



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

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

**MANOEL ALVES DE SOUZA NETO**

**DESENVOLVIMENTO DE SERVIDOR WEB DE ALTO DESEMPENHO PARA SOLUÇÕES**  
**DE RECONSTRUÇÃO DE ÁRVORES FILOGENÉTICAS: IGRAFUWEB**

**PPGMC-UESC**

**ILHÉUS – BA**

**2015**

**MANOEL ALVES DE SOUZA NETO**

**DESENVOLVIMENTO DE SERVIDOR WEB DE ALTO DESEMPENHO PARA  
SOLUÇÕES DE RECONSTRUÇÃO DE ÁRVORES FILOGENÉTICAS:  
IGRAFUWEB**

**PPGMC – UESC**

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 a obtenção do Grau de Mestre em Modelagem Computacional em Ciência e Tecnologia.

Orientadora: Prof.<sup>a</sup> Dra. Martha Ximena Torres Delgado

Ilhéus – BA

**2015**

S729

Souza Neto, Manoel Alves de.

Desenvolvimento de servidor web de alto desempenho para soluções de reconstrução de árvores filogenéticas : Igrafuweb / Manoel Alves de Souza Neto. - Ilhéus : UESC, 2015.

Xv, 133f. : 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.

1. Igrafuweb (Programa de computador).
  2. Serviços da web.
  3. Filogenia – Modelos matemáticos.
  4. Biologia molecular – Modelos matemáticos.
- I. Delgado, Martha Torres Delgado.  
II. Título.

CDD – 006.78

**MANOEL ALVES DE SOUZA NETO**

**DESENVOLVIMENTO DE SERVIDOR WEB DE ALTO DESEMPENHO PARA  
SOLUÇÕES DE RECONSTRUÇÃO DE ÁRVORES FILOGENÉTICAS:**

**IGRAFUWEB**

**PPGMC – UESC**

Ilhéus – BA, 8 de julho de 2015

Comissão Examinadora

*Martha X. M. Delgado*

**Prof.<sup>a</sup> Dra. Martha Ximena Torres Delgado**

UESC

(Orientadora)



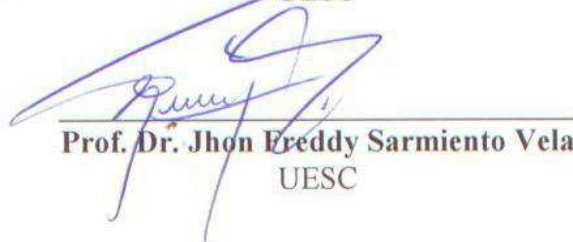
**Prof. Dr. Diego Gervasio Frías Suárez**

UNEB



**Prof. Dr. Luciano Ângelo de Souza Bernardes**

UESC



**Prof. Dr. Jhon Freddy Sarmiento Vela**

UESC

## DEDICATÓRIA

*A minha família...*

## **AGRADECIMENTOS**

- A minha professora e orientadora Martha Ximena, pela orientação e sapiência.
- Aos professores do curso, em especial a Francisco Bruno e Dany Sanches, pelo apoio e suporte nestes dois longos anos.
- A minha esposa que com amor e carinho me deu forças para enfrentar as dificuldades e superá-las.
- A minha mãe, a quem devo minha formação profissional, social e cultural.
- A meus amigos e familiares, por ter me incentivado nos momentos difíceis.
- A meus colegas do curso, pelo apoio em todo momento.
- Aos colegas do NBCGIB, pelo suporte e ajuda.

## PENSAMENTO

*“A menos que modifiquemos a nossa maneira de pensar, não seremos capazes de resolver os problemas causados pela forma como nos acostumamos a ver o mundo.”*

*Albert Einstein*

## RESUMO

As árvores filogenéticas são representações gráficas que explicam a história evolutiva das espécies. A construção destas árvores demanda muito recurso computacional, além de exigir a utilização de técnicas algorítmicas específicas que se comportem de forma adequada aos dados de entrada. Tais técnicas são representadas através de alguns métodos, associados a um modelo evolutivo de substituição de bases.

Existem diversos softwares destinados a inferir filogenias que utilizam tais métodos, porém, poucos foram projetados para utilização em um único ambiente. Além do mais, muitos deles exigem conhecimentos tecnológicos avançados, visto que, geralmente, não apresentam uma interface gráfica agradável ao usuário, sendo muitas vezes disponibilizados através de linhas de comandos. Por estes motivos, que se faz necessário o desenvolvimento de sistemas simples e compactos que possibilitem a integração de vários programas filogenéticos.

Sendo assim, idealizou-se a construção de um sistema Web destinado a Reconstrução de Árvores Filogenéticas (RAF), utilizando computação de alto desempenho, através dos métodos de Distância, Máxima Parcimônia, Máxima Verossimilhança e Inferência Bayesiana, com a possibilidade de visualização gráfica do resultado final.

**Palavras chaves:** Reconstrução de Árvores Filogenéticas, Métodos de Distância, Máxima Parcimônia, Máxima Verossimilhança, Inferência Bayesiana, Modelo Evolutivo.



## ABSTRACT

Phylogenetic trees are graphical representations explaining the evolutionary history of the species. The construction of these trees requires a lot of computational resources, and require the use of specific algorithmic techniques to behave appropriately to input data. Such techniques are represented by some methods, associated with an evolutionary model base substitution.

There are several software packages designed to infer phylogenies using such methods, however, few have been designed for use in a single environment. Moreover, many of them require advanced technological knowledge, as generally do not have a nice graphical user interface, often available through command lines. For these reasons, it is necessary to develop simple and compact systems that enable the integration of various phylogenetic programs.

Thus, the idealized construction was a Web system for Reconstruction of phylogenetic trees (LAR) using high-performance computing, by distance of methods, Maximum Parsimony, Maximum Likelihood and Bayesian inference, with the possibility of graphical display final result.

**Keywords:** Reconstruction of Phylogenetic Trees, Distance Methods, Maximum Parsimony, Maximum Likelihood, Bayesian Inference, Evolutionary Model.

## LISTA DE FIGURAS

Figura 1 - Exemplo descritivo de uma árvore filogenética. ....	6
Figura 2 - Exemplo de uma árvore filogenética sem raiz. ....	6
Figura 3 - Possíveis topologias sem raiz para 4 espécies. ....	7
Figura 4 - Possíveis topologias com raiz para 4 espécies. ....	8
Figura 5 - Representação gráfica da árvore filogenética (B,(A,C,E),D); padrão Newick..	8
Figura 6 - Representação do padrão Newick com comprimento dos ramos. ....	9
Figura 7 - Matriz de alinhamento de sequências de DNA, na qual as linhas representam as espécies e as colunas os sítios (base nucleotídica para determinada posição homóloga entre as espécies). ....	10
Figura 8 - Representação gráfica das mudanças de estados do modelo HMM. ....	26
Figura 9 - A árvore a é inicial, tem formato estrela e centro representado por G. A árvore b recebeu um novo nó, H, o qual liga os vizinhos A e D. ....	30
Figura 10 - Topologias (sem raiz) possíveis para 4 espécies (a, b, c e d). ....	34
Figura 11 - Identificação da Topologia A, exibindo as alterações nucleotídicas através dos traços que cortam os ramos (total de 7 mudanças). ....	35
Figura 12 - Representação gráfica da topologia B, totalizando 8 mudanças nucleotídicas através dos cortes nos ramos. ....	36
Figura 13 - Topologia C, com suas respectivas mudanças nucleotídicas, identificadas através dos traços que cortam os ramos, com o valor total de 8. ....	36
Figura 14 – Exemplo do Adição por passos. ....	38
Figura 15 – Exemplo do NNI. ....	39
Figura 16 – Exemplo do SPR. ....	39
Figura 17 – Exemplo do TBR. ....	40
Figura 18 - Árvore para exemplificar o cálculo de verossimilhança. ....	42
Figura 19 - Árvore inicial do PHYML. ....	54
Figura 20 - Otimização do galho U – V da árvore da Figura 17. ....	55
Figura 21 - Exemplo de arquivo de saída do MrBayes. ....	57
Figura 22 - Diagrama do MCMC implementado no MrBayes. ....	58
Figura 23 - Página inicial apresentado a estrutura organizacional do site e uma breve explicação do IgrafuWeb. ....	65
Figura 24 - Exemplo de arquivo (formato NEXUS) de sequências de DNA alinhadas....	65
Figura 25 – Campo da aba Sequence do MrBayes. ....	66
Figura 26 - Definição dos parâmetros para o modelo de DNA do MrBayes. ....	68
Figura 27 - Configurações avançadas do modelo do MrBayes para sequências de DNA. .....	70
Figura 28 - Definição dos parâmetros do modelo do MrBayes para sequências de Proteína. ....	71
Figura 29 - Campos da aba MCMC do MrBayes. ....	72
Figura 30 - Configurações avançadas do MCMC do MrBayes. ....	74
Figura 31 - Configurações da aba Tree do MrBayes. ....	75
Figura 32 - Configurações da aba Parameter (Sumamarization) do MrBayes. ....	77
Figura 33 - Configurações da aba Trees (Sumamarization) do MrBayes. ....	77
Figura 34 - Configurações da aba Sequence do PHYML. ....	79
Figura 35 - Configurações da aba DNA (Model) do PHYML. ....	80
Figura 36 - Configurações da aba Protein (Model) do PHYML. ....	81

<b>Figura 37 - Configurações da aba Tree do PHYML. ....</b>	<b>82</b>
<b>Figura 38 - Configurações da aba Bootstrap do PHYML. ....</b>	<b>83</b>
<b>Figura 39 - Configurações da aba Sequence do Digradu. ....</b>	<b>84</b>
<b>Figura 40 - Configurações da aba DNA (Model) do Digradu. ....</b>	<b>85</b>
<b>Figura 41 - Configurações da aba Protein (Model) do Digradu. ....</b>	<b>86</b>
<b>Figura 42 - Configurações da aba Sequence (Bootstrap-Seqboot) do Digradu. ....</b>	<b>88</b>
<b>Figura 43 - Configurações da aba Parameters (Bootstrap-Seqboot) do Digradu. ....</b>	<b>88</b>
<b>Figura 44 - Configurações da aba Bootstrap-Consense do Digradu. ....</b>	<b>89</b>
<b>Figura 45 - Configurações da aba Sequence do DNAPARS ou PROTPARS. ....</b>	<b>91</b>
<b>Figura 46 - Configurações da aba Tree do DNAPARS ou PROTPARS. ....</b>	<b>92</b>
<b>Figura 47 - Configurações da aba Bootstrap do DNAPARS ou PROTPARS. ....</b>	<b>93</b>
<b>Figura 48 - Configurações da aba Options do DNAPARS ou PROTPARS. ....</b>	<b>94</b>
<b>Figura 49 - Diagrama de Fluxo da execução do IgraduWeb. ....</b>	<b>95</b>
<b>Figura 50 - Filogenia da Solanum com sequências do gene COSII_At5g14320 para dados de batata. ....</b>	<b>117</b>
<b>Figura 51 - Resultado da filogenia da Solanum com sequências do gene COSII_At5g14320 utilizando o Método de Distância, mais precisamente, o Digradu, através do IgraduWeb. Os números nas linhas indicam o comprimento dos ramos. ....</b>	<b>118</b>
<b>Figura 52 - Resultado da filogenia da Solanum com sequências do gene COSII_At5g14320 utilizando o método de Máxima Parcimônia, mais precisamente, o DNAPARS (PHYLIP), através do IgraduWeb. Os números nas linhas indicam o comprimento dos ramos. ....</b>	<b>119</b>
<b>Figura 53 - Resultado da filogenia da Solanum com sequências do gene COSII_At5g14320 utilizando a máxima verossimilhança, mais precisamente, o PHYML através do IgraduWeb. Os números nos nós indicam valores de bootstrap. ....</b>	<b>121</b>
<b>Figura 54 - Resultado da filogenia da Solanum com sequências do gene COSII_At5g14320 utilizando a inferência bayesiana, mais precisamente, o MrBayes através do IgraduWeb. Os números nos galhos indicam os comprimentos dos ramos. ....</b>	<b>122</b>

## LISTA DE TABELAS

<b>Tabela 1 - Número de topologias possíveis sem e com raiz.....</b>	<b>7</b>
<b>Tabela 2 - Alinhamento de 4 sequências de DNA. ....</b>	<b>34</b>
<b>Tabela 3 - Comparação dos parâmetros do MrBayes.....</b>	<b>101</b>
<b>Tabela 4 - Comparação das opções dos parâmetros do PHYML disponibilizadas pelos Web Services estudados. ....</b>	<b>103</b>
<b>Tabela 5 - Comparação dos parâmetros do Digrafu.....</b>	<b>107</b>
<b>Tabela 6 - Comparação dos parâmetros do DNAPARS e PROTPARS.....</b>	<b>110</b>
<b>Tabela 7 - Tempo de Execução do PHYML para sequências genéticas de DNA com 20 espécies e 3768 sítios, utilizando o modelo HKY85 e bootstrap com valor igual a 100. ....</b>	<b>113</b>
<b>Tabela 8 - Tempo de Execução do MrBayes para sequências genéticas de DNA com 20 espécies e 3768 sítios, utilizando o modelo GTR, com 100.000 gerações e com taxa de heterogeneidade InvGamma. ....</b>	<b>114</b>
<b>Tabela 9 - Resumo dos parâmetros de execução do PHYML após o processamento no IgrafuWeb. ....</b>	<b>120</b>
<b>Tabela 10 - Resumo dos parâmetros de execução do MrBayes após o processamento no IgrafuWeb. ....</b>	<b>123</b>
<b>Tabela 11 - Valor de K retornado pelo Ktreedist ao comparar a árvore referência (Figura 50) com todas as árvores resultantes do processamento do IgrafuWeb.....</b>	<b>124</b>
<b>Tabela 12- Comparativo dos tamanhos dos galhos. ....</b>	<b>125</b>

## LISTA DE ABREVIATURAS E SIGLAS

AJAX	Asynchronous JavaScript and XML
BAMBE	Bayesian Analysis in Molecular Biology and Evolution
CACAU	Centro de Armazenamento de Dados e Computação Avançada da UESC
DNA	Ácido Desoxirribonucleico
DNAPARS	DNA Parsimony Program
EBI	The European Bioinformatics Institute
ENIAC	Electronic Numerical Integrator and Computer
FASTA	Fast Alignment
GENBANK	Genetic Sequence Database
GTR	General Reversible Time
HMM	Hidden Markov Model
JC	Jukes-Cantor
JSON	JavaScript Object Notation
JTT	Jones-Taylor-Thornton
MCMC	Markov Chain Monte Carlo
MCMCMC	Metropolis Coupled Markov Chain Monte Carlo
M-H	Metropolis-Hastings
MV	Máxima Verossimilhança
NBCGIB	Núcleo de Biologia Computacional e Gestão de Informações Biotecnológicas
NCBI	National Center for the Biotechnology Information
NJ	Neighbor Joining
NNI	Nearest Neighbor Interchange
OTUs	Operational Taxonomic Units
PHYLP	PHYLogeny Inference Package
RAF	Reconstrução de Árvores Filogenéticas
SPR	Subtree Pruning and Regrafting
SSH	Secure Shell
UESC	Universidade Estadual de Santa Cruz
UPGMA	Unweighted Pair Group Method with Arithmetic mean
WAG	Whelan And Goldman
WEIGHBOR	Weighted Neighbor Joining

## SUMÁRIO

<b>1</b>	<b>INTRODUÇÃO .....</b>	<b>1</b>
1.1	Motivação .....	3
1.2	Objetivos.....	3
<b>2</b>	<b>ÁRVORES FILOGENÉTICAS .....</b>	<b>5</b>
2.1	Representação dos Dados.....	10
2.2	Processos de Markov .....	11
2.3	Modelos evolutivos.....	14
2.3.1	Jukes-Cantor (JC69) .....	16
2.3.2	Kimura (K2P) .....	17
2.3.3	Felsenstein (F81) .....	18
2.3.4	General Time Reversible (GTR) .....	19
2.4	Taxas evolutivas de heterogeneidade.....	20
2.4.1	Distribuição Discreta .....	21
2.4.2	Distribuição Gama.....	22
2.4.3	Cadeia de Markov Oculta .....	24
<b>3</b>	<b>MÉTODOS DE RECONSTRUÇÃO DE ÁRVORES FILOGENÉTICAS .....</b>	<b>27</b>
3.1	Distância .....	27
3.1.1	Matriz de Distâncias .....	28
3.1.2	UPGMA.....	29
3.1.3	Neighbor-Joining (NJ).....	30
3.1.4	BIONJ.....	31
3.1.5	Weighbor .....	32
3.1.6	FastME .....	32
3.2	Máxima Parcimônia .....	32
3.2.1	Estratégia de busca da melhor árvore .....	36
3.3	Máxima Verossimilhança .....	40
3.3.1	Bootstrap.....	44
3.3.2	Jackknife.....	45
3.4	Inferência Bayesiana .....	46
3.4.1	Aplicação na Filogenética .....	47
3.4.2	Método de Monte Carlo via Cadeias de Markov.....	47
3.4.3	MCMCMC (MC <sup>3</sup> ) .....	49
<b>4</b>	<b>IMPLEMENTAÇÕES DOS MÉTODOS.....</b>	<b>51</b>
4.1	Digrafu.....	51
4.2	DNAPARS e PROTPARS.....	52

<b>4.3</b>	<b>PHYML</b> .....	54
<b>4.4</b>	<b>MrBayes</b> .....	56
4.4.1	Implementação do MCMC .....	57
4.4.2	Probabilidade de Transição .....	59
4.4.3	Deteção de Convergência das Cadeias .....	59
4.4.4	Amostragem, diagnóstico e sumarização de dados .....	61
<b>5</b>	<b>METODOLOGIA E DESENVOLVIMENTO</b> .....	62
<b>5.1</b>	<b>Metodologia</b> .....	62
5.1.1	Softwares de RAF.....	63
5.1.2	CACAU .....	63
5.1.3	PHP.....	63
5.1.4	Software de visualização de Árvores Filogenéticas .....	64
<b>5.2</b>	<b>Desenvolvimento</b> .....	64
5.2.1	MrBayes .....	66
5.2.2	PHYML .....	77
5.2.3	Digrafu.....	83
5.2.4	DNAPARS e PROTPARS.....	89
5.2.5	Execução.....	94
<b>6</b>	<b>TRABALHOS CORRELATOS</b> .....	97
<b>7</b>	<b>RESULTADOS E DISCUSSÕES</b> .....	112
7.1	Contribuições .....	112
7.2	Estudo de Caso.....	114
<b>8</b>	<b>CONCLUSÕES E TRABALHOS FUTUROS</b> .....	127
	<b>REFERÊNCIAS</b> .....	129

## 1 INTRODUÇÃO

Determinar as relações evolutivas de um conjunto de espécies, empregando sequências genéticas como dados de entrada, e apresentando como resultado tais relações em forma de árvores, mais precisamente árvores filogenéticas, é o que caracteriza a inferência filogenética. São essas árvores que auxiliam a desvendar os possíveis relacionamentos entre as espécies atuais e supor as histórias evolutivas das mesmas. Uma vez que as informações das espécies extintas são insuficientes, deve-se considerar cada árvore filogenética apenas como uma possibilidade hipotética.

Mais popularmente conhecidas como Filogenias, as árvores filogenéticas são compostas de folhas, que são as sequências genéticas que representam as espécies, galhos, que são as interligações entre os nós da árvore e representam o tempo de evolução entre eles, e nós internos, que correspondem aos seus ancestrais hipotéticos. Encontrar a árvore que melhor represente uma filogenia específica é um problema bastante complexo, pois o número de árvores a serem avaliadas cresce muito rapidamente conforme aumenta o número de espécies estudadas (FELSENSTEIN, 2004).

Neste contexto, segundo FELSENSTEIN (2004), os principais métodos utilizados para Reconstrução de Árvores Filogenéticas (RAF) são: Distância, Máxima Parcimônia, Máxima Verossimilhança e Inferência Bayesiana. Todos estes utilizam diferentes técnicas, cada um tratando de forma peculiar as hipóteses sobre o processo de evolução, como heurísticas próprias ou algoritmos estatísticos, para produzir a filogenia desejada. Alguns transformam a informação molecular em matrizes de distância para, enfim, aplicar um algoritmo específico que retorne a árvore. Outros empregam critérios de otimização, que permite avaliar cada árvore possível, com o objetivo de encontrar a melhor solução de acordo as escolhas iniciais. Vale ressaltar, que além da heurística de um desses métodos, a RAF exige a utilização de sequências genéticas devidamente alinhadas e um modelo de evolução de substituição de nucleotídeos (este trabalho não aborda os processos de evolução envolvendo aminoácidos). Os métodos citados serão analisados a seguir:

- Os métodos baseados em Distância foram pioneiros na RAF e ainda são muito utilizados devido a sua performance, embora já existam métodos mais exatos. Eles atuam sobre uma estrutura padrão de dados numéricos chamada matriz de distâncias (construída com base em modelos evolutivos), aplicando seus respectivos algoritmos para obter a melhor hipótese da filogenia. Dentre os métodos de distância mais conhecidos e utilizados, destacam-se o UPGMA (Unweighted Pair Grouping Method with Arithmetic Means) (SNEATH, 1973), o



NJ (Neighbor Joining) (SAITOU; NEI, 1987), o BIONJ (GASCUEL, 1997), o Weighbor (Weighted Neighbor Joining) (BRUNO; SOCCI; HALPERN, 2000) e o FastME (Fast Minimum Evolution) (DESPER; GASCUEL, 2002). Além desses citados, tem-se o DiGrafu (TORRES et al., 2011), que reúne todos os anteriores;

- O método de Máxima Parcimônia busca uma árvore que traduz o menor número possível de mudanças ocorridas nos caracteres das sequências genéticas submetidas. Sendo assim, ao se atribuir pesos na mudança de um caractere x para y, a hipótese da árvore mais parcimoniosa será aquela cujo somatório das mudanças informa o menor valor. Este método é considerado relativamente rápido e proporciona bons resultados desde que os galhos sejam curtos. Como exemplos de softwares que utilizam este método, pode-se citar: DNAPARS (FELSENSTEIN, 1993), PROTPARS (FELSENSTEIN, 1993), PARS (FELSENSTEIN, 1993), MIX (FELSENSTEIN, 1993) e TNT (GOLOBOFF; FARRIS; NIXON, 2005);
- O método de Máxima Verossimilhança atua de modo a buscar a árvore que maximize a probabilidade dos dados moleculares adequarem-se a um determinado modelo evolutivo. As implementações atuais deste método são baseadas em heurísticas que produzem árvores perto das ótimas, com a necessidade de um algoritmo de busca. Dos softwares que utilizam este método, pode-se citar: o PHYML (GUINDON; GASCUEL, 2003), FastDNAm1 (OLSEN et al., 1994), MorePhyML (CRISCUOLO, 2011), DNAML (FELSENSTEIN, 1993) e PROML (FELSENSTEIN, 1993);
- O método de Inferência Bayesiana oferece suporte estatístico baseado no teorema de Bayes para a certificação da árvore inferida. Analisando pelo âmbito computacional, este método utiliza um recurso numérico que é a amostragem de uma densidade de probabilidade pelo método de Monte Carlo via Cadeias de Markov (MCMC). O MrBayes (HUELSENBECK; RONQUIST, 2001) e o BAMBE (Bayesian Analysis in Molecular Biology and Evolution) (LARGET; SIMON, 1999), são exemplos de softwares que utilizam o método em questão.

## 1.1 Motivação

A utilização dos métodos mencionados para RAF estão disponíveis através de diversos softwares, dentre eles estão os citados, os quais apresentam características próprias, e muitas vezes não possuem interface gráfica: sua utilização é realizada através de comandos de texto, exigindo conhecimentos avançados da plataforma em questão. Além disso, dependendo da quantidade de espécies que se deseje analisar, a filogenia exige muito poder de processamento, o que nem sempre está disponível para os usuários. Além do mais, para um mesmo conjunto de espécies, pode-se inferir filogenias diferentes para cada método abordado. Isso implicaria na utilização de vários softwares separados, cada um de acordo às especificações do seu método.

Dessa maneira, fica evidente a necessidade de melhorar a usabilidade em relação a RAF. Sendo assim, o software proposto neste trabalho de pesquisa visa suprir todas as seguintes deficiências citadas:

- O usuário não precisará instalar, configurar e utilizar softwares que RAF em linha de comandos de textos, visto que, os parâmetros de cada software estarão devidamente organizados, de forma intuitiva, em uma interface gráfica na Web;
- O usuário não precisará utilizar  $n$  ambientes de softwares diferentes para RAF de acordo ao método de sua escolha. O ambiente em questão abrigará um software para cada método de RAF, além de um módulo para visualizar o resultado final (a árvore filogenética) em formato gráfico;
- Os usuários que não possuem um computador de alto desempenho para submeter as suas sequências genéticas, serão beneficiados com o cluster de computadores de alta performance. Todas as execuções dos softwares citados serão realizadas neste supercomputador, diminuindo assim, o tempo de resposta do resultado filogenético.

## 1.2 Objetivos

O principal objetivo deste trabalho é fornecer as informações necessárias para compreensão das particularidades da filogenia e da modelagem matemática utilizada nos programas mencionados anteriormente, além de disponibilizar uma estrutura computacional de alto desempenho na Web para RAF, com o intuito de facilitar o trabalho do usuário final. Mais precisamente, pretende-se montar um Web Service, denominado IgrafuWeb, de endereço eletrônico <http://nbcgib.uesc.br/igrafuweb>, para permitir a RAF utilizando o poder de processamento do CACAU (Centro de Armazenamento de

Dados e Computação Avançada da UESC) e os softwares MrBayes (HUELSENBECK; RONQUIST, 2001), PHYML (GUINDON; GASCUEL, 2003), Digrafu (TORRES et al., 2011) e DNAPARS/PROTPARS (FELSENSTEIN, 1993), além de possibilitar a visualização gráfica da(s) árvore(s) filogenética(s) resultado da filogenia no próprio sistema, através de um applet, o Archaeopteryx<sup>1</sup> (HAN; ZMASEK, 2009).

Nesse sentido, para atingir esses objetivos, o desenvolvimento deste projeto é feito da forma como segue. O capítulo 2 apresenta uma visão geral da área de árvores filogenéticas, apresentando todo o conceito inicial, até chegar na modelagem matemática dos modelos de evolução, bem como dos processos de Markov e das taxas evolutivas de heterogeneidade.

Logo a seguir apresenta-se o capítulo 3, que aborda as particularidades de cada método de RAF citado. Dando seguimento, chega-se ao capítulo 4, que faz um estudo minucioso das implementações dos métodos de RAF através dos softwares MrBayes, PHYML, Digrafu, DNAPARS e PROTPARS.

O capítulo 5 explica em detalhes os modos e formas utilizadas para desenvolver o IgrafuWeb, além de apresentar todas as funcionalidades do software implementado para RAF.

Embora existam alguns Web Services que facilitam a RAF, o IgrafuWeb é válido, pois pretende ser uma aplicação computacional de alto desempenho, sendo considerado um diferencial sobre os atualmente disponíveis. Como consequência disso, surgiu a necessidade de abordar os trabalhos correlatos existentes, fazendo um paralelo com o produto resultante desta pesquisa. O capítulo que aborda esta questão é o 6, nele foram pesquisados os trabalhos correlatos ao proposto, com o objetivo de conhecer as características funcionais mais adequadas para a modelagem e a construção da estrutura física do sistema.

Os resultados encontrados neste projeto foram confrontados com filogenias conhecidas na literatura, a fim de comprovar a veracidade e a qualidade das informações finais resultantes. Toda esta análise encontra-se no capítulo 7.

Por fim, o capítulo 8 apresenta as conclusões deste trabalho, bem como as sugestões de propostas para a elaboração de trabalhos futuros.

---

<sup>1</sup> Applet disponibilizado pelo software Archaeopteryx, acessado através do endereço eletrônico <https://sites.google.com/site/cmzmasek/home/software/archaeopteryx>.

## 2 ÁRVORES FILOGENÉTICAS

As árvores filogenéticas são representações gráficas, estruturadas na forma de uma árvore, que explicam a possível história evolutiva das espécies ou grupos de espécies, construídas a partir de sequências de atributos (tais como sequências de DNA) obtidos de diversos organismos. Como a árvore filogenética é uma hipótese sobre relações evolutivas, precisam-se utilizar caracteres que sejam indicadores confiáveis de ancestralidade comum para se construir a árvore (HYPÓLITO, 2005), (TICONA, 2008), (AMORIM, 2002). Desse modo, devem-se utilizar características herdadas de um ancestral comum, podendo ser qualquer uma recebida como herança.

Neste sentido, as árvores filogenéticas detêm informações úteis para uma grande variedade de questões biológicas e sociais. Podem auxiliar no controle e combate de parasitas responsáveis por doenças, no estudo epidemiológico, para criação de vacinas mais eficientes, na produção de novas drogas agrícolas e médicas, no estudo do desenvolvimento da biodiversidade, etc (AMORIM, 2002). Outra utilidade das árvores filogenéticas é fornecer subsídios importantes para decisões relativas a transplantes de órgãos ou de tecidos de outras espécies.

Compreender árvores filogenéticas (*Figura 1*) (MENDES, 2015) exige o entendimento de algumas características (HYPÓLITO, 2005), (TICONA, 2008):

- Nós: são pontos na árvore, que podem ser internos, representando o ancestral comum, ou terminais, que são os organismos estudados;
- Ramos: são linhas que ligam os nós e representam o tempo de evolução entre eles;
- OTUs - Operational Taxonomic Units: são organismos incluídos na análise, do qual se deseja inferir a história filogenética, também conhecidos como táxons;
- Topologia: uma representação gráfica unindo as OTUs através de ramos e nós.

Percebe-se através da *Figura 1*, que de cada nó interno ramificam-se exatamente dois ramos (galhos), levando a concluir que esta relação evolutiva entre as espécies é representada por árvores binárias. Além disso, é relevante salientar que todos os táxons são descendentes de um mesmo ancestral, o nó raiz, implicando uma direção de tempo de evolução. Logo, conclui-se que esta é uma árvore enraizada, com raiz.

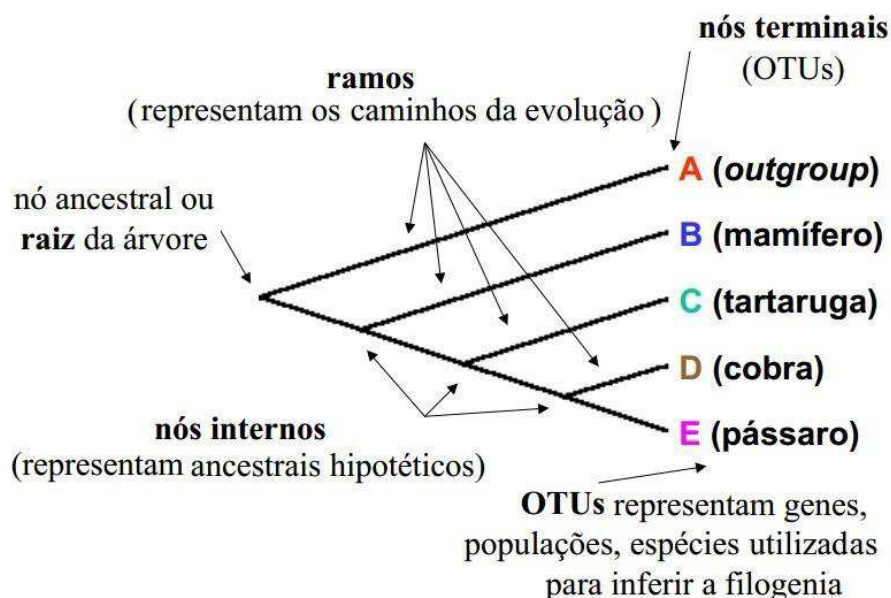


Figura 1 - Exemplo descritivo de uma árvore filogenética.

Por outro lado, existem também as árvores sem raiz, como é o caso da árvore representada na *Figura 2*, na qual não se tem a relação de ancestralidade. Ainda assim, é possível escolher um ponto qualquer da árvore sem raiz onde se insira um nó raiz, desencadeando em uma árvore com raiz. Dependendo do local de inserção deste nó raiz, pode-se gerar várias árvores enraizadas. Aplicar esta mesma inserção de um nó raiz em uma árvore enraizada não é válido, visto que pode-se comprometer o resultado original (a árvore foi constituída levando-se em conta que haveria um nó de origem).

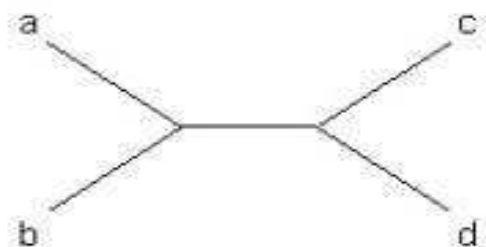


Figura 2 - Exemplo de uma árvore filogenética sem raiz.

As árvores filogenéticas representam hipóteses da história evolutiva das espécies. A inferência da árvore que se mostra mais adequada aos dados obtidos é uma tarefa complexa. Além disso, para um determinado conjunto de táxons, existem mais de uma possibilidade de árvores a serem analisadas, o que torna este processo ainda mais complicado.

Nesse sentido, dar-se o nome de topologia à forma como os nós internos se conectam uns com os outros e com as folhas. Quanto maior a quantidade de táxons, maior é o número de topologias (ver *Tabela 1* adaptada de HOLMES, 1998, onde  $s$  representa o número de táxons).

Tabela 1 - Número de topologias possíveis sem e com raiz.

$s$	Sem raiz	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

Este número varia de acordo ao tipo de árvore e ao número de táxons. Sendo assim, pode-se calcular as quantidades de topologias com raiz através da *Equação (1)*, e sem raiz através da (2) (HYPÓLITO, 2005).

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

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

Com efeito de exemplificação, foi feita uma análise de 4 espécies ( $a$ ,  $b$ ,  $c$  e  $d$ ), para ilustrar as possíveis topologias com e sem raiz para essas espécies (Figura 3 e Figura 4) (HYPÓLITO, 2005).

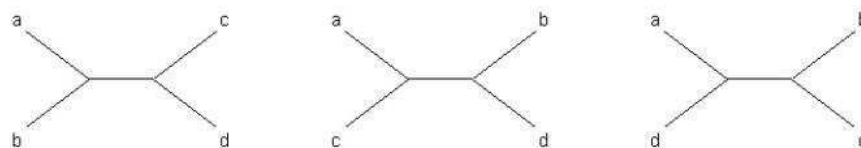


Figura 3 - Possíveis topologias sem raiz para 4 espécies.

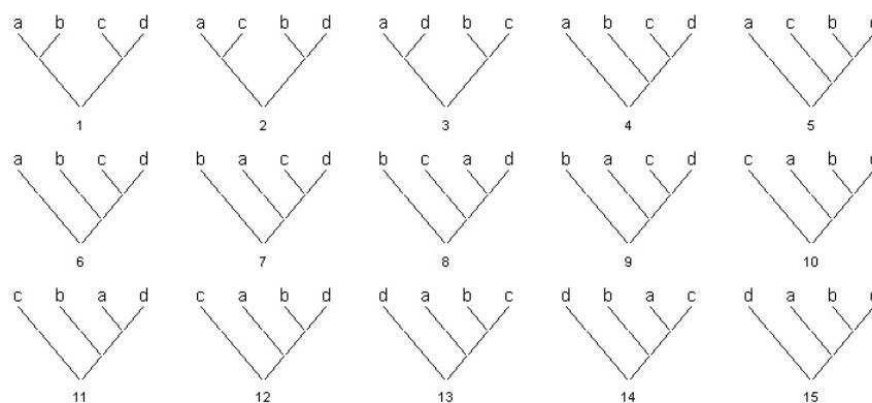


Figura 4 - Possíveis topologias com raiz para 4 espécies.

As representações gráficas de árvores filogenéticas, conforme vistas nas Figura 3 e Figura 4, são realizadas através de softwares que realizam estes desenhos de acordo aos dados de entrada (arquivo de texto). Os dados de saída destes softwares são organizados de uma forma peculiar, muitas vezes determinado por um formato específico conhecido como Newick. Este foi idealizado em 1857 pelo matemático inglês Arthur Cayley (1821 - 1895), sendo adotado como padrão em 1986 (VIANA, 2007).

Neste sentido, idealizou-se esta notação de forma a facilitar a 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 seguindo o padrão Newick da seguinte forma: (B,(A,C,E),D);. Esta é representada graficamente através da Figura 5 (SILVA, 2007).

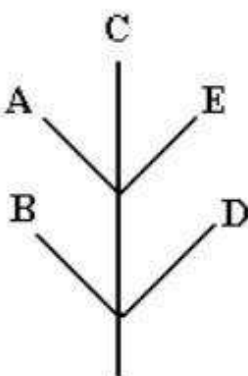


Figura 5 - Representação gráfica da árvore filogenética (B,(A,C,E),D); padrão Newick.

Sendo assim, de acordo à análise da Figura 5 e dos comentários acima, pode-se afirmar a respeito do padrão Newick (VIANA, 2007), (SILVA, 2007):

- 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), as quais 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. A Figura 6 exhibe um exemplo deste caso (VIANA, 2007).

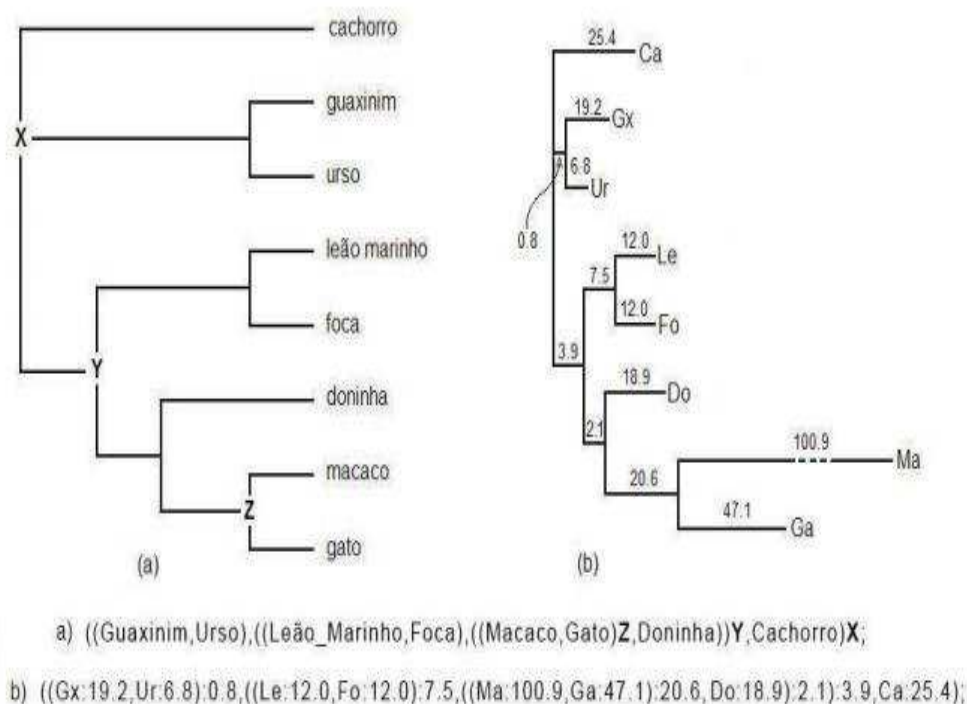


Figura 6 - Representação do padrão Newick com comprimento dos ramos.



## 2.1 Representação dos Dados

Os estudos de biologia molecular produzem uma grande quantidade de dados que são apresentados como sequências de diversos tipos. Dentre estes, destacam-se as de DNA e as Proteicas.

As sequências de DNA são compostas de uma sucessão de nucleotídeos, 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 proteicas consistem de uma sucessão de aminoácidos, que podem assumir 20 estados diferentes.

Para se reconstruir árvores filogenéticas, se faz 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:

- 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. A Figura 7 (GONÇALVES, 2008) ilustra a matriz citada.

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

Figura 7 - Matriz de alinhamento de sequências de DNA, na qual as linhas representam as espécies e as colunas os sítios (base nucleotídica para determinada posição homóloga entre as espécies).

Existem diversos softwares que realizam o processo de alinhamento de sequências genéticas, dentre os quais, pode-se citar Clustalw2<sup>2</sup> (LARKIN et al., 2007). Este será utilizado na apresentação dos resultados deste trabalho.

A seguir serão abordados os processos de Markov, bem como os modelos matemáticos relacionados com a evolução de sequências e as taxas de heterogeneidade inseridas nestes modelos.

## 2.2 Processos de Markov

Um processo é considerado de Markov se as probabilidades futuras não dependem do passado (dependendo apenas do estado presente), sendo chamado de cadeias de Markov caso seu espaço de estados seja discreto ou enumerável. Sendo assim, supondo um processo de Markov com espaço de estados  $S$  discreto, com tempo definido nos reais positivos, e  $X$  a variável que assume valores em  $S$ , a característica principal deste processo implica que, dado um estado  $X(t - h) = i$  em algum tempo  $t$ , a probabilidade de que  $X(t) = j$ , em um tempo futuro  $t$ , não depende dos valores de  $X$  antes do tempo  $t$ . Esta afirmação é representada pela equação a seguir (CYBIS, 2009), (HYPÓLIO, 2005).

$$P(X(t) = j | X(t - h) = i) \quad (3)$$

Sendo assim, considerando as análises das cadeias de Markov, a probabilidade condicional  $P(X(t) = j | X(t - h) = i)$  é independente de  $t$ , e, portanto pode ser definida como uma função de transição conforme equação a seguir (HYPÓLIO, 2005):

$$p_{ij}(s, t) = P[X(t) = j | X(s) = i] \quad s \leq t \quad (4)$$

Desse modo, determinar a função de transição, representada pela *Equação (4)*, é o ponto de partida para resolução das cadeias de Markov.

Nesse sentido, de acordo aos conceitos de probabilidade total, pode-se afirmar que, dados  $A, B_1$  e  $B_2$ , tais que  $B_1$  e  $B_2$  são mutuamente exclusivos, tem-se:

$$P(A) = \sum_{\forall i} P(A \wedge B_i) = \sum_{\forall i} P(A | B_i) P(B_i) \quad (5)$$

Condicionando  $[X(t) = j | X(s) = i]$  a  $[X(u) = r]$  para algum  $s \leq u \leq t$  e considerando  $A = [X(t) = j | X(s) = i]$  e  $B = [X(u) = r | X(s) = i]$ , pode-se encontrar a *Equação (6)*, conhecida como a equação de Chapman-Kolmogorov (GRIMMENT; STIRZAKER, 1992), (EWENS; GRANT, 2001):

---

<sup>2</sup> Consular o link <http://www.ebi.ac.uk/Tools/msa/clustalw2/help> para obter mais detalhes sobre o Clustalw2.

$$\begin{aligned}
p_{ij}(s, t) &= \sum_{\forall r} P[X(t) = j, X(s) = i | X(u) = r, X(s) = i] P[X(u) = r | X(s) = i] \\
p_{ij}(s, t) &= \sum_{\forall r} P[X(t) = j | X(u) = r] P[X(u) = r | X(s) = i] \\
p_{ij}(s, t) &= \sum_{\forall r} p_{rj}(u, t) p_{ir}(s, u) = \sum_{\forall r} p_{ir}(s, u) p_{rj}(u, t)
\end{aligned} \tag{6}$$

A *Equação (6)* pode ser reescrita na forma matricial, conforme descrição a seguir:

$$P(s, t) = P(s, u)P(u, t) \quad s \leq u \leq t \tag{7}$$

Para os instantes  $s \leq t \leq t + \Delta t$ , a *Equação (7)* fica da seguinte forma:

$$P(s, t + \Delta t) = P(s, t)P(t, t + \Delta t) \tag{8}$$

Subtraindo os dois termos da igualdade por  $P(s, t)$ , e posteriormente colocando-o em evidência no lado direito, encontra-se a *Equação (9)*.

$$\begin{aligned}
P(s, t + \Delta t) - P(s, t) &= P(s, t)P(t, t + \Delta t) - P(s, t) \\
P(s, t + \Delta t) - P(s, t) &= P(s, t)[P(t, t + \Delta t) - I]
\end{aligned} \tag{9}$$

Dividindo a *Equação (9)* por  $\Delta t$  e tomando o limite  $\Delta t \rightarrow 0$ , tem-se:

$$\begin{aligned}
\lim_{\Delta t \rightarrow 0} \frac{P(s, t + \Delta t) - P(s, t)}{\Delta t} &= P(s, t) \lim_{\Delta t \rightarrow 0} \frac{P(t, t + \Delta t) - I}{\Delta t} \\
\frac{\partial P(s, t)}{\partial t} &= P(s, t)Q(t) \quad s \leq t
\end{aligned} \tag{10}$$

Sabendo que a cadeia é homogênea, a função de transição independe dos valores absolutos de  $s$  e  $t$ , mas somente da diferença  $\tau = (t - s)$ . Neste caso, a *Equação (10)* fica na forma:

$$\frac{\partial P(\tau)}{\partial t} = P(\tau)Q \tag{11}$$

A solução desta equação, com condições iniciais:

$$p_{ij}(0) \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}, \quad \text{ou seja,} \quad P(0) = I \tag{12}$$

é dada por:

$$P(\tau) = e^{Q\tau} \tag{13}$$

Nesse sentido, para este trabalho foram utilizadas cadeias de Markov de tempo contínuo, cuja caracterização se baseia através de uma matriz  $Q$  geradora infinitesimal, cujo a 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  $t$  (CYBIS, 2009).

Ainda nesse contexto, a matriz infinitesimal  $Q$  pode determinar as probabilidades de transição em um determinado intervalo de tempo  $t$  de um processo de Markov, através da resolução da *Equação (13)*. Assim, a decomposição da matriz  $Q$  em seus autovalores e autovetores leva à matriz  $P$ . Logo, a solução da *Equação (13)* é:

$$P(t) = e^{Qt} = ADA^{-1} \quad (14)$$

onde  $D$  representa a matriz diagonal cujos elementos são os 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  $t$ . A soma dos elementos de uma determinada linha desta matriz  $P$  é igual a 1 (CYBIS, 2009).

Nesse sentido, conforme descrito nas afirmações acima, a soma dos elementos de cada linha da matriz  $Q$  é zero, enquanto a da matriz  $P$  é 1. Isto pode ser confirmado matematicamente, de acordo a explicação a seguir. Como a matriz  $P$  trata de probabilidades, pode-se afirmar que o valor da soma dos elementos de cada linha é 1, logo:

$$\sum_j p_{ij}(s, t) = 1 \quad (15)$$

Seguindo a mesma linha de raciocínio, agora analisando a matriz  $Q$ , sabe-se que  $q_{ii}(t)$  e  $q_{ij}(t)$  são encontrados através das equações a seguir:

$$q_{ii}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{ii}(t - \Delta t, t) - 1}{\Delta t} \quad (16)$$

$$q_{ij}(t) = \lim_{\Delta t \rightarrow 0} \frac{p_{ij}(t - \Delta t, t)}{\Delta t} \quad i \neq j \quad (17)$$

Logo, o somatório de cada linha da matriz  $Q$  é descrito matematicamente da seguinte forma:

$$\begin{aligned} q_{ii}(t) + \sum_{j \neq i} q_{ij}(t) &= \lim_{\Delta t \rightarrow 0} \left\{ \frac{p_{ii}(t - \Delta t, t) - 1}{\Delta t} + \sum_{j \neq i} \frac{p_{ij}(t - \Delta t, t)}{\Delta t} \right\} \\ \sum_j q_{ij}(t) &= \lim_{\Delta t \rightarrow 0} \left\{ \frac{\sum_j p_{ij}(t - \Delta t, t) - 1}{\Delta t} \right\} \\ \sum_j q_{ij}(t) &= 0 \end{aligned} \quad (18)$$

A distribuição estacionária é outro ponto relevante que precisa ser mencionado como pré-requisito para apresentar mais adiante os modelos que especificam o processo de evolução de cada sequência. Assim, para um processo de Markov com espaço de estados finito 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  $i$ ,

$$\lim_{t \rightarrow \infty} P(X_t = j | X_0 = i) = p_{0j} \quad (19)$$

onde  $p_{0j}$  representa a  $j$ -ésima componente do vetor  $p_0$  (CYBIS, 2009).

A afirmação acima é relevante, pois os modelos descritos a seguir assumem que a distribuição inicial dos processos é a própria distribuição estacionária, visto que as sequências em estudo estão evoluindo sob esse processo há bastante tempo. Desse modo, analisando a realidade atual, as frequências dos estados já estariam muito próximas da distribuição estacionária (CYBIS, 2009).

Todos os conceitos apresentados acima são cruciais para a compreensão da seção a seguir, a qual aborda as características principais dos modelos evolutivos de substituição de nucleotídeos.

### 2.3 Modelos evolutivos

O modelo evolutivo é uma modelagem matemática, baseada em Cadeias de Markov, atribuída à transição dos nucleotídeos em uma provável mutação. O objetivo dos modelos de evolução é modelar o processo de substituição, ou seja, determinar dentro de uma mutação quais as probabilidades dos nucleotídeos, substituírem uns aos outros. Vale ressaltar que, estes modelos assumem que o único tipo de mutação que ocorre é a substituição de uma base por outra na mesma posição da sequência, de maneira que adições e inserções são ignoradas.

No estudo de sequências genéticas de nucleotídeos, observa-se apenas o estado atual do processo, não tendo como saber quantas mutações ocorreram na história evolutiva de um sítio. Se no passado a base de um sítio específico foi A, e atualmente observa-se um C, pode-se inferir que houve uma transição neste sítio, porém não se pode afirmar que a mutação foi de A para C, pois poderia ter ocorrido a sequência de mutação de A para G, e depois para C. Seguindo o mesmo raciocínio, se uma determinada base de um outro sítio foi A no passado e continua sendo A no presente, não se pode afirmar que não houveram mutações neste sítio, pois a sequência de A para T, depois para C, até chegar em A novamente, apresenta o mesmo resultado. A nomenclatura definida para este tipo de mutação é conhecida como silenciosa (CYBIS, 2009).

Sendo assim, os modelos de substituição discutidos neste trabalho corrigem estas mutações múltiplas e silenciosas. Eles tratam estes processos através de cadeias de Markov, onde se obtém a probabilidade de encontrar a base  $i$  em um tempo  $t$  para um determinado sítio, dado que no presente encontra-se a base  $j$ . Esta tal probabilidade é independente das mutações que aconteceram no passado (CYBIS, 2009).

Existem diversos modelos que relacionam a evolução da cadeia de DNA, permitindo a obtenção das probabilidades  $P(j|i, t)$  de que uma base  $i$  mude para outra base  $j$  em um determinado intervalo de tempo  $t$ . Estes modelos atribuem uma cadeia de Markov ao processo de evolução de cada sítio da sequência de DNA, visto que assumem que as probabilidades de mutação dependem apenas do estado atual da cadeia, sendo independente do passado do processo. 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  $E = \{A, G, C, T\}$ . 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 (CYBIS, 2009).

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. Desta forma, a matriz de taxas infinitesimais desses modelos é apresentada basicamente da forma a seguir, onde  $q_{ij}$  é a taxa de mutação da base  $i$  para base  $j$ , e os elementos da diagonal principal são definidos de tal forma que as linhas somam zero, ou seja,  $q_i = \sum_{j \neq i} q_{ij}$ , para  $i, j \in E$  (CYBIS, 2009).

$$\begin{pmatrix} -q_A & q_{A,G} & q_{A,C} & q_{A,T} \\ q_{G,A} & -q_G & q_{G,C} & q_{G,T} \\ q_{C,A} & q_{C,G} & -q_C & q_{C,T} \\ q_{T,A} & q_{T,G} & q_{T,C} & -q_T \end{pmatrix} \quad (20)$$

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 (21)*, onde  $\pi_i$  é a proporção da base  $i$  na molécula de DNA. Fica implícito na estrutura destes modelos, que o processo encontra-se próximo a seu estado de equilíbrio, de maneira que a distribuição estacionária  $p_0$  é também a distribuição inicial (CYBIS, 2009).

$$p_0 = (\pi_A, \pi_G, \pi_C, \pi_T) \quad (21)$$

Outro ponto relevante que precisa ser mencionado para esclarecer as principais características dos modelos evolutivos é a matriz de probabilidades de transição, chamada de  $P$ . Como mencionado no item anterior, cada modelo possui uma matriz  $Q$  de acordo as suas características,

consequentemente conterá também uma matriz  $P$ , visto que esta é encontrada a partir de  $Q$ , mediante a solução de uma equação diferencial, *Equação (14)*, conforme descrito anteriormente.

Cabe ainda mencionar que, como estes modelos apresentam distribuição estacionária, as probabilidades de substituição  $P(j|i, t)$  em longos intervalos de tempo se aproximam das frequências das bases na distribuição estacionária, conforme descrito na equação a seguir (CYBIS, 2009).

$$\lim_{t \rightarrow \infty} P(j|i, t) = \pi_j, \forall i, j \in E \quad (22)$$

Alguns exemplos de modelos de substituição de nucleotídeos serão abordados a seguir, de modo a facilitar as explicações descritas acima.

### 2.3.1 Jukes-Cantor (JC69)

Jukes e Cantor (1969) propuseram um modelo bastante simples, conhecido também pela sigla JC69, onde os nucleotídeos em uma sequência de DNA ocorrem em frequências iguais e as probabilidades de substituição de um nucleotídeo  $i$  para um nucleotídeo  $j$  são todas idênticas e ocorrem de maneira Markoviana (CYBIS, 2009), (HYPÓLIO, 2005).

Neste sentido, a matriz que fornece as taxas infinitesimais para as mutações da base  $i$  para a base  $j$ , sendo  $i$  e  $j$  pertencentes ao espaço de estados  $E = \{A, G, C, T\}$  do processo, é a  $Q$ , definida através da *Expressão (23)*, assim como, a distribuição inicial de probabilidades das bases da molécula de DNA, é  $p_0$ , exibido na *Expressão (24)*. Vale ressaltar que, conforme observa-se na *Expressão (23)*,  $3\alpha$  é a taxa de mutação do processo (CYBIS, 2009).

$$Q = \begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix} \quad (23)$$

$$p_0 = \left( \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4} \right) \quad (24)$$

Como já mencionado no item 2.3, em posse da matriz  $Q$ , pode-se encontrar a matriz  $P$  de probabilidades de substituição para este modelo. Para isto basta solucionar a *Equação (13)*, 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 (*Expressão (25)*) para o modelo em questão, na qual  $\alpha_t = \frac{1}{4}(1 - e^{-4\alpha t})$  (CYBIS, 2009).

$$P = \begin{pmatrix} 1 - 3\alpha_t & \alpha_t & \alpha_t & \alpha_t \\ \alpha_t & 1 - 3\alpha_t & \alpha_t & \alpha_t \\ \alpha_t & \alpha_t & 1 - 3\alpha_t & \alpha_t \\ \alpha_t & \alpha_t & \alpha_t & 1 - 3\alpha_t \end{pmatrix} \quad (25)$$

O modelo de JC foi um dos primeiros propostos na literatura para explicar a substituição de bases nucleotídicas, apresentando, contudo algumas limitações referentes aos fenômenos encontrados nas sequências genéticas de DNA, como as variações de taxas para tipos de substituições distintas e os diferentes números de nucleotídeos (CYBIS, 2009).

### 2.3.2 Kimura (K2P)

Este modelo foi idealizado por Kimura (1980), analisando a ocorrência de valores distintos entre as taxas de transição e transversão, representadas por  $\alpha$  e  $\beta$  respectivamente. A transição é a substituição de uma Purina (A ou G) por uma Purina ou de uma Pirimidina (T ou C) por outra Pirimidina e a transversão é a substituição de uma Purina por uma Pirimidina ou de uma Pirimidina por uma Purina (HYPÓLIO, 2005).

Deve-se observar que, para qualquer nucleotídeo, pode existir uma troca a uma taxa  $\alpha$  ou  $\beta$ , o que acarreta respectivamente uma transição ou transversão. A razão de transição/transversão do modelo é denotada por  $r$  e é igual a  $\alpha/(2\beta)$ . Vale ressaltar, que o modelo JC69 é um caso particular do modelo K2P, onde  $\alpha = \beta$  e  $r = 1/2$  (HYPÓLIO, 2005).

Nesse sentido, a matriz  $Q = \{q_{ij}\}$  de taxas instantâneas de substituição de nucleotídeos que representa o modelo é dada por:

$$Q = \begin{bmatrix} -\alpha - 2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha - 2\beta & \beta & \beta \\ \beta & \beta & -\alpha - 2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha - 2\beta \end{bmatrix} \quad (26)$$

Seguindo a mesma linha de raciocínio, pode-se calcular a matriz de probabilidades de transição,  $P = \{p_{ij}\}$ , através dos autovetores e autovalores de  $Q$ , como descrito na *Equação (13)*. Logo, encontra-se um sistema de equações, o qual tem como solução as probabilidades de transição (HYPÓLIO, 2005).

Dessa forma, para um comprimento de ramo  $t$  e um dado nucleotídeo  $i$ , pode-se dizer que:

$$p_{ii}(t) = 0.25 + 0.25e^{-4\beta t} + 0.5e^{-2(\alpha+\beta)t} \quad (27)$$



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

$$p_{ij}(t) = 0.25 + 0.25e^{-4\beta t} - 0.5e^{-2(\alpha+\beta)t} \quad (28)$$

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

$$p_{ij}(t) = 0.25 - 0.25e^{-4\beta t} \quad (29)$$

Logo, calculando o  $\lim_{t \rightarrow \infty} p_{ij}(t)$  encontra-se a probabilidade estacionária  $\pi_j = 1/4 \forall j \in \{A, C, G, T\}$ . Nota-se a semelhança do cálculo da probabilidade estacionária deste modelo com o do JC69.

### 2.3.3 Felsenstein (F81)

Este modelo foi idealizado por *Felsenstein* em 1981, sendo uma generalização do modelo JC69 (FELSENSTEIN, 1981). Diferentemente dos modelos já citados acima, ele não considera a igualdade de frequências de bases. As quatro frequências são calculadas de acordo com a porcentagem com que cada frequência aparece na sequência de nucleotídeos (HYPÓLIO, 2005).

A principal suposição deste modelo é que a probabilidade de substituição de um nucleotídeo por um dos outros três é proporcional à probabilidade estacionária do nucleotídeo substituído. Desse modo, a matriz de taxas instantâneas de substituição de nucleotídeos, onde  $k > 0$  é um parâmetro do modelo que representa a taxa total de transversão, é dada por (HYPÓLIO, 2005):

$$Q = \begin{bmatrix} -k(\pi_C + \pi_G + \pi_T) & k\pi_C & k\pi_G & k\pi_T \\ k\pi_A & -k(\pi_A + \pi_G + \pi_T) & k\pi_G & k\pi_T \\ k\pi_A & k\pi_C & -k(\pi_A + \pi_C + \pi_T) & k\pi_T \\ k\pi_A & k\pi_C & k\pi_G & -k(\pi_A + \pi_C + \pi_G) \end{bmatrix} \quad (30)$$

Com essa matriz especifica-se um sistema de equações, o qual tem como solução as probabilidades de transição a seguir.

$$p_{ij}(t) = \begin{cases} \pi_j(1 - e^{-kt}) & \text{se } i \neq j \\ e^{-kt} + \pi_j(1 - e^{-kt}) & \text{se } i = j \end{cases} \quad (31)$$

### 2.3.4 General Time Reversible (GTR)

Este modelo foi descrito pela primeira vez em 1986 por Simon Tavaré, sendo considerado o mais geral e independente possível. Além disso, é reversível no tempo, o que significa dizer que a probabilidade de se começar com  $i$  em uma das pontas de um ramo da árvore filogenética e de se terminar com  $j$  na outra é a mesma probabilidade de que ocorra ao contrário. A reversibilidade no tempo é uma propriedade matemática conveniente, apesar de não ser fundamentada em razões biológicas. Em termos práticos, muitos modelos que possuem esta propriedade se ajustam bem a dados reais (FELSENSTEIN, 2004). Isto pode ser expresso através da *Equação (32)*, na qual  $P_{x,y}(t)$  é a probabilidade de mudança do estado  $x$  para  $y$  no tempo  $t$ :

$$\pi_x P_{x,y}(t) = \pi_y P_{y,x}(t) \quad (32)$$

Este modelo atribui à evolução da sequência da molécula de DNA uma cadeia de Markov a tempo contínuo, com distribuição inicial dada pela *Expressão (21)* e matriz de taxas infinitesimal dada pela *Expressão (33)*, onde  $f_k$  é apresentado de tal forma que a soma dos elementos de uma determinada linha  $k$  seja 0, sendo  $k \in \{1,2,3,4\}$  (CYBIS, 2009).

$$Q_{GTR} = \begin{pmatrix} f_1 & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & f_2 & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & f_3 & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & f_4 \end{pmatrix} \quad (33)$$

Cabe mencionar que: a cadeia é estacionária,  $p_0$  é sua distribuição de equilíbrio e este modelo satisfaz a propriedade definida na *Equação (32)*. Além disso, o modelo GTR tem um grande número de parâmetros, o que inviabiliza a apresentação das expressões analíticas para as probabilidades de transição, no entanto, um exemplo numérico que ilustra a obtenção da matriz  $P_{GTR}(t)$  é apresentado a seguir (CYBIS, 2009).

Neste sentido, suponha o caso que se deseje obter a matriz  $P_{GTR}(t)$ , fixando os seguintes parâmetros:  $p_0 = (0.3, 0.2, 0.2, 0.3)$ ,  $\alpha = 1$ ,  $\beta = 0.2$ ,  $\gamma = 0.3$ ,  $\delta = 0.1$ ,  $\epsilon = 0.01$ , e  $\eta = 2.5$ . Logo, de acordo a definição da *Expressão (33)*, a matriz de taxas infinitesimais para o modelo em questão é dada por (CYBIS, 2009):

$$Q_{GTR} = \begin{pmatrix} -0.3300 & 0.2000 & 0.0400 & 0.0900 \\ 0.3000 & -0.3230 & 0.0200 & 0.0030 \\ 0.0600 & 0.0200 & -0.8300 & 0.7500 \\ 0.0900 & 0.0020 & 0.5000 & -0.5920 \end{pmatrix} \quad (34)$$

De acordo aos conceitos apresentados no item anterior, sabe-se que para encontrar a matriz de probabilidade de transição  $P_{GTR}$ , basta saber a matriz exponencial de  $Q_{GTR}$ , ou seja,  $P_{GTR}(t) = e^{Q_{GTR}t}$ , a qual tem como solução a decomposição da matriz  $Q$  em seus autovalores e autovetores, ficando:

$$P(t) = e^{Qt} = ADA^{-1} \quad (35)$$

Neste sentido,  $D$  representa a matriz diagonal cujos elementos são os autovalores de  $Q$  e  $A$  é a matriz cujas colunas são os autovetores diretos de  $Q$ .

$$D = \begin{pmatrix} -0.5768 & 0 & 0 & 0 \\ 0 & -1.3010 \times 10^{-18} & 0 & 0 \\ 0 & 0 & -0.1608 & 0 \\ 0 & 0 & 0 & -1.3356 \end{pmatrix} \quad (36)$$

$$A = \begin{pmatrix} 0.6474 & -0.5000 & -0.3778 & -0.0213 \\ -0.7500 & -0.5000 & -0.6306 & -0.0210 \\ -0.1154 & -0.5000 & 0.4818 & -0.8283 \\ -0.0705 & -0.5000 & 0.4770 & 0.5595 \end{pmatrix} \quad (37)$$

Uma vez encontrado  $A$  e  $D$ , o produto de ambas junto com a inversa de  $A$  retorna a matriz  $P$ , definida na *Expressão* (38). Vale ressaltar que o tempo  $t$  foi utilizado com valor 1.

$$P = \begin{pmatrix} 0.7448 & 0.1464 & 0.0387 & 0.0701 \\ 0.2196 & 0.7460 & 0.0178 & 0.0166 \\ 0.0580 & 0.0178 & 0.5290 & 0.3953 \\ 0.0701 & 0.0110 & 0.2635 & 0.6554 \end{pmatrix} \quad (38)$$

Todos os modelos evolutivos discutidos neste trabalho, determinam que os diferentes sítios na sequência evoluem da mesma maneira e com a mesma taxa. Esta suposição pode ser irrealista em dados reais, visto que a taxa de mutação pode variar entre os sítios e as mutações em locais 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. Desse modo, a inserção da variação de taxas evolutivas de heterogeneidade entre sítios no modelo, resulta em melhorias significativas para o resultado filogenético. Sendo assim, serão apresentados a seguir os modelos com variação nas taxas evolutivas de cada sítio.

## 2.4 Taxas evolutivas de heterogeneidade

A comparação de sequências homólogas revela que cada sítio evolui através de condições particulares, sendo que em alguns casos essas diferenças podem ser associadas a uma pressão seletiva devido à sua função. Isto determina que em alguns conjuntos de dados existam sítios tão

conservados ao ponto de não apresentarem substituições entre táxons relacionados, enquanto em outros sítios a variação é muito grande.

Desse modo, podem existir interferências nas inferências filogenéticas aumentando o número de substituições que não podem ser detectadas pela simples comparação das sequências. Logo, a incorporação da heterogeneidade das taxas evolutivas entre sítios levou a um novo conjunto de modelos que proporcionou um melhor ajuste dos dados observados, desencadeando em uma reconstrução filogenética mais adequada.

Vale ressaltar, que esses modelos informam como as diferentes taxas de mutação entre os sítios estão distribuídas, porém não descrevem sobre o processo de evolução da sequência em si. Deste modo, necessita-se aliar o modelo com a matriz  $Q$  de probabilidades de mutação entre as bases e um vetor  $p_0$  de probabilidades iniciais, representados pelos modelos de substituição (CYBIS, 2009). Logo, estes modelos devem ser utilizados juntamente com os modelos apresentados na seção 2.3. Alguns destes modelos serão apresentados a seguir.

#### 2.4.1 Distribuição Discreta

Este modelo caracteriza-se pela divisão dos sítios da sequência de DNA em  $C$  categorias, cada um com uma taxa de mutação distinta, sendo que a probabilidade de um determinado sítio pertencer a uma categoria  $l$  com taxa de mutação  $\mu_l$  é definida por  $q_l$ . Estas tais probabilidades estão sujeitas às restrições  $\sum_{l=1}^C q_l = 1$  e  $\sum_{l=1}^C \mu_l q_l = 1$ , onde a segunda fixa a média das taxas de mutação da sequência, e é equivalente às restrições feitas à taxa de mutação geral dos modelos de substituição de bases (seção 2.3), para que seja possível a estimação dos comprimentos dos ramos. Com estas restrições, o modelo com  $C$  categorias de taxas de mutação possui  $2C - 1$  parâmetros, além daqueles da matriz  $Q$  de transição de estados (CYBIS, 2009).

Ainda neste contexto, o processo de evolução segue o formato da filogenia, e as taxas  $\mu_l$  alongam ou encurtam os ramos para cada sítio. Assim, os sítios com taxa de mutação geral  $\mu_l$  têm taxas de mutação, de uma base para outra, definida por  $\mu_l Q$ . Em sequências reais, a taxa de mutação de cada sítio é desconhecida, levando o modelo a assumir que todos os sítios têm a mesma probabilidade de pertencer a uma determinada categoria.

As taxas de mutação  $u_l$  podem ser definidas previamente, o que acarreta em um processo mais complicado, pois depende apenas da sensibilidade e experiência do usuário. Por outro lado, a

definição destas taxas também pode ser realizada a partir dos dados, o que exige um número  $C$  de categorias prévias. A sugestão de YANG (1995) para  $C$  varia de 3 a 4 (CYBIS, 2009).

Um dos modelos, com distribuição discreta para as taxas de mutação, mais utilizados é o proposto por HASEGAWA et al. (1985): modelo de sítios invariáveis. Denotado pela simbologia I+ (por exemplo, JC69+I, HKY85+I, GTR+I, depende da matriz de transição escolhida), divide os sítios das sequências em 2 grupos (CYBIS, 2009):

- O primeiro segue o processo determinado por sua matriz de transição, com taxa de mutação  $\mu_i$ ;
- O segundo é invariável, com taxa de mutação  $\mu_0 = 0$ .

É relevante ressaltar que, se um sítio não é constante, ou seja, possui bases diferentes de acordo às diversas sequências, ele não pode pertencer à classe de sítios invariantes. Entretanto, se o sítio não pertence à classe dos sítios invariantes, ainda assim existe uma probabilidade de que ele seja constante (CYBIS, 2009).

Neste sentido, existem alguns trechos das sequências genéticas de DNA que são muito conservados, de maneira que, em termos práticos, funcionam como sítios invariáveis. Geralmente, isso ocorre devido alguma função específica desempenhada pela sequência, que seria perdida caso houvesse alteração na base. Como exemplo, pode-se citar os motivos de reconhecimento, os quais são fundamentais para o processamento da informação contida nas sequências (CYBIS, 2009).

Como vantagem da distribuição discreta, YANG (1996) destaca que os cálculos envolvidos nas análises de probabilidades são relativamente simples e rápidos. Entretanto, as estimativas das taxas são sensíveis à escolha do número de organismos, de modo que a interpretação do modelo se torna mais difícil. Além disso, não se pode fazer uma comparação dos resultados obtidos de sequências distintas quando são utilizados valores de  $C$  diferentes. Para muitos conjuntos de dados,  $C = 2$  não é suficiente para uma boa adequação do modelo, e valores maiores de  $C$  exigem a estimação de muitos parâmetros.

#### 2.4.2 Distribuição Gama

Outra maneira de considerar a variação de taxas de mutação entre os sítios é assumir que as taxas seguem uma distribuição contínua. Foram analisadas diversas distribuições, porém a mais amplamente utilizada para esse propósito é a gama. Não existe nenhuma razão biológica para a

escolha desta distribuição, mas o uso da gama decorre de sua versatilidade (YANG, 2007). O sufixo para representar este modelo é  $+\Gamma$  (exemplos: GTR $+\Gamma$ , HKY85 $+\Gamma$ , K80 $+\Gamma$ ).

Neste sentido, considere uma variável aleatória  $X$ , tal que  $X \sim \Gamma(a, b)$ , onde a função de densidade de probabilidade de  $X$  é dada por (CYBIS, 2009):

$$f(x) = \begin{cases} \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}, & x \geq 0 \\ 0, & \text{caso contrário} \end{cases} \quad (39)$$

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

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt. \quad (40)$$

Assim como no caso das taxas de distribuição discreta, não se sabe a taxa de mutação de cada sítio. Logo, para calcular a probabilidade dos dados, deve-se fazer uma média sobre todas as taxas possíveis. Como a distribuição é contínua, o resultado é obtido mediante integração (CYBIS, 2009).

YANG (1996) destaca que o modelo que designa uma distribuição gama para as taxas de mutação possui a vantagem de explicar, através de apenas 1 parâmetro, a variação das taxas, além de possuir uma interpretação simples e maior apelo biológico, devido a sua característica contínua. No entanto, o custo computacional envolvido é muito grande, a ponto de viabilizar amostras com até 6 sequências (CYBIS, 2009).

Além dos parâmetros do modelo de substituição de bases e dos comprimentos dos ramos, o parâmetro utilizado neste caso deve ser estimado a partir dos dados. Valores pequenos deste parâmetro sugerem grande variação nas taxas de mutação, enquanto que valores grandes indicam taxas próximas de 1.

Em YANG (1994), constata-se uma alternativa de qualidade para os dois modelos citados acima: o modelo da distribuição gama discretizada. Este apresenta uma fácil interpretação e uma boa aderência ao modelo de distribuição gama, além de um custo computacional compatível com o de taxas discretas. Definido pela simbologia  $+\Gamma_d$  utiliza  $C$  categorias, todas com probabilidade  $1/C$ , para aproximar a distribuição gama (CYBIS, 2009).

Os possíveis valores de  $\mu$  pertencentes ao intervalo  $(0, \infty)$  são divididos em  $C$  categorias por  $C - 1$  percentís  $(1/C, 2/C, \dots, (C - 1)/C)$ , onde a taxa de mutação  $\mu_i$  de cada categoria é representada pela média da distribuição gama, dentro dos limites determinados por cada categoria. Assim, a taxa de mutação da  $i$ -ésima categoria pode ser obtida como

$$\mu_i = \frac{\int_A^B xf(x)dx}{\int_A^B f(x)dx} = \frac{\int_A^B xf(x)dx}{1/C} \quad (41)$$

em que  $A$  é o percentil  $(i - 1)/C$  da distribuição gama,  $B$  é o percentil  $i/C$  da mesma distribuição e  $f(x)$  é a função de densidade da distribuição gama. Na prática, o usuário deve escolher com quantas categorias deseja trabalhar, de maneira a estimar apenas um único parâmetro (CYBIS, 2009).

### 2.4.3 Cadeia de Markov Oculta

Outra maneira de incorporar variação de taxas de mutação nos diferentes sítios das sequências de DNA é através de uma cadeia de Markov oculta. Conhecido pela abreviação HMM, do inglês *Hidden Markov Model*, este modelo foi apresentado por FELSENSTEIN e CHURCHILL em 1996, tendo como objetivo agregar um tratamento realista nas taxas de heterogeneidades dos sítios, através das seguintes propriedades (CYBIS, 2009):

- Permitir que as taxas de mutação variem entre os sítios;
- Não assumir que as taxas de mutação de cada sítio sejam conhecidas, mas sim inferi-las a partir dos dados;
- Permitir algum tipo de correlação entre os sítios adjacentes.

Os modelos discutidos nas seções 2.4.1 e 2.4.2 apresentam as duas primeiras propriedades citadas acima, porém ambos não permitem correlação entre taxas de sítios adjacentes, visto que a taxa de mutação de cada sítio é retirada de uma determinada distribuição de taxas em um processo independente (CYBIS, 2009).

O HMM é definido como um modelo onde a observação da formação do sistema se dá de forma indireta, como função probabilística da transição entre os estados definidos num espaço de estados discreto e finito. Por mais que todos os parâmetros do modelo sejam conhecidos, a evolução que demonstra a formação de tal sistema que governa este processo continua oculta. Em resumo, não se sabe qual o caminho ou sequência de passos exatos que levaram a uma determinada observação (IDALINO, 2010). Sendo assim, os elementos que constituem um HMM e como as sequências de observações são geradas, são apresentados a seguir.

Um HMM é caracterizado pelos seguintes elementos (SOUZA, 2013):

- $N$  representa o número de estados do modelo. Embora os estados sejam ocultos, em muitos problemas práticos, há um significado importante para os estados ou para uma sequência de estados;
- $S = \{S_1, S_2, \dots, S_N\}$  corresponde ao conjunto dos estados individuais do modelo;
- $q_t$  corresponde ao estado no tempo  $t$ ;
- $M$  representa o número de observações distintas por estado;
- $V = \{v_1, v_2, \dots, v_M\}$  corresponde ao conjunto de observações individuais;
- $A = \{a_{ij}\}$  corresponde a distribuição de probabilidades de transição dos estados, dada por:

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \quad 1 \leq i, j \leq N \quad (42)$$

- $B = \{b_j(k)\}$  corresponde a distribuição de probabilidades da observação no estado, dada por:

$$b_j(k) = P(v_k \text{ em } t | q_t = S_j) \quad 1 \leq j \leq N \quad 1 \leq K \leq M \quad (43)$$

- $\pi = \{\pi_i\}$  corresponde a distribuição inicial dos estados, dada por:

$$\pi_i = P(q_1 = S_i) \quad 1 \leq i \leq N \quad (44)$$

Dados os valores apropriados dos elementos do HMM, este pode ser usado para gerar uma sequência de observações  $O = O_1, O_2, \dots, O_T$  onde cada observação  $O_T$  é um dos símbolos de  $V$  e  $T$  é o número de observações da sequência. Sendo assim, os passos a seguir devem ser seguidos para gerar essa sequência de observações:

- 1) Escolha um estado inicial  $q_1 = S_i$  de acordo com a distribuição inicial dos estados  $\pi$ ;
- 2) Compute  $t = 1$ ;
- 3) Escolha  $O_t = v_k$  de acordo com a distribuição de probabilidade de observação do estado  $S_i$ , isto é,  $b_i(k)$ ;
- 4) Transite para um novo estado  $q_{t+1} = S_j$  de acordo com a distribuição de probabilidade de transição do estado para o estado  $S_i$ , isto é,  $a_{ij}$ ;
- 5) Compute  $t = t + 1$ ; retorne para o passo 3 se  $t < T$ ; caso contrário, termine o procedimento.

Constata-se através das observações acima, que o HMM pode ser definido pela distribuição de probabilidades de transição dos estados  $A$ , pela distribuição de probabilidade da observação no estado  $B$  e pela distribuição inicial dos estados  $\pi$ . Logo, é definida uma notação compacta para representar um HMM, dada por:



$$\lambda = (A, B, \pi) \quad (45)$$

Uma vez que as taxas de mutação dos sítios sejam atribuídas, cada sítio evolui de forma independente ao longo da filogenia de acordo com um modelo de substituição de base. Além do mais, assume-se que toda correlação entre os sítios é resultado do agrupamento de taxas altas ou baixas em sítios adjacentes (CYBIS, 2009). A Figura 8 (FELSENSTEIN; CHURCHILL, 1996) representa o processo que designa as taxas de mutação ao longo dos sítios.

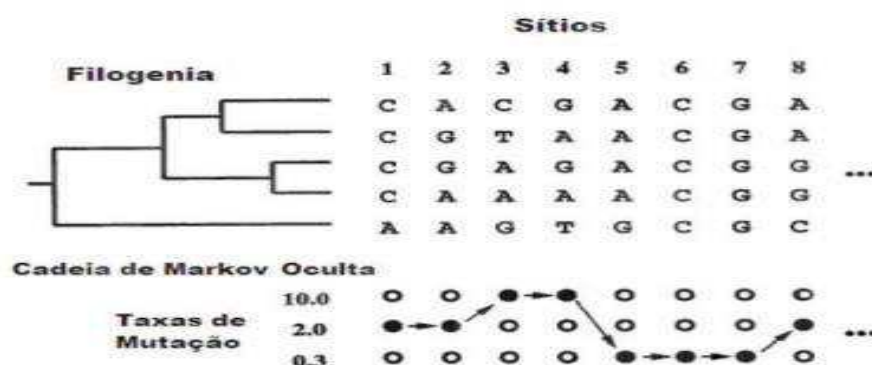


Figura 8 - Representação gráfica das mudanças de estados do modelo HMM.

### 3 MÉTODOS DE RECONSTRUÇÃO DE ÁRVORES FILOGENÉTICAS

Esta seção tem a finalidade de esclarecer os principais métodos de reconstrução de árvores filogenéticas, são eles: Distância, Máxima Parcimônia, Máxima Verossimilhança e Inferência Bayesiana.

#### 3.1 Distância

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. Os métodos de distância trabalham sobre uma estrutura padrão de dados numéricos chamada matriz de distâncias, a qual é construída com base em modelos evolutivos. Os respectivos algoritmos dos métodos são aplicados sobre a matriz de distância para se obter a melhor hipótese da árvore filogenética. Dentre os métodos de Distância mais conhecidos e utilizados, destacam-se o UPGMA (Unweighted Pair Grouping Method with Arithmetic Means) (SNEATH, 1973), o NJ (Neighbor Joining) (SAITOU; NEI, 1987), o BIONJ (GASCUEL, 1997), o Weighbor (Weighted Neighbor Joining) (BRUNO; SOCCI; HALPERN, 2000) e o FastME (Fast Minimum Evolution) (DESPER; GASCUEL, 2002). Além desses citados, tem-se o Digrafu, que reúne todos os anteriores (TORRES, 2011).

Na maioria das vezes, a entrada dos softwares de RAF consiste um arquivo contendo sequências genéticas (DNA, Proteínas, Códon) de um grupo de espécies, devidamente alinhadas. Os métodos baseados em distância não trabalham diretamente com as sequências das espécies, mas com a matriz de distância. Existem softwares que realizam a geração dessas matrizes para sequências de DNA e de proteína, como é o caso do Dnadist<sup>3</sup> e Prodist<sup>4</sup> (fazem parte do pacote PHYLIP<sup>5</sup>, PHYLogeny Inference Package), respectivamente (TORRES, 2011).

Em suma, os algoritmos dos métodos de distância selecionam um par de espécies (nós) a ser fundido a cada passo utilizando um critério específico. Nesse sentido, 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 (TORRES, 2011).

Nesse sentido, pode-se destacar dois passos principais na heurística destes métodos, que são repetidos até que a árvore esteja completa, são eles (TORRES, 2011):

---

<sup>3</sup> Consultar o endereço <http://evolution.genetics.washington.edu/phylip/doc/dnadist.html> para mais detalhes.

<sup>4</sup> Disponibilizado através do endereço <http://evolution.genetics.washington.edu/phylip/doc/protdist.html>.

<sup>5</sup> Página oficial do PHYLIP: <http://evolution.genetics.washington.edu/phylip.html>.

- 1º Passo – consiste na escolha de pares de nós a serem fundidos, ou seja, trocados por um novo nó simples representando o imediato ancestral comum deles;
- 2º Passo – as distâncias do novo nó para todos os outros nós são calculadas.

Esses dois passos são justamente o que diferencia cada um dos métodos. Sendo assim, as abordagens a seguir decifram as particularidades de cada método citado e a matriz de distância.

### 3.1.1 Matriz de Distâncias

Todos os métodos baseados em distância trabalham sobre uma estrutura padrão chamada matriz de distâncias, as quais 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 (PRADO 2001).

A matriz de distância é criada de acordo à análise de dois tipos de distância: a distância estimada ou observada e a distância evolutiva ou genética.

A distância observada  $p$  corresponde a proporção de sítios de nucleotídeos de duas sequências diferentes, onde este valor é obtido através da divisão de diferenças entre duas sequências ( $n_p$ ) e número de nucleotídeos comparados ( $n$ ), conforme equação a seguir (PEDROSA, 2013), (FELSENSTEIN, 2004).

$$p = \frac{n_p}{n} \quad (46)$$

Em suma, para construção de uma filogenia válida, as distâncias observadas devem ser corrigidas de maneira a levar em consideração as substituições múltiplas no mesmo sítio. Esta correção resulta nas distâncias evolutivas e é aplicada através de métodos que assumem diferentes hipóteses a respeito do processo de evolução. Um destes métodos é o de Jukes-Cantor (consultar item 2.3.1), o qual determina a distância genética entre duas sequências através da equação a seguir (PEDROSA, 2013), (FELSENSTEIN, 2004).

$$d = -\frac{3}{4} \ln\left(1 - \frac{4}{3}p\right) \quad (47)$$

A qualidade da árvore inferida está diretamente relacionada com a qualidade de estimação dessas distâncias. Desse modo, a construção de uma boa matriz de distância é crucial na implementação dos métodos baseados em distância, pois é através dos relacionamentos descritos nesta matriz que a árvore será construída (NEI; KUMAR, 2000; WEIR, 1996).

Após a criação da matriz de distâncias, o passo seguinte é utilizá-la em uma heurística qualquer dos métodos baseados em distância. Algumas dessas heurísticas serão analisadas a seguir.

### 3.1.2 UPGMA

É um método simples e rápido, muito utilizado no passado, que busca encontrar uma árvore que expresse os relacionamentos entre as espécies. É aplicável à construção de filogenias moleculares apenas quando as taxas de substituições de genes é basicamente constante. Sendo assim, este método ainda é utilizado por evolucionistas que se interessam apenas no relacionamento evolutivo, e não na semelhança gênica entre as espécies.

A seguir são descritos os passos de construção da árvore utilizando o UPGMA (PRADO, 2001):

- Entre com uma matriz de distâncias de tamanho  $l = n \times n$ ;
- Encontre o menor valor de distância entre os nós  $a$  e  $b$  da matriz;
- Crie 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:

$$d_{ui} = \frac{d_{ij}}{2} \quad (48)$$

- As distâncias do nó  $u$  aos demais nós da matriz pode, então, ser calculada:

$$d_{uk} = \frac{d_{ki} + d_{kj}}{2 * m * n} \quad (49)$$

onde  $k$  pode ser qualquer nó ainda não fundido na 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. Vale 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:

$$d_{AB} = \sum_{ij} \frac{d_{ij}}{rs} \quad (50)$$

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).

- Neste instante os nós  $a$  e  $b$  são retirados da matriz e o tamanho  $l$  da matriz é reduzido de um;
- Volte as passo 2, se ainda restem nós na matriz, caso contrário, a árvore está concluída.

### 3.1.3 Neighbor-Joining (NJ)

O Neighbor-Joining (SAITOU; NEI, 1987) é o mais popular dentre os métodos baseados em distância. Atua de acordo ao processo aglomerativo introduzido por Sattah e Tervsky (1977), no qual um par de taxon é fundido a cada passo, utilizando como critério o método matemático dos quadrados-mínimos (RANWEZ; GASCUEL, 2002). É aplicável a grandes conjuntos de dados devido a sua baixa complexidade de tempo computacional, e sua precisão têm sido demonstrada em simulações de computador.

Desse modo, seu processamento caracteriza-se por dois passos principais até o resultado final da árvore, são eles:

- Escolha dos pares de táxons a serem unidos, isto é, substituídos por um novo nó simples representando o imediato ancestral comum deles;
- As distâncias do novo nó para todos os outros é calculada.

Serão detalhados a seguir a heurística deste método para encontrar a árvore final.

- A árvore inicial tem o formato de estrela;

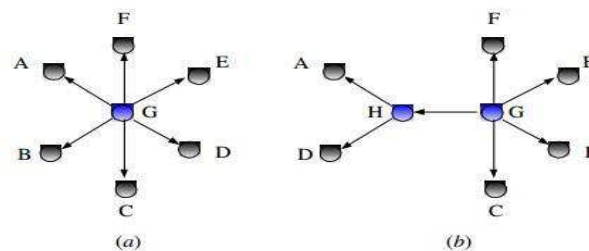


Figura 9 - A árvore a é inicial, tem formato estrela e centro representado por G. A árvore b recebeu um novo nó, H, o qual liga os vizinhos A e D.

- A matriz inicial é de distâncias corrigidas por qualquer modelo;
- Para cada nó terminal da árvore, calcular a divergência de todos os outros táxons. A distância entre os nós  $i$  e  $k$  é representada por  $d_{ik}$ ;

$$r_i = \sum_{k=1}^N d_{ik} \quad (51)$$

- Crie uma nova matriz de distância corrigida  $M$ , a qual é composta pelos elementos:

$$M_{ij} = d_{ij} - (r_i + r_j)/(N - 2) \quad (52)$$

- Um novo nó  $u$  é definido, ligados a  $i$ ,  $j$  e o restante da árvore por 3 ramos distintos. Os comprimentos dos ramos de  $u$  para  $i$  e  $j$  é definido da seguinte forma:

$$v_{iu} = d_{ij}/2 + (r_i + r_j)/2(N - 2) \quad (53)$$

$$V_{ju} = d_{ij} - v_{iu} \quad (54)$$

- Defina as distâncias de  $u$  para cada um dos outros nós terminais e insira-os na matriz de distâncias modificada

$$d_{ku} = (d_{ik} + d_{jk} - d_{ij})/2 \quad (55)$$

- Remova as distâncias para os nós  $i$  e  $j$  da matriz e decremente  $N$  em 1;
- Neste momento  $u$  é um nó terminal. Caso existam mais de dois nós terminais, volte ao passo 3. Caso contrário, siga para o próximo passo;
- O tamanho do último ramo que liga os dois táxons restantes é:

$$v_{ij} = d_{ij} \quad (56)$$

#### 3.1.4 BIONJ

O BIONJ (GASCUEL, 1997) é uma implementação melhorada do método NJ. A primeira parte do algoritmo NJ foi mantida, na qual dois táxons são escolhidos e unidos por outro ramo. Por outro lado, um fato inovador foi inserido no cálculo das distâncias estimadas para esse novo nó, atribuindo pesos às distâncias dos nós unidos. Nesse sentido, o método pretende aproximar-se mais da realidade, visto que as espécies envolvidas no caminho evolutivo dos seres ainda existentes ou já extintos, muito possivelmente não tiveram contribuições de taxas iguais para as mutações ocorridas.

A inovação do BIONJ é justamente trabalhar com pesos diferentes para cada um dos nós a serem assumidos. Esses pesos são atualizados a cada passo e armazenados em uma matriz auxiliar.

### 3.1.5 Weighbor

O Weighbor (BRUNO; SOCCI; HALPERN, 2000) é um método baseado em distância que utiliza critérios de funções de verossimilhança (*Item 3.3*). Substitui o critério de evolução mínima para a escolha dos nós a serem unidos em uma dada iteração por um critério de verossimilhança. Escolhe os nós a serem unidos de uma maneira diferente do NJ, mantendo o cálculo dos tamanhos dos ramos não muito diferente do BioNJ, que já o havia reformulado.

Em síntese, o critério de verossimilhança do Weighbor consiste em dois termos:

1. Aditividade: avalia divergências de aditividade dos ramos externos;
2. Positividade: garante que o tamanho dos ramos são positivos.

### 3.1.6 FastME

O FastME (DESPER; GASCUEL, 2002) é baseado no princípio de otimização através de trocas de árvores. Atua sobre uma árvore filogenética inicial, trocando os galhos dessa árvore entre si para poder analisar outras topologias próximas da topologia inicial. Existem dois métodos de geração de árvore inicial e dois métodos de otimização, são eles, respectivamente: FASTNNI e BNNI, e GME e BME. Assim, a inclusão do FASTME representa a inclusão de quatro combinações possíveis de métodos de otimização: GME+ FASTNNI, GME+BNNI, BME+FASTNNI e BME+BNNI (TORRES, 2011).

## 3.2 Máxima Parcimônia

O método de máxima parcimônia (FARRIS, 1972; FITCH, 1971) consiste em buscar a(s) árvore(s) que detém um número mínimo de mudanças evolutivas (substituições de nucleotídeos ou aminoácidos). De acordo aos princípios filosóficos, quando existem semelhanças entre as diversas respostas para uma determinada situação, a de hipótese mais simples tem preferência sobre as mais complexas. Este conceito é conhecido como a navalha de Occan (FELSENSTEIN, 2004), e caracteriza o método em estudo (TICONA, 2008).

Neste sentido, para cada possibilidade de topologia, determina-se o número de sequências em cada nó, de maneira a contabilizar o menor número de alterações, encontrando-se assim a mais parcimoniosa. Uma vez que existam múltiplas soluções, outros critérios de desempate devem ser empregados (SWOFFORD, 1996). A modelagem matemática deste método será destacada a seguir,

com o objetivo de exemplificar os conceitos citados e possibilitar o esclarecimento de potenciais dúvidas.

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  $\tau$  é determinada pela *Equação (57)*, na qual  $Par_j$  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  $\tau$  (TICONA, 2008).

$$Par(\tau) = \sum_{j=1}^{N_{sit}} Par_j \quad (57)$$

Sendo assim, o cálculo de  $Par_j$  pode ser definido através da *Equação (58)*, onde (TICONA, 2008):

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

$$Par_j = \sum_{(v,u) \in E} C_{v_j, u_j} \quad (58)$$

Vale ressaltar, que o cálculo do valor de parcimônia  $Par(\tau)$ , esclarecido através das explicações acima, correspondem a cada sítio individualmente, de acordo à topologia e aos estados dos nós (TICONA, 2008).

Um exemplo didático será exibido a seguir para facilitar a compreensão do método de máxima parcimônia. Nesse sentido, de forma hipotética, foi determinado a existência de 4 espécies (táxon), com as suas respectivas sequências genéticas de DNA, conforme descrito na

*Tabela 2* (adaptada de HYPÓLITO, 2005). Cabe observar, que as colunas formadas pelas sequências são conhecidas como sítios. Estes são classificados como informativos ou não informativos de acordo as diferenças nucleotídicas. Sendo assim, apenas os sítios 2, 3, 4 e 5 sofreram alterações, sendo portanto classificados como informativos. Já os sítios 1, 6 e 7 foram conservados, sendo estes não informativos.



Tabela 2 - Alinhamento de 4 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

Como descrito no item 2, mais precisamente através das *Equações (1) e (2)*, a quantidade de topologias de uma determinada árvore depende do número de espécies que se deseja analisar e do tipo de topologia (com ou sem raiz). Neste sentido, para facilitar o entendimento, optou-se pela sem raiz, a qual determina 3 topologias possíveis para 4 espécies (Figura 10) (HYPÓLITO, 2005). Sendo assim, a análise será feita com base nessas 3 topologias buscando-se identificar a mais parcimoniosa, ou seja, a que possui o menor número de eventos evolutivos para explicar as diferenças observadas entre as sequências genéticas em estudo.

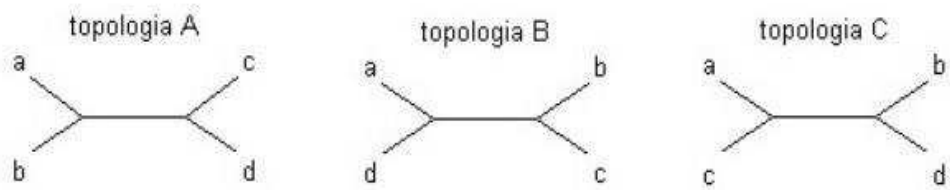


Figura 10 - Topologias (sem raiz) possíveis para 4 espécies (a, b, c e d).

Para calcular o número de mudanças é necessário identificar 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. Logo, como a ideia é descobrir a topologia com o menor número de mudanças evolutivas possíveis, o estado a ser assumido será o que minimizar este valor (HYPÓLITO, 2005).

Sendo assim, analisando a topologia A de acordo à *Figura 11* (HYPÓLITO, 2005), 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;

- O 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.

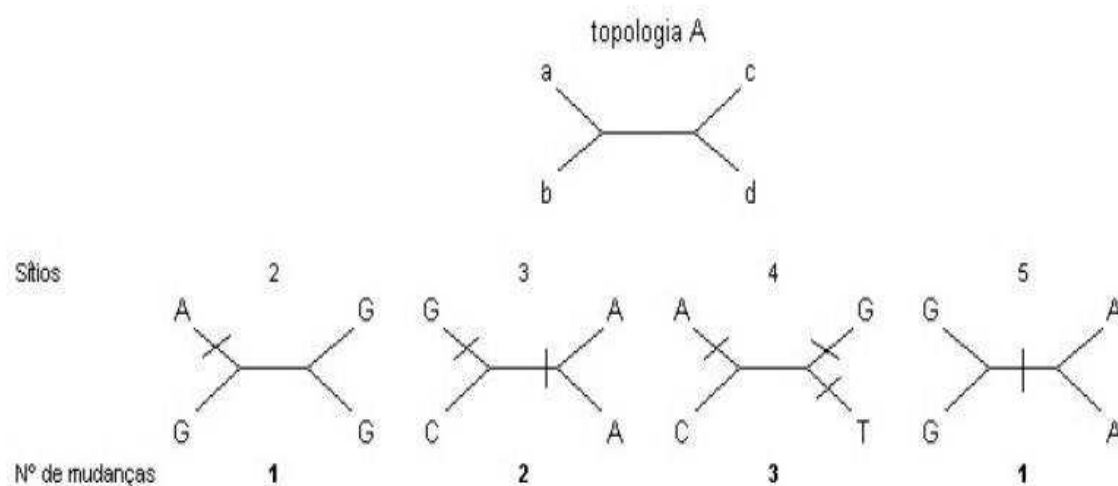


Figura 11 - Identificação da Topologia A, exibindo as alterações nucleotídicas através dos traços que cortam os ramos (total de 7 mudanças).

Fazendo a mesma análise para as topologias B e C (*Figura 12 e Figura 13*) (HYPÓLITO, 2005), constata-se um total de 8 mudanças para cada. Sendo assim, como a topologia A detém o menor número de mudanças nucleotídicas, pode-se afirmar que a mesma explica a sequências de dados com o menor número de passos possível e, portanto, será a árvore mais parcimoniosa. A seção a seguir aborda as técnicas utilizadas para busca da melhor árvore.

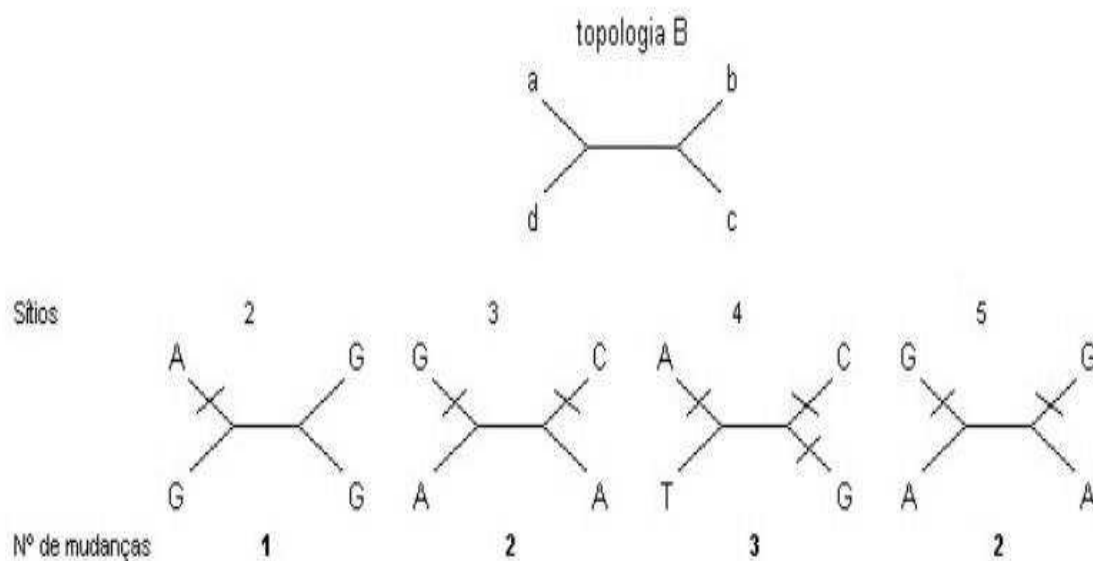


Figura 12 - Representação gráfica da topologia B, totalizando 8 mudanças nucleotídicas através dos cortes nos ramos.

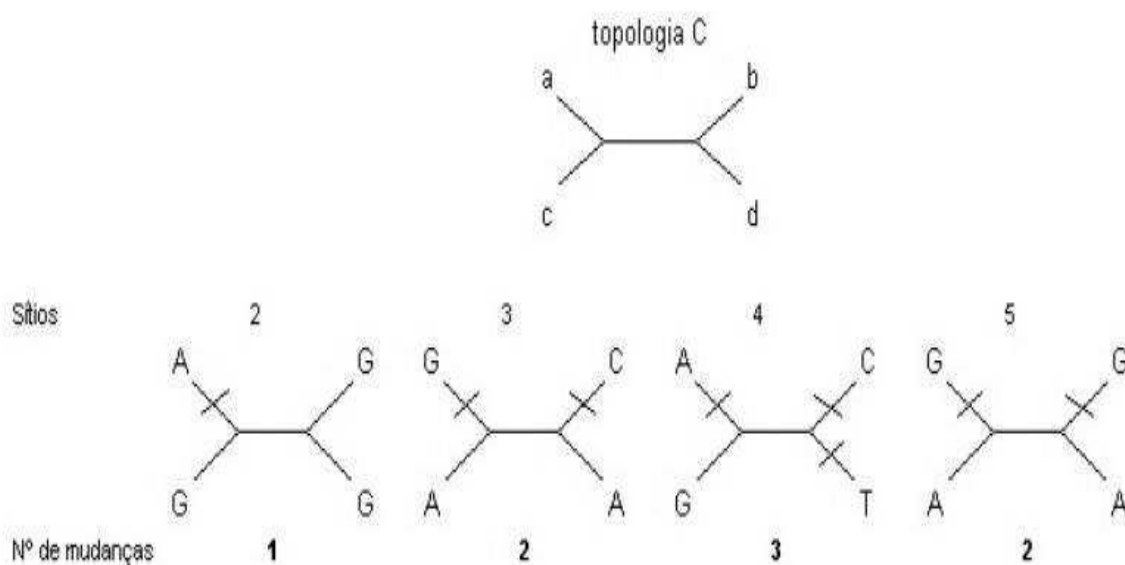


Figura 13 - Topologia C, com suas respectivas mudanças nucleotídicas, identificadas através dos traços que cortam os ramos, com o valor total de 8.

### 3.2.1 Estratégia de busca da melhor árvore

A busca pela árvore ótima, que otimize um determinado critério, é um problema muito complexo devido ao amplo espaço de alternativas possíveis. Em geral, utiliza-se duas técnicas de busca para resolver tal problema, são elas: busca exata ou busca heurística (TICONA, 2008).

As técnicas baseadas em busca exata consistem em procurar a solução ótima em toda a amplitude do espaço. Como exemplo desta, pode-se citar a busca exaustiva, na qual as espécies são adicionadas gradativamente, uma de cada vez, explorando todas as topologias possíveis, até encontrar a solução ótima. Um outra técnica, conhecida como *branch and bound* (HENDY; PENNY, 1982), avalia todas as topologias do espaço de busca, descartando regiões cuja exploração não levem a árvore ótima. Tais técnicas são vantajosas porque conseguem fornecer a topologia ótima, embora sejam adequadas apenas para conjunto de dados com poucas espécies. Na prática, para casos com grandes quantidades de espécies, tais técnicas necessitam de muito tempo computacional, sendo consideradas inviáveis (TICONA, 2008).

As técnicas de busca heurística começam com uma árvore inicial não ótima, sobre a qual são aplicadas várias trocas de ramos, de forma iterativa, visando melhorar tal solução. Na construção desta árvore inicial, os seguintes métodos podem ser empregados (NEI e KUMAR, 2000; SWOFFORD, 2000):

- Adição por passos (*stepwise addition*): inicia com uma árvore de 3 espécies, incluindo as demais de forma iterativa, até que todas estejam inseridas. O local onde a nova folha será inserida é escolhido analisando todos os ramos possíveis, escolhendo o melhor de acordo a algum critério de otimalidade. A Figura 14 (TICONA, 2008) exibe uma ilustração deste método;
- Decomposição estrela (*star decomposition*): este método inicia com uma topologia em estrela, na qual todas as espécies estão unidas a um nó interno. Posteriormente, duas espécies são agrupadas e separadas da estrela mediante a criação de um novo nó interno. A seleção de tais espécies pode ser realizada aleatoriamente ou analisando todas as alternativas possíveis. Este processo é repetido iterativamente até que seja formada uma árvore.

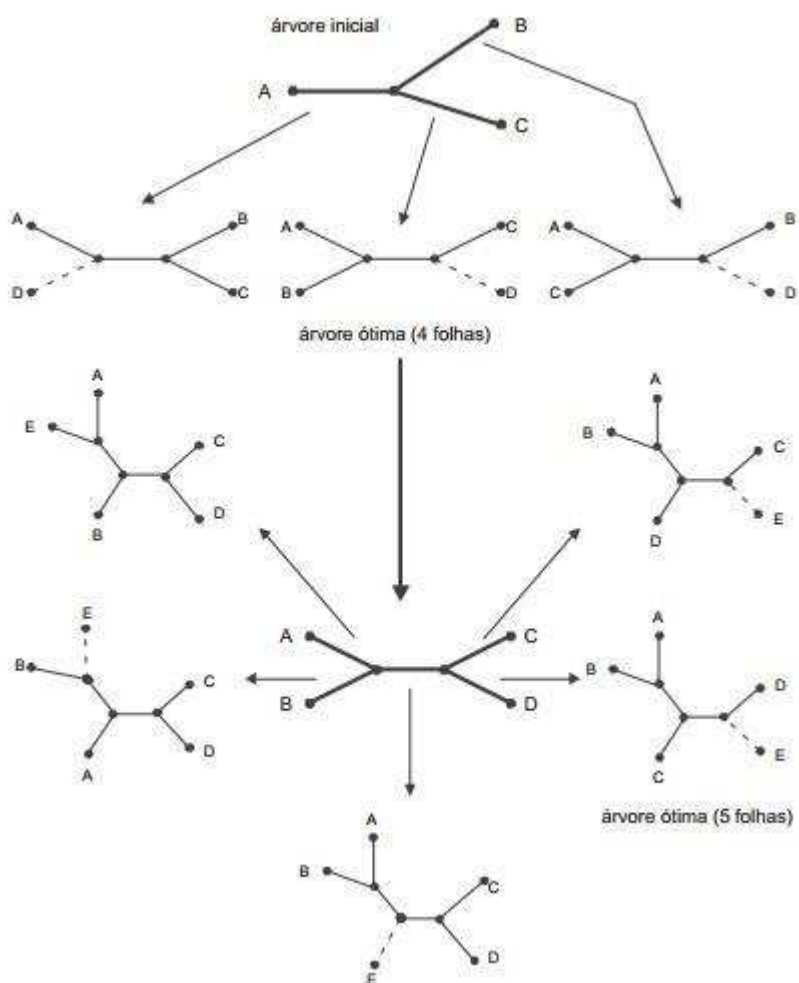


Figura 14 – Exemplo do Adição por passos.

Segundo SWOFFORD (1996), os métodos citados acima raramente levam à árvore ótima. Assim, outras modificações de árvore podem ser empregadas para melhorar as soluções fornecidas por tais métodos. Estas modificações são aplicadas na árvore original, de forma iterativa, com o intuito de encontrar a melhor solução. Existem três formas usuais de modificação topológica, são elas:

- Troca dos vizinhos mais próximos (NNI, do inglês *Nearest Neighbor Interchange*): esta técnica troca subárvores vizinhas de pares diferentes, modificando a árvore inicial. Todas as operações são efetuadas nos ramos da árvore original, até que a melhor solução seja retornada. A Figura 15 (TICONA, 2008) esclarece tal técnica;

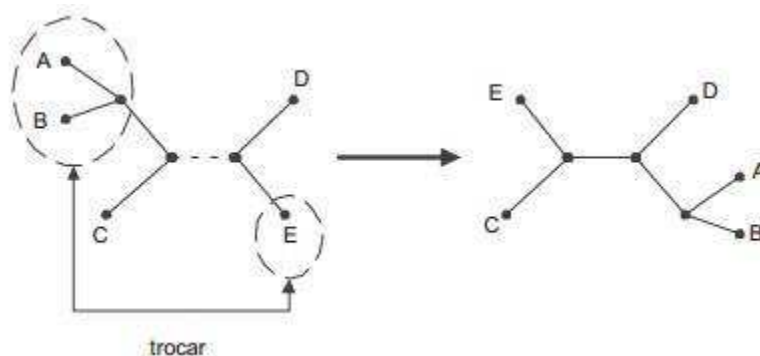


Figura 15 – Exemplo do NNI.

- Poda e inserção de subárvore (SPR, do inglês *Subtree Pruning and Regrafting*): esta operação separa uma subárvore da solução inicial e, posteriormente, é reinserida em todas as posições possíveis. Tal processo é repetido para todas as subárvores da solução inicial, retornando a melhor solução encontrada. O SPR aplica uma busca mais ampla do que o NNI, permitindo avaliar um maior número de árvores. A Figura 16 (TICONA, 2008) apresenta um exemplo deste processo;

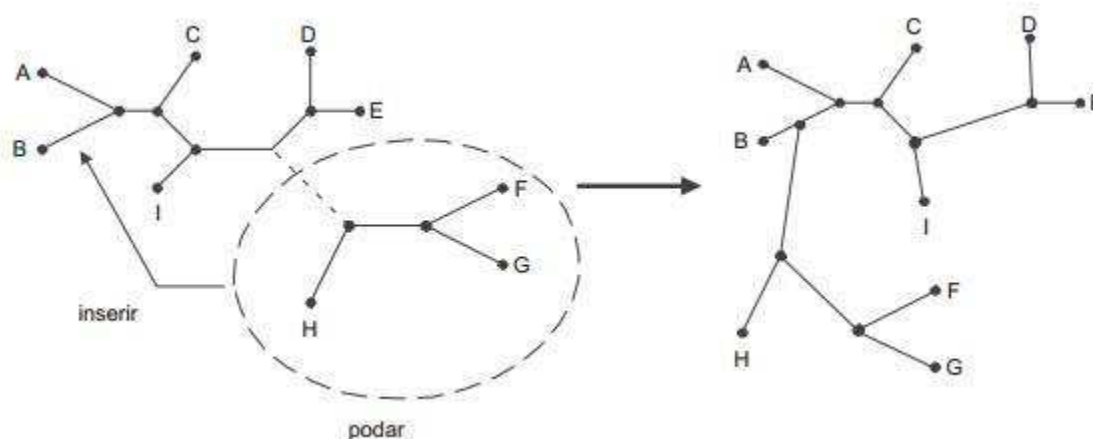


Figura 16 – Exemplo do SPR.

- Bisseção e reconexão de árvore (TBR, do inglês *Tree Bisection and Reconnection*): esta técnica consiste em eliminar um ramo interno da árvore original, originando duas subárvores. Posteriormente, tais subárvores são reconectadas através da criação de um novo ramo. Todas as subárvores e todas as reconexões possíveis são examinadas, retornando a melhor árvore encontrada. O TBR permite explorar um maior número de soluções que o SPR. A Figura 17 (TICONA, 2008) exhibe o processo do TBR.

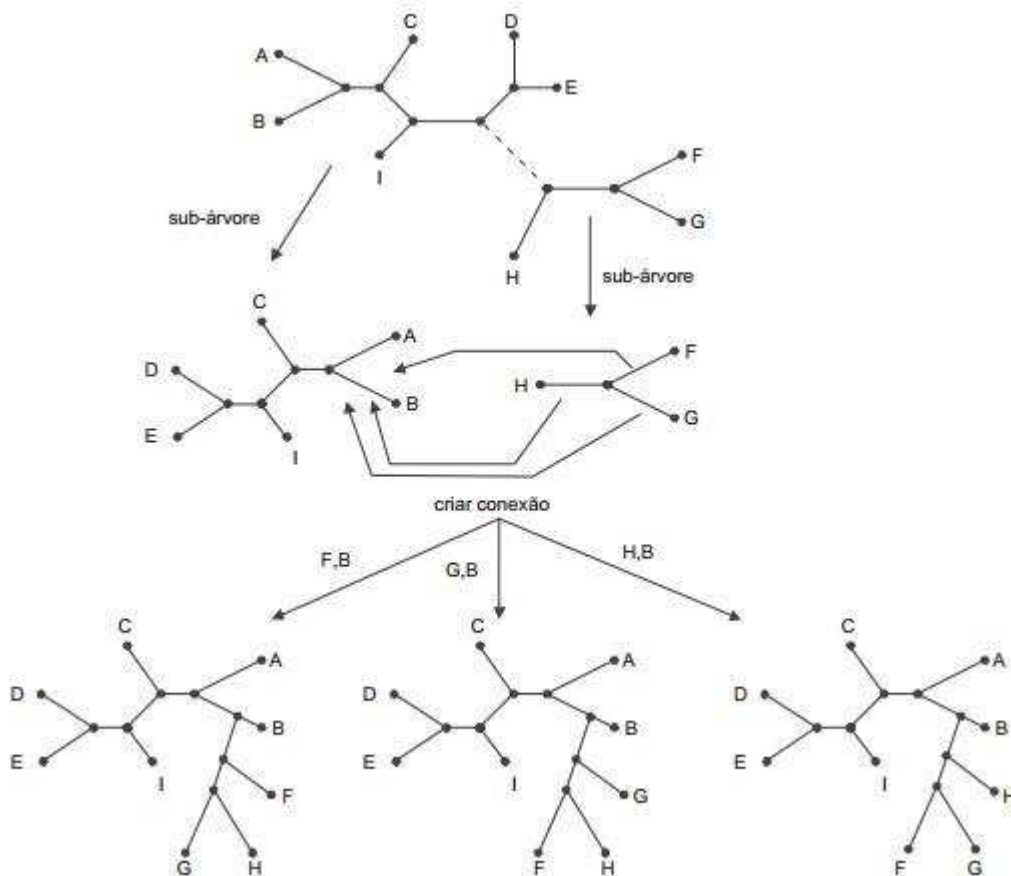


Figura 17 – Exemplo do TBR.

É relevante salientar, que as técnicas descritas nesta seção são independentes, podendo ser empregadas em outros métodos de RAF.

### 3.3 Máxima Verossimilhança

A verossimilhança determina a probabilidade  $\mathcal{P}(\mathcal{D}|\theta)$  de um conjunto de dados  $\mathcal{D}$  (sequências genéticas devidamente alinhadas) ajustar-se ao modelo  $\theta = \{\tau, \mathcal{B}, \mathcal{M}\}$ , no qual  $\tau$  é uma topologia da árvore,  $\mathcal{B}$  é o conjunto de comprimento dos ramos de  $\tau$  e  $\mathcal{M}$  é o modelo evolutivo de substituição de sequências genéticas. Portanto, o objetivo principal da verossimilhança é encontrar os parâmetros do modelo  $\theta$ , de modo que a função de verossimilhança, definida como  $L(\theta) = \mathcal{P}(\mathcal{D}|\theta)$ , seja maximizada (TICONA, 2008).

Para facilitar a compreensão do leitor, aplicou-se um exemplo ilustrativo para estimar a verossimilhança. Sendo assim, uma árvore meramente hipotética (Figura 18) (TICONA, 2008) foi utilizada, apresentando três espécies ( $u$ ,  $w$  e  $s$ ), com duas espécies ancestrais ( $v$  e  $r$ ), e com todos os seus respectivos comprimentos dos ramos.

Ainda neste contexto, suponha que as espécies ( $u$ ,  $w$  e  $s$ ) pertencem a um conjunto de dados  $\mathcal{D}$ , 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 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 de substituição de sequências genéticas, que permita calcular as probabilidades de transição entre os estados. Logo, as declarações acima servirão de base para esclarecer os conceitos descritos a seguir (TICONA, 2008).

Sendo assim, seguindo as premissas definidas por FELSENSTEIN (2004), pode-se afirmar que o cálculo da verossimilhança pode ser determinado através de um produto conforme a equação a seguir (TICONA, 2008):

$$L = \prod_{j=1}^{N_{sit}} \mathcal{P}(\mathcal{D}^j, \theta) \quad (59)$$

na qual,  $\mathcal{P}(\mathcal{D}^j, \theta)$  representa a verossimilhança no sítio  $j$ , a qual será chamada de  $L_j$  para facilitar a explicação. Como os nós internos são desconhecidos, o valor será igual a soma das probabilidades de cada cenário possível, levando em conta todos os possíveis estados. Sendo assim, de acordo a árvore descrita na Figura 18,  $L_j$  pode ser expressa da forma:

$$L_j = \sum_{r_j \in \Omega} \sum_{v_j \in \Omega} \pi_{r_j} P_{r_j s_j}(t_{rs}) P_{r_j v_j}(t_{rv}) P_{v_j u_j}(t_{vu}) P_{v_j w_j}(t_{vw}) \quad (60)$$

onde  $r_j$  e  $v_j$  representam os possíveis estados para os nós internos  $r$  e  $v$ ,  $t_{ij}$  é o comprimento do ramo que conecta os nós  $i$  e  $j$ ,  $\pi_{r_j}$  é a frequência do nucleotídeo correspondente ao estado  $r_j$  no conjunto de sequências  $\mathcal{D}$  e  $P_{x,y}(t)$  é a probabilidade da mudança do estado  $x$  para o estado  $y$  após um tempo  $t$  (TICONA, 2008). Vale ressaltar, que o termo  $P_{x,y}(t)$  foi explicado nos itens 2.2 e 2.3, sendo calculado de acordo ao modelo de evolução escolhido.



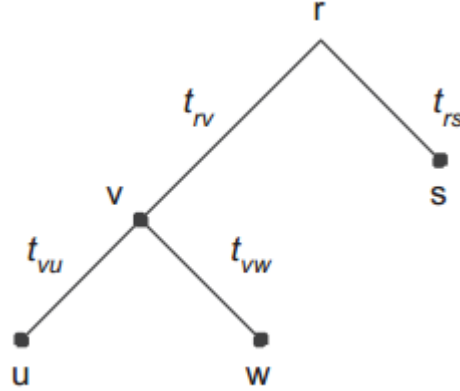


Figura 18 - Árvore para exemplificar o cálculo de verossimilhança.

O cálculo da verossimilhança também pode ser feito através da aplicação de recursividade nas subárvores. Para o caso da *Figura 18*, a verossimilhança 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, 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:

$$L_j^r(r_j) = \left[ \sum_{v_j \in \Omega} P_{r,v_j}(t_{rv}) L_j^v(v_j) \right] \times \left[ \sum_{s_j \in \Omega} P_{r,s_j}(t_{rs}) L_j^s(s_j) \right] \quad (61)$$

e para uma determinada folha  $a$ , no qual o estado  $a_j$  é fornecido por  $D$ , conforme a *Equação (62)* (TICONA, 2008).

$$L_j^a(x) = \begin{cases} 1, & \text{se } a_j = x \\ 0, & \text{caso contrário} \end{cases} \quad (62)$$

Logo, para o exemplo especificado pela *Figura 18* e pela *Equação (62)*, tem-se que:

$$L_j^v(v_j) = P_{v_j,u_j}(t_{vu}) P_{v_j,w_j}(t_{vw}) \quad (63)$$

$$L_j^s(s_j) = P_{r,s_j}(t_{rs}) \quad (64)$$

Assim, substituindo os termos nas *Equações (60)* e *(61)*, tem-se que:

$$L_j = \sum_{r_j \in \Omega} \pi_{r_j} L_j^r(r_j) \quad (65)$$

Conforme já mencionado, o cálculo da verossimilhança total é realizado através do produto dos valores  $L_j$  de todos os sítios (*Equação (59)*). Porém, estes valores são muito pequenos, podendo ocasionar erros de precisão numérica. Logo, a utilização de logaritmos naturais apresenta

resultados mais adequados. Sendo assim, aplicando o logaritmo natural em ambos os lados da Equação (59), têm-se (TICONA, 2008):

$$\ln L = \sum_{j=1}^{N_{sit}} \ln L_j \quad (66)$$

O modelo de substituição de sequências genéticas  $M$  define a suposição de que os sítios dos dados  $D$  evoluem a uma taxa constante. Porém, constata-se em dados de sequências reais, que os sítios evoluem com taxas diferentes, resultando em melhorias na análise de verossimilhança. Estas taxas são incorporadas ao modelo através de diversas formas, dentre elas pode-se citar: taxas de heterogeneidades específicas por sítio (item 2.4.1) e taxas de heterogeneidade Gama (item 2.4.2) (TICONA, 2008).

As taxas de heterogeneidades específicas por sítio são incorporadas ao modelo  $M$  através de vetor  $W = [w_1, w_2, \dots, w_{N_{sit}}]^T$ , no qual  $w_j$  representa a taxa de evolução correspondente ao sítio  $j$ . O cálculo da verossimilhança  $L$  é realizado da mesma forma descrito anteriormente, com uma ressalva: cada comprimento de ramo  $t_{ij}$  será multiplicado por  $w_j$  na obtenção das verossimilhanças das subárvores (Equação (61)). Esta abordagem é vantajosa, visto que o tempo gasto para realizar o cálculo de verossimilhança não sofre alterações significativas (STAMATAKIS, 2006). Porém, os valores de  $w_j$  devem ser fornecidos a priori, aumentando o número de parâmetros a serem estimados (TICONA, 2008).

Já na inserção da taxa de heterogeneidade Gama no modelo,  $w_j$  é uma variável aleatória obtida de uma distribuição Gama contínua ( $\Gamma$ ) (YANG, 1993). Assim, para este caso, a verossimilhança para um sítio  $j$  é dada por (TICONA, 2008):

$$L_j = \int_0^{\infty} \mathcal{P}(\mathcal{D}^{(j)} | \theta, w_j = x) f(x) dx \quad (67)$$

onde  $f$  é a função de densidade de probabilidade com distribuição  $\Gamma$ , e  $\mathcal{P}(\mathcal{D}^{(j)} | \theta, w_j = x)$  é a verossimilhança do sítio  $j$ , de tal forma que a taxa neste sítio seja  $x$ . Em termos práticos, o cálculo da integral é muito custoso computacionalmente. Sendo assim, uma função discreta  $\Gamma$  é inserida para aproximar tal valor (YANG, 1994):

$$L_j = \int_0^{\infty} \mathcal{P}(\mathcal{D}^{(j)} | \theta, w_j = x) f(x) dx \approx \sum_{k=1}^{N_{cat}} \rho_k \mathcal{P}(\mathcal{D}^{(j)} | \theta, w_j = x_k) \quad (68)$$

onde a distribuição  $\Gamma$  para as taxas dos sítios é discretizada em  $k = 1 \cdots N_{cat}$  categorias,  $x_k$  corresponde a taxa de evolução da categoria  $k$  e  $\rho_k$  é a probabilidade da categoria  $k$ . A *Equação (68)* pode ser escrita também da seguinte forma:

$$L_j = \sum_{k=1}^{N_{cat}} \sum_{r_j \in \Omega} \rho_k \pi_{r_j} L_j^r(r_j x_k) \quad (69)$$

onde  $L_j^r(r_j, w_j = x_k)$  é obtida da mesma forma que  $L_j^r(r_j)$  na *Equação (61)*, multiplicando por  $x_i$  os comprimentos dos ramos  $t_{rv}$  e  $t_{rs}$  (TICONA, 2008).

A incorporação de taxa de heterogeneidade Gama é mais vantajosa porque os valores  $w_j$  são obtidos a partir da distribuição  $\Gamma$ , a qual é composta por dois parâmetros:  $\alpha$  parâmetro de forma e  $\beta$  parâmetro de escala. Em termos práticos, apenas o parâmetro  $\alpha$  é empregado, visto que  $\beta$  é fixado em  $1/\alpha$ . De outro modo, o cálculo da *Equação (57)* torna-se mais lento utilizando este tipo de taxa, pois este é realizado para as  $N_{cat}$  categorias empregadas (TICONA, 2008).

Após obter a expressão da verossimilhança, a qual está em função das probabilidades e dos comprimentos dos ramos, deve-se achar os valores de  $j$  que maximizem a verossimilhança. Um dos métodos utilizados para maximizar a função  $L$  é o método de Newton-Raphson (CUNHA, 2003). Primeiro deriva-se  $L_j$  em função do número de ramos da árvore (os valores de  $j$ ) e a expressão iguala-se 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 que proporcionar a maior verossimilhança entre as testadas.

Diversos métodos foram propostos para estabelecer valores numéricos a ramos internos em árvores filogenéticas, com o objetivo de prover uma medida do grau de suporte daqueles ramos e dos grupos correspondentes. Em suma, estes métodos estimam a confiabilidade da árvore inferida. Dentre vários, destacam-se o Bootstrap e o Jackknife, os quais serão analisados a seguir.

### 3.3.1 Bootstrap

O Bootstrap é um método utilizado para estimar as incertezas estatísticas em situações em que a distribuição da amostra original é desconhecida ou é de difícil derivação analítica. Foi proposto por EFRON em 1979, e introduzido na análise filogenética através dos trabalhos de MUELLER e AYALA (1982) e FELSENSTEIN (1985), sendo considerado um recurso crucial para se investigar o

nível de suporte para determinada filogenia, devido à facilidade de implementação e interpretação (HYPÓLITO, 2005). Pode ser empregado nos principais métodos de reconstrução filogenética (Máxima Verossimilhança, Máxima Parcimônia e Distância), sem, contudo, fazer sentido utilizá-lo juntamente com a Inferência Bayesiana, pois o mesmo já possui tais características estatísticas em seu escopo original (mais detalhes sobre este método serão descritos mais a frente).

Este método consiste em aplicar, diversas vezes, uma perturbação aleatória no conjunto de dados em questão, de modo a originar árvores réplicas, com seus respectivos valores probabilísticos. Baseia-se na suposição de que, pequenas modificações nos dados não diminuirão a capacidade de encontrar os grupamentos, se estes estiverem bem representados pelos dados. Assim, grupos com baixos valores de Bootstrap são aqueles que deixaram de ser inferidos em muitas das réplicas, o que sugere baixo suporte dos dados. Contudo, observa-se que os valores definidos não indicam a probabilidade de que os grupos sejam corretos filogeneticamente, uma vez que diferentes sequências podem determinar diferentes grupos, de acordo ao caso analisado (ALVES, 2001).

Os conjuntos de sequências gerados pelo Bootstrap possuem o mesmo número de sítios das sequências iniciais, sendo que cada sítio é escolhido aleatoriamente de acordo aos dados originais. Com isso, um determinado conjunto pode conter sítios repetidos, enquanto outros ficarão de fora. Sendo assim, essas novas sequências geradas serão submetidas ao método de inferência filogenética utilizado, gerando  $n$  árvores réplicas. Por fim, calcula-se a proporção de cada clado da árvore inicial presente nas árvores réplicas. Tal valor mede a probabilidade de um clado ser recuperado no conjunto de réplicas (TICONA, 2008).

Embora o Bootstrap seja uma técnica simples e efetiva que mede a repetibilidade dos clados da árvore inferida, os graus de suporte calculados podem ser propensos a erros se o método de inferência não for empregado corretamente. Um outro inconveniente é o tempo necessário para realizar a análise de Bootstrap. Dado que um grande número de réplicas é recomendado (entre 200 e 2000), o tempo requerido de inferência de cada réplica pode ser inviável em termos práticos (TICONA, 2008).

### 3.3.2 Jackknife

Esta técnica foi inserida na análise filogenética a partir do trabalho de MUELLER e AYALA (1982), utilizando informações empíricas sobre a variação de um caractere para outro durante o processo de evolução. Consiste em eliminar, de forma aleatória, metade dos sítios de uma sequência, de modo que a nova sequência obtida seja exatamente a metade da original. Este procedimento será

repetido diversas vezes, gerando inúmeras novas amostras, sendo que cada nova amostra servirá de base para uma nova reconstrução filogenética. Sendo assim, as frequências das subárvores são contabilizadas a partir das árvores reconstruídas, de modo que, se uma subárvore estiver presente em todas as árvores reconstruídas, então valor de Jackknife será de 100%, demonstrando assim, uma maior confiabilidade (PINTO, 2004).

### 3.4 Inferência Bayesiana

A Inferência Bayesiana é determinada através do conceito de probabilidade posterior das hipóteses filogenéticas (HYPÓLITO, 2005). A probabilidade posterior de uma topologia pode ser interpretada como a possibilidade de representar de fato a história evolutiva do grupo de espécies em estudo. Desse modo, a topologia com a maior probabilidade posterior deve ser a mais provável estimativa da evolução do grupo em questão.

O teorema de Bayes permite calcular a probabilidade posterior, pois mescla a probabilidade *a priori* da topologia, com sua probabilidade de determinar a distribuição dos estados de caráter 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 valores dos parâmetros dos modelos de substituição de nucleotídeos. Como este cálculo é impossível de ser feito de forma analítica, mesmo com a inserção de buscas heurísticas, o método de amostragem de Monte Carlo via Cadeias de Markov (MCMC) pode ser utilizado, de forma a elucidar as árvores de distribuição de probabilidades posteriores.

As distribuições de probabilidades, utilizadas na análise filogenética bayesiana, representam o conhecimento a respeito da topologia, dos comprimentos dos ramos e dos parâmetros de substituição de nucleotídeos. O cálculo da probabilidade de uma árvore ocorrer 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.

### 3.4.1 Aplicação na Filogenética

A Inferência Bayesiana calcula a probabilidade de uma árvore ser verídica através dos dados observados (alinhamento de DNA), sendo determinada pela equação a seguir (GONÇALVES, 2008):

$$f(\tau_i|X) = \frac{f(X|\tau_i)f(\tau_i)}{\sum_{j=1}^{B(s)} f(X|\tau_j)f(\tau_j)} \quad (70)$$

Onde:

- $f(\tau_i|X)$  faz referência a probabilidade posterior (valor a posteriori) da i-ésima árvore ( $\tau_i$ );
- $f(X|\tau_i)$  é a Máxima Verossimilhança de  $\tau_i$ ;
- $f(\tau_i)$  corresponde à probabilidade prévia (valor a priori) de  $\tau_i$ ;
- No denominador encontra-se o somatório que avalia todas as possibilidades de árvores de acordo à quantidade de espécies  $S$ .

### 3.4.2 Método de Monte Carlo via Cadeias de Markov

A tarefa de calcular a probabilidade posterior a partir da *Equação (70)* é considerada complexa, uma vez que os valores para os tamanhos dos galhos  $v_j$  e para as substituições de DNA  $\theta$ , ao serem utilizados na inferência bayesiana, vão corresponder a todo universo de valores possíveis, como ilustra a equação a seguir (GONÇALVES, 2008).

$$f(X|\tau_i) = \int_v \int_{\theta} f(X|\tau_i, v, \theta) f(v, \theta) dv d\theta \quad (71)$$

Portanto, a probabilidade posterior não pode ser calculada analiticamente. No entanto, ela pode ser aproximada utilizando um método numérico que amostra árvores de acordo a uma função de distribuição de probabilidade. Assim, Monte Carlo via Cadeias de Markov, conhecido como MCMC, é utilizado (GONÇALVES, 2008).

O objetivo principal do método é construir uma cadeia de Markov com uma distribuição estacionária igual a distribuição posterior de interesse. Assim, esta cadeia é iniciada de um ponto escolhido arbitrariamente, dentro do universo de valores disponíveis, e é gerada sucessivamente até ser alcançado um valor estacionário, sendo que a probabilidade posterior é coletada a cada geração.

O método de Monte Carlo foi desenvolvido por Nicolas Metropolis, em trabalho realizado junto com outros pesquisadores, para testar o ENIAC, em 1945. O algoritmo para executá-lo via Cadeias de Markov foi publicado em METROPOLIS et al. (1953).

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 pela equação a seguir (GENTLE, 2007):

$$E(g(Y)) = \int_D g(y)p(y)dy = \int_D f(y)dy \quad (72)$$

sendo  $Y$  uma variável aleatória dentro do domínio  $D$  e  $p(y)$  a sua densidade de probabilidade.

Sendo assim, o problema de calcular a integral transforma-se na estimação da média  $E(g(X))$ , cujo procedimento padrão é coletar uma amostra aleatória, com densidade uniforme ao longo de  $D$ , que corresponde ao intervalo  $[a, b]$ , e calcular a média dessa amostra, como descrito na equação a seguir (GONÇALVES, 2008).

$$\hat{I} = (b - a)E(g(X)) \quad (73)$$

Portanto, em uma amostra de tamanho  $n$  a estimativa  $\hat{I}$  é:

$$\hat{I} = (b - a) \frac{1}{n} \sum_{i=1}^n g(x_i) \quad (74)$$

A Equação (74) pode ser generalizada, de tal forma a integrar não apenas em um intervalo finito  $[a, b]$ , mas em um domínio geral, como descrito a seguir (GONÇALVES, 2008):

$$\mu = \frac{1}{n} \sum_{i=1}^n f(X_i) \quad (75)$$

Na equação acima,  $\mu$  representa a estimativa de uma função  $f(X_i)$  de interesse.

De outro modo, segundo (TRIOLA, 2003), quando as amostras  $\{X_i\}$  são independentes, a lei dos grandes números certifica que  $\mu$  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 (GILKS; RICHARDSON; SPIEGELHALTER, 1996), esclarece que as amostras  $\{X_i\}$  não precisam ser totalmente independentes, mas podem ser geradas por algum mecanismo que determina  $\{X_i\}$  a partir da função de interesse. Portanto, resolve-se esta situação, através de uma cadeia de Markov tendo  $f$  como a distribuição estacionária.

Analisando ainda sob a ótica dos conceitos de processos estocásticos, uma cadeia de Markov acontece quando uma sequência de variáveis aleatórias  $\{X_0, X_1, X_2, \dots\}$  é amostrada a cada instante  $t \geq 0$  sob a distribuição  $P(X_{t+1}|X_t)$  (GILKS; RICHARDSON; SPIEGELHALTER, 1996). Assim, o próximo estado  $\{X_{t+1}\}$  depende apenas de  $\{X_t\}$  e de nenhum outro estado na história da cadeia  $\{X_0, X_1, X_2, \dots, X_{t-1}\}$ . Deste modo,  $P(\cdot|\cdot)$  é chamada de semente de transição da cadeia e assume-se que esta distribuição é homogênea no tempo porque ela não depende de  $t$ .

Nesse sentido, é relevante salientar o quanto o estado inicial, a variável  $X_0$ , influencia no comportamento da cadeia.

Assim, espera-se que uma cadeia de Markov esqueça gradualmente o valor do seu estado inicial e seja convergida para uma única distribuição estacionária independente de  $t$  ou  $X_0$ . O procedimento utilizado para isto é conhecido como *burn in*, o qual consiste em adotar uma quantidade  $n$  de amostras, suficientemente grandes para descartar  $m$  amostras iniciais. A equação a seguir, derivada da *Equação (63)*, demonstra este procedimento (GONÇALVES, 2008).

$$\mu = \frac{1}{n - m} \sum_{i=m+1}^n f(X_i) \quad (76)$$

### 3.4.3 MCMCMC (MC<sup>3</sup>)

Uma variante do MCMC denominada Metropolis Coupled Markov Chain Monte Carlo (MCMCMC ou MC<sup>3</sup>) é utilizada para acelerar o processo de exploração de árvores e para permitir explorar com maior eficiência o espaço de possibilidade de árvores. De acordo às características do algoritmo de Metropolis-Hastings (M-H), existe uma grande possibilidade de uma MCMC ficar presa em um determinado local do espaço em análise. Uma vez que a MCMC esteja presa em um determinado local, o método M-H não garante a exploração de uma outra região no espaço estudado. Dessa forma, foi proposta uma alteração do M-H para resolver essa limitação, o MCMCMC ou MC<sup>3</sup>.

O processo do MC<sup>3</sup> aborda a execução simultânea de  $n$  cadeias de Markov, no qual  $n - 1$  são consideradas cadeias aquecidas e 1 é considerada cadeia fria. Ele simula um mecanismo de aquecimento onde quanto maior for o índice da cadeia, maior é a sua probabilidade de aceitação. Logo, a primeira cadeia é considerada fria porque tem chance menor de ser aceita em uma transição e a última cadeia é a mais quente entre as demais. No entanto, para todas as cadeias explorarem de forma eficiente as possibilidades de árvores, periodicamente, duas são escolhidas aleatoriamente e seus estados de aquecimento são trocados.



Apenas a cadeia fria é amostrada, dentre as  $n$  cadeias, determinando que haverá uma árvore registrada para cada geração do MC<sup>3</sup>. Portanto, o processo é elaborado de maneira que as árvores de maior verossimilhança possuem probabilidade maior de serem movidas para uma cadeia fria e as de menor verossimilhança têm probabilidade maior de serem movidas para uma cadeia quente. Desse modo, trocas diferentes destas duas condições são raras de acontecerem.

Vale ressaltar que, uma troca de estados entre duas cadeias permite com que uma cadeia que esteja presa em um determinado local consiga explorar outras regiões e, conseqüentemente, consegue-se obter uma melhor aproximação da distribuição de probabilidade posterior. Nesse sentido, é relevante destacar que, a cadeia aquecida tem uma visão plana, o que acaba facilitando seu deslocamento, fazendo com que visite o ótimo global (região de probabilidade posterior máxima) quando uma proposição com a cadeia fria seja feita.

Portanto, é notório o avanço alcançado pelo MC<sup>3</sup> para aproximação correta da distribuição posterior dos parâmetros filogenéticos. No entanto, devido ao seu elevado custo computacional, de acordo ao número de cadeias utilizadas, alguns softwares de filogenia propuseram uma nova versão, baseada em computação paralela, que reduz drasticamente o tempo de execução.

Uma vez que os métodos de RAF foram analisados, assim como a modelagem matemática que os auxiliam, pode-se discorrer sobre algumas implementações de software que os reproduzem, as quais serão abordadas a seguir.

## 4 IMPLEMENTAÇÕES DOS MÉTODOS

Existem várias implementações de softwares dos métodos de inferências filogenéticas citados. Dentre os vários, os escolhidos, pela sua qualidade e eficiência foram: MrBayes, PHYML, Digrafu e DNAPARS ou PROTPARS. Estes serão analisados a seguir.

### 4.1 Digrafu

O Digrafu<sup>6</sup> é um software de domínio público para RAF baseado no método de distância, produzido na UESC utilizando algoritmos renomados (UPGMA, Weighbor, BIONJ e FastME), propondo uma melhora no tratamento dos dados de entrada e na aplicação dos mesmos. Feito em grande parte pela linguagem de programação Perl<sup>7</sup>, é um programa de código fonte aberto com o objetivo de aperfeiçoar as soluções existentes. Possui quatro características 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;
- Sua execução em si.

O início de execução do Digrafu se dá pela criação da matriz de distância, através do DnaDist ou ProDist, dependendo do tipo de sequência que se deseja trabalhar (Dna ou Proteína), presente no pacote PHYLIP. O Digrafu detém uma característica de manipulação (ajuste) das sequências a fim de otimizar ou mesmo ampliar o espaço de visão do método gerador de matrizes (DnaDist ou ProDist). Assim, existe a possibilidade de explorar os dados antes mesmo da geração da matriz.

Com a finalidade de aperfeiçoar os tipos de dados suportados pelo DnaDist e pelo ProDist, que trabalham apenas com formatos de sequências definidos no pacote PHYLIP (conhecidos como sequencial e intercalado), e aproveitando-se do fato de estar em poder das sequências genéticas, foi adicionado a possibilidade de se trabalhar com o formato NEXUS. Além disso, o Digrafu reconhece automaticamente o formato de sequências que está em questão, aplicando a cada formato suportado o seu tratamento específico.

---

<sup>6</sup> Disponível no seguinte endereço eletrônico <http://github.com/said/digrafu>.

<sup>7</sup> Para mais detalhes consultar <https://www.perl.org>.

Após a criação da matriz de distâncias, o Digrafu escolhe dinamicamente qual dos métodos (UPGMA, NJ, BIONJ, Weighbor ou FastME) deverá ser executado. Alguns fatores influenciam esta escolha, são eles:

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

A preferência citada acima corresponde a pesos concedidos na escolha do método, sendo que fica a critério do usuário optar por maior rapidez na saída dos resultados ou até mesmo maior exatidão do resultado gerado. Sendo assim, pode-se afirmar que a aplicação respeita as prioridades definidas pelo usuário.

## 4.2 DNAPARS e PROTPARS

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

O PHYLIP foi criado por Felsenstein na linguagem de programação C, apto para funcionar na maioria das plataformas computacionais possíveis. É composto de 35 programas que conseguem resolver diversos problemas filogenéticos da literatura atual. Cada um detém suas particularidades, com uma documentação específica, de acordo ao método filogenético que representa.

A execução é realizada através de comandos de texto, os quais são determinantes para definir configurações específicas do usuário de acordo aos 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.

De acordo à nomenclatura intuitiva de cada software analisado para o método de máxima parcimônia, constata-se a utilização de sequências de DNA para o DNAPARS, e de proteínas para o

<sup>8</sup> Consultar o site <http://evolution.genetics.washington.edu/phylip/doc/dnapars.html> para obter mais detalhes.

<sup>9</sup> O link <http://evolution.genetics.washington.edu/phylip/doc/protpars.html> apresenta detalhes deste software.

PROTPARS. Sendo assim, a seguir serão analisadas as principais características de cada um dos softwares citados. Para o caso do DNAPARS, tem-se:

- Cada sítio e diferentes linhagens evoluem de forma independente;
- A probabilidade de uma substituição de base em um determinado sítio é pequena durante os períodos de tempo envolvidos na construção de um ramo da filogenia;
- Não existem muitas alterações de valores nos ramos da filogenia, tanto que duas mudanças são mais prováveis em ramos de altas taxas, do que uma mudança em ramos de taxas baixas.

O DNAPARS calcula a topologia inicial empregando 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 (TICONA, 2008). Recomenda-se rodar o DNAPARS várias vezes, modificando a ordem com que as espécies são acrescentadas na árvore, dado que assim é possível obter diferentes resultados em cada execução e, possivelmente, escapar de ótimos locais.

Já o PROTPARS caracteriza-se por apresentar filogenias não enraizadas de sequências de proteínas, utilizando um método baseado nas abordagens de ECK, DAYHOFF (1966) e FITCH (1971). ECK e DAYHOFF (1966) definiu que qualquer aminoácido pode ser alterado por qualquer outro, contabilizando o número de mudanças necessárias para desenvolver as sequências de proteína em cada filogenia determinada, sem considerar o problema de permitir substituições que não são consistentes com o código genético, contabilizando-as da mesma forma das substituições que são consistentes. FITCH (1971), por outro lado, contabiliza o número mínimo de substituições de nucleotídeos que seriam necessários para atingir as sequências de proteínas dadas.

Sendo assim, seguindo as ideias apresentadas no manual do PROTPARS, as principais características deste são:

- Cada sítio e diferentes linhagens evoluem de forma independente;
- A probabilidade de uma substituição de base que altera a sequência do aminoácido é pequena durante os períodos de tempo envolvidos na construção de um ramo da filogenia;
- Não existem muitas alterações de valores nos ramos da filogenia, tanto que duas mudanças são mais prováveis em ramos de altas taxas, do que uma mudança em ramos de taxas baixas;
- A probabilidade de uma mudança para uma base considerada sinônima é muito maior do que a probabilidade de uma alteração para uma base não sinônima.

Todas estas características podem ser consultadas com mais detalhes através de uma série de artigos disponibilizados na página principal do site do PHYLIP.

### 4.3 PHYML

O PHYML é um programa de inferência de árvores filogenéticas que utiliza o método de Máxima Verossimilhança, tendo um resultado consideravelmente rápido e preciso devido a sua heurística. Para implementar o MV em um tempo aceitável, o PHYML aplica a seguinte heurística:

1. Começa calculando uma árvore inicial (Figura 19) com o método BIONJ (seção 3.1.4), que servirá de modelo nos processos seguintes;

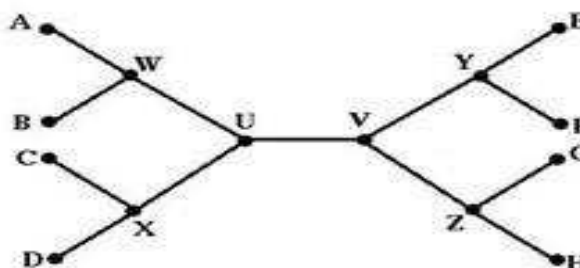


Figura 19 - Árvore inicial do PHYML.

2. A partir deste modelo, busca-se uma árvore ótima otimizando todos os galhos da árvore inicial;
3. A otimização consiste em calcular o tamanho dos galhos em função do melhor MV. Todas as trocas possíveis dos nós vizinhos ao galho são realizadas para se obter um galho ótimo (Figura 20);

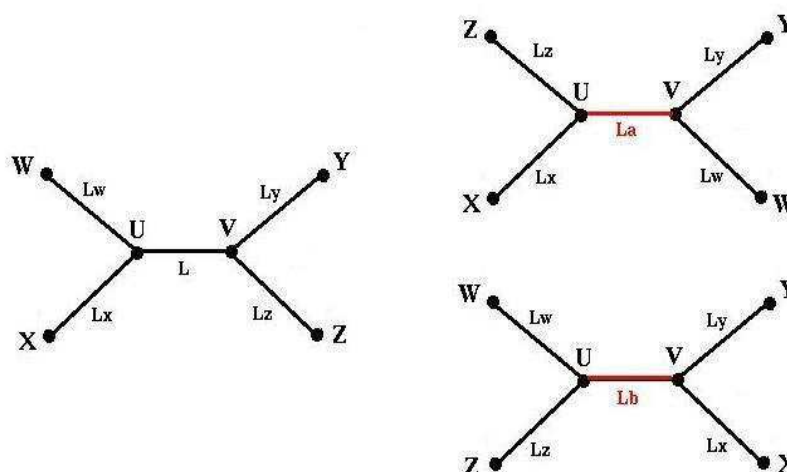


Figura 20 - Otimização do galho U – V da árvore da Figura 17.

4. Todas as topologias são descritas em uma lista, segundo o valor que obtiveram no cálculo do MV;
5. Após o processo anterior, a árvore original é modificada segundo a topologia que obteve o melhor MV;
6. Uma proporção  $\lambda$  das topologias restantes é utilizada a fim de calcular o MV total da árvore modificada;
7. Se o MV total da árvore modificada for menor que o da árvore inicial, a proporção  $\lambda$  é dividida por 2, e metade das trocas realizadas anteriormente são aplicadas com as topologias de melhores MV;
8. O passo 7 é repetido até que o MV da árvore modificada seja melhor que o da árvore inicial.

Segundo TICONA (2008), o PHYML emprega uma abordagem heurística que reduz consideravelmente o tempo de execução. Isto acontece devido a um forte relacionamento entre as modificações topológicas e a otimização dos parâmetros e dos comprimentos de ramos. Neste sentido, emprega-se o BIONJ para determinar a topologia inicial, além de algum método de otimização para estimar, inicialmente, os parâmetros do modelo de substituição. Em seguida, 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 independentemente calculadas 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 obtida é o novo ponto de partida para uma nova iteração do algoritmo, que continua até que não haja mais modificações a serem aplicadas. Por fim, os comprimentos de ramos e parâmetros do modelo são reotimizados.

O PHYML retorna o resultado de sua execução em arquivos de texto, com uma nomenclatura fixa de acordo ao nome do arquivo de sequência submetido. Sendo assim, para um arquivo de sequência de nome DnaIgrafuWeb, o retorno poderia resultar nos arquivos listados a seguir:

- DnaIgrafuWeb\_PHYML\_lk.txt: contém os valores de probabilidade para cada sítio analisado.
- DnaIgrafuWeb\_PHYML\_tree.txt: arquivo que possui o resultado da inferência filogenética submetida: contém a(s) árvore(s) filogenética(s); Sempre será criado;
- DnaIgrafuWeb\_PHYML\_stat.txt: armazena detalhes estatísticos da execução. Sempre será criado;
- DnaIgrafuWeb\_PHYML\_boot\_trees.txt: contém a(s) árvore(s) resultado da análise filogenética utilizando bootstrap. Será criado somente se a análise de bootstrap for utilizada;
- DnaIgrafuWeb\_PHYML\_boot\_stats.txt: arquivo que lista as estatísticas de execução do bootstrap. Será criado somente se a análise de bootstrap for utilizada;
- DnaIgrafuWeb\_PHYML\_trace.txt: exibem cada filogenia explorada durante o processo de execução.
- DnaIgrafuWeb\_rand\_trees.txt: lista as árvores de máxima verossimilhança provenientes de cada árvore aleatória de partida.

#### 4.4 MrBayes

O MrBayes<sup>10</sup> é um software para inferência bayesiana de árvores filogenéticas, que pode ser executado através de linhas de comandos ou em modo batch. Possui código fonte aberto, implementado na linguagem C, suportado pelas plataformas Windows, Macintosh e Unix. Atualmente encontra-se na versão 3.2.3, que é a mesma utilizada neste trabalho, empregando todos os conceitos apresentados nas seções anteriores. Além de apresentar uma documentação completa, bastante disseminada em trabalhos científicos, o MrBayes foi escolhido como objeto de estudo devido a utilização funcional do método bayesiano.

Para realizar a utilização do MrBayes é necessário definir os parâmetros de execução e o arquivo de sequências genéticas, devidamente alinhadas, do qual deseja-se inferir a árvore. A execução do MrBayes toma como base os parâmetros:

---

<sup>10</sup> Página oficial do MrBayes: <http://mrbayes.sourceforge.net>.

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

Durante a execução do MrBayes, a cada geração, as amostras do modelo de substituição são escritas, delimitados por tabulação, em um arquivo texto de extensão .p, conforme Figura 21.

```
[ID: 9409050143]
Gen      LnL      TL      r(A<->C) ... pi(G)      pi(T)      alpha      pinvar
1         -5723.498  3.357  0.067486 ... 0.098794 0.247609 0.580820 0.124136
10        -5727.478  3.110  0.030604 ... 0.072965 0.263017 0.385311 0.045880
....
9990      -5727.775  2.687  0.052292 ... 0.086991 0.224332 0.951843 0.228343
10000     -5720.465  3.290  0.038259 ... 0.076770 0.240826 0.444826 0.087738
```

Figura 21 - Exemplo de arquivo de saída do MrBayes.

O número descrito entre colchetes é uma marcação gerada aleatoriamente que permite identificar a análise. A segunda linha contém os cabeçalhos das colunas e as linhas seguintes contêm os valores amostrados. Eles estão organizados da esquerda para direita, representando:

1. O número de gerações (Gen);
2. A probabilidade de transição (LnL);
3. O comprimento total da árvore (a soma de todos os comprimentos dos ramos). (TL);
4. Os parâmetros de taxas do modelo GTR, são seis: (r(A <-> C), r(A <-> G), r(A <-> T), r(C <-> G), r(C <-> T), r(G <-> T));
5. As quatro frequências de nucleotídeos estacionários: (pi(A), pi(C), pi(G), pi(T));
6. O parâmetro que forma a taxa de variação da distribuição gama (alpha);
7. A proporção de sítios invariáveis (pinvar).

A seção a seguir descreve como os parâmetros citados são utilizados na implementação do MCMC.

#### 4.4.1 Implementação do MCMC

Como descrito anteriormente, o MrBayes implementa o MCMC através de um processo conhecido como algoritmo de Metrópolis-Hastings, que consiste de uma árvore filogenética que



passa por sucessivas alterações ou estados de transição até chegar ao estado de maior probabilidade posterior.

A Figura 22 (GONÇALVES, 2008) ilustra o ciclo seguido pelo MrBayes para geração da árvore. Os parágrafos a seguir descreverão em detalhes os procedimentos executados neste ciclo.

De início é necessário definir a quantidade de gerações e obter a árvore inicial, que corresponde ao estado inicial da cadeia de Markov. Esta árvore inicial pode ser gerada aleatoriamente, ou fornecida pelo usuário. Caso seja definida pelo usuário, a mesma árvore é atribuída para todas as cadeias, e então, inicia-se o ciclo de gerações de cadeias onde as atividades listadas na *Figura 20* acontecem em todas as cadeias da execução.

Em cada geração é proposto um movimento que significa modificar algum parâmetro em uma das três partes da árvore, escolhidas aleatoriamente: a topologia ( $\tau$ ), o tamanho dos galhos ( $v_i$ ) ou o modelo de substituição de DNA ( $\theta$ ). A realização de um movimento, mesmo que não seja na árvore, sempre vai interferir no comportamento da cadeia, já que um valor diferente para um parâmetro, implicará na variação do escore da cadeia, que contribuirá para ser mantida ou substituída na próxima geração.

Após a modificação citada acima, aplica-se uma função de distribuição de probabilidade que determina se a nova árvore é aceita ou recusada. Caso seja aceita, a cadeia assume a nova árvore, caso contrário, a árvore atual é mantida e o programa segue para a próxima geração. Este processo é repetido até o final da quantidade de gerações.

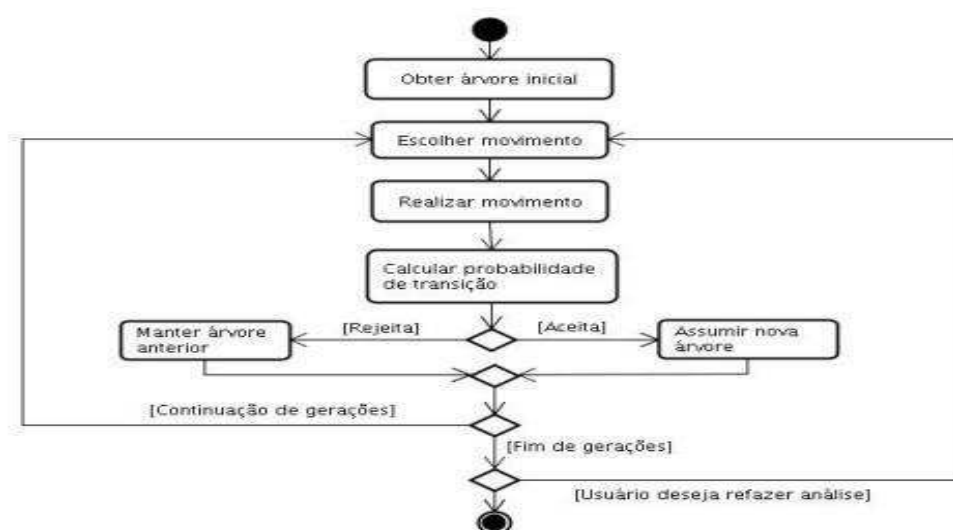


Figura 22 - Diagrama do MCMC implementado no MrBayes.

#### 4.4.2 Probabilidade de Transição

Segundo LARGET e SIMON (1999), o cálculo da probabilidade de transição consiste em determinar se acontecerá ou não uma transição de uma cadeia, ou seja, a substituição da árvore anterior por uma nova, e ocorre de acordo aos itens a seguir:

- Determina-se a probabilidade  $R$  de uma transição acontecer usando a equação abaixo, onde  $\psi'$  representa a árvore nova e  $\psi$  a árvore corrente:

$$R = \min \left( 1, \underbrace{\frac{f(X|\psi')}{f(X|\psi)}}_{\text{verossimilhança}} * \underbrace{\frac{f(\psi')}{f(\psi)}}_{\text{priori}} * \underbrace{\frac{f(\psi|\psi')}{f(\psi|\psi)}}_{\text{proposta}} \right) \quad (77)$$

- Um número aleatório  $n$ , distribuído de forma uniforme entre 0 e 1, é gerado;
- Caso  $n$  seja menor que  $R$ , acontecerá a transição para uma árvore nova, caso contrário, a anterior será mantida.

Pode-se observar na *Equação (77)* a utilização de 3 taxas, calculadas entre a árvore nova e a árvore atual, para se determinar a probabilidade de transição, são elas:

- Verossimilhança: calcula a probabilidade da árvore atual sobre a nova, observando as sequências genéticas fornecidas;
- Priori: calcula a probabilidade prévia, observando valores desejados, tais como distribuição uniforme, exponencial ou dirichlet, para tamanhos de galhos;
- Proposta: calcula um valor denominado taxa *Hasting*, que representa os movimentos necessários de transição da árvore atual para a árvore nova e vice-versa.

#### 4.4.3 Detecção de Convergência das Cadeias

O MrBayes utiliza amostras das cadeias frias para calcular o diagnóstico que indica o momento ideal de parada de sua execução. Este diagnóstico necessita de duas ou mais execuções do MC<sup>3</sup>, para comparar quão parecidas estão as amostras em cada execução, utilizando um cálculo conhecido como desvio padrão médio das frequências de partições das árvores.

Este cálculo é iniciado construindo uma lista de todas as partições (divisão das espécies de uma árvore em duas partes, obtida com o seccionamento de um galho) encontradas durante a amostragem. As partições que ainda não foram encontradas em amostras anteriores serão guardadas, e as que já foram encontradas, receberão o incremento de suas frequências. Logo, o programa

percorre a lista de partições assim que ocorre um diagnóstico e, para cada elemento desta lista, calcula o desvio padrão das frequências de cada execução. Após obter o desvio padrão de cada partição, o valor que expressará o diagnóstico é a média dos desvios padrão.

No entanto, um dos principais problemas do MCMC é detectar o estado de convergência devido a imprecisão das técnicas existentes. Durante a execução das cadeias, quando todas as árvores amostradas convergem para a mesma árvore, obtém-se o resultado da inferência bayesiana.

Dentre as diversas técnicas de detecção da convergência das cadeias markovianas, a mais simples é a observação do valor da Máxima Verossimilhança (MV) das árvores representadas em cada cadeia. Desse modo, cada cadeia inicia com uma árvore aleatória, sendo que nas primeiras gerações obtém-se uma taxa crescente de MV, estacionando após várias gerações. Isto pode indicar que a convergência foi alcançada, mas nada impede, após um tempo estacionado, que a taxa de determinada árvore volte a crescer.

De outro modo, a técnica apresentada pode informar taxas de MV parecidas ou iguais para representar árvores totalmente diferentes. Nesse sentido, para corrigir esta falha, o MrBayes utiliza uma técnica para atestar a convergência das cadeias baseadas na topologia de suas árvores. Assim, antes da execução do software é definida uma frequência de gerações em que o diagnóstico é aplicado em todos os conjuntos de amostras simultaneamente. A seguir são apresentados os passos que resumem todo este processo:

- Uma frequência  $f$  de iterações é determinada para cada execução amostrar a árvore filogenética de sua cadeia fria;
- Uma frequência  $d$  de iterações é determinada para diagnosticar a convergência das árvores amostradas;
- São aplicadas  $r$  execuções independentes de MC<sup>3</sup>, cada uma com  $c$  quantidade de cadeias e  $n$  iterações;
- Para verificar a convergência das árvores, observa-se o diagnóstico a cada  $d$  iterações;
- Se o diagnóstico é ruim as iterações continuam, caso contrário, são interrompidas.

Ao final deste processo, alcançando-se a convergência desejada, a árvore filogenética, que representa o resultado da análise bayesiana com o MrBayes, é obtida através dos trechos das árvores que aconteceram com maior frequência, montando uma árvore nova representando um consenso das amostras.

#### 4.4.4 Amostragem, diagnóstico e sumarização de dados

O MrBayes utiliza 3 recursos bastante relevantes na inferência filogenética bayesiana, são eles:

- Amostragem de dados: consistem em armazenar os dados lotados na cadeia fria da análise para, posteriormente, serem utilizados na etapa de sumarização. Os dados desta fase são as árvores e os demais parâmetros do modelo filogenético obtidos em uma determinada geração. O MrBayes amostra, por padrão, os dados de uma análise a cada 100 gerações de árvores, sendo que o usuário pode alterar este valor;
- Diagnóstico de dados: consiste no cálculo de algumas medidas estatísticas dos dados amostrados, de maneira a informar ao usuário se a execução chegou a um resultado ideal. A medida de similaridade das árvores amostradas em cada análise é a principal delas, visto que a existência de uma convergência indica que as análises alcançaram a distribuição estacionária do MCMC. Esta similaridade é obtida calculando o desvio padrão da frequência que alguns galhos aparecem nas amostras das análises. Pode-se utilizar esta medida como critério de parada em algumas execuções, principalmente nos casos onde não se tem ideia da quantidade de gerações a ser utilizada. O diagnóstico adota ainda, o descarte de amostras iniciais, denominado *burnin*, que visa não incluir nos cálculos as árvores iniciais cuja probabilidade é inferior às de distribuição estacionárias;
- Sumarização dos dados: é a fase final da inferência bayesiana e está fora dos ciclos de execução. Existem dois tipos de sumarização: a de parâmetros estimados e a de árvores. A primeira consiste em realizar uma análise estatística dos parâmetros evolutivos que foram amostrados nas análises. As médias calculadas são: a média, a variância, o menor valor, o maior valor, a mediana e um diagnóstico de convergência das árvores denominado Fator de Redução Escalar Potencial que mede se a quantidade de amostras foi suficiente para um bom suporte estatístico. A sumarização de árvores consiste em contabilizar as árvores amostradas nas análises para determinar aquelas que obtiveram a maior probabilidade posterior, ou seja, as suas estruturas topológicas apareceram com maior frequência na amostra.

## 5 METODOLOGIA E DESENVOLVIMENTO

Uma vez que os modelos matemáticos dos métodos de RAF, e os softwares (MrBayes, PHYML, Digrafu e DNAPARS/PROTPARS) disponibilizados para cada método foram estudados e entendidos nos capítulos anteriores, foi possível implementar de fato o desenvolvimento da ferramenta Web proposta neste trabalho.

Sendo assim, o objetivo deste trabalho é oferecer uma solução intuitiva para RAF baseado nos métodos Bayesiano, Distância, Parcimônia e Verossimilhança. Mais precisamente, permitir ao usuário determinar os parâmetros filogenéticos dos programas MrBayes, PHYML, Digrafu ou DNAPARS/PROTPARS em um ambiente gráfico, Web, executá-lo em uma plataforma de alto desempenho e visualizar o resultado final, a árvore filogenética em si, no próprio sistema. Em suma, o projeto tem o objetivo de facilitar a usabilidade do usuário final, visto que não precisará realizar nenhuma das tarefas citadas abaixo:

- Instalação e configuração do(s) software(s) que realiza(m) RAF;
- Montagem dos scripts necessários para utilização desses softwares, com uma sintaxe específica para cada, pois eles são utilizados em modo texto;
- Instalação e configuração dos softwares que permitem a visualização gráfica da(s) árvore(s), que são apresentadas em arquivos texto.

Todas essas tarefas estão disponibilizadas no software apresentado neste trabalho, além da performance oferecida pelo computador de alto desempenho da UESC, CACAU.

O software resultante deste trabalho está disponível através do link <http://nbcgib.uesc.br/igrafuweb>, de forma gratuita e sem restrição de acesso.

Para desenvolver o projeto foi necessário realizar diversas tarefas, que serão descritas a seguir.

### 5.1 Metodologia

Esta seção tem como objetivo principal apresentar as ferramentas técnicas utilizadas para o desenvolvimento do presente trabalho.

### 5.1.1 Softwares de RAF

Como já mencionado, os software estudados e disponibilizados foram os: MrBayes, PHYML, Digrafu e DNAPARS/PROTPARS. Todos foram devidamente estudados e analisados a fundo, pois foi necessário instalá-los no CACAU e entender a heurística e os parâmetros de cada, de forma a disponibilizar na ferramenta todos os recursos oferecidos por eles. Vale ressaltar a utilização de versões paralelas dos softwares citados.

### 5.1.2 CACAU

O Centro de Armazenamento de Dados e Computação Avançada da UESC (CACAU), que está localizado nas dependências do Núcleo de Biologia Computacional e Gestão de Informações Biotecnológica (NBCGIB<sup>11</sup>), foi adquirido justamente para oferecer à comunidade acadêmica um ambiente computacional de alto desempenho. Detém uma estrutura física composta por 20 nós, cada um composto de 2 processadores Intel(R) Xeon(R) E5430@ de 2.66 GHz e 16 GB de memória, totalizando 160 cores e 320 GB de memória.

O CACAU trabalha com um gerenciador de fila que controla a submissão de trabalhos para execução, o Slurm<sup>12</sup>. Um estudo detalhado e minucioso foi realizado a fim de compreender a utilização desta ferramenta, que teve fundamental importância na conclusão deste projeto.

Vale ressaltar, que todos os softwares aqui analisados foram devidamente instalados e configurados no CACAU para dar suporte à ferramenta proposta neste projeto.

### 5.1.3 PHP

A linguagem de programação utilizada para desenvolver o Web Services proposto neste trabalho foi o PHP (acrônimo recursivo para "PHP: Hypertext Preprocessor"), juntamente com o framework Yii<sup>13</sup> na versão 1.1.15. A escolha da linguagem e do framework citados foi realizada de acordo a estudos de usabilidade, e foi muito influenciada pela experiência apresentada pela equipe de trabalho com estas tecnologias.

---

<sup>11</sup> Página oficial do NBCGIB: <http://nbcgib.uesc.br/nbcgib>.

<sup>12</sup> Para mais detalhes consultar <https://computing.llnl.gov/linux/slurm>.

<sup>13</sup> Framework para desenvolvimento em PHP. Para mais detalhes consultar <http://www.yiiframework.com>.

#### 5.1.4 Software de visualização de Árvores Filogenéticas

O software utilizado no projeto para visualizar a árvore final foi o Archaeopteryx. Como o projeto é Web, o foco foi direcionado para uma ferramenta que apresentasse a possibilidade de visualização da árvore on-line. Isto só foi possível através de um Applet, software que executa uma atividade específica, dentro (do contexto) de outro programa maior (como por exemplo um Web browser, que é o caso deste projeto).

Analisando o contexto das plataformas Web, vale ressaltar que a grande maioria dos atuais navegadores de internet, dão suporte à plataforma Java. Para utilizá-la basta que esta esteja devidamente instalada na máquina do cliente. Desse modo, é necessário a instalação desta, visto que o software Archaeopteryx foi desenvolvido na linguagem Java. De maneira mais específica, o Archaeopteryx exige a versão 1.7 ou superior do Java.

É relevante salientar que o Java não é pré-requisito de funcionamento do IgrafuWeb, mas sim do Archaeopteryx, software que permite a visualização gráfica do resultado final da inferência filogenética. Sendo assim, se o cliente não esteja disposto a instalar o Java em sua máquina, ele pode visualizar a árvore final em um arquivo texto ou até mesmo utilizar uma outra plataforma.

## 5.2 Desenvolvimento

Esta seção explica em detalhes o desenvolvimento do projeto, a fim de elucidar todas as características e funcionalidades do IgrafuWeb.

O software apresentado neste projeto disponibiliza a RAF na Web. Nesse sentido, a heurística, os parâmetros e a sintaxe de cada um desses programas foram devidamente estudados, através do manual dos mesmos, com a intenção de disponibilizá-los no IgrafuWeb. Todos estes parâmetros representam a modelagem explicada, para cada método, em capítulos anteriores, além de alguns acréscimos particulares.

Como é notório no site, o mesmo detém um menu, que identifica justamente os métodos citados (*Figura 23*), são eles:

- Método de Inferência Bayesiana – disponibilizado através do software MrBayes;
- Método de Máxima Verossimilhança – oferecido pelo programa PHYML;
- Método de Distância – ofertado através do software Digrafu;
- Método de Máxima Parcimônia – ofertado através do software DNAPARS ou PROTPARS.

## IgrafuWeb

Bayesian Likelihood Distance Parsimony

### Welcome to IgrafuWeb

IgrafuWeb is a free and simple to use web service dedicated to reconstructing tree phylogenetic between molecular sequences.

IgrafuWeb connects various bioinformatics programs to reconstruct a robust phylogenetic tree from a set of sequences.

The software available in IgrafuWeb are:

- MrBayes representing the Inference Bayesian method
- PHYLML representing the Maximum Likelihood method
- Digradu representing the Distance method
- DNAPARS or PROTPARS representing the Maximum Parsimony method

To use it simply configures the parameters of each software, choose a sequence file and call the process execution via the button "Execute".

Figura 23 - Página inicial apresentado a estrutura organizacional do site e uma breve explicação do IgrafuWeb.

O passo inicial para realizar a RAF é a escolha de um dos métodos citados, de acordo ao menu do site, e o arquivo de entrada, contendo sequências de DNA ou de proteínas devidamente alinhadas (Figura 24). O sistema disponibiliza o uso de um exemplo de arquivo de sequência, caso o usuário assim deseje. Esta opção é habilitada através da marcação do campo “*Example File*” da aba *Sequence* de cada método.

```
#NEXUS
BEGIN DATA;
  DIMENSIONS NTAX=12 NCHAR=898;
  FORMAT DATATYPE=DNA INTERLEAVE=NO GAP=-;
  MATRIX
TARSIVS_SYRICHTA      AAGTTTCATTGGAGCCACCACCTCTTATAATTGCCCATGGCCTCACCTCCTCCCT
LEMUR_CATTA           AAGCTTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCAT
HOMO_SAPIENS          AAGCTTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATT
PAN                   AAGCTTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATT
GORILLA               AAGCTTTCACCGGCGCAGTTGTTCTTATAATTGCCACGGACTTACATCATCATT
PONGO                AAGCTTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCT
HYLOBATES            AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCCT
MACACA_FUSCATA       AAGCTTTTCCGGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCAT
M_MULATTA            AAGCTTTTCTGGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCAT
M_FASCICULARIS       AAGCTTCTCCGGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCAT
M_SYLVANUS           AAGCTTCTCCGGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCAT
SAIMIRI_SCIUREUS     AAGCTTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTAT
;
END;
```

Figura 24 - Exemplo de arquivo (formato NEXUS) de sequências de DNA alinhadas.

Ainda neste contexto, foi realizado um tratamento para identificar as quantidades de sequências e de sítios, com o intuito de utilizar os números adequados de nós (máquinas) do CACAU, mantendo assim a performance do mesmo. Caso as sequências e os sítios possuam números maiores que 4 e 50, respectivamente, o processamento será executado com a utilização de 4 nós, caso contrário, apenas 1 nó será utilizado.



É importante ressaltar a exigência do sistema enquanto a submissão do arquivo de sequência, caso o usuário assim escolha, além de impedir a submissão de um arquivo vazio. Uma vez escolhido a opção do arquivo de sequência, fica a critério do usuário determinar os parâmetros do software escolhido de acordo aos seus conhecimentos. A seguir será feita uma análise dos principais parâmetros de cada método.

### 5.2.1 MrBayes

Os parâmetros do MrBayes foram devidamente organizados em 4 abas com o intuito de melhorar a usabilidade do site (*Figura 25*). Os nomes das abas são ***Sequence***, ***Model***, ***MCMC***, ***Tree*** e ***Summarization***, e serão analisadas a seguir.

#### Sequence

Esta aba contém apenas as opções de definição do arquivo e do tipo de sequência que se deseja trabalhar, através dos campos “***File***” e “***Data Type***”, respectivamente. Vale ressaltar, a possibilidade de utilização de um arquivo de sequência sugerido pelo sistema através da opção “***Example File***”. Todas essas informações podem ser consultadas através da *Figura 25*.

MrBayes

Sequence Model MCMC Tree Summarization

Sequence

Example File File Escolher arquivo Nenhum arquivo selecionado

Data Type  
DNA ▼

Execute Clear Custom Command Block Visualize Download

Figura 25 – Campo da aba Sequence do MrBayes.

#### Model

Esta seção foi subdividida em Protein e DNA de acordo as especificações definidas pelo MrBayes. Desse modo, serão descritos a seguir, os parâmetros do modelo que estão devidamente organizados na *Figura 26* e na *Figura 28*.

O MrBayes implementa os seguintes modelos de substituição de DNA: GTR, SYM, HKY, K2P, F81 e JC (observar item 2.3). Estes pertencem a 3 estruturas de modelos de substituição, são elas:

- **“4by4”**: são modelos simples utilizados para analisar a evolução de nucleotídeos (opção padrão);
- **“doublet”**: modelo que se destina a analisar regiões do tronco emparelhados do DNA ribossomal;
- **“codon”**: modelo que analisa a sequência de DNA em termos de seus códons. Esta opção possibilita a definição do código genético e da variação ômega.

Além disso, taxas de substituição entre os sítios podem ser modeladas com uma proporção de sítios invariáveis, com uma variação de acordo a uma distribuição gama ou uma combinação das duas. Estas taxas serão explicadas com mais detalhes a seguir.

Conforme observado na *Figura 26* e na *Figura 28*, as opções que o MrBayes descreve para as taxas de heterogeneidade entre sítios são (observar item 2.4):

- **“equal”**: não existem variações entre os sítios;
- **“gamma”**: taxas de distribuição gama entre os sítios. A taxa é definida através de uma distribuição gama determinada pelo parâmetro *“lset rates=gama”*. A distribuição gama é contínua, porém é virtualmente impossível de calcular probabilidades sob a distribuição gama contínua. Assim, uma aproximação à gama contínua é utilizada: a distribuição gama é dividida em categorias *“ngammacat”* discretas de igual peso ( $1/ngammacat$ ). A taxa média para cada categoria representa a taxa para todas as categorias. Esta opção permite a escolha de quantas categorias de taxas serão utilizadas para usar quando aproximar a gama. A aproximação é melhor quanto *ngammacat* é maior. Na prática, *“ngammacat=4”* faz um trabalho razoável de aproximação da gama contínua. Pode-se alterar essa configuração utilizando o comando *“lset ngammacat”*. Por exemplo, caso deseja-se utilizar oito categorias de taxas distintas, o comando apropriado seria *“lset ngammacat=8”*. Uma análise com quatro categorias de gama discretos é quatro vezes mais lento do que uma análise sem variação entre sítios e duas vezes mais rápida do que uma com oito categorias;

- **“*adgamma*”**: taxas auto correlacionadas entre sítios. Neste caso, as taxas variam entre os sítios de acordo com um modelo de gama auto correlacionado, em que a taxa em cada sítio depende, em certa medida, das taxas em sítios adjacentes. A auto correlação é medida pelo parâmetro  $r$ , o qual varia de -1 (auto correlação negativa, isto é, sítios adjacentes tendem a ter muitas taxas diferentes) para 1 (sítios adjacentes têm taxas muito semelhantes). A probabilidade prévia padrão para  $r$  é uma distribuição uniforme cobrindo todo o intervalo (-1,1). O parâmetro do MrBayes que representa este intervalo é o *ratecorrpr*, sendo que este possui as opções *uniform(<number>,<number>)* ou *fixed(<number>)*. Tal como acontece com o modelo de gama, a distribuição gama auto correlacionada é aproximada com um certo número de taxas de categorias discretas determinadas pelo comando “*lset ngammacat*”;
- **“*propinv*”**: proporção de sítios invariáveis. Este modelo é chamado usando o comando “*lset rates=propinv*”. A proporção dos sítios invariáveis é referido como “*pinvarpr*”, que pode variar de 0 (nenhum sítio invariável) a 1 (todos os sítios são invariáveis). Este valor anterior é uma distribuição uniforme no intervalo (0,1). O comando “*pinvarpr*” possui as opções “*uniform(<number>,<number>)*” ou “*fixed(<number>)*”;
- **“*invgamma*”**: a proporção de sítios invariáveis provenientes de uma distribuição gama. É referenciado pelo comando “*lset rates=invgamma*”. Este modelo também é aproximado com um certo número de taxas de categorias discretas determinadas pelo comando “*lset ngammacat*”.

## MrBayes

The screenshot shows the MrBayes software interface with the 'Model' tab selected. The 'DNA' sub-tab is active. The 'Model Type' is set to '4by4'. The 'Substitution Model' is 'GTR'. The 'Rate Variation' is 'equal'. The 'Gamma Categories' is set to '4'. The 'Advanced Settings' button is visible. Under 'Advanced Settings', the 'Codon' section shows 'Genetic Code' as 'Universal' and 'Omega Variation' as 'equal'. The 'Adgamma prior' section has 'Uniform' and 'Fixed' options, both with a value of '0'. The 'Propinv prior' section also has 'Uniform' and 'Fixed' options, both with a value of '0'.

Figura 26 - Definição dos parâmetros para o modelo de DNA do MrBayes.

Além disso, a análise Bayesiana com o MrBayes oferece a especificação de uma distribuição de probabilidade prévia para os parâmetros do modelo de probabilidade (botão *Advanced Settings* da aba modelo de DNA). Nesse sentido, vale ressaltar a relevância dos comandos do MrBayes que define as especificações para o modelo filogenético. Estes comandos permitem a personalização das

suposições prévias, as quais representam as crenças anteriores sobre o parâmetro, antes da observação dos dados. As opções destes comandos relacionados aos modelos citados anteriormente são (Figura 27):

- **“*Tratiopr*”**: parâmetro que define a prévia para a razão da taxa de transição / transversão. Pode ser fixo, um único valor para cada taxa, com a sintaxe *fixed(<number>)*, ou dois valores iguais ou distintos representando as taxas de transição e transversão respectivamente, com a sintaxe *beta(<number>, <number>)*;
- **“*Revmatpr*”**: parâmetro que define a prévia para as taxas de substituição do modelo *GTR* para dados de nucleotídeos. Assume seis valores representando os tipos de substituição dos nucleotídeos na ordem A <-> C, A <-> G, A <-> T, C <-> G, C <-> T e G <-> T. Possui as opções **“*dirichlet*”** e **“*fixed*”**;
- **“*Covswitchpr*”**: essa opção define a prévia para as taxas de comutação covarion. As taxas podem assumir valores individuais, uniforme ou exponencialmente distribuídos. A opção **“*Covswitchpr*”** só é válida se a opção **“*model covarion*”** estiver marcada. O **“*model covarion*”** obriga a utilização de um modelo covarion semelhante ao de substituição de nucleotídeos ou de aminoácidos. O modelo covarion permite que a taxa em um sítio possa mudar ao longo da história evolutiva. As opções para os sítios são: ligado ou desligado. Quando está desligado, as substituições não são possíveis. Quando o processo estiver ligado, substituições ocorrerão de acordo com um modelo de substituição especificado.

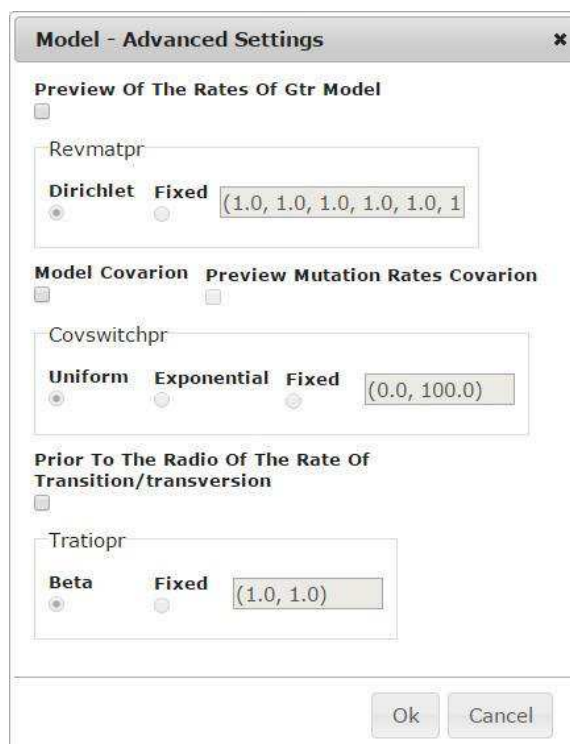


Figura 27 - Configurações avançadas do modelo do MrBayes para sequências de DNA.

Os modelos de proteínas implementados pelo MrBayes divide-se em duas categorias: os pré-fixados e os modelos de taxa variável. Dentre os quais, destacam-se (Figura 28):

- São diversos os modelos de taxas fixas implementados pelo MrBayes, sendo que cada um é apropriado para um tipo particular de proteínas. Os modelos são: Poisson, Dayhoff, Jones, Mtrev, Mtmam, WAG, Rtmrev, Ctrev, Vt e o Blosum62. Todos estes modelos são opções do campo “**Fixed Rate Models**”. O MrBayes define essa opção através do comando *prset aamodelpr=fixed(<modelo>)*. O modelo de Poisson é o mais simples, assumindo frequências de estados estacionários iguais e taxas de substituição também iguais. O restante dos modelos têm frequências de estados estacionário desiguais, mas fixas, e taxas de substituição que refletem as estimativas de evolução da proteína.
- O MrBayes permite uma maneira conveniente de se estimar o modelo de taxa fixa para os seus dados de aminoácidos, conhecido como modelo de salto entre os modelos de aminoácidos de taxa fixa. Este modelo caracteriza-se pela exploração de todos os modelos de taxa fixa listados acima, incluindo o modelo de Poisson, propondo regularmente novos modelos. Quando o procedimento MCMC convergir, cada modelo vai contribuir para os resultados na proporção de sua probabilidade posterior. Esta possibilidade é definida na interface através do campo “**Model Jumping Fixed-rate**”. O comando do MrBayes que define essa opção é o *prset aamodelpr=mixed*.

- O *Equalin* e o GTR são os dois modelos de taxa variável implementados pelo MrBayes para proteínas. O Equalin permite frequências de estados estacionários desiguais, mas assume a mesma taxa de substituição. O GTR permite que todas as frequências de estados estacionários e as taxas de substituição possam variar. Estes modelos podem ser selecionados através do campo “**Variable Rate Models**”, representando o comando *prset aamodelpr=fixed(<valores>)* do MrBayes.

## MrBayes

The screenshot shows the MrBayes Model configuration window. The 'Model' tab is selected, and within it, the 'Protein' sub-tab is active. The 'Matrix Rate' section shows 'Fixed Rate Models' with 'Poisson' selected and 'Variable Rate Models' with 'Equalin' selected. The 'Rate Variation' dropdown is set to 'equal'. The 'Gamma Categories' input field contains the number '4'. Below these, the 'Adgamma prior' and 'Propinv prior' sections each have 'Uniform' and 'Fixed' radio buttons, with the 'Fixed' option selected and its value set to '0'.

Figura 28 - Definição dos parâmetros do modelo do MrBayes para sequências de Proteína.

## MCMC

Esta seção apresenta a tela de configuração de parâmetros do MCMC, conforme observado na *Figura 29*. Os itens 3.4.2 e 4.4 apresentaram em detalhes a teoria por utilizada por traz destes parâmetros. A seguir serão listados os campos apresentados nesta tela:

- “**Number of cycles**”: representa a definição do número de ciclos para o algoritmo MCMC. O parâmetro do MrBayes relacionado com este campo é o *ngen* (número de gerações);
- “**Number of runs**”: determina a quantidade de análises independentes que serão iniciadas simultaneamente. Representa o parâmetro *nruns* no MrBayes;
- “**Number of chains**”: especifica quantas cadeias serão executadas para o MCMC. O valor padrão é 4, com 1 cadeia fria e 3 aquecidas. Representa o parâmetro *nchains* do MrBayes;
- “**Swap frequency**”: valor que determina a quantidade de tentativas de troca de estados da cadeia. O padrão é 1, o que resulta em uma análise para cada geração. Se o valor for 10, então “**Number of swaps**” serão analisados para cada décima geração de cadeias. Representa o parâmetro *swapfreq* no MrBayes;

- “**Number of swaps**”: número analisado por cada geração de troca de cadeia. Este campo representa o parâmetro *nswaps* no MrBayes;
- “**Sample frequency**”: determina quantas vezes a cadeia de Markov é amostrada. O parâmetro do MrBayes que este campo representa é o *samplefreq*;
- “**Diagn. Frequency (diagnfreq)**”: especifica o número de gerações do cálculo de diagnósticos do MCMC;
- “**Min. part. Frequency (minpartfreq)**”: a frequência mínima necessária para uma divisória ser incluída nos cálculos da convergência topológica de diagnóstico;
- “**Stop rule (stoprule)**”: se esta opção estiver desmarcada, a cadeia será executada o número de gerações determinado pelo valor de *ngen*. Se estiver marcada, e os diagnósticos de convergências topológicas estiverem calculados (*mcmcdiagn* definido como “sim”), a cadeia será interrompida antes do número pré-determinado de gerações, isso se as quedas de diagnósticos de convergências for abaixo do valor de parada;
- “**Critical value converfence (Stopval)**”: determina o valor para a convergência topológica de diagnóstico. Apenas usado quando *stoprule* e *mcmcdiagn* estão definidas para “sim” e mais de uma análise é executado simultaneamente (*nruns* > 1). No caso deste projeto, o valor do parâmetro *mcmcdiagn* está fixo para “sim”.

## MrBayes

The screenshot shows the 'MCMC' tab in the MrBayes software interface. It contains several input fields and a button for 'Advanced Settings'. The fields are organized into three main sections: 'Number Of Cycles', 'Number Of Runs', and 'Sample Frequency'. Below these are 'Chains' settings, 'Diagnostics' settings, and 'Stop Rule' settings.

Number Of Cycles	Number Of Runs	Sample Frequency
1000000	2	100

**Chains**

Number Of Chains	Swap Frequency	Number Of Swaps
4	1	1

**Diagnostics**

Diagnostics Frequency	Min Part Frequency
1000	0.1

**Stop Rule**

Stop Rule	Critical Value Convergence
<input type="checkbox"/>	0.01

Advanced Settings

Figura 29 - Campos da aba MCMC do MrBayes.

O botão **Advanced Settings** da Figura 29 abre a tela representada na Figura 30. Os campos apresentados nesta tela são:

- “**seed**”: define o número de sementes para o gerador de números aleatórios. O número aleatório é inicializado no início de cada sessão do MrBayes. Esta opção permite definir a semente para um valor específico, permitindo repetir exatamente a análise. Se a análise usa a

troca entre as cadeias de frios e quentes, deve-se definir a semente *swap* (campo a seguir) para repetir exatamente a análise;

- **“*swap seed*”**: define a semente usada para gerar a sequência de trocas quando correntes aquecidas do Metropolis acoplados são usados;
- **“*Save Branch Lenght(savebrlens)*”**: determina se as informações de comprimento dos ramos são armazenadas nas árvores;
- **“*Order taxa (ordertaxa)*”**: determina se a taxa deve ser ordenada antes das árvores serem impressas no arquivo. Se estiver marcada, os terminais nas árvores amostradas serão reordenadas para coincidir com o fim da taxa na matriz de dados o mais próximo possível. Por padrão, as árvores serão impressos sem reordenamento da taxa;
- **“*Reweight (Decrease, Increase e Increment)*”**: três números que representam, respectivamente, a porcentagem de caracteres para diminuir de peso, a porcentagem de caracteres para aumentar de peso e o incremento;
- **“*Proportional burning (Relburnin)*”**: se este campo estiver marcado, uma parte dos valores amostrados serão descartado no cálculo da convergência de diagnóstico. A proporção deste descarte é definido com o parâmetro *Burninfrac* representado pelo campo **“*Burnin Fraction*”**;
- **“*Specific burning (Relburnin)*”**: se este campo estiver marcado, um número específico de amostras devem ser descartadas. Este número é definido por *burnin* representado pelo campo **“*Burnin*”**;
- **“*Burnin Fraction (Burninfrac)*”**: determina a fração de amostras que serão descartadas quando os diagnósticos de convergências são calculados. O valor desta opção só é relevante quando *relburnin* está definido como *YES*;
- **“*Burnin*”**: determina qual parte da cadeia será descartada. Pode demorar um pouco para a cadeia alcançar a estacionariedade. Amostras retiradas quando a cadeia não é de estacionariedade (fase inicial da cadeia) deve ser descartada. Este parâmetro é uma maneira de simplesmente descartar amostras iniciais.



**MCMCMC - Advanced Settings** [X]

Seed: 1116335510

Swap Seed: 1116335510

Sabe Branch Lenght: ☒

Order Taxa: ☐

Reweight

Decrease(%): 0

Increase(%): 0

Increment: 1

Burnin

Proportional Burning: ☒

Burnin Fraction: 0.25

Specific Burnin: ☐

Burnin: 0

Ok Cancel

Figura 30 - Configurações avançadas do MCMC do MrBayes.

## Tree

Esta seção apresenta parâmetros referentes à definição da árvore de entrada (*Figura 31*). Esta pode ser randômica ou definida pelo usuário, através de um formato e um número randômico de perturbações específicas. A prioridade de topologia da árvore também pode ser definida.

A árvore de partida para a cadeia pode ser selecionada de forma aleatória ou definida pelo usuário. Escolhendo a opção “**Random Tree**” (parâmetro *startingtree* com a opção *random*; exemplo *startingtree=random*) a forma aleatória será utilizada. Já se a escolha for pela opção “**User Tree**” (parâmetro *startingtree* com a opção *user*; exemplo *startingtree=user*), a árvore de partida, que necessariamente precisa ser binária, deve ser definida pelo usuário através do campo “**Newick format tree**” (parâmetro *UserTree* com a árvore digitada pelo usuário; exemplo *UserTree=binary tree*). Ainda analisando a árvore de partida através da escolha do usuário, existe a possibilidade, através do campo “**Num. of random pertubations**”, de definir um número de perturbações aleatórias para ser aplicado nesta árvore. Este campo representa o parâmetro *nperts* no MrBayes.

Ainda nesta seção, pode-se definir a distribuição de probabilidade prévia para o comprimento dos ramos. Para o caso de se considerar o comprimento dos ramos anteriores para árvores sem restrição e sem raiz, a opção “**Non-Clock**” deve ser marcada, especificando um valor para a função exponencial através do campo “**Branch lengths**”. Já para o caso de árvores com restrição de relógio e enraizadas, pode-se utilizar as opções “**Simple Clock**”, “**Coalescence**” e “**Birth-death**” através do

campo “**Clock**”. Cada uma destas 3 opções possuem particularidades específicas que são listadas a seguir:

- “**Simple Clock**”: esta opção necessita de um valor para especificar a distribuição de probabilidade prévia sobre a altura da árvore. Este valor é representado pelo campo “**Tree Total Height**” e corresponde ao parâmetro *treeheightpr* no MrBayes;
- “**Coalescence**”: pode-se definir a prévia deste parâmetro através da opção *thetapr* do MrBayes, representado pelo campo “**Theta**”. Além disso, pode-se especificar também a ploidia do organismo (quantidade de lotes cromossômicos presentes em determinada célula) através do campo “**Ploidy level**” com as opções “**haploid**” (apenas um lote cromossômico) e “**diploid**” (dois lotes cromossômicos);
- “**Birth-death**”: as opções deste parâmetro são: a definição prévia sobre a taxa de especiação representada pelo campo “**Speciation**” (este campo corresponde ao parâmetro *speciationpr* do MrBayes); a definição prévia sobre a taxa de extinção representada pelo campo “**Extinction**” (campo correspondente ao parâmetro *extinctionpr* do MrBayes); e a definição da fração de espécies que serão amostrados na análise representada pelo campo “**Sampling probability**”, que corresponde ao *sampleprob* no MrBayes.

## MrBayes

Figura 31 - Configurações da aba Tree do MrBayes.

## Summarization

Após conclusão da análise do MCMC, um certo número de gerações, anteriores à convergência, deve ser descartado e as árvores e parâmetros devem ser sumarizados. Sendo assim, esta seção foi dividida nas abas parâmetros e árvores para definição das configurações de acordo à necessidade do usuário. Todos os campos que serão citados nesta seção estão ilustrados através das

imagens exibidas na *Figura 32* e na *Figura 33*. Consultar o item 4.4.4 para obter mais detalhes sobre esta seção.

A seguir serão listados os campos da aba parâmetros, são eles:

- O parâmetro *burnin* sumariza a amostragem posterior dos valores dos parâmetros, descartando as primeiras amostras, de acordo ao valor definido pelo usuário;
- Se o campo “**Likelihood Plot**”, que representa o parâmetro *plot*, estiver marcado, os valores das probabilidades deverão conter no arquivo de saída;
- O campo “**Marginal Model Likelihood**” corresponde ao parâmetro *marglike* do MrBayes, que determina se as estimativas de probabilidades dos modelo marginais devem ser calculadas;
- Se o usuário desejar que a saída seja uma tabela que contenha um resumo das amostras dos parâmetros, o campo “**Table**” deve ser marcado;
- Os resultados podem ser impressos em um arquivo definido pelo usuário. Esta definição pode ser feita através do campo *outputname*.

A outra aba desta seção é a “árvores”, que é responsável pela configuração da produção de estatísticas de resumo das árvores amostradas durante uma análise Bayesiana. Caso o usuário deseje descartar as estatísticas das árvores amostradas na parte inicial da análise, o campo *burning* deve ser preenchido. O usuário também pode escolher se as probabilidades das árvores devem ser calculadas, através do campo “**Calc. Tree Probabilites**” (parâmetro *calctreeprobs* do MrBayes). Além disso, pode-se escolher a probabilidade mínima das partições através do campo “**Min. Probability partitions**” (parâmetro *displaygeq* do MrBayes) e o tipo de árvore de consenso através das opções “**Halfcompat**” ou “**Allcompat**” (opções do parâmetro *contype* do MrBayes).

## MrBayes

The screenshot shows the MrBayes web interface with the 'Summarization' tab selected. Within this tab, the 'Parameters' sub-tab is active. The 'Summarize Parameters Samples' checkbox is unchecked. Under the 'Sump' section, the 'Burnin' value is set to 0. In the 'Output' section, three checkboxes are checked: 'Likelihood Plot', 'Marginal Model Like', and 'Table'.

Figura 32 - Configurações da aba Parameter (Sumamarization) do MrBayes.

## MrBayes

The screenshot shows the MrBayes web interface with the 'Summarization' tab selected. Within this tab, the 'Trees' sub-tab is active. The 'Samples Summarize Trees' checkbox is unchecked. Under the 'Sumt' section, the 'Burnin' value is set to 0. The 'Min Probability Partitions' value is also set to 0. The 'Calc Tree Probabilites' checkbox is checked. In the 'Consensus Tree' section, the 'Halfcompat' radio button is selected, and the 'Allcompat' radio button is unselected.

Figura 33 - Configurações da aba Trees (Sumamarization) do MrBayes.

### 5.2.2 PHYML

Assim como no MrBayes, os comandos do PHYML também foram divididos basicamente pela sua funcionalidade, sendo organizados em opções (campos) do site, dispostos em 4 abas, de nomes ***Sequence***, ***Model***, ***Tree*** e ***Bootstrap***, as quais serão analisadas a seguir. Cabe frisar, com o objetivo de esclarecimento, que as instruções do PHYML são precedidas por hífen (-) ou duplo hífen (--), e são organizadas de forma peculiar para execução do mesmo. Um exemplo será descrito abaixo para elucidar potenciais dúvidas.

## Sequence

Esta aba apresenta campos relevantes de definição inicial para análise filogenética (*Figura 34*). Nela encontra-se o campo para submissão do arquivo de sequência ou a utilização de um arquivo exemplo sugerido pela aplicação. Para este caso, este arquivo será de acordo à escolha do campo “**Data Type**” (Dna ou Aminoacid). Os parâmetros que representam os atributos citados são: *-i* ou *-input* para o “**File Sequence**” e *-d* ou *--datatype* para o “**Data Type**”.

Os campos possuem descrições intuitivas que informam a sua função na inferência filogenética em estudo. No entanto, a análise detalhada de cada campo é crucial para manter o padrão e a qualidade do software apresentado. A seguir serão descritos os outros parâmetros pertencentes a esta aba, são eles:

- “**Sequence file**”: representado pelo comando *-q* ou *-sequential*, identifica os tipos de sequência do arquivo: *interleaved* ou *sequential*;
- “**Number Of Data Sets**”: corresponde ao número de conjunto de dados a analisar. É definido pelo comando *-n* ou *-multiple*;
- “**Initiate The Random Number Generator**”: é a semente utilizada para iniciar o gerador de números aleatórios. O comando que o representa é o *--r\_seed*.

O PHYML retorna ao final de sua execução uma série de arquivos de textos contendo resultados descritivos de seu processamento. Dentre estes, estão os que listam as probabilidades para cada sítio e os que exibem cada filogenia explorada durante o processo de pesquisa da árvore. Esses dois arquivos só são criados através da marcação das opções “**Print The Likelihood**” e “**Print Each Phylogeny Explored**”, representados, respectivamente, pelos comandos *--print\_site\_lnl* e *--print\_trace*.

## Phyml

The screenshot shows the 'Sequence' tab of the PhyML web interface. It features a tabbed menu at the top with 'Sequence', 'Model', 'Tree', and 'Bootstrap'. The 'Sequence' tab is active and contains several configuration options:

- Sequence:** Includes radio buttons for 'Example File' and 'File'. The 'File' option is selected, with a text input field containing 'Escolher arquivo' and a status message 'Nenhum arquivo selecionado'.
- Data Type:** Includes radio buttons for 'Dna' and 'Aminoacid'. The 'Dna' option is selected.
- Sequence file:** Includes radio buttons for 'Interleaved' and 'Sequential'. The 'Interleaved' option is selected.
- Number Of Data Sets:** A text input field containing the value '1'.
- Initiate The Random Number Generator:** A text input field containing the value '3'.
- Print Likelihood:** An unchecked checkbox.
- Print Each Phylogeny Explored:** An unchecked checkbox.

Figura 34 - Configurações da aba Sequence do PHYML.

### Model

Como visto na aba *Sequence*, os tipos de dados de sequências suportados pelo PHYML são os de Dna e Proteína. Dessa forma, por convenção, a aba *Model* foi subdividida nos tipos citados, recebendo portanto, seus parâmetros correspondentes, sendo a aba DNA definida como a opção padrão (*Figura 35 e Figura 36*).

Ainda neste contexto, fazendo um paralelo rápido entre os parâmetros de cada aba, fica notório a diferença apenas das opções dos modelos de substituição, representado pelo campo “***Substitution Model***”, e da taxa de transição/transversão, que equivale ao campo “***Transition/Transversion Ratio***”. Todos os outros campos são comuns às duas abas.

Sendo assim, os modelos de substituição disponibilizados pela aba DNA são: HKY85, K80, JC69, F81, F84, TN93 e GTR (consultar o item 2.3 para obter mais detalhes sobre estes modelos). Como dito anteriormente, outro campo que é utilizado somente por esta aba, não sendo possível utilizá-lo para os modelos JC69 e F81, é o “***Transition/Transversion Ratio***”, que representa o comando *-t* ou *--ts/tv*. Para o caso da utilização do modelo K80, este campo representa o item *r* conforme explicado na seção 2.3.2.

Em contrapartida, para aba Protein, os modelos apresentados são: LG, WAG, Dayhoff, mtREV, JTT, DCMut, RtREV, CpREV, TV, Blosom62 e MtMam.

Vale ainda ressaltar, que os modelos de substituição são representados pelo comando *-m* ou *-model*, sendo o modelo HKY85 definido como a opção padrão para DNA, e o LG para proteína.

Os outros parâmetros restantes da aba *Model*, comum tanto para DNA como para Proteína, são:

- **“Number Of Substitution Rate Categories”**: representa o comando *-c* ou *-nclasses*, que informa a categoria de taxas de substituição. Este campo caracteriza-se pela divisão dos sítios da sequência em *C* categorias, conforme descrito no 2.4.1;
- **“Proportion of invariable sites”**: equivale ao comando *-v* ou *-pin*, o qual identifica a proporção de sítios invariáveis. Este campo determina a aplicação de uma porcentagem fixa e outra variante, como discutido no item 2.4.1;
- **“Equilibrium Frequencies”**: determina as frequências de caracteres através das opções *Empirical*, *Estimated* e *Equal*. Este campo é representado pelo comando *-f* com as opções *e*, *d*, ou *fA fC fG fT* de acordo à ordem citada;
- **“Gamma shape parameter”**: campo que equivale a forma do parâmetro gama, representado pelo comando *-a* ou *-alpha*.

## Phyml

The screenshot shows the PhyML software interface. At the top, there are four tabs: "Sequence", "Model", "Tree", and "Bootstrap". The "Model" tab is selected. Below it, there are two sub-tabs: "Protein" and "DNA". The "DNA" sub-tab is active. The "DNA" sub-tab contains several configuration options:

- Substitution Model**: A dropdown menu showing "HKY85".
- Number Of Substitution Rate Categories**: A text input field containing the number "4".
- Equilibrium Frequencies**: A dropdown menu showing "Empirical".
- Transition/Transversion Ratio**: Two radio buttons, "Estimated" (selected) and "Fixed".
- Proportion of invariable sites**: Two radio buttons, "Estimated" and "Fixed" (selected). A text input field next to "Fixed" contains the number "0".
- Gamma shape parameter**: Two radio buttons, "Estimated" (selected) and "Fixed". A text input field next to "Fixed" contains the number "0".

Figura 35 - Configurações da aba DNA (Model) do PHYML.

The image shows the 'PhymI' software interface. At the top, there are four tabs: 'Sequence', 'Model', 'Tree', and 'Bootstrap'. The 'Model' tab is selected. Below this, there are two sub-tabs: 'Protein' and 'DNA', with 'Protein' selected. The main configuration area for the 'Protein' model includes:

- Substitution Model:** A dropdown menu set to 'WAG'.
- Number Of Substitution Rate Categories:** A text input field containing the number '4'.
- Equilibrium Frequencies:** A dropdown menu set to 'Empirical'.
- Proportion of invariable sites:** A section with two radio buttons, 'Estimated' (unselected) and 'Fixed' (selected). Next to the 'Fixed' button is a text input field containing '0'.
- Gamma shape parameter:** A section with two radio buttons, 'Estimated' (unselected) and 'Fixed' (selected). Next to the 'Fixed' button is a text input field containing '0'.

Figura 36 - Configurações da aba Protein (Model) do PHYML.

### Tree

Esta seção apresenta as opções do PHYML para otimizar a árvore resultado da análise filogenética (Figura 37). Sendo assim, uma das alternativas é representada pelo campo “**Starting Tree**”, que equivale ao comando *-u* ou *-inputtree*, o qual é utilizado como árvore de partida pelo algoritmo de MV. Por padrão, utiliza-se uma árvore de acordo aos princípios do programa BIONJ, o qual é baseado no método de Distância, mas pode-se utilizar árvores de partidas aleatórias, incrementada pelas opções SPR e NNI (consultar item 3.2.1). É possível também fornecer um arquivo texto contendo uma ou mais árvores, uma por linha, em formato Newick (consultar item 2).

Outra opção apresentada nesta aba são os tipos de melhorias de busca da topologia, que poderão ser aplicados no ato de reconstrução da árvore. O campo do site que representa esta opção é o “**Type Of Tree Improvement**”, o qual exerce a função do comando *-s* ou *-search*. Esta operação de busca da topologia da árvore pode ser do tipo NNI ou SPR, ou ambos. O NNI é a opção definida como padrão, sendo esta é mais rápida do que a SPR.

Nesse sentido, caso a opção de busca da topologia seja a SPR, o PHYML implementa uma opção para definir a árvore inicial como aleatória (comando *--rand\_start*), bem como a quantidade utilizada. Esta opção é representada no site pelo campo “**Number Of Random Starting Tree**”, que representa o comando *--n\_rand\_starts* seguido do valor digitado.

Caso o usuário não escolha a opção **File** do campo “**Starting Tree**”, existe a possibilidade de utilização de uma árvore mínima de parcimônia, através do campo “**Minim Parsimony Starting Tree**”, que representa o comando *-p* ou *-pars*.



O PHYML se encarrega de otimizar as opções de parâmetros específicos, de acordo as marcações dos campos “*Optimise Branch Lengths*”, “*Rate Parameters Optimization*” e “*Optimise Topology*”, os quais acrescentam opções no comando `-o` da seguinte forma:

- Todos os 3 marcados: o comando `-o` é acrescido das opções `tlr`;
- Opções “*Optimise Branch Lengths*” e “*Optimise Topology*” marcadas: opção `tl` é acrescentada;
- Opções “*Optimise Branch Lengths*” e “*Rate Parameters Optimization*” marcadas: `lr` é acrescentada;
- Apenas a opção “*Optimise Branch Lengths*” marcada: `l` é acrescentada ao comando `-o`;
- Apenas a opção “*Rate Parameters Optimization*” marcada: `r` é acrescentada ao comando `-o`;
- Nenhuma opção marcada: `n` é acrescentada ao comando `-o`.

## Phyml

Figura 37 - Configurações da aba Tree do PHYML.

## Bootsrap

A última aba do IgrafuWeb para o PHYML é a Bootstrap (Figura 38). Esta apresenta dois campos que referenciam o mesmo comando: `-b` ou `-bootstrap`. As opções para esses comandos são fixas, exceto para a representada pelo campo “*Enter The Number Of Bootstrap To Be Used*”, que obrigatoriamente tem que possuir valor maior do que zero, informando o número de repetições do bootstrap. O item 3.3.1 apresenta mais detalhes sobre o bootstrap. As outras opções são representadas pelo campo “*Approximate Likelihood Ratio Test*” com valores:

- **“Neither approximate nor bootstrap values are computed”**: esta opção equivale ao valor “0” do comando citado, e indica que não serão computados testes aproximados de verossimilhança, muito menos de bootstrap;
- **“aLRT statistic”**: esta opção aplica um teste de verossimilhança retornando estatísticas aLRT (teste estatístico para computar o suporte dos ramos). Representa a opção de valor “-1”;
- **“Chi2-based parametric branch support”**: opção que equivale o valor “-2”. Representa um teste aproximando de verossimilhança retornando suporte dos ramos parametrizados por Chi2-based;
- **“SH-like branch supports alone”**: opção que representa o valor “-4”. Esta opção indica que o ramo suporta somente Sh-like.

### Phyml

Figura 38 - Configurações da aba Bootstrap do PHYML.

#### 5.2.3 Digrafu

Do mesmo modo do MrBayes e do PHYML, os comandos do Digrafu também foram divididos de acordo à sua função, sendo apresentados em opções (campos) do site, disponibilizados em 4 abas de nomes: **Sequence**, **Model**, **Bootstrap – Seqboot** e **Bootstrap – Consense**. Todos estes campos possuem nomenclaturas intuitivas, de forma a passar uma ideia de sua função na análise filogenética em questão. Sendo assim, as análises a seguir demonstram as características principais do Digrafu, bem como a sua sintaxe particular, que detém palavras chaves, seguidas dos valores de fato da escolha do usuário.

#### Sequence

Esta aba inicial do Digrafu apresenta campos cruciais para inferência filogenética, como observado na *Figura 39*. Nela encontram-se os campos para definição do tipo de arquivo de

sequência, “**Data Type**”, e da utilização do bootstrap, “**Want to use bootstrap**”. O primeiro campo citado representa o comando TYPE, enquanto o segundo indica a utilização de bootstrap (esta opção será descrita mais abaixo). Além disso, existe o campo para inserção do arquivo de sequência ou a utilização de um exemplo sugerido pelo IgrafuWeb, de acordo a escolha do campo “**Data Type**”: Dna ou Protein. Este campo é o “**Sequence**”, que disponibiliza as opções “**Example File**” e “**File**”, o qual representa o comando INPUT.

Além disso, esta aba ainda possui o campo que vai definir a escolha de um dos algoritmos de distância citados no item 3.1. Este está sendo representado pela identificação “**Time**”, com as possibilidades “**Execution**” e “**Accuracy**”. Quando o usuário seleciona “**Execution**”, o Digrafu executa o BioNJ, que é o mais eficiente em quase todos os casos. Quando o usuário escolhe ambas as opções, a execução ocorrerá de acordo ao método que proporciona maior exatidão, exceto nos casos nos quais o método de maior exatidão seja o Weighbor (seu tempo de execução é muito alto comparado aos outros métodos), nesses casos será usado o método FastME. Se a escolha do usuário for “**Accuracy**”, o DiGrafu escolhe o método com maior exatidão (TORRES et al., 2011).

## Digrafu

The screenshot shows the 'Digrafu' web interface with the 'Sequence' tab selected. The interface includes several configuration options:

- Sequence**: A section with two radio buttons, 'Example File' (selected) and 'File'. A button 'Selecionar arquivo...' is next to the 'File' option, and the text 'Nenhum arquivo selecionado.' is displayed.
- Data Type**: A dropdown menu currently set to 'DNA'.
- Want to use bootstrap?**: Two radio buttons, 'Yes' and 'No' (selected).
- Time**: Two checkboxes, 'Execution' and 'Accuracy'.

At the top of the interface, there are four tabs: 'Sequence', 'Model', 'Bootstrap - Seqboot', and 'Bootstrap - Consense'.

Figura 39 - Configurações da aba Sequence do Digrafu.

## Model

A seção Model, assim como no MrBayes e no PHYML, apresenta duas sub abas de acordo aos tipos de sequências suportados: DNA ou Proteína (Figura 40 e Figura 41, respectivamente). Cada uma delas possui peculiaridades, porém existem campos em comum, como é o caso dos campos:

- “**Substitution Model**”: representa o comando MODEL e disponibiliza as opções F84, JC69, Kimura ou LogDet para Dna, e as opções “**Jones-Taylor-Thornton matrix**”,

“*Henikoff/Tillier PMB matrix*”, “*Dayhoff PAM matrix*” ou “*Kimura formula*” para proteína. A teoria utilizada pelos modelos de evolução de nucleotídeos pode ser consultada no item 2.3;

- “***Gamma Distribution***”: representa a distribuição gama (item 2.4.2), não podendo ser utilizado para as sequências de DNA e proteína, sob o modelo LogDet e Kimura, respectivamente. O comando representado por este campo é o GAMMA;
- “***Weights For Sites***”: campo que aplica pesos sobre os sítios ou os índices dos sítios que serão considerados nas sequências em questão. Representa o comando WEIGHT.

Além disso, existem outros campos que só são disponíveis para sequências de DNA, são eles:

- “***Transition/ Transversion Ratio***”: campo que representa a taxa de transição/transversão, o qual equivale ao comando RATIO. Não aplicado para sequências de DNA utilizando os modelos JC e LogDet. Para o caso da utilização do modelo K80, este campo representa o item  $r$  conforme explicado na seção 2.3.2;
- “***Fraction Of Invariantes Sites***”: campo que define a porcentagem de sítios invariáveis, representando o comando ISITE. Este campo determina a aplicação de uma porcentagem fixa e outra variante, como discutido no item 2.4.1;
- “***Empirical Frequencies***”: campo que define as frequências de bases empíricas, equivalente ao comando FREQUE. Caso a escolha seja pela não utilização de bases, o comando terá a opção “n”, caso contrário terá valores para todos os nucleotídeos, A, C, G, T. Não é aplicado para sequências de DNA utilizando os modelos Kimura, JC e LogDet.

## Digrafu

The screenshot shows the Digrafu software interface with the following configuration for the DNA (Model) tab:

- Sequence** | **Model** | Bootstrap - Seqboot | Bootstrap - Consense
- Protein** | **DNA**
- Substitution Model**: F84
- Gamma Distribution**: 1
- Transition/ Transversion Ratio**: 1
- Empirical Frequencies**: ☒ 4 4 4 4
- Weights For Sites**: ☒ 0
- Fraction Of Invariantes Sites**: 0

Figura 40 - Configurações da aba DNA (Model) do Digrafu.

## Digrafu

The screenshot shows the Digrafu software interface with the 'Protein' tab selected under the 'Model' category. The 'Substitution Model' is set to 'Jones-Taylor-Thornton matrix'. The 'Gamma Distribution' is set to '0'. The 'Weights For Sites' checkbox is checked, and the value is set to '0'.

Figura 41 - Configurações da aba Protein (Model) do Digrafu.

### Bootstrap-Seqboot

Esta seção apresenta o conjunto de dados da ferramenta Seqboot, utilizada para realizar bootstrap, representada através da *Figura 42* e da *Figura 43*. De acordo a análise de cada campo, ficou definido a inserção de abas alternativas para facilitar o entendimento e usabilidade do software, são elas: *Sequence* e *Parameters*.

### Aba Sequence

Esta aba apresenta as características iniciais para a execução do Seqboot, dentre as quais pode-se citar:

- **“Data Type”**: indica o tipo de sequência avaliada. Possui as opções DNA, RNA e PRO (representa proteína);
- **“Sequence Type”**: indica o tipo da sequência. As opções disponíveis são **“Molecular sequences”** (SEQU), **“Discrete Morphology”** (MORF), **“Restriction Sites”** (REST) e **“Gene Frequencies”** (FREQ);
- **“Model Permutation”**: indica o modelo de permutação utilizado, pode assumir uma das opções: **“BOOTSTRAP”** (boot), **“JACKKNIFE”** (jack), **“PERMUTE CHARACTER”** (perm), **“PERMUTE CHARACTER ORDER”** (pord), **“PERMUTE WHITH SPECIES”** (pspec). Consultar os itens 3.3.1 e 3.3.2 para obter mais detalhes sobre essas opções;
- **“Format”**: define o formato do arquivo de entrada, tendo **“INTERLEAVED”** (i) ou **“SEQUENCIAL”** (s) como opções;

- **“Categories Of Sites”**: arquivo de informações extras sobre as categorias de sítios, fatores ou alelos. Assume dois parâmetros: o primeiro a palavra chave ALL, CAT ou FAC, e o segundo o diretório seguido do nome do arquivo;
- **“Seed”**: campo que define o valor da semente, aceitando somente valores ímpares;
- **“Replicates”**: campo que representa a quantidade de réplicas;
- **“Block”**: tamanho do bloco que divide a sequência. Representa o comando BLOCO, o qual recebe como opção o tamanho que será dividido as sequências;
- **“Samples”**: campo que indica a porcentagem das amostras. Representa o comando FRACAO seguido do valor digitado;
- **“Enzyme”**: campo que informa a utilização de enzimas no arquivo de entrada. A marcação deste campo resulta na inserção do comando ENZIMA.

### **Aba Parameters**

Esta aba apresenta os campos:

- **“Weights”**: arquivo que contém os pesos dos caracteres. Representa o comando WEIGHT, seguido do caminho e nome do arquivo selecionado;
- **“Mixture Files”**: representa o comando MIX seguido do caminho e nome do arquivo selecionado;
- **“Ancestral Files”**: representa o comando ANC seguido do caminho e nome do arquivo selecionado;
- **“Exit Format”**: campo que define o formato de saída. Representa o comando RESC, seguido das palavras PHY, indicando o formato PHYLIP, NEX, formato nexus ou XML, formato xml;
- **“Multiples Exit Files”**: campo que define os arquivos múltiplos de saída, através das opções **“Data”** ou **“Weights”**, representando, respectivamente, os comandos: JUSTW e OUTD.

## Digrafu

Figura 42 - Configurações da aba Sequence (Bootstrap-Seqboot) do Digrafu.

## Digrafu

Figura 43 - Configurações da aba Parameters (Bootstrap-Seqboot) do Digrafu.

### Bootstrap-Consense

Os campos disponibilizados nesta seção são (Figura 44):

- **“Tree”**: campo que oferece a possibilidade de inserção de um arquivo para receber a árvore resultado da análise filogenética. Esta opção representa o comando FILE, seguido do caminho e nome do arquivo;
- **“Root”**: campo que define se a árvore será tratada com ou sem raiz. Representa o comando “R” seguido da opção “S”, de sim, ou “N”, de não;
- **“Evaluation”**: campo que define o tipo de avaliação. As opções são: **“EXTENDED MAJORITY RULE”**, **“STRICT”**, **“MAJORITY RULE”** e **“ML”**. A escolha de uma dessas opções resulta na inserção de um dos comando a seguir, seguindo a ordem de apresentação das opções: MRE, STR, MR ou ML;



- **“Ancestral Higher”**: opção para definir o nó raiz (maior ancestral). Representa o comando ROOT, seguido da indicação da posição da espécie no arquivo em questão considerando a ordem. Como exemplo, pode-se citar que o valor 3 representa a terceira espécie;
- **“Species In Exit File (incate)”**: campo que indica as espécies serão escritas no arquivo de saída. Caso esta opção esteja marcada, o comando PRINT será inserido;
- **“Execution (incate)”**: campo que define a execução. A marcação deste campo resulta na inserção do comando RUN;
- **“Tree In Exit File (draw)”**: campo que informa se a árvore resultado será desenhada no arquivo de saída. Esta opção implica na inserção do comando TREE;
- **“Time Fraction”**: opção disponibilizada apenas para os casos que o campo **“Evaluation”** tenha valor “ML”. Indica a fração de vezes que o ramo deve ser analisado. Representa o comando FRACTION, seguido do valor digitado.

## Digrafu

The screenshot shows the 'Bootstrap - Consense' tab of the Digrafu software interface. It contains several configuration options:

- Tree**: A button labeled 'Escolher arquivo' with the text 'Nenhum arquivo selecionado' next to it.
- Root**: Two radio buttons, 'Yes' (selected) and 'No'.
- Evaluation**: A dropdown menu currently set to 'EXTENDED MAJORITY RULY'.
- Ancestral Higher**: A text input field containing the value '0'.
- Species In Exit File(incate)**: A checkbox that is currently unchecked.
- Execution(incate)**: A checkbox that is currently unchecked.
- Tree In Exit File(draw)**: A checkbox that is currently unchecked.
- Time Fraction**: A text input field containing the value '0.5'.
- Use**: A checkbox that is currently unchecked.

Figura 44 - Configurações da aba Bootstrap-Consense do Digrafu.

### 5.2.4 DNAPARS e PROTPARS

Os parâmetros utilizados pelo DNAPARS e pelo PROTPARS são quase todos iguais, o que resultou em apenas uma interface gráfica, suficientemente organizada para abrigar ambos os softwares, fazendo o devido tratamento para algumas diferenças pontuais. Neste sentido, cabe ressaltar uma peculiaridade relevante destes softwares, que é o fato de utilizar apenas letras ou números, de maneira simbólica, para identificar os seus comandos de configuração. Como exemplo, pode-se citar a utilização da letra “i” maiúscula (I) para representar a utilização de sequências genéticas intercaladas. Sendo assim, as explicações a seguir dão ênfase à fase de desenvolvimento, explicando todas as características e divisão das configurações destes dois



softwares no IgrafuWeb. É relevante mencionar, a utilização de nomenclaturas intuitivas para tornar mais claro a identificação de cada tipo de parâmetro na interface, de acordo aos requisitos definidos no manual dos softwares em questão.

De início, assim como nos outros softwares analisados, ficou evidente a necessidade de acoplar todos os parâmetros em abas, com o objetivo claro de manter o padrão e a qualidade apresentados. Logo, foram criadas 4 abas (serão analisadas em detalhes a seguir), são elas: ***Sequence, Tree, Bootstrap e Options.***

### **Sequence**

Esta aba (*Figura 45*) apresenta os campos essenciais para iniciar a RAF utilizando o método de máxima parcimônia, através dos softwares DNAPARS ou PROTPARS tendo o IgrafuWeb como interface gráfica. É nesta aba que acontece a definição de qual tipo de sequência será analisado: DNA ou Proteína. De acordo a esta definição, que é feita através do campo “***Type Sequence***”, utiliza-se o DNAPARS ou PROTPARS. Além disso, os campos “***Use Transversion Parsimony***” e “***Use Which Genetic Code***” são habilitados ou não de acordo à definição deste tipo de sequência. Esses campos representam, respectivamente, os comandos “***T***” e “***C***” dos softwares apresentados, de acordo à ordem de citação.

Além do mais, esta aba disponibiliza a submissão do arquivo de sequência a ser analisado. Assim como foi feito para as outras implementações dos métodos de RAF, pode-se utilizar um arquivo exemplo ofertado pela aplicação de acordo ao tipo de sequência que se deseja trabalhar. Esta opção é representada pelo campo “***Sequence***”, fazendo a devida escolha de um arquivo exemplo (opção “***Example File***”) ou a submissão de outro arquivo de sequência qualquer (opção “***File***”).

Neste contexto, de acordo a análise da *Figura 41*, os campos que faltam ser mencionados são: “***Format Sequence File***” e “***Weight Options***”. Logo, percebe-se pela análise do primeiro, que os formatos de sequências suportados são o intercalado ou o sequencial, os quais são representados, respectivamente, pelas opções “***Interleaved***” e “***Sequential***”. Para este caso, apenas a opção “***Interleaved***” representa o parâmetro “***I***”. Caso a escolha seja “***Sequential***”, nenhum parâmetro é adicionado. Seguindo a mesma linha raciocínio, de maneira a analisar o parâmetro “***Weight Options***”, constata-se a possibilidade de submissão de um arquivo de pesos de acordo a escolha das opções “***No***” ou “***Yes***”. Caso deseja-se utilizar a “***Yes***”, o parâmetro “***W***” será utilizado juntamente com o arquivo submetido.

## DnaPars or ProtPars (PHYLIP)

Figura 45 - Configurações da aba Sequence do DNAPARS ou PROTPARS.

### Tree

Esta aba, representada pela *Figura 46*, possibilita a definição de alguns parâmetros que auxiliam a busca pela melhor árvore. Esta busca pode ser feita com ou sem o auxílio de um arquivo, dependendo da marcação da opção “*No*” ou “*Yes*” do campo “*Search for best tree*”.

Optando pela opção “*No*”, o usuário tem a possibilidade de definir um arquivo de texto contendo informações das árvores, de maneira a auxiliar a pesquisa pela melhor resposta que represente a filogenia em questão. Ainda neste contexto, pode-se definir a quantidade de árvores contidas neste arquivo através do campo “*Number of trees in the file*”.

A utilização da opção “*Yes*” possibilita a definição do campo “*Search Option*”, o qual oferece um mecanismo diferente para busca da melhor árvore, através das opções: “*More thorough search*”, “*Less thorough search*” ou “*Rearrange on one best tree*”. Destas, apenas as duas últimas determinam a inserção de parâmetro, são eles, respectivamente: “*Y*” ou “*N*”. Estes dois parâmetros precisam ser precedidos pelo parâmetro “*S*”, o qual informa a utilização de um controle de busca (campo “*Search Option*”). A opção padrão é a “*More thorough search*”, a qual não exige a utilização parâmetro. Vale ressaltar, que estas configurações só estão disponíveis para sequências de proteínas.

Além disso, existe a possibilidade de salvar algumas árvores através do campo “*Number Of Trees To Save*”, o qual determina a inserção do parâmetro “*V*” juntamente com o valor numérico digitado neste campo.

## DnaPars or ProtPars (PHYLIP)

The screenshot shows the 'Tree' tab of the DnaPars or ProtPars (PHYLIP) interface. It features four tabs at the top: 'Sequence', 'Tree' (selected), 'Bootstrap', and 'Options'. The main area is divided into two sections. The top section, titled 'Search for best tree?', contains radio buttons for 'Yes' (selected) and 'No', a 'Selecionar arquivo...' button, the text 'Nenhum arquivo selecionado.', and a 'Number Of Trees In The File' input field with the value '1'. The bottom section, titled 'Pars Options', contains a 'Search Option' dropdown menu set to 'More thorough search' and a 'Number Of Trees To Save?' input field with the value '10000'.

Figura 46 - Configurações da aba Tree do DNAPARS ou PROTPARS.

### Bootstrap

Os comandos de Bootstrap (item 3.3.1) ofertados pelos programas DNAPARS e PROTPARS foram basicamente divididos em 2 campos: “*Analyze multiple data sets*” e “*Randomize input order of sequences*”.

O primeiro apresenta as opções “*No*”, “*Yes, multiple data sets*” ou “*Yes, multiple sets of weight*”. Caso deseje-se utilizar uma das opções “*Yes*”, o parâmetro “*M*” é inserido, além da definição da quantidade de conjuntos, atribuído através do campo “*Number Of Sets*”. Este campo representa um valor que será inserido na linha seguinte ao parâmetro “*M*”.

O segundo campo disponibiliza as opções “*No, use input order*” ou “*Yes*”. Optando pela “*Yes*”, o parâmetro “*J*” será inserido, além de possibilitar a definição dos campos “*Random number seed*” e “*Number of times to jumble*”. Estes campos são numéricos, e seus respectivos valores serão inseridos logo após a definição do parâmetro “*J*”.

A opção “*No*”, comum aos 2 campos em análise, não representa nenhum parâmetro de configuração, e define a utilização do software no formato padrão. Todos estes campos estão devidamente organizados e distribuídos na aba Bootstrap, conforme exibe a *Figura 47*.

## DnaPars or ProtPars (PHYLIP)

Figura 47 - Configurações da aba Bootstrap do DNAPARS ou PROTPARS.

### Options

Esta aba, representada pela *Figura 48*, caracteriza-se basicamente pelas configurações dos arquivos de saída. Dentre as opções, tem-se:

- Impressão dos dados no início da execução. Esta opção representa a inserção do comando “*I*”, e está disponível na interface através do campo “***Print Out The Data At Start Of Run***”;
- Impressão de uma imagem considerada gráfica no arquivo de saída. O campo que representa esta opção é o “***Print Out Tree***”, indicando a inserção do comando “*3*”;
- Impressão de etapas da execução para cada sítio. O comando que representa esta opção é o “*4*”, sendo disponibilizado na interface através do campo “***Print Out Steps In Each Site***”;
- Impressão dos caracteres em todos os nós da árvore. Representado na interface pelo campo “***Print Character At All Nodes Of Tree***”, indicando a utilização do comando “*5*”;
- Salvar a árvore resultado da execução em arquivo texto. Opção que representa o comando “*6*”, referenciada através do campo “***Write Out Trees Onto Tree File***”.

Esta aba também possibilita a definição de um grupo externo, através do campo “***Outgroup Root?***”, o qual disponibiliza as opções “*No, use as outgroup species 1*” e “*Yes*”. Uma vez que a segunda opção foi escolhida, abre-se a possibilidade de digitação do valor deste grupo, através do campo “***Outgroup Value***”. Estes campos representam o comando “*O*”, seguido do valor digitado.

Por fim, esta aba possibilita a definição de limiares para parcimônia. O campo para isso é o “***Use Threshold?***”, marcado com a opção “*Yes*”. Sendo assim, o campo “***Threshold Value***” é habilitado, possibilitando a inserção de um valor para este caso. Estas opções representam o comando “*T*”, seguido do valor digitado.

## DnaPars or ProtPars (PHYLIP)

The screenshot shows the 'Options' tab of the DnaPars or ProtPars (PHYLIP) interface. The interface is divided into four tabs: 'Sequence', 'Tree', 'Bootstrap', and 'Options'. The 'Options' tab is selected. It contains two main sections: 'Output' and 'Parsimony'. The 'Output' section has five options: 'Print Out The Data At Start Of Run' (No), 'Print Out Tree' (Yes), 'Print Out Steps In Each Site' (No), 'Print Character At All Nodes Of Tree' (No), and 'Write Out Trees Onto Tree File?' (Yes). The 'Parsimony' section has two options: 'Use Threshold?' (No) and 'Threshold Value' (0). The 'Other' section has two options: 'Outgroup Root?' (No, use as outgroup species 1) and 'Outgroup Value' (1).

Figura 48 - Configurações da aba Options do DNAPARS ou PROTPARS.

### 5.2.5 Execução

Após a escolha do método e dos seus respectivos parâmetros, assim como o arquivo de sequência, pode-se realizar a RAF clicando no botão *Execute*. A partir daí o sistema dispara uma série de rotinas cruciais para a conclusão do processo, são elas (Figura 49):

- Montagem de um arquivo texto, de acordo à sintaxe do programa escolhido, que servirá de base para execução do método escolhido no CACAU;
- Montagem do arquivo texto do Slurm, que levará em conta também o software escolhido, assim como as quantidades de sequência e de sítios envolvidos na análise filogenética em questão;
- Transferência dos dois arquivos citados acima, assim como do arquivo de sequência escolhido pelo usuário. Os detalhes desta transferência serão explicados abaixo;
- Execução de fato do programa escolhido. Esta execução é realizada no CACAU através do gerenciador de fila Slurm.

O CACAU detém uma estrutura organizacional de tal forma que o servidor (nó) que hospeda os sites se mantém em uma rede distinta da rede dos “nós” que realizam de fato o processamento da análise filogenética. Sendo assim, foi necessário construir um procedimento de troca de arquivos entre o servidor Web e o servidor que controla as execuções dos procedimentos executados no CACAU.

Neste contexto, assim como os arquivos são transferidos do servidor Web para o nó central de execuções, o procedimento inverso também é realizado após a conclusão do processo submetido. Porém, os arquivos que são submetidos a esta transferência, são os gerados pelos softwares

escolhidos pelo usuário, contendo o resultado final na análise filogenética. Esta transferência dos arquivos com o resultado obtido foi necessário para o perfeito funcionamento da visualização da árvore no próprio site, visto que esta é realizada através de um Applet, como descrito no capítulo 5. A Figura 49 ilustra as transferências de arquivos citadas.

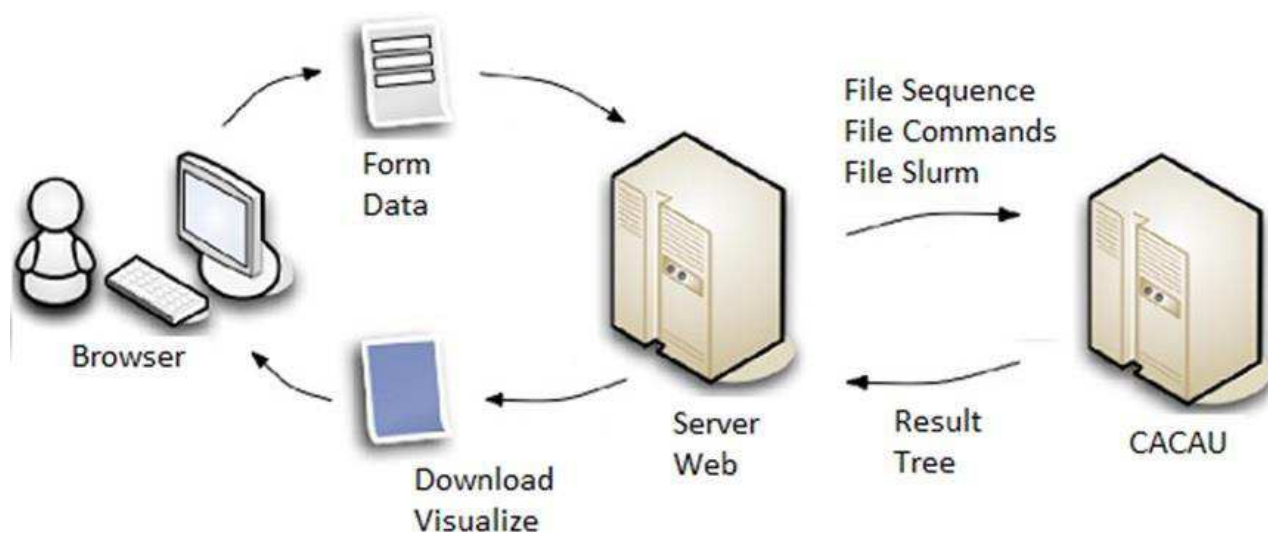


Figura 49 - Diagrama de Fluxo da execução do IgrafuWeb.

Ainda sob a ótica da análise do CACAU, faz-se necessário abordar as questões do usuário e do acesso para execução. Para realizar qualquer submissão de trabalhos no CACAU é necessário autenticação, ou seja, usuário e senha, além de ter que acessar remotamente, neste caso utilizando o SSH<sup>14</sup> (Secure Shell). Como este trabalho ainda não exige autenticação, foi criado um perfil específico para esse fim. Além do mais, foi necessário a elaboração de um mecanismo de acesso remoto para cada execução realizada pelo IgrafuWeb. Isto só foi possível graças a uma biblioteca de acesso remoto, disponibilizada pelo PHP.

Uma vez que os arquivos com o resultado final da inferência filogenética estejam disponíveis no servidor Web, em um diretório específico de uma determinada execução, os botões “*Visualize*” e “*Download*” (Figura 23) são habilitados.

O clique do botão “*Visualize*” inicia a execução do Applet Archaeopteryx, o qual recebe como parâmetro o(s) arquivo(s) contendo a(s) árvore(s) filogenética(s) em análise, fazendo com que esta(s) fique(m) disponível(is) em formato gráfico para visualização do usuário.

<sup>14</sup> Protocolo de rede que permite a conexão com outro computador.

O botão “*Download*” é responsável por disponibilizar a visualização dos arquivos de entrada (sequência, Slurm, parâmetros do software escolhido) e saída (árvore(s) filogenética(s), estatística e bootstrap dependendo da escolha do usuário) da inferência filogenética.

O software resultado deste projeto foi planejado para atender todos os públicos, desde um simples usuário, até um especialista da área biológica interessado em realizar RAF. Neste sentido, para facilitar o uso dos usuários avançados, que consigam montar seus blocos de comandos sem auxílio gráfico, foi criado um campo para receber esses comandos, é o “*Use custom control block*”. Caso esse campo esteja preenchido, os parâmetros definidos nas telas são ignorados e a chamada do programa escolhido será feita passando os dados definidos no campo citado. Sendo assim, pode-se dar seguimento à execução do programa de acordo às explicações dadas anteriormente.

Vale destacar também, o botão “*Clear*”, que forme a grafia, faz uma espécie de limpeza dos parâmetros do método em questão: todos os campos voltam ao estado inicial. Todos esses botões citados, assim como as suas respectivas funcionalidades, estão disponíveis para cada método, disparando rotinas de acordo à necessidade do software utilizado. Estes podem ser consultados através da Figura 25.

## 6 TRABALHOS CORRELATOS

Esta seção tem o objetivo de apresentar os resultados das pesquisas realizadas para obter fundamentos dos trabalhos relacionados ao apresentado nesta dissertação. O foco desta busca foi direcionado na aquisição de conhecimento das características e mecanismos visuais gráficos de cada software explorados anteriormente. A princípio, foram analisadas as opções disponibilizadas para RAF que operam a nível de usuário. Dentre os vários serviços Web disponíveis na literatura, os que mais se destacam são: *Phylemon*<sup>15</sup> (SÁNCHEZ et al., 2001), *Phylogeny*<sup>16</sup> (DEREEPER et al., 2008), *Mobyle @Pasteur*<sup>17</sup> (NÉRON et al., 2009) e *T-Rex*<sup>18</sup> (BOC; DIALLO; MAKARENKOV, 2012). Estes serão analisados a seguir.

### *Phylemon*

É uma suíte de ferramentas Web para evolução molecular e filogenias. Foi projetado de acordo as prerrogativas de rede de computadores (Cliente-Servidor), onde o lado do servidor foi implementado em Java, enquanto o lado do cliente foi desenvolvido em AJAX (Asynchronous JavaScript and XML). O JSON (JavaScript Object Notation) foi empregado para fazer a troca de dados entre cliente e servidor. O *Phylemon* pode ser acessado por usuários anônimos ou registrado. A diferença está no fato de que os registrados podem ter muitos projetos diferentes e usar o servidor para armazenar até 1GB de dados. Os anônimos têm seus arquivos apagados após 24 horas. Possui uma estrutura sofisticada de gerenciamento dos trabalhos submetidos, onde o usuário pode alterar, parar, deletar, e até mesmo acompanhar o andamento da execução do processo. Além disso, pode-se visualizar os resultados das execuções na própria ferramenta. Apresenta em sua interface, softwares que representam os 4 métodos de RAF citados neste trabalho, conforme descrição a seguir:

- Método de Distância: este método é representado pelos softwares DNADIST e PROTDIST, pertencentes ao pacote PHYLIP na versão 3.68;
- Método de Máxima Parcimônia: são fornecidos, através do pacote PHYLIP na versão 3.68, os softwares DNAPARS e PROTPARS;
- Método de Máxima Verossimilhança: a análise pode ser realizada utilizando DNAML e PROML, pertencentes ao PHYLIP na versão 3.68;
- Método de Inferência Bayesiana: representado pelo MrBayes na versão 3.1.2.

<sup>15</sup> O Phylemon pode ser acessado através do link <http://phylemon.bioinfo.cipf.es/phylogeny.html>.

<sup>16</sup> O Phylogeny está disponível do link: <http://www.phylogeny.fr>.

<sup>17</sup> O Mobyle @Pasteur possui o seguinte endereço eletrônico: <http://mobyle.pasteur.fr/cgi-bin/portal.py#welcome>.

<sup>18</sup> O T-Rex é disponibilizado através do endereço eletrônico: <http://www.trex.uqam.ca>.



### ***Phylogeny***

É um simples serviço gratuito da Web, que não exige autenticação, dedicado a oferecer soluções de bioinformática, com algumas limitações relacionadas com o tamanho das sequências em análise. Estas limitações variam de acordo ao software escolhido: para o caso do MrBayes, por exemplo, apenas arquivos com 30 sequências são aceitos. Possibilita a visualização gráfica e um resumo simplificado do resultado filogenético. Os softwares disponibilizados em sua interface, de acordo ao método de inferência filogenética, são:

- Método de Distância: apresenta o software BIONJ;
- Método de Máxima Parcimônia: disponibiliza o software TNT na versão 1.1;
- Método de Máxima Verossimilhança: oferece o PHYML na versão 3.1;
- Método de Inferência Bayesiana: apresenta o MrBayes na versão 3.2.3.

### ***Mobyle @Pasteur***

É um portal Web destinado a oferecer soluções relacionadas com a bioinformática, com acessos identificados ou anônimos. Detém uma grande variedade de softwares, os quais podem ser pesquisados através de um mecanismo de busca muito sofisticado. Além disso, possui uma estrutura organizacional em abas, onde os trabalhos em execução ou executados podem ser consultados, que facilita bastante o trabalho do usuário. Os resultados das análises ficam divididos em blocos bastante intuitivos, onde o usuário pode consultar todas as informações inseridas, com possibilidade de reprocessamento, download, deleção e visualização gráfica do resultado filogenético. Apresenta os seguintes softwares de RAF de acordo aos seus respectivos métodos:

- Método de Distância: oferece os softwares Dnadist, Fitch, Kitsch, Neighbor e Protdist, todos pertencentes ao PHYLIP na versão 3.67, além do BIONJ, do Distmat (pertence ao EMBOSS na 6.3.1), do QuickTree na versão 1.1 e o Weighbor na versão 1.2.1;
- Método de Máxima Parcimônia: possui os softwares DNAPARS, PROTPARS, PARS, MIX, todos pertencentes ao pacote PHYLIP na versão 3.67;
- Método de Máxima Verossimilhança: oferta os software FastDNaml na versão 1.2.2, MorePhyML na versão 1.14, PHYML na versão do ano de 2012, SeqGen na versão 1.3.2 e o Tree-Puzzle na versão 5.2;
- Método de Inferência Bayesiana: disponibiliza o BAMBE na versão 4.01.

### ***T-Rex***

É um serviço Web dedicado a RAF, de acesso livre e com possibilidade de visualização gráfica do resultado filogenético. Foi desenvolvido em C++, tendo em seu escopo, os seguintes softwares, de acordo aos seus respectivos métodos de inferência filogenética:

- Método de Distância: oferece os softwares Neighbor-Joining, NINJA large-scale Neighbor Joining, ADDTREE, Unweighted Neighbor Joining, Circular order reconstruction, Weighted least-squares (MW), BIONJ, FITCH (PHYLIP);
- Método de Máxima Parcimônia: disponibiliza os softwares DNAPARS, PROTPARS, PARS e DOLLOP, todos pertencentes ao pacote PHYLIP;
- Método de Máxima Verossimilhança: disponibiliza os softwares DNAML, DNAMLK, PROML, PROMLK, todos acoplados ao PHYLIP, PHYML e RAxML;
- Método de Inferência Bayesiana: não oferece software de RAF utilizando tal método.

A seguir serão feitas algumas análises comparativas entre os Web Services citados acima e o IgrafuWeb. Com o intuito de ser o mais criterioso possível e facilitar a compreensão do leitor, as comparações foram feitas por tipos de métodos de acordo ao software relacionado. Além disso, foram criadas tabelas, a fim de elucidar as diferenças de opções de comandos ofertados por cada Web Service.

Foram criadas 4 tabelas, uma para cada software disponibilizado para RAF no IgrafuWeb (MrBayes, PHYML, Digrafu e DNAPARS ou PROTPARS), de maneira a facilitar a compreensão das comparações efetuadas. Sendo assim, a estrutura das tabelas têm: uma coluna (a primeira) que representa os comandos estipulados por cada software, e as restantes, correspondem às opções destes comandos ofertados pelos Web Services.

Vale ressaltar que, algumas tabelas não apresentam a especificação de coluna que corresponda a um determinado Web Services, devido a não disponibilização do software em questão. Como exemplo, pode-se citar a *Tabela 3*, a qual não possui a coluna destinada à comparação dos valores dos comandos para as plataformas Web *Mobyle @Pasteur* e *Trex*, pois estas não apresentam a utilização do MrBayes.

### **MrBayes**

O IgrafuWeb apresenta vários campos que são relevantes para utilização do MrBayes, mas não são utilizados pelas ferramentas citadas. Pode-se citar como exemplo o *Phylogeny*, que não oferece ao usuário a possibilidade de definição dos campos: *Ngammacat*, *Nruns*, *Swapfreq*,

*Printfreq*, *Nchains*, *Savebrlens* e *Ordertaxa* (campos encontrados na aba MCMC). O campo *Printfreq* não foi inserido na interface, mas é utilizado em rotinas internas com valor 1. Além disso, o *Phylogeny* também não apresenta o tipo correlacionado de taxa de heterogeneidade entre os sítios: opção *adgamma* (campo da aba Model).

O IgrafuWeb apresenta um diferencial relevante em relação as outras plataformas ao oferecer a possibilidade de definição da topologia e da árvore de entrada, através da aba *Tree* do site. Outro ponto crucial de destaque é a aba Sumarize, a qual não é apresentada pelo *Phylogeny*, e o *Phylemon* disponibiliza apenas os campos *Burning*, *Contype* e *showtreeprobs* caso o modo de execução seja não iterativo. Vale ressaltar, que o campo *showtreeprobs* não está disponibilizado na interface, mas está sendo utilizado internamente pelo sistema.

Seguindo a análise das particularidades do IgrafuWeb, o MrBayes oferece alguns recursos extras para o modelo de probabilidade, que são disponibilizados apenas neste projeto, é o caso dos parâmetros *Tratiopr*, *Revmatpr* e *Covswitchpr*.

É relevante ressaltar que os Web Services Mobylyle @Pasteur e *Trex* não disponibilizam a interface gráfica para utilização do MrBayes. Além disso, cabe observar que o *Phylemon* não suporta a RAF de proteínas. Todas estas informações citadas nesta seção podem ser consultadas através da *Tabela 3*.

Portanto, constata-se que o IgrafuWeb é uma ferramenta completa, que corresponde às expectativas de inferência filogenética baseada no método bayesiano utilizando o MrBayes.

Tabela 3 - Comparação dos parâmetros do MrBayes.

MrBayes	PhyIemon	Phylogeny	IgrafuWeb
Nst	1 (J69 or F81), 2 (K80 or HKY85) ou 6 (GTR)	1 (J69 or F81), 2 (K80 or HKY85) ou 6 (GTR)	GTR, SYM, HKY, K2P, F81, JC
nucmodel DNA	codon, 4by4 ou Doublet	4by4, doublet ou códon	4by4, doublet ou códon
nucmodel Proteína	Não oferece suporte para arquivos de proteína	Poisson, Dayhoff, Blosun62, WAG, Mtrev, Mtmam, Rtrev, Cprev ou Vt	Poisson, Jones, Dayhoff, Mtrev, Mtmam, WAG, Rtrev, Cprev, Vt ou Blosun62
Code	Universal, Vert. mit. DNA, Meta. mit. DNA, Mycoplasma, Yeast ou Ciliate	Universal, Vertmt, Mycoplasma, Yeast, Ciliates ou Metmt	Universal, Vert. mit. DNA, Mycoplasma, Yeast, Ciliates ou Meta. mit. DNA
Rates	Equal, Gamma, Propinv, Invgamma ou Adgamma	Equal, Gamma, Propinv ou Invgamma	Equal, Gamma, Adgamma Propinv ou Invgamma
Ngammacat	Default, longer run ou even longer run	Não possui	O campo é numérico com valor default 4
Ngen	Campo numérico: o valor default é 1000	Opções: 1000, 10000(default) ou 100000	Campo numérico: o valor default é 1000000
Nruns	Opções: 1, 2(default), 5 ou 10	Não possui	Campo Numérico: o valor Default é 2
Swapfreq	Opções: 1(default), 5 ou 10	Não possui	Numérico: o valor default é 1
Samplefreq	Opções: 100(default), 1000 ou 5000	Opções: 10(default), 100 ou 1000	Numérico: Default 100 e Mínimo 1
Printfreq	Opções: 100(default), 1000, 5000 ou 10000	Não possui	Utilizado em rotinas internas com valor fixo default 1
Nchains	Opções: 1 ou 4(default)	Não possui	Numérico: Default 4 e Mínimo 1
Savebrlens	Opções: Yes(default) ou no	Não possui	Opções: Yes (default) ou no
Ordertaxa	Opções: Yes(default) ou no	Não possui	Opções: Yes ou no (default)
Outgroup	Campo Numérico	Não possui	Não possui
Burning	Não possui	Opções: 10, 25, 50, 100, 250(default), 500, 1000, 2500 e 5000	Campo numérico
Contype	Não possui	Não possui	Opções: Allcompat ou Halfcompat
Showtreeprobs	Não possui	Não possui	Utilizado em rotinas internas com valor fixo default “no”

## PHYML

O IgrafuWeb abrange todos os parâmetros disponibilizados pelo PHYML de acordo às especificações do manual.

Após analisar o *Phylogeny* e fazer uma comparação com o IgrafuWeb, ficou evidente as diferenças entre ambos, as quais serão descritas a seguir.

Analisando as possibilidades de configurações do modelo, o *Phylogeny* não apresenta o parâmetro de equilíbrio de frequências e peca em relação aos modelos de substituição. No caso de DNA, não oferece os modelos 'JC69', 'K80', 'F81', 'F84' e 'TN93'. Para as sequências de proteína, não disponibiliza os modelos mtrev, mtmam, dcmut, rtrev, cprev, vt, mtart, hivw, hivb e blosum.

Analisando os dados da aba Sequence, o *Phylogeny* deixa de oferecer os parâmetros relacionados ao tipo de sequência (Interleaved ou Sequential), ao número de conjunto de dados, ao gerador de números aleatórios, a impressão do likelihood e a impressão da filogenia analisada.

Ainda analisando o *Phylogeny*, constata-se que nenhum parâmetro relacionado à árvore (aba Tree do IgrafuWeb) está disponibilizado em sua interface.

Os outros 3 Web Services, o *Mobyle @Pasteur*, o *Phylemon* e o *Trex* possuem interfaces compatíveis, com as algumas diferenças pontuais, em termos de configuração de parâmetros, comparadas com o IgrafuWeb. Portanto, pode-se afirmar a veracidade das informações apresentadas no IgrafuWeb relacionadas ao PHYML. Estas análises podem ser consultadas através da *Tabela 4*.

Tabela 4 - Comparação das opções dos parâmetros do PHYML disponibilizadas pelos Web Services estudados.

PHYML	Phylemon	Phylogeny	Mobyle @Pasteur	T-Rex	IgrafuWeb
-i ou --input	Utilizado em rotinas internas	Utilizado em rotinas internas	Utilizado internamente	Utilizado em rotinas internas	Utilizado em rotinas internas
-d ou --datatype	Opções: DNA ou Amino Acid	Opções: auto-select, protein ou DNA/RNA	Opções: DNA (nt) ou Amino Acid (aa)	Opções: DNA ou Amino Acids	Opções: Dna ou Aminoacid
-q ou --sequential	Não possui	Não possui	Não possui	Opções: Interleaved ou Sequential	Opções: Interleaved ou Sequential
-n ou --multiple	Campo numérico	Não Possui	Campo numérico	Não Possui	Campo numérico
-p ou --pars	Opções: Yes ou No	Não Possui	Não Possui	Não Possui	Opções: Yes ou No
-b ou --bootstrap	b = -1: "approximate likelihood ratio test returning aLRT statistics"; b = -4: "SH-like branch supports alone" b = -2: "approximate likelihood ratio test returning Chi2-based para-metric branch supports" b = 0: "neither approximate likelihood ratio test nor bootstrap values are computed" b > 0: Campo numérico	b = -1: "Minimum of SH-like and Chi2-based"; b = -2: "Chi2-based parametric". b = -4: "SH-like"; b > 0: Campo numérico	Campo numérico para b > 0	b = 0: "No"; b = -1: "aLRT statistics" b = -2: "Chi2-based supports" b = -4: "SH-like supports" b > 0: Campo numérico	b = 0: "Neither approximate nor bootstrap values are computed"; b = -1: "aLRT statistics" b = -2: "Chi2-based parametric branch supports" b = -4: "SH-like branch supports alone" b > 0: Campo numérico
-m ou --model (DNA)	Opções: HKY85, JC69, K80, F81, F84, TN93, GTR ou custom	Opções: HKY85 ou GTR	Opções: HKY85, JC69, K80, F81, F84, TN93 ou GTR	Opções: JC69, K80, F81, F84, HKY85, TN93 ou GTR	Opções: HKY85, JC69, K80, F81, F84, TN93 ou GTR
-m ou --model (Proteína)	Opções: LG, WAG, JTT, MtREV, Dayhoff, DCMut, RtREV, CpREV, VT, Blosum62, MtMam, MtArt, HIVw ou HIVb	Opções: WAG, JTT ou Dayhoff	Opções: LG, WAG, JTT, MtREV, Dayhoff, DCMut, RtREV, CpREV, VT, Blosum62, MtMam, MtArt, HIVw ou HIVb	Opções: LG, WAG, Dayhoff, JTT, Blosum62, MtREV, RtREV, DCMut, VT, MtMam, MtArt, HIVw ou HIVb	Opções: WAG, JTT, MtREV, Dayhoff, DCMut, RtREV, CpREV, VT, Blosum62, MtMam, MtArt, HIVw ou HIVb
-f	Opções: <b><i>"Use empirical frequencies", "Frequencies estimated using maximum</i></b>	Não possui	Opções: <b><i>"e", "m" ou "campos numéricos para frequências de A, C, G e T"</i></b>	Não possui	Opções: <b><i>"Empirical", "Estimated" ou "Equal"</i></b>

PHYML	PhyML	Phylogeny	MoBYE @ Pasteur	T-Rex	IgrafuWeb
	<i>likelihood for DNA, and model for amino acid</i> " ou " <i>Equal Frequencies (A=0.25, C=0.25, G=0.25, T=0.25)</i> "				
-t ou --ts/tv	Opções: Yes ou No (definição de valor numérico)	Opções: Estimated ou Fixed (definição de valor numérico)	Opções: Yes ou No (definição de valor numérico)	Opções: Estimated ou Fixed (definição de valor numérico)	Opções: Estimated ou Fixed (definição de valor numérico)
-v ou --pinv	Opções: Yes ou no (definição de valor numérico)	Opções: Estimated ou Fixed (definição de valor numérico)	Opções: Yes ou no (definição de valor numérico)	Opções: Estimated ou Fixed (definição de valor numérico)	Opções: Estimated ou Fixed (definição de valor numérico)
-c ou --nclases	Campo numérico com valor padrão 4	Campo numérico com valor padrão 4	Opções: Yes ou no (definição de valor numérico)	Campo numérico com valor padrão 4	Campo Numérico com valor padrão 4
-a ou --alpha	Opções: Yes ou no (esta opção exige definição de valor numérico)	Opções: Estimated ou Fixed (definição de valor numérico)	Opções: Yes ou no (exige definição de valor numérico)	Opções: Estimated ou Fixed (definição de valor numérico)	Opções Estimated ou Fixed (definição de valor numérico)
--use_median	mean Utilizada por padrão	mean Utilizada por padrão	mean Utilizada por padrão	Opções: mean ou median	mean Utilizada por padrão
-s	Opções: Nearest Neighbor Interchange (fast), Subtree Pruning and Regrafting (slower) ou Best of NNI and SPR search	Não possui	Opções: NNI, SPR (a bit slower than NNI) ou BEST (best of NNI and SPR search)	Opções: NNI moves (fast, approximate), SPR moves (slow, accurate) ou Best of NNI and SPR	Opções: NNI, SPR ou NNI & SPR
-u ou --inputtree	Não possui	Não possui	Upload de arquivo	Opções: BIONJ, Parsimony ou Use tree (upload de arquivo)	Opções: BIONJ ou File
-o	Opções: " <b>Branch length optimization</b> ", " <b>Rate parameters optimization</b> " e/ou " <b>Tree topology optimization</b> "	Não possui	Opções: " <b>tree topology (t), branch length (l) and rate parameters (r) are optimized</b> ", " <b>tree topology and branch lengths are optimized</b> ", " <b>branch lengths and rate parameters are optimized</b> ", " <b>branch lengths are optimized</b> ",	Não possui	Opções: " <b>Optimise Branch Lengths</b> ", " <b>Rate Parameters Optimization</b> " e/ou " <b>Optimise Topology</b> "

PHYML	Phylemon	Phylogeny	Mobyle @Pasteur	T-Rex	IgrafuWeb
			<i>“rate parameters are optimized”</i> ou <i>“no parameter is optimised”</i>		
--rand_start	Não possui	Não possui	Opções: No e Yes (exige valor numérico)	Não possui	Opções: No e Yes (exige valor numérico)
--n_rand_starts	Não possui	Não possui	Campo Numérico	Não possui	Campo Numérico
--r_seed	Não possui	Não possui	Campo Numérico	Não possui	Campo Numérico
--print_site_lnl	Não possui	Não possui	Opções: Yes ou No	Não possui	Opções: Yes ou No
--print_trace	Não possui	Não possui	Opções: Yes ou No	Não possui	Opções: Yes ou No



## Digrafu

O Digrafu é um software de RAF baseado no método de distância, que tem acoplado ao seu escopo os algoritmos UPGMA, NJ, BIONJ, Weighbor e FastMe, bem como os programas Dnadist e Protdist, os quais atuam de acordo do tipo de sequência analisado (DNA ou Proteína), para criar a matriz de distâncias. A utilização de um deste não fica explícito na tela para escolha do usuário: um deles será selecionado em tempo de execução, dependendo dos parâmetros definidos nas telas do site (observar item 4.3). Já nos Web Services aqui analisados, estes algoritmos são disponibilizados visualmente em locais separados, e serão utilizados de acordo à necessidade do usuário, sem critérios de escolha de qual algoritmo utilizar. Além dos programas citados, o Digrafu possui também o Seqboot e Consense acoplado ao seu escopo

As análises foram feitas para as opções de comandos dos Web Services, direcionadas para os softwares Dnadist, Protdist, Seqboot e Consense. Sendo assim, de acordo aos estudos realizados, pode-se afirmar que: o *Phylogeny* e o *T-rex* não disponibilizam nenhum dos softwares citados, enquanto que o *Mobyle @Pasteur* e o *Phylemon* apresentam todos eles. Logo, a *Tabela 5* detém apenas as opções do *Phylemon* e do *Mobyle @Pasteur*.

Conforme descrito na *Tabela 5*, as opções de comandos do Dnadist e do Protdist, que vão do comando TYPE até o GAMMA, são totalmente compatíveis com os do IgrafuWeb.

Fazendo a análise do Seqboot, o *Mobyle @Pasteur* possui apenas as opções do modelo de permutação, e as quantidades de sementes e de réplicas. Já para o *Phylemon*, apenas as opções dos arquivos de ancestral e mistura, bem como o formato de saída, não são ofertadas.

De acordo a análise do Consense, o *Mobyle @Pasteur* não disponibiliza a opção que marca a definição do nó raiz, assim como, deixa de ofertar a opção que especifica a fração de vezes que o ramo será analisado. O *Phylemon* não apresenta as opções referentes a: definição do nó raiz, indicação das espécies que serão escritas no arquivo de saída, definição da execução e informação se a árvore será desenhada no arquivo de saída.

Tabela 5 - Comparação dos parâmetros do Digrafu.

Digrafu – DnaDist /ProtDist	Phylemon	Mobyle @Pasteur	IgrafuWeb
TYPE	Opções: dnadist ou protdist	Opções: dnadist ou protdist	Opções: <i>“DNA”</i> ou <i>“Protein”</i>
MODEL DNA	Opções: <i>“F84”</i> , <i>“Kimura 2-parameter”</i> , <i>“Jukes-Cantor”</i> , <i>“LogDet”</i> ou <i>“Similarity Table”</i>	Opções: <i>“F84”</i> , <i>“Kimura 2-parameter”</i> , <i>“Jukes-Cantor”</i> , <i>“LogDet”</i> ou <i>“Similarity Table”</i>	Opções: <i>“F84”</i> , <i>“JC69”</i> , <i>“Kimura”</i> ou <i>“LogDet”</i>
MODEL Prot	Opções: <i>“Jones-Taylor-Thornton matrix”</i> , <i>“Henikoff/Tillier PMB matrix”</i> , <i>“Dayhoff PAM matrix”</i> , <i>“Kimura formula”</i> , <i>“Similarity Table”</i> ou <i>“Categories model”</i>	Opções: <i>“Jones-Taylor-Thornton matrix”</i> , <i>“Henikoff/Tillier PMB matrix”</i> , <i>“Dayhoff PAM matrix”</i> , <i>“Kimura formula”</i> , <i>“Similarity Table”</i> ou <i>“Categories model”</i>	Opções: <i>“Jones-Taylor-Thornton matrix”</i> , <i>“Henikoff/Tillier PMB matrix”</i> , <i>“Dayhoff PAM matrix”</i> ou <i>“Kimura formula”</i>
ISITE	Campo numérico	Campo numérico	Campo numérico
FREQUE	Opções: <i>“Yes”</i> ou <i>“No”</i> (campo numérico para definição de valor para frequências dos nucleotídeos A, C, G, e T)	Opções: <i>“No”</i> ou <i>“Yes”</i> (campo numérico para definição de valor para frequências dos nucleotídeos A, C, G, e T)	Opções: <i>“No”</i> ou <i>“Yes”</i> (campo numérico para definição de valor para frequências dos nucleotídeos A, C, G, e T)
RATIO DNA	Campo numérico	Campo numérico	Campo numérico
WEIGHT	Campo numérico	Campo numérico	Campo numérico
GAMMA	Opções: <i>“No”</i> , <i>“Yes”</i> ou <i>“Gamma+Invariant”</i> (campo numérico para definição de valor para os comandos RATIO e WEIGHT)	Opções: <i>“No”</i> , <i>“Yes”</i> ou <i>“Gamma+Invariant”</i>	Campo numérico
Digrafu – Seqboot	Phylemon	Mobyle @Pasteur	IgrafuWeb
DNA, RNA ou PRO	Utilizado internamente.	Utilizado internamente.	Opções: <i>“DNA”</i> , <i>“RNA”</i> ou <i>“PRO”</i>
SEQU, MORF, REST ou FREQ	Opções: <i>“Molecular sequences”</i> , <i>“Discrete Morphology”</i> , <i>“Restriction Sites”</i> ou <i>“Gene Frequencies”</i> .	Não possui	Opções: <i>“Molecular sequences”</i> , <i>“Discrete Morphology”</i> , <i>“Restriction Sites”</i> ou <i>“Gene Frequencies”</i> .
BOOT, JACK, PERM, PORD ou PSPEC	Opções: <i>“Bootstrap”</i> , <i>“Delete-half jackknife”</i> , <i>“Permute species for each character”</i> , <i>“Permute character order”</i> , <i>“Permute within species”</i> ou <i>“Rewrite data”</i>	Opções: <i>“Bootstrap”</i> , <i>“Delete-half jackknife”</i> , <i>“Permute species for each character”</i> , <i>“Permute character order”</i> ou <i>“Permute within species”</i>	Opções: <i>“BOOTSTRAP”</i> , <i>“JACKKNIFE”</i> , <i>“PERMUTE CHARACTER”</i> , <i>“PERMUTE CHARACTER ORDER”</i> ou <i>“PERMUTE WITHIN SPECIES”</i> .
I ou S	Opções: <i>“Yes”</i> ou <i>“No”</i>	Não possui	Opções: <i>“INTERLEAVED”</i> ou <i>“SEQUENCIAL”</i>

Digrafu – Seqboot	Phylemon	Mobyle @Pasteur	IgrafuWeb
ALL, CAT ou FAC	Submissão de arquivo.	Não possui	Submissão de arquivo.
SEED	Campo numérico	Campo numérico	Campo numérico
REPLICATES	Campo numérico	Campo numérico	Campo numérico
BLOCO	Campo numérico	Não possui	Campo numérico
FRACAO	Opções: “ <b>Yes</b> ” ou “ <b>No</b> ”. Opção “ <b>No</b> ” habilita campo numérico	Não possui	Campo numérico
ENZIMA	Opções: “ <b>Yes</b> ” ou “ <b>No</b> ”.	Não possui	Opções: “ <b>Yes</b> ” ou “ <b>No</b> ”
WEIGHT	Opções: “ <b>Yes</b> ” ou “ <b>No</b> ”. Opção “ <b>Yes</b> ” habilita submissão de arquivo.	Não possui	Submissão de arquivo
MIX	Não possui	Não possui	Submissão de arquivo
ANC	Não possui	Não possui	Submissão de arquivo
RESC	Não possui	Não possui	Opções: “ <b>PHY</b> ”, “ <b>NEXUS</b> ” ou “ <b>XML</b> ”
OUTD ou JUSTW	Opções: “ <b>Data</b> ” ou “ <b>Weights</b> ”	Não possui	Opções: “ <b>Data</b> ” ou “ <b>Weights</b> ”
Digrafu – Consense	Phylemon	Mobyle @Pasteur	IgrafuWeb
FILE	Submissão de arquivo	Submissão de arquivo	Submissão de arquivo
R	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”
MRE, STR, MR ou ML	Opções: “ <b>Majority rule (extended)</b> ”, “ <b>Strict</b> ”, “ <b>Majority rule</b> ” ou “ <b>MI</b> ”	Opções: “ <b>Majority rule (extended)</b> ”, “ <b>Strict</b> ”, “ <b>Majority rule</b> ” ou “ <b>MI (M-sub-L)</b> ”	Opções: “ <b>EXTENDED MAJORITY RULE</b> ”, “ <b>STRICT</b> ”, “ <b>MAJORITY RULE</b> ” ou “ <b>ML</b> ”
ROOT	Não possui	Não possui	Campo numérico
PRINT	Não possui	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”
RUN	Não possui	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”
TREE	Não possui	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”
FRACTION	Campo numérico	Não possui	Campo numérico
OUTGROUP	Opções: “ <b>No</b> ” ou “ <b>Yes</b> ”. Opção “ <b>Yes</b> ” habilita campo numérico	Campo numérico	Não possui

## DNAPARS e PROTPARS

O DNAPARS e o PROTPARS são softwares de RAF baseado no método de Máxima Parcimônia, que pertencente à suíte de pacotes PHYLIP. Dos Web Services analisados, apenas o *Phylogeny* não disponibiliza a utilização destes aplicativos. Sendo assim, o escopo desta seção será baseado na análise dos portais Web *Phylemon*, *Mobyle @Pasteur* e *Trex*. Esta seção esta representada pela *Tabela 6*.

Sendo assim, de acordo à análise da tabela citada, comparando o IgrafuWeb com os Web Services, constata-se as seguintes diferenças:

- O *Phylemon* não disponibiliza a utilização de entradas randômicas para a ordem das sequências (parâmetro “*J*”), além de não apresentar nenhuma das opções de impressão representadas pelos parâmetros “*1, 3, 4, 5 e 6*”;
- O *Mobyle @Pasteur* não apresenta as opções “*More Thorough*”, “*Rearrange on one best tree*” ou “*Less Thorough*” relacionadas ao parâmetro “*U*”, que especifica a busca da melhor árvore. Além disso, a opção de escolha do tipo de sequência, intercalada ou sequencial, não é disponibilizada em sua interface.
- O *Trex* apresenta deficiência quanto às opções de impressão, representadas pelos parâmetros “*1, 3 e 6*”;

Tabela 6 - Comparação dos parâmetros do DNAPARS e PROTPARS.

DNAPARS / PROTPARS	Phylemon	Mobyle @Pasteur	T-Rex	IgrafuWeb
U	Opções <b><i>More Thorough, Rearrange on one best tree</i></b> ou <b><i>Less Thorough</i></b>	Não possui	Opções <b><i>More Thorough, Rearrange on one best tree</i></b> ou <b><i>Less Thorough Search</i></b>	Opções <b><i>More Thorough Search, Rearrange on one best tree</i></b> ou <b><i>Less Thorough Search</i></b>
S	Definição de arquivo	Definição de arquivo	Definição de arquivo	Definição de arquivo
V	Campo numérico	Campo numérico	Campo numérico	Campo numérico
J	Não Possui	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b><i>"No, use input order"</i></b> ou <b><i>"Yes"</i></b>	Opções: <b><i>"No, use input order"</i></b> ou <b><i>"Yes"</i></b>
O	Opções: <b>Yes</b> ou <b>No</b>	Campo Numérico	Opções: <b><i>"No, use as outgroup species 1"</i></b> ou <b><i>"Yes"</i></b>	Opções: <b>Yes</b> ou <b>No</b>
T	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b><i>"Yes"</i></b> ou <b><i>"No, use ordinary parsimony"</i></b>	Opções: <b>Yes</b> ou <b>No</b>
C	Opções: <b><i>Universal, Mitochondrial, Vertebrate mitochondrial, Fly mitochondrial e Yeast mitochondrial</i></b>	Opções: <b><i>Universal (U), Mitochondrial, Vertebrate mitochondrial (M), Fly mitochondrial (F) e Yeast mitochondrial (Y)</i></b>	Opções: <b><i>Universal, Mitochondrial, Vertebrate mitochondrial, Fly mitochondrial e Yeast mitochondrial</i></b>	Opções: <b><i>Universal, Mitochondrial, Vertebrate mitochondrial, Fly mitochondrial e Yeast mitochondrial</i></b>
N	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b><i>"No, count all steps"</i></b> ou <b><i>"Yes, count only transversions"</i></b>	Opções: <b>Yes</b> ou <b>No</b>
W	Opções: <b><i>Yes(upload arquivo)</i></b> ou <b>No</b>	Opções: <b><i>Yes(upload arquivo)</i></b> ou <b>No</b>	Opções: <b><i>Yes(upload arquivo)</i></b> ou <b>No</b>	Opções: <b><i>Yes(upload arquivo)</i></b> ou <b>No</b>
M	Opções: <b><i>"No"</i></b> ou <b><i>"Multiple data sets"</i></b>	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b><i>"No"</i></b> , <b><i>"Yes, multiple data sets"</i></b> ou <b><i>"Yes, multiple sets of weight"</i></b>	Opções: <b><i>"No"</i></b> , <b><i>"Yes, multiple data sets"</i></b> ou <b><i>"Yes, multiple sets of weight"</i></b>
I	Opções: <b>Yes</b> ou <b>No</b>	Não possui	Opções: <b><i>"Yes"</i></b> ou <b><i>"No, sequentiel"</i></b>	Opções: <b><i>"Interleaved"</i></b> ou <b><i>"Sequential"</i></b>
0	Não possui	Não possui	Não possui	Não possui
1	Não possui	Opções: <b>Yes</b> ou <b>No</b>	Não possui	Opções: <b>Yes</b> ou <b>No</b>
2	Não possui	Não possui	Não possui	Não possui
3	Não possui	Opções: <b>Yes</b> ou <b>No</b>	Não possui	Opções: <b>Yes</b> ou <b>No</b>
4	Não possui	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b>Yes</b> ou <b>No</b>
5	Não possui	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b>Yes</b> ou <b>No</b>	Opções: <b>Yes</b> ou <b>No</b>
6	Não possui	Opções: <b>Yes</b> ou <b>No</b>	Não possui	Opções: <b>Yes</b> ou <b>No</b>

Após todas estas análises realizadas através dos trabalhos correlatos, pode-se determinar que o IgrafuWeb detém todas os atributos necessários para ser considerado um Web Services de renome para RAF. Além dos diferenciais citados, o bloco de comando personalizado também caracteriza-se como uma funcionalidade singular do projeto. Portanto, de acordo ao estudo das ferramentas citadas acima, constatou-se que todas as características técnicas (parâmetros dos softwares) que as soluções Web oferecem são também ofertadas pelo software apresentado neste trabalho.

Neste sentido, este capítulo teve como foco a análise de alguns trabalhos relacionados com este projeto, visando situar o leitor da grande diversidade de ferramentas da modelagem de RAF e seus derivados. Sendo assim, pode-se afirmar, que essa variedade de aplicações destinadas a inferir filogenias foi um dos motivos e razões para a realização deste estudo.

## 7 RESULTADOS E DISCUSSÕES

Esta seção tem o objetivo de apresentar os resultados obtidos através das análises filogenéticas submetidas ao IgrafuWeb, organizadas através de um estudo de caso, e as contribuições efetivas apresentadas em termos de performance e incremento dos parâmetros de configurações dos softwares apresentados.

### 7.1 Contribuições

Este projeto de pesquisa apresenta um ambiente Web para RAF. Mesmo que existam outros ambientes internacionais com este mesmo intuito, o IgrafuWeb possui contribuições consideráveis em relação a estes, que são: quantidades de parâmetros ofertados pelos softwares MrBayes, PhymI, Digrafu e DNAPARS ou PROTPARS, e o ganho de desempenho apresentado pela computação de alta performance para apresentar os resultados filogenéticos.

Neste sentido, o capítulo anterior apresentou os trabalhos correlatos a este, que analisa os Web Services internacionais disponíveis para RAF. Logo, percebe-se pelas análises realizadas, que o IgrafuWeb oferta uma série de parâmetros de configurações dos softwares MrBayes, PhymI, Digrafu e DNAPARS ou PROTPARS que são omitidos pelos Web Services *Phylemon*, *Phylogeny*, *Mobyle @Pasteur* e *Trex*. Esta é uma das contribuições apresentada por este trabalho e estes valores quantitativos serão descritos a seguir analisando cada plataforma e software em questão.

Sendo assim, pode-se constatar, em termos quantitativos, de acordo a análise da Tabela 3, que o IgrafuWeb possui 13 parâmetros a mais de configurações do MrBayes comparados com as plataformas *Phylogeny* e *Phylemon*, divididos em 3 e 10, respectivamente.

Seguindo esse mesmo raciocínio, em termos numéricos, pode-se afirmar, de acordo a análise da Tabela 4, que o IgrafuWeb possui 30 parâmetros a mais de configuração do PhymI que os outros Web Services, divididos da seguinte forma: *Phylemon* com 7, *Phylogeny* com 12, *Mobyle @Pasteur* com 2 e o *Trex* com 9 parâmetros a menos.

Ainda neste contexto, constata-se, de acordo à análise da Tabela 5, que o IgrafuWeb possui 20 parâmetros a mais de configuração do Digrafu, sendo que este valor, divide-se entre as plataformas *Phylemon* e *Mobyle @Pasteur*, com 7 e 13 parâmetros a menos, respectivamente.

Neste sentido, pode-se afirmar, de acordo a análise da Tabela 6, que o IgrafuWeb possui 17 parâmetros a mais de configuração do DNAPARS e PROTPARS comparados com os outros Web

Services. Esta quantidade de parâmetros divide-se em valores 8, 4 e 5, representando a quantidade de parâmetros a menos que as plataformas *Phylemon*, *Mobyle @Pasteur* e *Trex* deixam de disponibilizar, respectivamente, para a utilização do DNAPARS e do PROTPARS.

Uma vez que a contribuição relacionada com a quantidade numérica de parâmetros de configuração dos softwares de RAF foi apresentada, pode-se analisar o desempenho do IgrafuWeb, confrontando os tempos de execução de cada Web Service citado.

Neste contexto, os esforços foram destinados a encontrar sequências genéticas grandes que comprovassem o alto desempenho computacional apresentado pelo IgrafuWeb. Dentre as diversas análises realizadas, a que apresentou um resultado mais satisfatório foram as sequências genéticas de DNA com 20 espécies e 3768 sítios. Essas sequências foram submetidas ao IgrafuWeb, utilizando os softwares PHYML e MrBayes, sendo que os tempos de execução destas submissões foram confrontados com as execuções destas mesmas sequências nos respectivos softwares, utilizando os Web Services em análise.

Seguindo esta linha de raciocínio, as submissões do PHYML foram processadas e os tempos de execução podem ser consultados através da Tabela 7.

Tabela 7 - Tempo de Execução do PHYML para sequências genéticas de DNA com 20 espécies e 3768 sítios, utilizando o modelo HKY85 e bootstrap com valor igual a 100.

	IgrafuWeb	<i>Mobyle @Pasteur</i>	<i>Phylemon</i>	<i>Phylogeny</i>
Tempo de Execução	00h:32m:02s	00h:33m:57s	01h:52m:19s	01h:34m:35s

Com os valores apresentados na Tabela 7, pode-se tecer algumas observações, são elas:

- O IgrafuWeb apresentou o melhor resultado comparado com os outros Web Services. Isso comprova a computação de alto desempenho apresentada por este ambiente Web de RAF;
- O *Mobyle @Pasteur* apresentou um resultado muito semelhante ao IgrafuWeb, corroborando, também, a ideia da utilização de um ambiente de alta performance utilizando computação paralela;
- O *Phylemon* e o *Phylogeny* não apresentaram resultados satisfatório a ponto de competir com os dados do IgrafuWeb e do *Mobyle @Pasteur*. Com isso, pode-se afirmar que os dois Web Services em questão não possuem uma plataforma de alta performance para executar seus processos filogenético;



- Como o *Trex* constrói a árvore filogenética utilizando recursos diferentes do tradicional, ele não serviu de parâmetro para confrontar os tempos de execuções dos outros Web Services. Para mais detalhes sobre a RAF utilizando o *Trex*, recomenda-se a leitura do trabalho BOC, DIALLO e MAKARENKOV (2012).

Além do comparativo do tempo de execução do PHYML, foram feitas também análises de tempo de execução para o MrBayes, utilizando as mesmas sequências, com 20 espécies e 3768 sítios.

Embora o *Phylemon* disponibilize a utilização do MrBayes em sua interface, a execução das sequências citadas não procedeu como esperado, e o resultado filogenético não foi retornado, impossibilitando o comparativo.

Neste contexto, de acordo a análise dos parâmetros disponibilizados pelo *Phylogeny* para executar o MrBayes, constatou-se que as possibilidades de configuração são limitadas: são ofertados apenas 6 parâmetros. Mesmo com essa deficiência, o comparativo foi realizado, fazendo a execução do IgrafuWeb de acordo aos parâmetros disponibilizados pelo *Phylogeny*, tornando as 2 execuções o mais semelhante possível em termos de configuração do MrBayes. Com isso foi possível confrontar esses dados e comprovar a alta performance apresentada pelo IgrafuWeb, onde o mesmo precisou de apenas 4 minutos para retornar a árvore, enquanto o *Phylogeny* necessitou de 7 minutos e 49 segundos. Estes dados podem ser consultados através da Tabela 8. É relevante mencionar, que o *Mobyle @Pasteur* e o *Trex* não apresentam o MrBayes para RAF em sua interface.

Tabela 8 - Tempo de Execução do MrBayes para sequências genéticas de DNA com 20 espécies e 3768 sítios, utilizando o modelo GTR, com 100.000 gerações e com taxa de heterogeneidade InvGamma.

	IgrafuWeb	<i>Phylogeny</i>
Tempo de Execução	00h:04m:00s	00h:07m:49s

## 7.2 Estudo de Caso

Esta seção tem o objetivo de apresentar os resultados obtidos através da utilização de sequências genéticas conhecidas e que pudessem servir de parâmetros para avaliar a qualidade e a veracidade das informações resultantes do processamento do software apresentado neste trabalho.

Sendo assim, dentre as diversas análises realizadas, as sequências escolhidas foram as de *Solanum* do gene COSII\_At5g14320, retiradas do banco de dados do Genbank do NCBI<sup>19</sup>. Estas

<sup>19</sup> Link para download das sequências analisadas: <http://www.ncbi.nlm.nih.gov/popset/224980695?report=fasta>.

sequências foram devidamente alinhadas, e posteriormente submetidas ao IgrafuWeb, com o claro objetivo de analisar e validar o resultado da inferência filogenética em questão. Isso será discutido a seguir.

Como o alinhamento é uma etapa crucial do processo de inferência filogenética, as sequências nucleotídicas analisadas foram devidamente alinhadas antes de serem submetidas ao IgrafuWeb. Este alinhamento foi realizado através do programa Clustalw2, utilizando o EBI<sup>20</sup>. O resultado desse alinhamento foi salvo nos formatos NEXUS e PHYLIP, sendo posteriormente submetidos ao IgrafuWeb de acordo à explicação a seguir: o arquivo no formato NEXUS<sup>21</sup> foi utilizado para realizar o processo filogenético através do MrBayes, enquanto o do PHYLIP<sup>22</sup> intercalado foi utilizado através do PHYML e do Digrafu, e o do PHYLIP<sup>23</sup> sequencial foi utilizado pelo DNAPARS.

Cabe mencionar, que estes arquivos foram editados com o objetivo de melhorar a visualização gráfica da árvore através do applet. Os nomes de todas as sequências foram devidamente resumidos, ficando basicamente da seguinte forma: para uma espécie identificada originalmente como “gi|224980695|gb|FJ599363.1| *Solanum Bulbocastanum* COSII\_At5g14320 gene”, a alteração resultou apenas em *Solanum Bulbocastanum*. Estas alterações foram meramente para facilitar a ilustração da árvore em si, não alterando em nada o resultado final do alinhamento.

Uma vez que as sequências foram alinhadas, o passo seguinte é submetê-las ao IgrafuWeb e comparar o resultado com uma árvore conhecida no meio científico. Como estas mesmas sequências, que por sinal são de DNA, contendo 11 espécies e 652 caracteres, foram submetidas à análises filogenéticas no trabalho do RODRIGUEZ et al. (2009), resultando em algumas árvores, ficou evidente que uma das filogenias deste trabalho (Figura 50) servirá de parâmetro para avaliar o resultado da inferência filogenética do software resultado deste projeto. Esta pesquisa realizada por RODRIGUEZ et al. em 2009 destinou-se a tentar fornecer as relações filogenéticas entre tomates e batatas, que pudessem ajudar as marcações apropriadas para estudos futuros, utilizando espécies adicionais.

De acordo com o trabalho do RODRIGUEZ et al. (2009), a filogenia apresentada na Figura 50 foi resultado de várias análises que resultou na “concordância” entre os métodos de Máxima Parcimônia, Máxima Verossimilhança e Inferência Bayesiana: todos retornaram a mesma árvore.

---

<sup>20</sup> Link do software utilizado para realizar o alinhamento das sequências: <http://www.ebi.ac.uk/Tools/msa/clustalw2>.

<sup>21</sup> Link do arquivo sequencial no formato NEXUS: <http://nbcgib.uesc.br/igrafuweb/tree/filesResults/file.nex>.

<sup>22</sup> Link do arquivo intercalado no formato PHYLIP: <http://nbcgib.uesc.br/igrafuweb/tree/filesResults/phylip.interleaved>.

<sup>23</sup> Link do arquivo sequencial no formato PHYLIP: <http://nbcgib.uesc.br/igrafuweb/tree/filesResults/phylip.sequential>.

Sendo assim, analisando as execuções com base no modelo de evolução, de maneira geral, os modelos de probabilidade que permitiram a variação da taxa de distribuição gama entre os sítios, resultou no maior aumento da verossimilhança máxima. Em mais de 90% dos casos, o modelo HKY, que inclui cinco parâmetros, melhor se ajustou aos dados.

As configurações dos parâmetros, bem como os softwares de RAF, de acordo ao método de inferência filogenética específico, utilizados no trabalho do RODRIGUEZ et al. (2009), são apresentados a seguir.

Para as análises filogenéticas com base na Máxima Parcimônia foram utilizadas pesquisas heurísticas sobre os critérios de Fitch (FITCH, 1971), com pesos iguais para todos os caracteres, aplicadas através do PAUP<sup>24</sup>.

Já para MV, a filogenia foi estimada utilizando o software RAxML<sup>25</sup>, versão 7.0.3, que permite que cada partição possa ter seu próprio modelo e parâmetros. Para avaliar a estabilidade dos clados na árvore ótima, uma análise de bootstrap foi executada com 100 repetições.

Para Inferência Bayesiana foi utilizado o MrBayes, versão 3.1.2, com execuções de 4 cadeias com 1,1 ou 2,2 milhões de gerações, com árvores amostradas a cada 100 gerações.

Portanto, os parágrafos a seguir descrevem e comparam o resultado das filogenias submetidas ao IgrafuWeb com a árvore exibida na Figura 50, seguindo os padrões e configurações de execuções citados acima.

---

<sup>24</sup> Página oficial do PAUP: <http://paup.csit.fsu.edu>.

<sup>25</sup> Página oficial do RAxML: <http://sco.h-its.org/exelixis/web/software/raxml/index.html>.

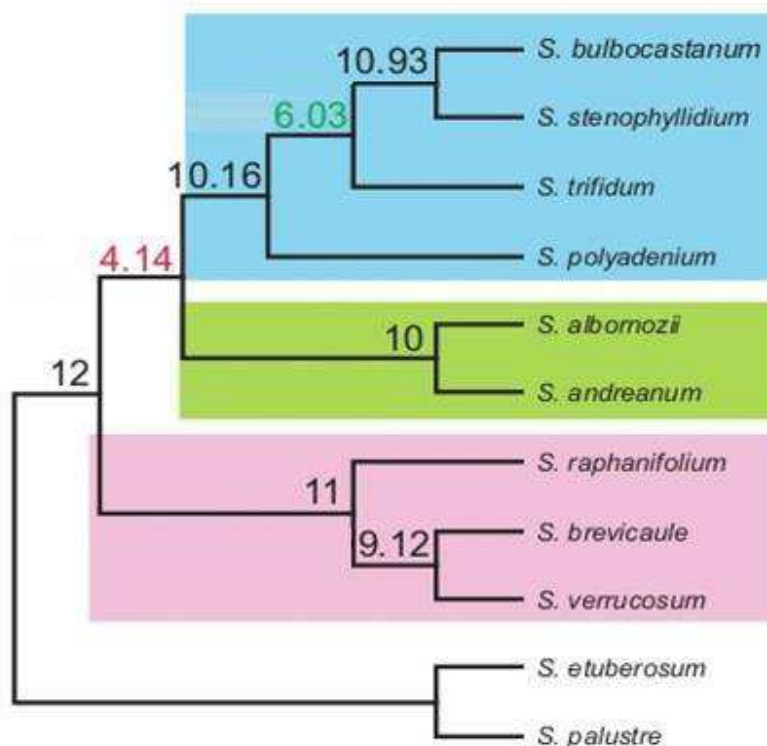


Figura 50 - Filogenia da *Solanum* com sequências do gene COSII\_At5g14320 para dados de batata.

Para o método de Distância foram realizadas buscas heurísticas utilizando o modelo F84. Após a análise, a árvore filogenética gerada foi visualizada através do applet Archaeopteryx, gerando posteriormente uma imagem exibida na Figura 51. Sendo assim, analisando a presente filogenia com da Figura 50, constata-se que o grupo monofilético (grupo que inclui o ancestral e todos os seus descendentes) formado pelas espécies *Raphanifolium*, *Verrucosum* e *Brevicaule* foi conservado, assim como, o clado (grupo de organismos originados de um único ancestral comum exclusivo) formado pelas espécies *Etuberosum* e *Palustre*. Além disso, o grupo monofilético que contém as espécies *Stenophyllidium*, *Bulbocastanum*, *Polyadenium* e *Trifidum* também foi mantido, porém as duas últimas espécies citadas passaram a formar um clado. Outra diferença constatada é a monofilia das espécies *Albornozii* e *Andreanum*.

Vale ressaltar, que o trabalho do RODRIGUEZ et al. (2009) não apresentou resultados analisando o método de Distância, porém para manter o padrão e a qualidade deste projeto de pesquisa, a análise deste método foi devidamente realizada de acordo às configurações padrões do Digrafu.

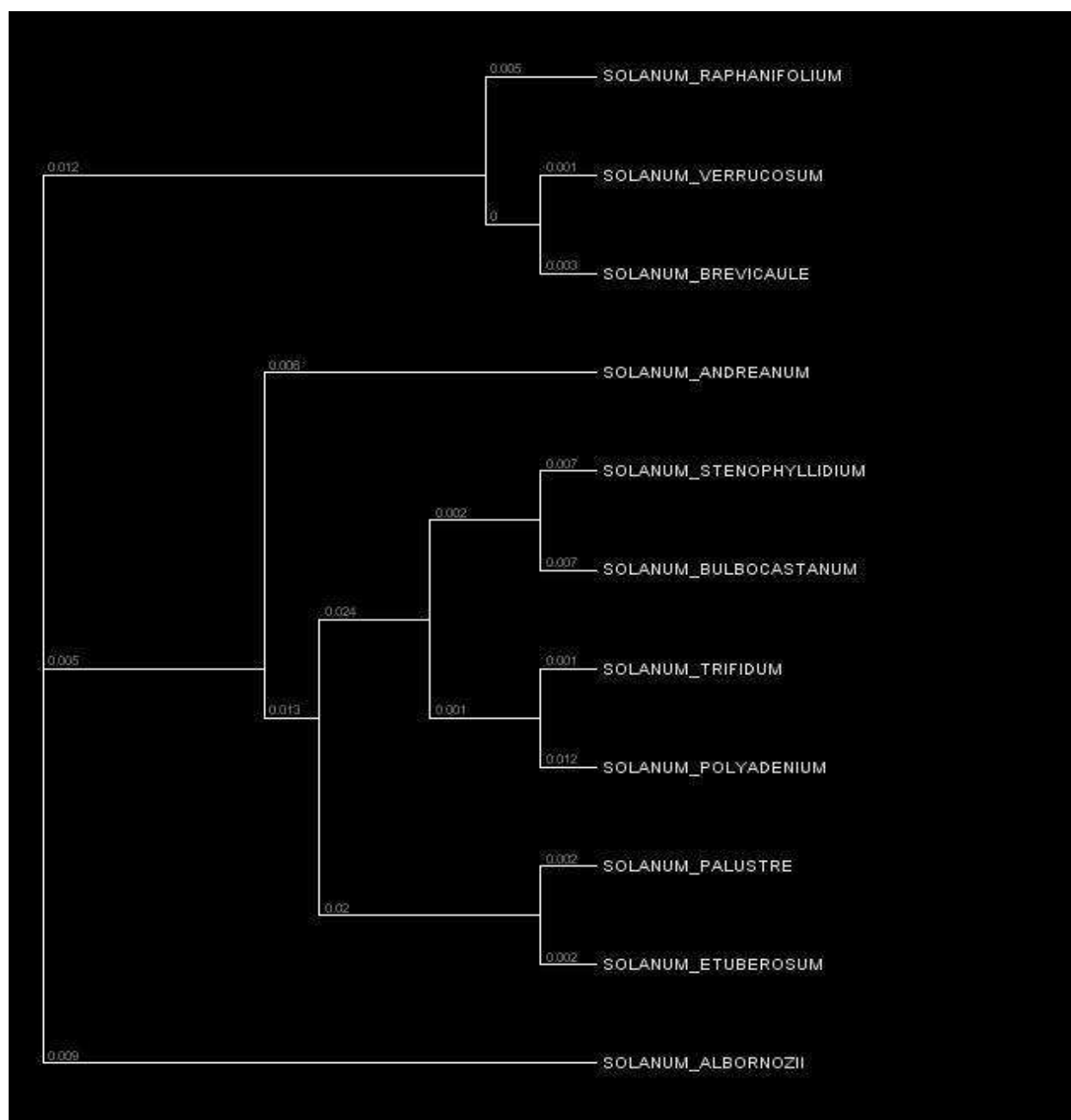


Figura 51 - Resultado da filogenia da *Solanum* com sequências do gene COSII\_At5g14320 utilizando o Método de Distância, mais precisamente, o Digrafu, através do IgrafuWeb. Os números nas linhas indicam o comprimento dos ramos.

As sequências genéticas do gene COSII\_At5g14320, quando submetidas à inferência por parcimônia apresentou uma topologia (Figura 52), na qual constata-se que o grupo monofilético formado pelas espécies *Stenophyllidium*, *Bulbocastanum*, *Polyadenium* e *Trifidum* foi conservado, assim como, o clado formado pelas espécies *Etuberosum* e *Palustre*. Entretanto, as espécies *Verrucosum* e *Brevicaule* deixaram de formar um clado e passaram a atuar como grupo monofilético. Outras diferenças constatadas são a monofilia da espécie *Albornozii* e a formação do clado entre as espécies *Andreanum* e *Raphanifolium*.

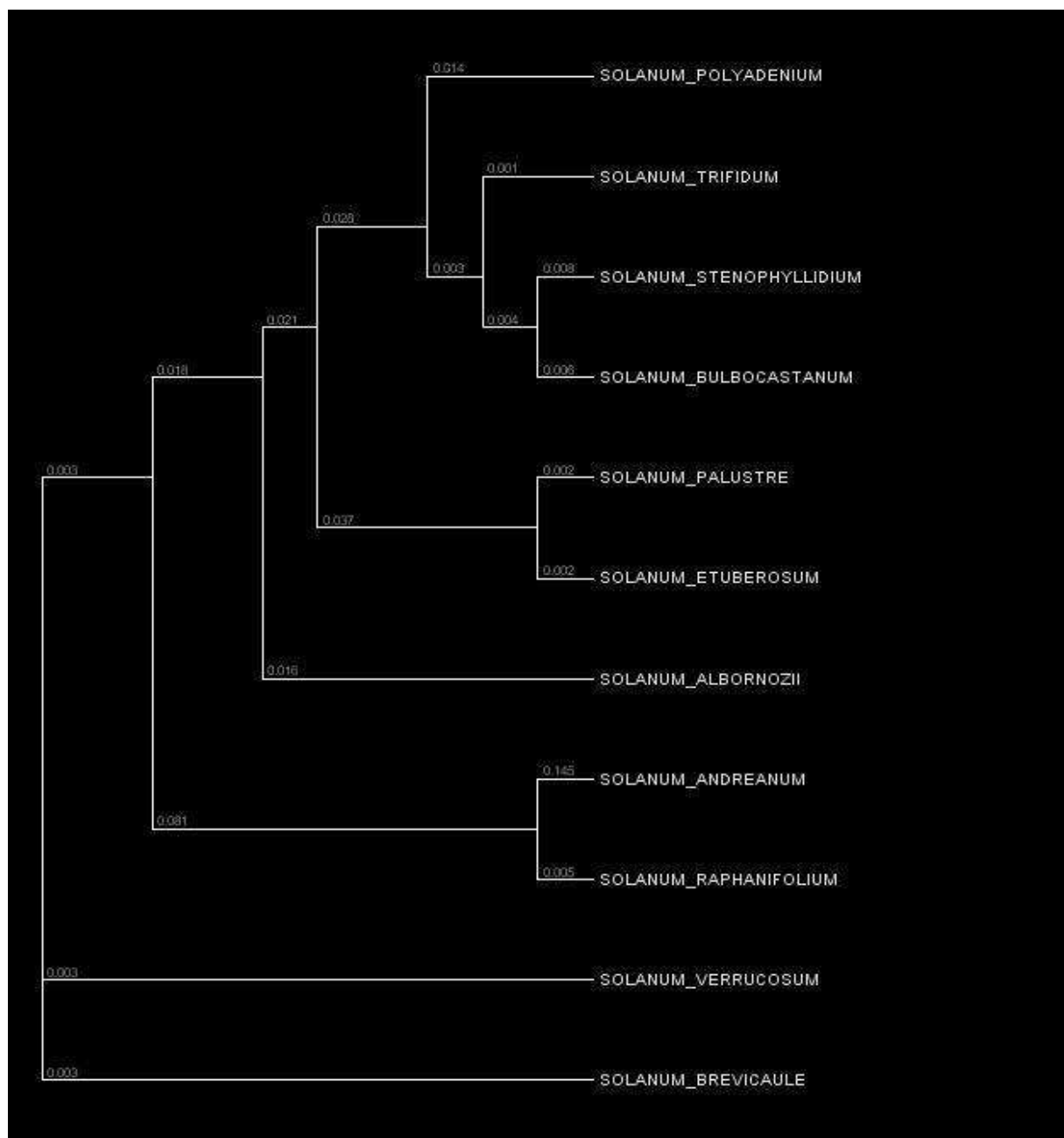


Figura 52 - Resultado da filogenia da *Solanum* com seqüências do gene COSII\_At5g14320 utilizando o método de Máxima Parcimônia, mais precisamente, o DNAPARS (PHYLIP), através do IgrafuWeb. Os números nas linhas indicam o comprimento dos ramos.

Já para o método de Máxima Verossimilhança, as buscas realizadas foram concentradas na utilização do modelo HKY85 e com o valor de número 100 para o bootstrap. Após a análise, foram obtidas as árvores filogenéticas geradas, e estas foram visualizadas através do applet, gerando posteriormente algumas imagens. Dentre as várias árvores geradas, a que tem mais semelhança com a do artigo (Figura 50) é a da Figura 53. Vale ressaltar, que os números contidos nos nós são do teste de confiança (bootstrap) e representam a frequência que os grupos foram amostrados da forma

apresentada nos ramos. Em relação ao resultado da Figura 50, pode-se citar as mudanças do clado formado pelas espécies *Brevicaule* e *Verrucosum*, para *Brevicaule* e *Raphanifolium*, e também as alterações relacionadas com as espécies *Palustre* e *Etuberosum*, as quais deixaram de formar um clado e passaram a atuar como grupo monofilético. Entretanto, constata-se que o grupo monofilético formado pelas espécies *Stenophyllidium*, *Bulbocastanum*, *Polyadenium* e *Trifidum* foi conservado, assim como, o clado formado pelas espécies *Albornozii* e *Andreanum*. O resumo da execução do PHYML pode ser observado na Tabela 9, a qual mostra o resultado dos parâmetros de execuções utilizados.

Tabela 9 - Resumo dos parâmetros de execução do PHYML após o processamento no IgrafuWeb.

<b>Modelo Evolutivo</b>	HKY85
<b>Número de Taxas</b>	11
<b>Logaritmo de Probabilidade</b>	-1423.26752
<b>Probabilidade Irrestrita</b>	-1791.80346
<b>Parcimônia</b>	88
<b>Busca da topologia da árvore</b>	NNIs
<b>Árvore inicial</b>	BIONJ
<b>Tamanho da árvore</b>	0.14967
<b>Taxa de Transição/Transversão</b>	2.788
<b>Modelo de Gama Discreto</b>	
<b>Número de Classes</b>	4
<b>Forma do parâmetro Gama</b>	0.244
<b>Taxa relativa da classe 1</b>	0.00185
<b>Taxa relativa da classe 2</b>	0.06243
<b>Taxa relativa da classe 3</b>	0.48878
<b>Taxa relativa da classe 4</b>	3.44695
<b>Frequências de Bases</b>	
<b>A</b>	0.24707
<b>C</b>	0.17046
<b>G</b>	0.20982
<b>T</b>	0.37264

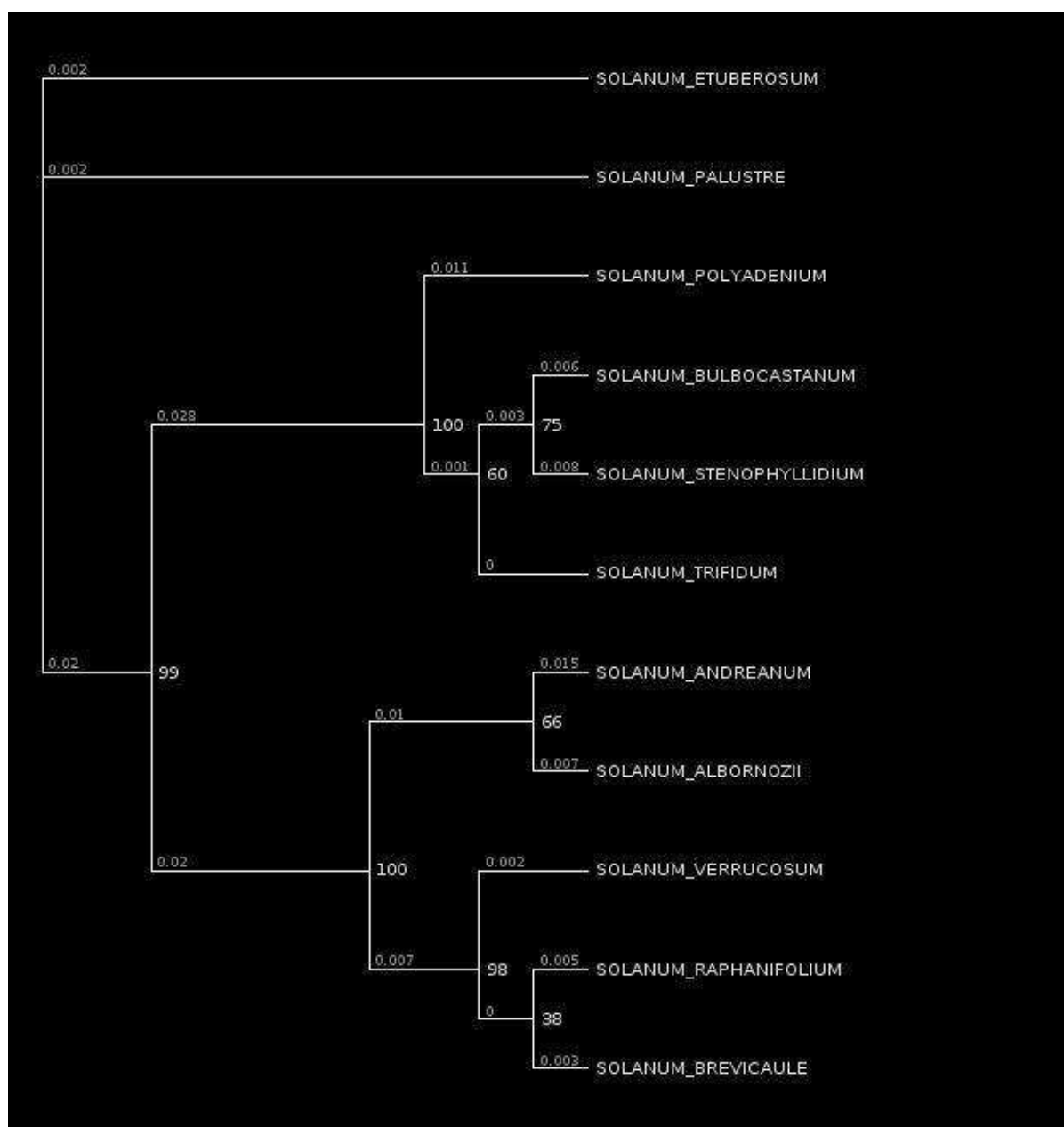


Figura 53 - Resultado da filogenia da *Solanum* com seqüências do gene COSII\_At5g14320 utilizando a máxima verossimilhança, mais precisamente, o PHYML através do IgrafuWeb. Os números nos nós indicam valores de bootstrap.

O MrBayes foi executado várias vezes através do IgrafuWeb, resultando em diversas filogenias, as quais foram analisadas e serão discutidas a seguir. As execuções procederam de forma normal, seguindo as configurações padrões do programa: o modelo evolutivo GTR e o valor de 1.000.000 para o número de ciclos. Após o término do processo de execução, o MrBayes disponibiliza uma série de arquivos de saída contendo informações referentes ao estado de execução do algoritmo, probabilidade posterior das árvores iteradas, as árvores filogenéticas e os parâmetros do modelo evolutivo utilizado. Um resumo destas informações pode ser observado através da análise



da Tabela 10. Como a execução foi definida em 1 milhão de ciclos, foi apresentado apenas os valores médios, mínimos e máximos dos parâmetros processados. O item 4.4 apresentou em detalhes o significado destes parâmetros.

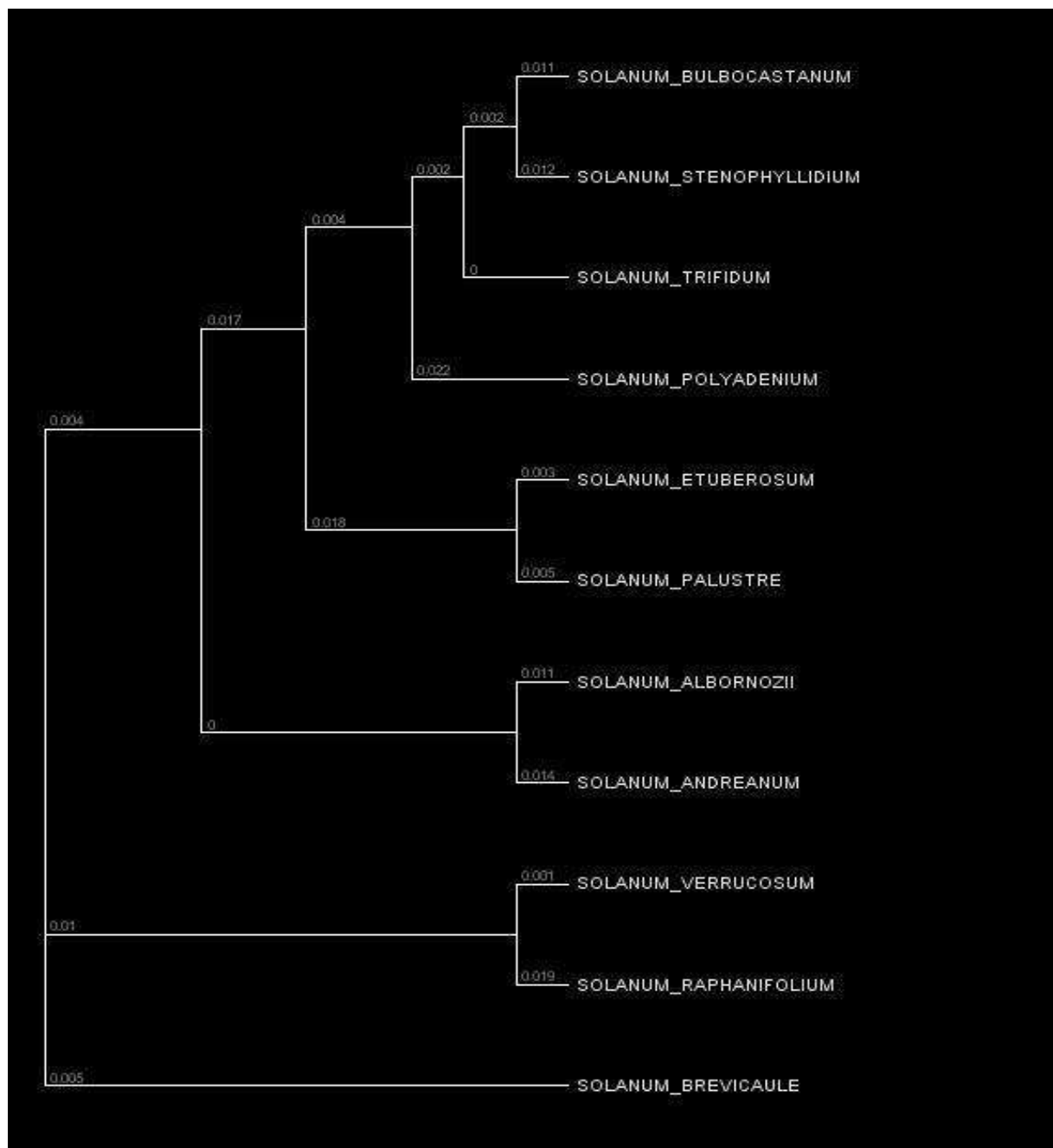


Figura 54 - Resultado da filogenia da *Solanum* com seqüências do gene COSII\_At5g14320 utilizando a inferência bayesiana, mais precisamente, o MrBayes através do IgrafuWeb. Os números nos galhos indicam os comprimentos dos ramos.

A filogenia apresentada pelo MrBayes, através do IgrafuWeb, é exibida na Figura 54, na qual foram encontradas algumas diferenças pontuais que serão analisadas a seguir.

Na formação dos clados, a única diferença encontrada é na espécie *Verrucosum* que forma um clado com *Raphanifolium*, e não com *Brevicaule* como exibido na figura original do artigo. Ademais, todos os clados são os mesmos, alternando apenas os números dos comprimentos dos ramos e os seus respectivos ancestrais hipotéticos. Vale ressaltar, que estas comparações foram realizadas de forma visual, sendo que o comparativo numérico serão feitos mais a frente neste capítulo.

Tabela 10 - Resumo dos parâmetros de execução do MrBayes após o processamento no IgrafuWeb.

	LnL	TL	r(A<->C)	r(A<->G)	r(A<->T)	r(C<->G)	r(C<->T)	r(G<->T)
<b>Média</b>	-1,43E+09	1,40E+05	3,63E+04	4,77E+05	1,66E+05	4,43E+04	8,93E+04	1,70E+05
<b>Mínimo</b>	-2383586000	-2,4E+09	858,5205	858,5205	40831,75	204,9582	204,9582	11214,04
<b>Máximo</b>	-1,31E+09	1,90E+06	1,76E+05	7,14E+05	3,93E+05	1,96E+05	2,49E+05	4,24E+05
	pi(A)	pi(C)	pi(G)	pi(T)	Alpha	Pinvar		
<b>Média</b>	1,48E+05	6,46E+05	4,96E+04	1,53E+05	3,44E+07	9,98E+05		
<b>Mínimo</b>	43588,84	46412,85	5398,69	5398,69	62510,24	717523,8		
<b>Máximo</b>	2,78E+05	7,67E+05	2,50E+05	4,18E+05	7,47E+07	1,00E+06		

O modelo evolutivo utilizado na filogenia apresentada pelo MrBayes foi o GTR, o qual leva em consideração a taxa de substituição de um nucleotídeo para cada um dos demais, levando-se em conta a sua frequência e taxa de substituição. Além disso, a Inferência Bayesiana emprega o cálculo de verossimilhança e, quando as informações a priori não influenciam significativamente, os resultados concordam com os da Máxima Verossimilhança.

O MrBayes apresentou soluções intermediárias não condizentes com a filogenia definida como parâmetro de teste. Para cada critério utilizado, o IgrafuWeb foi capaz de apresentar árvores alternativas que compõe o escopo de melhores soluções. Sendo assim, as árvores disponibilizadas foram percorridas e analisadas, a fim de encontrar a que mais se encaixe no perfil desejado.

Todas as execuções apresentadas até aqui podem ser processadas pelo leitor a fim de comprovar a veracidade das informações. Para isto, basta submeter o arquivo de sequência apresentado (observar notas de rodapé 21, 22 e 23), escolher os parâmetros de acordo ao tipo de método escolhido, e iniciar o processo. Após o final da execução, o usuário tem a possibilidade de visualizar a(s) árvore(s) filogenética(s) resultante(s), além dos arquivos de saída para cada software analisado.

Como existem várias formas de representar as relações de evolução para um determinado conjunto de dados, gera-se o problema de dedução de qual delas é a melhor. Tal problema pode ser

resolvido através de testes estatísticos que indiquem qual árvore é melhor de acordo a determinado critério, ou se mais de uma árvore explica igualmente bem os dados. Esses testes são aplicados através de softwares que comparam topologias e comprimentos dos ramos de árvores filogenéticas. Sendo assim, com o objetivo de quantificar as semelhanças e diferenças entre as árvores citadas, foi necessário aplicar a comparação da árvore referência (Figura 50) com todas as árvores resultantes do processamento do IgrafuWeb. Essa comparação foi realizada através do software Ktreedist (SORIA-CARRASCO et al. 2007).

O Ktreedist faz o comparativo da topologia de duas árvores retornando um valor  $K$  que determina o grau de diferenças entre estas. Quanto maior o valor de  $K$ , maior é a diferença entre a árvore referência e a comparada. Logo, este comparativo foi realizado levando-se em conta a árvore referência (Figura 50) e as árvores a serem comparadas que representam a execução do IgrafuWeb para cada software do seu escopo. Este resultado encontra-se na Tabela 11.

Tabela 11 - Valor de  $K$  retornado pelo Ktreedist ao comparar a árvore referência (Figura 50) com todas as árvores resultantes do processamento do IgrafuWeb.

	$K$
PHYML	22.01846
Digrafu	22.01894
MrBayes	25.87989
DNAPARS	26.49937

De acordo aos dados da Tabela 11, pode-se concluir que a árvore retorna pelo PHYML é a mais congruente possível da árvore original apresentada no trabalho do RODRIGUEZ et al. (2009), visto que apresenta o menor valor de  $K$ .

O comparativo dos comprimentos dos galhos pode ser feito através da medida do grau de dissimilaridade. Estas medidas analisam o conjunto  $(P_1, P_2, \dots, P_n)$  de todas as partições (representação dos nós terminais divididos por galhos) possíveis de uma árvore, sendo que para cada árvore pode-se definir um vetor  $B$  de valores reais não negativos  $(b_1, b_2, \dots, b_n)$ , onde  $b_i$  é o tamanho do galho correspondente à partição  $P_i$ . Quando uma partição não existir em uma árvore, o valor 0 é atribuído para o seu galho  $b_i$ . (GONÇALVES, 2007).

Desta maneira, para mensurar a dissimilaridade de galhos entre 2 árvores, obtêm-se os respectivos vetores de tamanho de partições ( $B$  e  $B'$ ), calcula-se o somatório do quadrado da

diferença entre eles, e por fim aplica-se a raiz quadrado deste valor final (GONÇALVES, 2007). A Equação (78) ilustra este cálculo.

$$Dissimilaridade(B, B') = \sum_{j=1}^n (b_i - b'_i)^2 \quad (78)$$

Neste sentido, os valores de  $b_i$  foram calculados através do Ktreedist, substituídos na equação acima, e assim obteve-se o valor de dissimilaridade dos comprimentos dos galhos da árvore referência comparadas com os resultados do IgrafuWeb (Tabela 12).

Tabela 12- Comparativo dos tamanhos dos galhos.

	Dissimilaridade Galhos
DNAPARS	26.88415
PHYML	26.88693
Digrafu	26.8908
MrBayes	26.9014

Sendo assim, analisando os valores da Tabela 12, constata-se que os comprimentos dos galhos apresentados pelo resultado do DNAPARS teve o menor valor de dissimilaridade expressando pouca divergência comparada com a árvore de referência. No entanto, o valor apresentado pelo MrBayes foi o maior, indicando uma divergência maior entre os comprimentos dos galhos.

Considerando-se todas as hipóteses e resultados analisados, obtidos através do IgrafuWeb, para apresentar as filogenias pelos métodos de Distância, Máxima Parcimônia, Máxima Verossimilhança e Inferência Bayesiana, pode-se afirmar a semelhança das topologias das árvores. No entanto, algumas alterações foram constatadas, e as afirmações a seguir podem explicar as mudanças encontradas.

Nesse sentido, vale ressaltar, que as informações disponibilizadas no trabalho do RODRIGUEZ et al. (2009), relacionadas com as configurações dos parâmetros filogenéticos utilizados pelos softwares citados para construir a filogenia apresentada pela Figura 50, são insuficientes para possibilitar uma execução semelhante utilizando o IgrafuWeb. Seria relevante processá-lo aplicando todas as configurações utilizadas no trabalho do RODRIGUEZ et al. (2009) para construir a filogenia em questão. Isso implicaria em resultados mais semelhantes. Além do mais, tem a questão da diferença dos softwares utilizados: apenas o MrBayes é comum em ambos os trabalhos.

Além disso, as sequências genéticas foram editadas e intervenções manuais foram realizadas no ato do alinhamento do trabalho do RODRIGUEZ et al. (2009). Tais alterações não foram realizadas nas sequências submetidas ao IgrafuWeb. Com isso, os alinhamentos podem ter apresentados resultados distintos, implicando em alterações na filogenia.

Mesmo com as diferenças apontadas entre as filogenias analisadas, esses resultados indicam a veracidade filogenética dos dados utilizados, confirmando a ideia difundida pela utilização e qualificação do IgrafuWeb. Logo, pode-se afirmar a consistência do software apresentado neste projeto, visto que, resultou em uma árvore correta de acordo aos dados de entrada, além de mostrar eficiência e robustez, permitindo a compreensão dos fatores envolvidos por cada método em tempo hábil.

## 8 CONCLUSÕES E TRABALHOS FUTUROS

A partir das análises de alguns projetos e softwares de RAF, percebeu-se a necessidade de produzir soluções mais simples e acessíveis, com capacidade de poder de processamento, para realizar pesquisas relacionadas à evolução das espécies. Deste modo, foram disponibilizados neste trabalho, serviços Web para inferir filogenias em supercomputadores, tendo como dados de entrada, sequências de DNA ou de Proteínas, devidamente alinhadas.

Neste sentido, a principal contribuição deste trabalho de pesquisa foi proporcionar a pesquisadores a oportunidade de realizar filogenias, em um ambiente Web de alta performance, composto dos principais métodos de inferência filogenética (Distância, Máxima Parcimônia, Máxima Verossimilhança e Inferência Bayesiana), com possibilidade de visualização gráfica das árvores resultantes. Analisando de forma mais específica, foram desenvolvidas rotinas que possibilitam a execução dos programas de inferência filogenética MrBayes, PHYML, Digradu ou DNAPARS/PROTPARS, em uma plataforma de computadores de alto desempenho (CACAU), através de um sistema Web, o IgraduWeb.

Para atingir os objetivos deste trabalho foi necessário apresentar uma revisão bibliográfica a respeito da problemática da filogenética, da modelagem matemática utilizada pelos modelos de evolução, bem como das implementações e modelagem dos métodos de inferência filogenética, possibilitando ao leitor a compreensão e a utilização do sistema resultante deste trabalho.

Procurou-se utilizar o que há de mais simples e rápido na implementação dos algoritmos do sistema. Para isso, utilizou-se como instrumento de trabalho um framework de alta performance em PHP, próprio para grandes aplicações Web, que conseguiu atender todos os requisitos necessários para o desenvolvimento do projeto, o Yii. Todas as páginas do sistema foram resultantes da análise da modelagem dos softwares citados de acordo às características de cada um.

Um dos focos deste projeto foi direcionado para difundir a utilização de RAF em ambiente Web com alta performance. Isto só foi possível graças ao poder de processamento disponibilizado pelo cluster da UESC, CACAU. Sendo assim, foi necessário realizar uma série de estudos para entender a logística de submissão de trabalhos a serem executados e troca de arquivos entre os servidores que compõe este supercomputador. Nele também foram instalados e configurados todos os softwares de RAF aqui mencionados.

Uma seção desta dissertação foi destinada para apresentar os trabalhos correlatos a este projeto de pesquisa. Constatou-se a partir desse estudo, que os Web Services analisados não possuem

uma metodologia padrão de desenvolvimento, sendo observado deficiências significativas relacionadas à disponibilização de campos dos softwares estudados. Por outro lado, as virtudes encontradas serviram de suporte para incrementar e fortalecer adequadamente a construção do sistema. Logo, constata-se a relevância das ferramentas analisadas de forma a auxiliar a compreensão de um sistema real de RAF na Web, desencadeando em um produto simples e ao mesmo tempo eficiente.

Para avaliar a performance e a qualidade da(s) árvore(s) filogenética(s) retornadas pela execução do sistema produzido neste trabalho, foi realizado um estudo comparativo com uma solução de filogenia conhecida no ambiente científico, resultado de estudos apresentado no trabalho RODRIGUEZ et al. (2009), que permitiu concluir que o resultado final do IgrafuWeb apresenta soluções consistentes e significativas.

Espera-se que as contribuições apresentadas, para resolução de problemas de filogenias, possa despertar interesse aos pesquisadores da área de bioinformática e que novas soluções possam ser desenvolvidas a partir dos conceitos definidos neste trabalho. Em suma, pode-se afirmar que a pesquisa proposta foi realizada de acordo às suas metas, visto que o IgrafuWeb tem obtido resultados condizentes com o esperado.

Este projeto de pesquisa proporciona uma abordagem considerada principiante em âmbito nacional, para RAF em um ambiente Web, utilizando computadores de alta performance. Com isto, pode-se afirmar que muito ainda pode ser feito para aperfeiçoar as pesquisas nesta linha de trabalho. Sendo assim, para trabalhos futuros, com base no assunto desta dissertação, pode-se propor:

- Incorporação de um módulo para identificar e sugerir o melhor método de inferência filogenética a utilizar de acordo aos dados de entrada (sequências genéticas), além do modelo evolutivo e seus respectivos parâmetros;
- Criação de perfis de usuário mediante autenticação (login e senha): ao realizar o cadastro o pesquisador seria avaliado, e posteriormente liberado para utilização do sistema. Isto permitiria a possibilidade de criação de uma rotina para visualização do histórico de submissão daquele determinado usuário. Com isso, os seus trabalhos de filogenias realizados poderiam ser editados, consultados e/ou deletados. Além do mais, abre-se também a prerrogativa de envio por email das soluções filogenéticas apresentados pelo IgrafuWeb;
- Inserção de rotina para avaliar, mediante critérios de testes, a qualidade e as informações apresentadas pelo IgrafuWeb.

## REFERÊNCIAS

- ALVES, J. M. P. Caracterização e Filogenia Moleculares de *Acanthamoeba*. Dissertação (Tese de Doutorado) – Universidade de São Paulo (USP), São Paulo – SP, Brasil, 2001.
- AMORIM, D. S. Fundamentos de sistemática filogenética. 1. ed, Ribeirão Preto, SP. Holos, 2002.
- BOC, A.; DIALLO, A. B.; MAKARENKO, V. T-REX: a Web server for inferring, validating and visualizing phylogenetic trees and networks, *Nucleic Acids Research*, v. 40, 2012.
- BRUNO, W.; SOCCI, N.; HALPERN, A. Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction, *Molecular biology and evolution*, v. 17, e. 1, p. 189–197, 2000.
- CRISCUOLO, A. morePhyML: improving the phylogenetic tree space exploration with PhyML 3. *Molecular Phylogenetics and Evolution*, 2011.
- CUNHA, M. C. C. Métodos Numéricos, e. 2, p. 276, Campinas: Editora Unicamp, 2003.
- CYBIS, G. B. Teste da Razão de Verossimilhança e seu Poder em Árvores Filogenéticas. Dissertação (Mestrado em Matemática) - Universidade Federal do Rio Grande do Sul - UFRGS, Rio Grande do Sul - RS, Brasil, Julho 2009.
- DEREEPER, A.; GUIGNON, V.; BLANC, G.; AUDIC, S.; BUFFET, S.; CHEVENET, F.; DUFAYARD, J. F.; GUINDON, S.; LEFORT, V.; LESCOT, M.; CLAVERIE, J. M.; GASCUEL, O. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, 2008.
- DESPER, R.; GASCUEL, O. Fast and accurate phylogeny reconstruction algorithms based on the Minimum-Evolution principle, *Journal of computational biology*, v. 9, n. 5, p. 687–705, 2002.
- ECK, R. V.; DAYHOFF, M. O. Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Spring, MD, 1966.
- EWENS, W. J.; GRANT, G. R. Statistical Methods in Bioinformatics-An Introduction. Springer-Verlag New York, Inc. 2001.
- FARRIS, J. Estimating phylogenetic trees from distance matrices. *American Naturalist*, v. 106, n. 951, p. 645-668, 1972.
- FELSENSTEIN, J. Evolutionary trees from DNA sequences: A maximum like-lihood approach. *Journal of Molecular Evolution*, v. 17, p. 368-376, 1981.
- FELSENSTEIN, J. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, v. 39, n. 4, p. 783-791, 1985.



- FELSENSTEIN, J. PHYLIP (Phylogeny Inference Package) version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle, 1993.
- FELSENSTEIN, J.; CHURCHILL, G. A. A hidden Markov model approach to variation among sites in rate of evolution. *Molecular biology and evolution*, v. 13, p. 93–104, 1996.
- FELSENSTEIN, J. *Inferring Phylogenies*. Sunderland, Massachusetts: Sinauer, 2004.
- FITCH, W. M. Toward defining the course of evolution: minimum change for a specific tree topology. *Systematic Zoology*, v. 20, n. 4, p. 406–416, 1971.
- GASCUEL, O. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution*, v. 14, e. 7, p. 685–695, 1997.
- GENTLE, J. E. *Statistics and Computing - Random number generation and Monte Carlo methods*, c. 7, p. 231, 2007.
- GILKS, W. R.; RICHARDSON, S.; SPIEGELHALTER, D. J. *Markov chain Monte Carlo in practice*. Chapman and Hall, London, UK, ed. 1996.
- GOLOBOFF, P. A.; FARRIS, J. S.; NIXON, K. TNT: Tree analysis using New Technology, *Systematic Biology*, v. 54, n. 1, p. 176–178, 2005.
- GONÇALVES, G. D. *Análise de desempenho de programas para reconstrução de árvores filogenéticas*. Ilhéus: UESC - Universidade Estadual de Santa Cruz, 2007.
- GONÇALVES, G. D. *Estudo de técnicas para melhorar o desempenho da reconstrução de árvores filogenéticas por análise bayesiana*. Ilhéus: UESC - Universidade Estadual de Santa Cruz, 2008.
- GRIMMENT, G. R.; STIRZAKER, D. R. *Probability and Random Processes*. Oxford University Press Inc., New York. 1992.
- GUINDON, S; GASCUEL, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, v. 52, e. 5, p. 696–704, 2003.
- HASEGAWA, M.; KISHINO, H.; YANO T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22:160–174. 1985.
- HAN, M. V.; ZMASEK, C. M. phyloXML: XML for evolutionary biology and comparative genomics *BMC, Bioinformatics*, 10:356, 2009.
- HENDY, M. D.; PENNY D. Branch and bound algorithms to determine minimal evolutionary trees, *Mathematical Biosciences*, v. 60, p. 133–142, 1982.
- HUELSENBECK, J. P.; F. RONQUIST. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, v. 17, e. 8, p. 754–755, 2001.

HYPÓLITO, E. B. Uma resposta bayesiana ao Paradoxo de Suzuki. Dissertação (Mestrado em Estatística) - Instituto de Matemática da Universidade Federal do Rio de Janeiro, Rio de Janeiro – RJ, Brasil, p. 3-35, Abril 2005.

IDALINO, R. C. L. Homologias em genes relacionados à resistência à mastite em vacas, ovelhas e cabras. Dissertação (Mestrado em Biometria e Estatística Aplicada) – Universidade Federal Rural de Pernambuco, Recife – PE, Brasil, Dezembro 2010.

KIMURA, M. A simple model for estimating evolutionary rates of base sub-stitutions through comparative studies of nucleotides sequences. *Journal of Molecular Evolution*, v. 16, p. 111-120, 1980.

LARGET, B.; SIMON, D. L. Markov Chain Monte Carlo Algorithms or the Bayesian Analysis of Phylogenetic Trees. *Molecular biology and evolution*, v. 16, n. 6, p. 750-579, 1999.

LARKIN, M. A.; BLACKSHIELDS, G.; BROWN, N. P.; CHENNA, R.; MCGETTIGAN, P. A.; MCWILLIAM, H.; VALENTIN, F.; WALLACE, I. M.; WILM, A.; LOPEZ, R.; THOMPSON, J. D.; GIBSON, T. J.; HIGGINS, D. G. Clustal W and Clustal X version 2.0. *Bioinformatics*, v. 23, e. 21, p. 2947-2948, 2007.

MENDES, R. Reconstrução Filogenética. Disponível em: <<http://rodrigomendes.tripod.com/bioinformatics.html>>. Acesso em: 4/2/2015

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, v. 21, p. 1087-1092, 1953.

MUELLER, L. D.; AYALA, F. J. Estimation and interpretation of genetic distance in empirical studies. *Genetical Research*, v. 40, e. 2, p. 127-37, 1982.

NÉRON, B.; MÉNAGER, H.; MAUFRAIS, C.; JOLY, N.; MAUPETIT, J.; LETORT, S.; CARRERE, S.; TUFFERY, P.; LETONDAL, C. Mobyle: a new full Web bioinformatics framework. *Bioinformatics*, 2009.

OLSEN, G. J.; MATSUDA, H.; HAGSTROM, R.; OVERBEEK, R. fastDNAm1: A tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.*, v. 10, e. 1, p. 41-48, 1994.

PEDROSA, L. A. Comparação entre distância entrópica e distância genética para análise de sequências de dna. Dissertação (Mestrado em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco (UFRPE), Recife – PE, Brasil, Julho 2013.

PINTO, J. F. C. Epidemiologia molecular do vírus da imunodeficiência humana do tipo I: Métodos de Inferência Filogenética. Dissertação (Mestrado em Saúde Pública) - Escola Nacional de Saúde Pública Sérgio Arouca (Fundação Oswaldo Cruz), Rio de Janeiro – RJ, Brasil, Julho 2004.

PRADO, O. G. Computação evolutiva empregada na reconstrução de Árvores Filogenéticas. Dissertação (Mestrado em Engenharia Elétrica) - Faculdade de Engenharia Elétrica e de Computação (FEEC/Unicamp), Campinas – SP, Brasil, Dezembro 2001.

RANWEZ, V; GASCUEL, O. Improvement of distance based phylogenetic methods by a local maximum likelihood approach using triplets. *Molecular biology and evolution*, v. 19, p. 1952–1963, 2002

RODRIGUEZ, F.; WU, F.; ANÉ, C.; TANKSLEY, S.; SPOONER, D. M. Do potatoes and tomatoes have a single evolutionary history, and what proportion of the genome supports this history? *BMC Evolutionary Biology*, v. 9, 2009.

SAITOU, N.; NEI, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, v. 4, e. 4, p. 406–425, 1987.

SÁNCHEZ, R.; SERRA, F.; TÁRRAGA, J.; MEDINA, I.; CARBONELL, J.; PULIDO, L.; MARÍA, A.; CAPELLA-GUTIÉRREZ, S.; HUERTA-CEPAS, J.; GABALDÓN, T.; DOPAZO, J.; DOPAZO, H. Phylemon 2.0: a suite of Web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic acids research*, 2001.

SILVA, A. E. A. Uma abordagem multi-objetivo e multimodal para reconstrução de Árvores Filogenéticas. Dissertação (Tese de Doutorado) - Faculdade de Engenharia Elétrica e de Computação (FEEC/Unicamp), Campinas – SP, Brasil, Dezembro 2007.

SORIA-CARRASCO, V.; TALAVERA, G.; Igea, J.; CASTRESANA, J. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees, v. 23, e. 21, p. 2954-2956, 2007.

SOUZA, D. M. Modelos ocultos de Markov: uma Abordagem em Controle de Processos, trabalho de conclusão de curso (Graduação em Estatística), Universidade Federal de Juiz de Fora, Juiz de Fora – MG, Brasil, 2013.

SNEATH, P.H.; SOKAL, R.R. Numerical taxonomy: the principles and practice of numerical classification. San Francisco: W. H. Freeman, 1973.

STAMATAKIS, A. Phylogenetic Models of Rate Heterogeneity: A High Performance Computing Perspective. In: Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International, 2006.

SWOFFORD, D. L.; OLSEN, G. J.; WADELL, P. J.; HILLIS, D. M. Phylogeny inference. *Molecular systematics*. Sinauer, Sunderland, MA, p. 407-514, 1996.

SWOFFORD, D. L. PAUP\* Phylogenetic Analysis Using Parsimony, CSIT Florida State University, 2000.

TICONA, W. G. C. Algoritmos evolutivos multi-objeto para reconstrução de árvores filogenéticas. USP/São Carlos, 2008.

TORRES, M.; DIAS, G.; GONÇALVES, G.; VIEIRA, C. Tool that Integrates Distance Based Programs for Reconstructing Phylogenetic Trees. *Revista IEEE América Latina*, v. 9, e. 5, 2011.

TRIOLA, M. F. Elementary Statistics. In: 9th edition. New York USA: Addison Wesley, p. 259-267, 2003.

YANG, Z. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular biology and evolution*, v. 10, n. 6, p. 1396–1401, 1993.

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. A Space-time process model for the evolution of DNA sequences. *Genetics*, 139:993–1005, 1995.

YANG, Z. Among-site rate variation and its impact on phylogenetic analyses. *TREE*, v. 11, n. 9, p. 367-372, 1996.

YANG, Z. *Computational Molecular Evolution*. New York: Oxford University Press, 2007.

VIANA, G. V. R. Técnicas para Construção de Árvores Filogenéticas. Dissertação (Tese de Doutorado) — Universidade Federal do Ceará - UFC, Fortaleza - CE, Brasil, Abril 2007.

WEIR, B. S. *Genetic Data Analysis II*. Sinauer, Sunderland, MA, 1996.