



**UNIVERSIDADE ESTADUAL DE SANTA CRUZ
PRO-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO**

**PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL
EM CIÊNCIA E TECNOLOGIA**

MURILO SILVA SANTANA

**DESENVOLVIMENTO DE UMA FERRAMENTA AUTOMÁTICA PARA SELEÇÃO DE
MÉTODOS E MODELOS EVOLUTIVOS DE DNA PARA A RECONSTRUÇÃO DE
ÁRVORES FILOGENÉTICAS**

**ILHÉUS-BA
2016**

MURILO SILVA SANTANA

**DESENVOLVIMENTO DE UMA FERRAMENTA
AUTOMÁTICA PARA SELEÇÃO DE MÉTODOS E MODELOS
EVOLUTIVOS DE DNA PARA A RECONSTRUÇÃO DE
ÁRVORES FILOGENÉTICAS**

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional em Ciência e Tecnologia da Universidade Estadual de Santa Cruz, como parte das exigências para obtenção do título de Mestre em Modelagem Computacional em Ciência e Tecnologia.

Orientador: Prof^a. Dra. Martha Ximena Torres Delgado

ILHÉUS-BA
2016

S232

Santana, Murilo Silva.

Desenvolvimento de uma ferramenta automática para seleção de métodos e modelos evolutivos de DNA para a reconstrução de árvores filogenéticas / Murilo Silva Santana. – Ilhéus, BA: UESC, 2016.

92 f. : Il.

Orientadora: Martha Ximena Torres Delgado.

Dissertação (mestrado) – Universidade Estadual de Santa Cruz. Programa de Pós-graduação em Modelagem Computacional em Ciência e Tecnologia.

Inclui referências e apêndices.

1. Filogenia – Metodologia. 2. Análise cladística – Matemática. 3. Bioinformática. 4. Evolução (Biologia) – Programas de computador. 5. Árvores – Classificação. 6. Modelos matemáticos. I. Título.

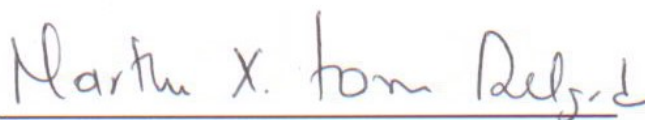
CDD 578.012

MURILO SILVA SANTANA

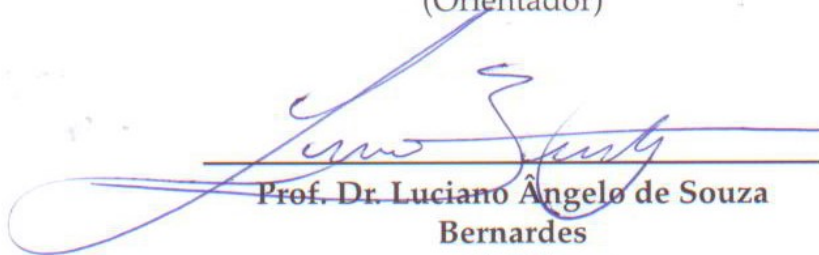
**DESENVOLVIMENTO DE UMA FERRAMENTA
AUTOMÁTICA PARA SELEÇÃO DE MÉTODOS E MODELOS
EVOLUTIVOS DE DNA PARA A RECONSTRUÇÃO DE
ÁRVORES FILOGENÉTICAS**

Ilhéus-BA, 22/07/2016

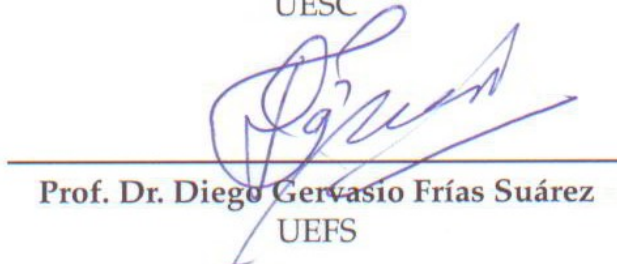
Comissão Examinadora



Prof^a. Dra. Martha Ximena Torres
Delgado
UESC
(Orientador)



Prof. Dr. Luciano Ângelo de Souza
Bernardes
UESC



Prof. Dr. Diego Gervasio Frías Suárez
UEFS

À minha amada esposa Andressa, que me apoiou nessa jornada, aos meus pais, pois a eles eu devo tudo em minha vida, à minha orientadora, Martha Ximena Torres Delgado, que com seu compromisso e competência me orientou, ao professor Francisco Bruno, pelo excelente trabalho como coordenador do programa, ao professor Dany Sanchez, por ser uma referência para mim como professor e profissional e aos meus amigos Thiago Paim e Luenne Nailam pelo companheirismo e amizade, assim como a quem é brilhante e ilustre, dedico.

"The affinities of all the beings of the same class have sometimes been represented by a great tree... As buds give rise by growth to fresh buds, and these if vigorous, branch out and overtop on all sides many a feebler branch, so by generation I believe it has been with the great Tree of Life, which fills with its dead and broken branches the crust of the earth, and covers the surface with its ever branching and beautiful ramifications."
Charles Darwin

Desenvolvimento de uma ferramenta automática para seleção de métodos e modelos evolutivos de DNA para a Reconstrução de Árvores Filogenéticas

Resumo

A reconstrução de árvores filogenéticas (RAF) é o processo de inferência que visa representar as relações evolutivas existentes entre as espécies. Para fazer RAF, depois de ter o conjunto de sequências alinhadas, é necessário escolher o modelo evolutivo e seus parâmetros que melhor se adeque aos dados. Além disso, é necessário escolher o método apropriado para gerar a árvore correta. No caso de RAF de DNA, tem-se pelo menos 230 modelos evolutivos à disposição. Ademais, é preciso escolher entre os 4 principais métodos para RAF (distância, parcimônia, inferência bayesiana e máxima verossimilhança), todos com suas particularidades. Pesquisas bibliográficas apontam que é fundamental fazer a escolha correta, tanto do modelo evolutivo, quanto do método, para se obter a reconstrução mais aproximada para um determinado conjunto de dados. Sendo assim, uma ferramenta que permita realizar a escolha adequada de métodos e modelos evolutivos de forma automática é útil para realizar RAF. Para o desenvolvimento deste trabalho, foi necessário realizar uma pesquisa bibliográfica das metodologias utilizadas em RAF, no ano de 2015, além do embasamento teórico sobre as ferramentas de seleção de modelos evolutivos e as aplicações dos principais métodos de RAF para cada conjunto de sequências de alinhadas. De tal modo, esta dissertação apresenta uma ferramenta automática de seleção de modelos evolutivos de DNA e dos métodos, fazendo a sua escolha baseada na sequência de entrada, assim como na finalidade da RAF pretendida. Essa ferramenta será integrada ao *vebservice* IgrafuWeb.

Palavras-chave: Bioinformática, RAF, Ferramentas de seleção de modelos evolutivos, Modelos evolutivos, Métodos de RAF

Development of an automated tool for automatic selection of the evolutionary models and methods for Reconstruction of Phylogenetic Trees

Abstract

The phylogenetic tree reconstruction (PTR) is the process of inference that aims to represent the existing evolutionary relationships among species. To realize PTR, after the aligned sequence, you must choose the evolutionary model that best fits the data and its parameters. Furthermore, it is necessary to select the appropriate method to generate the correct tree. In the case of PTR on DNA, exists at least 230 evolutionary designs to be chosen. Moreover, it is necessary to choose between the 4 main methods for PTR (distance, parsimony, maximum likelihood and Bayesian inference), all with its peculiarities. Bibliographic researches indicates that is essential to make the right choice, both of the evolutive model and the PTR method for obtaining the most approximate reconstruction for a given set of data. Therefore, a tool that allows to perform the suitable choice of methods and evolution models automatically is useful to perform PTR. To develop this work, it was necessary to perform a literature review of the methodologies used in PTR in year of 2015, in addition to the theoretical basis of the models of evolutionary selection tools and applications of the main PTR methods for each input sequence. Therefore, this work presents an automatic tool for selection of evolutionary models of DNA and methods, making your choice based on the input sequence, as well as the intended purpose of the PTR and is integrated to IgrafuWeb webservice.

Keywords: Bioinformatics , PTR , Evolutionary models Selection tools , Evolutionary models , PTR Methods

Lista de figuras

Figura 1 – Árvore filogenética	1
Figura 2 – Métodos e suas vantagens e desvantagens	4
Figura 3 – Matriz de alinhamento	9
Figura 4 – Alinhamento de sequências	9
Figura 5 – Terminologias em uma árvore	11
Figura 6 – Exemplo de árvore sem raiz	11
Figura 7 – Ilustração da árvore ((A,B)((C,D),E));	12
Figura 8 – Possíveis árvores com raiz para 4 sequências	14
Figura 9 – Possíveis árvores sem raiz para 4 sequências	14
Figura 10 – Parte de sequências de três diferentes espécies	15
Figura 11 – Parte de sequências de três diferentes espécies	15
Figura 12 – Árvore sem raiz obtida através do método NJ.	18
Figura 13 – (a) Árvore sem raiz em formato estrela; (b) árvore sem raiz com os elementos 1 e 2 agrupados.	18
Figura 14 – Possíveis topologias sem raiz possíveis para 4 espécies	21
Figura 15 – Identificação da Topologia A, onde as mudanças são representadas por traços na árvore	22
Figura 16 – Identificação da Topologia B, onde as mudanças são representadas por traços na árvore	22
Figura 17 – Identificação da Topologia C, onde as mudanças são representadas por traços na árvore	23
Figura 18 – Árvore exemplo para o cálculo de Máxima Verossimilhança	24
Figura 19 – Tela inicial do JModelTest, modo gráfico	46
Figura 20 – Página inicial apresentado a estrutura organizacional do site e uma breve explicação do IgrafuWeb.	52
Figura 21 – Importância da adequação de modelos	57
Figura 22 – Reinos estudados	58
Figura 23 – Quantificação dos métodos utilizados pelos autores	58
Figura 24 – Quantidade de Métodos utilizados em cada artigo	59
Figura 25 – Valores de bootstrap e probabilidade posterior de diferentes métodos para uma mesma sequência	60

Figura 26 – Quantificação dos modelos Default, F81 (Felsenstein 81), GTR (<i>General-Time-Reversible</i>), HKY (Hasegawa-Kino-Yano), JC (Jukes-Cantor), K80 (Kimura), nenhum modelo, SYM (<i>symmetrical model</i>), T92 (Tamura92), TIM (<i>Transitional Model</i>), TN93 (Tamura93), TPM (<i>Three Parameters model</i>), TVM(<i>Transversional Substitution Model</i>), utilizados pelos autores. O modelo default é o padrão dos softwares de RAF utilizado pelos autores.	60
Figura 27 – Utilização dos modelos evolutivos de acordo com os seus parâmetros.	61
Figura 28 – ferramenta para seleção de modelo evolutivo utilizadas pelos autores	62
Figura 29 – Técnicas para seleção de modelos evolutivos utilizadas pelos autores	62
Figura 30 – Ferramentas utilizadas para a RAF	63
Figura 31 – Métodos utilizados em RAFs que não usaram nenhuma ferramenta de escolha de modelo evolutivo	64
Figura 32 – Modelos utilizados em RAFs que não usaram nenhuma ferramenta de escolha de modelo evolutivo e utilizaram o método de inferência bayesiana	64
Figura 33 – Modelos utilizados em RAFs que não usaram nenhuma ferramenta de escolha de modelo evolutivo e utilizaram o método de MV	65
Figura 34 – Fluxograma detalhando a metodologia criada	69
Figura 35 – Tela inicial do IgrafuWeb	70
Figura 36 – Fluxo de execução e dados do IgrafuWeb	71
Figura 37 – Exemplo de arquivo de entrada no formato NEXUS	71

Lista de tabelas

Tabela 1 – Quantidade de topologias	13
Tabela 2 – Exemplo de alinhamento de sequências de DNA.	20
Tabela 3 – Modelos evolutivos usados JModelTest (alguns dos modelos possíveis). Qualquer um desses modelos podem incluir os parâmetros I, G ou os dois (+I+G).	48

Lista de abreviaturas e siglas

CACAU	Centro de Armazenamento de Dados e Computação Avançada da UESC
DCET	Departamento de Ciências Exatas e Tecnológicas
DNA	Ácido Desoxirribonucleico
DNAPARS	DNA Parsimony Program
F81	Felsenstein 81
FASTA	Fast Alignment
GTR	General Reversible Time
HKY	Hasegawa Kino Yano
IB	Inferência Bayesiana
IC	Intervalo de Confiança
JC69	Jukes-Cantor
K2P	Kimura 2 Parameter
MCMC	Markov Chain Monte Carlo
MP	Máxima Parcimônia
MV	Máxima Verossimilhança
NBCGIB	Núcleo de Biologia Computacional e Gestão de Informações Biotecnológicas
NJ	Neighbor Joining
NNI	Nearest Neighbor Interchange
OTUs	Operational Taxonomic Units
PHYLP	PHYLogeny Inference Package
RAF	Reconstrução de Árvores
RNA	RiboNucleic Acid

SPR	Subtree Pruning Regrafting
SSH	Secure Shell
SYM	Symmetrical Model
TIM	Transitional Model
TPM	Three Parameters model
TVM	Transversional Substitution Model
UESC	Universidade Estadual de Santa Cruz
UPGMA	Unweighted Pair Grouping Method with Arithmetic Means

Sumário

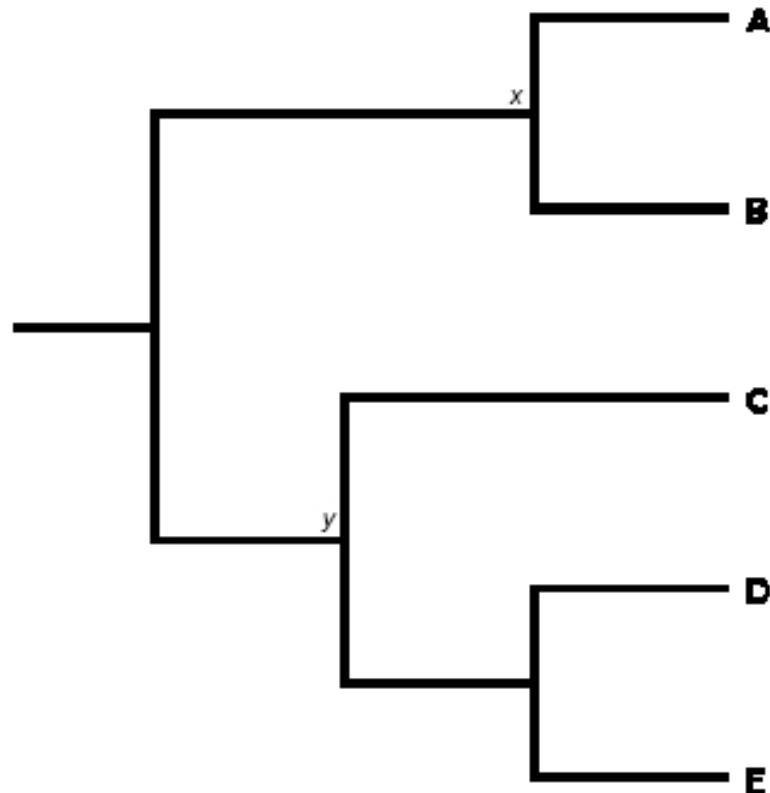
1 – Introdução	1
1.1 Motivação	2
1.2 Objetivos	6
1.3 Organização do trabalho	7
2 – Referencial Teórico	8
2.1 Princípios da filogenia	8
2.1.1 Representação de dados	8
2.1.2 Determinação de presença de sinal filogenético	10
2.1.3 Árvores filogenéticas	10
2.2 Métodos para RAF	13
2.2.1 Distância	13
2.2.1.1 Método UPGMA	16
2.2.1.2 <i>Neighbor-Joining</i> (NJ)	17
2.2.2 Parcimônia	19
2.2.3 Máxima Verossimilhança	23
2.2.4 Inferência Bayesiana	26
2.2.4.1 Método de Monte Carlo via Cadeias de Markov	27
2.3 Processos de Markov	29
2.4 Modelos evolutivos	32
2.4.1 O modelo Jukes Cantor	34
2.4.2 Modelo Kimura dois parâmetros (K2P)	34
2.4.3 O modelo F81 (Felsenstein 1981)	35
2.4.4 O modelo HKY (Hasegawa et al. 1984, 1985)	36
2.4.5 O modelo TN93 (Tamura and Nei 1993)	36
2.4.6 O modelo <i>General Time Reversible</i> (GTR)	37
2.4.7 Taxas evolutivas de heterogeneidade	37
2.4.7.1 Modelos com proporções de Sítios Invariantes	38
2.4.7.2 Distribuição gama para taxas de mutação	38
2.5 Métodos e Ferramentas para a escolha de modelos evolutivos	40
2.5.1 <i>Hierarchical likelihood ratio test</i> (HLRT)	41
2.5.2 <i>Akaike Information Criteria</i> (AIC)	41
2.5.3 <i>Bayesian Information Criteria</i> (BIC)	42
2.5.4 <i>Decision Theory</i> (DT)	42
2.5.5 <i>Bayes Factor</i>	43
2.6 Diferenças e vantagens dos métodos de escolha de modelo evolutivo	43

2.7	Intervalo de confiança	44
2.8	JModelTest	45
3	– Metodologia	49
3.1	Pesquisa dos fundamentos de RAF	49
3.2	Busca e leitura de artigos sobre filogenia	49
3.3	Levantamento e cruzamento de dados obtidos nos artigos	50
3.4	Desenvolvimento da metodologia utilizada na ferramenta	50
3.5	Implementação da ferramenta	51
3.5.1	IgrafuWeb	51
3.5.1.1	DNAPARS e PROTPARS	52
3.5.1.2	Digrafu	53
3.5.1.3	PhyML	53
3.5.1.4	MrBayes	54
3.5.2	CACAU	54
4	– Desenvolvimento	56
4.1	O problema da escolha do modelo	56
4.2	Análise da pesquisa bibliográfica	57
4.3	Cálculo de distância para escolha simplificada de métodos e modelos evolutivos	66
4.4	Descrição da metodologia proposta	68
4.5	Desenvolvimento da ferramenta	69
4.5.1	Implementação do cálculo de distância entre sequências	72
4.5.2	Módulo de seleção de modelo evolutivo	72
4.5.3	Chamada aos métodos do IgrafuWeb	73
4.5.4	Execução	73
4.6	Testes de validação da ferramenta	74
5	– Conclusão e trabalhos futuros	76
	Referências	77
	Apêndices	81
	APÊNDICE A – Lista das revistas e seus respectivos artigos	82

1 Introdução

Na biologia, a filogenia pode ser definida como o estudo da relação evolutiva entre grupos de organismos. Essa relação é feita através da análise de dados de sequências de aminoácidos, nucleotídeos e códonos. Entender filogenia é bem parecido com ler uma árvore genealógica. A raiz da árvore representa a linhagem ancestral, e as pontas das ramificações representam os descendentes desse ancestral. Em uma árvore filogenética, conforme se avança da raiz para as pontas, se avança no tempo. A filogenia se fundamenta na Teoria da Evolução, que afirma que grupos com organismos que apresentam atributos similares descendem de um ancestral comum. A figura 1 ilustra uma árvore filogenética. O estudo filogenético normalmente objetiva testar a validade de grupos e sua taxonomia. Seguindo esse ponto de vista, apenas são aceitos como naturais os grupos confirmadamente monofiléticos, ou seja, que descendem de um ancestral comum([MATIOLI, 2013](#))

Figura 1 – Árvore filogenética



A divisão de uma linhagem (especiação) é representada como uma ramificação na árvore. Quando um evento de especiação ocorre, uma única linhagem ancestral dá origem a duas linhagens filhas. As árvores filogenéticas traçam padrões de ancestralidade entre espécies. Cada espécie tem uma parte de sua história que é única e outra

parte que é compartilhada com outras espécies.

As árvores filogenéticas representam a estrutura básica para se realizar dois procedimentos de estudo:

- expressar em um único modelo as diferenças relativas entre espécies;
- permitir a validação de modelos estatísticos para o processo evolutivo.

Segundo (SILVA, 2007), a reconstrução de árvores filogenéticas (RAF) é o processo de inferência que visa representar as relações evolutivas existentes entre espécies tomando como parâmetro as bases nitrogenadas contidas no DNA e RNA e proteínas. Essa inferência se dá pelo uso de algoritmos computacionais que agrupam as espécies sob a forma de uma árvore filogenética, partindo da hipótese de que todos os seres vivos descendem de um ancestral comum e que cada espécie conhecida ocupará uma folha na árvore filogenética.

Baseada na sua utilidade, as árvores filogenéticas contém informações de extrema utilidade para uma grande diversidade de questões biológicas e sociais. São de inconteste importância no controle e combate de parasitas responsáveis por diversas doenças, no estudo epidemiológico, na criação de vacinas mais eficientes, na criação de novas drogas farmacológicas (AMORIM; SOUZA, 2002). Uma outra utilidade das árvores filogenéticas é fornecer subsídios de extrema importância para protocolos de transplante de órgãos e tecidos inter-espécies (SILVA, 2016)

Árvores filogenéticas são estruturas hierárquicas (como árvores genealógicas) que expressam os relacionamentos entre indivíduos de espécies diferentes. Elas não podem ser vistas como genealogias porque geralmente os indivíduos ancestrais de duas espécies relacionadas na árvore são desconhecidos. É importante saber que os tamanhos dos ramos de uma árvore filogenética representam o tempo evolutivo entre uma espécie ancestral e a espécie descendente, o que dá à árvore uma aparência de grafo.(VIEIRA, 2007)

1.1 Motivação

Para a análise filogenética os cientistas analisam caracteres moleculares, ou seja, sequências de genes. Esses caracteres possuem um poder comparativo muito abrangente, pois todos os organismos vivos compartilham da mesma organização genética. Neste caso, cada sítio observado em uma dada sequência é considerado um caracter. Para se fazer a análise, as sequências genéticas de diferentes organismos devem primeiramente ser alinhadas. Com o advento de métodos de sequenciamento genético cada vez mais baratos, hoje em dia existe enorme quantidade de organismos com seus genomas completamente sequenciados (HENNIG et al., 1999).

Em todas as metodologias de filogenética molecular, cada posição ocupada na sequência (nucleotídeo ou aminoácido) é considerada como um caráter do tipo multiestado (podendo ser um dos quatro nucleotídeos ou um dos vinte aminoácidos) e cada caráter é considerado independente dos demais. A variação dos estados de caracteres fornecerá informações filogenéticas. (MATIOLI, 2013)

A reconstrução de árvores filogenéticas confiáveis envolve um esforço significativo, em razão das dificuldades inerentes ao processo de aquisição de dados biológicos e, posteriormente, à complexidade computacional associada ao problema de reconstrução (HOLMES, 2003). As árvores filogenéticas representam hipóteses da história evolutiva das espécies. No entanto, a inferência da árvore que se mostra mais adequada aos dados das sequências é uma tarefa difícil e exige um alto custo computacional. Além disso, para um determinado conjunto de táxons existe mais de uma possibilidade de árvores a serem analisadas, o que torna esse processo muito mais complicado.

Segundo (SILVA, 2007), a complexidade computacional do procedimento de reconstrução filogenética e a dificuldade de se chegar a um resultado inquestionável se devem a quatro fatores principais, os quais estão interligados:

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

Conforme afirma (PINTO, 2004) existem três abordagens possíveis frente a problemas multi-objetivo:

- Priorizar um objetivo obtendo-se a solução com base em uma formulação mono-objetivo para o problema;
- Considerar simultaneamente um sub-conjunto de objetivos ponderados em única função-objetivo gerando um problema mono-objetivo que procura compor a importância relativa de cada objetivo em uma função matemática;

- Tratar como um problema multi-objetivo em sua forma original onde se buscam soluções não-dominadas entendendo-se a impossibilidade de melhorar o atendimento a qualquer objetivo sem piorar outro.

Os métodos que são normalmente usados para estimar as árvores filogenéticas podem ser agrupados em quatro classes principais: parcimônia, distância, máxima verossimilhança e Inferência Bayesiana. Segundo (SILVA, 2007), os métodos de reconstrução de árvores filogenéticas se dividem em métodos fenéticos (independentes de modelo) e métodos cladísticos (dependentes de modelo), que utilizam diferentes técnicas, cada um tratando de forma peculiar as hipóteses sobre o processo de evolução utilizando heurísticas próprias ou algoritmos estatísticos, para produzir a filogenia desejada (NETO, 2015). Com exceção da parcimônia, o restante dos métodos dependem de um modelo matemático responsável por descrever a evolução (modelo evolutivo), para descrever o processo a partir do qual um sítio em uma sequência pode se modificar.

A figura 2 ilustra as principais vantagens e desvantagens dos métodos de RAF:

Figura 2 – Métodos e suas vantagens e desvantagens

Método	Tipo	Vantagens	Desvantagens	Softwares
Parcimônia	Fenético	Suficientemente rápido para a análise de centenas de sequências; robusto se ramos são curtos (sequências estreitamente relacionadas ou amostragem densa)	Pode ser inexato, se há uma variação substancial nos comprimentos dos ramos	PAUP; MEGA; PHYLIP; NONA.
Distância	Cladístico	Rápido	Informação é perdida ao comprimir sequências em distâncias; Estimativas confiáveis de distâncias entre pares pode ser difícil de obter para sequências divergentes	PAUP; MEGA; PHYLIP; Digrafu.
Máxima Verossimilhança	Cladístico	A MV capta perfeitamente o que os dados nos dizem sobre a filogenia sob um determinado modelo	Pode ser proibitivamente lento (dependendo do rigor da pesquisa e acesso a recursos computacionais)	PAUP; MEGA; PHYLIP; PhyML.
Inferência Bayesiana	Cladístico	Tem uma forte ligação com o método de máxima verossimilhança; Pode ser uma maneira mais rápida para avaliar o suporte para bootstrapping	As distribuições a priori para os parâmetros devem ser especificados; Pode ser difícil determinar se a aproximação por cadeia de Markov Monte Carlo (MCMC) foi executada por tempo suficiente	MrBayes; BAMBE.

adaptado e traduzido de (HOLDER; LEWIS, 2003)

De acordo com (HOLDER; LEWIS, 2003), a escolha do método de RAF é de suma importância, tendo em vista o desempenho, a adequação do método com o tipo de sequências, assim como seu comprimento, o que pode levar a um gasto excessivo de

tempo na RAF ou árvores com topologias incorretas, se selecionado um método menos adequado para a inferência.

Conforme afirma ([KELCHNER; THOMAS, 2006](#)), a maioria das RAFs dependem de modelos evolutivos formais e conceituais. Portanto, as técnicas ou aplicações filogenéticas exigem um modelo que abrange adequadamente os processos mutacionais que geram os dados. A escolha de diferentes modelos podem resultar em topologias diferentes, além de diferentes comprimentos de galhos que consequentemente podem alterar o resultado de uma RAF, o que pode então alterar a nossa inferência sobre a quantidade de mudança que ocorreu entre duas sequências. Qualquer aplicação em filogenia se baseia em uma escolha correta da quantidade de mudanças entre os sítios das sequências, ou seja, precisam de modelos adequados.

Baseada na evolução independente entre sítios, a filogenia utiliza modelos evolutivos baseados em cadeias de Markov para calcular a probabilidade de mudança de um caractere para outro, considerando os caracteres encontrados em duas sequências quaisquer num determinado sítio. Os modelos baseados em cadeias de Markov determinam dentro de uma mutação quais as probabilidades dos caracteres substituírem uns aos outros ([SILVA, 2016](#)).

Os modelos evolutivos podem ser definidos por dois tipos de parâmetros que determinam seu comportamento:

- frequência de cada nucleotídeo: parâmetro que mede a frequência de nucleotídeos na matriz de dados;
- tipos de substituições e suas taxas de substituição: as taxas de substituição são as taxas relativas de mudança de um nucleótido para outro em um intervalo de tempo t_0 a um tempo t_1 ;

A partir das combinações possíveis destes parâmetros podem ser identificados cerca de 203 modelos de substituição de nucleótidos. Os modelos mais simples são aqueles que incluem menos parâmetros.

Além desses possíveis modelos existem também outros parâmetros, como a distribuição gama (G), que serve como aproximação do quanto as taxas de substituição são variáveis dentro de cada gene, sendo o seu valor inversamente proporcional à variação e proporção de sítios invariáveis (I), no qual uma determinada proporção de sítios é assumida como incapaz de sofrer substituição e que pode também ser usada para modelar a heterogeneidade entre os sítios.

Os modelos existentes não conseguem modelar com exatidão a maneira como os seres vivos evoluem. O que pode-se buscar, mesmo sabendo que não existe um modelo

perfeito ou exato, é o modelo que melhor se adequa ao conjunto de dados (KELCHNER; THOMAS, 2006; POSADA; CRANDALL, 2001).

Apesar de existirem ferramentas para a seleção de modelos evolutivos e seus parâmetros, como JModelTest (POSADA; CRANDALL, 1998) e MrModelTest (IRESTEDT et al., 2004), que se baseiam em técnicas estatísticas como *Akaike Information Criteria* (AIC) e *Bayesian Information Criteria* (BIC), essas ferramentas apenas estudam o conjunto de sequências de entrada desconsiderando a finalidade da RAF. Segundo (KELCHNER; THOMAS, 2006) a importância do ajuste do modelo pode ser visto como um processo contínuo dependendo da técnica filogenética ou aplicação a ser utilizada e as questões que estão sendo investigadas.

Embora (KELCHNER; THOMAS, 2006) tenha feito uma pesquisa bibliográfica que analisa através de publicações feitas em tal data, verificando quais são os modelos evolutivos mais utilizados e qual é a técnica de seleção de modelos mais utilizada. Não foi encontrada uma pesquisa nos dias atuais que nos permita conhecer qual é metodologia que os pesquisadores realizam para fazer RAF desde o conjunto de sequências alinhadas até obter a árvore filogenética.

Existem muitos estudos sobre a importância de realizar corretamente a escolha do modelo evolutivo e também do método apropriado, porém eles não fazem nenhuma análise da natureza da aplicação. Outro problema existente é que não existe nenhuma ferramenta que automatize a RAF, tendo o usuário que fazer todas as escolhas nesse processo, o que geralmente leva a escolha de modelos menos adequados para as sequências analisadas. Dessa maneira, fica evidente a necessidade de automatizar a escolhas dos modelos evolutivos e métodos para a RAF.

1.2 Objetivos

O presente trabalho tem como objetivo desenvolver uma ferramenta que permita realizar a RAF, tendo como entrada as sequências alinhadas e o tipo de aplicação para qual será utilizada a RAF, fornecendo de maneira automática as árvores e escolhendo automaticamente o método, o modelo evolutivo e seus parâmetros de maneira que melhor se adequem aos dados e à aplicação.

O objetivo geral será alcançado pela execução dos seguintes objetivos específicos:

1. Realizar uma pesquisa bibliográfica atualizada sobre como é realizada a RAF, para comparar os dados publicados, o que vai permitir saber se de fato estão sendo utilizadas todas as ferramentas atualmente disponíveis e se estão sendo realizadas corretamente de acordo às exigências teóricas;

2. Criar a metodologia que a ferramenta a ser implementada utilizará para a escolha dos métodos, modelos e seus parâmetros;
3. desenvolver uma ferramenta automática para escolha do modelo e parâmetros baseada nos programas existentes.
4. desenvolver uma ferramenta automática que escolha um programa que implementa o método, realize o fundamento estatístico se necessário e gere a resposta.

1.3 Organização do trabalho

O restante trabalho está organizado da seguinte maneira:

- Capítulo 2: Referencial teórico: apresenta a revisão de literatura, abordando uma visão a respeito de filogenia, métodos para RAF, cadeias de Markov, Modelos evolutivos e ferramentas e metodologias para escolha de modelos, por serem objetos de estudos deste trabalho;
- Capítulo 3: Metodologia: apresenta os aspectos metodológicos do trabalho, materiais, métodos e recursos computacionais utilizados para o desenvolvimento do trabalho;
- Capítulo 4: Resultados: apresenta a conclusão da pesquisa bibliográfica, a metodologia criada e a descrição da proposta do sistema e o sistema criado;
- Capítulo 5: Considerações finais: apresenta as conclusões deste trabalho, bem como as sugestões de propostas para a elaboração de trabalhos futuros.

2 Referencial Teórico

2.1 Princípios da filogenia

Segundo ([PINTO, 2004](#)), o processo de análise filogenética se divide em cinco passos:

1. Alinhamento das sequências
2. Determinação da presença de um sinal filogenético
3. Escolha do método de RAF
4. Construção da árvore filogenética
5. Avaliação da árvore filogenética

2.1.1 Representação de dados

Os estudos de biologia molecular têm produzido uma extensa quantidade de dados filogenéticos que podem ser apresentados como sequências de diversos tipos. Dentre os principais tipos, destacam-se as sequências de DNA, RNA e sequências de proteínas. A inferência filogenética tem como base esses três tipos de sequências, os quais representam o mesmo gene em cada um dos organismos.

Os dados das sequências são compostos por uma série de caracteres desenhados a partir de um alfabeto limitado. As sequências de DNA são compostas que por sua vez são de quatro tipos: Adenina (A), Citosina (C), Timina (T) e Guanina (G). São chamados de purinas os nucleotídeos dos tipos A e G, e pirimidinas os tipos C e T. Já as sequências protéicas consistem de uma sucessão de aminoácidos, que podem assumir 20 estados diferentes. Para se reconstruir árvores filogenéticas é necessário que as sequências sejam homólogas, ou seja, devem ter ancestrais em comum. Nesse sentido, analisando-se uma determinada sequências de DNA, considera-se que ocorrem mutações nos nucleotídeos durante a evolução. Tais mutações podem ser divididas em ([FELENSTEIN, 2004](#)):

- Substituições: um caractere é trocado pelo outro;
- Deleções: deleção de uma quantidade específica de caracteres;
- Inserções: inserção de qualquer quantidade de caracteres.

As sequências de DNA ou de aminoácidos das espécies em análise são alinhadas em forma de uma matriz, e serão utilizadas na reconstrução da árvore filogenética. A matriz é obtida através de técnicas de alinhamento de caracteres homólogos, agregando-os na maior quantidade de colunas possíveis (FELENSTEIN, 2004). A Figura 3 ilustra a matriz citada.

Figura 3 – Matriz de alinhamento

$$X = \{x_{ij}\} = \begin{matrix} \text{Species 1} \\ \text{Species 2} \\ \text{Species 3} \\ \vdots \\ \text{Species s} \end{matrix} \left\{ \begin{matrix} A & A & C & C & T \\ A & A & C & G & G \\ A & C & C & C & T \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A & C & C & C & T \end{matrix} \right\}$$

fonte: (GONÇALVES, 2008)

O alinhamento de sequências consiste no processo de comparar duas ou mais sequências (de nucleotídeos ou aminoácidos) de forma a se observar seu nível de similaridade. O alinhamento de sequências é uma forma de organizar sequências de DNA, RNA ou proteína para identificar regiões similares que possam ser consequência de relações funcionais, estruturais ou evolucionárias entre elas. Espaçamentos (gaps) podem ser inseridos entre os resíduos para que caracteres semelhantes (por algum critério) sejam alinhados em colunas sucessivas. Sequências alinhadas de nucleotídeos ou resíduos de aminoácidos são representadas tipicamente como linhas de uma matriz (BRITO, 2003).

Antes de serem alinhadas as sequências só podem ser analisadas isoladamente. O alinhamento de sequência é o ponto de partida para análises biológicas diversas, dentre elas como a derivação da sequência de medidas de similaridade, a identificação dos sítios homólogos, reconstrução filogenética, identificação de domínios funcionais, etc. É o início de todas as análises que envolvem comparação de dados moleculares, como pode ser visto na figura 4 (MULAN, 2002).

Figura 4 – Alinhamento de sequências

```
Vaca    ATG---ACTAACATTCGAAAGTCCACCCCACTAATAAAAAATTGTAAC
Ovelha  ATG---ATCAACATCCGAAAAACCCACCCCACTAATAAAAAATTGTAAC
Cabra   ATG---ACCAACATCCGAAAGACCCACCCCACTAATAAAAAATTGTAAC
Cavalo  ATG---ACAAACATCCGAAATCTCACCCCACTAATAAAATCATCAAT
Burro   ATG---ACAAACATCCGAAATCCACCCGCTAATAAAATCATCAAT
Ostra   ATGGCCCCCAACATTCGAAATCGCACCCCTGCTCAAAATATCAAC
Emu     ATGGCCCCTAACATCCGAAATCCACCCCTCACTCAAAATCATCAAC
Peru    ATGGCACCCCAATATCCGAAATCACACCCCTATTAAAAACATCAAC
```

Adaptado e traduzido de (MULAN, 2002)

2.1.2 Determinação de presença de sinal filogenético

As características das espécies, sejam elas morfológicas, fisiológicas ou comportamentais são certamente vinculados ao nicho e suscetíveis ao processo de evolução. Certamente, a modelagem das características do nicho, podem ser encaradas como um fenômeno de evolução. Essa evolução pode ser através da análise relacionada à distância funcional e filogenética da sequência. Logo, a estrutura filogenética pode ser definida por diferentes medidas, dentre elas, a diversidade filogenética, a distância média do táxon mais próximos e a variabilidade, riqueza e equitabilidade do táxon mais próximo. (SOBRAL; CIANCIARUSO, 2012)

2.1.3 Árvores filogenéticas

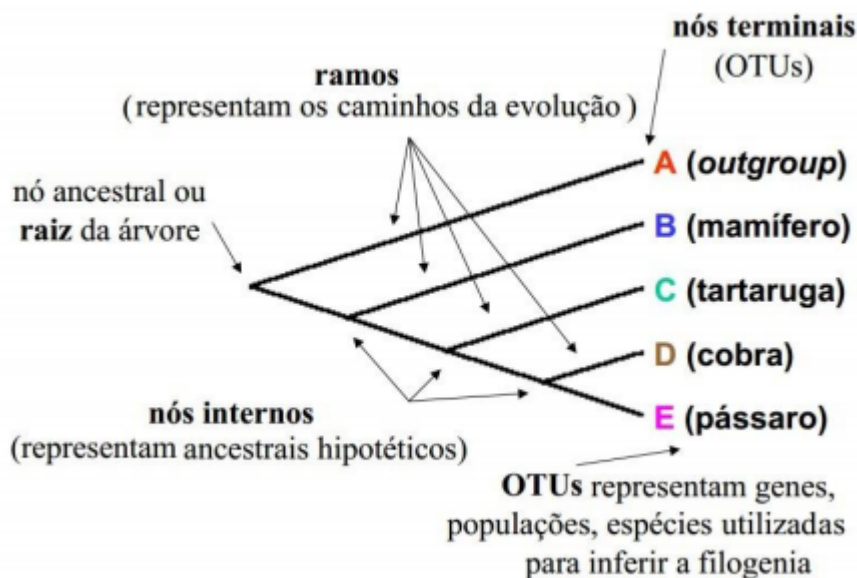
As árvores filogenéticas são representações gráficas estruturadas em forma de árvore que explicam a possível história evolutiva das espécies ou grupo de espécies. Essas árvores são construídas a partir de sequências de caracteres (tais como sequenciamento de DNA, códons, proteínas) obtidos de organismos diversos. Por ser considerada uma hipótese sobre relações evolutivas, se faz necessária a utilização de caracteres que sejam indicadores confiáveis de ancestralidade comum para a construção da árvore (HYPÓLITO, 2005; TICONA, 2008; AMORIM; SOUZA, 2002), podendo ser qualquer dessas recebidas como herança (NETO, 2015).

A compreensão de árvores filogenéticas exige o entendimento de algumas características (HYPÓLITO, 2005; TICONA, 2008)

1. Nós: são pontos nas árvores, que podem ser internos, representando hipoteticamente o ancestral comum ou terminais, que são os organismos estudados;
2. Ramos: são linhas que ligam os nós e representam os pontos de evolução entre eles;
3. OTUs - *Operational Taxonomic Units*: são espécies incluídas na análise, também conhecido como táxons, dos quais se deseja inferir a história filogenética;
4. Topologia: representação gráfica unindo as OTUs, através de ramos e nós.

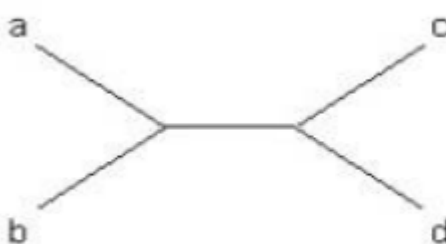
Tais características e terminologias são representadas na figura 5. Baseada nessa ilustração, percebe-se que de cada nó interno ramifica-se exatamente dois ramos, deixando claro que a relação evolutiva entre as espécies é representada por árvores binárias, isto é, no máximo duas ramificações saindo de cada nó interno.

Figura 5 – Terminologias em uma árvore



Existem também as árvores sem raiz, representada na figura 6, onde inexiste a relação de ancestralidade. Embora, seja possível escolher um ponto qualquer da árvore sem raiz onde se insira um nó raiz, gerando uma árvore raiz. A depender do lugar onde esses nós são inseridos, pode-se gerar várias outras árvores enraizadas. Não há validade em se aplicar um nó em uma árvore enraizada, já que pode haver comprometimento do resultado inicial.

Figura 6 – Exemplo de árvore sem raiz

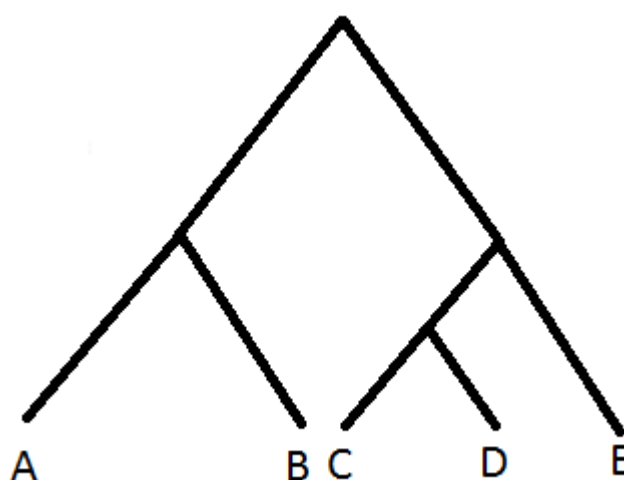


Nesse sentido, de acordo com (HYP6LITO, 2005; NETO, 2015) dá-se o nome de topologia à árvore que apresenta somente as relações de parentesco interespecíes, isto é, a forma como os nós internos são conectados uns com os outros e com as folhas, sem levar em consideração os valores de comprimento de galhos. Sendo assim, quanto maior a quantidade de táxons, maior o número de topologias a serem analisados.

As representações gráficas de árvores filogenéticas são realizadas através de softwares que organizam esses desenhos de acordo com os dados de entrada (arquivo de texto). Os dados de saída desses softwares são organizadas de uma forma peculiar,

na maioria das vezes determinado por um formato específico conhecido como Newick, idealizado pelo matemático inglês Arthur Cayley (1821-1895) em 1857, sendo adotado como padrão em 1986 (PHYLP, 2016). Dessa forma, idealizou-se esta notação como facilitador da manipulação computacional, baseado em uma lista de atributos correspondentes entre uma árvore e caracteres delimitados por parênteses aninhados. Logo, para as espécies hipotéticas A, B, C, D e E, pode-se determinar uma árvore segundo o padrão Newick da seguinte forma ((A,B)((C,D),E));, representada graficamente pela figura 7.

Figura 7 – Ilustração da árvore ((A,B)((C,D),E));



Sendo assim, de acordo à análise da Figura 5 e dos comentários acima, pode-se afirmar a respeito do padrão Newick (PHYLP, 2016):

- Os nós externos ou folhas são identificados pelos seus próprios nomes;
- O fim da árvore é determinado através do caractere ponto e vírgula;
- O par de parênteses representam um nó interno ou a raiz da árvore;
- Dentro desse par de parênteses, ficam os nós que são imediatamente descendentes desse nó separados por vírgulas. Logo, os descendentes imediatos são: B, um outro nó interno e D. Um novo nó interno é representado por um par de parênteses incluindo as representações de seus descendentes imediatos: A, C e E.

A notação Newick pode conter mais de uma identificação para uma determinada árvore. Como exemplo, pode-se citar as formas (A,(B,C),D) e (A,(C,B),D), que representam a mesma árvore. Além disso, existem situações de representação da mesma árvore diferenciando apenas na inserção ou não de raiz. Como é o caso da árvore com raiz

$(B,(A,D),C)$, que é a mesma da sem raiz $((A,D),(C,B))$ (SILVA, 2007). Os comprimentos dos ramos da árvore também são representados pela notação Newick, através da inclusão de um número real colocado depois de um nó e precedido pelo símbolo de dois pontos, sendo representada da seguinte maneira: $(B:6.0,(A:5.0,C:3.0,E:4.0):5.0,D:11.0);$.

De acordo com (SILVA, 2007), a equação que representa a quantidade (B) de possíveis árvores é calculada através da quantidade de sequências (S) é dada por:

$$B = \frac{(2s - 3)!}{2^{(s-3)}(s - 3)!}, \quad (1)$$

para árvores com raiz e

$$B = \frac{(2s - 5)!}{2^{(s-3)}(s - 3)!}, \quad (2)$$

para árvores sem raiz.

Por conta disso, o número de possíveis topologias, dado um conjunto de S sequências cresce de acordo com a tabela 1:

Tabela 1 – Quantidade de topologias

S	Quantidade de possíveis árvores sem Raiz	Quantidade de possíveis árvores com raiz
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10.395
8	10.395	135.135
9	135.135	2.027.025
10	2.027.025	34.459.425

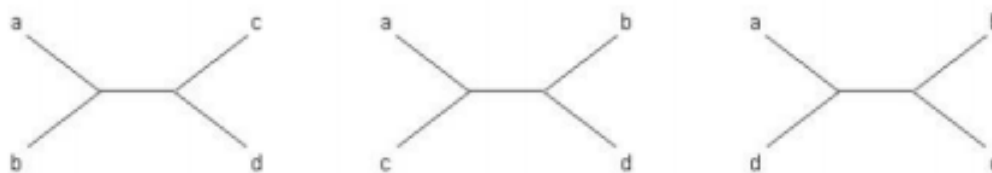
Levando em consideração o afirmado acima, um conjunto de 4 sequências pode gerar 3 diferentes árvores sem raiz (figura 8) e 15 árvores distintas com raiz (figura 9).

2.2 Métodos para RAF

2.2.1 Distância

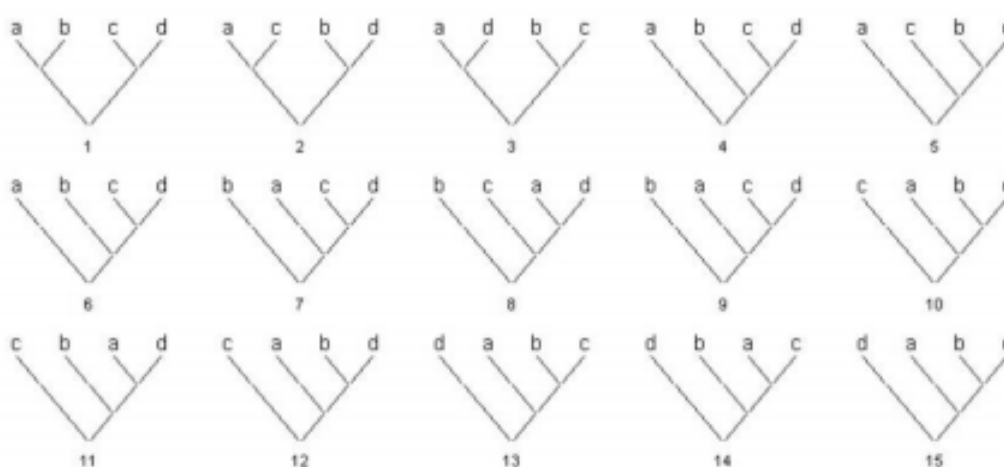
Foram os primeiros métodos utilizados para RAF. Os dados das sequências são primeiramente transformados em uma matriz de distâncias entre pares de sequências,

Figura 8 – Possíveis árvores com raiz para 4 sequências



Fonte: (SILVA, 2016)

Figura 9 – Possíveis árvores sem raiz para 4 sequências



Fonte: (SILVA, 2016)

com o comprimento total dos ramos necessários para ajustar esta matriz para cada topologia possível sendo calculado, com a topologia escolhida sendo aquela com o menor comprimento total de ramos (SILVA, 2016).

Os métodos baseados em distância foram pioneiros na RAF e ainda são muito utilizados devido a sua enorme velocidade, apesar de já existirem métodos mais exatos.

O processo de busca da melhor árvore filogenética também pode ser entendido como um processo de agrupamento, ou clusterização, ou seja, processo de organizar objetos em grupos cujos membros são similares de alguma forma. Os métodos que utilizam matrizes de distâncias como entrada podem ser conceituados, de forma resumida pelos seguintes aspectos(SILVA, 2007):

- existe uma matriz de distâncias que possui as distâncias observadas entre as espécies (OTUs);
- qualquer árvore que determina o comprimento de seus ramos acaba por prever um conjunto de distâncias d_{ij} , o qual pode ser calculado pela soma dos

comprimentos dos ramos entre as espécies i e j .

A matriz de distâncias pode ser gerada pela observação das diferenças existentes entre sequências de bases nucleotídicas, como as sequências ilustradas na figura 10:

Figura 10 – Parte de sequências de três diferentes espécies

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

fonte: (SILVA, 2007)

A matriz de distâncias gerada pela observação das diferenças existentes entre as bases de cada par de espécies da Figura 10 é mostrada pela Figura 11. Nesse exemplo, a distância é dada pelo número de vezes que uma base difere da sua base correspondente em cada posição de cada par de espécies comparadas. Existem formas mais sofisticadas de se obter essas distâncias par a par (FELENSTEIN, 2004).

Figura 11 – Parte de sequências de três diferentes espécies

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

fonte: (SILVA, 2007)

Em suma, os algoritmos dos métodos de distância selecionam um par de nós a serem fundidos utilizando um critério específico e estes dois nós selecionados são substituídos por um novo nó simples e a matriz de distância é reduzida, por substituir as distâncias relativas aos dois nós unidos por este novo nó (DIAS et al., 2011).

As matrizes de distâncias são geradas de acordo com as alterações sofridas entre cada par de espécies. Os relacionamentos entre essas distâncias são levados em conta no ato de criação da árvore. Ela é criada de acordo à análise de dois tipos de distância: distância estimada (observada) e distância evolutiva (genética) (PRADO, 2001).

A distância observada (p) corresponde à proporção das diferenças entre os sítios de duas sequências diferentes, cujo valor é obtido através da seguinte fórmula:

$$p = \frac{n_p}{n} \quad (3)$$

Onde n_p corresponde ao número de diferenças entre as duas sequências analisadas e n corresponde ao tamanho da sequência (FELENSTEIN, 2004).

Para a construção de uma filogenia válida, as distâncias observadas precisam ser corrigidas para levar em consideração as múltiplas substituições no mesmo sítio. Essa correção resulta nas distâncias evolutivas e é aplicada através de modelos evolutivos (FELENSTEIN, 2004).

A qualidade da árvore inferida está diretamente ligada à qualidade da estimação das distâncias. Assim sendo, a construção de uma boa matriz de distância é fundamental nos métodos baseados em distância, pois a árvore é construída a partir dos relacionamentos descritos na matriz (NEI; KUMAR, 2000).

Depois de criada a matriz de distância, o próximo passo para a inferência é utilizar uma heurística dos métodos baseados em distância. Algumas dessas heurísticas serão abordadas a seguir.

2.2.1.1 Método UPGMA

O método mais simples deste grupo (baseado em matriz de distâncias) é conhecido como *unweighted pair-group method using an arithmetic average*.

A árvore construída baseada neste método também é conhecida como fenograma, pois foi originalmente utilizada para representar a extensão das similaridades fenotípicas de um grupo de espécies em taxonomia numérica. Entretanto, ela também pode ser usada para construir filogenias. Particularmente, quando dados de frequência genética são usados para reconstrução filogenética, este modelo produz resultados bons e confiáveis, quando comparados com outros métodos baseados em matriz de distâncias (PRADO, 2001).

NEY; KUMAR (2000) afirma que para construir uma árvore utilizando UPGMA, os passos são necessários:

O primeiro passo é entrar com uma matriz de distâncias, de tamanho $l = m \times n$. Depois disso, é necessário encontrar o menor valor de distância entre os nós a e b da matriz.

O próximo passo é Criar um novo nó u na matriz de distâncias, que será a junção dos nós a e b , cuja distância entre u e o nó i , sendo $i = a$ ou $i = b$ e $j = a$ ou $j = b$ com $i \neq j$, são dadas por:

$$d_{ui} = \frac{d_{ij}}{2} \quad (4)$$

As distâncias entre o nó u e os demais nós da matriz pode, então, ser calculada

através da fórmula:

$$d_{uk} = \frac{d_{ki} + d_{kj}}{2mn}, \quad (5)$$

onde k pode ser qualquer nó ainda não fundido à matriz, m é a quantidade de nós unidos por u e n é a quantidade de nós possivelmente unidos por k , se este for uma junção de taxa. Cabe ressaltar que i ou j já podem ser junção de taxas unidos em u . Neste caso, a nitidez do cálculo das distâncias é notória, ficando da forma (NEI; KUMAR, 2000):

$$d_{ab} = \sum_{ij} \frac{d_{ij}}{rs} \quad (6)$$

onde r e s são os números de elementos nos agrupamentos A e B , respectivamente, e d_{ij} é a distância entre o elemento i no agrupamento A e o elemento j no agrupamento B (NEI; KUMAR, 2000).

Depois disso, os nós a e b são retirados da matriz de distância e o tamanho l da matriz é reduzido em um.

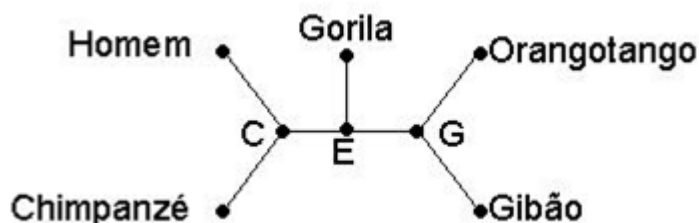
2.2.1.2 Neighbor-Joining (NJ)

Desenvolvido por (SAITOU; NEI, 1987), é um eficiente método de construção de árvores filogenéticas sem raiz, baseado no princípio da evolução mínima. Ao final de sua execução, o método fornece a topologia da árvore filogenética e também os comprimentos dos seus ramos. Este método não examina todas as possíveis topologias, mas em cada estágio de ramificação da árvore o princípio da evolução mínima é utilizado. Tem como objetivo identificar os pares mais próximos de elementos ou vizinhos (*neighbors*), de forma a minimizar o comprimento total da árvore construída. Um par de vizinhos é definido como sendo formado por dois elementos conectados por um ramo em uma árvore sem raiz bifurcada (dois ramos unidos por um nó interior) (PRADO, 2001).

SAITOU; NEI (1987) mostra que este método produz a uma boa árvore por pura adição de dados, onde a distância entre cada par de espécies é a soma dos comprimentos dos ramos que os unem na árvore.

Na Figura 12 (WEIR, 1996), homem e chimpanzé são vizinhos, mas homem e gorila não são. Se homem e chimpanzé forem combinados em um único nó, então esta combinação e gorila se tornam vizinhos. Em geral é possível determinar a topologia de uma árvore por uniões sucessivas de pares de vizinhos (PRADO, 2001).

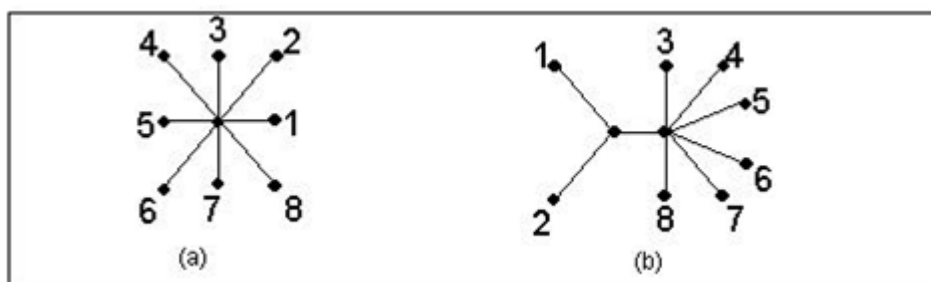
Figura 12 – Árvore sem raiz obtida através do método NJ.



Fonte: (WEIR, 1996)

Este método começa com uma árvore em forma de estrela, ilustrada na Figura 13. Os vizinhos são os pares de espécies que quando combinados resultam em uma árvore de menor comprimento total. Eles devem ser unidos para formar uma nova unidade (composta pelos dois vizinhos identificados). Este processo de identificação de vizinhos é repetido até que existam apenas três elementos (combinados) na estrutura.

Figura 13 – (a) Árvore sem raiz em formato estrela; (b) árvore sem raiz com os elementos 1 e 2 agrupados.



Fonte: (WEIR, 1996)

Segundo (SAITOU; NEI, 1987), os passos para a RAF utilizando o método NJ são:

- escolher os pares de táxons a serem unidos, substituindo-os por um novo nó simples representando o imediato ancestral comum deles;
- calcular a distância do novo nó para os demais;

A seguir será detalhada a heurística do método NJ para encontrar a árvore final:

- Primeiramente, deve-se tomar a árvore estrela contendo todos os OTUs;
- Depois, é feita a correção da matriz inicial de distância utilizando um modelo evolutivo;

- Para cada nó terminal, fazer o cálculo da divergência de todos os táxons. A distância entre dois nós (i e k) é representada por d_{ik}

$$r_i = \sum_{k=1}^N d_{ik} \quad (7)$$

- A partir daí cria-se uma nova matriz de distância M , composta pelos elementos:

$$M_{ij} = d_{ij} - \frac{(r_i + r_j)}{(N - 2)} \quad (8)$$

- Então, um novo nó u é definido e ligado a i e j e ao restante da árvore por três ramos distintos. Os comprimentos dos ramos entre o nó u e os nós i e j são definidos por:

$$v_{ui} = d_{ij}/2 + (r_i + r_j)/2(N - 2) \quad (9)$$

$$V_{ij} = d_{ij} - v_{ij} \quad (10)$$

- Depois disso, deve-se remover as distâncias para os nós i e j e decrementar a matriz N em 1;
- Agora u é um nó terminal. Caso haja mais de um nó terminal, volte ao passo 3, caso contrário, siga para o próximo passo;
- O tomando do último ramo que liga os dois táxons remanescentes é: $V_{ij} = d_{ij}$.

2.2.2 Parcimônia

O método da máxima parcimônia tem por objetivo encontrar árvores para um grupo sequencial que possa ser demonstrada em menor número de mutações prováveis. A teoria que tem por base a parcimônia, se baseia no conceito conhecido como navalha de Occam, que onde havendo diversas respostas, opta-se pela hipótese mais simples, com relação às mais complexas. Dessa forma, significa que o método de máxima parcimônia tenta encontrar a árvore que necessite de um menor número de substituições entre as sequências a serem analisadas partindo de um conjunto de todas as árvores filogenéticas. Antes da realização da análise parcimoniosa, se faz necessária a definição da parcimônia localizada no alinhamento ([KEANE, 2006](#); [HOLDER](#); [LEWIS, 2003](#)).

O modelo matemático desse método será destacada abaixo, objetivando exemplificar os conceitos supracitados, possibilitando assim dirimir dúvidas que possam surgir.

Considere um conjunto de sequências D que possui n espécies e N_{sit} sítios para cada sequência em estudo. A contabilização do número de mudanças de estados para uma determinada árvore τ é determinada pela equação abaixo (TICONA, 2008):

$$Par(\tau) = \sum_{j=1}^{n_{sit}} Par_j \quad (11)$$

onde Par_j representa o valor de parcimônia para o sítio j . Este valor é calculado pela soma das diferenças dos estados entre cada par de nós conectados nos ramos de τ .

Sendo assim, o cálculo de Par_j pode ser definido pela Equação (TICONA, 2008):

$$Par(\tau) = \sum_{(u,v) \in E} C_{v_j, u_j} \quad (12)$$

onde E representa o conjunto de ramos (v, u) da árvore τ , v_j e u_j são os estados no sítio j para as sequências correspondentes aos nós v e u , respectivamente e C_{v_j, u_j} é o custo de mudança do estado v_j para u_j no sítio j .

Convém ressaltar que o cálculo do valor de parcimônia, Par_t , explanado acima, corresponde a cada sítio de maneira individual, conforme topologia e estados dos nós (TICONA, 2008).

Para facilitar o entendimento do método de máxima parcimônia, será exibido a seguir, um método didático. Dessa forma, hipoteticamente, foi determinada a existência de quatro táxons e suas respectivas sequências de DNA, conforme descrito na tabela 2, adaptada de (HYP6LITO, 2005).

Tabela 2 – Exemplo de alinhamento de sequências de DNA.

Táxon	Sítios						
	1	2	3	4	5	6	7
A	A	A	G	A	G	T	C
B	A	G	C	C	G	T	C
C	A	G	A	G	A	T	C
D	A	G	A	T	A	T	C

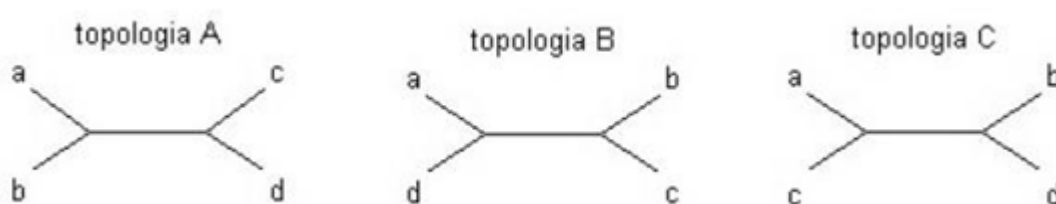
Fonte: (NETO, 2015)

Convém observar que as colunas formadas pelas sequências são conhecidas como sítios. Tais sítios podem ser classificados como informativos ou não informativos,

conforme diferenças nucleotídicas, por conseguinte, apenas os sítios 2,3,4 e 5 passaram por alterações, sendo, dessa forma classificados como informativos. Já os sítios 1,6 e 7, foram considerados como não informativos.

Conforme descrito anteriormente, a quantidade de topologias de uma determinada árvore depende do número de espécies a serem analisadas e do tipo de topologia (com ou sem raiz). Nessa perspectiva, para facilitar o entendimento, preferiu-se utilizar árvores sem raiz que determina 3 topologias possíveis para quatro Táxons, conforme mostrado na figura 14:

Figura 14 – Possíveis topologias sem raiz possíveis para 4 espécies



Fonte: (HYP6LITO, 2005)

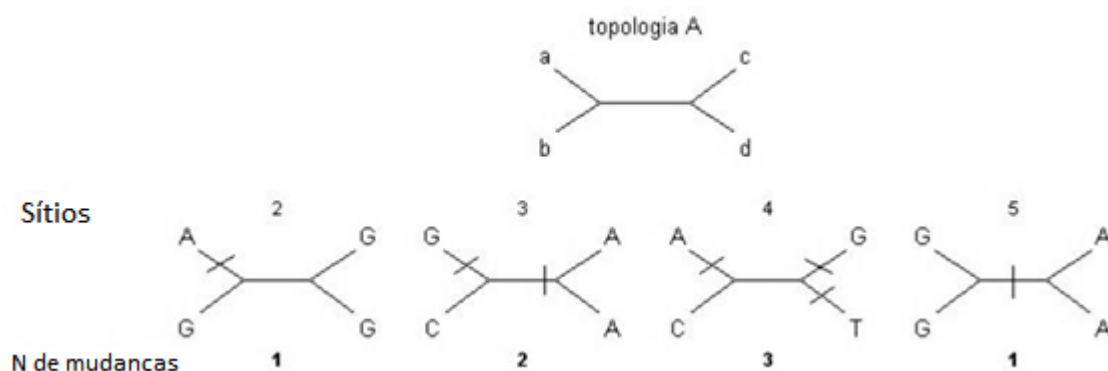
Dessa forma, a análise será feita com base nessas 3 topologias para identificar a mais parcimoniosa, ou seja, a que possui o menor número de eventos para explicar as diferenças observadas entre as sequências genéticas em estudo (FELENSTEIN, 2004).

Para calcular o número de mudanças é necessário que se identifique os estados dos nucleotídeos nos nós internos. Porém, estes são ancestrais hipotéticos, ou seja, podem assumir qualquer um dos 4 estados dos nucleotídeos possíveis: A, C, G ou T. Assim, como a ideia é descobrir a topologia com o menor número de mudanças evolutivas possíveis, o estado assumido será o que minimiza este valor (NETO, 2015) apud (FELENSTEIN, 2004)

Sendo assim, analisando a topologia A de acordo à Figura 15, pode-se afirmar que o total de mudanças é 7, pois:

- O sítio 2 sofreu apenas 1 mudança, assumindo que os nós internos possuem estados G;
- O sítio 3 sofreu 2 mudanças, sendo que os nós internos assumiram estados C e A, respectivamente;
- sítio 4 sofreu 3 mudanças, sendo que os 2 nós internos ficaram com o estado C;
- O sítio 5 sofreu apenas 1 mudança, sendo que os nós internos assumiram os estados G e A, respectivamente.

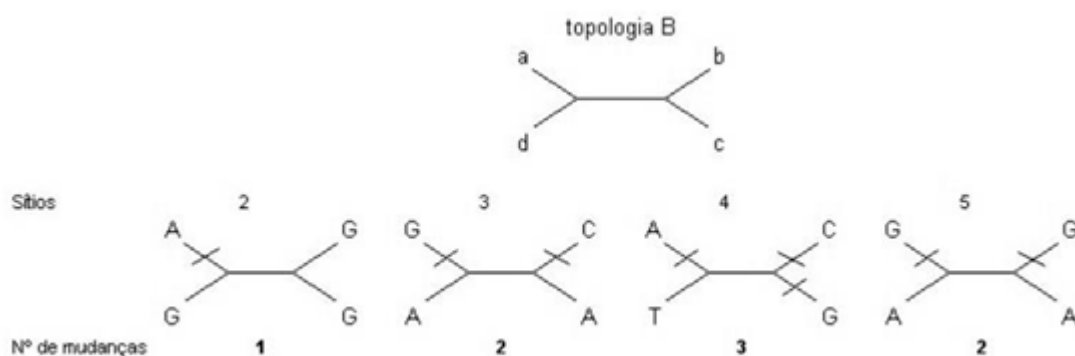
Figura 15 – Identificação da Topologia A, onde as mudanças são representadas por traços na árvore



Fonte: (HYP6LITO, 2005)

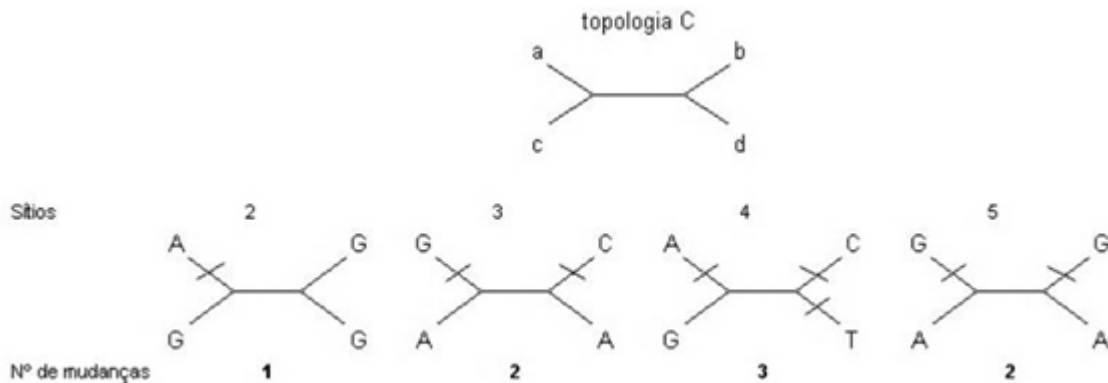
Analizando igualmente as formas topológicas B e C (Figura 16 e Figura 17), verifica-se um total de 8 mudanças para cada. Desse modo, como a topologia A detém o menor número de mudanças nucleotídicas, pode-se afirmar que a mesma explica as sequências de dados com o menor número de passos possíveis e, portanto, será a árvore mais parcimoniosa.

Figura 16 – Identificação da Topologia B, onde as mudanças são representadas por traços na árvore



Fonte: (HYP6LITO, 2005)

Figura 17 – Identificação da Topologia C, onde as mudanças são representadas por traços na árvore



Fonte: (HYP6LITO, 2005)

2.2.3 Máxima Verossimilhança

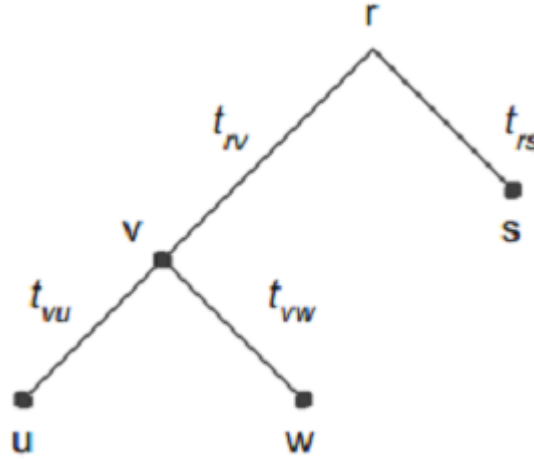
O método de máxima verossimilhança se baseia em técnicas estatísticas padronizadas para calcular as probabilidades de ocorrência de um determinado estado nos diferentes sítios de um conjunto de sequências e seus ancestrais. O método necessita de um modelo de substituição para avaliar a probabilidade de mutações particulares. Uma árvore que necessita de mais nós interiores para mostrar a filogenia em tela será avaliada como de probabilidade mais baixa. Nesse caso, o método precisa que a evolução em locais e linhagens distintas, devem ser estaticamente independentes. Apropriado para a análise de sequências distantemente relacionados (MATIOLI, 2013)

A verossimilhança determina a probabilidade $P(D|\theta)$ de um conjunto de dados D (sequências genéticas devidamente alinhadas) ajustar-se ao modelo $\theta = \tau, B, M$, onde τ é uma topologia da árvore, B é o conjunto de comprimento dos ramos de τ e M é o modelo evolutivo de substituição de sequências genéticas. Assim sendo, o objetivo principal da máxima verossimilhança é encontrar os parâmetros do modelo θ , de modo que a função de verossimilhança, definida como $L(\theta) = P(D|\theta)$, seja maximizada (TICONA, 2008; SILVA, 2016).

Para ilustrar o funcionamento da estimação de MV, será adotado um exemplo meramente ilustrativo, o qual foi desenvolvido por (TICONA, 2008) usando a ilustração da Figura 18 a seguir, onde são apresentadas três espécies (u, w e s) como fohas, e duas espécies ancestrais (v e r).

Dando continuidade ao exemplo, suponha que as espécies (u, w e s) pertencem a um conjunto de dados D, e que cada sequência possui N_{sit} sítios (colunas), tal que, u_j, w_j e s_j representam os estados das espécies u, w e s no sítio j, respectivamente. Para

Figura 18 – Árvore exemplo para o cálculo de Máxima Verossimilhança



Fonte: (TICONA, 2008)

uma análise realizada utilizando sequências de DNA, que é o caso deste exemplo, os estados citados estão definidos em um alfabeto de caracteres $\Omega = A, C, G, T$. Além disso, necessita-se supor a existência de um modelo evolutivo, que permita calcular as probabilidades de transição entre os estados.

Baseado na premissa que os sítios evoluem de maneira independente e idêntica, (FELENSTEIN, 2004) define a seguinte fórmula para calcular a verossimilhança:

$$L = \prod_{j=1}^{N_{sit}} P(D^j, \theta), \quad (13)$$

onde $P(D^j, \theta)$ representa a verossimilhança no sítio j , doravante chamada de L_j .

Como não se tem conhecimento dos nós internos, o valor de L_j será a soma de todas as probabilidades dos cenários, tendo em consideração o espaço de estados. Pelo fato de um estado em um processo markoviano depender apenas do estado anterior, o cálculo da árvore exemplo pode ser expresso conforme a equação:

$$L_j = \sum_{r_j \in \Omega} \sum_{v_j \in \Omega} \pi_{r_j} P_{r_j s_j}(t_{rs}) P_{r_j v_j}(t_{rv}) P_{v_j u_j}(t_{vu}) P_{v_j w_j}(t_{vw}), \quad (14)$$

onde r_j e v_j são os estados possíveis para os nós r e v , respectivamente; t_{ij} é o comprimento do galho que conecta os nós i e j ; π_{r_j} é a frequência da base correspondente ao estado r_j no conjunto de sequências D ; $p_{x,y}(t)$ é a probabilidade de mudança do estado x para o estado y após um tempo t (FELENSTEIN, 2004). O termo $p_{x,y}(t)$ será melhor abordado nas seções processos markovianos e modelos evolutivos.

O cálculo da verossimilhança também pode ser feito de maneira recursiva, empregando verossimilhanças condicionais de subárvores (TICONA, 2008; YANG, 2006). Nesse sentido, para o caso da árvore exemplo 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 de tal subárvore para os nós descendentes, dado que o estado do nó r no sítio j seja r_j . Logo, para o nó r , que tem os descendentes v e s , tem-se:

$$L_j^r(r_j) = \left[\sum_{v_j \in \Omega} P_{r_j v_j}(t_{rv}) L_j^v(v_j) \right] X \left[\sum_{s_j \in \Omega} P_{r_j s_j}(t_{rs}) L_j^s(s_j) \right] \quad (15)$$

e para um nó folha a , onde o estado a_j é fornecido por d , o termo $L_j^s(s_j)$ assume os seguintes valores:

$$L_j^a(X) = \begin{cases} 1, & \text{se } a_j = x \\ 0, & \text{caso contrário} \end{cases} \quad (16)$$

Assim, utilizando as equações anteriores na árvore exemplo, temos que:

$$L_j = \sum_{r_j \in \Omega} \pi_{r_j} L_j^r(r_j). \quad (17)$$

O processo realizado para a verossimilhança recursiva pode ser aplicado em qualquer nó interno de uma árvore e não apenas na raiz, mas começando pela raiz permite a propagação dos cálculos até as folhas e, de certa forma, condiciona que os cálculos sejam realizados cumulativamente das folhas para a raiz da árvore (SILVA, 2016).

Conforme mencionado anteriormente, o cálculo da verossimilhança total se realiza através do produto dos valores L_j de todos os sítios, porém, estes valores são muito pequenos de tal forma que costumam ocasionar erros de precisão numérica. Por esse motivo, a utilização de logaritmos naturais em ambos os lados da equação torna os resultados mais precisos (YANG, 1994). Aplicando o log natural em ambos os lados da equação de verossimilhança, temos:

$$\ln L = \sum_{j=1}^{N_{sit}} \ln L_j \quad (18)$$

Após obter a expressão da verossimilhança em função das probabilidades e dos comprimentos dos ramos, deve-se achar os valores de j de modo a maximizar a verossimilhança. Um dos métodos utilizados para maximizar a função L é o método de Newton-Raphson, onde primeiro deriva-se L_j , em função do número de ramos

da árvore (valores de j) e se iguala a expressão a zero. Assim, aplica-se o método de Newton-Raphson para achar os zeros do sistema. Desta forma, encontram-se os valores de j que serão substituídos na expressão de L_j para achar a máxima verossimilhança da topologia correspondente. O cálculo é realizado para diversas topologias diferentes e escolhe-se aquela com o maior valor da verossimilhança ([FELENSTEIN, 2004](#)).

2.2.4 Inferência Bayesiana

Parte do mesmo arcabouço matemático da Máxima Verossimilhança (MV), mas enquanto que nos métodos MV os parâmetros são consideradas constantes desconhecidas fixas nos métodos Bayesianos. Os parâmetros do modelo são considerados como variáveis aleatórias com distribuições estatísticas. Antes da análise dos dados são atribuídos aos parâmetros uma distribuição a priori, que é, então, combinada com os dados (ou suas verossimilhanças) para gerar a distribuição a posteriori que serve de base para as inferências sobre os parâmetros que devem ser estimados ([HUELSENBECK et al., 2002](#); [KEANE, 2006](#)).

O teorema calcula a probabilidade posterior mesclando a probabilidade a priori da topologia, com a probabilidade de determinar a distribuição dos estados nos táxons atuais com base em um modelo evolutivo. Portanto, o cálculo das probabilidades posteriores para todas as hipóteses é necessário porque este determina a escolha da melhor topologia, assim como sua integração com todas as possíveis combinações de tamanho de ramos e os valores dos parâmetros dos modelos evolutivos. Pelo fato de este cálculo ser impossível de ser feito de maneira analítica, mesmo com a inserção de buscas heurísticas, o método de amostragem de Monte Carlo via Cadeias de Markov (MCMC) é utilizado de forma a elucidar as árvores de distribuição de probabilidades posteriores([HOLDER; LEWIS, 2003](#)).

As distribuições de probabilidades representam o conhecimento a respeito da topologia, dos comprimentos dos ramos e dos parâmetros de substituição das bases. O cálculo da probabilidade de ocorrência de uma árvore é feito através da combinação da verossimilhança da árvore com a probabilidade a priori da mesma normalizada pela probabilidade dos dados terem ocorrido. O resultado é uma distribuição de probabilidade posterior que permite a escolha da árvore com maior chance de estar correta ([HYPÓLITO, 2005](#)).

Em suma, no âmbito do embasamento filogenético, o teorema de Bayes informa a probabilidade de que determinada ramificação da árvore esteja correta (probabilidade posterior), baseadas em várias gerações, sumarizadas ao final da análise e contabilizadas as suas frequências ([HUELSENBECK et al., 2002](#)).

A Inferência Bayesiana calcula a probabilidade de uma árvore ser verídica

através das sequências de alinhamento de DNA, determinada pela equação a seguir (GONÇALVES, 2008):

$$f(\tau|X) = \frac{f(X|\tau_i)f(\tau_i)}{\sum_{j=1}^{B(S)} f(X|\tau_j)f(\tau_i)}, \quad (19)$$

onde:

- $f(\tau|X)$ é a probabilidade posterior da i -ésima árvore (τ_i);
- $f(X|\tau_i)$ é a Máxima Verossimilhança de (τ_i);
- $f(\tau_i)$ corresponde ao valor a priori (τ_i);
- O denominador é o somatório de todas as possibilidades de árvores, de acordo com a quantidade de espécies S .

2.2.4.1 Método de Monte Carlo via Cadeias de Markov

Calcular a probabilidade posterior a partir da Equação de Inferência Bayesiana é considerada complexa e não pode ser calculada analiticamente, uma vez que os valores para os tamanhos dos galhos v_l e para as substituições de DNA θ , ao serem utilizados na inferência Bayesiana, corresponderão a todo universo de valores possíveis, como ilustra a equação a seguir (GONÇALVES, 2008):

$$f(X|\tau_i) = \int_v \int_\theta f(X|\tau_i, v, \theta) f(v, \theta) dv d\theta \quad (20)$$

A probabilidade posterior pode ser aproximada utilizando o método de Monte Carlo via Cadeias de Markov (MCMC), que amostra árvores de acordo a uma função de distribuição de probabilidade. O objetivo principal do método é construir uma cadeia de Markov com distribuição estacionária igual à distribuição posterior de interesse. Assim, esta cadeia é iniciada de um ponto arbitrário no universo de valores disponíveis, e é gerada sucessivamente até ser alcançado um valor estacionário sendo que a probabilidade posterior é coletada a cada geração (GRIMMENT; STIRZAKER, 1992).

De maneira geral o método consiste em obter o valor estimado de uma integral, sabendo que a média de uma função $g(y)$ pode ser obtida por (TRIOLA, 2014):

$$E(g(Y)) = \int_D g(y)p(y)dy = \int_D f(y)dy, \quad (21)$$

onde y é uma variável aleatória do domínio D e $p(y)$ a sua densidade de probabilidade.

Deste modo, o problema do cálculo da integral passa a ser a estimação da média $E(g(X))$, cujo procedimento padrão é fazer a coleta de uma amostra aleatória, com densidade uniforme em todo D , que corresponde ao intervalo $[a, b]$ e calcular a média dessa amostra, como descrito na equação a seguir (NETO, 2015):

$$= (b - a)E(g(X)) \quad (22)$$

Portanto, em uma amostra de tamanho n , a estimativa é dada por:

$$= (b - a) \frac{1}{n} \sum_{i=1}^n g(x_i) \quad (23)$$

A equação acima pode ser generalizada de modo a integrar todo o domínio geral, e não apenas o intervalo a, b , conforme descrito na equação a seguir:

$$\mu = (b - a) \frac{1}{n} \sum_{i=1}^n f(X_i), \quad (24)$$

onde μ representa a estimativa de uma função $f(x_i)$ de interesse.

(TRIOLA, 2014) afirma que quando as amostras X_i são independentes, a lei dos grandes números certifica que μ pode ser obtido tão exato quanto desejado à medida que o valor de n cresce, que por sua vez é uma variável sob controle do analista.

No entanto, as amostras X_i não precisam ser totalmente independentes, mas podem ser geradas através de algum mecanismo para determinar X_i a partir da função de interesse. Assim sendo, resolve-se esta situação através de uma cadeia de Markov tendo f como a distribuição estacionária (GRIMMENT; STIRZAKER, 1992).

Se analisarmos o conceito de processos estocásticos, veremos que uma cadeia de Markov se dá quando uma sequência de variáveis aleatórias X_0, X_1, X_2, \dots é amostrada a cada instante de tempo $t \geq 0$ sob a distribuição $P(X_{t+1}|X_t)$. Baseado na afirmação anterior, o estado no instante $t + 1$ (X_{t+1}) depende apenas do estado imediatamente anterior. Assim sendo, $P(.|.)$ é chamada somente de transição da cadeia e se assume que essa distribuição é homogênea no tempo, pois não depende de t .

Cabe salientar que o estado inicial (X_0) influencia no comportamento da cadeia. Sendo assim, se espera que uma cadeia de Markov esqueça gradualmente o valor do estado inicial e convirja para uma distribuição estacionária, independente do tempo ou do estado inicial. O procedimento utilizado para isso é chamado de *burn in*, o qual adota uma quantidade n de amostras, que são grandes o suficiente para descartar m

amostras iniciais e pode ser demonstrada pela equação a seguir (GONÇALVES, 2008):

$$\mu = \frac{1}{n-m} \sum_{i=m+1}^n f(X_i). \quad (25)$$

2.3 Processos de Markov

Os processos markovianos são processos estocásticos onde a probabilidade de um determinado estado em um determinado instante depende apenas do estado imediatamente anterior, independente dos estados passados. Por isso, os processos markovianos são classificados como processos sem memória (HILLIER; LIEBERMAN, 2001). Um processo estocástico pode ser classificado como markoviano se:

$$P\{X(t_{k+1}) \leq x_{k+1} \mid X(t_k) = x_k, X(t_{k-1}) = x_{k-1}, \dots, X(t_1) = x_1, X(t_0) = x_0\} = P\{X(t_{k+1}) \leq x_{k+1} \mid X(t_k) = x_k\}$$

para $t_0 \leq t_1 \leq \dots \leq t_k \leq t_{k+1} = 0, 1, \dots$ e toda sequência $k_0, k_1, \dots, k_{t-1}, k_t, k_{t+1}$

A expressão acima pode ser traduzida por: a probabilidade de qualquer evento futuro dado qualquer evento passado e o estado presente $X(t_k) = x_k$, é independente do evento passado e depende somente do estado presente. Em termos mais simples: um Processo Estocástico é dito ser um Processo Markoviano se o estado futuro depende apenas do estado imediatamente anterior e não dos estados passados (LEVIN et al.,).

Uma maneira conveniente de representar probabilidades de transição para n passos é utilizando a forma matricial (HILLIER; LIEBERMAN, 2001):

<i>Estado</i>	0	1	...	M
0	$P_{00}^{(n)}$	$P_{01}^{(n)}$...	$P_{0M}^{(n)}$
1	$P_{10}^{(n)}$	$P_{11}^{(n)}$...	$P_{1M}^{(n)}$
—
M	$P_{M0}^{(n)}$	$P_{M1}^{(n)}$...	$P_{MM}^{(n)}$

(26)

A matriz $P^{(n)}$ é denominada Matriz de Transição de Passo n . Quando $n = 1$, a matriz é denominada apenas Matriz de Transição.

Existem duas categorias de processos de Markov: processos de tempo discreto, que assumem valores inteiros não negativos para t e processos de tempo contínuo, onde $t \in [0, \infty]$. Em ambas categorias os estados são caracterizados por números inteiros não negativos, definidos a partir dos valores que a variável X pode assumir.

Sejam dois estados i e j , o instante de tempo u tal que $0 \leq u \leq t$, e $X(t)$ e $X(u)$ as variáveis aleatórias que contém os números dos estados nos tempos t e u ,

respectivamente, a probabilidade de transição $p_{ij}(u, t)$ é definida pela probabilidade condicional:

$$p_{ij}(u, t) = P(X(t) = j | X(u) = i), \quad (27)$$

sendo $i, j = 0, 1, 2, \dots$, e $P_{ij}(t, t) = 1$, se $i = j$ e $P_{ij}(t, t) = 0$, se $i \neq j$.

Assim sendo, determinar a função de transição representada pela equação anterior é o ponto de partida para resolução das cadeias de Markov (GRIMMENT; STIRZAKER, 1992).

Com base no teorema de probabilidade total expressamos os dados A , B_1 e B_2 , tendo B_1 e B_2 como mutuamente exclusivos:

$$P(A) = \sum_{\forall i} P(A \wedge B_1) = \sum_{\forall i} P(A|B_1)P(B_1), \quad (28)$$

Ao condicionarmos $[X(t) = j | X(s) = i]$ a $[X(u) = r]$ para algum $s \leq u \leq t$ e considerando $A = [X(t) = j | X(s) = i]$ e $B = [X(u) = r | X(s) = i]$, podemos encontrar a equação de Chapman-Kolmogorov. As equações de Chapman-Kolmogorov fornecem o cálculo de probabilidade de transição em n passos. Estas equações mostram que as probabilidades de transição de n etapas podem ser obtidas recursivamente a partir das probabilidades de transição de uma etapa (DURAND, 2013).

$$P_{ij}(s, t) = \sum_{\forall r} P_{ir}(s, u) p_{rj}(u, t) \quad (29)$$

A equação de Chapman-Kolmogorov pode ser reescrita na forma matricial, conforme descrição a seguir, juntamente com as condições pré-estabelecidas (LEVIN et al.,):

$$P(s, t) = P(s, u)P(u, t) \quad s \leq u \leq t \quad (30)$$

Quando adicionamos uma variação no instante t , onde $s \leq t \leq \Delta t + t$, a equação anterior fica da seguinte forma:

$$P(s, t + \Delta t) = P(s, t)P(t, t + \Delta t) \quad (31)$$

Ao subtrairmos $P(s, t)$ dos dois termos da igualdade e posteriormente colocando o mesmo em evidência, encontramos a seguinte equação:

$$P(s, t + \Delta t) - P(s, t) = P(s, t)[P(t, t + \Delta t) - 1] \quad (32)$$

Dividindo a equação anterior pela variação de tempo (Δt) e tomando o limite com $\Delta \rightarrow 0$, tem-se a equação diferencial matricial, que é representada por (LEVIN et al.,):

$$\lim_{\Delta t \rightarrow 0} \frac{P(s, t + \Delta t) - P(s, t)}{\Delta t} = P(s, t) \lim_{\Delta t \rightarrow 0} \frac{P(t, t + \Delta t) - 1}{\Delta t} \quad (33)$$

$$\frac{\partial P(s, t)}{\partial t} = P(s, t)Q(t) \quad s \leq t \quad (34)$$

Sabendo que a cadeia é homogênea, a função de transição não depende dos valores de s e t , mas somente da diferença entre eles $\tau = (t - s)$. Nesse caso, podemos reescrever a equação anterior como:

$$\frac{\partial P(\tau)}{\partial \tau} = P(\tau)Q \quad (35)$$

Assim sendo, de acordo com (LEVIN et al.,), considerando que a matriz de transição seja igual à matriz identidade, isto é, adotando as seguintes condições iniciais:

$$p_{ij}0 \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases} \quad \text{ou seja, } P(0) = I \quad (36)$$

é dada por:

$$P(\tau) = e^{Q\tau}$$

Para este trabalho foram utilizadas cadeias de Markov de tempo contínuo que utiliza a equação anterior para gerar uma matriz de probabilidades de transição em determinado intervalo de tempo $P(\tau)$ através da matriz de taxas de transição Q (geradora infinitesimal), cuja soma dos elementos de cada linha é zero. Para esta matriz, cada dado representa a taxa infinitesimal de uma determinada mudança de estado do processo markoviano, onde $P(j|i, t) \equiv P(X_t = j | X_0 = i)$ representa a probabilidade de transição do estado i para o estado j no tempo t . A partir da matriz geradora infinitesimal, podemos determinar todas as probabilidades de transição de um estado para o outro, em qualquer tempo (CYBS, 2009).

De acordo com (FELENSTEIN, 2004), podemos decompor a matriz Q em seus autovetores e autovalores, tendo a seguinte solução para a equação diferencial:

$$P(\tau) = e^{Q\tau} = ADA^{-1} \quad (37)$$

Onde D é a matriz diagonal cujos elementos são autovalores de Q , A é a matriz cujas colunas são os autovetores diretos de Q e $P(t)$ é a matriz de probabilidade de transição no tempo t .

Outro ponto relevante que precisa ser mencionado como pré-requisito para apresentar mais adiante os modelos evolutivos, é a distribuição estacionária. Assim, para um processo de Markov com estados finitos e com sua matriz geradora infinitesimal Q , existe um vetor estacionário (p_0) tal que $(p_0)Q = 0$. Além disso, se este vetor é único, tem-se que, para qualquer i (GRIMMENT; STIRZAKER, 1992):

$$\lim_{t \rightarrow \infty} P(X_t = j | X_0 = i) = P_{0j} \quad (38)$$

Onde P_{0j} é a j -ésima componente do vetor P_0 .

A afirmação acima é importante, pois os modelos evolutivos tomam a distribuição inicial dos processos como a própria distribuição estacionária, pelo fato de as OTUs analisadas estarem evoluindo há bastante tempo. Assim sendo, as frequências dos estados já estariam muito próximas da distribuição estacionária (CYBS, 2009).

2.4 Modelos evolutivos

Existe uma série de diferentes modelos de evolução de DNA baseados em cadeias de Markov que diferem entre si pela quantidade de parâmetros que cada modelo utiliza. Os modelos descrevem a evolução do DNA como uma sequência de estados discretos e as taxas relativas às diferentes modificações dos estados. As taxas de mudança são descritas por uma matriz de probabilidades. Ao expressarmos os modelos em termos de taxas de variação, podemos evitar um grande número de parâmetros para cada ramo da árvore filogenética, visto que basta um estado inicial (sequência) e uma matriz de probabilidades (BOTTU et al., 2009; DURAND, 2013).

Segundo (DURAND, 2013), os modelos de substituição sequência são usados para atender a uma ampla gama de questões que surgem em evolução molecular:

1. Correção de múltiplas substituições
2. Simulação de evolução de sequências
3. Estimar taxas de evolução
4. Derivar as taxas de substituições de matrizes
5. Estimar a probabilidade de observar um par de nucleotídeos alinhados, dado um modelo evolutivo.

No tocante da filogenia podemos apenas observar o estado atual do processo não tendo como mensurar quantas mudanças ocorreram na história evolutiva de um determinado sítio. Se o estado atual um sítio for A, e um estado passado foi C, não

podemos afirmar que a mudança ocorreu diretamente de A para C, podendo ter havido mudanças entre esses estados. Seguindo esse mesmo raciocínio, se um sítio tem seu estado atual como A e um estado passado A, não podemos afirmar que não houveram mudanças entre o estado passado e o estado atual (CYBS, 2009).

O espaço de estados destes processos é definido de acordo às quatro bases do DNA, Adenina (A), Guanina (G), Citosina (C) e Timina (T), denotado por $E = A, C, G, T$. As características químicas destas bases definiram a criação de dois grupos: as purinas, que contém as bases A e G, e as pirimidinas, C e T (CYBS, 2009). Quando temos uma mudança do tipo purina-purina ou pirimidina-pirimidina ($A \leftrightarrow G$ ou $C \leftrightarrow T$), essa mudança é chamada de transição. Quando a mudança ocorre entre uma purina e uma pirimidina ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$ ou $G \leftrightarrow T$) ela é chamada de transversão.

Segundo NETO (2015), cada modelo possui suas peculiaridades de acordo as definições de cada autor, porém todos possuem a sua matriz de taxas infinitesimais e o seu vetor de distribuição estacionária. Neste caso, a representação de sua matriz de taxas infinitesimais pode ser padronizada como segue:

$$Q = \begin{pmatrix} -q_A & q_{A,G} & q_{A,C} & q_{A,T} \\ q_{G,A} & -q_G & q_{G,C} & q_{G,T} \\ q_{C,A} & q_{C,G} & -q_C & q_{C,T} \\ q_{T,A} & q_{T,G} & q_{T,C} & -q_T \end{pmatrix} \quad (39)$$

onde $q_{i,j}$ é a taxa instantânea de mutação da base i para a base j , e os elementos da diagonal principal são definidos de tal forma que a soma de todos os elementos da linha seja igual a zero, ou seja, $q_i = \sum_{i \neq j} q_{ij}$, para $i, j \in \Omega$

Assim, como a matriz de taxas infinitesimais, o vetor de distribuição estacionária também possui uma forma que pode-se afirmar como genérica. Como nenhuma das taxas $q_{i,j}$ dos modelos apresentados são nulas, o processo possui uma distribuição estacionária denotada pela equação (CYBS, 2009):

$$p_0 = (\pi_A, \pi_G, \pi_C, \pi_T) \quad (40)$$

onde π_i é a proporção do nucleotídeo i na molécula de DNA.

Um ponto importante que precisa ser abordado é o fato de que cada modelo possui uma matriz Q , de acordo com suas características e, consequentemente, uma matriz P , que é calculada da maneira vista na seção anterior.

2.4.1 O modelo Jukes Cantor

O modelo de substituição mais simples é o Jukes-Cantor (JC69), de 1969, que assume frequências iguais entre as bases ($\pi A = \pi C = \pi G = \pi T = 1/4$) e todas as substituições entre bases ocorrem em igual taxa (α), tendo como consequência que a probabilidade de mudança de estado é dada por $\lambda = 3\alpha$. Existe apenas um parâmetro para esse modelo. A matriz de de taxas infinitesimais do modelo JC69 é ilustrada abaixo:([JUKES; CANTOR, 1969](#))

$$\begin{array}{c|cccc}
 & A & C & G & T \\
 \hline
 A & -3\alpha & \alpha & \alpha & \alpha \\
 C & \alpha & -3\alpha & \alpha & \alpha \\
 G & \alpha & \alpha & -3\alpha & \alpha \\
 T & \alpha & \alpha & \alpha & -3\alpha
 \end{array} \quad (41)$$

Matriz de taxas infinitesimais de transição do modelo JC69.

Como já foi mencionado, em posse da matriz Q , pode-se encontrar a matriz P de probabilidades de substituição para este modelo. Para isto basta solucionar a Equação de transição, a qual tem como solução a decomposição da matriz Q em seus autovalores e autovetores. Com este resultado encontra-se a matriz de probabilidades de transição, onde $\alpha_t = \frac{1}{4}(1 - e^{(-4\alpha t)})$:

$$\begin{array}{c|cccc}
 & A & C & G & T \\
 \hline
 A & 1 - 3\alpha_t & \alpha_t & \alpha_t & \alpha_t \\
 C & \alpha_t & 1 - 3\alpha_t & \alpha_t & \alpha_t \\
 G & \alpha_t & \alpha_t & 1 - 3\alpha_t & \alpha_t \\
 T & \alpha_t & \alpha_t & \alpha_t & 1 - 3\alpha_t
 \end{array} \quad (42)$$

2.4.2 Modelo Kimura dois parâmetros (K2P)

O modelo proposto por Kimura assume frequências iguais entre as bases, mas diferentemente do modelo JC69 a taxa de substituição entre transições ($A \leftrightarrow G$ ou $C \leftrightarrow T = \alpha$) ocorrem em uma taxa maior que em transversões ($A \leftrightarrow C$, $A \leftrightarrow T$, $G \leftrightarrow C$ ou $G \leftrightarrow T = \beta$)([YANG, 2006](#)).

Deve-se observar que, para qualquer nucleotídeo, pode existir uma troca a uma taxa α ou β . O que significa respectivamente uma transição ou transversão. A razão de transição/transversão do modelo é denotada por r e é igual a α/β . Vale ressaltar que o

modelo JC69 é um caso particular do modelo K2P, onde $\alpha = \beta$ e $r = 1/2$ (HYP6LITO, 2005). A matriz de taxas instantânea do modelo K80 é ilustrada abaixo:

$$\begin{array}{c|cccc}
 & A & C & G & T \\
 \hline
 A & -\alpha - 2\beta & \alpha & \beta & \beta \\
 C & \alpha & -\alpha - 2\beta & \beta & \beta \\
 G & \beta & \beta & -\alpha - 2\beta & \alpha \\
 T & \beta & \beta & \alpha & -\alpha - 2\beta
 \end{array} \quad (43)$$

Matriz de taxas instantâneas do modelo K80

Assim, como no modelo JC69, a distribuição de equilíbrio das bases é homogênea e as probabilidades de mutação podem ser obtidas da mesma maneira que no modelo de Jukes-Cantor. Assim sendo, podemos expressar para um comprimento de ramo t e dado uma base i :

$$p_{ii}(t) = 0.25 + 0.25e^{-4\beta t} + 0.5e^{-2(\alpha+\beta)t}. \quad (44)$$

A expressão para $p_{ij}(t)$, com $i \neq j$, depende das escolhas de i e j . Caso i e j sejam uma purina (respectivamente Pirimidina), então a probabilidade $p_{ij}(t)$ é dada por:

$$p_{ij}(t) = 0.25 + 0.25e^{-4\beta t} - 0.5e^{-2(\alpha+\beta)t}. \quad (45)$$

Caso o nucleotídeo i seja uma Purina (respectivamente Pirimidina) e j seja uma Pirimidina (respectivamente Purina), calcula-se a probabilidade por:

$$p_{ij}(t) = 0.25 + 0.25e^{-4\beta t}. \quad (46)$$

2.4.3 O modelo F81 (Felsenstein 1981)

O modelo F81, assume frequências diferentes entre as bases ($\pi A \neq \pi C \neq \pi G \neq \pi T$) e igual taxa de substituição entre os nucleotídeos, sejam transições ou transversões (FELENSTEIN, 2004).

A matriz de taxas infinitesimais do modelo F81 pode ser considerada como um

caso particular do modelo JC69 e é definida por:

$$Q_{F81} = \begin{matrix} & \begin{matrix} * & A & G & C & T \end{matrix} \\ \begin{matrix} A \\ G \\ C \\ T \end{matrix} & \begin{matrix} b_1 & \alpha\pi G & \alpha\pi C & \alpha\pi T \\ \alpha\pi A & b_2 & \alpha\pi C & \alpha\pi T \\ \alpha\pi A & \alpha\pi G & b_3 & \alpha\pi T \\ \alpha\pi A & \alpha\pi G & \alpha\pi C & b_4 \end{matrix} \end{matrix}, \quad (47)$$

onde b_k é a constante tal que a soma dos elementos da linha k é 0, para todo $k \in \{1, \dots, 4\}$, π_i representa a proporção da base i na amostra e α a taxa de mutação, que é a mesma para todos os estados. O vetor de probabilidade inicial desse modelo é dado por: $p_0 = (\pi_A, \pi_G, \pi_C, \pi_D)$.

Com essa matriz, podemos montar um sistema de equações, o qual tem como solução as probabilidades de transição a seguir (NETO, 2015):

$$P_{ij}(t) = \begin{cases} \pi_j(1 - e^{-kt}) & \text{se } i \neq j \\ e^{-kt} + \pi_j(1 - e^{-kt}) & \text{se } i = j \end{cases} \quad (48)$$

2.4.4 O modelo HKY (Hasegawa et al. 1984, 1985)

Segundo (CYBS, 2009), o modelo HKY é um caso particular do modelo TN93. Neste modelo temos que $\frac{\alpha_R}{\alpha_Y} = \frac{\pi_R}{\pi_Y}$, e, portanto, as probabilidades dos eventos do tipo I em cada grupo são proporcionais às frequências de bases daquele grupo. Além disso, a razão $\frac{\alpha_l}{\pi_l}$, com $l \in \{R, Y\}$, é igual para purinas e pirimidinas. Desta forma, a taxa instantânea de transição para a base j , pertencente à classe l , é dada por:

$$\alpha_l \frac{\pi_i}{\pi_j} + \gamma \pi_j = \left(\frac{\alpha_l}{\pi_l} + \gamma \right) \pi_j = \beta \pi_j \quad (49)$$

em que $l \in \{Y, R\}$; Y representa as pirimidinas e R , as purinas.

A matriz de taxas infinitesimais do modelo HKY é dada por:

$$Q_{HKY} = \begin{matrix} & \begin{matrix} * & A & G & C & T \end{matrix} \\ \begin{matrix} A \\ G \\ C \\ T \end{matrix} & \begin{matrix} d_1 & \beta\pi G & \gamma\pi C & \gamma\pi T \\ \beta\pi A & d_2 & \gamma\pi C & \gamma\pi T \\ \gamma\pi A & \gamma\pi G & d_3 & \beta\pi T \\ \gamma\pi A & \gamma\pi G & \beta\pi C & d_4 \end{matrix} \end{matrix}, \quad (50)$$

2.4.5 O modelo TN93 (Tamura and Nei 1993)

O modelo proposto por Tamura e Nei propõe um modelo que é uma extensão do modelo K80 (probabilidades distintas para transições e transversões) para a situação em que a distribuição das bases não é homogênea (TAMURA; NEI, 1993)

O modelo TN93 tem a seguinte matriz de taxas infinitesimais:

$$Q_{HKY} = \begin{array}{c|cccc} * & A & G & C & T \\ \hline A & c_1 & \alpha_R \frac{\pi_G}{\pi_R} + \gamma\pi_G & \gamma\pi_C & \gamma\pi_T \\ G & \alpha_R \frac{\pi_A}{\pi_R} + \gamma\pi_A & c_2 & \gamma\pi_C & \gamma\pi_T \\ C & \gamma\pi_A & \gamma\pi_G & c_3 & \alpha_Y \frac{\pi_T}{\pi_Y} + \gamma\pi_T \\ T & \gamma\pi_A & \gamma\pi_G & \alpha_Y \frac{\pi_C}{\pi_Y} + \gamma\pi_C & c_4 \end{array}, \quad (51)$$

2.4.6 O modelo *General Time Reversible* (GTR)

O modelo GTR é o modelo mais geral, independente possível. Os parâmetros do modelo GTR consistem em um equilíbrio do vetor de frequências de base, dando a frequência com que cada base ocorre em cada local, e a matriz de taxa (FELENSTEIN, 2004).

O modelo GTR é reversível no tempo, o que significa dizer que a probabilidade de se começar com j em uma das pontas de um ramo da árvore filogenética e de se terminar com k na outra é a mesma probabilidade de que ocorra ao contrário. Isso pode ser expresso através da equação: $\pi_x P_{x,y}(t) = \pi_y P_{y,x}(t)$, onde $P_{x,y}(t)$ é a probabilidade de mudança do estado x para o estado y no tempo t . A reversibilidade no tempo é uma propriedade matemática conveniente, apesar de não ser fundamentada em razões biológicas e se ajusta bem a dados reais.(FELENSTEIN, 2004).

Este modelo atribui à evolução da sequência da molécula de DNA a uma cadeia de Markov de tempo contínuo, com distribuição inicial dada por $p_0 = (\pi_A, \pi_G, \pi_C, \pi_T)$ e tem a seguinte matriz infinitesimal:

$$Q_{GTR} = \begin{array}{c|cccc} * & A & G & C & T \\ \hline A & f_1 & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ G & \alpha\pi_A & f_2 & \delta\pi_C & \epsilon\pi_T \\ C & \beta\pi_A & \delta\pi_G & f_3 & \eta\pi_T \\ T & \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & f_4 \end{array} \quad (52)$$

2.4.7 Taxas evolutivas de heterogeneidade

Os modelos evolutivos vistos até então determinam que os diferentes sítios em uma sequência evoluem da mesma maneira e com a mesma taxa. Esta suposição não é válida em dados reais, visto que a taxa de mutação pode variar entre os sítios de uma mesma sequência. As mutações em sítios distintos podem ser fixadas em taxas

diferentes, devido aos seus papéis na estrutura e na função do gene e, assim, sofrerem diferentes pressões seletivas que atuam sobre elas (YANG, 2006).

Assim sendo, não levar essas características em consideração pode levar a grandes impactos na análise filogenética, pois estas podem aumentar o número de substituições, as quais não podem ser detectadas por uma simples comparação dos dados das sequências (YANG, 2006). Desse modo, a inserção da variação de taxas de heterogeneidade entre sítios no modelo levou a um novo conjunto de modelos que proporcionou um melhor ajuste dos dados observados permitindo assim melhorias significativas para o resultado filogenético.

De acordo com (CYBS, 2009), os modelos de heterogeneidade informam como as diferentes taxas de mutação estão distribuídas entre os sítios. Eles não descrevem as características do processo evolutivo das sequências e, por isso, há a necessidade de aliar o modelo com a matriz de taxas de mutação Q e o vetor de probabilidades iniciais p_0 . No sentido da heterogeneidade, uma abordagem sensata é usar uma distribuição estatística para modelar a variação da taxa e que, em geral, são usadas duas abordagens estatísticas, a distribuição discreta e a distribuição contínua gama. Ambas as distribuições são abordadas nas subseções seguintes (SILVA, 2016).

2.4.7.1 Modelos com proporções de Sítios Invariantes

O modelo de sítios invariantes, proposto por (HASEGAWA et al., 1985) é um dos modelos mais utilizados. Esse modelo é denotado por +I (GTR +I, JC69 +I). Ele divide os sítios de uma sequência em dois grupos. O primeiro grupo segue o processo determinado pela matriz de transição com taxa de mutação geral (μ_1). O segundo grupo, invariante, com taxa de mutação μ_0 . A média da taxa de mutação sobre todos os sítios é 1, ou seja, $q_1\mu_1 + q_0\mu_0 = q_1\mu_1 = 1$. Dessa forma, $\mu_1 = q_1^{-1} = (1 - q_0)^{-1}$.

Observa-se que quando um sítio não é constante (tem bases diferentes nas diferentes sequências) ele não pode pertencer à classe dos sítios invariantes. Porém, se um sítio não pertence à classe de sítios invariantes, ainda pode existir a possibilidade de o mesmo ser constante (CYBS, 2009).

Existem diversos trechos nas sequências que funcionam como sítios invariantes, pela sua alta conservação. Esse fato decorre comumente de alguma função específica que a sequência desempenha e que seria perdida caso o sítio fosse alterado.

2.4.7.2 Distribuição gama para taxas de mutação

Outra abordagem para variar as taxas de mutação ao longo dos sítios é assumir que as taxas seguem uma distribuição contínua. Existem diversas distribuições, como por exemplo a log-natural, mas a distribuição mais utilizada para esse fim é a gama

(G ou Γ) (CYBS, 2009). Apesar de não existir nenhuma razão biológica para a sua utilização, o uso da gama decorre de sua versatilidade.

Considere uma variável aleatória X , tal que $X \sim \Gamma(a, b)$ então a função densidade de probabilidade de X é dada por:

$$f_{(X)} = \begin{cases} \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}, & x \geq 0 \\ 0, & \text{caso contrário} \end{cases} \quad (53)$$

onde a função gama $\Gamma(a)$ é definida por:

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt. \quad (54)$$

Assim como no caso das taxas de distribuição discreta, não sabemos a taxa de mudança de cada sítio. Portanto, para calcular a probabilidade dos dados devemos calcular uma média sobre todas as taxas possíveis. Como a distribuição é contínua o resultado é obtido mediante integração de 0 a infinito (CYBS, 2009), ou aplicando o somatório sobre todas as categorias para obter um valor aproximado.

Em (YANG, 1994) mostra-se que existe uma alternativa de qualidade para os dois modelos citados acima, o modelo da distribuição gama discretizada. Ele apresenta uma boa aderência ao modelo de distribuição gama, além de um custo computacional compatível com o de taxas discretas. Definido pela simbologia $+\Gamma_d$ e utiliza C categorias, todas com probabilidade $1/C$, para aproximar a distribuição gama. Na prática, o usuário deve escolher com quantas categorias deseja trabalhar, de maneira a estimar apenas um parâmetro (CYBS, 2009).

A seguir ilustra-se como se incorpora a heterogeneidade de taxa no cálculo da máxima verossimilhança visto anteriormente.

Neste sentido, as taxas de heterogeneidade sobre sítios serão incorporadas à função de verossimilhança (Gama) por resultar em inferências mais realistas (TICONA, 2008).

As taxas de heterogeneidade são incorporadas ao modelo através de um vetor $W = [w_1, w_2, \dots, w_{N_{sit}}]^t$, no qual w_j representa a taxa evolutiva do sítio j . Nesse caso, o cálculo da verossimilhança é executado da mesma maneira do cálculo da verossimilhança recursiva, tendo porém o comprimento de ramo t_{ij} sendo multiplicado por w_j (YANG, 1994).

Já a inserção da taxa de heterogeneidade Gama no modelo, assume uma variável aleatória w_j obtida de uma distribuição Gama contínua (Γ). Assim, a verossimilhança

para um sítio j é dada pela equação:

$$L_j = \int_0^\infty P(D^{(j)}|\theta, w_j = x)f(x)dx, \quad (55)$$

onde f é a função de densidade de probabilidade com distribuição Γ , e $P(D^{(f)}|\theta, w_j = x)$ é a verossimilhança do sítio j condicionado à taxa x . Em termos computacionais, o cálculo da integral é um processo dispendioso, por isso faz-se o uso de uma função discreta Γ para aproximar tal valor (FELENSTEIN, 2004):

$$L_j = \int_0^\infty P(D^{(f)}|\theta, w_j = x)f(x)dx \approx \sum_{k=1}^{N_{cat}} \rho_k P(D^{(j)}|\theta, w_j = x_k), \quad (56)$$

onde a distribuição Γ para as taxas dos sítios é discretizada em $k = 1 \dots N_{cat}$ categorias, x_k corresponde a taxa de evolução da categoria k e ρ_k é a probabilidade da categoria k . A equação anterior pode ser reescrita da seguinte maneira (FELENSTEIN, 2004):

$$\sum_{k=1}^{N_{cat}} \sum_{r_j \in \Omega} \rho_k \pi_{r_j} L_j^r(r_j x_k), \quad (57)$$

onde $L_j^r(r_j, w_j = x_k)$ é obtida da mesma forma que $L_j^r(r_j)$ da equação de verossimilhança recursiva, multiplicando t_{rv} e t_{rs} pela taxa da categoria x_i (FELENSTEIN, 2004).

A incorporação de taxa de heterogeneidade Gama é mais vantajosa porque os valores w_j são obtidos a partir da distribuição *varGamma*, a qual é composta por dois parâmetros: α parâmetro de forma e β parâmetro de escala. Em termos práticos, apenas o parâmetro α é empregado, visto que β é fixado em $1/\alpha$. De outro modo, o cálculo da Equação torna-se mais lento utilizando este tipo de taxa, pois este é realizado para as N_{cat} categorias empregadas (NETO, 2015).

2.5 Métodos e Ferramentas para a escolha de modelos evolutivos

Existem várias estratégias para escolha do modelo evolutivos no contexto da filogenia. Dentre as estratégias disponíveis para a seleção do modelo podemos citar *Hierarchical Likelihood Ratio Test*, o *Akaike Information Criteria*, o *Bayesian Information Criteria*, a *Decision Theory* e o *Bayes Factor*.

2.5.1 Hierarchical likelihood ratio test (HLRT)

A estratégia do HLRT consiste na realização de testes de razão de probabilidade de pares em uma sequência específica até que convirja para um modelo final. Os LRTs comparam os log do likelihood do modelo nulo (0) e do modelo alternativo (1). Se o valor P associado é menor do que o limiar pré-definido (geralmente 0,05), dizemos que modelo alternativo se ajusta melhor aos dados do que o modelo nulo (isto é, rejeitamos o modelo nulo) e vice-versa (HUELSENBECK; CRANDALL, 1997). O cálculo do HLRT é feito pela seguinte fórmula:

$$LRT = 2(l_1 - l_0), \quad (58)$$

onde l_1 é a máxima verossimilhança do modelo mais complexo e rico em parâmetro (hipótese alternativa) e l_0 é a máxima verossimilhança sob o modelo menos complexo (hipótese nula).

Quando os modelos comparados são aninhados (a hipótese nula é um caso especial da hipótese alternativa) e a hipótese nula é correta, a estatística LRT é assintoticamente distribuída como um χ^2 com graus q de liberdade, onde q é a diferença no número de parâmetros livres entre os dois modelos (POSADA; BUCLEY, 2002).

2.5.2 Akaike Information Criteria (AIC)

O Akaike Information Criteria (AKAIKE, 1987) é um estimador baseado em Teoria da Informação e é utilizado para selecionar um modelo a partir de um conjunto finito de modelos. Podemos pensar no AIC como a quantidade de informação perdida quando usamos um modelo específico para aproximar o verdadeiro processo de evolução molecular. Portanto, o modelo com o menor AIC é o preferido. O AIC é calculado através da seguinte fórmula:

$$AIC = -2l + 2k, \quad (59)$$

onde l é o valor do log da máxima verossimilhança dos dados sob dado modelo e K é o número de parâmetros livres no modelo incluindo comprimentos dos ramos se eles foram estimados. Quando o tamanho da amostra (n) é pequena em comparação com o número de parâmetros ($\frac{n}{k} < 40$) o uso de uma segunda ordem AIC, o AICc, é recomendável:

$$AIC_c = AIC + \frac{(2k(k+1))}{(n-k-1)} \quad (60)$$

O AIC compara vários modelos candidatos ao mesmo tempo. Ele pode ser usado para comparar modelos aninhados e não aninhados, e a seleção do modelo da incerteza pode ser facilmente quantificado utilizando as diferenças AIC. O AIC é uma medida útil, que premia modelos com bom ajuste, mas impõe uma penalidade por parâmetros desnecessários.

2.5.3 Bayesian Information Criteria (BIC)

Bayesian Information Criteria é um critério para seleção de modelos entre um conjunto finito de modelos, baseado na função de verossimilhança e que está intimamente relacionado com o critério de informação de Akaike (AIC). O cálculo do BIC é feito através da fórmula:

$$BIC = -2l + k \log(n), \quad (61)$$

onde k é o número de parâmetros estimáveis, n é o tamanho da amostra e l é o valor do log da máxima verossimilhança dos dados sob dado modelo. O BIC compara vários modelos candidatos ao mesmo tempo, escolhendo o modelo com o menor valor BIC (RAFTERY, 1999).

2.5.4 Decision Theory (DT)

Em (MININ et al., 2003) foi desenvolvida uma abordagem que seleciona os modelos com base no seu desempenho filogenético, medido através do erro esperado em comprimentos dos ramos estimativas ponderados pelo seu respectivo valor de BIC. Utilizando o framework da Teoria da Decisão (DT). O melhor modelo é aquele com que minimiza a função de risco:

$$C_i \approx \sum_{j=1}^n \|\hat{B}_i - \hat{B}_j\| \frac{e^{\frac{-BIC_j}{2}}}{\sum_{j=1}^R (e^{\frac{-BIC_i}{2}})} \quad (62)$$

onde

$$\|\hat{B}_i - \hat{B}_j\|^2 = \sum_{l=1}^{2t-3} (\hat{B}_{il} - \hat{B}_{jl})^2 \quad (63)$$

e onde t é o número de sequências, BIC é o valor BIC calculado. As simulações sugerem que os modelos selecionados com este resultado critério resultam em inferências com melhores estimativas de comprimento dos ramos, sendo mais precisos do que os modelos obtidos por HLRT.

2.5.5 Bayes Factor

Em filogenia, o fator de Bayes é uma técnica utilizada para se realizar a escolha entre dois modelos evolutivos. Segundo (DARRIBA et al., 2012), essa técnica é a mais indicada para efetuar a escolhas dos modelos que serão utilizados no método de Inferência Bayesiana.

Dado um problema em que temos de escolher entre dois modelos, com base em dados observados D , a plausibilidade dos dois modelos diferentes M_1 e M_2 , parametrizado pelo vetores de parâmetros pelo modelo θ_1 e θ_2 é avaliado pelo pelo Fator de Bayes K , onde o modelo que retorne o menor valor de k é escolhido. O fator de Bayes pode ser calculado através da fórmula:

$$K = \frac{Pr(D|M_1)}{Pr(D|M_2)} = \frac{\int Pr(\theta_1|M_1)Pr(D|\theta_1, M_1)d\theta_1}{\int Pr(\theta_2|M_2)Pr(D|\theta_2, M_2)d\theta_2} \quad (64)$$

onde $Pr(M|D)$ é o valor da probabilidade posterior de um modelo, dada por:

$$Pr(M|D) = \frac{Pr(D|M)Pr(M)}{Pr(D)} \quad (65)$$

2.6 Diferenças e vantagens dos métodos de escolha de modelo evolutivo

O uso de diferentes estratégias de seleção de modelo pode levar à escolha de diferentes modelos (POSADA; CRANDALL, 2001). Sabemos que o modelo escolhido afeta todos os aspectos da análise filogenética, ficando evidente a comparação e avaliação de diferentes estratégias de seleção de modelo em todos os métodos de RAF exceto Parcimônia.

O método HLRT não é o mais indicado, pois a obtenção de valores P corretos para o cálculo do HLRT pode ser difícil. LRTs implicitamente assumem que pelo menos um dos modelos de comparação é correto, e quando os modelos são mal especificados, estes testes podem muitas vezes ser incorretos (POSADA; BUCLEY, 2002).

Pelo fato do HLRT somente fazer a comparação entre dois modelos, existem situações em que um modelo ideal pode não existir para o procedimento. Isso pode ocorrer, por exemplo, se o modelo GTR não é significativamente melhor do que HKY, o HKY não é significativamente melhor do que JC69, mas GTR é significativamente melhor do que JC69. Mesmo que exista um modelo ideal, que será sempre uma função do nível de significância, e o resultado do procedimento de escolha pode variar. Além disso, a abordagem HLRTs executa múltiplos ensaios com os mesmos dados, o que

aumenta a taxa de falsos positivos (isto é, a rejeitar a hipótese nula quando ela é verdadeira) (POSADA; BUCLEY, 2002).

A partir do exposto anteriormente, fica claro que a AIC e BIC apresentam várias vantagens importantes sobre os HLRTs para seleção do modelo. Isso se dá principalmente pelo fato de eles serem capazes de comparar simultaneamente vários modelos aninhados ou não. Os critérios AIC e BIC são muito simples de calcular a partir da Máxima verossimilhança, embora eles dependem de estimativas pontuais e não levam em conta a incerteza topológica (BOLLBACK, 2002).

A utilização da técnica de *Decision Theory* ainda é nova na filogenia, não sendo muito utilizada, além de não existirem muitos estudos que confirmam sua eficácia.

O *Bayes Factor* é recomendado para RAFs que utilizam o método de Inferência Bayesiana, sendo restritivo apenas a esse método.

2.7 Intervalo de confiança

Em estatística, um intervalo de confiança é um valor estimado de um parâmetro estatístico. Em vez de estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis. Quão prováveis são estas estimativas é determinado pelo coeficiente de confiança e pela significância. Quanto maior a probabilidade de o intervalo conter o parâmetro, maior será o intervalo (MORESCHI, 2010).

O quanto essas estimativas são prováveis será determinado pelo coeficiente de confiança $(1 - \alpha)$, para $\alpha \in (0, 1)$, onde α é a significância. Pode-se afirmar que confiança é o quanto se deseja confiar no resultado de uma expressão matemática e a significância é o quanto se deseja desconfiar.

Para um melhor entendimento, tomemos U e V como funções de amostras, cuja distribuição de probabilidade dependa do parâmetro θ e $P(U < \theta < V|\theta) = 1 - \alpha$, então o intervalo aleatório (U, V) é um intervalo de confiança nível $100(1 - \alpha)\%$ para θ . Podemos portanto interpretar o intervalo de confiança como um intervalo que contém os valores "plausíveis" que o parâmetro θ pode assumir. Assim, a amplitude do intervalo está associada a incerteza que temos a respeito do parâmetro.

Considere X_1, X_2, \dots, X_n uma amostra aleatória retirada de uma população com distribuição f_θ que depende do parâmetro θ . Por exemplo, tomamos X_1, X_2, \dots, X_n uma amostra aleatória com distribuição normal com média μ desconhecida e desvio padrão conhecido $\sigma = 1$. Para propormos um intervalo de confiança para o parâmetro θ , vamos introduzir o conceito de quantidade pivotal. Uma função Q da amostra (X_1, X_2, \dots, X_n) e do parâmetro θ cuja distribuição de probabilidade não depende do parâmetro θ é

denominada quantidade pivotal. Desta forma, dado o nível de confiança $1 - \alpha$, tomamos

$$1 - \alpha = \mathbb{P}(q_1 \leq Q(X_1, X_2, \dots, X_n; \theta) \leq q_2) \quad (66)$$

Se a quantidade pivotal Q for inversível, podemos resolver a inequação acima em relação a θ e obter um intervalo de confiança.

2.8 JModelTest

Embora existam várias ferramentas para realizar escolha dos modelos e parâmetros como JModeltest, FindModel, ModelGenetator e MrAIC, neste item vamos falar somente de JModelTest por ser um dos programas mais utilizados para este fim além de estar em constante atualização.

O JModelTest é uma ferramenta de código aberto, desenvolvida por ([DARRIBA et al., 2012](#)) que realiza a seleção de modelos evolutivos que tenham o melhor ajuste nas sequências analisadas.

O JModelTest implementa cinco diferentes estratégias de seleção de modelos:

- *Hierarchical likelihood ratio test* (HLRT)
- *Dinamic likelihood ratio test* (DLRT)
- *Akaike Information Criteria* (AIC)
- *Bayesian Information Criteria* (BIC)
- *Decision Theory* DT

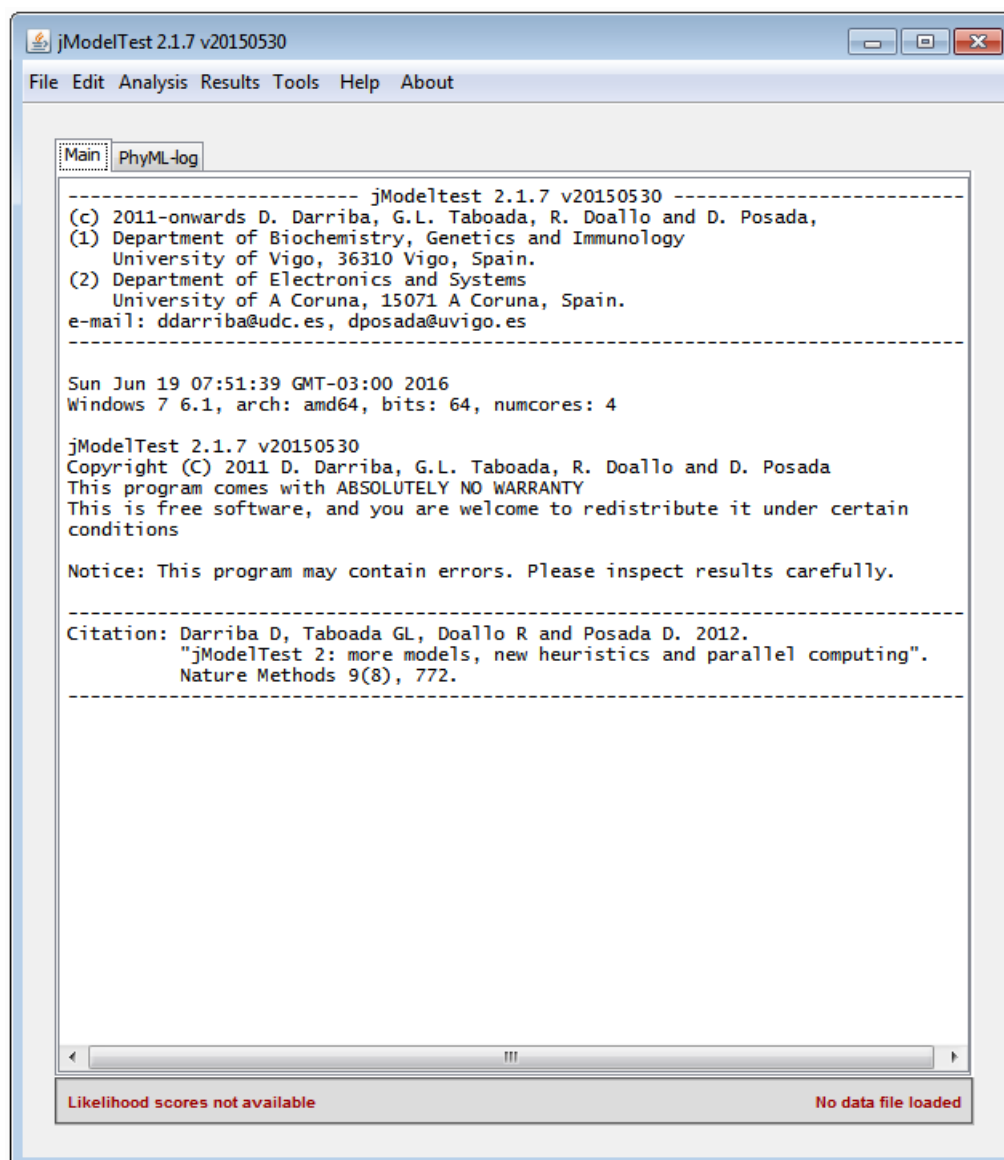
Pelo fato de ser desenvolvido na linguagem Java¹, ele pode ser executado em qualquer sistema operacional, desde que instalada a *Java Runtime Environment* (JRE).

Para o cálculo de máxima verossimilhança, o JModelTest utiliza o software PhyML, que vem incluso na biblioteca de seu projeto.

O JModelTest pode ser executado de duas maneiras diferentes: Prompt de comando ou interface gráfica, como observado na figura 19.

¹www.oracle.com/br/java

Figura 19 – Tela inicial do JModelTest, modo gráfico



No modo de prompt de comando, o JModelTest possui como parâmetros principais de entrada:

- -AIC: realiza a seleção de modelo através do AIC;
- -BIC: realiza a seleção de modelo através do BIC;
- -DT: realiza a seleção de modelo através do DT;
- -c Intervalo: define o intervalo de confiança para o processo de seleção de modelo;
- -d Arquivo: define o arquivo de entrada;
- -dLRT: realiza a seleção de modelo através do dLRT;

- -f: inclui modelos com frequências desiguais de bases;
- -g NúmeroDeCategorias: inclui modelos com taxa de variação entre sítios e define o número de categorias;
- -getPhylip converte a sequência no formato PHYLIP;
- -h Intervalo: define o intervalo de confiança para HLRT;
- -help: exibe uma mensagem de ajuda;
- -hLRT: realiza a seleção de modelo através do HLRT;
- -i: inclui modelos com proporção de sítios invariáveis;
- -o ArquivoSaida: redireciona os resultados para um arquivo;
- -s 3|5|7|11|203: define o número de modelos a serem avaliados, onde:
 - 3: JC/F81, K80/HKY, SYM/GTR (usados como padrão);
 - 5: JC/F81, K80/HKY, TrNef/TrN, TPM1/TPM1uf, SYM/GTR;
 - 7: JC/F81, K80/HKY, TrNef/TrN, TPM1/TPM1uf, TIM1ef/TIM1, TVMef/TVM, SYM/GTR;
 - 11: todos os modelos mostrados na tabela 2.8.
 - 203: todas as possíveis submatrizes do modelo GTR.
- -tr NumeroDeThreads: define o número de threads a serem utilizadas;
- -S NNI|SPR|BEST: define a heurística utilizada para encontrar a árvore inicial para o cálculo da máxima verossimilhança;

O JModelTest utiliza diversos modelos evolutivos, com diferente quantidade de parâmetros livres, frequências de bases iguais ou diferentes, assim como as taxas de substituição. A tabela 3 ilustra alguns modelos implementados no JModelTest e suas características.

Tabela 3 – Modelos evolutivos usados JModelTest (alguns dos modelos possíveis). Qualquer um desses modelos podem incluir os parâmetros I, G ou os dois (+I+G).

Modelo	Parâmetros Livres	Frequências de Base	Taxas de substituição	Código
JC	0	iguais	AC=AG=AT=CG=CT=GT	000000
F81	3	diferentes	AC=AG=AT=CG=CT=GT	000000
K80	1	iguais	AC=AT=CG=GT;AG=GT	010010
HKY	4	diferentes	AC=AT=CG=GT;AG=GT	010010
TrNef	2	iguais	AC=AT=CG=GT;AG;GT	010020
TrN	5	diferentes	AC=AT=CG=GT;AG;GT	010020
TPM1	2	iguais	AC=GT;AG=CT;AT=CG	012210
TPM1uf	5	diferentes	AC=GT;AG=CT;AT=CG	012210
TPM2	2	iguais	AC=AT;CG=GT;AG=CT	010212
TPM2uf	5	diferentes	AC=AT;CG=GT;AG=CT	010212
TPM3	2	iguais	AC=AT;AG=GT;AG=CT	012012
TPM3uf	5	diferentes	AC=CG;AT=GT;AG=CT	012012
TIM1	3	iguais	AC=GT;AT=CG;AG;CT	012230
TIM1uf	6	diferentes	AC=GT;AT=CG;AG;CT	012230
TIM2	3	iguais	AC=AT;CG=GT;AG;CT	010232
TIM2uf	6	diferentes	AC=AT;CG=GT;AG;CT	010232
TIM3	3	iguais	AC=CG;AT=GT;AG;CT	012032
TIM3uf	6	diferentes	AC=CG;AT=GT;AG;CT	012032
TVMef	4	iguais	AC;CG;AT;GT;AG=CT	012314
TVM	7	diferentes	AC;CG;AT;GT;AG=CT	012314
SYM	5	iguais	AC;CG;AT;GT;AG;CT	012345
GTR	8	diferentes	AC;CG;AT;GT;AG;CT	012345

O primeiro passo na execução do JModelTest é efetuar o cálculo da máxima verossimilhança dos modelos a serem analisados. Em posse desses valores, o JModelTest usa a técnica de seleção de modelo escolhida, para retornar o conjunto de modelos que se adequam às sequências de entrada, baseado no intervalo de confiança passado.

O JModelTest tem como saída:

- o valor do log da máxima verossimilhança de cada modelo testado;
- o conjunto de modelos indicados aos dados.

3 Metodologia

O desenvolvimento do trabalho se deu em cinco fases:

- Pesquisa dos fundamentos de RAF;
- Busca e leitura de artigos sobre filogenia;
- Levantamento e cruzamento de dados obtidos nos artigos;
- Criação da metodologia utilizada na ferramenta;
- Implementação da ferramenta.

3.1 Pesquisa dos fundamentos de RAF

Para o entendimento das técnicas utilizadas em RAF, foi feito um estudo sobre a filogenia. Os métodos que ela utiliza, suas vantagens, desvantagens, ferramentas que implementam os métodos de RAF. Depois foi estudado os processos markovianos. Feito isso, foi estudado os modelos evolutivos, baseados em cadeias de Markov, assim como seus parâmetros e as taxas de heterogeneidade entre sítios. O próximo passo foi analisar as técnicas utilizadas para a seleção de modelos evolutivos (HLRT, AIC e BIC), assim como suas vantagens e desvantagens. Tendo feito o levantamento das técnicas para escolha dos modelos evolutivos, foi feita uma pesquisa sobre as ferramentas que implementam as técnicas de escolha do modelo, em especial o JModelTest. O último passo para a conclusão dessa etapa foi fazer um estudo sobre a utilização das distâncias evolutivas entre sítios para a escolha do método de RAF simplificado.

3.2 Busca e leitura de artigos sobre filogenia

Para a coleta de dados, foram lidos 137 artigos científicos publicados no ano de 2015, baixadas através do portal de periódicos da CAPES, utilizando as seguintes palavras chaves: *jmodeltest*, *mr bayes*, *phyml*, *modeltest*, *phylogenetic reconstruction*, *mr-bayes+puma*, *phyml+m3l*.

Todas as referências bibliográficas utilizadas para a coleta dos dados encontram-se em anexo.

Foram analisadas, em cada publicação, a finalidade da reconstrução filogenética, os métodos utilizados, a ferramenta utilizada na RAF, os modelos evolutivos, os parâmetros dos modelos e se foi utilizada alguma ferramenta para a seleção de modelos.

3.3 Levantamento e cruzamento de dados obtidos nos artigos

Uma vez lidos os artigos, foi feito o cruzamento dos dados, com a finalidade de entender e quantificar os seguintes aspectos:

1. A finalidade da filogenia
2. Os reinos analisados nas RAF;
3. As revistas em que os artigos foram submetidos;
4. A ferramenta mais utilizada para a escolha de seleção de modelo evolutivo;
5. Os modelos evolutivos mais selecionados pelas ferramentas de seleção;
6. o método estatístico mais utilizado pelas ferramentas de seleção de modelo;
7. Os modelos mais utilizados para RAFs que não usaram um programa de seleção de modelos;
8. A ferramenta de reconstrução mais utilizada;
9. Os métodos mais utilizados em reconstruções que utilizaram de um programa de seleção de modelo evolutivo;
10. A quantidade de métodos utilizados em cada conjunto de dados;
11. Os métodos mais utilizados em reconstruções feitas sem a utilização de um programa de seleção de modelo evolutivo;
12. Os modelos mais utilizados para RAF que não usaram um programa de seleção de modelos e que utilizaram o método de ML;
13. Os modelos mais utilizados para RAF que não usaram um programa de seleção de modelos e que utilizaram o método de Inferência Bayesiana;

3.4 Desenvolvimento da metodologia utilizada na ferramenta

Depois de feita a análise quantitativa e qualitativa da pesquisa bibliográfica, foi realizado um estudo teórico sobre a melhor maneira de realizar RAF. Com base nisso, foi criada uma metodologia para a seleção automática dos métodos, a qual otimiza o desempenho da execução da RAF e a sua precisão. As conclusões metodológicas serão melhor abordadas na seção Desenvolvimento.

3.5 Implementação da ferramenta

Uma vez com a metodologia proposta, continuou-se com a implementação da ferramenta de seleção automática de métodos, modelos e seus parâmetros para RAF.

3.5.1 IgrafuWeb

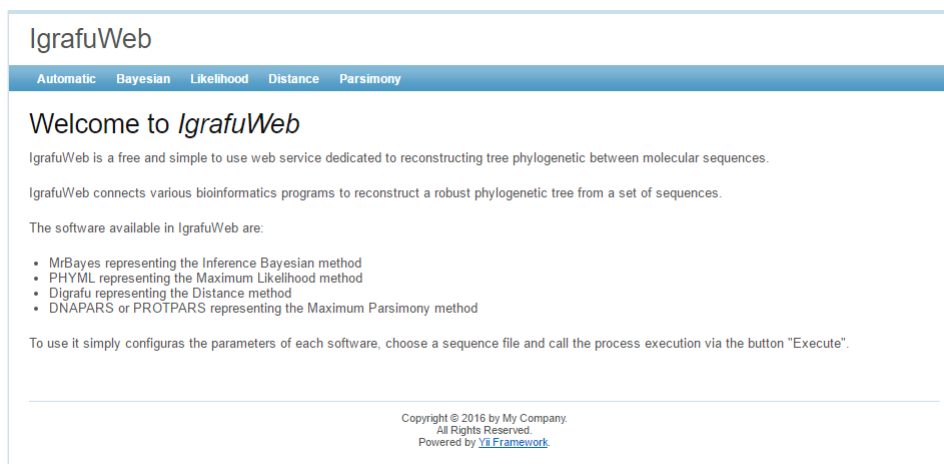
O presente trabalho incorporou o módulo de seleção automática ao IgrafuWeb ([NETO, 2015](#)), que utiliza em seu núcleo os programas: MrBayes, PHYML, Digrafu e DNAPARS/PROTPARS. O IgrafuWeb foi estudado e analisado em todas as suas características, a fim de entender a heurística, os parâmetros e o funcionamento do mesmo.

O IgrafuWeb (Figura 20) é uma ferramenta computacional de alto desempenho, desenvolvida por ([NETO, 2015](#)) que incorpora ferramentas de RAF dos quatro principais métodos de inferência filogenética:

- Método de Máxima Parcimônia – ofertado através do software DNAPARS ou PROTPARS ([FELENSTEIN, 1993](#));
- Método de Distância – ofertado através do software Digrafu (??);
- Método de Máxima Verossimilhança – oferecido pelo programa PHYML ([GUINDON; GASCUEL, 2003](#));
- Método de Inferência Bayesiana – disponibilizado através do software MrBayes ([HUELSENBECK; RONQUIST, 2001](#)).

O IgrafuWeb disponibiliza uma estrutura na Web para RAF e permite a RAF utilizando o poder de processamento do CACAU (Centro de Armazenamento de Dados e Computação Avançada da UESC) e disponibiliza a árvore gerada na Web ([NETO, 2015](#)).

Figura 20 – Página inicial apresentado a estrutura organizacional do site e uma breve explicação do IgrafuWeb.



3.5.1.1 DNAPARS e PROTPARS

Os softwares DNAPARS 8 (*DNA Parsimony Program*) e PROTPARS 9 (*Protein Sequence Parsimony Method*), que realizam RAF, com base no método de Máxima Parcimônia, foram os escolhidos para serem acoplados ao IgrafuWeb. Isto aconteceu basicamente pela qualidade e eficiência dos mesmos, e pela ampla utilização de ambos na literatura. Pertencem à suíte de programas de filogenética computacional PHYLIP (FELENSTEIN, 1993), de licença livre, para inferência de árvores evolucionárias.

O PHYLIP foi concebido por Felsenstein na linguagem de programação C, habilitado para funcionar na maioria das plataformas computacionais. É composto de 35 programas que conseguem dirimir diversos problemas filogenéticos da literatura atual. Cada um contém suas particularidades, com um protocolo específico, conforme o método filogenético que representa. A execução é realizada através de comandos de texto, que são determinantes para definir configurações específicas do usuário conforme os dados de entrada. Estes por sua vez são inseridos através de arquivo de texto, contendo na primeira linha as devidas informações sobre as quantidades de espécies e de sítios, assim como as sequências representativas de cada espécie relacionadas com o estudo (FELENSTEIN, 1993)

O DNAPARS calcula a topologia inicial utilizando o método de adição por passos, sendo que após a inserção de uma espécie e antes de adicionar uma outra, todas as modificações topológicas do tipo NNI são aplicadas sistematicamente e uma nova topologia é aceita desde que o seu valor de parcimônia seja menor que a melhor solução encontrada até o momento. Uma vez que todas as espécies foram adicionadas, o DNAPARS fornece uma opção para fazer modificações topológicas adicionais empregando SPR (poda e enxerto de sub-árvores)(TICONA, 2008, NETO, 2015, p.53).

3.5.1.2 Digrafu

O Digrafu ¹ é um software de inferência filogenética, baseada nos métodos de distância que utiliza os algoritmos UPGMA, Weighbor, BIONJ e FastME e propõe uma melhora no tratamento dos dados de entrada e na aplicação dos mesmos. O Digrafu é um programa de domínio e código fonte aberto, produzido na UESC (??).

A execução do Digrafu é dada seguindo os seguintes passos principais:

- Conversão de formatos para um suportado pelo DnaDist ou ProDist;
- Cálculo da matriz de distâncias;
- Escolha do melhor algoritmo de distância para realizar a RAF dadas as sequências;
- Execução do algoritmo escolhido para a criação da árvore.

O início da execução do Digrafu é dada pelo cálculo da matriz de distância, que é feito pelos programas DnaDist ou ProDist, ambos presentes no pacote PHYLIP.

Uma vez criada a matriz de distâncias, o Digrafu escolhe dinamicamente qual algoritmo de distância deverá ser executado. Existem alguns fatores que influenciam essa escolha, são eles:

- As sequências genéticas;
- A matriz de distâncias;
- Preferência do usuário, a qual está atrelada à finalidade que se pretende dar ao resultado do programa.

3.5.1.3 PhyML

O PhyML é um programa de inferência de árvores filogenéticas que estima a máxima verossimilhança de alinhamentos de nucleotídeos e aminoácidos e através de processos de Markov e estratégias de heterogeneidade sobre sítios. Implementa 8 modelos evolutivos de DNA e 12 modelos evolutivos de proteínas, levando a um resultado rápido e preciso, devido a sua heurística (NETO, 2015, p. 54).

O PhyML emprega uma abordagem heurística que reduz consideravelmente o tempo de execução. Isto se dá por conta de um forte relacionamento entre as modificações topológicas e a otimização dos parâmetros e dos comprimentos de ramos. Por essa razão, emprega-se o BIONJ para determinar a topologia inicial, juntamente com algum método de otimização para estimar inicialmente os parâmetros do modelo de

¹Disponível em: <http://github.com/said/digrafu>.

substituição. Seguidamente, todas as modificações topológicas do NNI são examinadas, sendo apenas otimizado o comprimento do ramo envolvido em tal operação. Dessa forma, todas as mudanças possíveis são calculadas de forma independente com um menor custo computacional. Aplica-se uma proporção das modificações que mais aumentaram a verossimilhança das árvores e, finalmente, recalculam-se os parâmetros do modelo de substituição. A nova topologia encontrada é o novo ponto de partida para uma nova iteração do algoritmo que perdura até que não haja mais modificações a serem aplicadas. Finalmente, os comprimentos de ramos e parâmetros do modelo são reotimizados (GUINDON et al., 2010).

O PHYML traz de volta o resultado de sua execução em arquivos de texto, com uma nomenclatura fixa de acordo com o nome do arquivo de sequência submetido.

3.5.1.4 MrBayes

O MrBayes (HUELSENBECK; RONQUIST, 2001) software de código fonte aberto, implementado na linguagem C, suportado pelas plataformas Windows, Macintosh e Unix e implementa o método inferência bayesiana para a reconstrução de árvores filogenéticas. O MrBayes pode ser executado através de linhas de comandos ou em modo batch.

A primeira versão do MrBayes foi publicada em 2001 por Huelsenbeck e colaboradores, e atualmente está na versão 3.2.5 de 8 de abril de 2015.

Para utilizar MrBayes é necessário definir os parâmetros de execução e o arquivo contendo os alinhamentos os quais deseja-se inferir a árvore. A execução do MrBayes toma como base os parâmetros:

- Número de gerações que representa quantidade de testes que será realizada na execução;
- Quantidade de cadeias que é a estrutura que irá conter uma árvore;
- Quantidade de análises que representa uma seção de testes independentes dos demais;
- Frequência de amostragem onde os dados obtidos nos testes são armazenados;
- Frequência de diagnóstico que é uma ferramenta para verificação parcial dos resultados obtidos.

3.5.2 CACAU

O Centro de Armazenamento de dados e Computação Avançada da Universidade Estadual de Santa Cruz-UESC (CACAU), localizado no Núcleo de Biologia

Computacional e Gestão de Informações Biotecnológicas (NBCGIB) é um ambiente computacional de alto desempenho que é composto por 20 nós com 2 processadores processadores Intel(R) Xeon(R) E5430@ de 2.66 GHz e 16 GB de memória, totalizando 160 cores e 320 GB de memória.

A ferramenta desenvolvida, assim como o IgrafuWeb estão instalados e configurados no CACAU.

O CACAU possui uma estrutura organizacional de onde o servidor que hospeda os sites (Sioux) se mantém em uma rede distinta da rede dos nós que realizam de fato o processamento da análise filogenética.

4 Desenvolvimento

Este capítulo está dividido em cinco seções: uma recompilação teórica sobre como resolver o problema de seleção de modelo evolutivo e suas implicações, outra seção sobre a análise de pesquisa bibliográfica sobre como atualmente realiza-se RAF de forma prática, depois explicação teórica de como calcular distância para escolha simplificada de métodos e modelos, na seção seguinte descreve-se a metodologia proposta e finalmente tem-se uma seção que explica o desenvolvimento da metodologia proposta.

4.1 O problema da escolha do modelo

A seleção de um modelo adequado é fundamental para a produção de boas análises filogenéticas. A escolha errada de modelos evolutivos pode ser feita de duas maneiras: escolhendo modelos subparametrizados (que utilizam menos parâmetros que o necessário para a inferência da sequência analisada) e escolhendo modelos superparametrizados (quando os parâmetros são excessivos para representar a evolução da sequência analisada).

A utilização de um modelo subparametrizado pode influenciar fortemente as estimativas de probabilidades posteriores e outros parâmetros do modelo. A tendência é especialmente grave quando a heterogeneidade de taxa é negligenciada e pode conduzir a uma subavaliação de comprimentos dos ramos (BUCKLEY; CUNNINGHAM, 2002).

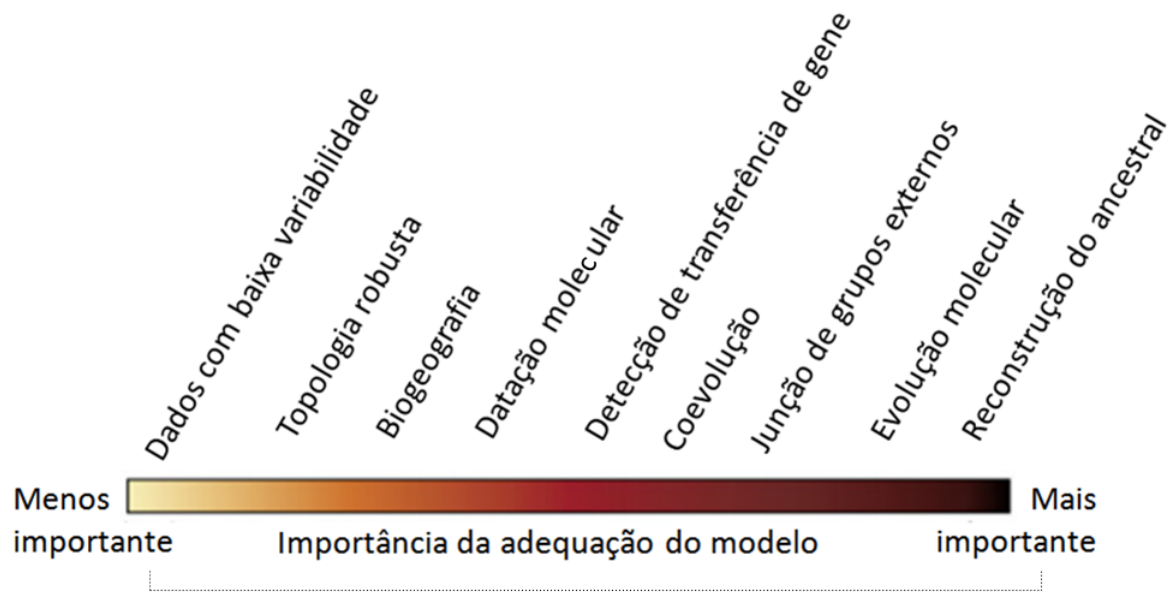
(KELCHNER; THOMAS, 2006) afirma que a importância do ajuste do modelo pode ser visto como um processo contínuo e que depende da aplicação a ser utilizada e as questões que estão sendo investigados. Para os estudos que se concentram apenas em relação entre os organismos, a exigência principal é uma topologia precisa em vez de uma estimativa precisa de quantas mudanças ocorreram na árvore. Em certos casos, a topologia permanece a mesma mesmo utilizando uma vasta gama de modelos sugerindo que quando se busca construir uma topologia robusta, pode-se utilizar modelos com menos parâmetros. Outro tipo de reconstrução que pode utilizar modelos menos complexos é quando os conjuntos de dados mostram pouca variação nas sequências.

Uma grande preocupação em determinados contextos é se o pesquisador deve escolher entre sub ou super parametrização em uma RAF. Observando os erros na escolha dos modelos nos últimos vinte anos de estudo de filogenia, (KELCHNER; THOMAS, 2006) afirma que em caso de dúvida, para resultados mais seguros, deve-se preferir modelos ricos em parâmetros.

Para a maioria dos outros usos da filogenia o comprimento dos ramos pode ter

um papel crucial. Quando o objetivo é estimar tempos de divergência entre linhagens em uma árvore, a estimativa precisa de comprimento ramo são mais importantes ou quando se necessita investigar diferenças significativas entre topologias concorrentes (coevolução ou transferência horizontal de genes) a necessidade de escolher modelos mais adequados e geralmente mais ricos em parâmetros para a sequência a ser analisada se torna mais importante (KELCHNER; THOMAS, 2006; SULLIVAN; SWOFFORD, 2001), como pode ser visto na figura 21.

Figura 21 – Importância da adequação de modelos



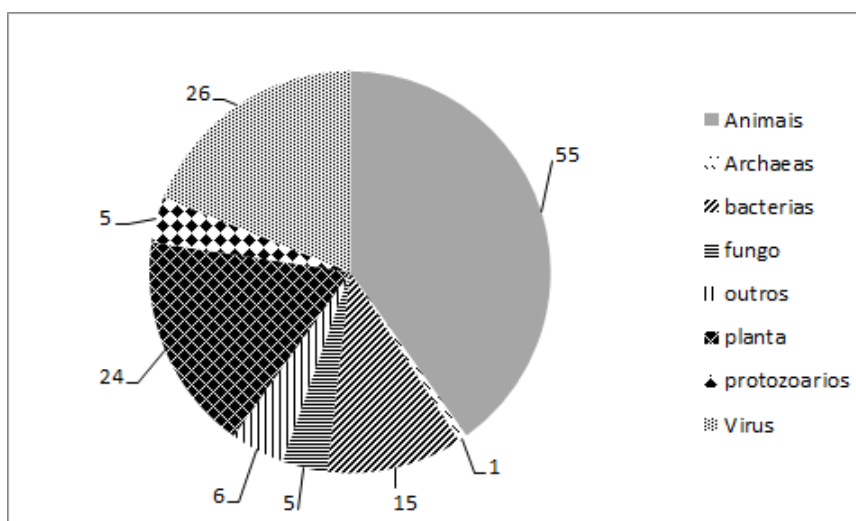
adaptado e traduzido de (KELCHNER; THOMAS, 2006)

4.2 Análise da pesquisa bibliográfica

A seguir, a análise da pesquisa bibliográfica realizada.

Dentre os grandes reinos analisados nos artigos, o reino mais estudado foi o animal, com 55 artigos, seguido pelos vírus, com 26 artigos, e as plantas, com 24 artigos. A figura 22 ilustra os reinos analisados por cada artigo:

Figura 22 – Reinos estudados

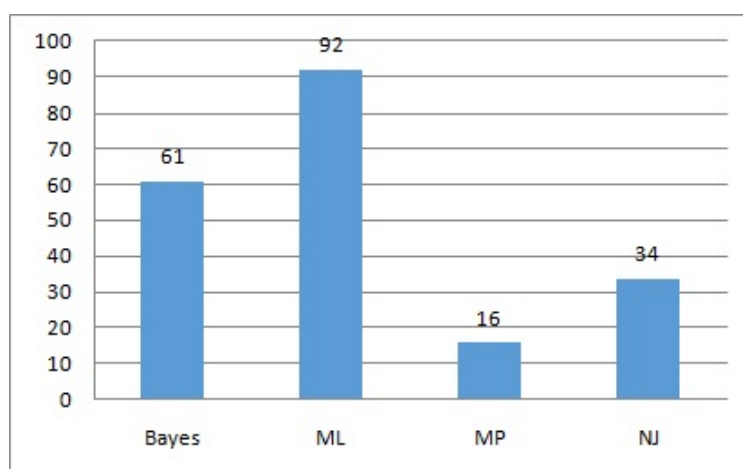


Isto mostra que a análise filogenética atual abrange grandes variedades de reinos, o que mostra que existe a necessidade de fazer uma análise mais aprofundada nas sequências de entradas, pois estas podem ser pouco homólogas entre si, ou seja, elas podem não ter a mesma origem embrionária e desenvolvimento semelhante em diferentes espécies que descendem de um ancestral em comum.

A pesquisa gerou os seguintes dados em relação aos métodos de RAF:

Do total dos 137 artigos analisados foi descoberto que das reconstruções feitas, como ilustrado na figura 23, o método mais utilizado foi o de máxima verossimilhança (92 vezes), seguido da inferência bayesiana (61 vezes), distância (34 vezes) e parcimônia (16 vezes).

Figura 23 – Quantificação dos métodos utilizados pelos autores

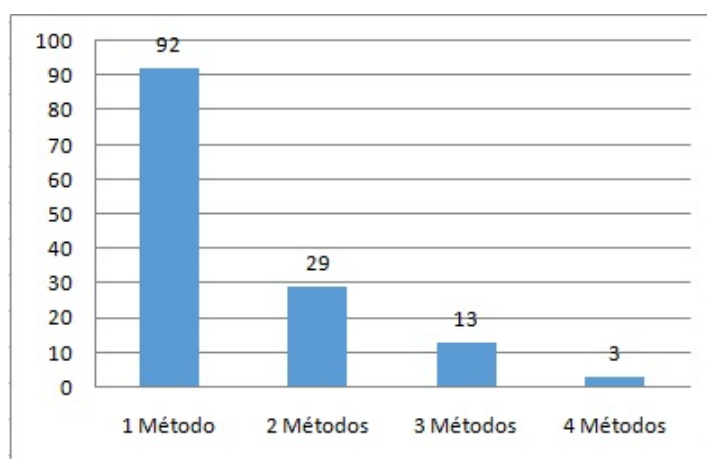


Dentre 89 artigos que utilizam programas de seleção de modelos matemáticos, os métodos mais utilizados foram os de ML (65 vezes) e IB (54 vezes), sendo que 50 destes artigos fizeram a RAF utilizando apenas 1 método (26 ML, 18 IB, 6 distância).

Isso mostra que nos casos onde a inferência é feita com maior critério, os autores optam por métodos mais complexos.

Em relação à quantidade de diferentes métodos utilizados em cada artigo, constatou-se que 92 deles utilizaram apenas um método para a RAF, enquanto 29 autores utilizaram 2 métodos, 13 utilizaram 3 métodos e apenas 3 autores utilizaram os 4 métodos em seus artigos, como pode ser visto na figura 24:

Figura 24 – Quantidade de Métodos utilizados em cada artigo



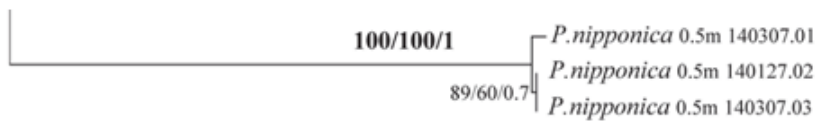
Dos artigos que fizeram as RAFs através de mais de um método, 22 deles apresentaram uma árvore de um método específico, colocando em seus galhos os valores de bootstrap dos métodos e o valor de probabilidade posterior, geralmente representado por barras, onde os primeiros valores são os valores de bootstrap dos métodos MV e distância e o último valor é o valor de probabilidade posterior do método de IB como pode ser observado na figura 25. Essas RAFs foram feitas utilizando bootstrap e uma árvore consensual, mas apenas para cada método utilizado. Não é citado que se faz uma árvore consenso entre as árvores geradas por diferentes métodos. Quatro artigos exibiram apenas a árvore gerada por um método, três deles por identidade absoluta entre as árvores geradas pelos diferentes métodos de RAF e um por escolha do autor. Seis artigos exibiram as árvores geradas por diferentes métodos de maneira independente, mostrando uma árvore para cada método. Dois autores criaram uma árvore a partir das árvores geradas, utilizando ferramentas como o TreeAnnotator¹.

O treeAnnotator é um software do pacote BEAST² que auxilia na síntese das informações a partir de uma amostra de árvores produzidas através de Inferência Bayesiana em uma única árvore "resumo". As informações da árvore resumo incluem as probabilidades posteriores dos nós na árvore de destino, as estimativas posteriores e limites das alturas de dos seus nós.

¹disponível em: <http://beast.bio.ed.ac.uk/downloads>

²<http://beast.bio.ed.ac.uk>

Figura 25 – Valores de bootstrap e probabilidade posterior de diferentes métodos para uma mesma sequência

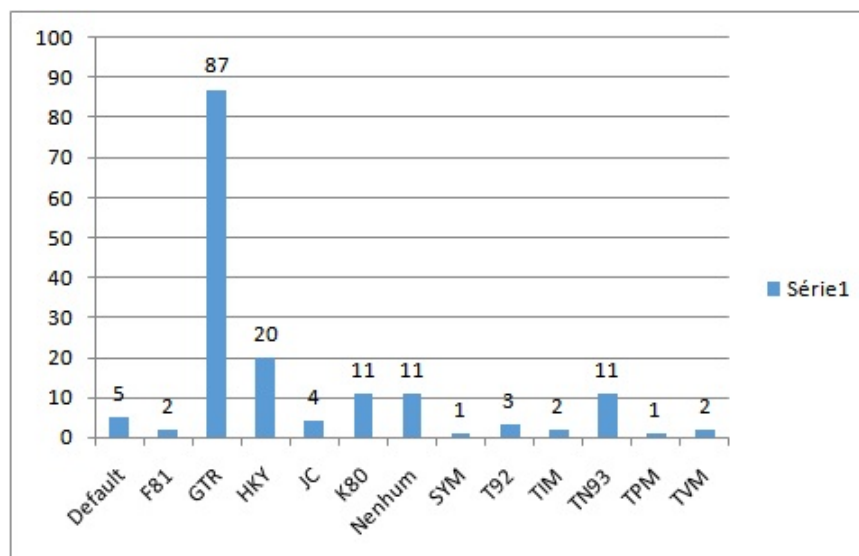


Doravante será quantificado como os modelos evolutivos são analisados. Dos 45 artigos que usam mais de um método de RAF, 28 (62,2%) deles utilizam o mesmo modelo evolutivo para todos os métodos, um artigo (2,2%) não cita quais os modelos utilizados e 16 (35,5%) utilizam diferentes modelos evolutivos, o que mostra que a maioria dos autores utilizam os modelos baseados nas sequências, não no método.

Dos 22 artigos que usam dois ou mais modelos evolutivos nas RAFs, cinco (22,7%) o fazem utilizando o mesmo método, porém, utilizando diferentes conjuntos de dados. Os outros 17 (77,3%) utilizam diferentes modelos para diferentes métodos, com o mesmo conjunto de dados.

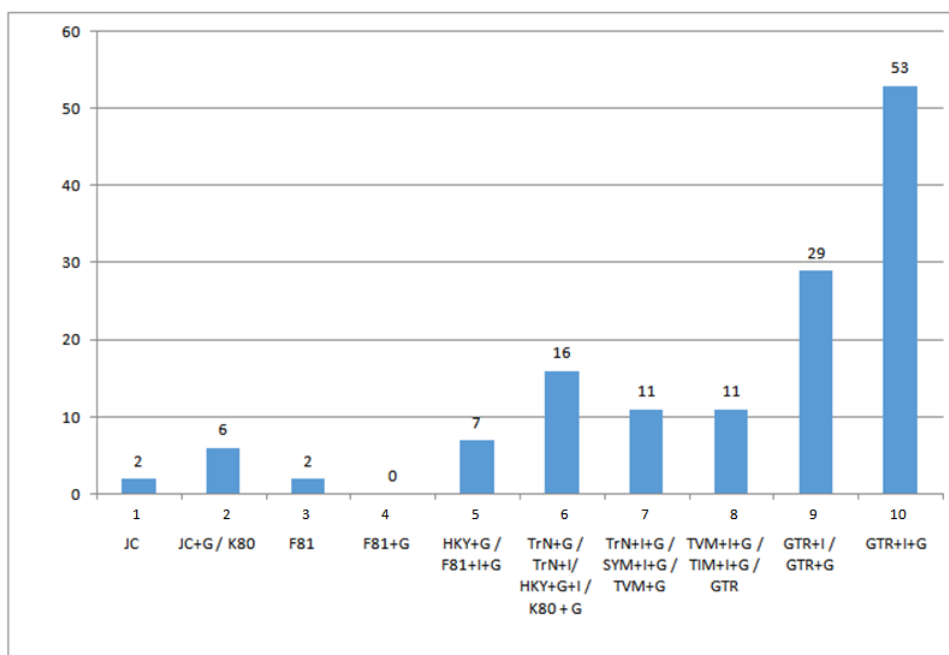
O modelo evolutivo mais utilizado pelos autores foi o modelo GTR, com 87 ocorrências, seguido pelo modelo HKY (que foi usado 20 vezes). Os demais modelos estão ilustrados na figura 26.

Figura 26 – Quantificação dos modelos Default, F81 (Felsenstein 81), GTR (*General-Time-Reversible*), HKY (Hasegawa-Kino-Yano), JC (Jukes-Cantor), K80 (Kimura), nenhum modelo, SYM (*symmetrical model*), T92 (Tamura92), TIM (*Transitional Model*), TN93 (Tamura93), TPM (*Three Parameters model*), TVM (*Transversional Substitution Model*), utilizados pelos autores. O modelo default é o padrão dos softwares de RAF utilizado pelos autores.



Em relação ao número de parâmetros utilizados, a maior parte dos modelos utilizados pode ser considerado rico em parâmetros, visto que 67,88% dos artigos utilizaram modelos com mais de 8 parâmetros (GTR + I + G, GTR + I, GTR + G, TVM + I + G, TIM + I + G e GTR). Não foi medido o grau de similaridade entre sequências em cada conjunto de dados dos artigos estudados, se forem geralmente baixos, poderia se esperar uma inclinação das utilizações em direção a modelos mais simples (mais à esquerda na figura 27).

Figura 27 – Utilização dos modelos evolutivos de acordo com os seus parâmetros.

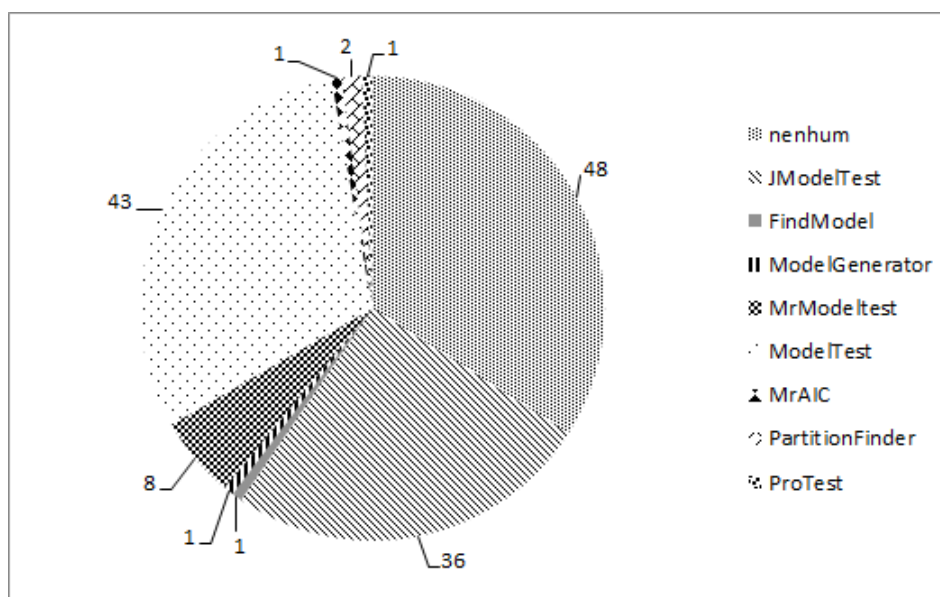


A seguir, será analisada o uso ou não de ferramentas para escolha do melhor modelo evolutivo.

Com base na pesquisa realizada, pode-se constatar que 48 (34 %) dos autores não utilizaram nenhuma ferramenta para escolha do melhor modelo evolutivo. As ferramentas para a seleção de modelos mais utilizadas foram ModelTest (43 vezes), JModeltest (36 vezes) e MrModelTest, que foi usada por 8 autores. As demais ferramentas, a quantidade de artigos que as utilizaram e as porcentagens estão descritos na figura 28.

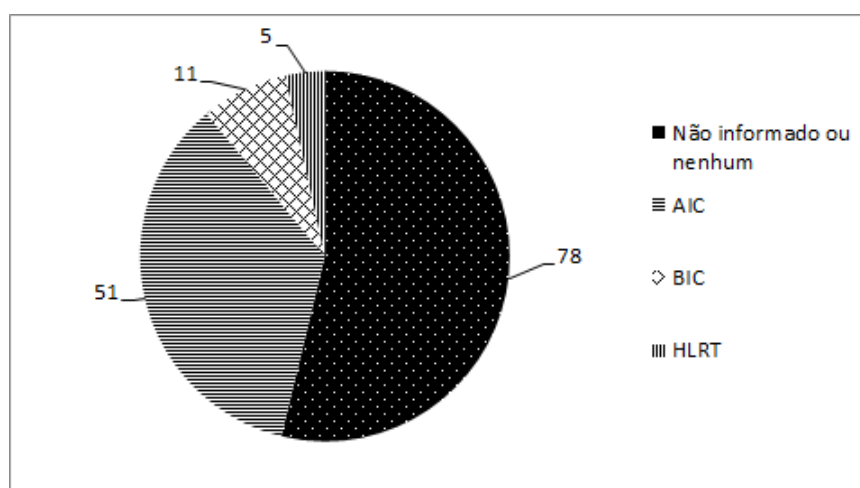
Dos quatro artigos que utilizaram mais de um programa para seleção do modelo evolutivo, três utilizaram um programa para IB (MrBayes) e um para ML (Modeltest), o outro artigo utilizou os programas para conjuntos diferentes de dados. Os programas, retornaram diferentes modelos para os conjuntos de dados.

Figura 28 – ferramenta para seleção de modelo evolutivo utilizadas pelos autores



Dentre os artigos pesquisados, 78 não utilizaram nenhuma técnica para seleção de modelo evolutivo ou não informaram qual técnica utilizaram, mesmo usando alguma ferramenta para a seleção de modelos. Dos artigos que utilizaram alguma técnica para a escolha do modelo evolutivo, 51 usaram AIC, 11 utilizaram BIC e 5 autores optaram pelo método HLRT (figura 29).

Figura 29 – Técnicas para seleção de modelos evolutivos utilizadas pelos autores

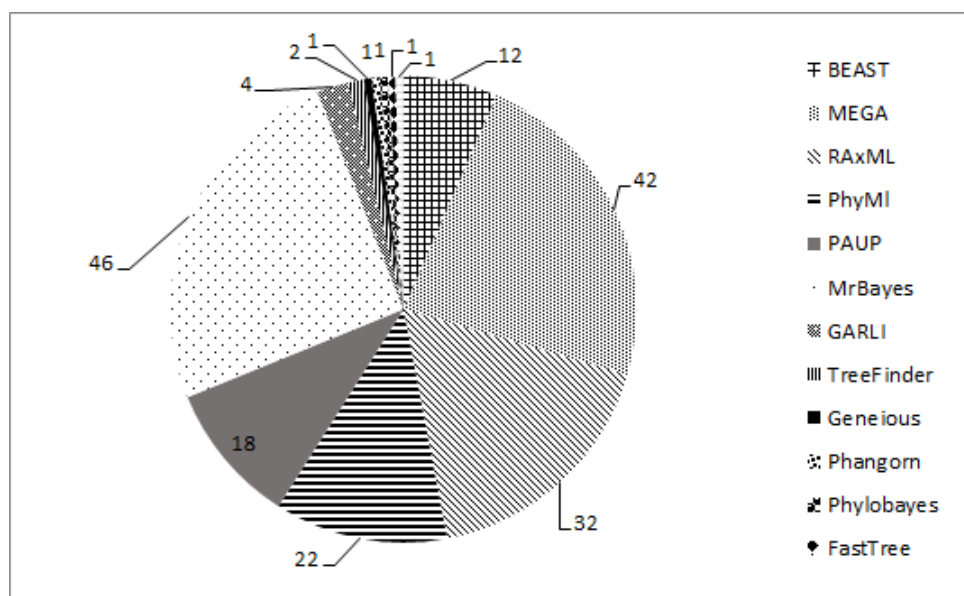


Outra informação importante é quais são os programas mais utilizados nas RAFs. Como ilustrado na figura 30, o software mais utilizado foi o MrBayes (46 ocorrências), seguido pelo MEGA (42 ocorrências), pelo RAxML (32 ocorrências) e o PhyML (22 ocorrências), que implementam o método de máxima verossimilhança.

MEGA (*Molecular Evolutionary Genetics Analysis*) é um software livre para inferência filogenética que implementa os métodos de Parcimônia, Distância, Máxima

Verossimilhança e Inferência Baiesiana.

Figura 30 – Ferramentas utilizadas para a RAF



Dos Artigos que fizeram a RAF sem utilizar nenhuma ferramenta para a seleção do melhor modelo evolutivo e que utilizaram apenas um método, 3 utilizaram Inferência Bayesiana, 21 utilizaram ML, 2 utilizaram MP e 17 utilizaram NJ. 3 artigos fizeram a inferência pelos métodos ML e IB, 1 artigo utilizou NJ e MP, 1 usou ML, IB e MP e 1 ML, MP e NJ (figura 31). Cabe ressaltar que 6 artigos utilizaram mais de um método. Desses artigos, três utilizaram AIC ou BIC, mas não informaram se foi usada alguma ferramenta para tal. Os modelos evolutivos utilizados na Inferência Bayesiana foram: HKY (2), GTR (1), GTR + G (2), GTR + I + G (1) e um artigo não citou qual foi o modelo evolutivo utilizado (figura 32). Já nas RAF que utilizaram o método de Máxima Verossimilhança (figura 33), 1 artigo usou o modelo JC, 2 o K2P, 1 o T92, 1 o TrN, 3 o HKY, 11 o GTR + G, 5 o GTR + I + G, 2 não informaram qual foi o modelo utilizado. O modelo GTR puro não foi utilizado em nenhuma RAF. A escolha dos modelos pelos autores não foi citada, quando se utilizou métodos que dependem de modelos evolutivos.

Figura 31 – Métodos utilizados em RAFs que não usaram nenhuma ferramenta de escolha de modelo evolutivo

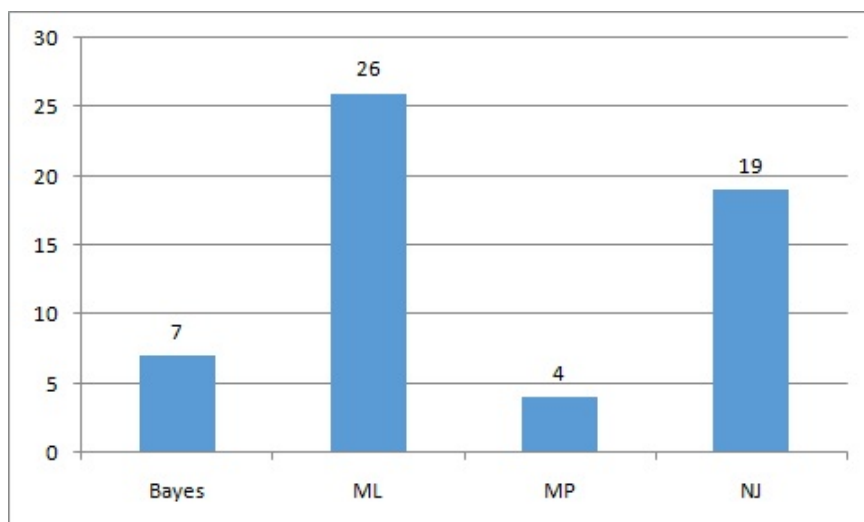


Figura 32 – Modelos utilizados em RAFs que não usaram nenhuma ferramenta de escolha de modelo evolutivo e utilizaram o método de inferência bayesiana

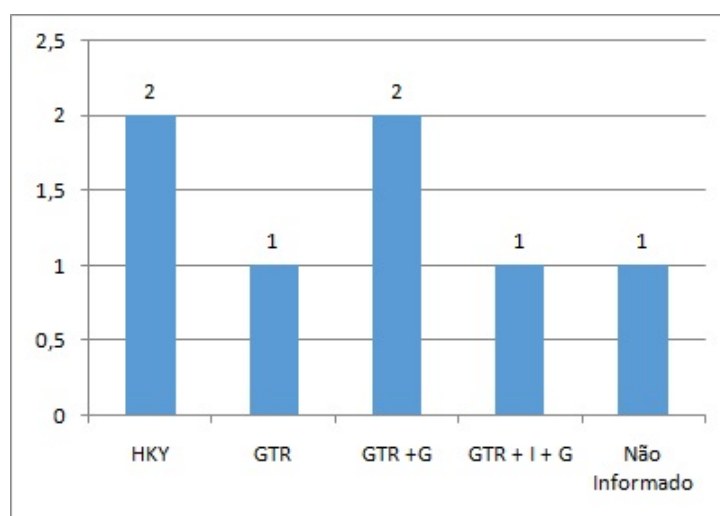
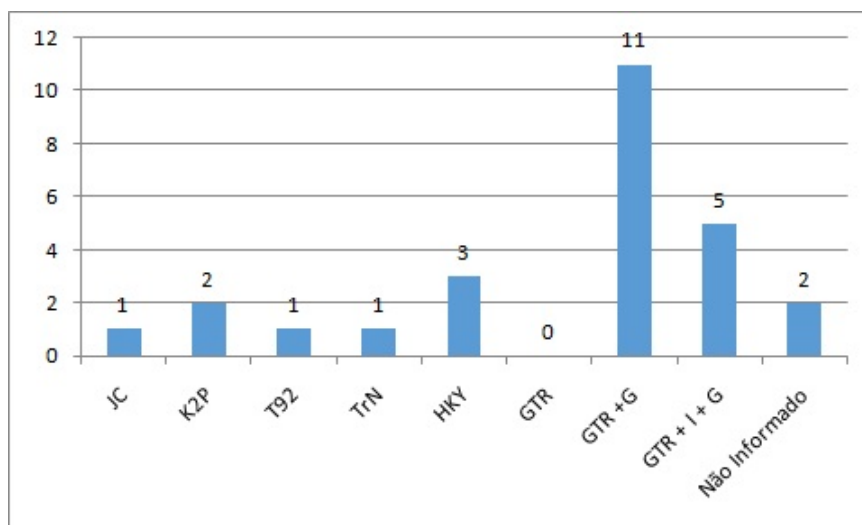


Figura 33 – Modelos utilizados em RAFs que não usaram nenhuma ferramenta de escolha de modelo evolutivo e utilizaram o método de MV



Com base nos dados levantados, constatou-se que os métodos mais aplicados são os de Máxima Verossimilhança, com 45% das RAFs e Inferência Bayesiana, com 30%. Os métodos de distância são mais utilizados quando a finalidade da RAF é somente criar uma topologia simples, corroborando com o que afirma (KELCHNER; THOMAS, 2006). Nos casos onde foram utilizados os métodos de distância não houve a preocupação com o modelo evolutivo, visto que não foi feita nenhuma análise prévia de qual o modelo mais adequado para aquele tipo de sequência. Nota-se também que a maioria das filogenias são feitas utilizando somente um método de RAF.

Em relação às publicações, notou-se que as revistas que não abordam diretamente a filogenia como tema principal tenderam a utilizar RAFs simples, onde geralmente utilizavam as topologias apenas com a finalidade de georreferenciamento, principalmente nas revistas que abordam doenças infecto contagiosas, ao passo que revistas que tem como tema central a genética ou evolução, tendem a utilizar inferências mais completas, estudando o tempo evolutivo das espécies e seus ancestrais.

A pesquisa bibliográfica realizada constatou que embora seja sinalizado que é fundamental utilizar uma ferramenta para obter o modelo evolutivo que melhor se adeque aos dados ainda em 2015, 35% dos autores não utilizam nenhum tipo de ferramenta para fazer isto.

Outro ponto importante é que em nenhum artigo foi feito o cálculo das distâncias entre as sequências para verificar se os dados são de baixa variabilidade.

Embora seja popularizado, o método AIC (76% de utilização, dentre as técnicas), apontado como uma boa alternativa pelos teóricos da filogenia, sempre orientou a escolha para modelos evolutivos mais complexos (o modelo GTR foi escolhido em 78% das inferências que utilizaram a técnica AIC para a escolha do modelo) sem levar em

consideração o tipo de aplicação que será utilizada para essa filogenia, que pode ser simples (onde a meta da inferência é uma topologia simples, não sendo importante o comprimento dos galhos ou a análise ancestral) ou complexa (que leva em consideração o tempo evolutivo e a análise ancestral).

Verificou-se também a existência de uma predileção por modelos mais ricos em parâmetros (GTR + G principalmente) quando os autores não utilizavam ferramentas para escolha de modelos evolutivos e a finalidade da filogenia era mais complexa, como por exemplo, coevolução. Isso indica que quando não se utiliza técnicas de seleção de modelos, existe uma tendência a escolher modelos mais ricos em parâmetros, para evitar a subparametrização.

Em (FELENSTEIN, 2004) afirma-se que quando se utilizam métodos diferentes para o mesmo dado não deve ser utilizado um programa consenso para realizar a comparação das árvores, existem outras alternativas para realizar esta comparação mas ainda continua um problema em aberto. Isto explica porque foram utilizadas diferentes maneiras para juntar os resultados.

Embora (RIBEIRO et al., 2012) aconselhe que para obter uma melhor RAF é necessário utilizar mais de um método para os mesmos dados, 67,15% somente utilizou um método para publicar seus resultados. Além disso, constatou-se que daqueles que utilizaram mais de um método, a maioria aplicaram o mesmo modelo evolutivo para métodos diferentes, fato que a literatura de filogenia (POSADA; BUCLEY, 2002) não aconselha pois a técnica *Bayes factor* terá melhor resultados ao aplicá-la a um método Bayesiano e a técnica AIC ou BIC terá melhor efeito ao ser aplicada em métodos baseados em MV.

4.3 Cálculo de distância para escolha simplificada de métodos e modelos evolutivos

Como foi dito anteriormente, no caso dos dados conter baixa variabilidade, a importância da adequação dos modelos não é tão grande, podendo ser utilizados modelos menos complexos (KELCHNER; THOMAS, 2006), o que torna o processo de RAF mais rápido, visto que retira o passo de utilizar os métodos para a escolha do modelo evolutivo, que é muito caro computacionalmente.

Uma maneira de verificar se os dados possuem baixa variabilidade é através do cálculo entre as distâncias das sequências e da taxa de transição/transversão. Segundo (JIN; NEI, 1990; RZHETSKY; NEI, 1995; RZHETSKY; SITNIKOV, 1996), o resultado desses cálculos podem informar o modelo mais simples que se adeque à sequência a ser realizada a inferência.

Esse cálculo pode ser inferido através dos diversos modelos evolutivos, porém (RZHETSKY; SITNIKOV, 1996) diz que a utilização do modelo Jukes-Cantor é suficiente para o cálculo da distância entre duas sequências.

(FELENSTEIN, 2004) descreve como calcular variabilidade (medida através da distância entre as sequências) das sequências alinhadas. O cálculo da distância entre duas sequências A e B (d_{AB}), para o modelo Jukes-Cantor se dá através da fórmula:

$$d_{AB} = -\frac{3}{4} \ln \left(1 - \frac{4}{3}d \right) \quad (67)$$

Onde d é a distância p entre duas sequências, dada por $d = i/c$, sendo i a quantidade de diferenças em sítios da mesma posição, entre duas sequências e c corresponde ao comprimento total da sequência.

Outro valor importante, segundo (JIN; NEI, 1990) é a taxa de transição/transversão, que é o cálculo da proporção de transições e transversões entre diferentes sítios em uma sequência. Esse valor pode ser obtido através da fórmula:

$$B = \alpha / (\alpha + \beta_1 + \beta_2) \quad (68)$$

Sendo α correspondente ao total de mudanças $A \leftrightarrow G, T \leftrightarrow C$ entre as mesmas posições duas sequências diferentes, β_1 correspondente às mudanças entre $A \leftrightarrow T, G \leftrightarrow C$ e β_2 corresponde às mudanças $A \leftrightarrow C, T \leftrightarrow G$ entre os sítios de duas sequências.

Cabe ressaltar que o valor de distância (d) utilizado é o maior valor de distância encontrado entre as sequências analisadas.

Em posse dos resultados, podemos verificar se existe a necessidade de se utilizar métodos e modelos mais complexos baseados nos seguintes valores (JIN; NEI, 1990):

- Quando a estimativa da distância utilizando o modelo Jukes-Cantor for menor que 0.1, pode-se utilizar o modelo JC69 e o método de distância é suficientemente bom para realizar a inferência.
- Se a D for maior que 0.1 e menor que 0.3, pode-se usar o modelo JC69, a menos que a taxa de transição/transversão B seja menor que 0.5, caso contrário, deve-se utilizar o modelo K2P.
- Quando a distância d é maior que 0.3, existe a necessidade de utilizar um modelo mais complexo.

A seguir será descrita a metodologia proposta para fazer RAF automática a partir da sequência alinhada.

4.4 Descrição da metodologia proposta

Com base no dito anteriormente a metodologia proposta possui como entrada as sequências alinhadas, o e-mail do usuário (será explicado na parte de implementação) e a escolha entre duas opções: filogenia simples ou filogenia complexa. Quer dizer o usuário vai ter a oportunidade de definir na entrada dos dados qual é o tipo de filogenia que ele quer. Se está interessado em uma filogenia que tenha a topologia e comprimentos de galhos mais aproximada possível significa que quer uma filogenia complexa. No caso de querer uma filogenia para fins que o anterior não seja importante então se encaixa em uma filogenia simples.

Na literatura de filogenia ([RZHETSKY; NEI, 1995](#); [RZHETSKY; SITNIKOV, 1996](#)), está claramente demonstrado que em sequências com baixa variabilidade entre os sítios, qualquer método é tão exato quanto outro. Ou seja basta simplesmente utilizar o método mais simples (pois é mais eficiente) que não vai ter perda de informação da RAF, corroborando com o que afirma ([KELCHNER; THOMAS, 2006](#)).

Nesse caso, se faz necessário verificar as distâncias entre cada sequência e tomar o maior valor entre elas. Sequências menos variáveis têm uma menor distância entre si e, por consequência maior homologia.

Nesse sentido, o passo inicial da metodologia proposta é calcular a distância d entre as sequências, do mesmo modo que a taxa de transição/transversão, para verificar se existe a necessidade de se utilizar um método de RAF mais complexo, bem como os modelos evolutivos, baseada na variabilidade dos sítios entre as sequências.

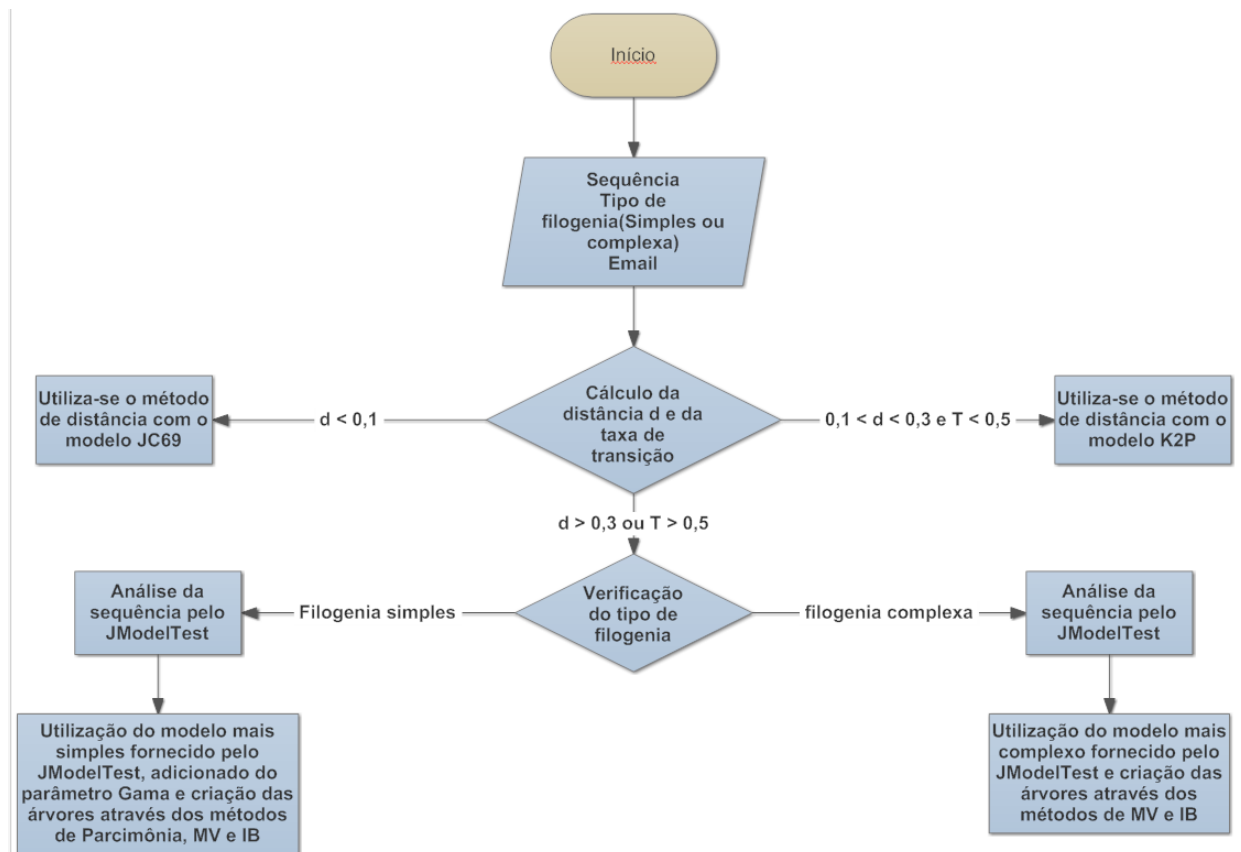
([KELCHNER; THOMAS, 2006](#)) afirma que caso a RAF não possua uma finalidade complexa, ou seja, se o usuário pretende fazer a inferência onde o comprimento dos galhos, assim como os ancestrais não sejam importantes, pode-se fazer o uso de modelos evolutivos menos complexos, nesse caso o sistema utilizará o modelo menos complexo do conjunto fornecido pela ferramenta de escolha de modelos, acrescido da taxa Gama. A utilização da taxa Gama, segundo ([TICONA, 2008](#); [POSADA; CRANDALL, 2001](#)) é necessária para melhor adequar o modelo escolhido ao conjunto de dados. De acordo com ([ASSIS, 2015](#)), em alguns casos onde a homologia entre as sequências é grande, o método de Parcimônia leva vantagem em relação aos métodos de MV e IB. Assim sendo, se faz necessário fazer reconstrução utilizando os métodos de MV, IB e Parcimônia, fornecendo ao usuário as três árvores geradas, para que o mesmo possa fazer a escolha que mais se adequa às suas necessidades.

Ainda de acordo com ([KELCHNER; THOMAS, 2006](#)), caso o comprimento dos galhos ou a análise ancestral seja importante, se faz necessário utilizar um modelo evolutivo mais rico em parâmetros. Nesse caso, o sistema utilizará o modelo evolutivo mais complexo dentre os fornecidos pela ferramenta de seleção de modelo e a RAF será

feita através de MV e IB.

Em todos os casos, o sistema fornecerá todas as árvores geradas, cabendo ao usuário escolher a que achar mais correta, dentre as árvores geradas pelos diferentes métodos. A metodologia é melhor ilustrada através da figura 34.

Figura 34 – Fluxograma detalhando a metodologia criada



4.5 Desenvolvimento da ferramenta

A ferramenta apresentada neste projeto disponibiliza as RAFs no email fornecido pelo usuário. A heurística, os parâmetros e a sintaxe do IgrafuWeb e de cada programa utilizado pelo IgrafuWeb foram devidamente estudados, através do manual dos mesmos e da documentação escrita por (NETO, 2015).

Cabe ressaltar que, com exceção do JModelTest, todos os softwares utilizados já estavam devidamente instalados e configurados no CACAU, sendo utilizados pela versão manual do IgrafuWeb.

Além dos métodos manuais (Inferência Bayesiana – disponibilizado através do software MrBayes, Máxima Verossimilhança – oferecido pelo programa PHYML, Distância – ofertado através do software Digradu e Máxima Parcimônia – ofertado através do software DNAPARS ou PROTPARS) disponíveis no menu principal, foi

adicionado um novo menu, que possibilita o usuário a fazer a RAF automaticamente, conforme figura 35.

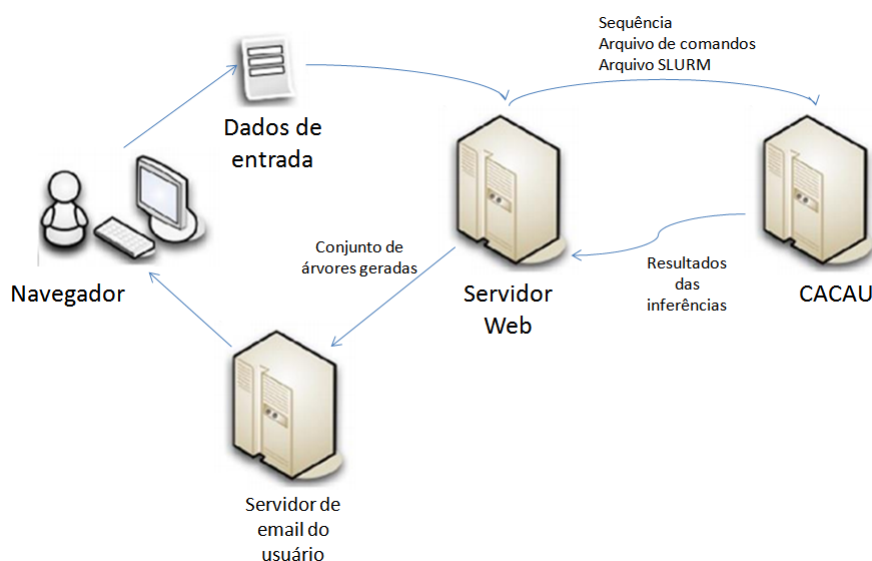
Figura 35 – Tela inicial do IgrafuWeb

The screenshot displays the IgrafuWeb web application interface. At the top, there is a navigation bar with tabs for 'Automatic', 'Bayesian', 'Likelihood', 'Distance', and 'Parsimony'. Below this, a breadcrumb trail shows 'Home » Automatic Selection'. The main heading is 'Automatic Selection of Method and Models'. A large box contains the following elements: a label 'Choose the kind of phylogeny' above a dropdown menu; a 'Sequence' section with two radio buttons, 'Simple Phylogeny' (selected) and 'Complex Phylogeny'; a button labeled 'Escolher arquivo' next to the text 'Nenhum arquivo selecionado'; an 'Email' label above a text input field; and at the bottom of the box, 'Execute' and 'Clear' buttons. The footer of the page includes copyright information: 'Copyright © 2016 by My Company. All Rights Reserved. Powered by [J8 Framework](#)'.

Pelo fato de o servidor web do CACAU (SIOUX) e os nós de processamento de dados do CACAU estarem em redes diferentes, foi necessário construir um procedimento para a troca de arquivos e comandos entre o servidor web e o servidor de controle dos procedimentos executados no CACAU. Assim sendo, uma vez carregados os dados no servidor web eles serão transferidos para o nó central de processamento e, quando terminada a execução são devolvidos os resultados ao servidor contendo as árvores geradas pelos métodos de RAF.

Uma vez devolvidos ao SIOUX, as árvores geradas são compactadas e enviadas ao email informado na tela inicial do módulo de seleção automática do IgrafuWeb. A figura 36 ilustra a transferência dos arquivos citados.

Figura 36 – Fluxo de execução e dados do IgrafuWeb



O passo inicial para realizar a RAF é a escolha da finalidade da filogenia e o arquivo de entrada, contendo sequências alinhadas (figura 37) de DNA, juntamente com o email.

Figura 37 – Exemplo de arquivo de entrada no formato NEXUS

```

#NEXUS
Begin data;
Dimensions ntax=12 nchar=898;
Format datatype=dna gap=-;
Matrix
Lemur_catta      AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGC
Homo_sapiens     AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGG
Pan              AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGA
Gorilla          AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCACGGA
Pongo            AAGCTTCACCGGCGCAACCACCCTCATGATTGCCATGGA
Hylobates        AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGA
Macaca_fuscata   AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGA
M._mulatta       AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGA
M._fascicularis  AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGG
M._sylvanus      AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCATGGA
Saimiri_sciureus AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGG
Tarsius_syricha  AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGC;
End;
  
```

Para acessar o CACAU via SIOUX é necessário se realizar uma autenticação de usuário via SSH (*Secure Shell*). Para esse fim, foi utilizado o mesmo perfil criado para a

utilização manual das ferramentas, criado por (NETO, 2015).

Uma vez passado os dados para a RAF, o sistema cria uma pasta temporária (nomeada com o email fornecido + a data e hora da submissão), contendo o arquivo de entrada e o log das atividades realizadas pelo sistema. Por exemplo, se o usuário fornece o email murilossantana@ymail.com às 11:00 do dia 12/05/2016, a pasta criada terá o nome murilossantana@ymail.com110012052016.

Depois de criada a pasta, o sistema faz a chamada da classe de calculo de distâncias entre sequências

4.5.1 Implementação do cálculo de distância entre sequências

Independentemente da escolha do tipo de filogenia, as sequências são enviadas para a classe CalcDistController, onde é feito o cálculo da distância entre as sequências, assim como as taxas de transição/transversão. Esse cálculo é feito no SIOUX, pelo fato de ser um cálculo simples e rápido.

Uma vez calculada a homologia entre as sequências (dada pela distância), o sistema pode fazer a escolha do modelo evolutivo, usando o JModelTest ou fazer a RAF utilizando o método de distância, como descrito na seção de proposta de metodologia.

4.5.2 Módulo de seleção de modelo evolutivo

O módulo de seleção de modelo evolutivo recebe o arquivo de sequência e faz a chamada via SSH do JModelTest, que analisa a sequência e devolve ao sistema o arquivo contendo os modelos evolutivos escolhidos. Os parâmetros utilizados na execução do JModelTest são:

- -AIC, que estima o o melhor modelo através do método AIC;
- -f, que inclui modelos com diferentes taxas de frequências;
- -g, que inclui os modelos com variação da taxa entre os locais e define o número de categorias;
- -i, que inclui modelos com proporção de sítios invariáveis;
- -d, que define o arquivo de entrada;
- -S NNI, que define a topologia inicial para o cálculo de Máxima Verossimilhança através de NNI;
- -o, que redireciona a saída para um arquivo;

Uma vez selecionado o conjunto de modelos para a sequência de entrada, o arquivo de saída do JModelTest é enviado para a pasta criada anteriormente no servidor web.

De posse do arquivo contendo o conjunto de modelos, o sistema fará uma chamada dos métodos já implementados no IgrafuWeb, de acordo com o tipo de filogenia pretendida.

4.5.3 Chamada aos métodos do IgrafuWeb

A utilização dos métodos de RAF, implementados no IgrafuWeb se deu através da chamada dos métodos da classe dos controladores de cada ferramenta. Nesse caso, ao invés de preencher os parâmetros da função de execução dos controladores através dos formulários das telas de cada método, os parâmetros são passados de maneira automática, dependendo do modelo escolhido e dos demais parâmetros a serem utilizados em cada RAF.

Cabe ressaltar que em inferências onde se utiliza mais de um método, as chamadas das diferentes ferramentas serão feitas em fila, para que não haja sobrecarga do processamento do CACAU. Nesse caso, outro método só pode ser executado depois o anterior enviar a sua árvore para o servidor web e a mesma já esteja salva na pasta temporária.

4.5.4 Execução

Após a análise da sequência e a escolha dos métodos e modelos, o sistema dispara uma série de rotinas para a conclusão do processo, são elas:

1. Montagem dos arquivos texto, de acordo à sintaxe dos programas, que servirão de base para execução dos métodos escolhido no CACAU;
2. Criação do arquivo Slurm, que levará em conta também o software escolhido, assim como a quantidade de sequência e seus tamanhos;
3. Transferência dos dois arquivos citados acima, junto com arquivo de sequência escolhido pelo usuário;
4. Execução de fato do programa escolhido. Esta execução é realizada no CACAU através do gerenciador de fila Slurm;
5. Retorno da árvore gerada através do método escolhido;
6. Envio do conjunto de árvores geradas para o email do usuário.

4.6 Testes de validação da ferramenta

Pelo fato de a metodologia proposta estar devidamente provados na literatura, não é necessário fazer testes exaustivos para verificar se a metodologia está correta. O que foi necessário fazer foi um teste de funcionamento da ferramenta implementada, para verificar se a mesma é executada corretamente de acordo com os parâmetros de entrada.

Para a validação da ferramenta desenvolvida, foram criados seis arquivos de entrada no formato fasta, contendo cada um deles duas sequências com 100 sítios cada uma. Para cada arquivo de entrada foi feito o cálculo da distância entre as sequências e das taxas de transição (fórmula 67 e fórmula 68, respectivamente) de maneira analítica e depois os arquivos criados foram executados pela ferramenta desenvolvida, tendo os seus resultados comparados para verificar a corretude do cálculo do programa e se o mesmo faz as chamadas dos métodos de maneira correta.

O primeiro arquivo de entrada preserva total identidade entre as sequências, que contém apenas adeninas. Esse arquivo serviu para testar arquivos com sequências idênticas. O valor da distância entre as sequências calculado analiticamente é 0, assim como o valor calculado pela ferramenta desenvolvida. O valor da taxa de transição calculado pela ferramenta é 0, assim como o valor calculado analiticamente. O método selecionado pela ferramenta foi o de Distância, utilizando o modelo JC69.

O segundo arquivo de entrada tem 10 sítios diferentes entre as duas sequências, sendo a primeira formada apenas por adeninas e a segunda sendo formada por 90 adeninas e 10 timinas. O valor da distância entre as sequências calculado analiticamente é 0.107325632, assim como o valor calculado pela ferramenta desenvolvida. O valor da taxa de transição calculado pela ferramenta é 0, o valor calculado analiticamente também é 0. O método selecionado pela ferramenta foi o de Distância, utilizando o modelo K2P.

O terceiro arquivo de entrada tem 23 sítios diferentes entre as duas sequências, sendo a primeira formada apenas por adeninas e a segunda sendo formada por 77 adeninas e 23 timinas. O valor da distância entre as sequências calculado analiticamente é 0.2746832962, assim como o valor calculado pela ferramenta desenvolvida. O valor da taxa de transição calculado pela ferramenta é 0, o valor calculado analiticamente também é 0. O método selecionado pela ferramenta foi o de Distância, utilizando o modelo K2P.

O quarto arquivo de entrada tem 25 sítios diferentes entre as duas sequências, sendo a primeira formada apenas por adeninas e a segunda sendo formada por 75 adeninas e 25 timinas. O valor da distância entre as sequências calculado analiticamente é 0.3040988311, assim como o valor calculado pela ferramenta desenvolvida. O valor

da taxa de transição calculado pela ferramenta é 0 e o valor calculado analiticamente também é 0. Conforme a metodologia proposta, foi verificada a finalidade da filogenia (simples ou complexa) e depois foi feita a chamada da ferramenta para a seleção do melhor modelo evolutivo para o conjunto de sequências.

O quinto arquivo de entrada tem 25 sítios diferentes entre as duas sequências, sendo a primeira formada apenas por adeninas e a segunda sendo formada por 80 adeninas, 10 timinas e 10 guaninas. O valor da distância entre as sequências calculado analiticamente é 0.2326161962, assim como o valor calculado pela ferramenta desenvolvida. O valor da taxa de transição calculado pela ferramenta é 0.5, o valor calculado analiticamente também é 0.5. Conforme a metodologia proposta, foi verificada a finalidade da filogenia (simples ou complexa) e depois foi feita a chamada da ferramenta para a seleção do melhor modelo evolutivo para o conjunto de sequências.

Depois de selecionados os métodos e os modelos evolutivos, a ferramenta fez a chamada dos programas que implementam os métodos de RAF determinados pela metodologia proposta e, por fim, faz o envio das árvores geradas para o email do usuário. Conforme a metodologia proposta, foi verificada a finalidade da filogenia (simples ou complexa) e depois foi feita a chamada da ferramenta para a seleção do melhor modelo evolutivo para o conjunto de sequências.

Foi verificado o perfeito funcionamento da ferramenta implementada, onde a mesma executa todos os cálculos com resultados idênticos aos resultados obtidos de forma analítica, com precisão de 10 casas decimais.

A ferramenta desenvolvida ainda não está disponível no site por haverem problemas de compatibilidade com a versão atual do IgrafuWeb e com os diversos formatos de sequências alinhadas existentes.

5 Conclusão e trabalhos futuros

Neste trabalho foi proposta uma metodologia para inferências filogenéticas de maneira automática que leva em conta a distância entre as sequências e a finalidade da filogenia.

Foi realizado um levantamento bibliográfico sobre como se realiza RAF atualmente. A análise dos dados permitiu concluir que apesar de existirem diversas ferramentas para a escolha de modelos, os pesquisadores ainda tendem a utilizar modelos superestimados, com o intuito de não incorrer em subestimação de modelo. Além disso, a maioria das publicações utiliza apenas um método para a inferência filogenética. Outra conclusão foi que nenhum autor analisou a finalidade da filogenia para a RAF, bem como a análise da distância entre as sequências, utilizando por muitas vezes modelos e métodos mais complexos do que o necessário.

O trabalho fornece uma ferramenta que realiza a escolha de métodos e modelos transparente aos usuários para realizar a RAF em um ambiente web de alta performance. Espera-se que as contribuições auxiliem e facilitem o processo de reconstrução de árvores filogenéticas e que novas soluções possam ser desenvolvidas a partir dos conceitos apresentados nesse trabalho.

Cabe salientar que o autor desconhece uma proposta similar à descrita neste trabalho, a qual abrange todas as fases de RAF, desde as sequências alinhadas até as árvores geradas e leva em consideração os principais métodos de RAF.

Sendo assim, para trabalhos futuros, com base no assunto desta dissertação, pode-se propor:

- Inserção de um módulo que avalie a qualidade das árvores geradas pela ferramenta;
- Incorporação da análise de sequências de proteínas no módulo de automatização de escolhas de métodos e modelos.
- Criação de perfis de usuários, com login e senha. O que permite a visualização do histórico de submissões daquele usuário, dando a possibilidade de o mesmo editar, consultar ou deletar as RAFs anteriormente realizadas;
- Incluir uma ferramenta que faça a junção de árvores criadas a partir de métodos diferentes.

Referências

- AKAIKE, H. Factor analysis and aic. **Psychometrika**, v. 52, p. 317–332, 1987.
- AMORIM; SOUZA, D. de. **Fundamentos de Sistemática Filogenética**. Brasil: HOLOS, 2002.
- ASSIS, L. C. S. Homology assessment in parsimony and model-based analyses: two sides of the same coin. **Cladistics**, v. 31, p. 315–320, 2015.
- BOLLBACK, J. P. Bayesian model adequacy and choice in phylogenetics. **Mol. Biol. Evol.**, 2002.
- BOTTU, G.; VAN-RANST, M.; LEMEY, P. **The Phylogenetic Handbook**. Cambridge, UK: Cambridge University Press, 2009. 723 p.
- BRITO, R. T. de. **Alinhamento de Sequências Biológicas**. 163 p. Mestrado — Instituto de Matemática e Estatística da Universidade de São Paulo, 2003.
- BUCKLEY, T. R.; CUNNINGHAM, C. W. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. **Mol. Biol. Evol.**, v. 4, n. 19, p. 394–405, 2002.
- CYBS, G. B. **Teste da Razão de Verossimilhança e seu Poder em Árvores Filogenéticas**. 225 p. Mestrado — Programa de Pós-Graduação em Matemática - Universidade Federal do Rio Grande do Sul, Porto Alegre - Rio Grande do Sul, 2009.
- DARRIBA, D.; DAOLLO, R.; POSADA, D. jmodeltest 2: more models, new heuristics and parallel computing. **Nature Methods**, 2012.
- DIAS, G.; TORRES, M.; GONÇALVES, G.; VIEIRA, C. Tool that integrates distance based programs for reconstructing phylogenetic trees. **Revista IEEE América Latina**, v. 9, n. 5, 2011.
- DURAND, D. The markov model of sequence evolution. **Computational Genomics and Molecular Biology**, p. 6, 2013.
- FELENSTEIN, J. software, **PHYLIP (Phylogeny Inference Package)**. 1993. Disponível em: <<http://evolution.genetics.washington.edu/phylip.html>>.
- FELENSTEIN, J. **Inferring Phylogenies**. Massachusetts - EUA: Sinauer Associates Inc., 2004.
- GONÇALVES, G. D. **Estudo de técnicas para melhorar o desempenho da reconstrução de árvores filogenéticas por análise bayesiana**. 73 p. Especialização — Universidade Estadual de Santa Cruz, 2008.
- GRIMMENT, G. R.; STIRZAKER, D. R. **Probability and Random Processes**. New York: Oxford University Press Inc., 1992.

- GUINDON, S.; DUFAYARD, J. F.; LEFORT, V.; ANISIMOVA, M.; HORDIJK, W.; GASCUEL, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of phym 3.0. **Systematic Biology**, v. 59, n. 3, 2010.
- GUINDON, S.; GASCUEL, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. **Systematic biology**, 2003.
- HASEGAWA, M.; KISHINO, H.; T., Y. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. **J. Mol. Evol.**, p. 160–174, 1985.
- HENNIG, W.; DWIGHT, D.; ZANGERL, R. **Phylogenetic Systematics**. Illinois - USA: University of Illinois Press, 1999. 263 p.
- HILLIER, F. S.; LIEBERMAN, G. J. **Introduction to operations research**. 7. ed. New York: McGraw-Hill Higher Education, 2001. 1214 p. ISBN 0072321695.
- HOLDER, M.; LEWIS, P. O. Phylogeny estimation: traditional and bayesian approaches. **Nature Reviews**, v. 4, p. 275–284, 2003.
- HOLMES, S. Statistics for phylogenetic trees. **Theoretical Population Biology**, v. 63, p. 17–32, 2003.
- HUELSENBECK, J. P.; CRANDALL, K. A. Phylogeny estimation and hypothesis testing using maximum likelihood. **Annu. Rev. Ecol. Syst.**, v. 28, p. 437–466, 1997.
- HUELSENBECK, J. P.; LARGET, B.; MILLER, R. E.; RONQUIST, F. Potential applications and pitfalls of bayesian inference of phylogeny. **Syst. Biol**, v. 51, n. 4, p. 673–688, 2002.
- HUELSENBECK, J. P.; RONQUIST, F. Mrbayes: Bayesian inference of phylogenetic trees. **Bioinformatics**, v. 17, n. 8, p. 754–755, 2001.
- HUELSENBECK, J. P.; RONQUIST, F. Mrbayes: Bayesian inference of phylogenetic trees. **Bioinformatics Applications Note**, v. 17, n. 8, p. 754–755, 2001.
- HYPÓLITO, E. B. **Uma resposta bayesiana ao Paradoxo de Suzuki**. Mestrado — Instituto de Matemática, Universidade Federal do Rio de Janeiro, 2005.
- IRESTEDT, M.; FJELDSA, J.; NYLANDER, J.; ERICSON, P. Phylogenetic relationships of typical antbirds (thamnophilidae) and test of incongruence based on bayes factors. **BioMed Central**, v. 4, n. 16, p. 16, 2004.
- JIN, L.; NEI, M. Limitations of the evolutionary parsimony method of phylogenetic analysis. **Mol. Biol. Evol.**, v. 7, n. 1, p. 82–102, 1990.
- JUKES, T.; CANTOR, C. Evolution of protein molecules. In **Mammalian protein metabolism**, p. 142–180, 1969.
- KEANE, T. M. **Computational methods for statistical phylogenetic inference**. 155 p. Doutorado — National University of Ireland, 2006.
- KELCHNER, S. A.; THOMAS, M. A. Model use in phylogenetics: nine key questions. **TRENDS in Ecology and Evolution**, v. 22, n. 02, p. 87–94, 2006.
- LEVIN, D. A.; PERES, Y.; WILMER, E. L. **Markov Chains and Mixing Times**. Oregon - USA: University of Oregon.

- MATIOLI, F. M. de C. F. Noções de filogenética molecular. **Biológico**, São Paulo-SP, p. 37–38, 2013.
- MININ, V.; ABDO, Z.; JOYCE, P. Performance-based selection of likelihood models for phylogeny estimation. **Systematic Biology**, v. 53, p. 674–683, 2003.
- MORESCHI, C. R. O que vem a ser o intervalo de confiança? **Metrologia e Instrumentação**, p. 78–82, 2010.
- MULAN, L. J. Multiple sequence alignment—the gateway to further analysis. **Brief Bioinform**, v. 2, n. 2, p. 303–305, 2002.
- NEI, M.; KUMAR, S. **Molecular Evolution and Phylogenetics**. New York: Oxford University Press, 2000.
- NETO, M. A. D. S. **Desenvolvimento De Servidor Web De Alto Desempenho Para Soluções De Reconstrução De Árvores Filogenéticas: Igrafuweb**. Dissertação (Mestrado) — Programa de Pós Graduação em Modelagem Comptacional - Universidade Estadual De Santa Cruz, 2015.
- PHYLIP. **The Newick Format**. 2016. Disponível em: <<http://evolution.genetics.washington.edu/phylip/newicktree.html>>.
- PINTO, J. F. da C. **Epidemiologia molecular do vírus da imunodeficiência humana do tipo i: métodos de inferência filogenética**. 69 p. — Escola Nacional De Saúde Pública Sérgio Arouca, Rio de Janeiro - RJ, 2004.
- POSADA, A.; CRANDALL, K. A. Modeltest:testing the model of dna substitution. **Bioinformatics Applications Note**, USA, p. 817–818, 1998.
- POSADA, D.; BUCLEY, T. R. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. **Syst. Biol.**, v. 53, n. 05, p. 793 – 808, 2002.
- POSADA, D.; CRANDALL, K. A. Selecting models of nucleotide substitution: an application to human immunodeficiency virus 1 (hiv-1). **Mol. Biol. Evol.**, p. 897–906, 2001.
- POSADA, D.; CRANDALL, K. A. Selecting the best-fit model of nucleotide substitution. **Syst. Biol**, USA, p. 580–601, 2001.
- PRADO, O. G. **Computação evolutiva empregada na reconstrução de Árvores Filogenéticas**. Mestrado — Faculdade de Engenharia Elétrica e de Computação (FEEC/Unicamp), Campinas – São Paulo – Brasil, 2001.
- RAFTERY, A. E. Bayes factors and bic: Comment on “a critique of the bayesian information criterion for model selection.”. **Sociol. Methods Res**, v. 27, p. 411–427, 1999.
- RIBEIRO, P. L.; RAPINI, A.; SILVA, U. C. S. e; BERG, C. van den. Using multiple analytical methods to improve phylogenetic hypotheses in minaria (apocynaceae). **Molecular Phylogenetics and Evolution**, 2012.
- RZHETSKY, A.; NEI, M. Tests of applicability of several substitution models for dna sequence data. **Mol. Biol. Evol.**, v. 12, n. 01, p. 131–151, 1995.

- RZHETSKY, A.; SITNIKOV, T. When is it safe to use an oversimplified substitution model in tree-making? **Mol. Biol. Evol.**, v. 13, n. 09, p. 1255–1265, 1996.
- SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. **Molecular biology and evolution**, v. 4, p. 406–425, 1987.
- SILVA, A. E. A. da. **Uma Abordagem Multi-Objetivo E Multimodal Para Reconstrução De Árvores Filogenéticas**. 178 p. Mestrado — Universidade Estadual De Campinas, Campinas – São Paulo – Brasil, 2007.
- SILVA, J. O. da. **Solução de alto desempenho para reconstrução de árvores filogenéticas usando o método de máxima verossimilhança**. 156 p. — Programa de Pós Graduação em Modelagem Comptacional - Universidade Estadual De Santa Cruz, Ilhéus - Bahia - Brasil, 2016.
- SOBRAL, F. L.; CIANCIARUSO, M. V. Estrutura filogenética e funcional de assembléias: (re)montando a ecologia de comunidades em diferentes escalas espaciais. **Biosci. J.**, v. 28, n. 4, p. 617 – 631, 2012.
- SULLIVAN, J.; SWOFFORD, D. L. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution patterns are violated? **Syst. Biol.**, v. 50, p. 723–729, 2001.
- TAMURA, K.; NEI, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial dna in humans and chimpanzees. **Mol. Biol. Evol.**, v. 3, n. 10, p. 512 – 526, 1993.
- TICONA, W. C. G. **Algoritmos evolutivos multi-objetivo para a reconstrução de árvores filogenéticas**. 134 p. Doutorado — Instituto de ciências Matemáticas e de Computação, Universidade de São Paulo, 2008.
- TRIOLA, M. F. **Elementary Statistics**. 9. ed. Boston - USA: Pearson, 2014. 259-267 p.
- VIEIRA, C. M. **Análise E Paralelização Do Modelgenerator: Um Programa Para Seleção De Modelos Evolutivos Para Reconstrução De Árvores Filogenéticas**. 48 p. — UNIVERSIDADE ESTADUAL DE SANTA CRUZ, Ilhéus - Bahia - Brasil, 2007.
- WEIR, B. S. **Genetic Data Analysis 2: Methods for Discrete Population Genetic Data**. 2. ed. Massachusetts - USA: Sinauer Associates, 1996.
- YANG, Z. Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: Approximate methods. **Journal of Molecular evolution**, v. 39, n. 3, p. 306, 314, 1994.
- YANG, Z. **Computational Molecular Evolution**. Oxford - New York: Oxford University Press, 2006.

Apêndices

APÊNDICE A – Lista das revistas e seus respectivos artigos

- AIDS Research and Human Retroviruses

Sequencing and Phylogenetic Analysis of Near Full-Length HIV-1 Subtypes A, B, G and Unique Recombinant AC and AD Viral Strains Identified in South Africa
- American Journal of Botany

The Roles Of History And Ecology In Chloroplast Phylogeographic Patterns Of The Bird-Dispersed Plant Parasite *Phoradendron Californicum* (Viscaceae) In The Sonoran Desert
- Applied and Environmental Microbiology

Implications of Genome-Based Discrimination between *Clostridium botulinum* Group I and *Clostridium sporogenes* Strains for Bacterial Taxonomy

Marine Cyanophages Demonstrate Biogeographic Patterns throughout the Global Ocean

Phylogenetically Distinct Phylotypes Modulate Nitrification in a Paddy soil

The Role of Host Phylogeny Varies in Shaping Microbial Diversity in the Hindguts of Lower Termites
- Arch Microbiol

Characterization of ‘*Candidatus Syngnamydia salmonis*’ (Chlamydiales, Simkaniaceae), a bacterium associated with epitheliocystis in Atlantic salmon (*Salmo salar* L.)
- Arch Virol

Molecular epidemiology of coxsackievirus A6 associated with outbreaks of hand, foot, and mouth disease in Tianjin, China, in 2013
- Biochemical Systematics and Ecology

Higher substitution rates and lower dN/dS for the plastid genes in Gnetales than other gymnosperms

The phylogeographic history of the self-pollinated herb *Tacca chantrieri* (Dioscoreaceae) in the tropics of mainland Southeast Asia

- Biochemistry genetic

Mitochondrial DNA-Based Analyses of Relatedness Among Turkeys, Meleagris Gallopavo
- BioMed Research International

Identification of Type II Interferon Receptors in Geese: Gene Structure, Phylogenetic Analysis, and Expression Patterns
- BioOne

Two Myxozoans from the Urinary Tract of Topsmelt, Atherinops affinis
- BMC Bioinformatics

leBIBIQBPP: a set of databases and a webtool for automatic phylogenetic analysis of prokaryotic sequences
- BMC Genomics

Plastome organization and evolution of chloroplast genes in Cardamine species adapted to contrasting habitats

Carotenoid biosynthetic genes in Brassica rapa: comparative genomic analysis, phylogenetic analysis, and expression profiling

Characterization and expression profiling of glutathione S-transferases in the diamondback moth, Plutella xylostella (L.)

Full genome SNP-based phylogenetic analysis reveals the origin and global spread of Brucella melitensis

Molecular and phylogenetic characterization of the homoeologous EPSP Synthase genes of allohexaploid wheat, Triticum aestivum (L.)

Next generation genome sequencing reveals phylogenetic clades with different level of virulence among Salmonella Typhimurium clinical human isolates in Hong Kong
- BMC Medicine

Emergence and potential for spread of Chikungunya virus in Brazil
- BMC Microbiology

Diversity and phylogenetic analysis of endosymbiotic bacteria of the date palm root borer Oryctes agamemnon (Coleoptera: Scarabaeidae)

In silico evolutionary analysis of Helicobacter pylori outer membrane phospholipase A (OMPLA)

- British Infection Association
Genotype shift in human coronavirus OC43 and emergence of a novel genotype by natural recombination
- Chinese Journal of Oceanology and Limnology
Extensive genetic divergence among *Diptychus maculatus* populations in northwest China*
- Comptes Rendus Biologies
Sequencing of the mitochondrial genome of the avocado lace bug *Pseudacysta perseae* (Heteroptera, Tingidae) using a genome skimming approach
- Current Microbiology Journal
Yeast Biogeography and the Effects of Species Recognition Approaches: The Case Study of Widespread Basidiomycetous Species from Birch Forests in Russia
- Elsevier
Asymmetry in genitalia does not increase the rate of their evolution
Cryptic species of hairworm parasites revealed by molecular data
Molecular phylogenetics and biogeography of the Neotropical skink genus *Mabuya* Fitzinger (Squamata: Scincidae) with emphasis on Colombian populations
- Enfermedades Infecciosas y Microbiología Clínica
Los estudios de resistencias a antirretrovirales como herramienta para el análisis de los clusters de transmisión del virus de la inmunodeficiencia humana
- European Journal of Protistology
Morphology, morphogenesis and molecular phylogeny of a soil ciliate, *Pseudouroleptus caudatus caudatus* Hemberger, 1985 (Ciliophora, Hypotricha), from Lhalu Wetland, Tibet
- Evolucionary Biology
A genome-scale mining strategy for recovering novel rapidly-evolving nuclear single-copy genes for addressing shallow-scale phylogenetics in *Hydrangea*
Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian *Eugongylus* group scincid lizards
Evolutionary history and leaf succulence as explanations for medicinal use in aloes and the global popularity of *Aloe vera*

Genetic and morphological variation in sexual and asexual parasitoids of the genus *Lysiphlebus* – an apparent link between wing shape and reproductive mode

- Fisheries Research

Genetic structure and consequences of stock exploitation of *Chrysoblephus puniceus*, a commercially important sparid in the South West Indian Ocean

- Florida Entomological Society

DNA Barcoding and Phylogenetic Relationships of *Spodoptera litura* and *S. exigua* (Lepidoptera: Noctuidae)

- Fungal Diversity

Fragiliporiaceae, a new family of Polyporales (Basidiomycota)

- fungal Ecology

Marine fungal communities in water and surface sediment of a sea cucumber farming system: habitat-differentiated distribution and nutrients driving succession

- Gene

Complete mitochondrial genome of *Cuora trifasciata* (Chinese three-striped box turtle), and a comparative analysis with other box turtles

Divergence and population traits in evolution of the genus *Pisum* L. as reconstructed using genes of two histone H1 subtypes showing different phylogenetic resolution

Genome-wide identification and analysis of the MADS-box gene family in apple

Identification of a CONSTANS homologous gene with distinct diurnal expression patterns in varied photoperiods in ramie (*Boehmeria nivea* L. Gaud)

Internal transcribed spacer (ITS) evolution in populations of the hyperparasitic European mistletoe pathogen fungus, *Sphaeropsis visci* (Botryosphaeriaceae): The utility of ITS2 secondary structures

Phylogenetic and stress-responsive expression analysis of 20 WRKY genes in *Populus simonii* x *Populus nigra*

The mitochondrial genome of *Dastarcus helophoroides* (Coleoptera: Bothridiidae) and related phylogenetic analyses

Transcriptome analysis of the plateau fish (*Triplophysa dalaica*): Implications for adaptation to hypoxia in fishes

What's behind these scales? Comments to "The complete mitochondrial genome of Temminck's ground pangolin (*Smutsia temminckii*; Smuts, 1832) and phylogenetic position of the Pholidota (Weber, 1904)"

- Genome Biology and evolution

Draft Genomes, Phylogenetic Reconstruction, and Comparative Genomics of Two Novel Cohabiting Bacterial Symbionts Isolated from *Frankliniella occidentalis*

- Global Ecology and Biogeography

Phylogenetic relatedness within Neotropical fern communities increases with soil fertility

Explaining the variation in impacts of non-native plants on local-scale species richness: the role of phylogenetic relatedness

- Harmful Algae

Cyanotoxin production and phylogeny of benthic cyanobacterial strains isolated from the northeast of Brazil

- Helgoland Marine Research

Description of two free-living nematode species of *Halomonhystera disjuncta* complex (Nematoda: Monhysterida) from two peculiar habitats in the sea

- Hindawi Publishing Corporation

Sequence Variation in HSP40 Gene among 16 *Toxoplasma gondii* Isolates from Different Hosts and Geographical Locations

- Ichthyological Research Journal

No genetic deviation between two morphotypes of the snipefishes (Macroramphosidae: *Macroramphosus*) in Japanese waters

- Infection, Genetics and Evolution

Molecular diagnosis and genotype analysis of *Giardia duodenalis* in asymptomatic children from a rural area in central Colombia

- Infectious Diseases

Molecular epidemiology and phylogenetic analysis of Hepatitis B virus in a group of migrants in Italy

Phylogenetic analysis of dengue virus reveals the high relatedness between imported and local strains during the 2013 dengue outbreak in Yunnan, China: a retrospective analysis

Spatial, temporal and genetic dynamics of highly pathogenic avian influenza A (H5N1) virus in China

- International Journal for Parasitology

Morphology and phylogeny of *Agmasoma penaei* (Microsporidia) from the type host, *Litopenaeus setiferus*, and the type locality, Louisiana, USA

- Invertebrate Microbiology

Microbial Associates of the Vine Mealybug *Planococcus ficus* (Hemiptera: Pseudococcidae) under Different Rearing Conditions

- Journal of Applied Phycology

Endemic *Pyropia* species (Bangiales, Rhodophyta) from the Gulf of California, Mexico

- Journal of Clinical Microbiology

Differential Single Nucleotide Polymorphism-Based Analysis of an Outbreak Caused by *Salmonella enterica* Serovar Manhattan Reveals

Epidemiological Details Missed by Standard Pulsed-Field Gel Electrophoresis

- Journal of Experimental Zoology

Which Came First: The Lizard or the Egg? Robustness in Phylogenetic Reconstruction of Ancestral States

- Journal of Genetics

The Dicer-like, Argonaute and RNA-dependent RNA polymerase gene families in *Populus trichocarpa*: gene structure, gene expression, phylogenetic analysis and evolution

- Journal of Medical Virology

Genetic Changes in Influenza A(H3N2) Viruses Circulating During 2011 to 2013 in Northern India (Lucknow)

- Journal of Molecular Neuroscience

Vertebrate Paralogous CRMPs in Nervous System: Evolutionary, Structural, and Functional Interplay

- Journal of Shellfish Research

Phylogeny and phylogeography of the geoduck *panopea* (bivalvia: hiatellidae)

- Malaria Journal

Diversity of malaria parasites in great apes in Gabon

- Mammalian Biology

Mitochondrial sequences yield new insight into the Quaternary history of the edible dormouse on the landbridge Adriatic islands

- Marine Micropaleontology

Re-discovery of a “living fossil” coccolithophore from the coastal waters of Japan and Croatia

- Memórias do Instituto Oswaldo Cruz

Detection of Oropouche virus segment S in patients and in *Culex quinquefasciatus* in the state of Mato Grosso, Brazil

- Molecular Phylogenetics and Evolution

Underground evolution: New roots for the old tree of lumbricid earthworms

Multilocus phylogenetic analysis reveals the monophyly of a recircumscribed papilionoid legume tribe Diocleae with well-supported generic relationships

Applying species-tree analyses to deep phylogenetic histories: Challenges and potential suggested from a survey of empirical phylogenetic studies

Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased

Multilocus phylogenetic analyses of Hispaniolan and Bahamian trunk anoles (distichus species group)

Phylogenetic analyses provide insights into the historical biogeography and evolution of *Brachyrhaphis* fishes

Phylogenetic analysis of molecular and morphological data highlights uncertainty in the relationships of fossil and living species of Elopomorpha (Actinopterygii: Teleostei)

- Nature

Dissemination, divergence and establishment of H7N9 influenza viruses in China

- Naunyn-Schmiedeberg's Arch Pharmacol

Nme family of proteins—clues from simple animals

- New Phytologist

Evolution of selenium hyperaccumulation in *Stanleya*

- Organisms Diversity & Evolution

New Ichthyophis species from Indochina (Gymnophiona, Ichthyophiidae): 1. The unstriped forms with descriptions of three new species and the redescriptions of *I. acuminatus* Taylor, 1960, *I. youngorum* Taylor, 1960 and *I. laosensis* Taylor, 1969

- PaleoWorld

Early divergence dates of demosponges based on mitogenomics and evaluated fossil calibrations

- Plant Systematics and Evolution

Evaluation of DNA barcode candidates for the discrimination of the large plant family Apocynaceae

Phylogeny, karyotype evolution and taxonomy of *Crocus* series Verni (Iridaceae)

- Plos One

A Molecular Phylogeny of the Chalcidoidea

Barcoding of Ancient Lake Ostracods

Efficient Detection of Novel Nuclear Markers for Brassicaceae by Transcriptome Sequencing

Evaluation of a Phylogenetic Marker Based on Genomic Segment B of Infectious Bursal Disease Virus: Facilitating a Feasible

Incorporation of this Segment to the Molecular Epidemiology Studies for this Viral Agent

Evolutionary History of the Live-Bearing Endemic *Allotoca diazi* Species Complex (Actinopterygii, Goodeinae): Evidence of Founder Effect Events in the Mexican Pre-Hispanic Period

Exploring Genomic, Geographic and Virulence Interactions among Epidemic and NonEpidemic St. Louis Encephalitis Virus (Flavivirus) Strains

Historical Isolation of the Galápagos Carpenter Bee (*Xylocopa darwini*) despite Strong Flight Capability and Ecological Amplitude

Initial Evidence for Adaptive Selection on the NADH Subunit Two of Freshwater Dolphins by Analyses of Mitochondrial Genomes

Molecular Phylogeny of the Cliff Ferns (Woodsiaceae: Polypodiales) with a Proposed Infrageneric Classification

Phylogenetic and Molecular Clock Analysis of Dengue Serotype 1 and 3 from New Delhi, India

Reconstruction of the Evolutionary Dynamics of A(H3N2) Influenza Viruses Circulating in Italy from 2004 to 2012

Signatures of Natural Selection at the FTO (Fat Mass and Obesity Associated) Locus in Human Populations

The Complete Genome Phylogeny of Geographically Distinct Dengue Virus Serotype 2 Isolates (1944-2013) Supports Further Groupings within the Cosmopolitan Genotype

Genetic Diversity of the Invasive Gall Wasp *Leptocybe invasa* (Hymenoptera: Eulophidae) and of its *Rickettsia* Endosymbiont, and Associated Sex-Ratio Differences

Comparative and Evolutionary Analyses of *Meloidogyne* spp. Based on Mitochondrial Genome Sequences

Genetic and Genomic Diversity Studies of Acacia Symbionts in Senegal Reveal New Species of *Mesorhizobium* with a Putative Geographical Pattern

Mitochondrial DNA Genomes Organization and Phylogenetic Relationships Analysis of Eight Anemonefishes (Pomacentridae: Amphiprioninae)

Molecular and Morphological Analyses Reveal Phylogenetic Relationships of Stingrays Focusing on the Family Dasyatidae (Myliobatiformes)

Molecular Phylogenetic Analysis of NonSexually Transmitted Strains of *Haemophilus ducreyi*

Molecular Phylogenetic Analysis of Ballistoconidium-Forming Yeasts in Trichosporonales (Tremellomycetes): A Proposal for *Takashimella* gen. nov. and *Cryptotrichosporon tibetense* sp. nov.

Morphological Characters Are Compatible with Mitogenomic Data in Resolving the Phylogeny of Nymphalid Butterflies (Lepidoptera: Papilionoidea: Nymphalidae)

Phylogenetic Analysis of Hepatitis B Virus Genotypes Circulating in Different Risk Groups of Panama, Evidence of the Introduction of Genotype A2 in the Country

Spatiotemporal Evolution of *Calophaca* (Fabaceae) Reveals Multiple Dispersals in Central Asian Mountains

The Combination of Phylogenetic Analysis with Epidemiological and Serological Data to Track HIV-1 Transmission in a Sexual Transmission Case

The Complete Mitochondrial Genome of *Corizus tetraspilus* (Hemiptera: Rhopalidae) and Phylogenetic Analysis of Pentatomomorpha

The Influence of Hepatitis C Virus Genetic Region on Phylogenetic Clustering Analysis

The Mitochondrial Genomes of *Aquilafasciata* and *Buteo lagopus* (Aves, Accipitriformes): Sequence, Structure and Phylogenetic Analyses

Whole Genome-Sequencing and Phylogenetic Analysis of a Historical Collection of *Bacillus anthracis* Strains from Danish Cattle

- PNAS

Error, signal, and the placement of Ctenophora sister to all other animals

The two-domain tree of life is linked to a new root for the Archaea

- Primates

Molecular phylogenetics and phylogeography of all the *Saimiri* taxa (Cebidae, Primates) inferred from mt COI and COII gene sequences

- Scientia Horticulturae

Utility of ITS2 sequence data of nuclear ribosomal DNA: Molecular evolution and phylogenetic reconstruction of *Lathyrus* spp.

- Scientific Reports

Multiple Sources of Infection and Potential Endemic Characteristics of the Large Outbreak of Dengue in Guangdong in 2014

- Springer International Publishing

Phylogeographic analysis of genus *Herichthys* (Perciformes:Cichlidae), with descriptions of *Nosferatu* new genus and *H. tepehua* n. sp.

- Springer Science

Morphological and molecular characterisation of *Ditrachybothridium macrocephalum* Rees, 1959 (Cestoda: Diphyllidea) from *Galeus melastomus* Rafinesque in the Western Mediterranean

- The Journal of Eukaryotic Microbiology

Phylogenetic Analysis and the Evolution of the 18S rRNA Gene Typing System of *Acanthamoeba*

- Transboundary and Emerging Diseases

Evolutionary and Ecological Dynamics of Transboundary Disease Caused by H5N1 Virus in Southeast Asia

- Tree Genetics & Genomes

Chloroplast haplotypes suggest preglacial differentiation and separate postglacial migration paths for the threatened North American forest tree *Juglans cinerea* L.

Phylogeography of *Quercus glauca* (Fagaceae), a dominant tree of East Asian subtropical evergreen forests, based on three chloroplast DNA interspace sequences

- Veterinary Microbiology

Host-specificity of *Staphylococcus aureus* causing intramammary infections in dairy animals assessed by genotyping and virulence genes

- Veterinary Research

Molecular characterization and phylogenetic analysis of transmissible gastroenteritis virus HX strain isolated from China

- Virology Journal

Insights into the evolutionary history of Japanese encephalitis virus (JEV) based on whole-genome sequences comprising the five genotypes

Phylogenetic analysis of avian infectious bronchitis virus S1 glycoprotein regions reveals emergence of a new genotype in Moroccan broiler chicken flocks

Phylogenetic analysis of eight sudanese camel contagious ecthyma viruses based on B2L gene sequence

Phylogenetic and recombination analysis of Tobacco bushy top virus in China

Sequencing and phylogenetic analysis of the gp51 gene from Korean bovine leukemia virus isolates

- Zoologica Scripta

Multiple reversals of chirality in the land snail genus *Albinaria* (Gastropoda, Clausiliidae)

Molecular phylogeny of the genus *Helix* (Pulmonata:Helicidae)