Evolutiva organiza o conhecimento sobre a diversidade biológica através da história evolutiva dos grupos de organismos. A Escola Cladista (RA et al.,) ordena a diversidade biológica a partir do relacionamento genealógico entre grupos naturais, isto é, com base nas relações de parentesco entre grupos formados por organismos que possuem um mesmo ancestral comum exclusivo.

O relacionamento genealógico é fundamental, uma vez que quando não se comparam característica homólogas, pode-se incidir no erro de considerar similaridades sem origem comum e, portanto, com histórias evolutivas diferentes. Uma das formas de avaliar esta escolha é incluir nas análises, sequências de grupos externos (organismos com história evolutiva conhecida em relação ao grupo estudado), que funcionam como controles no processo de reconstrução de ancestralidade das espécies.

Homologia, Plesiomorfia e Apomorfia

A similaridade resultante da herança de um ancestral comum é chamada de *homologia*. Estruturas *homólogas*, podem ser idênticas entre si, como os braços de dois indivíduos de uma mesma espécie *semelhantes*, como os braços de um homem e de um chimpanzé; ou *diferentes*, como os bicos de um papagaio e de um beija-flor, caso em que certamente o ancestral comum às duas espécies também tinha bico (STEARNS, 2003).

Há três métodos básicos para inferir relações de homologia entre espécies diferentes (AMORIM, 2005): primeiro, quando as estruturas homólogas comparadas têm formas parecidas; segundo, quando ocupam a mesma posição relativa a outras estruturas do corpo e, por último, quando elas se formam a partir de células que ocupavam posição similar em estágios embrionários iniciais e fazem parte da mesma sequência de modificações.

Dentro do paradigma evolutivo, quando se afirma a existência de homologia entre uma estrutura de grupos distintos, está implícita a suposição de que essa estrutura es-

teve presente na espécie ancestral comum mais recente entre os grupos envolvidos. As características primitivas ou ancestrais são chamadas de **plesiomórficas** e quando uma delas é partilhada por todos os membros de um mesmo grupo, denomina-se de **simplesiomorfia**. Já as características derivadas ou avançadas designam-se como **apomorfias** e quando pertencem apenas ao grupo a ser estudado, denominam-se **autapomorfias**.

Grupos Monofiléticos, Parafilético e Polifilético

Como resultado da definição de relacionamento biológico, Hennig (1999) identificou três tipos de agrupamentos taxonômicos: **monofilético, parafilético e polifilético**. Se duas espécies têm um ancestral comum e exclusivo, e isso significa que existiu uma espécie ancestral que se dividiu e gerou essas duas espécies filhas e somente elas. Um táxon monofilético consiste de um agrupamento que inclui uma espécie ancestral e todas as suas espécies descendentes. Trata-se de um grupo natural, ou seja, um produto do processo evolutivo (AMORIM, 2005).

Um grupo taxonômico parafilético é formado por uma espécie ancestral, mas não inclui todos os descendentes. Assim, os descendentes possuem um ancestral em comum, mas este não é exclusivo. Um grupo polifilético consiste de um grupo de espécies que apresentam dois ou mais ancestrais diferentes, ou seja, não apresentam um ancestral comum.

Homoplasias ou Analogias

As modificações ocorridas em uma espécie ancestral podem ser compartilhadas pelos seus descendentes. Organismos diferentes podem compartilhar tanto o estado plesiomórfico de um caráter *(simplesiomorfia)*, quanto o estado apomórfico *(sinapomorfia)*. Estados apomórficos semelhantes podem surgir independentemente em animais distintos sendo denominados *homoplasias (convergência, paralelismo e reversão)* (AVANCINI; AGUIAR; VOGT,).

O esperado acerca da evolução de um caráter é que sua modificação do estado primitivo ocorra uma única vez durante a anagênese (processo que representa uma evolução progressiva de um grupo de espécies que surgem ou se modificam, alterando a frequência genética de uma população (AMORIM, 2005)) de uma espécies. Assim, em dois ou mais momentos distintos da evolução biológica, um determinado caráter pode se modificar em seu estado derivado a partir da mesma ou de diferentes condições primitivas. Esses casos são denominados *homoplasias*. Quando isso ocorre, pode gerar uma interpretação errônea acerca da evolução da espécie, pois uma origem múltipla independente do estado derivado (homoplasia) pode ser entendida como uma origem única compartilhada (sinapomorfia) e, desse modo, táxons não diretamente relacionados evolutivamente podem ser equivocadamente agrupados, supondo-se para eles uma ancestralidade comum exclusiva.

Segundo Stearns (2003), existem três tipos de homoplasias, descritas a seguir:

O **paralelismo**, é o caso em que a partir de um mesmo estado primitivo surgem duas ou mais vezes o mesmo estado derivado. A **convergência** é o caso em que a partir de diferentes estados primitivos surgem duas ou mais vezes o mesmo estado derivado. Por fim, a **reversão** é o caso em que a partir do estado derivado ocorre uma reversão para o estado primitivo.

Árvores Filogenéticas

Uma árvore filogenética é um grafo, isto é, um sistema descrito por elementos representados por pontos e relações/ligações entre os elementos desenhados por arestas. A árvore é um grafo em que cada par de nós possui uma única sequência de nós e arestas conectando o par (um grafo conexo e acíclico). As árvores filogenéticas são árvores de grafo que refletem as relações evolutivas entre um conjunto de espécies. Os nós folha (ligados a apenas uma aresta) representam as espécies que são ligadas a um nó interno, representando o ancestral comum mais recente desses táxons. Em outras palavras, a topologia de uma árvore filogenética pode ser definida como uma representação gráfica unindo nós folha através de arestas a nós internos. A Figura 2.4 ilustra uma árvore filogenética, em que os nós folha (A,B,C,D,E), representam as espécies atuais unidas por arestas aos nós internos (F,G,H,I), que representam os ancestrais comuns mais recentes aos nós folha.

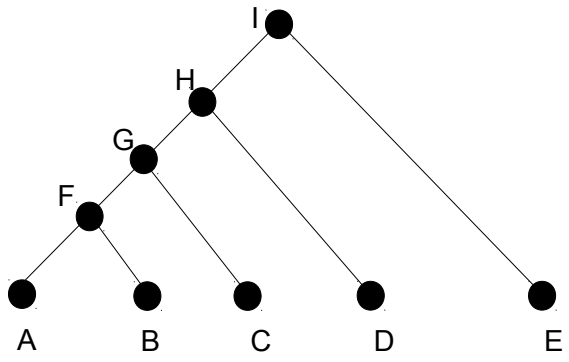


Figura 2.4: Representação de uma árvore filogenética.

2.5 Considerações Finais

Um aspecto importante para o entendimento da biodiversidade dos seres vivos é a identificação, denominação e classificação das espécies. A Taxonomia facilita a organização dessas espécies, utilizando uma estrutura hierárquica (classe, ordem, gênero e demais grupos) para esse fim.

Com as Teorias da Evolução propostas por Darwin (1859), Lamarck (1839) e pelas Leis da Genética propostas por Mendel (1866), foi possível entender como ocorre o processo evolutivo entre a espécie, e como uma espécie pode transmitir suas características hereditárias para outras (de geração para geração). Para inferir as relações evolutivas de um determinado grupo de espécies, considerando sequências de nucleotídeos, é necessário a escolha de um método filogenético e de espécies que apresentem ancestralidade em comum, ou seja, que sejam homólogas.

Quando não se comparam característica homólogas, pode-se incidir no erro de considerar similaridades sem origem comum e, portanto, com histórias evolutivas diferentes. Uma das formas de avaliar esta escolha é incluir nas análises sequências de grupos externos, que são organismos com história evolutiva conhecida em relação ao grupo estudado, funcionando como controles para o processo de reconstrução filogenética.

O Capítulo 3 apresenta os principais métodos computacionais utilizados para reconstrução filogenética, destacando seus princípios e aspectos computacionais relevantes na obtenção de filogenias verossímeis. Observa-se que a qualidade das filogenias afeta diretamente a qualidade dos os clados em que cada espécie é associada.

3

Métodos Computacionais Filogenéticos

Os métodos de reconstrução filogenética podem ser organizados em dois grupos: métodos de agrupamento e métodos de otimalidade. Esta pesquisa investiga um método de determinação de clados, que usa principalmente filogenias obtidas por métodos de agrupamento. Esses métodos baseiam-se em métricas de distância entre espécies. Assim, e capítulo começa apresentando as métricas de disesttância (Seção 3.1) para posteriormente, descrever os métodos de agrupamento (Seção 3.2). Em seguida, os métodos de otimalidade (Seção 3.3) são apresentados, dentre os quais encontra-se a Inferência Bayesiana, que é utilizada para gerar filogenias de referência na validação dos resultados obtidos pelo *CladeNet* (Capítulo 5). Por fim, a Seção 3.4 apresenta considerações sobre os métodos descritos neste capítulo.

3.1 Métricas de Distância Biológicas

Métodos computacionais eficientes para inferir filogenias têm sido desenvolvidos para determinar as relações evolutivas entre diversos grupos de espécies. Conjuntos de dados biológico de espécies, tais como sequências de nucleotídeos, aminoácidos ou dados morfológicos, internos ou externos, podem ser utilizados na construção de filogenias.

Sequências nucleotídicas ou aminoácidos alinhados (observe que técnicas de alinhamento de sequências (KUMAR; TAMURA; NEI, 2004) não são abordadas neste trabalho) podem ser comparados através de modelos que possibilitam salientar semelhanças e diferenças. Em sequências de nucleotídeos as diferenças são geradas por mutações que ocorreram em espécies provenientes de um ancestral comum que foram se diversificando no decorrer do tempo. A diferença que prevalece entre as espécies pode ser expressa na forma de distâncias, que podem ser utilizadas para fim de inferência filogenética. A distância entre duas sequências é calculada utilizando-se modelos evolutivos, que podem determinar a variação de sequências de nucleotídeos, de frequências de genes ou de caracteres quantitativos, dentre outros.

Há também métricas capazes de comparar diferentes tipos de dados, utilizadas em diversas áreas das Ciências e que não utilizam conhecimento a priori sobre modelos evolutivos. Tais métricas são estudadas principalmente na área de Teoria da Informação (RISSANEN, 2007). Algumas das principais métricas dessa área são apresentadas no Apêndice A. Conforme apresentado na Seção 3.2, qualquer métrica capaz de gerar uma matriz de distância pode ser utilizada para se construir uma filogenia por método de agrupamento.

Esta Seção descreve somente métricas que se baseiam em modelos evolutivos. Na sequência são descritos os principais tipos de distância utilizados na literatura para dados biológicos.

Jukes-Cantor

O modelo de Jukes e Cantor (JUKES T.H E CANTOR, 1969) é bastante simples, pois considera que os nucleotídeos aparecem em igual frequência em uma sequência de nucleotídeos e todas as substituições ocorrem na mesma taxa. O modelo de substituição assume que mudanças de nucleotídeos na sequência ocorrem com igual probabilidade, independentemente das bases nitrogenadas trocadas. Esse modelo (como também os de *Kimura* e *Felsenstein*) baseia-se em uma medida de distância relativamente simples, representada por p, que mede a proporção de posições em que duas sequências diferem entre si, assim:

$$p = n_p/n, \tag{3.1}$$

em que n_p é o número de nucleotídeos diferentes em cada posição de duas sequências e n é o número total de nucleotídeos comparados. A sua variância pode ser estimada pela seguinte equação:

$$v(p) = p(1-p)/n. \tag{3.2}$$

Com base em p, a Equação 3.3 descreve o modelo de *Jukes-Cantor*, representado pela distância d.

$$d = -\frac{3}{4}\log_e\left(1 - \frac{4p}{3}\right), \tag{3.3}$$

com variância

$$v(d) = \frac{p(1-p)}{[(1-\frac{4p}{3})^2 n]}, \tag{3.4}$$

em que p é estimado pela Equação 3.1.

Distância Kimura 2 – Parâmetros

Baseado no fato de que substituições do tipo transições ocorrem com mais frequência do que transversões, Kimura (1980) desenvolveu um modelo que considera essas diferenças de frequência. As transições são as substituições de uma base purínica por outra purínica, bem como aquelas de uma base pirimídica por outra base pirimídica. Por outro lado, as transversões são substituições de uma purina por uma pirimidina e vice-versa.

Com substituições de nucleotídeos ocorrendo puramente ao acaso, taxa de substituições seria de 0,50. No entanto, na maioria das vezes quando é observada uma análise de sequências homólogas, essa taxa não é de 0,50. Observando somente o número de possíveis transversões (A<->C, A<->T, G<->C, G<->T) e de possíveis transições (A<->G, C<->T), tem-se 4 tipos de transversões e 2 tipos de transições, isto é, há 4 em 6 possibilidades de ocorrer uma transversão.

No entanto, as *purinas* possuem dois anéis de carbono, enquanto as pirimidinas possuem apenas um anel de carbono. Dessa forma, as substituições de *purina* por *purina* ou *pirimidina* por *pirimidina* são mais simples em termos de alterações na estrutura bioquímica molécula de DNA e, portanto, passam mais facilmente pelo crivo dos mecanismos de reparo da célula (DIECKERSON, 1971). Assim, ocorrem mais transições que transversões.

O modelo de Kimura (1980) considera esse desvio para transições. Por exemplo, se o nucleotídeo original for A, então a probabilidade do A ser substituído por G é maior do que a do A ser substituído por C ou T. A Equação 3.5 sintetiza essa correção, utilizando taxa de transição duas vezes mais provável que a de transversão (mas que pode variar conforme o grupo taxonômico).

$$d = -\log_e(1 - 2q - r) - \log_e(1 - 2r), \tag{3.5}$$

em que q e r são as taxas de transições e transversões entre as sequências comparadas respectivamente. A variância de d é dada por:

$$v(d) = [c_1^2 q + c_3^2 r - (c_1 q + c_3 r)^2]/n, \tag{3.6}$$

em que $c_1 = \frac{1}{(1-2q-r)}$, $c_2 = \frac{1}{(1-2r)}$ e $c_3 = (c_1 + c_2)$.

Modelo Proporcional - F84

Partindo do fato de que a frequência dos quatro nucleotídeos nem sempre são similares, Felsenstein elaborou um modelo para calcular a probabilidade de substituição considerando a frequência de cada uma das bases de nucleotídeos (FELSENSTEIN, 1985). Se

o conteúdo GC (proporção de bases G e C na sequência, é diferente entre as sequências, pode ocorrer distorção na medida de distâncias. Isso porque a probabilidade de ocorrer substituições é afetada pela diferença de conteúdo GC. A métrica proposta por Felsenstein corrige esse efeito considerando conteúdo GC e as taxas de transição e transversão (ver Equação 3.9). Essa métrica pode ser calculada somente quando o conteúdo GC é significativamente diferente de 0,50 e pode ser sintetizada na Equação 3.7.

$$d = -b \log_e(1 - \frac{p}{b}), \tag{3.7}$$

em que b é o conteúdo GC, a variância de d é dada por

$$v(d) = p\frac{(1-p)}{(1-\frac{p}{b})^2 n}, \tag{3.8}$$

e b é definido como segue:

$$b = \frac{1}{2}\left(1 - \sum_{i=1}^{4} g_i^2 + \frac{p^2}{c}\right), c = \sum_{i=1}^{3}\sum_{j=i+1}^{4} \frac{x_{ij}}{2} g_i g_j. \tag{3.9}$$

em que g_i e g_j são frequências dos nucleotídeos i e j, considerando (i,j= A, T, G, C), e x_{ij} é a frequência relativa do par de nucleotídeos ij, na comparação entre duas sequências. Dessa forma, x_{ij} inclui o correspondente às taxas de substituições estimadas a partir das duas sequências.

3.2 Métodos de agrupamento

Métodos filogenéticos que se baseiam em uma matriz de distâncias entre espécies, em geral, são métodos de agrupamento. Isto decorre pelo fato de que eles formam grupos de espécies em níveis hierárquicos por meio da associação de um nó ancestral até que não haja mais nós para agrupar, construindo uma árvore. Essa abordagem possui algumas vantagens como facilidade de implementação e eficiência computacional dos programas (Page, 1998). Tais métodos produzem sempre uma única árvore, não possibilitando a avaliação de outras topologias de árvores com potencial como hipótese filogenética. Na sequência são apresentados os dois principais métodos de agrupamento que utilizam o agrupamento por pares, o *Neighbor Joining* (NJ) (SAITOU; NEI, 1987; STUDIER; KEPPLER, 1988) e o *Unweighted Pair Group Method with Arithmetic* (UPGMA) (SOKAL; MICHENER, 1975).

3.2.1 *Neighbor Joining*

O NJ é um método de agrupamento utilizado para a construção de árvores filogenéticas sem raiz (SAITOU; NEI, 1987). Seus criadores buscavam por meio desse algoritmo uma heurística para a construção (ou uma aproximação) da árvore de Evolução Mínima. É importante observar que a Evolução Mínima é um método de otimalidade, ver Seção 3.3.3, apesar de utilizar distâncias.

O NJ baseia-se em distâncias entre pares de espécies (em geral pares de sequências de nucleotídeos) na forma de uma matriz de distâncias. A partir dessa matriz, busca-se construir uma árvore de forma que, no caminho de um nó folha i até um outro folha j, a soma das distâncias associadas às arestas neste caminho tem o mesmo valor da distância na posição i, j da matriz de distâncias fornecida para construir a árvore.

A construção da árvore pelo NJ ocorre como descrito na sequência. O algoritmo inicia com uma árvore-estrela como mostra na Figura 3.1(a). Procura-se então o par de vizinhos (b e a) que possui a menor distância. Em seguida, é realizada uma junção desses dois vizinhos da árvore-estrela, aos quais é associado um nó adicional u que representa um ancestral em comum entre eles (Figura 3.1(b)). Então, as distâncias de u para todos os demais nós da árvore são calculadas. Esse processo de determinação do par de vizinhos, associação de um ancestral e o cálculo de distâncias é repetido enquanto a árvore não estiver com todos os nós com grau 3 ou 1 (o grau de um nó é o número de arestas conectadas diretamente a este nó).

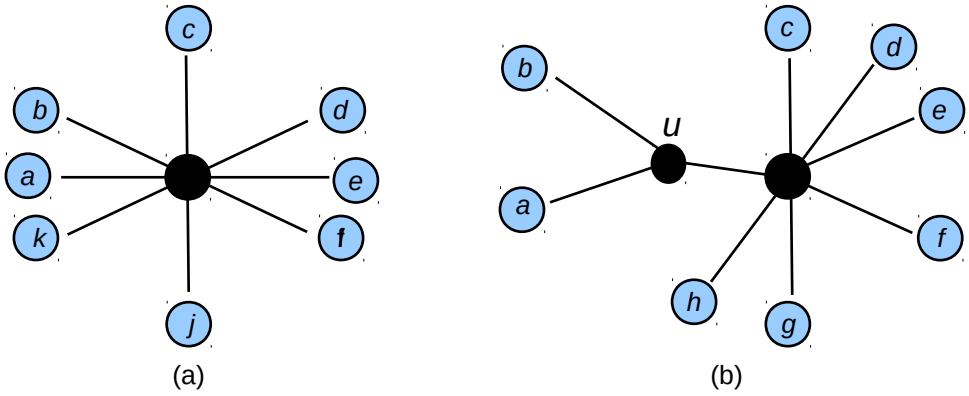


Figura 3.1: Primeira iteração do NJ: (a) Árvore-estrela e (b) Árvore após a inclusão do nó ancestral u.

Na sequência detalham-se os passos do NJ. Primeiramente calcula-se a matriz de distância D usando uma das métricas d descritas na Seção 3.1 (Equações **??**, 3.5 e 3.7). A partir dos valores da matriz distância D, uma matriz M é calculada conforme descreve a Equação 3.10.

$$M_{ij} = D_{ij} - \frac{R_i + R_j}{n-2}, \tag{3.10}$$

em que R_i, dado pela Equação 3.11, é a divergência de rede em relação ao nó i, definida como a soma de todas as distâncias referentes a i:

$$R_i = \sum_{j, i \neq j} D_{ij}. \tag{3.11}$$

A partir de M, escolhe-se o par de nós i e j com menor distância M_{ij}. Então retiram-se os nós i e j de D e inclui-se o nó u em D. Para essa inclusão os comprimentos dos ramos dos dois nós i e j na árvore são recalculados da seguinte forma:

$$S_{iu} = \frac{D_{ij}}{2} + \frac{R_i - R_j}{2(n-2)}, \tag{3.12}$$

$$S_{ju} = \frac{D_{ij}}{2} + \frac{R_j - R_i}{2(n-2)}. \tag{3.13}$$

Em seguida a matriz distância D é recalculada usando a (Equação 3.14), com redução de sua dimensão para $(n-1)$x$(n-1)$, uma vez que o novo nó u agora substitui em D os dois nós agrupados na árvore.

$$D_{ku} = \frac{D_{ik} + D_{jk}}{2} - \frac{D_{ij}}{2}. \tag{3.14}$$

O procedimento descrito até aqui contém instruções para a realização de uma única junção de dois nós (cada atualização da matriz distância D diminuindo sua dimensão em uma unidade). A matriz D modificada pode ser usada como um novo argumento de entrada para o NJ, que determinará um novo par que terá um novo ancestral comum e reduzirá novamente a dimensão da matriz em uma unidade. Esse processo é repetido sucessivamente até restarem apenas dois nós na matriz. Por fim, esses nós são agrupados, dando origem ao último nó a ser adicionado à árvore final.

A Figura 3.2(b) mostra uma filogenia produzida pelo NJ a partir de uma matriz distância apresentada na Figura 3.2(a).

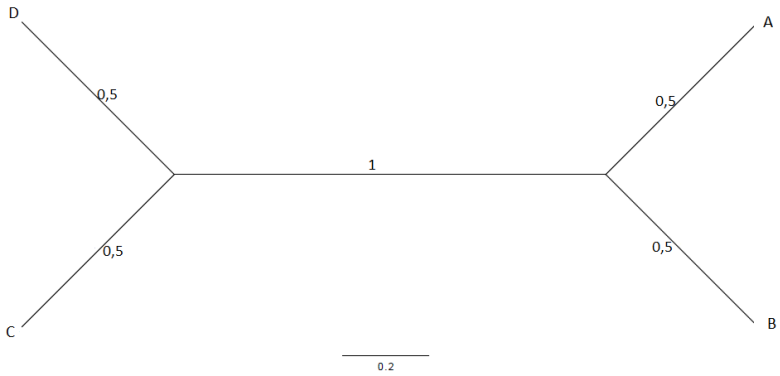
3.2.2 *UPGMA*

O *Unweighted Pair Group Method with Arithmetic Mean* (UPGMA) assume o relógio molecular para a construção da topologia, ou seja, assume que todas as linhagens evoluem a uma taxa constante, uma árvore com essa propriedade é chamada de árvore ultramétrica, em que as distâncias da raiz a qualquer folha são iguais.

O procedimento de cálculo do UPGMA é parecido com o do NJ. Partindo de uma árvore-estrela inicial com todos os nós interligados, executam-se os seguintes passos:

	A	B	C	D
A	0	1	2	2
B	1	0	2	2
C	2	2	0	1
D	2	2	1	0

(a)



(b)

Figura 3.2: Matriz distância e árvore filogenética correspondente obtida pelo NJ.

1. Usando diretamente da matriz distância D, os nós i e j mais próximos são ligados a um nó interno u, ancestral comum;

2. As novas distâncias entre i e j correspondem à metade da distância anterior entre eles;

3. A substituição dos nós i e j por u em D requer o cálculo de distâncias de u para os demais nós da árvore, conforme a Equação 3.15.

$$d_{uk} = \frac{d_{ik} + d_{jk}}{2}, \tag{3.15}$$

em que k é uma coluna de D, tal que $k \neq i$ e j.

A Figura 3.3(b) mostra a filogenia produzida pelo UPGMA a partir da matriz distância na Figura 3.3(a).

3.3 Métodos de Otimalidade

Os métodos de agrupamento produzem uma única árvore, impossibilitando a avaliação de outras topologias de árvores que podem resultar em uma melhor hipótese filogenética. A reconstrução filogenética a partir de critério de ótimo escolhe a melhor filogenia dentro de um conjunto de possíveis árvores candidatas. Este critério é utilizado para atribuir a cada árvore uma nota de qualidade, obtida por uma função que avalia o quão bem uma árvore ajusta-se aos dados. Tal função pode ser, por exemplo, baseada em um modelo de

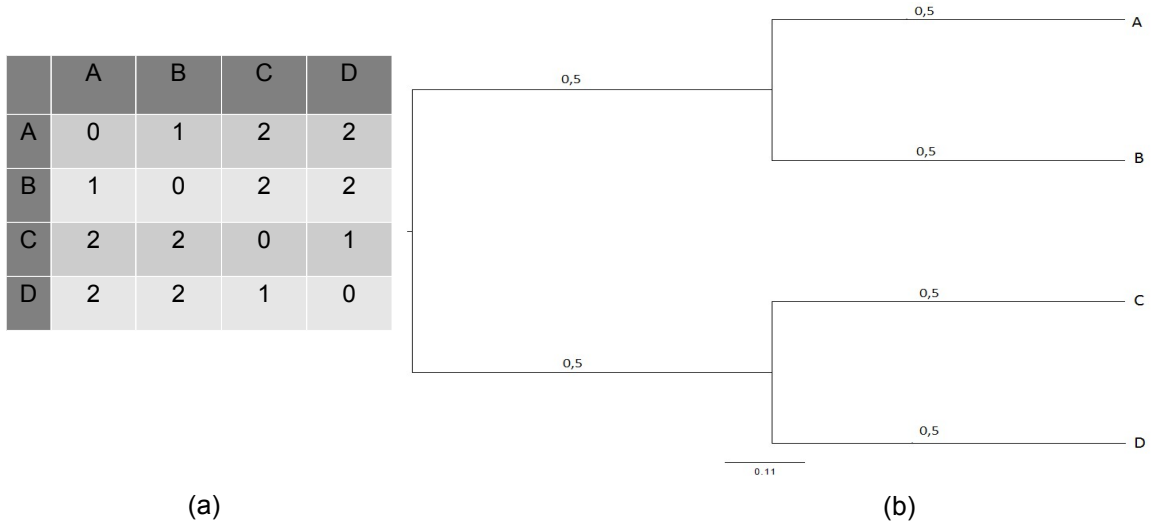


Figura 3.3: Matriz distância e filogenia correspondente obtida pelo UPGMA.

como as sequências evoluem. O critério de otimalidade pode ser aplicado a um conjunto de árvores quaisquer, a árvore com maior nota é a hipótese que melhor explica os dados.

Esses métodos requerem sempre:

1. A geração de uma diversidade de topologias;
2. A avaliação de cada topologia de acordo com o critério de otimalidade escolhido.

Cada método de otimalidade basicamente difere de outro pelas estratégias de gerar as topologias e principalmente pelo critério de avaliação.

Dos métodos de otimalidade na literatura, são apresentados a Máxima Parcimônia (FITCH, 1981) por ter sido um dos primeiros métodos para reconstrução de filogenias, a Evolução Minima que utiliza distâncias, a Máxima Verossimilhança que apresenta resultados significativamente melhores (FELSENSTEIN, 1981a) e a Inferência Bayesiana, que tem sido considerado o método que realiza a melhor inferência. Dentre esses métodos, a Máxima Parcimônia é explicada mais detalhadamente como forma de ilustrar como os métodos de otimalidade funcionam em geral. Para os demais métodos, apresentam-se os aspectos fundamentais que os diferenciam da Máxima Parcimônia. As Seções 3.3.1, 3.3.2, 3.3.3 e 3.3.4 descrevem esses métodos.

3.3.1 Máxima Parcimônia

O método de Máxima Parcimônia baseia-se no princípio de solução parcimoniosa para um problema, proposto pelo filósofo inglês Ockam no século XVII. Conhecido como a "Navalha de Ockam" ou "Lei da Parcimônia", esse princípio dizia o seguinte: se duas

hipóteses explicam os dados com igual eficiência, deve-se adotar a mais simples. Na análise dos dados moleculares, são considerados apenas os sítios informativos (WEIR, 1996), porque nem toda variabilidade em um sítio das sequências resulta em avaliação diferente para alguma topologia de árvore não contribuindo para distinguir uma árvore que melhor explique os dados. A reconstrução de árvores filogenéticas baseadas na máxima parcimônia explica o processo evolutivo que infere o menor número de mudanças ocorridas. Após serem selecionados os sítios informativo, é calculado o número mínimo de alterações (*score*) necessárias para explicar os dados inferidos em cada árvore. A árvore que oferecer melhor *score* é escolhida como a árvore parcimoniosa.

Pequeno Problema de Parcimônia

O pequeno problema de parcimônia consiste em avaliar o quão adequada é uma árvore, determinando o custo da mesma (FELSENSTEIN, 1988). Existem algoritmos polinomiais (LEISERSON et al., 2001) capazes de resolver este problema. Entre eles, se destacam o algoritmo de Fitch, que utiliza como entrada a topologia de uma árvore filogenética enraizada.

Algoritmo de Fitch

O algoritmo de Fitch calcula a parcimônia analisando sítio por sítio de uma sequência de dados para inferir o número mínimo de mudanças necessárias para explicar o processo evolutivo seguindo a estrutura de uma certa árvore. Para isso, o algoritmo constrói um conjunto de possíveis nucleotídeos para cada nó interno de uma árvore, que é computado a partir dos estados de seus filhos. O cálculo é iniciado pelas folhas da árvore filogenética, para cada folha é atribuído o nucleotídeo do sítio em análise. Então, a árvore é percorrida em pós-ordem (todos os filhos do nó atual são visitados primeiro), aplicando-se o operador de parcimônia. Este operador funciona como a intersecção dos conjuntos de características se esta intersecção não for vazia; caso contrário esse operador funciona, como a união dos dois conjuntos.

Grande Problema de Parcimônia

O grande problema consiste em encontrar qual de todas as topologias de árvores possíveis é a mais parcimoniosa. Este é um problema de otimização NP-completo (LEISERSON et al., 2001), no qual técnicas exatas, como busca exaustiva, são computacionalmente inviáveis, mesmo para conjuntos de dados de tamanho relativamente moderado. Dessa forma, técnicas de busca heurísticas têm sido amplamente utilizados. As técnicas de busca heurística, começam com uma árvore inicial não ótima, sobre a qual são aplicadas várias

formas de troca de ramos (arestas da árvore) de forma iterativa buscando melhorar a solução.

A construção da árvore inicial pode empregar os seguintes métodos (SWOFFORD et al., 1996):

1. Adição por passos (*stepwise addition*): inicia com uma árvore de 3 espécies. As demais espécies são adicionadas iterativamente. A posição em que a nova folha será inserida é escolhida analisando todos os ramos onde a nova folha pode ser inserida. Assim, escolhe-se a melhor posição de inserção conforme algum critério de otimalidade. Esse processo é repetido até que todas as espécies estejam adicionadas na árvore;

2. Decomposição de estrela (*star decomposition*): essa técnica inicia com uma árvore-estrela, em que todas as espécies estão unidas a um nó interno. Posteriormente, duas espécies são agrupadas e separadas da estrela mediante a criação de um novo nó interno (assemelha-se ao processo do NJ para construção de topologia). A seleção de tais espécies pode ser realizada aleatoriamente ou analisando todas as alternativas possíveis. Tal processo é repetido iterativamente até que seja formada uma árvore.

Os métodos de adição por passos e decomposição por estrela raramente resultam na árvore ótima. Outras modificações de árvore podem ser aplicadas para melhorar as soluções fornecidas por tais métodos. Swofford et al. (1996) descrevem três formas usuais de modificação topológica:

1. Troca dos vizinhos mais próximos (NNI, *Nearest Neighboor Interchange*): essa técnica trabalha com os ramos interiores da árvore. Cada ramo define 4 subárvores vizinhas conectadas aos seus extremos. Cada um dos extremos possui um par de subárvores. O NNI troca subárvores vizinhos de pares diferentes modificando a árvore inicial. Todas as operações NNI são efetuadas nos ramos da árvore original;

2. Poda e inserção de subárvore (SPR, *Subtree Prunning and Regrafting*): esta operação separa uma subárvore da solução inicial e, posteriormente, é reinserida em todas as posições possíveis. Tal processo é repetido para todas as subárvores da solução inicial. O SPR realiza uma busca mais abrangente que o NNI, possibilitando avaliar um maior número de árvores;

3. Bisseção e reconexão de árvore (TBR, *Tree Bisection and Reconnection*): essa técnica elimina um ramo interno da árvore original, separando-a em duas subárvores. Em seguida, tais subárvores são reconectadas, criando um novo ramo que conecta as duas subárvores. Todas as subárvores e todas as reconexões possíveis são examinadas. O TBR permite explorar um maior número de soluções que o SPR.

Caso seja encontrada uma melhor solução, esta é usada como novo ponto de partida para as próximas modificações. Este processo é repetido iterativamente até que não seja possível encontrar melhores soluções.

A estratégia de obter uma árvore por busca heurística e, em seguida, aplicar NNI, SPR ou TBR é implementada na maioria dos programas de inferência filogenética, tais como PHYLIP (FELSENSTEIN, 2000), PAUP (SWOFFORD, 2000) e PAML (YANG, 1997) indicados para problemas de até 100 amostras. Para problemas maiores os programas NONA (GOLOBOFF, 1999) e TNT (GOLOBOFF; FARRIS; NIXON, 2004) que utilizam novas abordagens de busca topológicas, podem ser mais adequados. Uma descrição mais detalhada dessas técnicas pode ser vista em (CANCINO, 2008).

3.3.2 Máxima Verossimilhança

A máxima verossimilhança é um dos estimadores mais empregados na inferência estatística (GASCUEL, 2007). O conceito de verossimilhança lida com situações em que hipóteses ou modelos referentes a um conjunto de dados são avaliados, com base em um determinado modelo de evolução, para representar um processo evolutivo. Assim, para cada topologia é computada a probabilidade desta ocorrer segundo um modelo evolutivo, selecionando o comprimento dos ramos de maneira a maximizar a probabilidade dos dados analisados em função da árvore em questão.

Como outros métodos probabilísticos, a Máxima Verossimilhança compara os valores esperados e os estimados a partir de certos parâmetros. O método verifica o quanto o conjunto formado por dados (sequências) mais hipótese (árvore estimada) concorda com o que seria esperado (modelo evolutivo).

A verossimilhança fornece a probabilidade $Prob(D|\theta)$ de um conjunto de sequências D ajustar-se ao modelo $\theta = \{T, B, M\}$, em que T é uma topologia da árvore, B é o conjunto de comprimento de ramos de T e M é o modelo de substituição de sequências. O objetivo do critério de máxima verossimilhança é encontrar os parâmetros do modelo θ, tal que a função de verossimilhança definida como $L = Prob(D|\theta)$ seja maximizada. Para evitar números grandes difíceis de representar e computar, usa-se $log_e L$ ($\ln L$). O algoritmo para o cálculo de $\ln L$ para cada árvore é sintetizado no Apêndice D.

Os mesmos programas de busca topológica para Máxima Parcimônia podem ser usados para obter as topologias no método de Máxima Verossimilhança. Para verossimilhança, existem programas como o PHYML (GUINDON; GASCUEL, 2003a) e RAxML (STAMATAKIS, 2005) que produzem resultados de forma relativamente eficiente.

3.3.3 Evolução Mínima

O método de Evolução Mínima estima, para cada possível árvore alternativa, o comprimento de cada ramo da árvore a partir das distâncias evolutivas entre os táxons, calculando o somatório de todos os comprimentos de ramos. A árvore escolhida pelo critério de evolução mínima é a que apresenta menor valor do somatório. O procedimento para cálculo dos comprimentos dos ramos usados para se obter o somatório de uma topologia é explicado na Equação 3.16.

$$\hat{B} = (A^t A)^{-1} A^t d, \tag{3.16}$$

em que $\hat{B}$ é um vetor coluna com as estimativas calculadas dos comprimentos de ramos de uma árvore, A é a matriz topológica que descreve a estrutura da árvore, d é o vetor de estimativas de distâncias evolutivas entre as sequências, t e -1 representam as operações de matriz de transposição e inversão respectivamente. Os comprimentos de ramo pode ser estimados de forma computacionalmente mais eficiente sem utilizar álgebra de matrizes conforme mostrado em (RZHETSKY; NEI, 1993).

No entanto, o método de evolução mínima descrito originalmente é relativamente lento (para número grande de sequências), uma vez que todas as possíveis árvores tinham que ser construídas (MATIOLI, 2001). Esse aspecto motivou o desenvolvimento do NJ (Seção 3.2.1). Observa-se que a Evolução Mínima poderia ser mais eficiente se heurísticas para geração de topologias como as apresentadas na Seção 3.3.1 fossem utilizadas. Por outro lado, a qualidade da inferência filogenética à princípio tenderia a ser similar à filogenia obtida pelo NJ, que também usa distâncias. De fato, não se encontrou na literatura estudo que confirmasse tais evidências.

3.3.4 Inferência Bayesiana de Filogenias

A Inferência Bayesiana é um método estatístico mais recentemente utilizado para inferência filogenética. O principal aspecto da Inferência Bayesiana é o cálculo da probabilidade condicional *a posteriori* das diferentes árvores e parâmetros do modelo evolutivo em função de um conjunto de dados. Sejam D um conjunto de dados, τ_i a i-ésima topologia do conjunto de N_T topologias possíveis para n espécies, e φ= (B,M) um modelo em que B é o conjunto dos comprimentos de ramos de τ_i e M corresponde aos parâmetros do modelo de substituição de sequências. A probabilidade posterior de τ_i dado D é expressada pela Equação 3.17, que requer o somatório, para todas as topologias existentes, da integral sobre todos os valores possíveis dos parâmetros em φ.

$$P\left(\tau_i|D\right) = \frac{P\left(D|\tau_i,\varphi\right)P\left(\tau_i,\varphi\right)}{\sum\limits_{j=0}^{N_T}\int P\left(D|\tau_j,\varphi\right)P\left(\tau_j,\varphi\right)d\varphi}, \tag{3.17}$$

em que $P\left(D|\tau_i,\varphi\right)$ é a verossimilhança da árvore τ_i e $P\left(\tau_i,\varphi\right)$ é o produto das probabilidades *a priori* da topologia τ_i e dos parâmetros em φ.

Os mesmos modelos de substituição de nucleotídeos usados na máxima verossimilhança podem ser usados na análise Bayesiana. O programa MrBayes usa as cadeias Markovianas para gerar uma variedade de topologias e estimar uma distribuição das probabilidades das árvores filogenéticas (METROPOLIS et al., 1953), possibilitando o cálculo da probabilidade *a posteriori*.

O algoritmo de Monte Carlo em Cadeias Markovianas (MCMC) trabalha do seguinte modo: *i*) um novo estado para a cadeia é proposto usando um modelo estocástico; *ii*) a probabilidade de aceitação para este novo estado é calculada; *iii*) uma variável aleatória cujo valor varia de 0 a 1 é criada. Se o número for menor do que a probabilidade de aceitação, o novo estado é aceito e o estado da cadeia é atualizado. Caso contrário, a cadeia permanece no estado antigo. Esse processo de propor e aceitar novos estados é repetido milhares ou milhões de vezes e a proporção com que uma árvore é visitada durante o curso da cadeia é uma aproximação válida das suas probabilidades *a posteriori*. MrBayes utiliza uma variante do MCMC (KERAMIDAS; KAUFMAN, 1992).

O MCMC executa *n* cadeias, *n-1* das quais são aquecidas (cadeia que tem a probabilidade *a posteriori* aumentada por um valor. As inferências são baseadas apenas nos estados amostrados pelas cadeias frias. Após todas as cadeias completarem um passo, duas cadeias escolhidas ao acaso tentam uma troca de estado frio para quente. Se a troca ocorre, então a cadeia cujo estado antigo era frio pode mais facilmente explorar o espaço das árvores filogenéticas.

3.4 Considerações Finais

Neste capítulo foram apresentados os principais métodos da área de inferência filogenética. A reconstrução de uma filogenia que melhor explique as relações evolutivas entre as espécies analisadas é um problema complexo, devido ao grande espaço de busca e a falta de informação sobre as especies ancestrais.

Os métodos de otimalidade tem sido os métodos de reconstrução mais utilizados pelos pesquisadores, possibilitando melhores inferências. Por outro lado, estes métodos exigem um esforço computacional significativo. Enquanto métodos de distância tendem a gerar resultados satisfatórios com eficiência computacional significativamente menor.

O Capítulo 5 propõe um novo método, o *CladeNet*, que busca encontrar clados para conjuntos de dados utilizando tanto um conjunto de filogenias obtidas por método de distância, quanto uma filogenia obtida por método de otimalidade (no caso, a Inferência Bayesiana). O *CladeNet* consegue obter clados significativamente similares tanto usando filogenias de métodos de distância quanto de otimalidade.

Naturalmente, a obtenção dos clados usando filogenia de métodos de distância resulta em um método computacionalmente mais eficiente. Dessa forma, o *CladeNet* consegue um compromisso interessante entre qualidade da inferência e tempo de computação, pelo menos em relação à obtenção de clados.

4

Testes de Confiabilidade

Uma vez obtidos os clados filogenéticos empregando um método de reconstrução filogenética baseada em distâncias, é importante utilizar alguma medida de confiabilidade no resultado. Neste capítulo são descritas brevemente técnicas para avaliação dos clados obtidos, estes mesmos testes estatísticos podem ser utilizados para avaliar árvores filogenéticas.

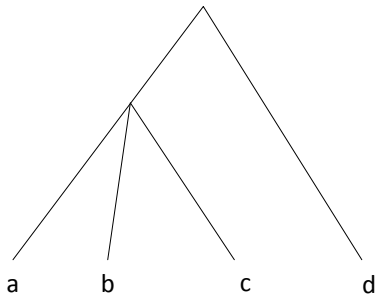
4.1 Análise de *Bootstrap* e *Jackknife*

O *bootstrap* é uma técnica de reamostragem estatística empregada frequentemente na avaliação de clados de uma árvore inferida. Felsenstein foi o primeiro a sugerir tal técnica para o método de máxima verossimilhança (FELSENSTEIN, 1988). Entretanto, o *bootstrap* pode ser empregado com outros métodos de reconstrução filogenética como a máxima parcimônia e os métodos de agrupamento (YANG, 2006).

O método de *bootstrap* consiste em gerar um determinado número de banco de dados a partir de uma amostragem dos sítios dos dados originais. As réplicas de sequências geradas pelo *bootstrap* possuem o mesmo número de sítios que as sequências originais. Na construção de uma réplica, cada sítio é selecionado de forma aleatória a partir dos dados originais, sem ser eliminado do conjunto original de dados (amostragem, com reposição). Com isso, alguns sítios dos dados originais podem aparecer duplicados na réplica e outros podem não estar presentes na réplica. O *Bootstrap* também pode ser realizado com um número menor do que o número total de sítios presentes nos dados originais (ZHARKIKH; LI, 1995).

Após isso, cada réplica gerada é empregada como entrada para o método de reconstrução filogenética considerado e cada árvore é inferida por cada réplica. Finalmente, é calculada a proporção de cada clado da árvore inicial presente nas árvores das réplicas. Este valor mede a probabilidade de um clado ser recuperado em um conjunto de réplicas.

(a) Duas topologias distintas de árvore com raiz e quatro folhas.



(b)

Figura 4.1: Exemplo de consenso pelo método Consenso Estrito.

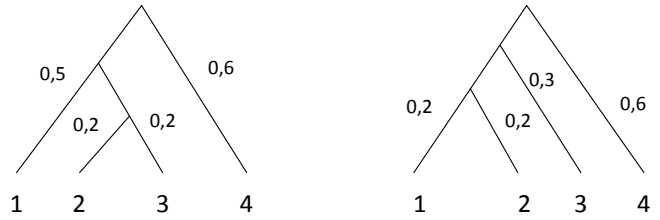
Recomenda-se utilizar no mínimo 1000 réplicas (FELSENSTEIN, 1985). A partir destas réplicas, árvores são construídas, e no final, uma árvore de consenso de todas as réplicas é feita e os valores percentuais são anotados para cada nó interno da árvore. O porcentual obtido para cada nó representa a proporção de vezes que o nó permaneceu intacto nas 1000 réplicas.

O *bootstrap* e o *jackknife* diferem apenas no modo com que as substituições são feitas. No *jackknife*, o conjunto de dados é reduzido de uma só vez para uma certa quantidade sem reposição (em geral 50% do seu tamanho original) e não ocorrem sítios duplicados. Então uma árvore filogenética é inferida a partir desse conjunto menor de dados.

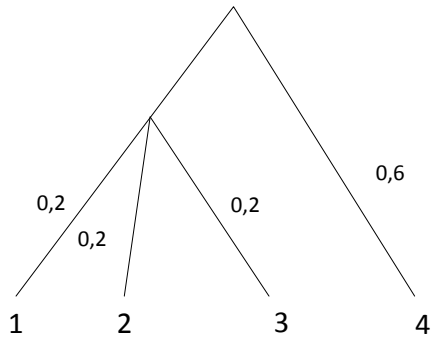
4.2 Técnicas de Consenso

4.2.1 Consenso Estrito

O Consenso Estrito é uma técnica que consiste em construir uma árvore filogenética de consenso por meio da reunião de todos os grupos ou clados que são comuns a todas as árvores. É considerado o tipo de consenso mais simples de todos os métodos de consenso e não considera o comprimento dos ramos. A Figura 4.1 mostra as filogenias da Figura 4.1(a) um exemplo de Consenso Estrito para 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 4.1(b).



(a) Duas topologias distintas de árvore com raiz e quatro folhas, incluindo o comprimento dos ramos.



(b) Árvore resultante do método de Consenso Estrito entre as árvores da Figura 4.2(a)

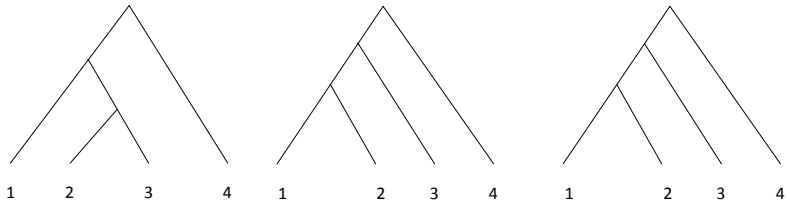
Figura 4.2: Exemplo de consenso pelo método Consenso Estrito utilizando comprimento de ramos.

4.2.2 Consenso estrito utilizando comprimento dos ramos

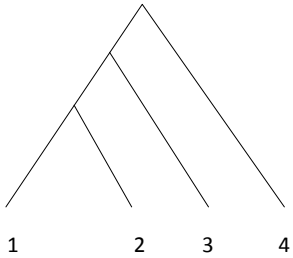
Esse método, além da topologia obtida pelo Consenso Estrito, recalcula os comprimentos dos ramos com base nos valores deles em todas as árvores. Cada comprimento na árvore de consenso é o comprimento mínimo do respectivo grupo ou clado comum. Para as árvores da Figura 4.2(a), aplicando a técnica de consenso estrito e considerando o comprimento dos ramos, a árvore de consenso seria aquela mostrada na Figura 4.2(b)

4.2.3 Consenso da regra majoritária

A árvore de consenso também pode ser construída considerando grupos ou clados que são comuns a uma determinada porcentagem das árvores analisadas. Essa porcentagem recebe o nome de regra majoritária (FELSENSTEIN, 2003). Por exemplo, a regra majoritária de 60% implica que a necessidade de se obter grupos ou clados comuns em 60% ou mais do total das árvores. Considere o conjunto de três árvores com raiz, conforme apresentado na Figura 4.3(a). Os grupos *1,2*, *1,2,3* e *1,2,3,4* aparecem em duas das três árvores. Dessa forma, a árvore com regra majoritária de 60% ou mais corresponde a



(a) Três topologias distintas de árvores com raiz e quatro folhas.



(b) Consenso Majoritário.

Figura 4.3: Exemplo de consenso pelo método da Regra Majoritária.

árvore apresentada na Figura 4.3(b).

4.2.4 Consenso de regras majoritárias considerando comprimento de ramos

Pode-se também considerar o consenso da regra majoritária com o uso dos comprimentos dos ramos das árvores analisadas. A técnica consiste em calcular o comprimento de determinado ramo utilizando a média de todos os comprimentos daquele ramo nas árvores analisadas. Quando não houver o ramo para determinada árvore, seu comprimento é considerado como zero para efeito do cálculo da média.

4.3 Considerações sobre as técnicas de consenso

Neste capítulo foram apresentados os principais métodos de confiança utilizados na filogenia, cabe ao usuário escolher qual método é mais apropriado para cada tipo de dado e de análises a serem realizadas, uma vez que nem todas as técnicas são apropriadas para todas as situações.

Um caso importante que pode dificultar a utilização de determinada técnica é a existência ou não de raiz em uma árvore. A extensão para os casos de árvores sem raiz, a partir de técnicas adequadas para árvores com raiz, pode produzir diferentes resultados de árvores de consenso (BRYANT, 2003).

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 consideram apenas os aspectos topológicos das árvores, sem considerar a distância que separa a evolução dos táxons.

5

Proposta do Método *CladeNet*

5.1 Introdução

Neste capítulo é apresentado o método *CladeNet*, uma abordagem capaz de detectar automaticamente clados a partir de uma filogenia e também de uma União de Filogenias em Rede *(UFR)*. Diferentes filogenias podem ser obtidas, por exemplo, considerando métricas de distância que se baseiam em modelos de substituição distintas, dentre as quais podem ser destacadas: *Jukes-Cantor* (JUKES T.H E CANTOR, 1969), *Kimura 2-Parâmetros* (KIMURA, 1980) e o modelo proporcional de *Felsenstein F84* (KISHINO; HASEGAWA, 1989a; FELSENSTEIN, 1985) (ver Capítulo 3). Filogenias obtidas por diferentes métodos (Evolução Mínima, Parcimônia, entre outras, ver Capítulo 3) também podem ser utilizadas.

Foram investigadas duas abordagens de geração de diversas filogenias para se determinar clados pelo *CladeNet*. A primeira consiste na utilização do método do consenso por regra majoritária para se construir uma filogenia que representa o consenso de informações de várias filogenias (*Consense-Phylip*, Seção 5.2). O outro método para extração de informação consensual de várias filogenias é a *UFR*, uma proposta desta pesquisa. As demais seções deste capítulo estão organizadas como descrito a seguir. A Seção 5.3 detalha como são detectados os clados usando o *CladeNet* e a Seção 5.4 apresenta as considerações finais sobre o conteúdo deste capítulo.

5.2 Análise de Várias Filogenias

Esta seção descreve os dois métodos utilizados nesta pesquisa para se extrair informações consensuais de várias filogenias: *Consense-Phylip* e *UFR*.

Método *Consense-Phylip*

O *Consense-Phlyip* disponibiliza dois tipos de métodos de consenso: o Consenso Estrito e o Consenso por Regra Majoritária (ver Capítulo 4), que tem sido mais aplicado segundo a literatura e que é o método utilizado nesta pesquisa. Primeiramente, o *Consense-Phylip* lê um arquivo de réplicas de árvores gerado por uma análise de reamostragem, no caso deste trabalho, a *Bootstrap*. Basicamente a árvore de consenso é construída considerando grupos que são comuns em mais de 60% das árvores analisadas. A árvore de consenso tem impresso em cada grafo um número que indica quantas vezes o grupo ou a espécie ocorreu para todas as árvores. Assim dada uma árvore de consenso gerada a partir de 32 réplicas de árvores, se o grupo ocorreu em todas as árvores, então o número 32 aparece associado ao grupo na árvore de consenso.

Método *UFR*

A *UFR* é um método novo, proposto nesta pesquisa junto ao *CladeNet*. Para encontrar um consenso entre filogenias, a *UFR* interliga filogenias através de seus nós folhas. Considere, por exemplo, que 3 filogenias são geradas para a construção da *UFR*, sendo elas, 1 filogenia referente à métrica *Jukes-Cantor*, 1 filogenia referente à métrica *Kimura* e 1 filogenia referente a métrica *F84*. Essas 3 filogenias são combinadas fixando os nós folhas (associados às espécies), que são comuns em todas as 3 árvores filogenéticas e, em seguida, adicionando os demais nós internos e arestas das 3 árvores relativas às métricas *Jukes-Cantor, Kimura* e *F84*. A Figura 5.1 ilustra a construção de uma *UFR* a partir das 3 filogenias usadas no exemplo.

A partir dessas 3 filogenias outras combinações podem ser feitas formando mais estruturas de rede, por exemplo, além da combinação de *JC+KI+F84* podem ser realizadas as combinações *JC+KI*, *JC+F84* e *KI+F84*. Além disso, combinações de árvores filogenéticas geradas por outros métodos filogenéticos (ver Capítulo 3) também podem ser adicionados à rede. Dessa forma, a *UFR* pode combinar informações de várias filogenias, todas obtidas a partir de um mesmo conjunto de espécies. Com base nessas filogenias (uma para cada métrica ou método selecionado), pode-se salientar os clados mais confiáveis a partir da rede utilizando o *CladeNet*.

5.3 Identificação de clados

As saídas tanto do método *Consense-Phylip* quanto da *UFR* podem ser vistas como uma Rede Complexa (NEWMAN, 2004), isto é, um grafo esparso de larga-escala. Em outras palavras, uma filogenia de consenso é uma rede e uma composição de filogenias pela

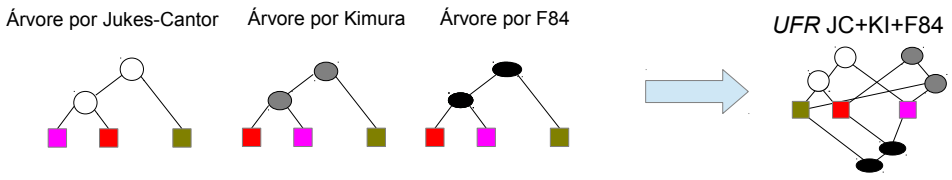


Figura 5.1: *UFR JC+KI+F84.*

UFR também é uma rede que, com quantidade relativamente grande de nós (larga-escala), corresponde a uma Rede Complexa.

Grupos de nós mais fortemente conectados na rede podem ser identificados por métodos de agrupamento. Esses grupos são chamados de comunidade na área de Redes Complexas. Uma comunidade com nós folhas, (espécies) corresponde a um potencial clado, um agrupamento forte de espécies segundo propriedades de Redes Complexas.

Há vários algoritmos para detecção de comunidades, por exemplo, *Adaptive Clustering* (ROESE, 1969), *Extreme Optimization* (DUCH; ARENAS, 2005) (ver Apêndice B) e o *Fast Algorithm* (NEWMAN, 2004). Este último foi o método utilizado para esta proposta de trabalho, uma vez que este tem sido aplicado em diversas pesquisas do grupo Sistemas Embarcados e Evolutivos (SEE) do ICMC-USP em problemas envolvendo agrupamento e Redes Complexas (NEWMAN, 2004). Além disso, a tese de (CROCOMO, 2012) mostra que as demais técnicas não apresentam desempenho superior ao *Fast Algorithm* para agrupamento em Redes Complexas.

O *Fast Algorithm* é baseado na ideia de modularidade, que é uma métrica para qualificar as estruturas das comunidades (NEWMAN, 2004). Para isso, define-se o termo e_{ij} como a fração de arestas que conectam as comunidades i e j em relação ao total de arestas da rede. O termo e_{ii} corresponde a fração de arestas dentro da comunidade i em relação ao total de arestas. A fração das arestas conectadas a nós da comunidade i é dada por $a_i = \sum_j e_{ij}$. Uma estimativa da fração de arestas internas à comunidade i, se a estrutura de conexões da rede não for considerada, é dada por a_i^2. Em outras palavras, esse é o valor esperado de arestas internas de as a_i fossem escolhidas de forma aleatória. Com isso, $e_{ii} - a_i^2$ estima o quanto o nível de conectividade de i distancia-se de ser aleatório. Com base nessa diferença, define-se o índice de modularidade Q conforme mostra a Equação 5.1. A Figura 5.2 ilustra os conceitos descritos acima.

$$Q = \sum_i (e_{ii} - a_i^2). \tag{5.1}$$

De posse de um índice para quantificar a qualidade das comunidades, pode-se construir um algoritmo que otimize o valor de Q sobre todas as possíveis divisões da rede em

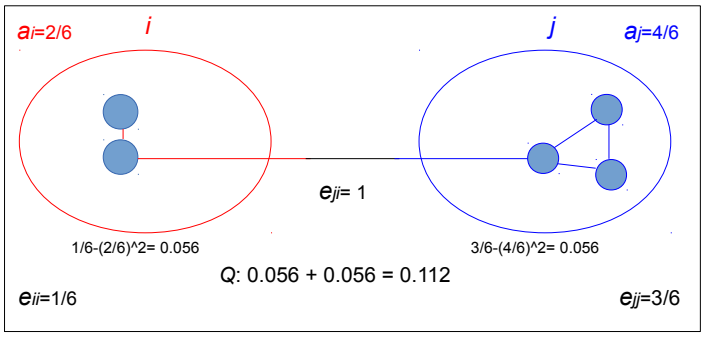


Figura 5.2: Rede com exemplo de cálculo de modularidade do *Fast Algorithm*.

comunidades. Foi proposto em (NEWMAN, 2004) um algoritmo guloso para esse fim. Inicialmente, esse algoritmo considera cada nó da rede como uma comunidade. Repetidamente, as comunidades são agrupadas em pares considerando todos os pares possíveis. Então, o agrupamento com maior valor de Q é preservado. A Figura 5.3 mostra o gráfico que ilustra as possíveis divisões do *Fast Algorithm* diminuindo o número de comunidade.

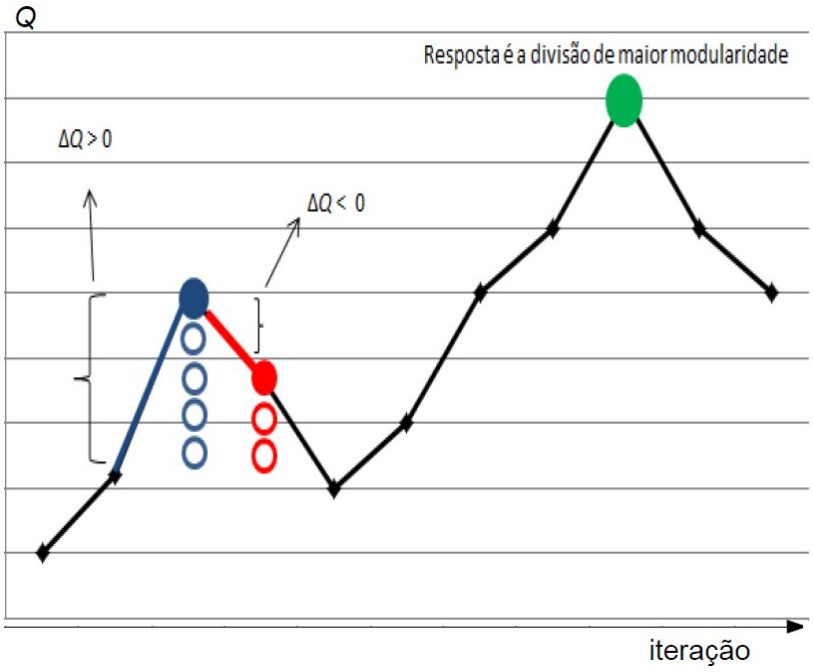


Figura 5.3: Possíveis valores de modularidade Q para divisões do *Fast Algorithm*, conforme o algoritmo agrupa comunidades.

É importante ressaltar que apenas comunidades que possuam pelo menos uma aresta ligando seus nós podem ser possivelmente unidas pelo algoritmo. Isso limita a um máximo de p pares de comunidades, em que p é o número de arestas do grafo. A variação em Q ao se unir duas comunidades é dada pela Equação 5.2.

$$\Delta Q = e_{ij} + e_{ji} - 2a_i a_j = 2(e_{ij} - a_i a_j). \tag{5.2}$$

O valor inicial de e_{ij} é igual à metade do grau de cada nós, uma vez que inicialmente cada comunidade é formada por apenas um nós. Após a junção de duas comunidades, os valores de e_{ij} devem ser atualizados. O Algoritmo 1 apresenta o funcionamento do *Fast Algorithm*.

Algorithm 1: *Fast Algorithm*

Separe os nós em n comunidades **enquanto** *número de comunidades > 1* **faça**
  Junte as duas comunidades i e j cuja aglomeração forneça o maior acréscimo (ou menor decréscimo) ΔQ no valor da modularidade Salve a divisão resultante e sua modularidade
fim
Escolha a divisão com a melhor modularidade encontrada

A combinação do procedimento descrito na Seção 5.2 com o *Fast Algorithm* apresentado nesta Seção para identificação de clados compõem o método *CladeNet*. As Figuras 5.4 5.5 5.6 mostram diagramas que sintetizam o funcionamento do *CladeNet*, respectivamente, para três procedimentos diferentes de obtenção de rede à qual o *Fast Algorithm* é aplicado: *i)* as filogenias são organizadas em uma rede pela *UFR*, *ii)* a rede corresponde a uma filogenia de consenso gerada usando o *Consense-Phylip* e, finalmente, *iii)* a filogenia inferida usando o método *MrBayes* é a rede utilizada pelo *CladeNet*.

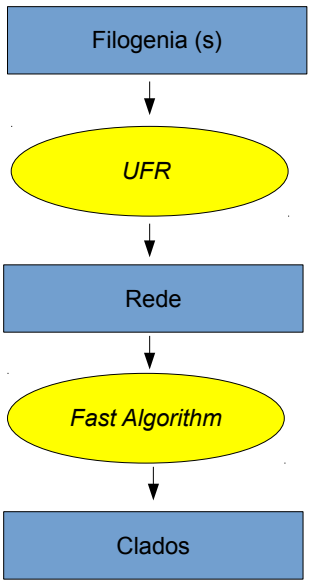


Figura 5.4: Diagrama contendo as principais etapas do *CladeNet* usando a *UFR*.

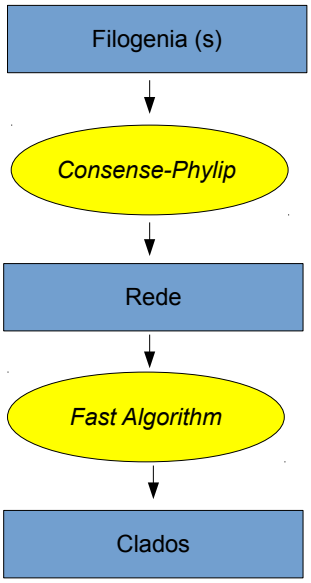


Figura 5.5: Diagrama contendo as principais etapas do *CladeNet* usando o *Consense-Phylip*.

5.4 Considerações Finais

O principal objetivo do *CladeNet* é encontrar clados mais confiáveis e robustos através de uma filogenia ou uma *UFR*, ambas constituídas por filogenias reconstruídas através de diversos métodos ou métricas. O *CladeNet* é capaz de detectar clados consistentes utilizando o *Fast Algorithm*, conforme apresentado no Capítulo 6.

Outros algoritmos de detecção de comunidades também podem ser empregados ao *CladeNet*. A investigação de outra técnica de agrupamento para o *CladeNet* pode possibilitar um desempenho superior, o que é interessante uma vez que as pesquisas em Redes Complexas tem gerado um número significativo de novos algoritmos de detecção de comunidades recentemente. Além disso, o *CladeNet* é flexível em relação às formas (métodos) para se obterem uma diversidade de filogenias (entrada do algoritmo), bem como, é flexível em relação a quantidade de filogenias a serem utilizadas.

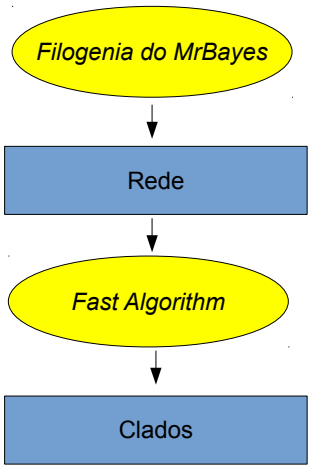


Figura 5.6: Diagrama contendo as principais etapas do *CladeNet* usando o *MrBayes*.

6

Procedimentos Experimentais e Resultados

Neste capítulo são apresentados os resultados da aplicação do *CladeNet* para conjuntos de dados biológicos. O objetivo dos experimentos é mostrar o desempenho do *CladeNet* para encontrar e avaliar a qualidade dos clados. As demais seções deste capítulo estão organizadas como descrito a seguir. A Seção 6.1 apresenta os conjuntos de teste empregados nos experimentos. A Seção 6.2 detalha os procedimentos experimentais e os parâmetros utilizados do *CladeNet*. A Seção 6.3 mostra os resultados de determinação de clados obtidos pelo *CladoNet* a partir de filogenias do*MrBayes*, do método de consenso do *Phylip* e da *UFR* para cada conjunto de dados. Além disso, a Seção 6.3 também apresenta comparação desses resultados com um conjunto de clados usado como referência (obtido a partir de filogenia construída pelo *MrBayes*). Finalmente, a Seção 6.4 apresenta considerações finais sobre o conteúdo deste capítulo.

6.1 Conjuntos de Dados

O *CladeNet* foi testado com 7 conjuntos de dados biológicos. Esses dados foram obtidos de pesquisas anteriores realizadas no grupo de pesquisa de Sistemas Embarcados e Evolutivos (SEE) do ICMC-USP. Esses conjuntos de dados possuem grupos de tamanhos diferentes (de algumas dezenas a várias centenas) implicando em construção de árvores com complexidade variável. Conforme mostrado em (ROKAS et al., 2003), para se obter inferência filogenética com alta qualidade é importante lidar com um significativo número de espécies, assim, testes com variedade no tamanho das sequências mostram-se fundamentais. Por outro lado, é um desafio conseguir métodos computacionais eficientes (rápidos) e capazes de manter a qualidade da inferência para grandes instâncias. As características de cada conjunto são descritos a seguir:

- O primeiro conjunto é o menor, possui 12 espécies de primatas com 898 caracteres (Hayasaka, Gojobori e Horai (1988) *apud* Vittori, Delbem e Pereira (2008));

- O segundo conjunto possui 19 espécies de aves com 1.757 caracteres (Pereira e Baker (2006) *apud* Vittori, Delbem e Pereira (2008));
- O terceiro conjunto envolve 20 espécies de mamíferos com 18.478 caracteres (Cao et al. (1998); Perretto (2010); Perretto e Lopes (2005) *apud* Vittori, Delbem e Pereira (2008));
- O conjunto *rbcL_55* contém 55 sequências de 1.314 nucleotídeos cada uma do gen rbcl de cloroplasto de plantas (Lemmon e Milinkovitch (2002), Lewis (1998) *apud* Cancino e Delbem (2007));
- O conjunto *mtDNA_186* possui 186 sequências de DNA mitocrondrial humano de 16.608 nucleotídeos cada uma obtido do banco mtDB (The Human Mitochondrial genome Database) (Ingman e Gyllensten (2006) *apud* Cancino e Delbem (2007);
- O conjunto *RDPII_218* envolve 218 sequências de RNA procariótico de 4.182 nucleotídeos obtido do projeto *RDPII (The Ribosomal Database Project II)* (Cole et al. (2005); Guindon e Gascuel (2003b); Stamatakis (2005); Vinh e Haeseler (2004) *apud* Cancino e Delbem (2007));
- O conjunto *ZILLA_500* possui 500 sequências de 1.428 nucleotídeos cada uma do gen rbcl de plantas (Chase et al. (1993); Goloboff e Farris (2001); Guindon e Gascuel (2003a); Stamatakis (2005); Vinh e Haeseler (2004) *apud* Cancino e Delbem (2007))).

6.2 Procedimentos Experimentais

Inicialmente são descritos na Seção 6.2.1 o procedimento de reamostragem e aspectos importantes do uso de cada método. Por fim, o processo de cálculo do nível de incongruência entre os conjuntos de clados é apresentado na Seção 6.2.2.

6.2.1 Reamostragem e obtenção de Clados

Para uma análise estatística dos resultados é fundamental utilizar várias amostras. Para isso, foi utilizada uma técnica de reamostragem para se obter réplicas de um mesmo conjunto de dados, conforme mostrado no Capítulo 4. Nesse sentido em todos os experimentos, são utilizados os seguintes passos:

- **Passo 1:** São geradas 32 réplicas para cada um dos 7 conjuntos de dados;

- **Passo 2:** Essas 32 réplicas são aplicadas no MrBayes para geração de 32 filogenias e então é empregado o *CladeNet* para detecção de clados, gerando assim, 32 conjuntos de clados. A partir dos 32 conjuntos de clados é determinado o *conjunto de referência*, que consiste no conjunto que mais se repete entre os 32 conjuntos de clados identificados pelo *CladeNet*;

- **Passo 3:** As 32 réplicas são aplicadas ao método de distância, para cada uma das métricas *Jukes-Cantor*, *Kimura* e *F84*, gerando assim 32 matrizes de distância para cada uma das métricas empregadas. A cada uma das matrizes de distância geradas, o *NJ* é aplicado para a construção de filogenias. Isso resulta no total de 96 filogenias para cada um dos 7 conjunto de dados.

Cada conjunto de 32 filogenias é então processado pelo *CladeNet* para a identificação de clados por meio de 2 procedimentos descritos a seguir:

1. O *CladeNet* usa o *Consense-Phylip* (ver Figura 5.5) para gerar uma árvore consenso para cada um dos 3 conjuntos (*Jukes-Cantor, Kimura e F84*) de 32 filogenias. Além disso, 16 filogenias (sempre escolhidas aleatoriamente) do conjunto referente à métrica *Jukes-Cantor* são utilizadas com 16 filogenias do conjunto referente à métrica *Kimura*, gerando uma filogenia de consenso, denominada *JC+KI*. Essa estratégia é repetida de forma similar gerando os consensos *JC+F84* e *KI+F84*. Em seguida, 11 filogenias de cada conjunto obtidas pela métrica *Jukes-Cantor*, 11 filogenias do conjunto gerado usando a métrica *Kimura* e 11 filogenias do conjunto construído com base na métrica *F84* são utilizadas para produzir o consenso *JC+KI+F84*. Em Resumo, a partir dos 3 conjuntos de filogenias obtidos usando as métricas *Jukes-Cantor*, *Kimura* e *F84*, são construídas 7 filogenias de consenso *(JC, KI, F84, JC+KI, KI+F84 e JC+KI+F84)* para um mesmo conjunto de dados biológicos. Por fim, o *CladeNet* identifica clados para cada uma das 7 filogenias de consenso;

2. O *CladeNet* combina filogenias em uma rede usando a *UFR* para cada conjunto de filogenias *(JC, KI, F84, JC+KI, KI+F84 e JC+KI+F84)*, ver Figura 5.4. Assim, clados são identificados para esses 7 conjuntos de filogenias. Em resumo, são gerados 7 conjuntos dos clados produzidos pelo procedimento do Item 1 e outros 7 pelo procedimento do Item 2 para cada conjunto de dados. Por fim, cada um desses 7 conjuntos de clados são comparados com o *conjunto de referência* (ver Passo 2).

6.2.2 Avaliação de Incongruências entre Clados

Os clados gerados pelo *CladeNet* usando tanto a árvore *Consense-Phylip* (Item 1 da Seção 6.2.1) quanto da *UFR* (Item 2 Seção 6.2.1) foram comparados com os clados do

conjunto de referência gerado pelo *MrBayes*. As análises são feitas a partir do reconhecimento de diferenças entre os clados de cada conjunto, assim, clados que não são iguais aos clados do *conjunto de referência* do *MrBayes* são considerados incongruentes. A Tabela 6.1 mostra como os clados gerados pelo *Consense-Phylip* para o conjunto de dados de 12 espécies de primatas. As espécies são rotuladas com números de 0 a 11 e cada clado é rotulado por um número aleatório (de 0 a 11) de forma que espécies em um mesmo clado possuem o mesmo rótulo. A Tabela 6.1 mostra os clados obtidos pelo *CladeNet* a partir de 3 filogenias geradas pelo *Consense-Phylip* utilizando as métricas *Jukes-Cantor*, *Kimura* e *F84*. Além disso, a Figura 6.1 apresenta na segunda coluna o *conjunto de referência* (ver Seção 6.2.1).

Tabela 6.1: Exemplo de conjunto de clados gerados pelo *CladeNet* a partir de filogenias do *Consense-Phylip*, usando as métricas *Jukes-Cantor*, *Kimura* e *F84*. Espécies com o mesmo rótulo (número) pertencem ao mesmo clado.

	Conjunto de Referência	*Consense – Phylip*	*Consense - Phylip*	*Consense – Phylip*
		Jukes-Cantor	*Kimura*	*F84*
Espécies	Clados	Clados	Clados	Clados
0	0	0	0	0
10	0	4	4	4
11	0	0	0	0
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	2	2	2
5	2	2	2	2
6	3	3	3	3
7	3	3	3	3
8	3	3	3	3
9	3	3	3	3

A Tabela 6.2 destaca em cinza as incongruências em cada clado mostrado no exemplo da Tabela 6.1. As linhas destacadas em cinza evidenciam as espécies que não estão no mesmo clado no *conjunto de referência*. A partir delas, verifica-se que as espécies 10 e 4 estão associadas ao clado errado para as colunas relativas ao *Consense-Phylip*.

A Tabela 6.3 mostra um conjunto um pouco mais difícil para se calcular as incongruências, pois mais espécies foram acrescentadas ao grupo. A partir dos destaques em cinza, verifica-se que as espécies 6, 9, 13, 12 e 18 para a métrica *Jukes-Cantor*, as espécies 6, 7, 8, 13, 15, 16 para a métrica de *Kimura* as espécies 6, 9, 13, 15, 16 do *Consense-Phylip* estão associadas ao clado errado.

Em casos de conjuntos grandes é difícil calcular as incongruências pois se torna mais complicado determinar a espécie que está correta em um clado. Por exemplo, no caso

Tabela 6.2: Totais de incongruências do conjunto de dados de 12 primatas para o conjunto de clados gerado pelo *CladeNet* usando o *Consense-Phylip*.

	Conjunto de Referência	*Consense - Phylip*	*Consense - Phylip*	*Consense - Phylip*
		Jukes-Cantor	*Kimura*	*F84*
Espécies	Clados	Clados	Clados	Clados
0	0	0	0	0
10	0	4	4	4
11	0	0	0	0
1	1	1	1	1
2	1	1	1	1
3	1	1	1	1
4	1	2	2	2
5	2	2	2	2
6	3	3	3	3
7	3	3	3	3
8	3	3	3	3
9	3	3	3	3
	Totais de Incongruências	2	2	2

da Tabela 6.3 na coluna de *Kimura* aparecem 3 clados diferentes para as espécies 7, 8 e 9, no caso apenas um deles estaria correto quando comparado aos clados do *conjunto de referência*. Nesse exemplo, a espécie 9 do clado 1 foi considerada correta, mas poderia ter sido a espécie 7 ou 8 a correta. Um outro exemplo acontece na coluna referente à métrica *F84* em que as espécies 15, 16, 17 e 18 estão divididas em 2 clados, sendo considerados 2 clados errados e 2 clados corretos, assim para quaisquer um desses clados escolhido como correto o nível de incongruência não muda. Em casos em que tal escolha muda o valor de incongruências no clado, deve-se escolher como corretas para um clado as espécies que resultem em menos incongruências no clado.

6.3 Resultados

Esta seção inicialmente descreve os resultados obtidos realizando os experimentos conforme descrito na Seção 6.2.1. Resultados referentes aos clados obtidos pelo *CladeNet* através das filogenias geradas pelo *MrBayes* são agrupadas na Seção 6.3.1. As incongruências relativas aos clados obtidos pelo *CladeNet* a partir das filogenias do *Consense-Phylip* estão organizadas na Seção 6.3.2 e, finalmente, os resultados relativos aos clados determinados pelo *CladeNet* através da *UFR* são apresentadas na Seção 6.3.3.

6.3.1 Clados obtidos pelo *CladeNet* usando MrBayes

Para gerar as filogenias pelo *MrBayes* é necessário a alteração de vários parâmetros utilizados por esse *software* (ver Apêndice C). Para estes resultados foram utilizados os

Tabela 6.3: Totais de incongruências do conjunto de dados de 19 aves para o conjunto de clados gerado pelo *CladeNet* usando o *Consense-Phylip*.

	Conjunto de Referência	*Consense-Phylip*	*Consense-Phylip*	*Consense-Phylip*
		Jukes-Cantor	*Kimura*	*F84*
Espécies	Clados	Clados	Clados	Clados
0	0	0	0	0
1	0	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
5	1	0	0	0
6	1	1	1	1
7	2	0	0	2
8	2	0	2	2
9	2	2	1	1
10	3	1	1	1
13	3	2	3	3
14	3	1	1	1
11	4	1	2	2
12	4	2	2	2
15	5	1	1	3
16	5	1	1	3
17	5	1	2	2
18	5	2	2	2
	Totais de incongruências	5	6	5

Parâmetros *default* do *MrBayes*, que consiste no parâmetro ***nst=6 rates=invgamma***. Este parâmetro considera o modelo de substituição com distribuição gamma para uma proporção diversificada entre os sítios. A Tabela 6.4 apresenta os valores dos parâmetros utilizados do *MrBayes* para a geração de filogenia a partir de cada conjunto de dados.

Tabela 6.4: Parâmetros utilizados para gerar as filogenias pelo *MrBayes* para cada conjunto de dados.

Parâmetros utilizados para geração de filogenias – MrBayes								
	Nset	rates	ngen	samplefreq	printfreq	diagnfreq	Relburnin	Conformat
12 primatas	6	invgamma	1000000	100	100	10000	no	Simple
19 aves	6	invgamma	1000000	100	100	10000	no	Simple
20 mamíferos	6	invgamma	1000000	100	100	10000	no	Simple
55 rbcL	6	invgamma	1000000	150	100	10000	no	Simple
186 mtDNA	6	invgamma	1000000	350	100	10000	no	Simple
218 RDPII	6	invgamma	1000000	400	100	10000	no	Simple
500 ZILLA	6	invgamma	1000000	10000	100	10000	no	Simple

Como observado na Tabela 6.4 dependendo do tamanho do conjunto de dados é necessário modificar o parâmetro **samplefreq** que controla a frequência de amostragem relativas as cadeias Markovianas (Capítulo 3). Os resultados dos experimentos evidenciam que o dobro da quantidade de espécies em um conjunto de dados em geral é um valor adequado. Observa-se que 100 é um valor mínimo sugerido na literatura Ronquist

e Huelsenbeck (2003) para esse parâmetro. Para os demais parâmetros foram utilizados os valores padrões recomendadas da literatura.

Todos os conjuntos de clados, gerados pelo *CladeNet* a partir das 32 filogenias do *MrBayes* foram analisados isoladamente cada um, comparando os clados de cada conjunto com os clados do *conjunto de referência*. Na comparação de cada conjunto de clados, os clados que não são iguais aos clados do *conjunto de referência* são considerados incongruentes. Para cada clado desse tipo, contam-se quantas espécies diferem do clados do *conjunto de referência*, assim como descrito na Seção 6.2.2.

A Figura 6.1 apresenta o número médio de incongruências por conjunto de dados, isto é, divide-se o total de incongruências por 32, referente às 32 réplicas. Conforme pode ser observado, o conjunto de 55 espécies foi o primeiro a demonstrar incongruências, iniciando com um total de 9 incongruências por árvore. O conjunto de 186 espécies de DNA mitocondrial humano apresenta ter 31 incongruências por árvore, valor 3,5 vezes maior que o do conjunto de 55 espécies. O conjunto de 218 espécies apresentou 39 incongruências. Por outro lado, para o conjunto de 500 espécies, observou-se um número de incongruências bem mais alto, com 108 incongruências por árvore.

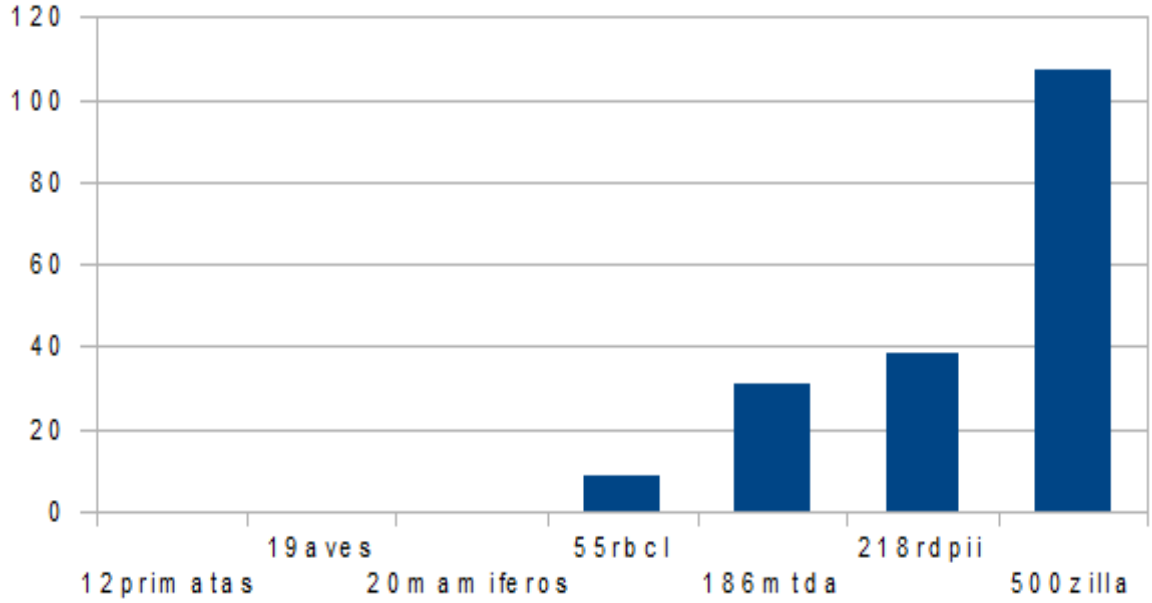


Figura 6.1: Incongruência Média, usando conjunto de clados obtido pelo *CladeNet* a partir de filogenias do *MrBayes*.

A Figura 6.2 apresenta a incongruência média normalizada, a normalização corresponde à divisão da incongruência média pelo número de quantidade de espécies de cada conjunto de dados. Naturalmente, observa-se que as incongruências são nulas para os

conjunto de 19 aves, 20 primatas, e 12 mamíferos, enquanto começa a ocorrer incongruências a partir do conjunto de 55 espécies aumentando suavemente conforme aumenta o número de espécies por conjunto. Um aumento um pouco maior ocorre para o conjunto de 500 espécies. Esses resultados evidenciam que, como mostrado na Figura 6.2, a taxa de incongruência por espécies adicionada (a um conjunto) aumenta nos clados obtidos pelo *MrBayes*, porém, esse aumento é pequeno em relação à quantidade de espécies adicionadas.

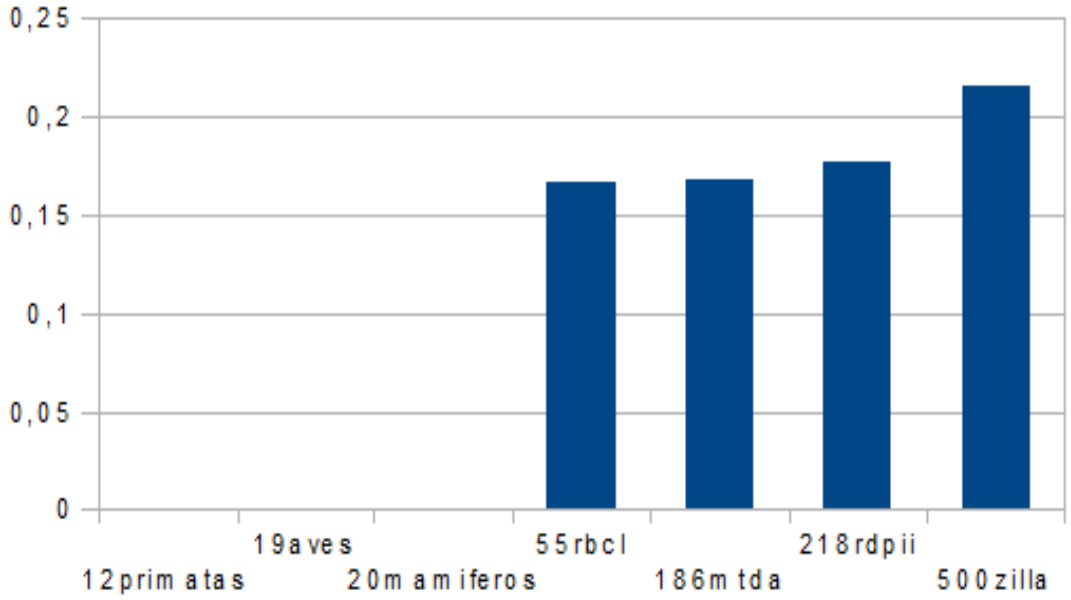


Figura 6.2: Incongruência Média Normalizada, usando conjunto de clados obtido pelo *CladeNet* a partir de filogenias do *MrBayes*.

6.3.2 Clados obtidos pelo *Consense-Phylip*

De forma similar ao apresentado na Seção 6.3.1 todos os conjuntos de clados (gerados pelo *CladeNet* a partir das 3 filogenias de consenso geradas pelo *Consense-Phylip*) foram analisados isoladamente, comparando os clados de cada conjunto com os clados do *conjunto de referência*. Os resultados apresentados nas Figuras 6.3(a) e 6.3(b) mostram incongruências já nos menores conjuntos de dados, o conjunto de 12 primatas apresenta 2 incongruências para cada métrica. No conjunto de 19 aves, essas incongruências diminuem considerando a métrica *Jukes-Cantor* (não ocorreu nenhuma incongruência); em relação às métricas de *Kimura* e *F84* ocorreu 1 incongruência para cada uma. O conjunto de 20 mamíferos manteve-se estável, apresentando 2 incongruências relativas a cada métrica. Observa-se que para os menores conjuntos, as incongruências foram relativamente baixas

e estáveis em relação à variação dos tipos de métricas utilizadas.

Os conjuntos maiores iniciaram com um aumento relevante de incongruências. Por exemplo, para o conjunto de 55 espécies usando a métrica *Jukes-Cantor* ocorreram 11 incongruências, praticamente um valor 5 vezes maior que o conjunto de 20 mamíferos. Com a métrica de *Kimura* ocorreram 13 incongruências, um valor 6 vezes mais do que no conjunto de 20 mamíferos. Para a métrica de *F84* houve apresentou 15 incongruências, também um valor maior do que o ocorrido no conjunto de 20 mamíferos. Observa-se que as incongruências aumentam conforme aumenta o número de espécies em cada conjunto de dados.

O conjunto de 186 espécies de DNA mitocondrial humano apresentou 44 incongruências para a métrica de *Jukes-Cantor*, 36 incongruências para *Kimura* e 47 incongruências para *F84*. Observa-se que os melhores resultados ocorreram para a métrica *Kimura*.

O conjunto de conjunto de 218 espécies apresentou 68 incongruências para *Jukes-Cantor*, 67 incongruências para *Kimura* e 65 incongruências para *F84*. Observa-se que neste conjunto de espécies os melhores resultados ocorreram usando métrica *F84*.

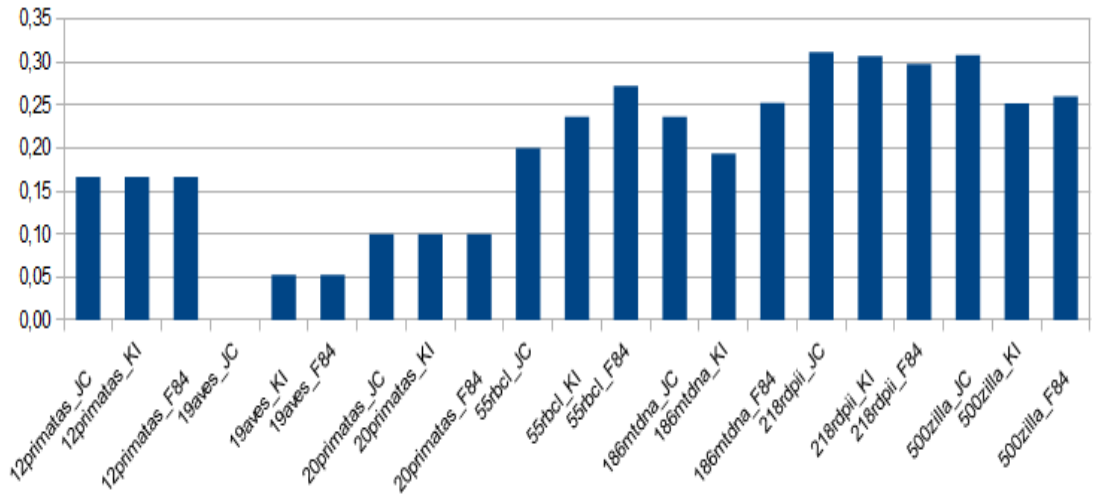
O conjunto de 500 espécies apresentou 154 incongruências para a métrica *Jukes-Cantor*, 126 incongruências para *Kimura* e 130 incongruências para a métrica *F84*. Para esse conjunto de espécies, os melhores resultados foram obtidos usando a métrica de *Kimura*.

A Figura 6.4(a) e 6.4(b) apresenta, além das incongruências média normalizada, as incongruências líquidas, que corresponde ao total de incongruências média normalizada descontado o total de incongruências média normalizada obtida usando o *MrBayes* (Seção 6.3.1). Com isso, busca-se estimar a contribuição em termos de incongruências exclusivamente do *CladeNet* usando o *Consense-Phylip*, pois desconta-se o valor estimado de incongruências geradas pelo próprio *MrBayes* (o qual é utilizado para construir o *conjunto de referência*).

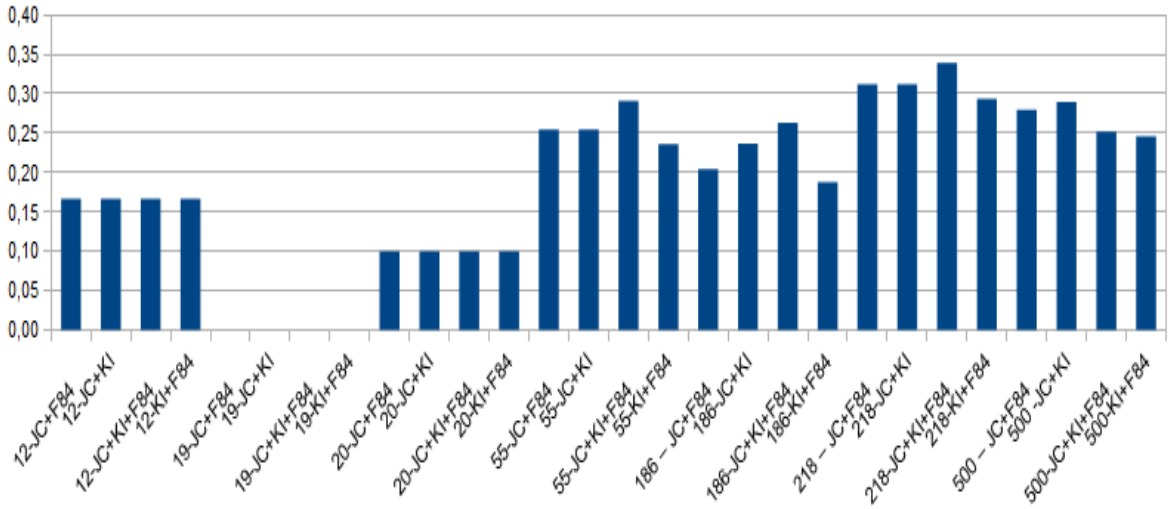
Na Figura 6.4 observa-se que as incongruências aumentam suavemente conforme aumentava o número de espécies de cada conjunto. A incongruência líquida apresenta valores mais estáveis uma vez que todas são inferiores a 0,20 para todos os conjuntos de dados. Em outras palavras, apesar de ocorrer variação dos níveis de incongruências líquidas para diferentes conjuntos de dados, esses níveis são limitados em um patamar baixo, próximos ou inferiores às incongruências líquidas dos conjuntos de dados com menos espécies.

Esses resultados motivam o uso do *CladeNet* para conjunto de dados de diferentes tamanhos, uma vez que o *CladeNet* apresenta nível de incongruências estável conforme aumenta o número de espécies em um conjunto de dados (limitado por um patamar relativamente baixo).

Além disso, esse método possui a vantagem de requerer tempo computacional signi-

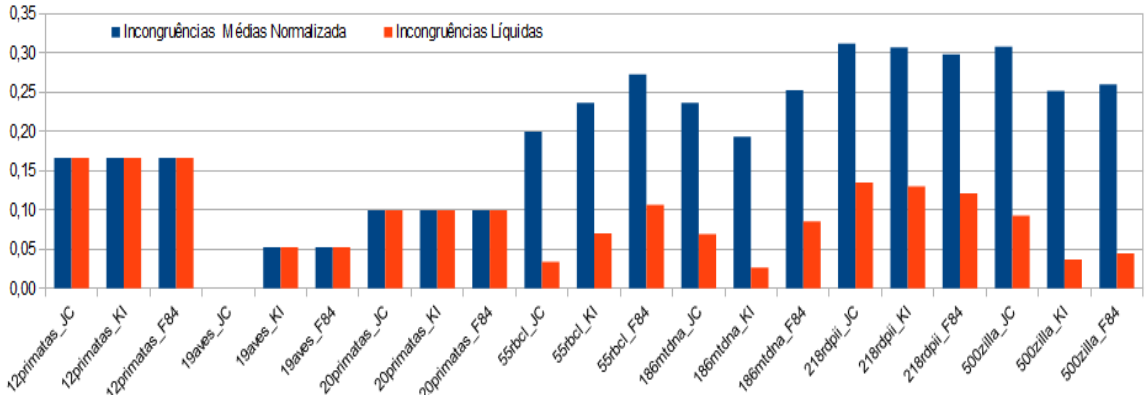


(a) Sem combinação de métricas

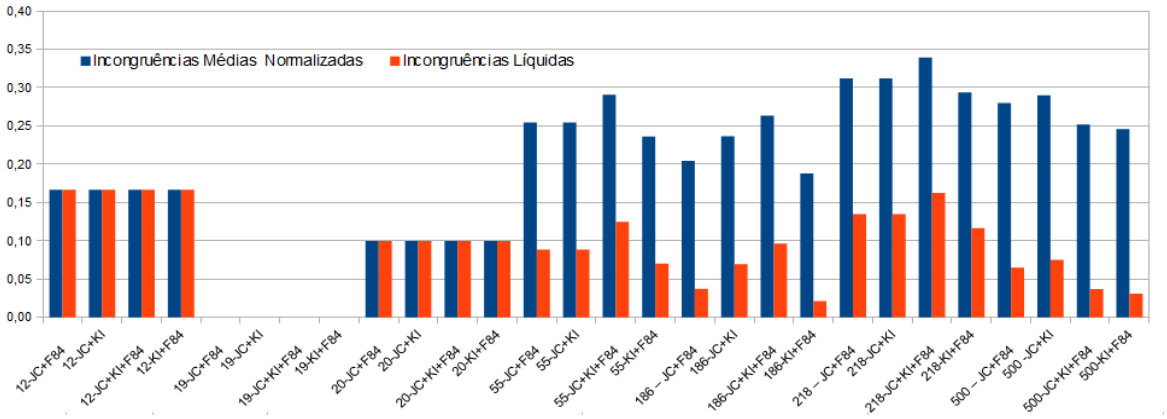


(b) Com combinação de métricas

Figura 6.3: Incongruências Médias Normalizadas, usando conjunto de clados obtido pelo *CladeNet* a partir de filogenias de consenso obtidas pelo *Consense-Phylip*.



(a) Sem combinação de métricas



(b) Com combinação de métricas

Figura 6.4: Incongruências Médias Normalizadas e Incongruências Líquidas, usando conjunto de clados obtido pelo *CladeNet* a partir de filogenias de consenso obtidas pelo *Consense-Phylip*.

ficativamente inferior ao método de Inferência Bayesiana usando a abordagem *MrBayes* seguido do *CladeNet (MrBayes+CladeNet)*. A Tabela 6.5 mostra os tempos computacionais requeridos em cada conjunto de dados testado com o NJ com diferentes métricas seguido do *CladeNet (NJ+CladeNet)* e com o *MrBayes+CladeNet*. Esses experimentos foram executados em um computador Intel(R) Core(TM) i7-2600 CPU @ 3.40GHz com 32GB de Memória RAM MB Intel DP67DE. Os valores de tempo apresentados na Tabela 6.5 correspondem ao tempo médio de 32 execuções.

Tabela 6.5: Tempos de execução para cada conjunto de dados para as abordagens *MrBayes+CladeNet* e *NJ+CladeNet*.

Conjunto de dados	*MrBayes+CladeNet* (minutos)	*NJ+CladeNet* (segundos)
12 primatas	30	7,5
19 aves	40	8
20 mamíferos	60	11,5
55 rbcL	90	17
186 mtDNA	420	128
218 RDPII	360	123
500 ZILLA	540	180

É interessante observar também que para os conjuntos maiores (com 55 ou mais espécies) o conjunto com mais incongruências líquidas é o RDPII (218 espécies), com cerca de 0,13 incongruências. Observa-se que para o conjunto ZILLA (500 espécies) a incongruência líquida é de valor 0,05. O RDPII diferentemente dos outros conjuntos é de espécies de procariotos e RNAs, que são moléculas menos estáveis que DNAs. Além disso, as sequências deste conjunto possuem muitos sítios com *gap* (falhas nas sequências) prejudicando a reconstrução das filogenias. Assim, o resultado apresentado na Figura 6.3 foi reanalisado retirando o conjunto RDPII, produzindo a Figura 6.5. A Figura 6.5 mostra a curva ajustada aos resultados usando a combinação de *KI+F84* (ver justificativa na Seção 6.3.4) que destaca que as incongruências líquidas diminuem significativamente conforme o tamanho dos conjuntos aumentam. A Figura 6.6 mostra as curvas de tendência ajustadas aos resultados de todas as combinações de métricas. Esses resultados indicam que os clados obtidos pelo *CladeNet* usando o *Consense-Phylip* aproxima-se significativamente dos clados obtidos pelo *CladeNet* principalmente para conjuntos maiores.

6.3.3 Clados obtidos pela *UFR*

De forma similar ao apresentado na Seção 6.3.1 e na Seção 6.3.2 todos os conjuntos de clados (gerados pelo *CladeNet* a partir de filogenias geradas pela *UFR*) foram anali-

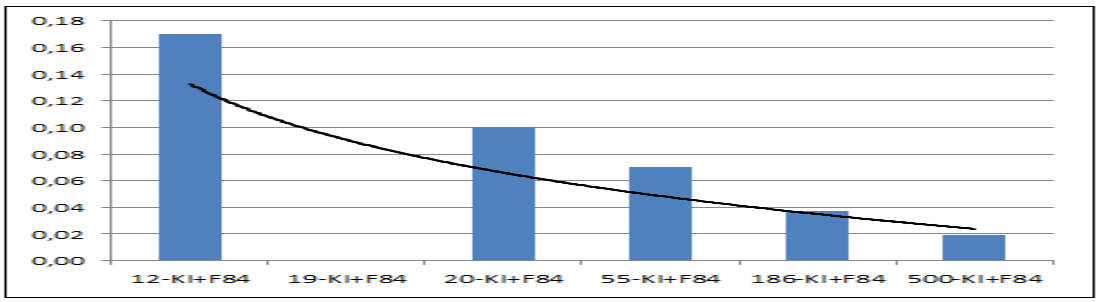


Figura 6.5: Curva de tendência de queda de incongruência usando conjunto de clados obtido pelo *CladeNet* a partir do *Consense-Phylip* para a combinação *KI+F84*.

sados isoladamente, comparando os clados de cada conjunto com os clados do *conjunto de referência*. Os resultados usando a *UFR* a partir de uma única filogenia são idênticos aos obtidos pelo *CladeNet* com *Consense-Phylip*, por isso as Figuras desta Seção apresentam somente os resultados envolvendo combinação de filogenias referentes às diferentes métricas.

Analisando os clados da *UFR* os conjuntos de dados pequenos demonstraram estar estáveis quanto a incongruências. Como mostrado na Figura 6.7, o conjunto de 12 primatas apresentou 2 incongruências para cada combinação de métricas, por outro lado, o conjunto de 19 aves manteve nulas a quantidade de incongruência em todas as combinações das métricas e o conjunto de 20 mamíferos obteve 2 incongruências para cada combinação.

Nos conjuntos maiores as incongruências aumentaram, para o conjunto de 55 espécies obteve-se 14 incongruências referente a união de *JC+F84*, 14 também para a união de *JC+KI*, 15 para a união de *JC+KI+F84*. O melhor resultado com 13 incongruências ocorreu para a união de *KI+F84*. Todas as combinações apresentaram um valor de incongruências praticamente maior que o conjunto de 20 mamíferos.

O conjunto de 186 espécies de DNA mitocondrial humano apresentou 33 incongruências para a união de *JC+F84*, 44 incongruências para a união de *JC+KI*, 30 incongruências para a união de *JC+KI+F84* e 40 incongruências para a união de *KI+F84*. Observa-se que a união de métricas *KI+F84* obteve os melhores resultados neste conjunto.

Por outro lado, o conjunto de 218 sequências de RNA procariótico apresentou 94 incongruências para a união métricas *JC+F84*, 67 incongruências para a união de *JC+KI*, 80 incongruências para a união de *JC+KI+F84* e 75 incongruências para a união de *JC+KI*. Observa-se que os melhores resultados desse conjunto foi da união de métricas *JC+KI*.

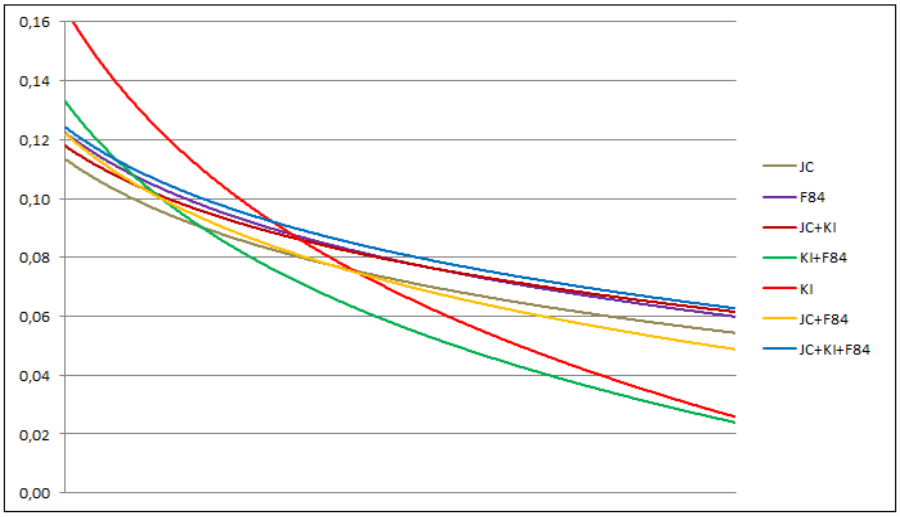


Figura 6.6: Curva de tendência de queda de incongruência usando conjunto de clados obtido pelo *CladeNet* a partir do *Consense-Phylip* para todas as combinações de métricas.

O conjunto de 500 espécies apresentou 130 incongruências para união de *JC+F84*, 121 incongruências para união de *JC+KI*, 123 incongruências para união de *JC+KI+F84*, e 117 incongruências para união de *KI+F84*. Observa-se que neste conjunto de espécies os melhores resultados foram da união de métricas de *KI+F84*.

A Figura 6.8 apresenta as incongruências médias normalizadas e as incongruências líquidas (ver Seção 6.3.2). Os níveis de incongruências possuem similaridades com o obtido pelo *CladeNet* usando o *Consense-Phylip* (ver Figura 6.4). Porém, o *CladeNet* usando a *UFR* possui casos com níveis de incongruências zero e, ao mesmo tempo, seu pico de incongruência líquida supera o pico obtido pelo *CladeNet* usando o *Consense-Phylip*.

Todas as incongruências líquidas na Figura 6.8 são limitadas pelo valor de 0,3. Como 0,3 está próximo de 0,2 (limite de incongruências obtido pelo *CladeNet* usando o *Consense-Phylip*, ver Figura 6.4) seria possível à princípio melhorar esses resultados com a inclusão na Rede de mais filogenias de outros métodos como, por exemplo, Evolução Mínima (RZHETSKY; NEI, 1992), Máxima Verossimilhança (FELSENSTEIN, 1981b), Máxima Parcimônia (KISHINO; HASEGAWA, 1989b). Conforme mostrado em (CROCOMO, 2012), quanto mais houver reamostragem de árvores (dezenas de árvores ou até centenas) melhores são as informações contidas na Rede.

Assim como realizado na Seção 6.3.2, houve também a retirada do conjunto RDPII de 218 espécies para os experimentos utilizando o *CladeNet* com a *UFR* produzindo a Figura 6.9 (que mostra os resultados obtidos para a métricas *KI+F84*) e a Figura 6.10 (que mostra as curvas de tendência ajustadas para todas as combinações de métricas).

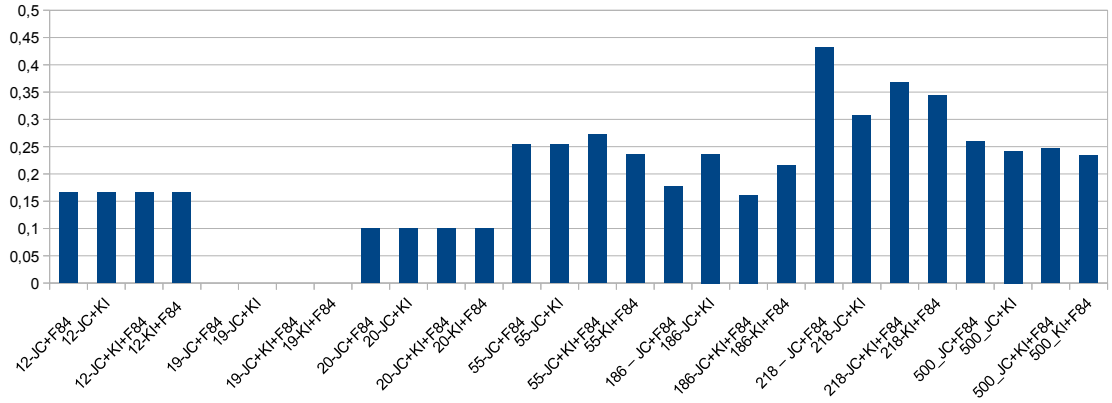


Figura 6.7: Incongruências Médias, usando conjunto de clados obtido pelo *CladeNet* a partir do *UFR*.

As curvas evidenciam novamente que as incongruências líquidas diminuem significativamente conforme o tamanho dos conjuntos aumentam. Esse resultado indica que os clados obtidos pelo *CladeNet* usando a *UFR* também (ver Seção 6.3.2) tendem a aproxima-se significativamente dos clados obtidos pelo *CladeNet* usando *MrBayes*.

6.3.4 Ranqueamento de Dominância

O ranqueamento de dominância (BRASIL, 2012) é um procedimento de comparação de abordagem que busca encontrar a que obtém em geral os melhores resultados para um conjunto de casos de teste. O processo de comparação ocorre, como se ocorressem várias competições. As Tabelas de 6.6 a 6.12 mostram os resultados da competição entre o *CladeNet* usando o *Consense-Phylip* para cada uma das 7 combinações de métricas. Nessas Tabelas, o resultado **'P' ('G', 'E')** indica que o procedimento de uma linha perde (ganha, empata) para (do, com) o procedimento da respectiva coluna **'P' ('G', 'E')** .

Tabela 6.6: Ranqueamento de Dominância para o conjunto de dados de 12 Primatas.

	JC	KI	F84	JC+F84	JC+KI	JC+KI+F84	KI+F84
JC		E	E	E	E	E	E
KI	E		E	E	E	E	E
F84	E	E		E	E	E	E
JC+F84	E	E	E		E	E	E
JC+KI	E	E	E	E		E	E
JC+KI+F84	E	E	E	E	E		E
KI+F84	E	E	E	E	E	E	

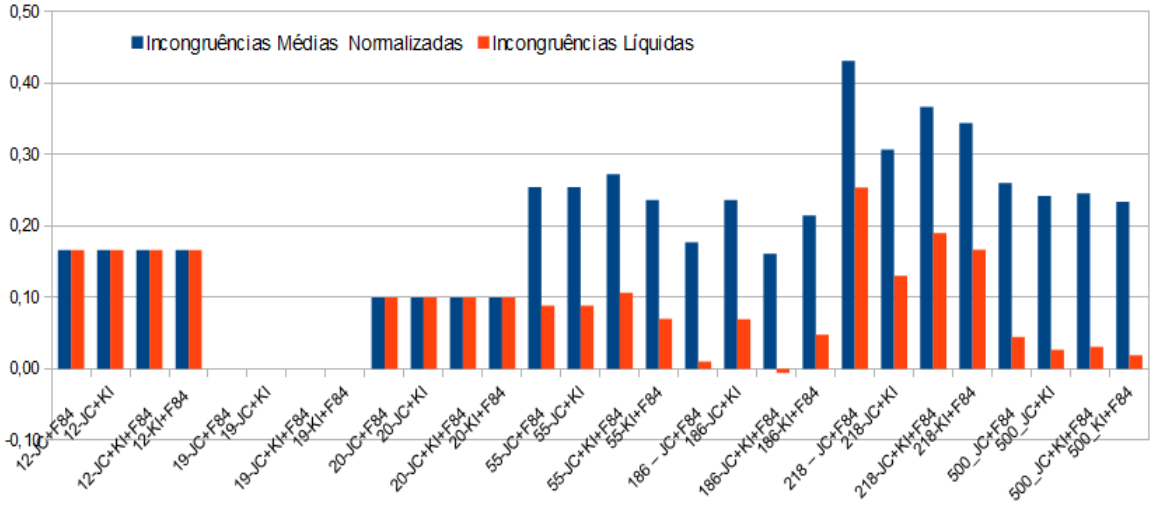


Figura 6.8: Incongruências Médias Normalizadas e Incongruências Líquidas, usando conjunto de clados obtido pelo *CladeNet* a partir do *UFR.*

Tabela 6.7: Ranqueamento de Dominância para o conjunto de dados de 19 aves.

	JC	KI	F84	JC+F84	JC+KI	JC+KI+F84	KI+F84
JC		P	P	E	E	E	E
KI	G		E	G	G	G	G
F84	G	E		G	G	G	G
JC+F84	E	P	P		E	E	E
JC+KI	E	P	P	E		E	E
JC+KI+F84	E	P	P	E	E		E
KI+F84	E	P	P	E	E	E	

Tabela 6.8: Ranqueamento de Dominância para o conjunto de dados de 20 mamíferos.

	JC	KI	F84	JC+F84	JC+KI	JC+KI+F84	KI+F84
JC		E	E	E	E	E	E
KI	E		E	E	E	E	E
F84	E	E		E	E	E	E
JC+F84	E	E	E		E	E	E
JC+KI	E	E	E	E		E	E
JC+KI+F84	E	E	E	E	E		E
KI+F84	E	E	E	E	E	E	

Por fim, a Tabela 6.13 apresenta o ranqueamento de dominância contabilizados os resultados das Tabelas 6.6 a 6.12. A Tabela 6.13 destaca que a combinação de métricas

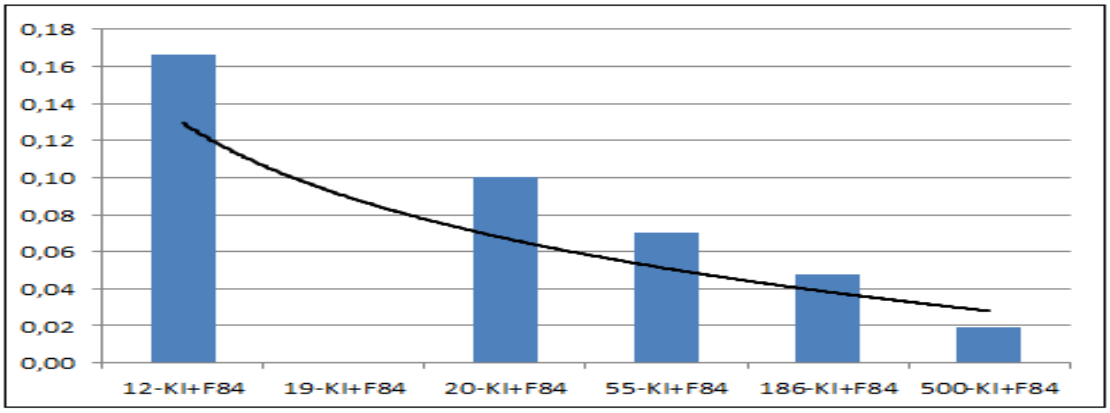


Figura 6.9: Reta de tendência de queda de incongruência usando conjunto de clados obtido pelo *CladeNet* a partir do *UFR* para a combinação *KI+F84*.

Tabela 6.9: Ranqueamento de Dominância para o conjunto de dados 55 rbcL.

	JC	KI	F84	JC+F84	JC+KI	JC+KI+F84	KI+F84
JC		P	P	P	P	P	P
KI	G		P	P	P	P	G
F84	G	G		G	E	E	G
JC+F84	G	G	P		P	P	G
JC+KI	G	G	P	E		P	G
JC+KI+F84	G	G	E	G	G		G
KI+F84	G	G	P	P	P	P	

Tabela 6.10: Ranqueamento de Dominância para o conjunto de dados de 186 mtDNA.

	JC	KI	F84	JC+F84	JC+KI	JC+KI+F84	KI+F84
JC		G	P	G	E	G	G
KI	P		P	G	P	G	P
F84	G	G		G	G	G	G
JC+F84	P	P	P		P	G	P
JC+KI	E	G	P	G		G	G
JC+KI+F84	P	P	P	P	P		P
KI+F84	P	G	P	G	P	G	

KI+F84 é, ao mesmo tempo, a que menos perde e a que mais ganha. Assim, para um conjunto de dados quaisquer, à principio, essa poderia ser a combinação de métricas recomendada.

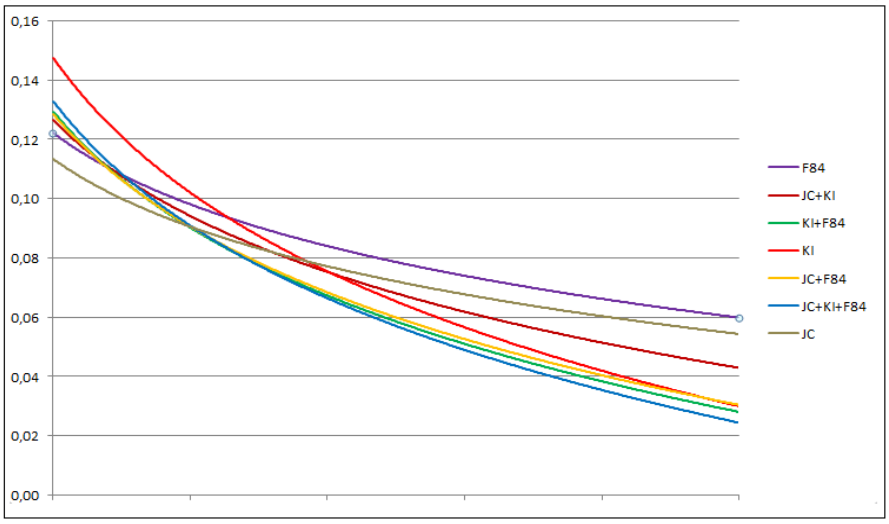


Figura 6.10: Curva de tendência de queda de incongruência usando conjunto de clados obtido pelo *CladeNet* a partir da *UFR* para todas as combinações de métricas.

Tabela 6.11: Ranqueamento de Dominância para o conjunto de dados de 218 RDPII.

	JC	KI	F84	JC+F84	JC+KI	JC+KI+F84	KI+F84
JC		G	G	P	P	P	P
KI	P		G	P	E	P	P
F84	P	P		P	P	P	P
JC+F84	G	G	G		G	G	G
JC+KI	G	E	G	P		P	P
JC+KI+F84	G	G	G	P	G		G
KI+F84	G	G	G	P	G	P	

Tabela 6.12: Ranqueamento de Dominância para o conjunto de dados de 500 ZILLA.

	JC	KI	F84	JC+F84	JC+KI	JC+KI+F84	KI+F84
JC		G	G	G	G	G	G
KI	P		P	P	G	G	G
F84	P	G		E	G	G	G
JC+F84	P	G	E		G	G	G
JC+KI	P	P	P	P		P	G
JC+KI+F84	P	P	P	P	G		G
KI+F84	P	P	P	P	P	P	

A existência de uma combinação de métricas que se destaca em relação às demais (para o *CladeNet* usando o *Consense-Phylip*), indica que o conceito de combinação de filogenias

obtidas por métodos distintos pode ser uma estratégia promissora para determinação de clados. Em outras palavras, esse resultado motiva a extensão do *CladeNet* para lidar com mais árvores, assim como sugerido no trabalho de (CROCOMO, 2012).

Tabela 6.13: Ranqueamento de Dominância considerando todos os conjuntos de dados testados.

	P	E	G	SUBTOTAL	
JC	12	17	13	42	
KI	12	14	16	42	
F84	20	15	7	42	
JC+F84	14	18	10	42	
JC+KI	11	19	12	42	
JC+KI+F84	12	17	13	42	
KI+F84	8	16	18	42	
MAX	20	19	18	294	TOTAL

6.4 Considerações Finais

O *CladeNet*foi utilizado para determinar clados de filogenias obtidas por 3 métodos: *MrBayes, Consense-Phylip* e a *UFR*. Sete conjuntos de dados biológicos foram usados para avaliar o *CladeNet* (ver Seção 6.1), usando a técnica de reamostragem *Bootstrap* de 50% do total de sítios (ver Capítulo 4). Para os métodos baseados em distância, geraram-se filogenias usando as métricas KI *(Kimura)*, JC *(Jukes-Cantor)* e F84, além das combinações das correspondentes filogenias nas seguintes formas: *KI+F84*, *JC+F84*, *JC+KI*, e *JC+KI+F84*.

Para o *CladeNet* utilizando o *Consense-Phylip*, as incongruências líquidas apresentaram valores estáveis, atingindo um patamar de incongruências inferior a 0,2 para todos os conjuntos de dados. Esse mesmo patamar é considerado baixo, pois se encontra próximo às incongruências obtidas para os conjuntos de dados com menos espécies, indicando assim, que este é um método eficaz, uma vez que os níveis de incongruências não tendem as ser muito altos.

Quando o *CladeNet* emprega a *UFR*, as incongruências líquidas apresentam valores limitados a 0,3, valores acima do que os obtidos pelo *Consense-Phylip*. Por outro lado, este resultado poderia ser melhorado incluindo na rede mais filogenias de outros métodos filogenéticos encontrados na literatura.

Outro resultado que merece destaque foi gerado pelas curvas de tendências de decaimento das incongruências líquidas conforme se aumentam os tamanhos dos conjuntos de

dados. Os experimentos com o *CladeNet* usando tanto o *Consense-Phylip* quanto a *UFR* indicam que o nível de incongruência líquida está entre 0,02 a 0,06 para 500 espécies, com tendência de diminuição para conjuntos maiores, por meio de uma estimativa de projeção das curvas das Figuras 6.6 e 6.10.

A análise por ranqueamento de dominância, indica que a combinação de métricas *KI+F84* é a que menos perde e também a que mais ganha. Além disso, as incongruências líquidas para o conjunto maior é de 0,02 quando se utiliza *KI+F84*. Essa combinação obteve os melhores resultados possivelmente por serem métricas que exigem mais parâmetros (mais informações) para geração de filogenias. Por exemplo, a métrica *Kimura* considera que substituições do tipo transições ocorrem mais do que transversões, enquanto a métrica *F84* considera que: *i)* cada base possui uma determinada frequência; *ii)* as taxas de frequência do conteúdo GC afetam o cálculo das distâncias; *iii)* as taxas de transição e transversão são semelhantes. Esses aspectos podem ser indicativos de que ambas as métricas quando combinadas podem se completar, refinando os resultados da filogenia gerada e, consequentemente, tornando os clados obtidos pelo *CladeNet* mais consistentes.

O *CladeNet* usando *Consense-Phylip* e a *UFR* apresentam sensibilidade diferente em relação a técnica de reamostragem. Enquanto o *CladeNet* usando o *Consense-Phylip* obteve resultados significativamente melhores com Bootstrap de 50%, o *CladeNet* usando a *UFR* atingiu resultados similares tanto para o Bootstrap de 50% quanto para o Bootstrap regular (réplicas com todos os sítios dos dados originais).

Por fim, destaca-se a vantagem do método *CladeNet* usando *Consense-Phylip* quanto a *UFR* requerer um tempo computacional significativamente inferior (cerca de 130 vezes mais rápido para as instâncias maiores) ao tempo computacional exigido pelo método pelo *CladeNet* usando *MrBayes* para as mesmas instâncias.

7

Considerações Finais e Trabalhos Futuros

7.1 Considerações Finais

Um aspecto importante para o entendimento da biodiversidade dos seres vivos é a identificação, denominação e classificação das espécies. A Taxonomia facilita a organização dessas espécies, utilizando uma estrutura hierárquica (classe, ordem, gênero e demais grupos) para esse fim.

A qualidade das filogenias implica diretamente na conclusão sobre os clados atribuídos a cada espécie. Entre os métodos filogenéticos mais utilizados, estão os métodos de distância e os métodos de otimalidade, estes têm sido bastante utilizados, para reconstrução de filogenias, possibilitando melhores inferências. Por outro lado, esses métodos exigem um esforço computacional significativo. Enquanto métodos de distância tendem a gerar resultados satisfatórios com eficiência computacional significativamente menor.

O *CladeNet* proposto neste trabalho, busca encontrar clados para conjuntos de dados utilizando tanto um conjunto de filogenias obtidas por método de distância, quanto uma filogenia obtida por método de otimalidade (no caso, a Inferência Bayesiana). O *CladeNet* consegue obter clados significativamente similares tanto usando filogenias de métodos de distância quanto de otimalidade. Naturalmente, a obtenção dos clados usando filogenia de métodos de distância resulta em um método computacionalmente mais eficiente. Dessa forma, o *CladeNet* consegue um compromisso interessante entre qualidade da inferência e tempo de computação, pelo menos em relação à obtenção de clados.

Para validação dos resultados obtidos pelo *CladeNet* utilizaram-se testes de confiança para todos os conjuntos de clados obtidos. Existem vários métodos de confiança para validação de filogenias, cabe ao usuário escolher qual método é mais apropriado para cada tipo de dado e análises a serem realizadas, assim, nem todas as técnicas são apropriadas para todas as situações. Para este trabalho foi utilizada a técnica de *Bootstrap*, que implica na geração de várias réplicas a partir de um mesmo conjunto de dados seguida

da técnica de Consenso por regra majoritária, em que é gerada uma árvore de consenso baseada em todas as réplicas de conjuntos de dados.

O principal objetivo do *CladeNet* é encontrar clados mais confiáveis e robustos através de uma filogenia ou uma *UFR*, ambas constituídas por filogenias reconstruídas através de diversos métodos ou métricas. O *CladeNet* é capaz de detectar clados consistentes utilizando o *Fast Algorithm*, conforme apresentado no Capítulo 6. Outros algoritmos de detecção de comunidades também podem ser empregados ao *CladeNet*. Além disso, o *CladeNet* é flexível em relação às formas (métodos) para se obterem uma diversidade de filogenias (entrada do algoritmo), bem como, é flexível em relação a quantidade de filogenias a serem utilizadas.

O *CladeNet* foi utilizado para determinar clados de filogenias obtidas por 3 métodos: *MrBayes, Consense-Phylip* e a *UFR*. Sete conjuntos de dados biológicos foram usados para avaliar o *CladeNet*, usando a técnica de reamostragem *Bootstrap* de 50% do total de sítios. Para os métodos baseados em distância, geraram-se filogenias usando as métricas KI *(Kimura)*, JC *(Jukes-Cantor)* e F84, além das combinações das correspondentes filogenias nas seguintes formas: *KI+F84*, *JC+F84*, *JC+KI*, e *JC+KI+F84*.

Para o *CladeNet* utilizando o *Consense-Phylip*, as incongruências líquidas apresentaram valores estáveis, atingindo um patamar de incongruências inferior a 0,2 para todos os conjuntos de dados. Esse mesmo patamar é considerado baixo, pois se encontra próximo às incongruências obtidas para os conjuntos de dados com menos espécies, indicando assim, que este é um método eficaz, uma vez que os níveis de incongruências não tendem as ser muito altos.

Quando o *CladeNet* emprega a *UFR*, as incongruências líquidas apresentam valores limitados a 0,3, valores acima do que os obtidos pelo *Consense-Phylip*. Por outro lado, este resultado poderia ser melhorado incluindo na rede mais filogenias de outros métodos filogenéticos encontrados na literatura.

Outro resultado que merece destaque foi gerado pelas curvas de tendências de decaimento das incongruências líquidas conforme se aumentam os tamanhos dos conjuntos de dados. Os experimentos com o *CladeNet* usando tanto o *Consense-Phylip* quanto a *UFR* indicam que o nível de incongruência líquida está entre 0,02 a 0,06 para 500 espécies, com tendência de diminuição para conjuntos maiores.

Para a interpretação de todos os resultados obtidos foi utilizada a técnica de análise por ranqueamento de dominância, que consiste em encontrar a técnica que obtém em geral os melhores resultados quanto comparada a outras. Esse processo ocorre como se fossem realizadas várias competições entre as técnicas usadas. O ranqueamento de distância indica que a combinação de métricas *KI+F84* é a que menos perde e também a que mais ganha.

Além disso, as incongruências líquidas para o conjunto maior é de 0,02 quando se utiliza *KI+F84*. Essa combinação obteve os melhores resultados possivelmente por envolver métricas que exigem mais parâmetros (mais informações) para geração de filogenias que quando combinadas, podem se completar, tornando mais consistentes os clados obtidos pelo *CladeNet*.

O *CladeNet* usando *Consense-Phylip* e a *UFR* apresentam sensibilidade diferente em relação a técnica de reamostragem. Enquanto o *CladeNet* usando o *Consense-Phylip* obteve resultados significativamente melhores com Bootstrap de 50%, o *CladeNet* usando a *UFR* atingiu resultados similares tanto para o Bootstrap de 50% quanto para o Bootstrap regular.

Por fim, destaca-se a vantagem do método *CladeNet* usando *Consense-Phylip* quanto a *UFR* requerer um tempo computacional significativamente inferior (cerca de 130 vezes mais rápido para as instâncias maiores) ao tempo computacional exigido pelo método pelo *CladeNet* usando *MrBayes* para as mesmas instâncias.

Este trabalho propôs e avaliou mais de um método para determinação de clados. O *CladeNet* é um método que, utilizando algoritmo de detecção de comunidade, pode encontrar uma solução para a busca de clados mais consistentes e confiáveis em uma árvore filogenética ou uma *UFR*.

7.2 Trabalhos Futuros

A seguir são apresentadas sugestões para continuação da pesquisa em trabalhos futuros:

- Acrescentar à *UFR* topologias geradas por outros métodos filogenéticos, por exemplo: Método de Parcimônia, Verosimilhança, Evolução Minima;
- Testar o *CladeNet* com conjuntos de dados biológicos montados pelo NCBI (SHERRY et al., 2001);
- Testar o *CladeNet* para conjuntos de dados mitocondriais; com conjuntos biológicos de DNA e conjuntos de RNA, ambos isoladamente e também testar com dados morfológicos, com todos os conjuntos variando de tamanho;
- Avaliar o efeito de conjuntos de referência possivelmente diferentes obtidos usando o *MrBayes* ajustado como modelos evolutivos coerentes com as métricas *Jukes-Cantor, Kimura e F84*;

- Aperfeiçoar a organização dos clados identificados pelo *CladeNet* por nível de ancestralidade das espécies (analisar níveis filogenéticos mais altos para identificar clados);
- Elaborar um comparativo com o algoritmo de Evolução Mínima que use a geração de topologias da Máxima Parcimônia e verificar o quanto essas topologias são melhores ou piores que a Evolução Mínima atual e o NJ;
- Testar o *CladeNet* associado a outros métodos de agrupamento, por exemplo: *Extreme Optimization* e *Adaptive Clustering*;
- Acrescentar métricas da Teoria da Informação (ver Apêndice A) ao *CladeNet* para geração das topologias;
- Realizar experimentos no qual se possa extrair informações de similaridade a partir de dados que não são sequências de nucleotídeos (por exemplo, dados morfológicos, dados fisiológicos como níveis de atividade de sistema imunológico, etc) combinados com dados que são sequências de nucleotídeos.

Referências[1]

AGHAGOLZADEH, M. et al. A hierarchical clustering based on mutual information maximization. In: *IEEE International Conference on Image Processing, 2007. ICIP 2007*. [S.l.: s.n.], 2007. v. 1. 86

AMORIM, D. de S. *Fundamentos de Sistemática Filogenética*. Holos, 2005. ISBN 9788586699368. Disponível em: <http://books.google.com.br/books?id=EYWmPgAACAAJ>. , 16, 24, 25

APORNTEWAN, C. et al. Gene hunting of the genetic analysis workshop 16 rheumatoid arthritis data using rough set theory. In: BIOMED CENTRAL LTD. *BMC proceedings*. [S.l.], 2009. v. 3, p. S126. 86

AVANCINI, M. M.; AGUIAR, R. S. de; VOGT, C. Ao redor de charles robert darwin. 25

BAROSSI-FILHO, M.; ACHCAR, J. A.; SOUZA, R. M. de. Modelos de volatilidade estocÁstica em sÉries financeiras: Uma aplicaÇÃo para o ibovespa. *Economia Aplicada*, v. 14, p. 25–40, 2010. 87

BAUHIN, C. 1623. *Pinax theatri botanici*, p. 310–313. 15

BOETTCHER, S.; PERCUS, A. G. Optimization with extremal dynamics. *Complex.*, John Wiley & Sons, Inc., New York, NY, USA, v. 8, n. 2, p. 57–62, 2002. ISSN 1076-2787. 90

BRASIL, C. R. S. *Algoritmo evolutivo de muitos objetivos para predição ab initio de estrutura de proteínas*. Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2012. 68

BRYANT, D. *A Classification of Consensus Methods for Phylogenetics*. 2003. 45

CANCINO, W.; DELBEM, A. Inferring phylogenies by multi-objective evolutionary algorithms. *international journal of information technology and intelligent computing*, v. 2, p. 1–26, 2007. 55

CANCINO, W. G. T. *Algoritmos evolutivos multi-objetivo para a reconstrucao de arvores filogeneticas*. Tese (Doutorado) — Universidade de Sao Paulo (USP) - Instituto de Ciencias Matematicas e de Computacao (ICMC), 2008. 11, 12, 38, 95

CAO, Y. et al. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *Journal of Molecular Evolution*, Springer, v. 47, n. 3, p. 307–322, 1998. 55

CARLIN, B.; LOUIS, T. *Bayes and Empirical Bayes Methods for Data Analysis, Second Edition*. Taylor & Francis, 2010. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9781584881704. Disponível em: <http://books.google.com.br/books?id=484r1P5o0hQC>. 86

[1]De acordo com a Associação Brasileira de Normas técnicas NBR6023.

CHASE, M. W. et al. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene rbcl. *Annals of the Missouri Botanical Garden*, JSTOR, p. 528–580, 1993. 55

CHATTON, É. P. L. *Les Péridiniens parasites: morphologie, reproduction, ethologie.* [S.l.]: Otto Koeltz Science Publishers, 1975. 16

CILIBRASI, R.; VITáNYI, P. M. B. Clustering by compression. *IEEE Transactions on Information Theory*, v. 51, p. 1523–1545, 2005. 86

COLE, J. R. et al. The ribosomal database project (rdp-ii): sequences and tools for high-throughput rrna analysis. *Nucleic Acids Research*, Oxford Univ Press, v. 33, n. suppl 1, p. D294–D296, 2005. 55

CROCOMO, M. K. *Algoritmo de otimização bayesiano com detecção de comunidades.* Tese (Doutorado) — Universidade de São Paulo, São Carlos, 2012. 49, 67, 72

DARLINGTON, P. A pratical criticism of hennig-brundin "phylogenetic systematics"and antartic biogeography. *Systematic Zoology*, v. 19, p. 1–18, 1970. 24

DARWIN, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favored Races in the Struggle for Life. (First Edition).* [S.l.]: London: John Murray, 1859. ISBN 1-4353-9386-4. , 17, 19, 23, 27

DELBEM, A. C. B.; MELO, V. V.; VARGAS, D. V. Algoritmo filo-genético. In: APDIO. *2ª Escola Luso-Brasileira de Computação Evolutiva (ELBCE).* [S.l.], 2010. 85

DELSUC, B. . P. "phylogenomics and the reconstruction of the tree of life". *Nature Reviews/Genetics.*, v. 6, p. 361–375, 2005. 23

DENNETT, D. C.; RODRIGUES, T. M. *A perigosa idéia de Darwin: a evolução e os significados da vida.* [S.l.]: Rocco, 1998. 17

DIECKERSON, R. The structure of cytochrome c and the rates of molecular evolution. *J.Mol.Evol*, v. 1, p. 26–45, 1971. 30

DIONíSIO, A.; MENEZES, R.; MENDES, D. Entropy-based independence test. *Nonlinear Dynamics*, Springer, v. 44, n. 1, p. 351–357, 2006. 86

DUCH, J.; ARENAS, A. Community detection in complex networks using extremal optimization. *Physical Review E*, American Physical Society, v. 72, n. 2, p. 027104+, ago. 2005. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.72.027104>. 49, 90

DUQUE, T. S. P. C.; GOLDBERG, D. E. Clustermi: Building probabilistic models using hierarchical clustering and mutual information. In: HIOT, L. M.; ONG, Y. S.; CHEN, Y.-p. (Ed.). *Exploitation of Linkage Learning in Evolutionary Algorithms.* [S.l.]: Springer Berlin Heidelberg, 2010, (Adaptation, Learning, and Optimization, v. 3). p. 123–137. ISBN 978-3-642-12834-9. 85

FELSENSTEIN, J. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, Springer, v. 17, n. 6, p. 368–376, 1981. 12, 35

FELSENSTEIN, J. Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, Springer, v. 17, n. 6, p. 368–376, 1981. 67, 96

FELSENSTEIN, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, JSTOR, v. 39, n. 4, p. 783–791, 1985. 30, 43, 47

FELSENSTEIN, J. Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics*, Annual Reviews, v. 22, n. 1, p. 521–565, 1988. 13, 36, 42

FELSENSTEIN, J. *PHYLIP (Phylogeny Inference Package)*. 2000. Disponível em: <http://evolution.genetics.washington.edu/phylip.html>. 38

FELSENSTEIN, J. *Inferring Phylogenies*. 2. ed. Sinauer Associates, 2003. Paperback. ISBN 0878931775. Disponível em: <http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20&path=ASIN/0878931775>. 44, 94

FITCH, W. M. A non-sequential method for constructing trees and hierarchical classifications. *Journal of Molecular Evolution*, Springer New York, v. 18, p. 30–37, 1981. ISSN 0022-2844. 10.1007/BF01733209. Disponível em: <http://dx.doi.org/10.1007/BF01733209>. 35

GASCUEL, O. *Mathematics of Evolution and Phylogeny*. New York, NY, USA: Oxford University Press, Inc., 2007. ISBN 0199231346, 9780199231348. Disponível em: <http://portal.acm.org/citation.cfm?id=1557209>. 38

GENES, O. O gene ea célula 67. *Introdução à genética médica*, Fundo Editorial Procienx, p. 67, 1968. 22

GOLOBOFF, P. *NONA*. 1999. Disponível em: <http://www.cladistics.com/aboutNona.htm>. 38

GOLOBOFF, P.; FARRIS, J. S.; NIXON, K. *TNT*. 2004. Disponível em: <http://www.cladistics.com/aboutTNT.html>. 38

GOLOBOFF, P. A.; FARRIS, J. S. Methods for quick consensus estimation. *Cladistics*, Wiley Online Library, v. 17, n. 1, p. S26–S34, 2001. 55

GREUTER, W.; MCNEILL, J. *Synopsis of Proposals on Botanical Nomenclature. Tokyo 1993. A Review of the Proposals concerning the International Code of Botanical Nomenclature Submitted to the XV International Botanical Congress*. [S.l.]: International Association for Plant Taxonomy (IAPT), 1993. 191–271 p. 16

GRIFFITHS, A. *Introdução à genética*. Guanabara Koogan, 1998. ISBN 9788527704632. Disponível em: <http://books.google.com.br/books?id=NmhEAQAACAAJ>. 22

GUINDON, S.; GASCUEL, O. *PHYML Phylogenetic Maximum-Likelihood*. 2003. Disponível em: <http://atgc.lirmm.fr/phyml/>. 38, 55

GUINDON, S.; GASCUEL, O. A simple, fast and accurate algorithm to estimate large phylogenies by maximum-likelihood. *Systematic Biology*, v. 5, p. 696–704, 2003. 55

HAYASAKA, K. T.; GOJOBORI; HORAI. Molecular phylogeny and evolution of primate mitochondrial DNA. *Molecular Biology and Evolution*, v. 5, p. 626–644, 1988. 54

HENNIG, W. *Phylogenetic systematics*. [S.l.]: University of Illinois Press, 1999. 25

HO, M.; SAUNDERS, P. *Beyond neo-Darwinism: An introduction to the new evolutionary paradigm [Más allá del neodarwinismo: Introducción al nuevo paradigma evolutivo]*. [S.l.]: Orlando, Florida: Academic Press, 1984. 19

INGMAN, M.; GYLLENSTEN, U. mtdb: Human mitochondrial genome database, a resource for population genetics and medical sciences. *Nucleic acids research*, Oxford Univ Press, v. 34, n. suppl 1, p. D749–D751, 2006. 55

JUKES T.H E CANTOR, C. Evolution of protein molecules. *Mammalian protein metabolism*, v. 3, p. 21–132, 1969. 29, 47

KERAMIDAS, E.; KAUFMAN, S. *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, Seattle, Washington, April 21-24, 1991*. Interface Foundation of North America, 1992. Disponível em: <http://books.google.com.br/books?id=Bfc1NAAACAAJ>. 40

KIMURA, M. A simple method for estimating evolutionary rate of base substituitions through comparative studies os nucleotide sequences. *J.Mol.Evol*, v. 16, p. 111–120, 1980. 30, 47

KISHINO, H.; HASEGAWA, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *Journal of Molecular Evolution*, Springer New York, v. 29, p. 170–179, 1989. ISSN 0022-2844. 10.1007/BF02100115. Disponível em: <http://dx.doi.org/10.1007/BF02100115>. 47

KISHINO, H.; HASEGAWA, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *Journal of molecular evolution*, Springer, v. 29, n. 2, p. 170–179, 1989. 67

KRASKOV, A. Synchronization and interdependence measures and their application to the electroencephalogram of epilepsy patients and clustering of data. *Report Nr.: NIC series; 24*, 2008. 86

KRASKOV, A. et al. Hierarchical clustering based on mutual information. *Arxiv preprint q-bio/0311039*, 2003. 85

KUMAR, S.; TAMURA, K.; NEI, M. Mega3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in bioinformatics*, Oxford Univ Press, v. 5, n. 2, p. 150–163, 2004. 28

LAMARCK, J.-B. d. M. *Histoire naturelle des animaux sans vertebres,...* [S.l.]: Meline, Cans et Compagnie, 1839. 17, 27

LEISERSON, C. E. et al. *Introduction to algorithms*. [S.l.]: The MIT press, 2001. 36

LEMMON, A. R.; MILINKOVITCH, M. C. The metapopulation genetic algorithm: an efficient solution for the problem of large phylogeny estimation. *Proceedings of the National Academy of Sciences*, p. p. 10516 a 10521, 2002. 55

LEWIS, P. O. A genetic algorithm for maximum-likelihood phylogeny inference using nucleotide sequence data. *Molecular Biology and Evolution*, v. 15, n. 3, p. 277–283, 1998. Disponível em: <http://mbe.oxfordjournals.org/content/15/3/277.abstract>. 55

LINNAEUS, C. *Species plantarum*. [S.l.]: Impensis GC Nauk, 1800. 16

MATIOLI, S. R. *Biologia Molecular e Evolução*. [S.l.: s.n.], 2001. 24, 39

MAZZAROLO, L. A. *Conceitos básicos de Sistemática Filogenética*. [S.l.: s.n.], 2005. 17

MENDEL, G. Versuche über pflanzen-hybriden. *Verhandlungen des naturforschenden Vereines*, v. 4, p. 3–47., 1866. 19, 27

METROPOLIS, N. et al. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical Physics*, AIP, v. 21, n. 6, p. 1087–1092, 1953. ISSN 00219606. Disponível em: <http://dx.doi.org/10.1063/1.1699114>. 40

NELSON, D. L.; COX, M. M.; CUCHILLO, C. *Lehninger princípios de bioquímica*. [S.l.]: Omega, 2001. 20, 22

NEWMAN, M. Fast algorithm for detecting community structure in networks. *PHYS.REV.E*, v. 69, p. 066133, 2004. Disponível em: <doi:10.1103/PhysRevE.69.066133>. 13, 48, 49, 50, 88

NUSSBAUM, R. *THOMPSON & THOMPSON GENÉTICA MÉDICA*. Elsevier Health Sciences Brazil, 2008. ISBN 9788535245752. Disponível em: <http://books.google.com.br/books?id=vjOET7ul_R0C>. 20, 22

PEREIRA, S. L.; BAKER, A. J. A molecular timescale for galliform birds accounting for uncertainty in time estimates and heterogeneity of rates of dna substitutions across lineages and sites. *Molecular phylogenetics and evolution*, Elsevier, v. 38, n. 2, p. 499–509, 2006. 55

PEREZ, E. A. DARWIN O EL FALSO CONFLICTO ENTRE LA TEORÍA DE LA SELECCIÓN NATURAL Y LA HIPÓTESIS DE LA PANGÉNESIS. *Acta biol.Colomb*, scieloco, v. 14, p. 63 – 76, 12 2009. ISSN 0120-548X. Disponível em: <http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-548X2009000400003&nrm=iso>. 18

PERRETTO, M. Aplicação do algoritmo de otimização por colônia de formigas aos problemas de reconstrução de árvores filogenéticas e dobramento de proteínas. Curitiba, 2010. 55

PERRETTO, M.; LOPES, H. S. Reconstruction of phylogenetic trees using the ant colony optimization paradigm. *Genetics and Molecular Research*, v. 4, n. 3, p. 581–589, 2005. 55

RA, P. et al. Nt236 teoria e prática da sistemática filogenética (cladística). 24

RAVEN, P.; EVERT, R.; EICHORN, S. Biologia vegetal. 6a. edição. *Rio de Janeiro, Editora*, 2001. 16, 17

RISSANEN, J. *Information and complexity in statistical modeling*. [S.l.]: Springer Publishing Company, Incorporated, 2007. 29

ROESE, J. *Application of adaptive clustering techniques in multivariate normal data*. Information and Computer Science)–University of California, San Diego, 1969. Disponível em: <http://books.google.com.br/books?id=RS9GAQAAIAAJ>. 49, 88

ROKAS, A. et al. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, Nature Publishing Group, v. 425, n. 6960, p. 798–804, 2003. 13, 54

RONQUIST, F.; HUELSENBECK, J. P. Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, Oxford Univ Press, v. 19, n. 12, p. 1572–1574, 2003. 13, 60

RZHETSKY, A.; NEI, M. A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol*, v. 9, n. 5, p. 945–967, 1992. 67

RZHETSKY, A.; NEI, M. Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, SMBE, v. 10, n. 5, p. 1073–1095, 1993. 39

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, Center for Demographic and Population Genetics, University of Texas Health Science Center, Houston 77225., v. 4, n. 4, p. 406–425, July 1987. ISSN 0737-4038. Disponível em: <http://mbe.oxfordjournals.org/content/4/4/406.abstract>. 31, 32

SANTOS, C. M. D.; CALOR, A. R. Ensino de biologia evolutiva utilizando a estrutura conceitual da sistemática filogenética-ii. *Ciência & Ensino (ISSN 1980-8631)*, v. 2, n. 1, 2008. 24

SCHOR, N.; BOIM, M.; SANTOS, O. dos. *Medicina celular e molecular: bases moleculares da biologia, da genética e da farmacologia*. Atheneu, 2003. (Clínica médica). ISBN 9788573796377. Disponível em: <http://books.google.com.br/books?id=n1hmAAAACAAJ>. 23

SCHWARZ, G. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Disponível em: <http://scholar.google.com/scholar.bib?q=info:sskrpr5jlLwJ:scholar.google.com/&output=citation&hl=en&as_sdt=2000&ct=citation&cd=0>. 86

SHERRY, S. T. et al. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, Oxford Univ Press, v. 29, n. 1, p. 308–311, 2001. 76

SOKAL, R.; MICHENER, C. A statistical method for evaluating systematic relationships. *Multivariate Statistical Methods, Among-groups Covariation*, John Wiley & Sons Inc, p. 269, 1975. 31

SPIEGELHALTER, D. J. et al. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Wiley Online Library, v. 64, n. 4, p. 583–639, 2002. 87

STAMATAKIS, A. *RAxML: Randomized Axelerated Maximum Likelihood*. 2005. Disponível em: <http://www.phylo.org/news/RAxML>. 38, 55

STEARNS, S. *Evolução: uma introdução*. Atheneu, 2003. ISBN 9788574540771. Disponível em: <http://books.google.com.br/books?id=MURIAAAACAAJ>. 24, 26

STUDIER, J.; KEPPLER, K. A note on the neighbor-joining algorithm of saitou and nei. *Molecular Biology and Evolution*, SMBE, v. 5, n. 6, p. 729, 1988. 31

SWOFFORD, D. L. *PAUP* Phylogenetic Analysis Using Parsimony*. 2000. Disponível em: <http://paup.csit.fsu.edu/>. 38

SWOFFORD, D. L. et al. Phylogenetic inference. *Molecular Systematics*, Sinauer, p. pp 407–514, 1996. 37

VARGAS, D. V.; DELBEM, A. C. B.; MELO, V. V. Algoritmo filo-genético. In: APDIO. *2ª Escola Luso-Brasileira de Computação Evolutiva (ELBCE)*. [S.l.], 2010. 85

VINH, L. S.; HAESELER, A. von. IQPNNI: Moving fast through tree space and stopping in time. *Molecular Biology and Evolution*, v. 21, p. 1565–1571, 2004. 55

VITTORI, K.; DELBEM, A. C.; PEREIRA, S. L. Ant-based phylogenetic reconstruction (abpr): A new distance algorithm for phylogenetic estimation based on ant colony optimization. *Genetics and Molecular Biology*, SciELO Brasil, v. 31, n. 4, p. 974–981, 2008. 54, 55

WATSON, J.; BERRY, A. *DNA: o segredo da vida*. Companhia das Letras, 2005. ISBN 9788535907162. Disponível em: <http://books.google.com.br/books?id=2cbfuq2RZyEC>. 22

WATSON, J. D.; CRICK, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, v. 171, n. 4356, p. 737–738, abr. 1953. ISSN 0028-0836. Disponível em: <http://www.ebi.ac.uk/citexplore/citationDetails.do?externalId=13054692&dataSource=MED>. , 20, 21

WEIR, B. *Genetic Data Analysis II*. [S.l.]: Sinauer Associates, 1996. 36

WEISMANN, A. *The germ-plasm: a theory of heredity*. [S.l.]: C. Scribner's sons, 1893. 18

WHITTAKER, R. H. Evolution and measurement of species diversity. *Taxon*, JSTOR, p. 213–251, 1972. 17

WOESE, C. R.; FOX, G. E. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 74, n. 11, p. 5088–5090, 1977. 17

YANG, Z. *PAML A Program Package for Phylogenetic Analysis by Maximum-Likelihood*. 1997. Disponível em: <http://abacus.gene.ucl.ac.uk/software/paml.html>. 38

YANG, Z. *Computational Molecular Evolution*. Oxford University Press, USA, 2006. Disponível em: <http://www.amazon.com/Computational-Molecular-Evolution-Oxford-Ecology/dp/0198567022/ref=pd_sim_b_title_5/002-8056378-2973632>. 42

YE, Z.; HU, S.; YU, J. Adaptive clustering algorithm for community detection in complex networks. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, APS, v. 78, n. 4, p. 046115+, 2008. Disponível em: <http://dx.doi.org/10.1103/PhysRevE.78.046115>. 89, 91

ZHARKIKH, A.; LI, W.-H. Estimation of confidence in phylogeny: the complete-and-partial bootstrap technique. *Molecular phylogenetics and evolution*, Elsevier, v. 4, n. 1, p. 44–63, 1995. 42

A
Apêndice

Este Apêndice apresenta outras métricas de distância que têm sido bastantes utilizadas na literatura da área de Teoria da Informação, Exatas e Engenharia, entre outras. O uso de métricas distintas pode produzir filogenias diferentes ou reforçar relações de clados encontrados usando métricas usuais em filogenia. Tais métricas poderiam também ser empregadas para gerar matriz de distância utilizando algoritmos de clusterização. Nesse contexto, a *UFR* possibilita que as relações presentes em filogenias geradas utilizando diversas métricas sejam identificadas como clados relevantes pelo *CladeNet*.

A.1 Métricas de Distância da Teoria de Informação

Informação Mutua (MI)

Informação Mútua (MI, *Mutual Information*), como descrito na Teoria da Informação (KRASKOV et al., 2003), é uma medida de similaridade entre duas variáveis que tem alguns casos de sucesso com Algoritmo de Estimação de Distribuição - EDAs (VARGAS; DELBEM; MELO, 2010; DELBEM; MELO; VARGAS, 2010; DUQUE; GOLDBERG, 2010). Essa medida pode ser calculada pela Equação A.1.

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p_{xy}(X,Y) \log_2 \left(\frac{p_{xy}(X,Y)}{p_x(X) p_y(Y)} \right), \tag{A.1}$$

em que X e Y são variáveis aleatórias, $p_{xy}(X,Y)$ é a Probabilidade Conjunta para $X = x$ e $Y = y$; enquanto que $p_x(X)$ e $p_y(Y)$ são as probabilidades marginais para $X = x$ e $Y = y$, respectivamente.

A probabilidade marginal $p_x(X = x)$ é definida como a soma de ocorrências de valor x para a variável X dividida pelo número de amostras sobre X. A probabilidade conjunta é calculada de forma semelhante, considerando que a contagem de ocorrências é feita em relação ao par ($X = x$ e $Y = y$).

Entretanto a MI satisfaz essa probabilidade de desigualdade triangular estrita. Por outro lado, $D(X, Y)$ como descrito na Equação A.2 é uma métrica com significativa importância na literatura (KRASKOV, 2008; DIONíSIO; MENEZES; MENDES, 2006; AGHAGOLZADEH et al., 2007; APORNTEWAN et al., 2009). Além da desigualdade triangular satisfaz todas as características fundamentais de métrica: desigualdade triangular; não-negatividade; propriedades simétricas e indiscernibilidade.

$$D(X, Y) = H(X, Y) - I(X, Y), \tag{A.2}$$

em que $H(X, Y)$ pode ser a entropia do par (X, Y) calculado pela Equação A.3.

$$H(X, Y) = -\sum_{x \in X} \sum_{y \in Y} p_{xy}(X, Y) \log_2(p_{xy}(X, Y)), \tag{A.3}$$

A métrica $D(X, Y)$ pode ser chamada de métrica universal, no sentido que se alguma medida de distância não trivial as variáveis X e Y também estabelece como relativamente próximas, então a $D(x, y)$ também apresentará X e Y como próximas. Em geral o seu custo computacional para X e Y é relativamente grande.

Distância Normalizada de Compressão (NCD)

A *Distância Normalizada de Compressão (NCD)* (CILIBRASI; VITáNYI, 2005) é uma métrica universal de similaridade que é computada a partir da compressão dos dados. Como a NCD pode ser aplicada a qualquer tipo de dado, desde que haja algoritmo capaz de comprimir as amostras de conjunto de dados. A métrica é definida da seguinte forma, supondo um compressor $C(X)$, que retorne o tamanho comprimido das n amostras da variável X e $C(XY)$ retorne o tamanho comprimido da concatenação nas n amostras das variáveis X e Y.

A NCD já foi aplicada a diversos contextos de aprendizado de máquina (CILIBRASI; VITáNYI, 2005). Esse aspecto da *NCD* é internamente para construção de filogenias a partir do *CladeNet*, uma vez que possibilitaria obter *UFRs* de filogenias de conjuntos de dados de sequencias de nucleotídeos, sequencias de aminoácidos e dados morfológicos de uma mesma espécie usando a mesma métrica.

Critério de Informação Bayesiana

O *Bayesian Information Criterion (BIC)* é um critério de seleção de modelo proposto por Schwarz (1978) e modificado por Carlin e Louis (2010) para ser aplicado utilizando a densidade a posteriori dos parâmetros do modelo ajustado. Este critério pondera entre

a função log-verossimilhança maximizada e o logaritmo do número de parâmetros do modelo. Assim, o melhor modelo é aquele que apresenta o maior valor do BIC, dado por:

$$BIC = E\left[\ln L(\theta)\right] - \frac{1}{2}p\ln(N), \tag{A.4}$$

em que $E\left[\ln L(\theta)\right]$ é o valor esperado tomado com relação a densidade a posteriori, da função log-verossimilhança, p é a dimensão do vetor de parâmetros e N é o tamanho amostral.

Critério Desvio-Informação

O *Deviance Information Criterion (DIC)* pode ser vista como uma extensão do *BIC*. Este critério é adequado para problemas Bayesianos de seleção de modelos em que se constrói amostras da distribuição a posteriori dos parâmetros dos modelos, em geral obtidas por simulação de Monte Carlos em Cadeias de Markov *(MCMC).* Semelhante ao *BIC*, este critério possui aproximação assintótica como aumento de número de amostras e requer a distribuição a posteriori adequada para aproximação de uma distribuição normal multivariada. Define-se o desvio como:

$$D(\theta) = -2\ln L(\theta) + C \tag{A.5}$$

em que (θ) é o valor de parâmetros desconhecidos do modelo; $L(\theta)$ a função de verossimilhança e C é uma constante que se cancela e, portanto não precisa ser conhecida na comparação de dois modelos. O critério *DIC* introduzido por Spiegelhalter et al. (2002) é dado por:

$$DIC = D(\hat{\theta}) + 2p_D \tag{A.6}$$

sendo $D(\hat{\theta})$ o desvio avaliado na média a posteriori e p_D é o número efetivo de parâmetros no modelo. Valores menores para *DIC* indicam melhores modelos, podendo estes serem negativos. O *DIC* é usado para a discriminação de diferentes modelos propostos para a análise de séries financeiras. (BAROSSI-FILHO; ACHCAR; SOUZA, 2010)

B

Apêndice

Adaptive Clustering

O algoritmo Adaptive Clustering (adClust) Roese (1969) muda cada vértice para um módulo (comunidade candidata) vizinho que exerça uma maior atração do que seu módulo atual. Esse procedimento é realizado até que um equilíbrio das mudanças seja atingido. Esse equilíbrio é verificado por meio da medida de modularidade Q proposta em Newman (2004). As forças de atração para cada um dos vértices de uma rede são definidas pelas Equações B.1 e B.2:

$$F_{in}^{(s)} = e_{in}^{(s)} - \frac{k(d_{in}^{(s)} - k)}{2E} \tag{B.1}$$

$$F_{out}^{(t)} = \max_{t \in \{vizinhos\}} \left\{ e_{out}^{(t)} - \frac{k d_{out}^{(t)}}{2E} \right\} \tag{B.2}$$

em que:

- $F_{in}^{(s)}$ é a força de atração para o módulo s em que o vértice está;
- $F_{out}^{(t)}$ é a força de atração sobre o vértice exercida pelo módulo vizinho t;
- $e_{in}^{(s)}$ é o número de arestas que conectam o vértice a seu módulo atual s;
- $e_{out}^{(t)}$ é o número de arestas que conectam o vértice ao módulo vizinho t;
- $d_{in}^{(s)}$ é o grau total dos vértices dentro do módulo s;
- $d_{out}^{(t)}$ é o grau total dos vértices dentro do módulo t;
- k é o grau do vértice;
- E é o número total de arestas do grafo;

- $\{vizinhos\}$ representa o conjunto de módulos com pelo menos uma aresta conectada ao vértice, com exceção do módulo ao qual o vértice pertence.

Em outras palavras, as Equações B.1 e B.2 calcula as diferenças entre o número real de arestas que ligam o vértice a uma comunidade e o número esperado de arestas ligadas à comunidade que um vértice de mesmo grau teria. Com isso estima a probabilidade do vertice pertencer a comunidade em questão. Se F_{out} for maior do que F_{in}, o vértice muda para a comunidade que exerce maior atração.

O Algoritmo 2 descreve o pseudocódigo do AdClust:

Algorithm 2: *Adaptive Clustering.*

1. Separe os vértices em n módulos;
2. Calcule F_{in} e F_{out} para cada vértice. 3. Movimente os vértices para o módulo com maior força de atração;
4. Repita os passos 2 e 3 até que todos os vértices não mudem de módulo. Salve a partição encontrada e o valor de Q para a mesma;
5. Calcule o aumento de Q para todos os possíveis pares de módulo encontrados no passo 4;
6. Escolhe o par com o maior Q e una os respectivos módulos;
7. Repita Passos 2 a 5 até que todos os vértices formem um único módulo;
8. Escolha a partição com o maior Q encontrada durante o algoritmo;

É importante destacar que no Passo 2 do Algoritmo 2, as forças são calculadas para todos os vértices da rede e, somente após isso, os vértices são deslocados para o módulo de maior atração. Em comparação com o FA, esse método introduz uma rotina adicional (representada pelos passos 2 e 4 do Algoritmo 2) de ordem $O(n)$. No artigo em que a técnica é apresentada Ye, Hu e Yu (2008), os autores afirmam que as estruturas de comunidade detectadas pelo AdClust, em geral apresentam melhor modularidade Q do que as estruturas encontradas pelo *FA*.

Extreme Optimization

O algoritmo Otimização Extrema (EO, do inglês *Extreme Optimization*) busca determinar melhores estruturas de comunidades. Essa técnica baseia-se em uma medida que verifica a contribuição de cada vértice para o valor de modularidade total. Tal contribuição é dado pela Equação B.3.

$$q_i = k_{r(i)} - k_i a_{r(i)}, \tag{B.3}$$

em que:

- r_i é a i-ésima comunidade;
- q_i é a contribuição do vértice i para a modularidade Q;
- $k_{r(i)}$ é o número de arestas que o vértice i possui para vértices da mesma comunidade;
- k_i é o grau do vértice i;
- $a_{r(i)}$ é a fração das arestas com pelo menos um vértice pertencente a comunidade $r(i)$.

$$Q = \frac{1}{2L} \sum_i q_i \tag{B.4}$$

A modularidade total pode ser apresentada a partir da Equação B.3 como um somatório das contribuições individuais de cada vértice i, conforme a Equação B.4 mostra.

A contribuição relativa λi de cada vértice i para o valor de modularidade total Q, q_i é dividido pelo grau do vértice i, conforme Equação B.5.

$$\lambda_i = \frac{q_i}{k_i} = \frac{k_{r(i)}}{k_i} - a_{r(i)}. \tag{B.5}$$

A busca heurística proposta por Duch e Arenas (2005) para encontrar uma estrutura de comunidades significativamente representativa conforme o Algoritmo 3 mostra.

Algorithm 3: *EO.*

1. Divida os vértices em duas partições aleatórias;

`// cada grupo de vértices conexos em uma partição representa uma comunidade`

2. Enquanto Q estiver sendo melhorado, mude o vértice λ para o outro grupo;
3. Salve Q para a partição atual;
4. Remova as arestas entre todas as comunidades;
5. **se** Q *estiver sendo melhorado* **então** repita recursivamente o processo para cada grupo conectado;

;

Os desenvolvedores do EO avaliam também a substituição da seleção do vértice com menor λ, (ver Passo 2 do Algoritmo 3) por uma probabilidade para se selecionar cada vértice Boettcher e Percus (2002). Essa probabilidade é dada pela Equação B.6:

$$P(q) \alpha q^{-\tau}, \tag{B.6}$$

em que:

- q é o *ranque* do vértice;
- $\tau = 1 + \frac{1}{ln(n)}$;
- n é o número de vértices da rede.

Os autores afirmam que essa técnica encontra os melhores valores de modularidade para problemas de *benchmark* testados até a data de publicação do artigo (2005). No entanto, Ye, Hu e Yu (2008) mostra que o adClust apresenta melhores valores de modularidade para alguns desses mesmos problemas.

C
Apêndice

Parâmetros do Programa *MrBayes*

O programa *MrBayes* usa o formato padrão Nexus para descrição dos dados de entrada. Este trabalho utilizou a versão 3.2, disponível em http://http://mrbayes.sourceforge.net/. A seguir é descrito os principais parâmetros usados do MrBayes.

Comando *Set*

São disponibilizados três opções de *Set* que são responsáveis por definir os aspectos das partições, onde podem ocorrer substituições.

- *set autoclose* (yes/no)
- *set partition* (número ou nome)
- *set warnings* (yes/no)

Comando *Lset*

Esse comando configura os parâmetros de verossimilhança. As opções principais são:

- *Lset applyto* esta opção permite que configurações diferentes sejam aplicadas a diferentes partições;
- *Lset applyto* esta opção permite que configurações diferentes sejam aplicadas a diferentes partições;
- *Lset Nucmodel* este especifica a forma geral do modelo de substituição nucleotídica. O modelo padrão usa 4 nucleotídeos;
- *Lset Nst* configura o número de substituição:
  - (1) todas as taxas são iguais;

  - (2) taxas para transcrição e transversão;
  - (6) todas as taxas podem ser diferentes;

- *Lset rates* configura o modelo para variação de taxas de substituição de sítios. A opção é: Lse rates=equall gammal "invgamma"

Comandos de Configuração das Cadeias Markovianas

Comando MCMCP

Esse comando configura os parâmetros para as cadeias MCMC sem iniciar a análise. As opções de comandos usados no MCMC são os seguintes:

- *Seed* semente para o gerador de números randômicos, se a mesma semente for usada a mesma análise pode ser repetida;
- *Ngen* essa opção estabelece o número de gerações para o algoritmo MCMC. É recomendado utilizar um número maior que um milhão;
- *Samplefreq* especifica a frequência de amostragem das cadeias;
- *Swapfreq* especifica as frequências de tentativas de mudança de estado da cadeias;
- *Printfreq* especifica a frequência com que as informações sobre a cadeia é mostrada em tela;
- *Nchains* especifica quantas cadeias estão rodando para a variante MCMC;
- *Temp* especifica o parâmetro para aquecimento das cadeias;
- *RelBurnin* estabelece o número de cadeias que serão descartadas;
- *"Conformat"* obtém a árvore consenso como saída, pois, a análise duas árvores como saída, esta opção pode determinar á árvore consenso entre as duas;
- *Diagnfreq* mostra a frequência obtida entre as espécies;

Alguns conjuntos de dados variam pela quantidade de espécie e quantidade de genes contidos no grupo, assim, podem ter maior importância conforme o conjunto utilizado.

$\mathcal{D}$

Apêndice

Cálculo de verossimilhança para uma árvore

A verossimilhança fornece probabilidade $P(D|\theta)$ de o conjunto de dados D ajustar-se ao modelo θ=(τ,B,M), em que τ é uma topologia da árvore, B é o conjunto de comprimento dos ramos de τ e M é o modelo de substituição de sequências. O objetivo do critério de máxima verossimilhança é encontrar os parâmetros do modelo θ, tal que a função de verossimilhança definida como $L(\theta)$=$P(D|\theta)$ seja maximizada.

A estimação de verossimilhança é ilustrada a seguir empregando um exemplo. Seja D um conjunto de dados correspondentes às sequências de três espécies (u, w e s). Cada sequência possui N_s sítios (colunas) tal que u_j, w_j, s_j representam os estados da espécies u, w e s no sítio j, respectivamente. Tais estados estão definidos em um alfabeto de caracteres de DNA Ω=(A, G, C, T). A Figura D.1 mostra uma árvore com as três espécies atuais (u, w e s), duas espécies ancestrais (v e r) e os respectivos comprimentos de ramo.

Além disso, supõe-se o uso de um modelo de substituição de sequências que possibilite o cálculo computacional das probabilidades de transição de estados. O cálculo de verossimilhança precisa de duas premissas (FELSENSTEIN, 2003):

- Os sítios das sequências evoluem de forma idêntica e independente;
- A ramificação da árvore é um processo de Markov, ou seja, a probabilidade de um nó possuir um determinado estado é função apenas do estado anterior.

A primeira suposição possibilita decompor a verossimilhança em que produto conforme a Equação D.1.

$$L = \prod_{j=1}^{N_s} P(D^{(j)}|\theta), \tag{D.1}$$

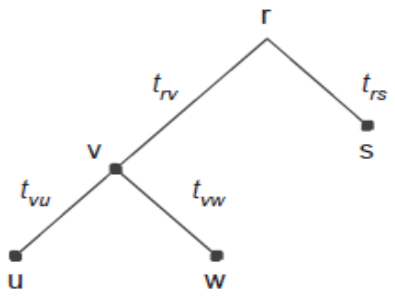


Figura D.1: Exemplo de topologia de árvore com comprimentos de ramo (CANCINO, 2008).

em que, $P(D^{(j)}|\theta)$ representa a verossimilhança no sítio j, também denotada de agora em diante por L_j. Essa quantidade é igual a soma de probabilidades de cada cenário possível considerando todos os possíveis estados dos nós internos (que são desconhecidos) para o sítio j. A independência da ramificação da árvore permite que L_j seja expressada pela Equação D.2 para a topologia da Figura D.1.

$$L_j = \sum_{r_j \epsilon \Omega} \sum_{vj \epsilon \Omega} \pi_{rj} P_{r_j,s_j}(t_{rs}) P_{r_j,v_j}(t_{rv}) P_{v_j,u_j}(t_{vu}) P_{v_j,w_j}(t_{vw}) \tag{D.2}$$

em que r_j, v_j representa os possíveis estados para os nós internos r e v, t_{ij} é o comprimento do ramo que conecta os nós i e j, π_{r_j} é a frequência do nucleótido correspondente ao estado r_j no conjunto de sequências D e $P_{x,y}(t)$ é a probabilidade da mudança de estado x para o estado y para um comprimento, isto é, após um tempo t, $Px,y(t)$ e $\Pi_r j$ são fornecidas pelo modelo de substituição M.

O cálculo de verossimilhança pode ser efetuado recursivamente empregando as verossimilhanças condicionais de subárvores. A verossimilhança condicional da subárvore cuja raiz é o nó r, denotada como, $L_j^r(r_j)$, é a probabilidade dos eventos observados a partir da tal subárvore, dado que o estado do nó r raiz da subárvore no sítio j seja r_j. Assim, se o nó r tem descendentes v e s, temos que:

$$L_j^r(r_j) = \left[\sum_{v_j \epsilon \Omega} P_{rf,vj}(t_{rv}) L_j^v(v_j) \right] X \left[\sum_{s_j \epsilon \Omega} P_{r_j s_j}(t_{rs}) L_j^s(s_j) \right], \tag{D.3}$$

Para as folhas a, em que o estado a_j é fornecido por D, tem se:

$$g(x,y) = \begin{cases} 1, & se \quad a_j = x \\ 0, & caso \quad contrário. \end{cases} \tag{D.4}$$

Para o exemplo relativo à árvore mostrada na Figura D.1, tem-se pela Equação D.4 que $L_j^v(v_j) = P_{v_j,u_j}(t_{vu})P_{v_j,w_j}(t_{vw})$ e $L_j^s(s_j) = P_{r_j s_j}(t_{rs})$. Dessa forma, pode-se escrever:

$$L_j = \sum_{r_j \epsilon \Omega} \pi_{r_j} L_j^r(r_j). \tag{D.5}$$

Para calcular a verossimilhança total conforme a Equação D.1 é necessário fazer o produtos dos valores L_j para todos os sítios. Dado que tais valores são números muito pequenos podem ocorrer erros de precisão numérica. Aplicando logaritmo natural a ambos lados da Equação D.1, tem-se obtendo

$$lnL = \sum_{j=1}^{N_{sit}} lnL_j. \tag{D.6}$$

As Equações D.5 e D.3 definem uma forma recursiva de calcular a verossimilhança para árvores filogenéticas, em qual as verossimilhanças condicionais de cada subárvore podem obter realizando um percurso pós-ordem. Tal procedimento foi proposto por (FELSENSTEIN, 1981b), é apresentado no Algoritmo a seguir.

Algorithm 4: Cálculo de verossimilhança.

Entrada:
τ, uma árvore com raiz r.
B, o conjunto de comprimentos de ramo de τ
D, um conjunto de dados contendo sequências de N_{sit} sítios.
M, o modelo de substituição de sequências que define a matriz probabilidades de transição de estados **P**
Saída: lnL
para cada *sítio j=1...N_{sit}* **hacer**
  para cada *nó interno v com filhos u e w percorrendo r em pós-ordem* **hacer**
    calcular $L_j^v(v_j)$, (Equação D.3);
  fin
  Calcular L_j (Equação D.5)
fin
Calcular ln$L$ (Equação D.6).;