UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

INSTITUTO DE MATEMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM MATEMÁTICA

Teste da Razão de Verossimilhança e seu Poder em Árvores Filogenéticas

Dissertação de Mestrado

Gabriela Bettella Cybis

Porto Alegre, 20 de Julho de 2009.

Dissertação submetida por Gabriela Bettella Cybis[1] como requisito parcial para a obtenção do grau de Mestre em Matemática pelo Programa de Pós-Graduação em Matemática do Instituto de Matemática da Universidade Federal do Rio Grande do Sul.

Professora Orientadora:
Drª. Sílvia Regina Costa Lopes

Banca Examinadora:
Dr. Alexandre Tavares Baraviera
Dr. Artur Oscar Lopes
Drª. Hildete Prisco Pinheiro

Data da Defesa: 10 de Agosto de 2009.

[1]Conselho Nacional de Desenvolvimento Científico e Tecnológico

ÍNDICE

Agradecimentos

Gostaria de expressar meu agradecimento aos meus pais, Helena Beatriz Bettella Cybis e Luiz Fernando de Abreu Cybis por todo o apoio e incentivo desde o início até a conclusão desta jornada. Vocês são minha base e meus modelos de vida.

À minha orientadora Dr^{a}. Sílvia Regina Costa Lopes pela confiança em mim depositada, pelas críticas e elogios nos momentos em que eles foram merecidos, e pelo incentivo de continuar batalhando. Teu papel foi fundamental para a construção de quem sou hoje.

À Dr^{a}. Hildete P. Pinheiro pela ajuda na definição dos rumos do trabalho, por todas as conversas extremamente ricas sobre o seu significado, e pela amizade.

Aos amigos Sergio Freitas, Pablo Guedes e Renan Machado pela ajuda com programação.

Aos professores do Programa de Pós-Graduação em Matemática da UFRGS pelo comprometimento na busca da qualidade de trabalho e conhecimento.

Aos colegas usuários do LCPM, pelo auxílio e amizade.

À banca examinadora pela atenção.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico pelo auxílio financeiro.

À Universidade Federal do Rio Grande do Sul pela oportunidade.

E ao meu namorado, por tudo.

Resumo

Muitas questões de importância biológica ligadas à evolução e filogenias podem ser abordadas por meio de análises estatísticas que utilizam a função de verossimilhança. Para que tais análises sejam feitas, são necessários modelos que designem probabilidades a eventos mutacionais, os modelos de substituição de bases. Como existem diversos destes modelos, critérios estatísticos para escolher entre eles são importantes nessa area. Assim, o objetivo desse trabalho é estudar as propriedades de um dos mais amplamente utilizados critérios para seleção desses modelos, o teste da razão de verossimilhança. Utilizando teoria assintótica, nós propomos um estimador consistente e de baixo custo computacional para o poder do teste. Nós também utilizamos Simulações de Monte Carlo para estudar a distribuição da estatística do teste. Além disso, estudamos propriedades dos estimadores de máxima verossimilhança para parâmetros do modelo, como sua distribuição assintótica em casos particulares e cotas inferiores para sua variância. A técnica do Jackknife é utilizada para a correção do vício destes estimadores, com bons resultados. Os modelos de substituição de bases mais utilizados tem pressupostos restritivos sobre o processo de evolução molecular, assim, nós também estudamos alguns modelos mais realistas que permitem variação das taxas de mutação e dependência entre sítios.

Abstract

Many important biological questions, especially regarding evolutionary history and phylogenies, may be tackled by statistical analysis of DNA sequences that use the likelihood function. In order to make these analysis, statistical models that assign probabilities to mutational events are needed. Since several of these base substitution models exist, statistical methods that choose between models are important in this field. The goal of this dissertation is to study the properties of likelihood ratio tests that compare base substitution models, the most widely used hypothesis tests for this purpose. Through asymptotic theory, we propose a low computational cost consistent estimator for the power of the likelihood ratio test. We also study the distribution of the test's statistic through Monte Carlo simulations. Properties of the maximum likelihood estimators for model parameters, such as its asymptotic distribution in special cases and lower bounds for it's variance are presented, as well as the suggestion of using the resampling method Jackknife for bias correction. The widely used base substitution models have strong assumptions about the process of evolution that are not generally valid, thus we also study models that alow for rate variation and dependency between sites.

Capítulo 1

Introdução

Neste trabalho estudamos o teste da razão de verossimilhança como ferramenta estatística para escolher entre diferentes modelos que explicam a evolução temporal de seqüências de DNA. O contexto no qual esta evolução se dá é o das árvores filogenéticas que relacionam as seqüências. Os modelos em questão são os chamados modelos de substituição de bases, que designam probabilidades ao tipo mais simples de mutação que ocorre no DNA, as substituições de bases. Através das mutações e filogenias, pode-se explicar a história de um gene, desde um único ancestral comum, em algum momento do passado, até as seqüências homólogas do gene em diversos organismos vivos hoje.

Diversas são as informações de interesse biológico que podem ser extraídas deste processo. Relação de parentesco entre grupos de organismos, histórias evolutivas de certos genes, tempo que separa uma espécie de seu mais recente ancestral comum e pressões evolutivas importantes na evolução das seqüências são alguns exemplos destas informações que podem ser obtidas a partir de conjuntos de seqüências de DNA. A maioria das ferramentas estatísticas utilizadas para obter estas informações fazem uso dos modelos de substituição de bases para o cálculo de probabilidades de eventos mutacionais. Em muitos casos, a utilização de modelos inapropriados pode levar a conclusões erradas acerca da evolução das seqüências. Assim, destaca-se a importância da abordagem utilizada neste trabalho para escolher, dentre os modelos de evolução das seqüências, o mais adequado.

No Capítulo 2, são apresentados alguns conceitos preliminares necessários para todo o trabalho. Estas informações são fundamentais para a compreensão de diversos modelos tratados nos capítulos seguintes, que muitas vezes são baseados em problemas biológicos específicos. As justificativas de diversos testes de hipóteses estudados neste trabalho residem no conhecimento de alguns detalhes técnicos, apresentados na seção de conceitos biológicos.

Além disso, no Capítulo 2, apresentamos os modelos de substituição de bases, que atribuem à evolução temporal de cada sítio do DNA uma cadeia de Markov. O que distingue esses modelos são os pressupostos biológicos para a evolução da seqüência. Entretanto, todos eles tem em comum o pressuposto de que os sítios evoluem independen-

temente, e todos de acordo com as mesmas taxas de mutação. Estes modelos servem de base para toda a teoria trabalhada no restante desta dissertação.

No terceiro capítulo, é introduzido o cálculo da função de verossimilhança para seqüências de DNA relacionadas por uma filogenia. Em seguida, é apresentado o teste da razão de verossimilhança para comparar modelos de substituição de bases. Estudamos um teorema que dá a distribuição assintótica da estatística do teste e verificamos que ele é válido para os modelos de substituição de bases apresentados no Capítulo 2. Consideramos, também, algumas críticas ao uso desta distribuição e apresentamos o método do bootstrap paramétrico como alternativa para obter a distribuição da estatística do teste. Observamos que o teste da razão de verossimilhança será o fio condutor do restante do trabalho.

Neste capítulo, também sugerimos dois métodos para a obtenção do poder o teste. O primeiro método é o bootstrap paramétrico, que geralmente é utilizado para obter o valor crítico, mas pode ser alterado para fornecer o poder do teste. Em seguida, nos reportamos à distribuição assintótica teórica para propor um segundo método para obter o poder do teste, um estimador baseado no estimador de máxima verossimilhança e na informação de Fisher observada. Este estimador possui a vantagem de, em uma área em que a maioria das análises requer grande custo computacional, não exigir cálculos adicionais além daqueles já realizados para a aplicação do teste. Além disso, nesse capítulo, estudamos algumas propriedades estatísticas do poder do teste e seu estimador. Note que o estudo do poder do teste da razão de verossimilhança que compara modelos de substituição de bases é um aspecto praticamente não contemplado na literatura.

Ao final do Capítulo 3 apresentamos, ainda, duas aplicações da teoria deste capítulo a seqüências reais. As mesmas seqüências são novamente utilizadas para aplicação de testes comparando alguns modelos apresentados no Capítulo 4.

O Capítulo 4 relaxa diversos pressupostos restritivos em relação à evolução dos diferentes sítios da seqüência de DNA. Consideramos modelos que permitem diferentes taxas de mutação nos sítios, designando distribuições para as taxas de mutação como a gama e a multinomial. Estudamos também um modelo que assume dependência entre as taxas de mutação nos diferentes sítios, através de uma cadeia de Markov oculta. Finalmente, consideramos alguns casos de modelos que assumem dependência entre os processos mutacionais que ocorrem nos diferentes sítios da seqüência, no caso de alguns problemas biológicos específicos. Assim, o quarto capítulo representa uma busca por modelos mais realistas para a evolução das seqüências.

Como no Capítulo 4 são relaxados alguns pressupostos utilizados para estabelecer os resultados apresentados no Capítulo 3, para cada um destes novos modelos, a teoria é revista, e alguns ajustes necessários são feitos.

Para a realização do teste da razão de verossimilhança, são necessários os estimadores de máxima verossimilhança para os parâmetros dos modelos de substituição de bases.

No quinto capítulo, estudamos propriedades estatísticas destes estimadores. Inicialmente, consideramos o caso mais simples, em que temos apenas duas seqüências. Nesta caso, a filogenia pode ser ignorada, de forma que obtém-se resultados acerca da distribuição assintótica destas estatísticas. Ainda na linha de redução das dificuldades impostas pela estrutura filogenética ao estudo dos estimadores de máxima verossimilhança, impomos algumas hipóteses restritivas, que nos permitem propor cotas inferiores para a variância dos estimadores. Utilizamos simulações computacionais para avaliar o comportamento destas cotas para um dos modelos de substituição de bases estudados.

Além disso, nesse capítulo, sugerimos uma abordagem de Jackknife para correção do vício dos estimadores de parâmetros do modelo e comprimentos dos ramos da filogenia. Essa abordagem é utilizada para propor uma nova versão mais precisa do algoritmo para a obtenção do poder do teste pelo bootstrap paramétrico.

O sexto capítulo é dedicado a avaliar os métodos estudados ao longo do trabalho por meio de simulações. Nele, utilizamos as simulações de Monte Carlo para obter informações importantes em relação à distribuição da estatística do teste e ao seu poder. Por meio das simulações, investigamos os tamanhos de amostra necessários para que a estatística do teste assuma os comportamentos assintóticos descritos no Capítulo 3. Além disso, apresentamos comparações entre o comportamento dos dois métodos propostos no Capítulo 3 para obter o poder do teste.

Este capítulo também apresenta um estudo para verificar a viabilidade da utilização do método Jackknife na correção do vício dos estimadores dos parâmetros dos modelos de substituição de bases, conforme sugerido no Capítulo 5.

Finalmente, no sétimo capítulo apresentamos as conclusões obtidas nesse trabalho. Além disso, comentamos alternativas para futuros trabalhos que complementam os resultados dessa dissertação.

Capítulo 2

Conceitos Preliminares

2.1 Conceitos Biológicos

O presente trabalho aborda técnicas matemáticas e estatísticas utilizadas para a análise de seqüências homólogas de DNA. Para uma compreensão dos objetivos e justificativas das análises aqui realizadas, é necessário o conhecimento de conceitos fundamentais de biologia molecular. Assim, nessa seção, serão apresentados alguns desses conceitos, juntamente com algumas informações sobre o processamento do DNA para que se possam utilizar as seqüências.

2.1.1 Seqüências de DNA

O DNA, ácido desoxirribonucléico, está presente em todas as células vivas[1] que se tem notícia. Ele é a molécula responsável pela informação genética que é transmitida para os descendentes. Nele, está contida quase toda a informação necessária para a formação de um ser vivo. Desse modo, estudando o DNA, podemos obter informações preciosas sobre o presente e o passado da vida na Terra. Mas para isso, primeiro precisamos conhecer algumas informações sobre o DNA.

A unidade básica do DNA é o *nucleotídeo*, que é uma molécula formada por uma pentose (açúcar), um grupo fosfato e uma base nitrogenada. A molécula de DNA é composta de duas fitas em forma de hélice, cada qual um polímero linear constituído destes nucleotídeos. Existem 4 tipos de nucleotídeos no DNA, que são distinguidos por suas bases nitrogenadas: *Adenina* (A), *Guanina* (G), *Citosina* (C) e *Timina* (T). É na seqüência das bases que está toda a informação do DNA. As bases ainda podem ser divididas em dois grupos: as *purinas*, A e G, cujas moléculas contém dois anéis carbônicos,

[1] Alguns vírus não tem sua informação genética em DNA, mas sim em RNA. Entretanto, vírus são seres acelulares, e há controvérsia sobre sua classificação como seres vivos. A posição mais aceita é que vírus não são seres vivos.

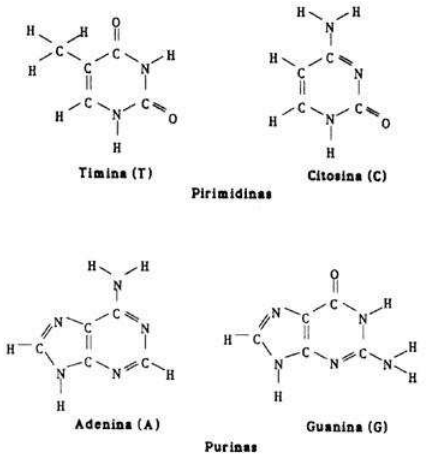


Figura 2.1: Estrutura Química das Bases do DNA.

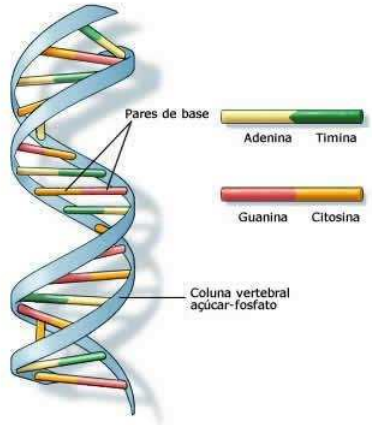


Figura 2.2: Representação de Dupla Hélice da Molécula de DNA.

e as *pirimidinas*, C e T, cujas moléculas contém apenas um anel carbônico (ver Alberts et al. 2004). A Figura 2.1 apresenta a estrutura química das bases púricas e pirimídicas.

As duas fitas do DNA são unidas através das bases, de forma complementar (A com T e C com G). Tal estrutura está representada na Figura 2.2. Assim, se em uma posição (sítio) da fita molde temos um A, certamente na mesma posição da fita complementar teremos um T. Essa característica faz com que uma fita contenha uma seqüência exatamente complementar à da outra. Desse modo, a estrutura de dupla fita é uma forma de conservar a informação genética, já que de apenas uma das fitas pode-se sintetizar uma nova molécula de DNA idêntica à anterior. Além disso, como as fitas são complementares, para a análise da molécula de DNA, precisamos apenas da seqüência de bases de uma delas.

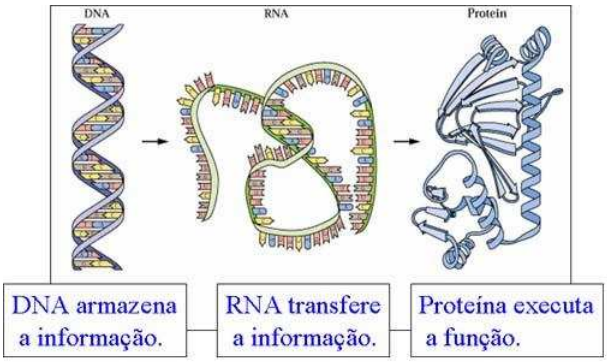


Figura 2.3: Dogma Central da Biologia Molecular.

Do DNA à Proteína

Mas como seqüências de bases podem determinar a transformação de uma única célula (zigoto) em um ser humano? De que forma o DNA age na construção do ser vivo? Nas seqüências de DNA encontramos os *genes*, que são porções do DNA que contêm a informação para a construção de proteínas, codificada nas seqüências das quatro bases. Essa informação é traduzida (copiada) para moléculas de RNA, que atuam como mensageiros temporários, levando a informação do DNA até os ribossomos, que são os responsáveis pela síntese proteica. É através das proteínas que a informação genética age sobre os organismos. Elas têm, além da função estrutural, a função de catalisar quase todas as reações que ocorrem no organismo.

A seqüência de eventos dada por informação genética contida no DNA, transcrita para o RNA mensageiro, e traduzida para proteína nos ribossomos, é conhecida como o dogma central da biologia molecular, por ser praticamente onipresente e fundamental para quase todos os processos celulares. Ela está representada na Figura 2.3.

Os ribossomos lêem as moléculas de RNA de três em três bases. O conjunto de três bases na seqüência é chamado de *códon*, e cada códon corresponde a um aminoácido na proteína. Proteínas são polímeros lineares de aminoácidos. É a ordem desses aminoácidos, e as interações químicas entre eles, que determinam a conformação tridimensional da proteína e, consequentemente, sua função. Assim, o estudo da seqüência de aminoácidos de uma proteína também é de grande importância e existem diversos trabalhos nessa área.

Os códons são compostos de três sítios, e existem apenas quatro diferentes bases no DNA. Assim, temos ao todo 64 diferentes códons. Por outro lado, encontramos em proteínas apenas 20 aminoácidos diferentes. Dessa forma, temos vários códons que codificam o mesmo aminoácido. Além disso, temos três códons de parada (que encerram a síntese proteica). A Figura 2.4 apresenta uma tabela com a correspondência entre

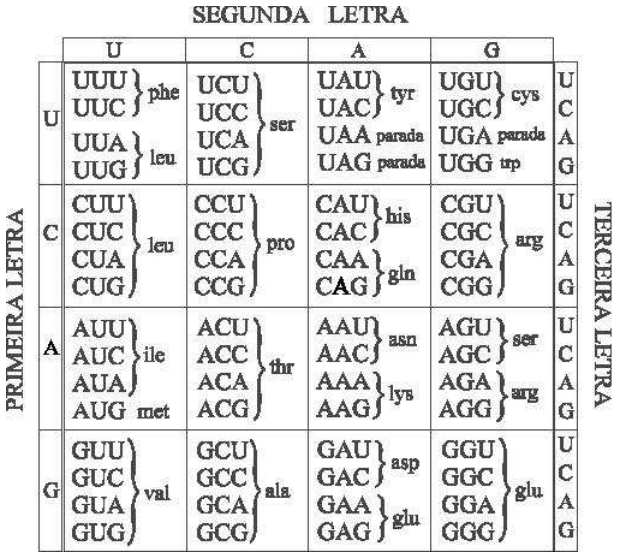


Figura 2.4: Código Genético.

códon e aminoácido, conhecida como *código genético*. Os códons na Figura 2.4 estão representados no RNA, assim, temos a substituição da base do DNA Timina (T) pela base do RNA Uracila (U).

Diz-se que o código genético é conservado. Uma característica é dita *conservada* se permanece inalterada em um grupo grande de organismos. Isso significa que não sofre muitas alterações na história evolutiva.

Note que, em geral, quando um aminoácido é codificado por vários códons, os primeiros dois nucleotídeos praticamente não variam, e a distinção entre os códons está no terceiro nucleotídeo. Assim, freqüentemente nos genes, os primeiros dois sítios dos códons são mais conservados do que terceiro.

Mutações

Conforme dito anteriormente, é através do DNA que a informação genética é transmitida para os descendentes. Essa transferência de informação se dá pela *replicação* do DNA, em que uma das fitas é utilizada como molde para síntese de uma nova molécula de DNA idêntica à anterior. Assim, quando a célula se divide, deixa uma cópia do seu DNA para cada descendente. Existem diversos mecanismos celulares que agem no sentido de garantir que as cópias sejam o mais perfeitas possível, além de evitar alterações na seqüência em outros momentos que não a replicação. Entretanto, devido ao enorme tamanho do *genoma* (informação genética total contida no DNA de uma célula), que nos humanos é de 3.2×10^7 pares de bases (ver Zaha et al., 2003), às vezes ocorrem alterações nas seqüências.

Qualquer alteração na seqüência de DNA é chamada de *mutação* e pode ser causada por

diversos fatores, como erro no processamento da molécula, vírus, exposição a substâncias mutagênicas, luz ultravioleta, entre outros. Existem diversos tipos de mutações como *adições* (em que trechos de DNA são inseridos no meio da seqüência), *deleções* (em que um pedaço da seqüência é perdido), *transposições* (em que um trecho do DNA é removido da seqüência e colocado em outro lugar do DNA) e *substituições de bases* (em que uma base é substituída por outra na mesma posição, sem alterar o comprimento da seqüência).

Existem modelos probabilísticos que abordam vários desses tipos de mutação, mas nesse trabalho consideramos apenas as mutações do tipo substituição de bases. Essas mutações são de abordagem mais direta e são as mais utilizadas para estudos filogenéticos.

2.1.2 Árvores Filogenéticas

Acredita-se que todos os seres vivos encontrados hoje na Terra descendem de um único ancestral comum. Assim, qualquer grupo de seres vivos pode ser relacionado por uma árvore filogenética (que representa os graus de parentesco entre eles). Em geral, organismos mais parecidos tem ancestrais comuns mais recentes, e isso é refletido na representação da filogenia.

Note, entretanto, que não temos muita informação direta sobre o passado das espécies. Assim, resta apenas inferir as filogenias. Para isso, além do registro fóssil (o qual tem registro de apenas alguns eventos isolados) temos apenas as informações que podemos obter dos organismos vivos hoje.

A classificação dos seres vivos em grupos, objeto de estudo da Sistemática, foi uma das primeiras preocupações da Biologia. Já a utilização de árvores filogenéticas é tão antiga quanto a teoria da evolução, que estabeleceu a necessidade da busca pelo parentesco. Originalmente, utilizavam-se caracteres anatômicos dos organismos para agrupá-los. Foi só no início da década de 1960 que os trabalhos de Sokal e Sneath inauguraram o uso de métodos numéricos na taxonomia (ver Sokal e Sneath, 1963). Nessa mesma época, com os avanços da Biologia Molecular, iniciou-se o uso de informações moleculares, como freqüências gênicas e seqüências proteicas, para a inferência de filogenias (ver Felsenstein, 2004).

Filogenias podem ser representadas de diversas formas. Entretanto, nesse trabalho, utilizaremos preferencialmente as filogenias dicotômicas (em que a separação das linhagens ocorre de dois em dois). A Figura 2.5 apresenta um exemplo de filogenia dicotômica com raiz. A raiz é a indicação de qual ponto da filogenia está o mais recente ancestral comum a todos os grupos da filogenia.

Quando tratando com filogenias, chamamos os organismos dos quais colhemos amostras de *nós externos da filogenia*, e os numeramos de 1 a N, em que N é o tamanho da amostra (número de organismos considerados). No caso da filogenia da Figura 2.5, as bolas pretas numeradas de 1 a 4 representam os nós externos. Os últimos ancestrais comuns entre duas

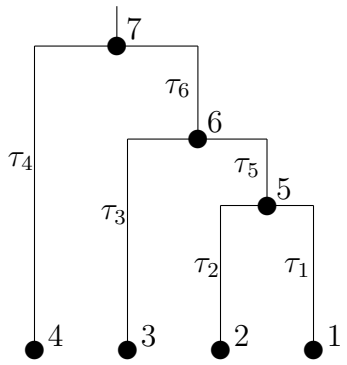


Figura 2.5: Exemplo de Árvore Filogenética.

linhagens são chamados de *nós internos da filogenia*, e numerados de $N+1$ até $2N-1$. No caso da Figura 2.5, os nós internos são representados pelas bolas pretas enumeradas de 5 a 7. O nó $2N-1$ é chamado de *raiz da filogenia*. A ligação entre dois nós é chamada de *ramo da árvore filogenética*, e representa o tempo evolutivo que separa um organismo de seu ancestral. Os ramos também são numerados de 1 a $2N-2$, e seus comprimentos são denotados por τ_l, $l \in \{1, \cdots, 2N-1\}$. No caso da Figura 2.5, temos seis ramos cujos comprimentos são denotados por $\tau_1, \cdots, \tau_6$.

Na reconstrução de filogenias, devem ser estimados a topologia da árvore filogenética (qual a relação de proximidade entre as linhagens) e os comprimentos dos ramos. Note que os comprimentos dos ramos representam um conjunto de parâmetros contínuos que devem ser todos positivos. Já a topologia da filogenia é de parametrização mais complicada. Assim, como nesse trabalho a construção da filogenia não é o principal foco, assumiremos, em geral, que a sua topologia é conhecida.

Existem diversos métodos de estimação de filogenias. Entre os mais utilizados estão os métodos por matriz de distâncias, por parcimônia e por máxima verossimilhança (ver Felsenstein, 2004). Todos eles têm vantagens e desvantagens. Entretanto, nesse trabalho preferimos a estimação por máxima verossimilhança. Apesar de exigir um tempo computacional maior do que os outros métodos, em estudos de simulações ele é o que apresenta maior eficiência na recuperação da filogenia original (ver Felsenstein, 2004).

Uma das formas mais utilizadas de representação de filogenias é o formato "newick", em que os grupos são separados por vírgulas e unidos por parênteses. Assim, a filogenia da Figura 2.5 é representada como "$(((1,2),3),4)$;". A essa representação ainda podem ser adicionados os comprimentos dos ramos, que são separados das linhagens por dois pontos, dessa forma, a mesma árvore pode ser escrita como "$(((1:\tau_1, 2:\tau_2):\tau_5, 3:\tau_3):\tau_6, 4:\tau_4)$;". Além de representar cada árvore de forma única, o formato "newick" tem a vantagem de ser de fácil interpretação. Em todas as aplicações e simulações computacionais

realizadas nesse trabalho, utilizamos esse formato para as árvores.

2.1.3 Preparação das Seqüências

Os métodos apresentados nesse trabalho têm como objetivo desvendar aspectos do passado evolutivo das seqüências de DNA. Para que isso possa ser realizado, trabalha-se com *seqüências homólogas*. Um conjunto de seqüências é dito *homólogo* se elas representam descendentes de uma seqüência ancestral em organismos diferentes. Em geral, seqüências homólogas são obtidas ao se estudar um mesmo gene (ou mesma região do DNA) em organismos diferentes.

Devido aos processos de mutação, em geral, seqüências homólogas não são idênticas. Um mesmo gene em espécies diferentes pode ter comprimentos diferentes, devido, por exemplo, às mutações do tipo adição e deleção. Entretanto, conforme comentado anteriormente, nesse trabalho consideramos apenas as mutações do tipo substituição de bases. Assim, é interessante eliminar das seqüências os efeitos das adições e deleções, de forma que possamos estabelecer correspondência entre bases que descendem de um mesmo sítio na seqüência ancestral. É só desta forma que podemos estudar isoladamente o processo de substituição de bases.

O processo que estabelece as correspondências entre os sítios das diferentes seqüências homólogas é o *alinhamento*. Note que, antes do alinhamento, em geral dispomos de N seqüências de DNA não necessariamente do mesmo comprimento, em que N é o número de organismos em estudo. No alinhamento é formada uma matriz que tem como linhas as N seqüências da amostra. Os sítios correspondentes são colocados todos na mesma coluna, e algumas bases são removidas (ou espaços em branco são adicionados) para que todas as seqüências fiquem com o mesmo comprimento S.

O alinhamento é um procedimento estatístico que atribui um escore para cada matriz de alinhamento das seqüências. Então, é feita uma busca no espaço de possíveis matrizes por aquela que apresenta o melhor escore. O alinhamento de bio-seqüências (DNA, RNA ou proteínas) é uma área de pesquisa ativa, de forma que existem diversas formas de se definir o escore das matrizes e diversos algoritmos de busca pelo melhor escore.

Nesse trabalho, os alinhamentos foram realizados utilizando-se um dos aplicativos mais amplamente difundidos para esse propósito, o Clustal X, disponível *on-line* em *http://www.clustal.org*. O sistema de escore desse aplicativo leva em consideração o fato de que alguns sítios são mais conservados do que outros, e que as seqüências não são independentes, mas sim relacionadas por uma filogenia. No aplicativo, ela é estimada pelo método de agrupamento de vizinhos ("Neighbour-Joining"; ver Saitou e Nei, 1987). Já a busca pelo melhor escore é realizada utilizando-se um algoritmo heurístico, para a redução do tempo computacional (ver Durbin et al., 2004).

2.2 Processos Estocásticos

Todos os modelos utilizados nesse trabalho envolvem processos estocásticos. Assim, apresentamos nessa seção alguns conceitos básicos de processos estocásticos e suas principais propriedades.

Definição 2.1. (Processo Estocástico). Seja $(\Omega, \mathcal{A}, \mathbb{P})$ um espaço de probabilidade, $(M, \mathcal{G})$ um espaço mensurável, e ainda uma família de variáveis aleatórias X_t indexadas por um parâmetro $t \in T$, onde $T \subseteq \mathbb{R}$ e cada $X_t : (\Omega, \mathcal{A}, \mathbb{P}) \to (M, \mathcal{G})$ é mensurável. Dizemos que $\{X_t\}_{t \in T}$ é um *processo estocástico*.

Observação 2.1. Nesse trabalho, M será sempre finito. Assim, temos que $\mathcal{G}$ é dado pelas partes de M.

Definição 2.2. (Espaço de Estados). O *espaço de estados* é o conjunto M, ou seja, o elenco dos possíveis valores de cada variável aleatória X_t, para $t \in T$.

Definição 2.3. (Espaço de Índices). O conjunto T é dito *espaço de índices* do processo.

O conjunto T possui uma ordem e dizemos que a variável X_t, para cada $t \in T$, descreve o que acontece com o processo no tempo t.

No que tange as seqüências de DNA, estamos interessados em dois processos estocásticos. O primeiro, representado por $\{X_t\}_{t \in T}$, é o processo que designa a combinação de bases em um sítio de todas as seqüências consideradas (uma coluna da matriz $\mathbf{X}$ obtida no alinhamento - ver Seção 2.1.3). O espaço de estados desse processo é o conjunto das possíveis combinações de bases em um sítio das N seqüências consideradas. Assim, temos exatamente 4^N elementos no espaço de estados, já que existem quatro bases no DNA. Este processo designa uma combinação para cada um dos S sítios considerados. Assim, dizemos que $\{X_t\}_{t \in T}$ é um processo à tempo discreto, pois $T \in \{1, 2, \cdots, S\}$. Este é o processo que será considerado para todas as análises realizadas a partir do Capítulo 3. Entretanto, devido a diversas peculiaridades do DNA, incluindo a estrutura filogenética que relaciona as seqüências, as probabilidades dos eventos deste processo não são simples de determinar. Portanto, recorremos a um processo auxiliar.

O processo auxiliar $\{Y_\tau\}_{\tau \in T}$, que será considerado na Seção 2.3, designa a base de um sítio de uma seqüência em cada instante de tempo τ. O espaço de estados desse processo, denotado por E, corresponde às quatro bases do DNA, isto é $E = \{A, G, C, T\}$. Este é um processo à tempo contínuo, em que $\tau \in T$, e $T = [0, \infty)$.

Definição 2.4. (Processo i.i.d.). Fixados $\Omega, \mathcal{A}$ e $\mathbb{P}$, dizemos que o processo $\{X_t\}_{t \in T}$ tomando valores em E (enumerável) e com parâmetro $t \in T \subseteq \mathbb{N}$ é *independente e identicamente distribuído*, denotado por *i.i.d.*, se

- X_t é independente, ou seja, para todo n e toda seqüência $t_1 < t_2 < \cdots < t_n$ e toda seqüência de conjuntos $B_1, B_2, \cdots, B_n$ contidos em M, vale que

$$\mathbb{P}(X_{t_1} \in B_1, X_{t_2} \in B_2, \cdots, X_{t_n} \in B_n) = \mathbb{P}(X_{t_1} \in B_1)\mathbb{P}(X_{t_2} \in B_2)\cdots\mathbb{P}(X_{t_n} \in B_n);$$

- e, para todo $t \in T$, vale

$$\mathbb{P}(X_t \in B) = \mathbb{P}(X_0 \in B),$$

para todo subconjunto B do espaço de estados M.

Na Seção 2.3 e em todo o Capítulo 3 trabalhamos com o pressuposto simplificador de que o processo $\{X_t\}_{t\in T}$ é i.i.d.. Já no Capítulo 4 essa hipótese será relaxada.

Definição 2.5. (Processo Estacionário). Dizemos que o processo $\{X_t\}_{t\in T}$, tomando valores em M, com índice $t \in T$, é *estacionário* se, para cada n, para cada seqüência de índices $t_1 < t_2 < \cdots < t_n$ e para cada seqüência de conjuntos $B_1, B_2, \cdots, B_n$ contidos em M, vale que

$$\mathbb{P}(X_{t_1} \in B_1, X_{t_2} \in B_2, \cdots, X_{t_n} \in B_n) = \mathbb{P}(X_{t_1+h} \in B_1, X_{t_2+h} \in B_2, \cdots, X_{t_n+h} \in B_n),$$

para todo $h \in T$.

Para todos os modelos considerados nesse trabalho assumimos processos estacionários.

2.2.1 Cadeias de Markov

O processo estocástico que mais utilizamos nesse trabalho é o processo de Markov à tempo contínuo. Ele é utilizado para descrever a evolução temporal de biomoléculas, pois é razoável assumir que estas possuem a propriedade de Markov: probabilidade de futuras mutações independe do passado da molécula, dependendo apenas do estado presente. Assim, temos a seguinte definição.

Definição 2.6. (Processo de Markov). Seja $\{X_t\}_{t\in T}$ um processo estocástico com espaço de estados M finito ou enumerável. Dizemos que $\{X_t\}_{t\in T}$ é um *processo de Markov* se vale a condição

$$\mathbb{P}(X_t = j | X_{t_1} = i_1, \cdots, X_{t_n} = i_n) = \mathbb{P}(X_t = j | X_{t_n} = i_n), \tag{2.1}$$

para todo $t, t_1, \cdots, t_n \in T$, e $j, i_1, \cdots, i_n \in M$, sempre que $t_1 < t_2 < \cdots < t_n < t$ e $\mathbb{P}(X_{t_1} = i_1, \cdots, X_{t_n} = i_n) > 0$.

Processos markovianos com espaço de estados discreto ou enumerável são geralmente chamados de cadeias de Markov. Nesse trabalho utilizamos, para descrever a cadeia de Markov a tempo contínuo, sua caracterização pela matriz geradora infinitesimal. A *matriz geradora infinitesimal* é uma matriz em que qualquer linha soma zero.

Definição 2.7. (Matriz Linha Soma Zero). Uma matriz $\mathbf{Q}$ quadrada com entradas reais é dita *linha soma zero* se satisfaz

- $\mathbf{Q}_{i,i} \leq 0$ para todo $i \in M$;
- $\mathbf{Q}_{i,j} \geq 0$ para todo $i \neq j$, $i, j \in M$;
- $\sum_{j \in M} \mathbf{Q}_{i,j} = 0$, para todo $i \in M$.

Cada entrada da matriz $\mathbf{Q}$ representa a taxa infinitesimal de uma dada transição (mudança de estado do processo). Assim, temos que

$$\mathbf{Q}_{i,j} = \lim_{t \to 0} \frac{\mathbb{P}(j|i,t)}{t} \qquad \text{e} \qquad \mathbf{Q}_{i,i} = \lim_{t \to 0} \frac{\mathbb{P}(i|i,t) - 1}{t}, \tag{2.2}$$

em que $\mathbb{P}(j|i,t) \equiv \mathbb{P}(X_t = j | X_0 = i)$ representa a probabilidade de transição do estado i para o estado j no tempo t.

Note que, a partir da matriz geradora infinitesimal, podemos determinar todas as probabilidades de transição de um estado para outro em qualquer tempo.

Teorema 2.1. *As probabilidades de transição em um determinado intervalo de tempo t do processo de Markov $\{X_t\}_{t \in T}$ podem ser determinadas a partir de sua matriz geradora infinitesimal $\mathbf{Q}$, resolvendo a seguinte equação diferencial matricial*

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}, \tag{2.3}$$

em que $\mathbf{P}(t)$ é a matriz das probabilidades de transição no tempo t, de forma que temos $\mathbf{P}_{i,j}(t) = \mathbb{P}(j|i,t) \equiv \mathbb{P}(X_t = j | X_0 = i)$.

A matriz $\mathbf{P}(t)$ é tal que a soma de suas linhas é 1, e é chamada de matriz estocástica.

No restante desse capítulo, apresentamos modelos que detalham o processo de evolução temporal de cada seqüência. Assim, estamos interessados no processo $\{Y_\tau\}_{\tau \in T}$ comentado anteriormente, e nas suas distribuições estacionárias. Dessa forma, temos a seguinte definição.

Definição 2.8. (Vetor Estacionário). Dizemos que o vetor $\mathbf{p}_0$ é o *vetor estacionário* do processo $\{X_t\}_{t \in T}$ se, para todo $t > 0$, temos $\mathbf{p}_0\mathbf{P}(t) = \mathbf{p}_0$.

Teorema 2.2. *Seja $\{X_t\}_{t\in T}$ um processo de Markov com espaço de estados M finito, e seja $\mathbf{Q}$ sua matriz geradora infinitesimal. Então, existe um vetor estacionário $\mathbf{p}_0$ tal que $\mathbf{p}_0\mathbf{Q} = 0$. Além disso, se esse vetor é único, temos que, para qualquer i,*

$$\lim_{t\to\infty} \mathbb{P}(X_t = j|X_0 = i) = \mathbf{p}_{0j}, \tag{2.4}$$

em que $\mathbf{p}_{0j}$ representa a j-ésima componente do vetor $\mathbf{p}_0$.

Esse resultado é importante, pois, nos modelos descritos na Seção 2.3, assumimos que a distribuição inicial dos processos é a própria distribuição estacionária. Como justificativa para esta opção está o fato de que as seqüências em estudo estão evoluindo sob esse processo há muito tempo. Assim, no presente as freqüências dos estados (bases do DNA) já estariam muito próximas da distribuição estacionária.

2.3 Modelos de Substituição de Bases de Seqüências de DNA

A análise de seqüências de DNA tem como objetivo responder a diversas questões relacionadas ao seu processo de evolução e aos organismos de onde elas foram obtidas. Dentre essas questões, podemos ressaltar a reconstrução biogeográfica da história de um grupo de organismos, a busca de relações entre espécies, a detecção de fenômenos como transferência gênica e coevolução e a datação do último ancestral comum a dois organismos. Para abordar essas questões, faz-se necessária uma forma de medir distâncias entre seqüências de DNA ou de encontrar as probabilidades $\mathbb{P}(j|i,\tau)$ de que uma base i mute para outra base j em um determinado intervalo de tempo τ.

Existem diversos modelos para a evolução da cadeia de DNA que permitem a obtenção das distâncias e probabilidades necessárias para fazer as análises mencionadas acima. Dentre estes, destacam-se os modelos de substituição de bases, que são simplificações probabilísticas da evolução de seqüências de DNA.

Os modelos de substituição de bases assumem que o único tipo de mutação que ocorre é a substituição de uma base por outra na mesma posição da seqüência, de forma que adições e deleções são ignoradas. Esta é uma simplificação relevante, que faz com que esses modelos não descrevam de forma adequada seqüências inteiras de DNA. Entretanto, os modelos de substituição de bases se prestam bem às análises moleculares desejadas graças à etapa de alinhamento, que faz parte da preparação das seqüências de DNA (ver Seção 2.1.3).

Os modelos que consideramos inicialmente assumem que os sítios do DNA evoluem de acordo com processos estocásticos independentes e identicamente distribuídos. Existem também modelos que permitem que uma proporção dos sítios sejam invariantes e modelos que consideram uma estrutura de dependência entre os sítios (ver Capítulo 4).

Os modelos que estudamos atribuem ao processo de evolução de cada sítio da seqüência de DNA uma cadeia de Markov, pois assumem que as probabilidades de mutação dependem apenas do estado (base) em que a cadeia se encontra no momento, sendo independente do passado do processo. O espaço de estados desses processos corresponde às quatro bases do DNA, Adenina (A), Guanina (G), Citosina (C) e Timina (T), denotado por $E = \{A, G, C, T\}$. Essas bases estão divididas em dois grupos devido a características químicas, as *purinas* $R = \{A, G\}$ e as *pirimidinas* $Y = \{C, T\}$.

Desta forma, a matriz de taxas infinitesimais desses modelos é dada por

$$\mathbf{Q} = \begin{pmatrix} -q_A & q_{A,G} & q_{A,C} & q_{A,T} \\ q_{G,A} & -q_G & q_{G,C} & q_{G,T} \\ q_{C,A} & q_{C,G} & -q_C & q_{C,T} \\ q_{T,A} & q_{T,G} & q_{T,C} & -q_T \end{pmatrix}, \tag{2.5}$$

onde $q_{i,j}$ é a taxa instantânea de mutação da base i para a base j, e os elementos da diagonal principal são tais que as linhas somam zero, ou seja $q_i = \sum_{i \neq j} q_{i,j}$, para $i, j \in E$.

Como, para todos os modelos considerados, nenhuma das taxas $q_{i,j}$ é nula, temos que a cadeia é irredutível e recorrente positiva, de forma que o processo possui uma distribuição estacionária denotada por

$$\mathbf{p}_0 = (\pi_A, \pi_G, \pi_C, \pi_T), \tag{2.6}$$

onde π_i é a proporção da base i no DNA. Um pressuposto implícito de alguns desses modelos é que o processo se encontra próximo a seu estado de equilíbrio, de forma que a distribuição estacionária $\mathbf{p}_0$ é também a distribuição inicial.

Partindo da matriz (2.5), é possível encontrar as probabilidades $\mathbb{P}(j|i, \tau)$ de mutação da base i para a base j em um intervalo de tempo τ. A matriz de transição $\mathbf{P}(\tau) = (\mathbb{P}(j|i, \tau))_{i,j \in E}$, que contém essas probabilidades, é uma matriz estocástica, de forma que a soma das linhas é 1. Ela pode ser obtida resolvendo a equação diferencial

$$\mathbf{P}'(\tau) = \mathbf{P}(\tau)\mathbf{Q}, \tag{2.7}$$

cuja solução geral é dada por

$$\mathbf{P}(\tau) = \exp(\tau\mathbf{Q}). \tag{2.8}$$

Nem sempre obteremos expressões analíticas para as probabilidades de mutação. Entretanto, estas sempre podem ser obtidas numericamente para um determinado intervalo de tempo τ.

Note que o "tempo" τ, considerado nas probabilidades de mutação, não corresponde necessariamente ao tempo real, mas sim ao "tempo" evolutivo, representado nas filogenias pelo comprimento dos ramos. É possível que dois ramos de uma filogenia tenham comprimentos diferentes, ainda que representem eventos ocorridos simultaneamente.

Observe ainda que, como esses modelos apresentam distribuição estacionária, as probabilidades de mutação $\mathbb{P}(i|j,\tau)$ em grandes intervalos de tempo se aproximam das freqüências das bases na distribuição estacionária, ou seja, $\lim_{\tau\to\infty}\mathbb{P}(i|j,\tau)=\pi_i$, para todo $i,j\in E$.

Note que, quando estudamos seqüências de DNA, temos acesso apenas ao estado presente do processo. Assim, não temos como saber quantas mutações ocorreram na história de um sítio. Por exemplo, se no passado a base de um determinado sítio foi A, e hoje observamos um C, podemos inferir que houve uma transição nesse sítio. Entretanto, não podemos afirmar com certeza se houve uma mutação $A\to C$, pois poderia ter ocorrido a seqüência de mutações $A\to G\to C$. Da mesma forma, se verificássemos que a base de outro sítio foi A no passado e continua sendo A no presente, não podemos afirmar com certeza que não houve mutação nesse sítio, pois a seqüência $A\to T\to C\to A$ apresenta o mesmo resultado. Esse tipo de mutação é chamado de *silenciosa*. O tratamento desse processo com cadeias de Markov, nesse aspecto se mostra adequado, pois $\mathbb{P}(i|j,\tau)$, obtida da matriz $\mathbf{P}$ dá a probabilidade de encontrarmos a base i em um tempo τ em determinado sítio do DNA, dado que no presente encontramos a base j nesse sítio. Tal probabilidade é independente das mutações que ocorreram no meio tempo. Assim, os modelos de substituição de bases corrigem para mutações múltiplas e silenciosas.

2.3.1 Jukes-Cantor (JC69)

O primeiro modelo que será considerado é o modelo Jukes-Cantor, denotado por JC69. Este modelo assume que as substituições de bases de seqüências de DNA ocorrem de maneira Markoviana, que as probabilidades de mutação são todas idênticas e que todas as bases aparecem com igual probabilidade (ver Jukes e Cantor, 1969). Dessa forma, seja $\mathbf{Q}_{\text{JC69}}$ a matriz de taxas infinitesimais do processo Jukes-Cantor e $\mathbf{p}_0$ a distribuição inicial de probabilidades das bases do DNA. Então, temos

$$\mathbf{Q}_{\text{JC69}}=\begin{array}{c}\begin{array}{cccc}A & G & C & T\end{array}\\ \begin{pmatrix}-3\alpha & \alpha & \alpha & \alpha\\ \alpha & -3\alpha & \alpha & \alpha\\ \alpha & \alpha & -3\alpha & \alpha\\ \alpha & \alpha & \alpha & -3\alpha\end{pmatrix}\end{array}, \tag{2.9}$$

em que 3α é a taxa de mutação do processo, e

$$\mathbf{p}_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right). \tag{2.10}$$

A matriz $\mathbf{Q}_{\text{JC69}}$ fornece as taxas infinitesimais para mutações da base i para a base j, sendo i e j pertencentes ao espaço de estados $E = \{A, G, C, T\}$ do processo. Observe que a distribuição inicial das bases também corresponde às freqüências de equilíbrio.

Com base na matriz apresentada acima, podemos encontrar a matriz $\mathbf{P}_{\text{JC69}}(\tau)$ das probabilidades de mutação em um intervalo de tempo τ resolvendo a equação diferencial

$$\mathbf{P'}_{\text{JC69}}(\tau) = \mathbf{P}_{\text{JC69}}(\tau)\mathbf{Q}_{\text{JC69}}. \tag{2.11}$$

Lema 2.1. *Considere o modelo Jukes Cantor, cuja matriz* $\mathbf{Q}_{\text{JC69}}$ *de taxas infinitesimais é dada por (2.9) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.10). Então, a matriz* $\mathbf{P}_{\text{JC69}}(\tau)$ *das probabilidades de transição no tempo* τ *é dada por*

$$\mathbf{P}_{\text{JC69}}(\tau) = \begin{pmatrix} 1-3\alpha_\tau & \alpha_\tau & \alpha_\tau & \alpha_\tau \\ \alpha_\tau & 1-3\alpha_\tau & \alpha_\tau & \alpha_\tau \\ \alpha_\tau & \alpha_\tau & 1-3\alpha_\tau & \alpha_\tau \\ \alpha_\tau & \alpha_\tau & \alpha_\tau & 1-3\alpha_\tau \end{pmatrix}, \tag{2.12}$$

em que

$$\alpha_\tau = \frac{1}{4}(1 - e^{-4\alpha\tau}). \tag{2.13}$$

Demonstração: A matriz $\mathbf{P}_{\text{JC69}}(\tau)$ das probabilidades de mutação no tempo τ pode ser encontrada através da equação diferencial (2.11). A solução desta equação é dada pela matriz $\exp(\tau\mathbf{Q}_{\text{JC69}})$. Notamos que $\tau\mathbf{Q}_{\text{JC69}} = \mathbf{T}\ \mathbf{Diag}(\tau)\ \mathbf{T^{-1}}$, em que $\mathbf{Diag}(\tau)$ é a matriz diagonal dos autovalores de $\tau\mathbf{Q}_{\text{JC69}}$, e $\mathbf{T}$ a matriz com os respectivos autovetores. Então temos $\exp(\tau\mathbf{Q}_{\text{JC69}}) = \mathbf{T}\ \exp(\mathbf{Diag}(\tau))\ \mathbf{T^{-1}}$. Como $\mathbf{Diag}(\tau)$ é matriz diagonal, temos que $\exp(\mathbf{Diag}(\tau))$ é dada pela matriz com o exponencial dos autovalores de $\tau\mathbf{Q}_{\text{JC69}}$ nas entradas da diagonal principal, e todas as outras entradas nulas. Portanto, temos

$$\begin{aligned} \mathbf{P}_{\text{JC69}}(\tau) &= \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix} \times \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{-4\alpha\tau} & 0 & 0 \\ 0 & 0 & e^{-4\alpha\tau} & 0 \\ 0 & 0 & 0 & e^{-4\alpha\tau} \end{pmatrix} \\ &\times \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ -1/4 & 3/4 & -1/4 & -1/4 \\ -1/4 & -1/4 & -1/4 & 3/4 \\ -1/4 & -1/4 & 3/4 & -1/4 \end{pmatrix}, \end{aligned}$$

e assim obtemos (2.12). □

O modelo JC69 foi um dos primeiros modelos de substituição de bases propostos e é o mais simples. Entretanto, este modelo tem hipóteses muito restritivas, que não explicam diversos dos fenômenos observados em seqüências de DNA como as diferentes taxas para tipos de mutações distintos e diferentes proporções entre as bases.

2.3.2 Kimura-2 Parâmetros (K80)

Assim como o modelo JC69, o modelo Kimura-2 Parâmetros (denotado por K80) atribui um processo Markoviano à evolução da seqüência de DNA e assume que as bases têm distribuição uniforme (ver Kimura, 1980). O modelo K80, entretanto, considera o fato de que as bases do DNA estão divididas em dois grupos, *purinas* $R = \{A\,,\, G\}$ e *pirimidinas* $Y = \{C\,,\, T\}$, de acordo com a semelhança química. Desta forma, no modelo K80, há distinção entre *transições* (mutações para a outra base da mesma categoria química) e *transversões* (mudança de *purina* para *pirimidina* ou de *pirimidina* para *purina*)[2]. Portanto, a matriz de taxas apresenta dois parâmetros. Observe que este modelo tem maior apelo biológico, pois esperamos que mutações entre estes compostos assemelhados sejam mais comuns.

A matriz de taxas infinitesimais que define o processo K80 é dada por

$$\mathbf{Q}_{\text{K80}} = \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} -\beta-2\gamma & \beta & \gamma & \gamma \\ \beta & -\beta-2\gamma & \gamma & \gamma \\ \gamma & \gamma & -\beta-2\gamma & \beta \\ \gamma & \gamma & \beta & -\beta-2\gamma \end{pmatrix} \end{array}, \tag{2.14}$$

em que β é a taxa de *transição* e 2γ é a taxa de *transversão*. Note que as bases são ordenadas de forma que *purinas* e *pirimidinas* fiquem reunidas.

Assim como no modelo Jukes-Cantor, a distribuição de equilíbrio das bases do modelo Kimura-2 Parâmetros é homogênea, e as probabilidades de mutação podem ser obtidas resolvendo uma equação similar à (2.11).

Lema 2.2. *Considere o modelo Kimura-2 Parâmetros, cuja matriz* $\mathbf{Q}_{\text{K80}}$ *de taxas infinitesimais é dada por (2.14) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.10). Então, a matriz* $\mathbf{P}_{\text{K80}}(\tau)$ *das probabilidades de transição no tempo* τ *é dada por*

[2]Observe a distinção do uso da palavra *transição* no contexto biológico, em que significa mutação dentro da mesma categoria química, e no contexto de processos estocásticos, em que transição significa uma mudança de um estado qualquer para outro.

$$\mathbf{P}_{\text{K80}}(\tau) = \begin{pmatrix} 1-\beta_\tau-2\gamma_\tau & \beta_\tau & \gamma_\tau & \gamma_\tau, \\ \beta_\tau & 1-\beta_\tau-2\gamma_\tau & \gamma_\tau & \gamma_\tau, \\ \gamma_\tau & \gamma_\tau & 1-\beta_\tau-2\gamma_\tau & \beta_\tau, \\ \gamma_\tau & \gamma_\tau & \beta_\tau & 1-\beta_\tau-2\gamma_\tau. \end{pmatrix}, \tag{2.15}$$

em que β_τ e γ_τ estão relacionados à β e γ através das transformações

$$\begin{aligned} \beta_\tau &= \frac{1}{4}(1+e^{-4\gamma\tau}-2e^{2(\beta+\gamma)\tau}), \\ \gamma_\tau &= \frac{1}{4}(1-e^{-4\gamma\tau}). \end{aligned} \tag{2.16}$$

Demonstração: Ver demonstração do Lema 2.4 e Observação 2.2. □

2.3.3 Kimura-3 Parâmetros (K81)

Outro modelo que consideramos é o modelo Kimura-3 Parâmetros (denotado por K81), que compartilha dos pressupostos do modelo K80 mas apresenta três probabilidades diferentes para as transições da cadeia de Markov (ver Kimura, 1981). Nesse modelo temos distinção de probabilidades para dois tipos de *transversões*: as *transversões* para a base complementar (A-T e C-G) e as *transversões* para a outra base; além de uma probabilidade para as *transições*. Assim como nos modelos JC69 e K80, a distribuição de equilíbrio das bases é homogênea. Desta forma, temos que a matriz de taxas infinitesimais desse modelo é dada por

$$\mathbf{Q}_{\text{K81}} = \begin{pmatrix} -\beta-\delta-\gamma & \beta & \gamma & \delta \\ \beta & -\beta-\delta-\gamma & \delta & \gamma \\ \gamma & \delta & -\beta-\delta-\gamma & \beta \\ \delta & \gamma & \beta & -\beta-\delta-\gamma \end{pmatrix}, \tag{2.17}$$

em que β representa a taxa infinitesimal para *transições*, δ a taxa de *transversões* para a base complementar, e γ a taxa das demais *transversões*.

Lema 2.3. *Considere o modelo Kimura-3 Parâmetros, cuja matriz $\mathbf{Q}_{\text{K81}}$ de taxas infinitesimais é dada por (2.17) e cujo vetor $\mathbf{p}_0$ de probabilidades iniciais é dado por (2.10). Então, a matriz $\mathbf{P}_{\text{K81}}(\tau)$ das probabilidades de transição no tempo τ é dada por*

$$\mathbf{P}_{\mathrm{K81}}(\tau) = \begin{pmatrix} a & \beta_\tau & \gamma_\tau & \delta_\tau \\ \beta_\tau & a & \delta_\tau & \gamma_\tau \\ \gamma_\tau & \delta_\tau & a & \beta_\tau \\ \delta_\tau & \gamma_\tau & \beta_\tau & a \end{pmatrix}, \tag{2.18}$$

onde $a = 1 - (\beta_\tau + \gamma_\tau + \delta_\tau)$*, de modo que a soma das linhas seja 1. Os parâmetros* β_τ*,* γ_τ *e* δ_τ *para o modelo K81 se relacionam com as entradas da matriz* $\mathbf{Q}_{\mathrm{K81}}$ *por*

$$\beta_\tau = \frac{1}{4}\left(1 - e^{-2(\gamma+\delta)\tau} + e^{-2(\gamma+\beta)\tau} - e^{-2(\beta+\delta)\tau}\right),$$

$$\gamma_\tau = \frac{1}{4}\left(1 + e^{-2(\gamma+\delta)\tau} - e^{-2(\gamma+\beta)\tau} - e^{-2(\beta+\delta)\tau}\right),$$

$$\delta_\tau = \frac{1}{4}\left(1 - e^{-2(\gamma+\delta)\tau} - e^{-2(\gamma+\beta)\tau} + e^{-2(\beta+\delta)\tau}\right). \tag{2.19}$$

Demonstração: A matriz $\mathbf{P}_{\mathrm{K81}}(\tau)$ das probabilidades de mutação no tempo τ pode ser encontrada utilizando o mesmo procedimento da demonstração do Lema 2.1. A solução da equação diferencial $\mathbf{P'}_{\mathrm{K81}}(\tau) = \mathbf{P}_{\mathrm{K81}}(\tau)\mathbf{Q}_{\mathrm{K81}}$ é dada pela matriz $\exp(\tau\mathbf{Q}_{\mathrm{K81}})$. Mas como $\tau\mathbf{Q}_{\mathrm{K81}} = \mathbf{T}\ \mathbf{Diag}(\tau)\ \mathbf{T^{-1}}$, em que $\mathbf{Diag}(\tau)$ é a matriz diagonal dos autovalores de $\tau\mathbf{Q}_{\mathrm{K81}}$, e $\mathbf{T}$ a matriz com os respectivos autovetores, temos que $\exp(\tau\mathbf{Q}_{\mathrm{K81}}) = \mathbf{T}\ \exp(\mathbf{Diag}(\tau))\ \mathbf{T^{-1}}$. Portanto, temos

$$\begin{aligned} \mathbf{P}_{K81}(\tau) &= \begin{pmatrix} -1 & 1 & -1 & 1 \\ -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \times \begin{pmatrix} e^{-2(\gamma+\delta)\tau} & 0 & 0 & 0 \\ 0 & e^{-2(\beta+\gamma)\tau} & 0 & 0 \\ 0 & 0 & e^{-2(\beta+\delta)\tau} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \\ &\times \begin{pmatrix} -1/4 & -1/4 & 1/4 & 1/4 \\ 1/4 & -1/4 & -1/4 & 1/4 \\ -1/4 & 1/4 & -1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}, \end{aligned}$$

e assim obtemos (2.18). □

2.3.4 Tamura-Nei (TN93)

Nos modelos que consideramos até o momento havia sempre o pressuposto de que a distribuição das bases na seqüência de DNA é homogênea. Entretanto, na maioria dos casos isso não corresponde à realidade. Em grande parte das seqüências de DNA alguma

das bases aparece com freqüência maior do que as outras, e essa informação pode ser incorporada ao modelo de substituição de bases.

Em 1993, Tamura e Nei propõem um modelo que representa uma extensão do K80 (que assume probabilidades diferentes para *transições* e *transversões*) para a situação em que a distribuição das bases não é homogênea (ver Tamura e Nei, 1993). Denotamos esse modelo por TN93. As taxas infinitesimais são dadas pela matriz

$$\mathbf{Q}_{\text{TN93}} = \begin{pmatrix} c_1 & \alpha_R \frac{\pi_G}{\pi_R} + \gamma\pi_G & \gamma\pi_C & \gamma\pi_T \\ \alpha_R \frac{\pi_A}{\pi_R} + \gamma\pi_A & c_2 & \gamma\pi_C & \gamma\pi_T \\ \gamma\pi_A & \gamma\pi_G & c_3 & \alpha_Y \frac{\pi_T}{\pi_Y} + \gamma\pi_T \\ \gamma\pi_A & \gamma\pi_G & \alpha_Y \frac{\pi_C}{\pi_Y} + \gamma\pi_C & c_4 \end{pmatrix}, \tag{2.20}$$

onde a constante c_k é tal que a soma das linhas seja 0, para todo $k \in \{1, \cdots, 4\}$, π_i é a proporção da base i, $i \in E$, nas seqüências, γ é o parâmetro relacionado à taxa de mutação do processo, $\pi_R = \pi_A + \pi_G$ é a proporção de *purinas* nas seqüências, $\pi_Y = \pi_C + \pi_T$ é a proporção de *pirimidinas* nas seqüências, e α_R e α_Y são parâmetros relacionados às *transições* entre *purinas* e *pirimidinas*, respectivamente.

O vetor de distribuição de probabilidade inicial deste processo é dado por

$$\mathbf{p}_0 = (\pi_A, \pi_G, \pi_C, \pi_T). \tag{2.21}$$

Assim como nos modelos anteriores, o processo é estacionário, de forma que esta também é a distribuição de equilíbrio.

Esse modelo pode ser interpretado em função de dois eventos distintos. O evento do tipo II pode ser visto como o processo mutacional básico, no qual as bases são substituídas proporcionalmente à sua distribuição de equilíbrio no DNA. Nele ocorrem tanto *transversões* quanto *transições*. Este evento tem probabilidade instantânea γ e também pode ser visto em termos de um processo de retirada de bolas de uma urna. A urna contém bolas representando as quatro bases do DNA nas suas proporções dadas pelo vetor $\mathbf{p}_0 = (\pi_A, \pi_G, \pi_C, \pi_T)$. O evento do tipo II corresponde à reposição da base do sítio considerado por outra retirada aleatoriamente da urna. Desta forma, existe a chance da base ser substituída por ela mesma. Já o evento do tipo I tem taxa α_R para *purinas* e α_Y para *pirimidinas* e representa apenas as *transições*. Este evento também pode ser interpretado em termos de um processo de retirada de bolas de uma urna. Neste caso, existem duas urnas, uma contendo apenas bolas representando *purinas* nas proporções encontradas no DNA, dadas por $\left(\frac{\pi_A}{\pi_R}, \frac{\pi_G}{\pi_R}\right)$, e outra contendo apenas bolas das *pirimidinas* nas proporções dadas por $\left(\frac{\pi_C}{\pi_Y}, \frac{\pi_T}{\pi_Y}\right)$. Se a base do sítio em questão é uma *purina*, o evento

do tipo I a repõe por uma base retirada da urna das *purinas*; caso contrário, a reposição é feita a partir da urna das *pirimidinas*.

Lema 2.4. *Considere o modelo Tamura-Nei, cuja matriz* $\mathbf{Q}_{\text{TN93}}$ *de taxas infinitesimais é dada por (2.20) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.21). Então, a matriz* $\mathbf{P}_{\text{TN93}}(\tau)$ *das probabilidades de transição no tempo* τ *é dada por*

$$\mathbf{P}_{\text{TN93}}(\tau) = \begin{pmatrix} c_\tau(A,A) & \beta_\tau(A,G) & \gamma_\tau(A,C) & \gamma_\tau(A,T) \\ \beta_\tau(G,A) & c_\tau(G,G) & \gamma_\tau(G,C) & \gamma_\tau(G,T) \\ \gamma_\tau(C,A) & \gamma_\tau(C,G) & c_\tau(C,C) & \beta_\tau(C,T) \\ \gamma_\tau(T,A) & \gamma_\tau(T,G) & \beta_\tau(T,C) & c_\tau(T,T) \end{pmatrix}, \tag{2.22}$$

em que

$$\gamma_\tau(i,j) = \pi_j(1 - e^{-\gamma\tau}), \tag{2.23}$$

$\forall i,j \in E$, *tais que* i *e* j *sejam de grupos químicos diferentes,*

$$\beta_\tau(i,j) = \pi_j \left(\frac{\pi_Y}{\pi_R} e^{-\gamma\tau} - \frac{e^{-(\alpha_R+\gamma)\tau}}{\pi_R} + 1 \right), \tag{2.24}$$

$\forall i,j \in R$, *e*

$$\beta_\tau(i,j) = \pi_j \left(\frac{\pi_R}{\pi_Y} e^{-\gamma\tau} - \frac{e^{-(\alpha_Y+\gamma)\tau}}{\pi_Y} + 1 \right), \tag{2.25}$$

$\forall i,j \in Y$. *Ainda, temos que a probabilidade de encontrar a mesma base em um determinado sítio após um tempo* τ *é dada por*

$$c_\tau(i,i) = \pi_i \frac{\pi_Y}{\pi_R} e^{-\gamma\tau} + \frac{\pi_j}{\pi_R} e^{-(\alpha_R+\gamma)\tau} + \pi_i \tag{2.26}$$

$\forall i \in R$, *e*

$$c_\tau(i,i) = \pi_i \frac{\pi_R}{\pi_Y} e^{-\gamma\tau} + \frac{\pi_j}{\pi_Y} e^{-(\alpha_Y+\gamma)\tau} + \pi_i, \tag{2.27}$$

$\forall i \in Y$. *Nas duas últimas expressões* j *representa a outra base da mesma categoria química de* i.

Demonstração: A matriz $\mathbf{P}_{\text{TN93}}(\tau)$ das probabilidades de mutação no tempo τ pode ser encontrada através da equação diferencial $\mathbf{P'}_{\text{TN93}}(\tau) = \mathbf{P}_{\text{TN93}}(\tau)\mathbf{Q}_{\text{TN93}}$. A solução desta equação é dada pela matriz $\exp(\tau\mathbf{Q}_{\text{TN93}})$. Inicialmente, notamos que $\tau\mathbf{Q}_{\text{TN93}} = \mathbf{T}\ \mathbf{Diag}(\tau)\ \mathbf{T^{-1}}$, em que $\mathbf{Diag}(\tau)$ é a matriz diagonal dos autovalores de $\tau\mathbf{Q}_{\text{TN93}}$, dada por

$$\mathbf{Diag}(\tau) = \begin{pmatrix} -\gamma\tau & 0 & 0 & 0 \\ 0 & -(\alpha_Y+\gamma)\tau & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -(\alpha_R+\gamma)\tau \end{pmatrix}, \tag{2.28}$$

e $\mathbf{T}$ é a matriz invertível cujas colunas são dadas pelos respectivos autovalores de $\tau\mathbf{Q}_{\text{TN93}}$. Assim, $\mathbf{T}$ é dada por

$$\mathbf{T} = \begin{pmatrix} -\frac{\pi_R}{\pi_Y} & 0 & 1 & -\frac{\pi_G}{\pi_A} \\ -\frac{\pi_R}{\pi_Y} & 0 & 1 & 1 \\ 1 & -\frac{\pi_T}{\pi_C} & 1 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix}. \tag{2.29}$$

Notamos que, como $\mathbf{Diag}(.)$ é matriz diagonal, temos

$$\exp(\mathbf{Diag}(\tau)) = \begin{pmatrix} e^{-\gamma\tau} & 0 & 0 & 0 \\ 0 & e^{-(\alpha_Y+\gamma)\tau} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & e^{-(\alpha_R+\gamma)\tau} \end{pmatrix},$$

e ainda que $\mathbf{P}_{\text{TN93}}(\tau) = \exp(\tau\mathbf{Q}_{\text{TN93}}) = \mathbf{T}\ \exp(\mathbf{Diag}(\tau))\ \mathbf{T^{-1}}$. □

Observação 2.2. Note que os modelos JC69, K80 e F81 são todos casos particulares do modelo TN93. Assim, fazendo a restrição adequada aos parâmetros dessa matriz encontramos as probabilidades de mutação para os demais modelos.

Por exemplo, se considerarmos $\pi_A = \pi_G = \pi_C = \pi_T$ e $\alpha_R = \alpha_Y = 0$ recaímos no modelo JC69, e temos que as entradas α da matriz $\mathbf{Q}_{\text{JC69}}$ correspondem a $\frac{\gamma}{4}$ na matriz $\mathbf{Q}_{\text{TN93}}$. Observe que, com esta restrição, temos que

$$\mathbb{P}(i|j,\tau) = \alpha_\tau = \frac{1}{4}\left(1 - e^{-\gamma\tau}\right) = \frac{1}{4}\left(1 - e^{-4\alpha\tau}\right), \tag{2.30}$$

para $i \neq j$, que coincide com as probabilidades de mutação do modelo JC69 dadas na expressão (2.13).

Dois outros casos particulares do modelo de Tamura-Nei, amplamente utilizados, são os modelos Hasegawa, Kishino e Yano, denotado por HKY85 (ver Hasegawa, Kishino e Yano, 1985) e Felsenstein 1984, denotado por F84 (ver Felsenstein e Churchill, 1996). Enquanto o modelo TN93 tem seis parâmetros, os modelos HKY85 e F84 apresentam cinco parâmetros.

2.3.5 Felsenstein 1981 (F81)

Felsenstein (1981) propôs um modelo que incorpora, de forma simples, a possibilidade de existirem diferentes proporções entre as bases. A matriz de taxas infinitesimais deste modelo, denotado por F81, que tem como caso particular o modelo Jukes-Cantor, é dada por

$$\mathbf{Q}_{\text{F81}} = \begin{pmatrix} b_1 & \alpha\pi_G & \alpha\pi_C & \alpha\pi_T \\ \alpha\pi_A & b_2 & \alpha\pi_C & \alpha\pi_T \\ \alpha\pi_A & \alpha\pi_G & b_3 & \alpha\pi_T \\ \alpha\pi_A & \alpha\pi_G & \alpha\pi_C & b_4 \end{pmatrix}, \tag{2.31}$$

onde a constante b_k é tal que a soma dos elementos da linha k é 0, para todo $k \in \{1, \cdots, 4\}$, π_i representa a proporção da base i na amostra, e α a taxa de mutação. O vetor de distribuição de probabilidade inicial deste processo é dado por

$$\mathbf{p}_0 = (\pi_A, \pi_G, \pi_C, \pi_T). \tag{2.32}$$

Assim como nos modelos anteriores, o processo é estacionário, de forma que esta também é a distribuição de equilíbrio.

Lema 2.5. *Considere o modelo Felsenstein 1981, cuja matriz* $\mathbf{Q}_{\text{F81}}$ *de taxas infinitesimais é dada por (2.31) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.21). Então, a matriz* $\mathbf{P}_{\text{F81}}(\tau)$ *das probabilidades de transição no tempo* τ *é dada por*

$$\mathbf{P}_{\text{F81}}(\tau) = \begin{pmatrix} \alpha_\tau(A,A) & \alpha_\tau(A,G) & \alpha_\tau(A,C) & \alpha_\tau(A,T) \\ \alpha_\tau(G,A) & \alpha_\tau(G,G) & \alpha_\tau(G,C) & \alpha_\tau(G,T) \\ \alpha_\tau(C,A) & \alpha_\tau(C,G) & \alpha_\tau(C,C) & \alpha_\tau(C,T) \\ \alpha_\tau(T,A) & \alpha_\tau(T,G) & \alpha_\tau(T,C) & \alpha_\tau(T,T) \end{pmatrix}, \tag{2.33}$$

em que

$$\alpha_\tau(i,i) = \pi_i + (1-\pi_i)e^{-\alpha\tau},$$

para todo $i \in E$, *e*

$$\alpha_\tau(i,j) = \pi_j\left(1 - e^{-\alpha\tau}\right),$$

para todo $i, j \in E$, $i \neq j$.

Demonstração: Ver demonstração do Lema 2.4 e Observação 2.2. □

2.3.6 Hasegawa, Kishino e Yano (HKY85)

Hasegawa, Kishino e Yano (1985) propõem um modelo que é caso particular do TN93, denotado por HKY85. Neste modelo, de acordo com a notação do modelo TN93, temos que $\frac{\alpha_R}{\alpha_Y} = \frac{\pi_R}{\pi_Y}$, e portanto, as probabilidades dos eventos do tipo I em cada grupo são proporcionais às freqüências das bases daquele grupo. Além disso, a razão $\frac{\alpha_l}{\pi_l}$, com $l \in \{R, Y\}$, é igual para *purinas* e *pirimidinas*. Desta forma, a taxa instantânea de transição para a base j, pertencente à classe l, é dada por

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

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

Assim, a matriz de taxas infinitesimais desse modelo pode ser escrita como

$$\mathbf{Q}_{\text{HKY85}} = \begin{pmatrix} d_1 & \beta\pi_G & \gamma\pi_C & \gamma\pi_T \\ \beta\pi_A & d_2 & \gamma\pi_C & \gamma\pi_T \\ \gamma\pi_A & \gamma\pi_G & d_3 & \beta\pi_T \\ \gamma\pi_A & \gamma\pi_G & \beta\pi_C & d_4 \end{pmatrix}, \tag{2.34}$$

onde a constante d_k é tal que a soma da linha k é 0, para todo $k \in \{1, \cdots, 4\}$, e β e γ são os parâmetros relativos a *transições* e *transversões*, respectivamente.

O modelo HKY85 é a extensão mais natural do modelo K80, e o contém como caso particular, quando qualquer base tem probabilidade $\frac{1}{4}$.

2.3.7 Felsenstein 1984 (F84)

No modelo proposto por Felsenstein em 1984 (ver Felsenstein e Churchill, 1996), denotado por F84, temos que $\alpha_Y = \alpha_R$, de forma que as probabilidades de eventos do tipo I são iguais para *purinas* e *pirimidinas*. Portanto, a matriz de taxas infinitesimais do modelo F84 é dada por

$$\mathbf{Q}_{\text{F84}} = \begin{pmatrix} e_1 & \alpha\frac{\pi_G}{\pi_R} + \beta\pi_G & \beta\pi_C & \beta\pi_T \\ \alpha\frac{\pi_A}{\pi_R} + \beta\pi_A & e_2 & \beta\pi_C & \beta\pi_T \\ \beta\pi_A & \beta\pi_G & e_3 & \alpha\frac{\pi_T}{\pi_Y} + \beta\pi_T \\ \beta\pi_A & \beta\pi_G & \alpha\frac{\pi_C}{\pi_Y} + \beta\pi_C & e_4 \end{pmatrix}, \tag{2.35}$$

onde e_k é tal que a soma dos elementos da linha k é 0, para todo $k \in \{1, \cdots, 4\}$, β é o parâmetro relacionado à taxa de mutação do processo, $\pi_R = \pi_A + \pi_G$ é a proporção de

purinas no DNA, $\pi_Y = \pi_C + \pi_T$ é a proporção de *pirimidinas* no DNA, e α é o parâmetro relacionado à taxa de *transições*. Observe que o modelo K80 também é um caso particular desse modelo.

O modelo F84 é muito parecido com o HKY85, mas apresenta a conveniência de que as suas probabilidades de mutação $\mathbb{P}(i|j,\tau)$ podem ser encontradas explicitamente com a ajuda da seguinte propriedade. Digamos que ocorra uma série de eventos, cujo último é do tipo II. Nesse caso, o passado da cadeia não importa, pois cada base terá uma probabilidade de aparecer proporcional à sua freqüência. Da mesma maneira, se ocorrer uma série de eventos do tipo I, independentemente de quantos forem, a probabilidade de aparecer A no final da cadeia é proporcional à $\frac{\pi_A}{\pi_R}$. Dessa forma, temos que eventos do tipo II anulam o efeito de todos os eventos anteriores, e eventos do tipo I anulam o efeito de eventos do tipo I anteriores.

Lema 2.6. *Considere o modelo Felsenstein 1984, cuja matriz* $\mathbf{Q}_{\mathrm{F84}}$ *de taxas infinitesimais é dada por (2.35) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.21). Então, a matriz* $\mathbf{P}_{\mathrm{F84}}(\tau)$ *das probabilidades de transição no tempo* τ *é dada por*

$$\mathbf{P}_{\mathrm{F84}}(\tau) = \begin{pmatrix} \mathrm{e}_\tau(A,A) & \beta_\tau(A,G) & \gamma_\tau(A,C) & \gamma_\tau(A,T) \\ \beta_\tau(G,A) & \mathrm{e}_\tau(G,G) & \gamma_\tau(G,C) & \gamma_\tau(G,T) \\ \gamma_\tau(C,A) & \gamma_\tau(C,G) & \mathrm{e}_\tau(C,C) & \beta_\tau(C,T) \\ \gamma_\tau(T,A) & \gamma_\tau(T,G) & \beta_\tau(T,C) & \mathrm{e}_\tau(T,T) \end{pmatrix}, \tag{2.36}$$

em que

$$\begin{aligned} \mathrm{e}_\tau(i,i) &= e^{-(\alpha+\beta)\tau} + \frac{\pi_i}{\pi_l} e^{-\beta\tau}(1-e^{-\alpha\tau}) + \pi_i(1-e^{-\beta\tau}) \\ \beta_\tau(i,j) &= \frac{\pi_j}{\pi_l} e^{-\beta\tau}(1-e^{-\alpha\tau}) + \pi_j(1-e^{-\beta\tau}) \\ \gamma_\tau(i,j) &= \pi_j(1-e^{-\beta\tau}), \end{aligned} \tag{2.37}$$

para $i,j \in E$*, em que* l *representa a categoria química,* R *ou* Y*, a qual* j *pertence.*

Demonstração: Suponhamos que deseja-se obter a probabilidade de uma mutação de A para G. Para esse modelo, temos que a probabilidade de que em um tempo τ não ocorra nenhum evento é dada por

$$\exp(-(\alpha+\beta)\tau). \tag{2.38}$$

Da mesma forma, a probabilidade de que ocorra pelo menos um evento do tipo I e nenhum do tipo II é dada por

$$(1-\exp(-\alpha\tau))\exp(-\beta\tau); \tag{2.39}$$

e a probabilidade de que ocorra pelo menos um evento do tipo II é dada por

$$1 - \exp(-\beta\tau). \tag{2.40}$$

Com base nas expressões (2.38), (2.39) e (2.40), temos que

$$\mathbb{P}(G|A,\tau) = [1 - \exp(-\alpha\tau)]\exp(-\beta\tau)\frac{\pi_G}{\pi_R} + [1 - \exp(-\beta\tau)]\pi_G.$$

De forma mais geral, podemos escrever as probabilidades de transição da seguinte maneira

$$\begin{aligned}\mathbb{P}(j|i,\tau) &= \exp(-(\alpha+\beta)\tau)\delta_{ij} + \exp(-\beta\tau)(1-\exp(-\alpha\tau))\left(\frac{\pi_j\epsilon_{ij}}{\sum_k \epsilon_{jk}\pi_k}\right) \\ &+ (1-\exp(-\beta\tau))\pi_j. \end{aligned} \tag{2.41}$$

Na expressão (2.41), δ_{ij} é a função delta de Kronecker, que tem valor 1, quando $i = j$, e 0, caso contrário. O valor ϵ_{ij} é a função Watson-Kronecker que tem valor 1, quando i e j são ambos *purinas* ou ambos *pirimidinas* e 0, caso contrário. □

O primeiro termo da soma da expressão (2.41), que representa a chance de não ocorrer nenhum evento, se anula quando tratamos de duas bases diferentes. O segundo termo da soma se refere à probabilidade de ocorrer um evento do tipo I seguido de transição e nenhum evento do tipo II (ver Felsenstein, 2004).

2.3.8 "General Time Reversible" (GTR)

O modelo, apresentado a seguir, é o mais geral possível que mantém a propriedade de reversibilidade no tempo, ou seja,

$$\pi_i\mathbb{P}(j|i,\tau) = \pi_j\mathbb{P}(i|j,\tau), \tag{2.42}$$

onde π_i é a proporção de bases do tipo i e $i, j \in E$.

Reversibilidade no tempo é uma propriedade matematicamente conveniente, apesar de não ser fundamentada em razões biológicas. Na prática, muitos modelos com essa propriedade se ajustam bem a dados reais (ver Felsenstein, 2004).

O modelo "General Time-Reversible" (denotado por GTR) atribui à evolução da seqüência de DNA uma cadeia de Markov a tempo contínuo com distribuição inicial dada por (2.21) e matriz infinitesimal dada por

$$\mathbf{Q}_{\text{GTR}} = \begin{pmatrix} f_1 & \alpha\pi_G & \beta\pi_C & \gamma\pi_T \\ \alpha\pi_A & f_2 & \delta\pi_C & \epsilon\pi_T \\ \beta\pi_A & \delta\pi_G & f_3 & \eta\pi_T \\ \gamma\pi_A & \epsilon\pi_G & \eta\pi_C & f_4 \end{pmatrix}, \tag{2.43}$$

onde f_k é tal que a soma da linha k é 0, para todo $k \in \{1, \cdots, 4\}$.

Observe que a cadeia é estacionária e que $\mathbf{p}_0$ é sua distribuição de equilíbrio. Note também que este modelo satisfaz a propriedade (2.42).

O modelo GTR é bastante geral e tem um grande número de parâmetros. Além disso, as expressões analíticas para as probabilidades de transição são mais complexas e não serão apresentadas aqui. Apresentamos, no entanto, um exemplo numérico que ilustra como obter a matriz $\mathbf{P}_{\text{GTR}}(\tau)$.

Exemplo: Considere o caso em que temos $\mathbf{p}_0 = (0.3, 0.2, 0.2, 0.3)$, $\alpha = 1$, $\beta = 0.2$, $\gamma = 0.3$, $\delta = 0.1$, $\epsilon = 0.01$ e $\eta = 2.5$. Desejamos obter a matriz $\mathbf{P}_{\text{GTR}}(\tau)$ para estes parâmetros fixados. Assim, pela expressão (2.43) a matriz de taxas infinitesimais para o modelo GTR é dada por

$$\mathbf{Q}_{\text{GTR}} = \begin{pmatrix} -0.3300 & 0.2000 & 0.0400 & 0.0900 \\ 0.3000 & -0.3230 & 0.0200 & 0.0030 \\ 0.0600 & 0.0200 & -0.8300 & 0.7500 \\ 0.0900 & 0.0020 & 0.5000 & -0.5920 \end{pmatrix}. \tag{2.44}$$

Tomando $\tau = 1$, e considerando a decomposição $\mathbf{Q}_{\text{GTR}} = \mathbf{T}\,\mathbf{Diag}(1)\,\mathbf{T}^{-1}$, em que $\mathbf{Diag}(1)$ é a matriz diagonal com os autovalores de $\mathbf{Q}_{\text{GTR}}$, e $\mathbf{T}$ a matriz com seus autovetores, temos que $\mathbf{P}_{\text{GTR}}(1) = \exp(\mathbf{Q}_{\text{GTR}}) = \mathbf{T}\exp(\mathbf{Diag}(1))\mathbf{T}^{-1}$, em que

$$\mathbf{Diag}(1) = \begin{pmatrix} -0.5786 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -0.1608 & 0 \\ 0 & 0 & 0 & -1.3356 \end{pmatrix}, \tag{2.45}$$

e

$$\mathbf{T} = \begin{pmatrix} 0.6474 & -0.5000 & -0.3778 & -0.0213 \\ -0.7500 & -0.5000 & -0.6306 & 0.0210 \\ -0.1154 & -0.5000 & 0.4818 & -0.8283 \\ -0.0705 & -0.5000 & 0.4770 & 0.5595 \end{pmatrix}. \tag{2.46}$$

Assim, temos que

$$\mathbf{P}_{\mathrm{GTR}}(1) = \begin{pmatrix} 0.7448 & 0.1464 & 0.0387 & 0.0701 \\ 0.2196 & 0.7460 & 0.0178 & 0.0166 \\ 0.0580 & 0.0178 & 0.5290 & 0.3953 \\ 0.0701 & 0.0110 & 0.2635 & 0.6554 \end{pmatrix}. \tag{2.47}$$

Já se o procedimento for repetido com $\tau = 0.01$, obtemos

$$\mathbf{P}_{\mathrm{GTR}}(0.01) = \begin{pmatrix} 0.9679 & 0.0194 & 0.0040 & 0.0087 \\ 0.0290 & 0.9685 & 0.0020 & 0.0005 \\ 0.0060 & 0.0020 & 0.9221 & 0.0699 \\ 0.0087 & 0.0003 & 0.0466 & 0.9443 \end{pmatrix}. \tag{2.48}$$

Note que como τ é pequeno $\mathbf{P}_{\mathrm{GTR}}(\tau)$ se aproxima da identidade, já que $\lim_{\tau\to 0} \mathbf{P}_{\mathrm{GTR}}(\tau)$ é a matriz identidade. Por outro lado, se escolhemos um valor grande para τ, temos que a entrada (i, j) da matriz se aproxima dos valores de equilíbrio π_j, pois $\lim_{\tau\to\infty} \mathbb{P}(j|i, \tau) = \pi_j$. Isso pode ser visto na matriz $\mathbf{P}_{\mathrm{GTR}}(50)$, dada por

$$\mathbf{P}_{\mathrm{GTR}}(50) = \begin{pmatrix} 0.3001 & 0.2001 & 0.2000 & 0.2998 \\ 0.3001 & 0.2001 & 0.1999 & 0.2999 \\ 0.2999 & 0.1999 & 0.2001 & 0.3001 \\ 0.2999 & 0.1999 & 0.2001 & 0.3001 \end{pmatrix}. \tag{2.49}$$

A Tabela 2.1 apresenta uma comparação entre os modelos de substituição de bases apresentados nesse capítulo, com base nos parâmetros da matriz $\mathbf{Q}$ de taxas infinitesimais, dada por

$$\mathbf{Q} = \begin{pmatrix} h_1 & \alpha_2\pi_G & \alpha_4\pi_C & \alpha_6\pi_T \\ \alpha_1\pi_A & h_2 & \alpha_8\pi_C & \alpha_{10}\pi_T \\ \alpha_3\pi_A & \alpha_7\pi_G & h_3 & \alpha_{12}\pi_T \\ \alpha_5\pi_A & \alpha_9\pi_G & \alpha_{11}\pi_C & h_4 \end{pmatrix}, \tag{2.50}$$

em que h_k é tal que a soma da linha k é 0, para todo $k \in \{1, \cdots, 4\}$.

Tabela 2.1: Comparação dos Parâmetros nos Modelos Estudados.

Modelo	$\mathbf{p}_0$	Taxas de Mutação
JC69	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 =$ $\alpha_7 = \alpha_8 = \alpha_9 = \alpha_{10} = \alpha_{11} = \alpha_{12}$
K80	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	$\alpha_1 = \alpha_2 = \alpha_{11} = \alpha_{12};$ $\alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = \alpha_9 = \alpha_{10}$
K81	$(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$	$\alpha_1 = \alpha_2 = \alpha_{11} = \alpha_{12};$ $\alpha_3 = \alpha_4 = \alpha_9 = \alpha_{10}; \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8$
F81	$(\pi_A, \pi_G, \pi_C, \pi_T)$	$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 =$ $\alpha_7 = \alpha_8 = \alpha_9 = \alpha_{10} = \alpha_{11} = \alpha_{12}$
HKY85	$(\pi_A, \pi_G, \pi_C, \pi_T)$	$\alpha_1 = \alpha_2 = \alpha_{11} = \alpha_{12};$ $\alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = \alpha_9 = \alpha_{10}$
TN93 e F84	$(\pi_A, \pi_G, \pi_C, \pi_T)$	$\alpha_1 = \alpha_2; \alpha_{11} = \alpha_{12};$ $\alpha_3 = \alpha_4 = \alpha_5 = \alpha_6 = \alpha_7 = \alpha_8 = \alpha_9 = \alpha_{10}$
GTR	$(\pi_A, \pi_G, \pi_C, \pi_T)$	$\alpha_1 = \alpha_2; \alpha_3 = \alpha_4; \alpha_5 = \alpha_6;$ $\alpha_7 = \alpha_8; \alpha_9 = \alpha_{10}; \alpha_{11} = \alpha_{12}$

Capítulo 3

Teste da Razão de Verossimilhança e seu Poder (Caso Homogêneo)

3.1 Função de Verossimilhança

Os modelos de substituição de bases apresentados na Seção 2.3 explicam como ocorrem as mudanças em uma seqüência de DNA ao longo do tempo. Entretanto, não se tem acesso ao passado da seqüência para verificar quantas e quais mutações ocorreram. Portanto, para fazer esse tipo de estudo, é necessário um conjunto de seqüências oriundas de diferentes organismos, referentes à mesma porção do DNA (seqüências homólogas). Traçando as relações de parentesco entre as seqüências de uma amostra, com base nas semelhanças entre elas, é possível fazer inferências sobre seu passado.

Para determinar o melhor modelo para um grupo de seqüências de DNA, diversos métodos utilizam a função de verossimilhança.

Definição 3.1. *Seja $f_X(\cdot)$ a função densidade da varável aleatória X, e $\boldsymbol{\theta}$ seu vetor de parâmetros. A função de verossimilhança é definida por $L(\boldsymbol{\theta}|X) = f_X(x)$, vista como uma função do parâmetro $\boldsymbol{\theta}$.*

Nesse caso, a função de verossimilhança pode ser vista como a probabilidade $\mathbb{P}(M|\mathbf{X})$ do modelo escolhido condicionada aos dados, onde $\mathbf{X}$ é o conjunto de dados, e M é o modelo que descreve o processo estocástico.

Para o cálculo da função de verossimilhança são importantes dois aspectos das árvores filogenéticas que descrevem a relação de parentesco entre as seqüências. O formato da árvore F determina quais seqüências são mais próximas, ou seja, quais delas têm um ancestral comum mais recente, e os tempos de coalescência $\bar{\tau} = \{\tau_1, \tau_2, \cdots, \tau_{2N-2}\}$ representam os tempos decorridos desde a divergência de cada ancestral comum.

Como exemplo, a Figura 3.1 apresenta a filogenia F que relaciona quatro espécies.

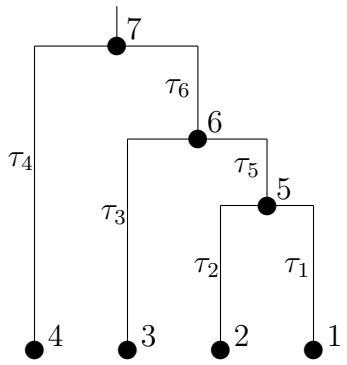


Figura 3.1: Exemplo de Árvore Filogenética.

Em geral, temos apenas informações sobre esses organismos, sendo o resto da filogenia resultado de inferência. Em um experimento hipotético, poderíamos ter quatro seqüências de genes homólogos, uma de cada espécie. Nesse caso chamamos os quatro organismos de nós externos da filogenia, e os numeramos de 1 a N, em que N é o tamanho da amostra (número de organismos considerados). As bolas pretas 5, 6 e 7 representam os últimos ancestrais comuns entre duas linhagens e são chamadas de nós internos e numeradas de $N+1$ até $2N-1$. O nó $2N-1$ é chamado de raiz da filogenia. A ligação entre dois nós é chamada de ramo da árvore filogenética, e representa o tempo evolutivo que separa um organismo de seu ancestral. Os ramos também são numerados de 1 a $2N-2$, e seus comprimentos[1] são denotados por τ_l, $l \in \{1, \cdots, 2N-1\}$.

Vamos supor conhecida a filogenia F que relaciona as seqüências a serem analisadas e seus tempos de coalescência. Observe que cada nó da filogenia corresponde a uma seqüência, mas não se tem acesso às seqüências dos nós internos (ancestrais). Sejam $X_u^1, \cdots, X_u^N$ as bases da posição u das seqüências $X^1, \cdots, X^N$; i^k as possíveis bases da seqüência ancestral k; e $h(k)$ o nó imediatamente anterior (ancestral) ao nó k. Então, a probabilidade de gerar essas N bases como resultado da filogenia conhecida (ver Durbin et al., 2004) é dada por,

$$\mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}) = \sum_{i^{N+1}, \cdots, i^{2N-1}} \pi_{i^{2N-1}} \prod_{k=N+1}^{2N-2} \mathbb{P}(i^k | i^{h(k)}, \tau_k) \prod_{l=1}^{N} \mathbb{P}(X_u^l | i^{h(l)}, \tau_l). \qquad (3.1)$$

Nessa expressão, $\pi_{i^{2N-1}}$ é a proporção, na distribuição de equilíbrio, da base que se encontra no nó $2N-1$ (raiz da filogenia). Dessa forma, $\pi_{i^{2N-1}}$ representa a probabilidade de encontrar a base i^{2N-1} no sítio u da seqüência na raiz da filogenia. Já a expressão

[1] O comprimento de um ramo da filogenia representa a quantidade de evolução ocorrida naquele ramo

$\mathbb{P}(i^k|i^{h(k)}, \tau_k)$ representa a probabilidade de que a base que se encontra na posição u do nó k seja gerada a partir da base que se encontra em sua seqüência ancestral $h(k)$ em um tempo τ_k. Essa mesma probabilidade é denotada por $\mathbb{P}(X_u^l|i^{h(l)}, \tau_l)$ quando a seqüência filha é parte da amostra $\mathbf{X}$, pois nesse caso tem-se conhecimento da base X_u^l. Essas probabilidades dependem do modelo de substituição de bases adotado.

Assim sendo, calcula-se a probabilidade de que uma determinada combinação de bases nas seqüências ancestrais gere o conjunto de dados da amostra multiplicando $\pi_{i^{2N-1}}$ pelas probabilidades de que cada base de seqüência ancestral gere as bases das suas seqüências filhas nos tempos $\bar{\tau}$ determinados pela filogenia. Entretanto, como não temos acesso às seqüências ancestrais, para o cálculo da função de verossimilhança temos que somar a probabilidade de obtermos os dados da amostra para todas as possíveis combinações de bases nos nós internos da filogenia F. Assim, obtemos a função de verossimilhança para um dos S sítios da seqüência.

Como assumimos independência entre os sítios, temos que a função de verossimilhança para a amostra $\mathbf{X}$ é dada por

$$L(F, \bar{\tau}|\mathbf{X}) = L(F, \bar{\tau}|X^1, \cdots, X^{2N-1}) = \prod_{u=1}^{S} \mathbb{P}(X_u^1, \cdots, X_u^{2N-1}|F, \bar{\tau}), \tag{3.2}$$

onde $\boldsymbol{\theta} = (F, \bar{\tau})$.

3.1.1 Algoritmo para a Obtenção da Função de Verossimilhança

Para simplificar o cálculo da função de verossimilhança, utiliza-se um algoritmo desenvolvido por Felsenstein (1981) que consiste em dividir a filogenia em árvores menores para facilitar o cálculo da função de verossimilhança.

Seja L_k a sub-árvore descendente do nó k, e seja

$$\mathbb{P}(L_k|i)_u = \mathbb{P}(\{X_u\}^k|X_u^k = i),$$

em que $\{X_u\}^k$ representa as bases encontradas na posição u dos nós externos (seqüências da amostra) descendentes do nó k. Note que tal probabilidade depende do modelo de substituição de bases adotado. Assim, $\mathbb{P}(L_k|i)_u$ representa a probabilidade de que o modelo de substituição de bases escolhido produza as bases encontradas na amostra nos nós externos descendentes da sub-árvore que tem o nó k como raiz, dado que a base encontrada no nó k é i.

O algoritmo proposto por Felsenstein (1981) utiliza os sucessivos $\mathbb{P}(L_k|i)_u$ para reduzir o número de cálculos necessários para a obtenção da função de verossimilhança. Inicia-se

o algoritmo calculando $\mathbb{P}(L_k|i)_u$ para nós externos ($k \in \{1, \cdots, N\}$), que é simplesmente dado pela função indicadora da base i, já que não há descendentes a serem considerados e temos conhecimento de qual base ocupa a posição u da seqüência k. Em seguida, calcula-se $\mathbb{P}(L_k|i)_u$ para árvores formadas pelos nós imediatamente anteriores (ancestrais) e seus descendentes. O procedimento de "subir" a filogenia, considerando árvores cada vez maiores, segue até que se atinja a raiz ($k = 2N - 1$). Em cada etapa, utilizam-se os valores de $\mathbb{P}(L_k|i)_u$ das árvores menores para simplificar o cálculo.

A função de verossimilhança para o sítio u é dada por $L(F, \bar{\tau}|X_u) = \sum_{i \in E} \mathbb{P}(L_{2N-1}|i)_u \pi_i$, em que π_i é interpretado como a probabilidade de encontrar a base i na posição u da seqüência da raiz da filogenia. A função de verossimilhança da amostra é dada pelo produto das funções de verossimilhança obtidas para cada sítio.

Dessa forma apresentamos o algoritmo de Felsenstein (1981):

1. Para cada sítio $u \in \{1, \cdots, S\}$ das seqüências:

(a) Para cada uma das seqüências da amostra ($k \in \{1, \cdots, N\}$), para cada $i \in E = \{A, G, C, T\}$ calcular:
$$\mathbb{P}(L_k|i)_u = I(X_u^k = i),$$
em que $I(A)$ representa a função indicadora do conjunto A.

(b) Para cada nó interno ($k \in \{N+1, \cdots, 2N-1\}$), para cada $i \in E = \{A, G, C, T\}$ calcular:
$$\mathbb{P}(L_k|i)_u = \sum_{j,l \in E} \mathbb{P}(j|i, \tau_n)\mathbb{P}(L_n|j)_u \mathbb{P}(l|i, \tau_m)\mathbb{P}(L_m|l)_u,$$
em que n e m são os nós descendentes do nó k.

(c) Calcular a função de verossimilhança no sítio u como
$$L(F, \bar{\tau}|X_u) = \sum_{i \in E} \mathbb{P}(L_{2N-1}|i)_u \pi_i.$$

2. Calcular a função de verossimilhança da árvore como $L(F, \bar{\tau}|\mathbf{X}) = \prod_{u=1}^{S} L(F, \bar{\tau}|X_u)$.

No caso dos modelos em que possuímos resultados analíticos para as probabilidades de mutação em função do tempo, podemos seguir explicitando as probabilidades $\mathbb{P}(j|i, \tau_k)$ que aparecem no algoritmo da função de verossimilhança.

Para o modelo Jukes-Cantor (JC69) utilizam-se as expressões

$$\mathbb{P}(j|i, \tau_k) = \frac{1}{4}(1 + 3e^{-4\alpha\tau_k}), \tag{3.3}$$

se $i = j$, e

$$\mathbb{P}(j|i,\tau_k) = \frac{1}{4}(1 - e^{-4\alpha\tau_k}), \tag{3.4}$$

se $i \neq j$.

Já para o modelo Kimura-2 Parâmetros (K80) essas mesmas probabilidades são dadas por

$$\mathbb{P}(j|i,\tau_k) = \frac{1}{4}(1 - e^{-4\gamma\tau_k}), \tag{3.5}$$

se i e j são ambos *purinas* ou ambos *pirimidinas*,

$$\mathbb{P}(j|i,\tau_k) = \frac{1}{4}(1 + e^{-4\alpha\tau_k} - 2e^{-2(\beta+\gamma)\tau_k}), \tag{3.6}$$

se i e j pertencem a categorias químicas diferentes, e

$$\mathbb{P}(j|i,\tau_k) = 1 - \beta_{\tau_k} - 2\gamma_{\tau_k}, \tag{3.7}$$

se $i = j$.

No modelo Kimura 3-Parâmetros (K81)

$$\mathbb{P}(j|i,\tau_k) = \frac{1}{4}\left(1 - e^{-2(\gamma+\delta)\tau_k} + e^{-2(\gamma+\beta)\tau_k} - e^{-2(\beta+\delta)\tau_k}\right), \tag{3.8}$$

se i e j são ambos *purinas* ou ambos *pirimidinas*,

$$\mathbb{P}(j|i,\tau_k) = \frac{1}{4}\left(1 - e^{-2(\gamma+\delta)\tau_k} - e^{-2(\gamma+\beta)\tau_k} + e^{-2(\beta+\delta)\tau_k}\right). \tag{3.9}$$

se i e j são bases complementares,

$$\mathbb{P}(j|i,\tau_k) = \frac{1}{4}\left(1 + e^{-2(\gamma+\delta)\tau_k} - e^{-2(\gamma+\beta)\tau_k} - e^{-2(\beta+\delta)\tau_k}\right), \tag{3.10}$$

se i e j pertencem a categorias químicas diferentes e não são bases complementares, e

$$\mathbb{P}(j|i,\tau_k) = 1 - \beta_{\tau_k} - \gamma_{\tau_k} - \delta_{\tau_k}, \tag{3.11}$$

se $i = j$.

Já no caso do modelo F84, utiliza-se a espressão

$$\begin{aligned} P(j|i,\tau_k) &= \exp(-(\alpha+\beta)\tau_k)\delta_{ij} + \exp(-\beta\tau_k)(1 - \exp(-\alpha\tau_k))\left(\frac{\pi_j\epsilon_{ij}}{\sum_k \epsilon_{jk}\pi_k}\right) \\ &+ (1 - \exp(-\beta\tau_k))\pi_j. \end{aligned}$$

Assim, temos como obter a função de verossimilhança para um conjunto de dados com filogenia F conhecida e parâmetros do modelo de substituição de bases definidos.

3.1.2 Cálculo do Máximo da Função de Verossimilhança

Assim como o cálculo da função de verossimilhança para uma determinada filogenia, a sua maximização deve ser feita computacionalmente. Como em grande parte das aplicações a filogenia que relaciona as seqüências é desconhecida, a maioria dos programas computacionais que maximizam a função de verossimilhança o fazem acompanhado de uma busca pela melhor árvore.

Entretanto, é praticamente impossível, até mesmo para conjuntos de dados relativamente pequenos, obter com certeza a melhor árvore. Dessa forma, na prática, os programas computacionais utilizam métodos heurísticos para obter árvores quase ótimas em tempos computacionais razoáveis (ver Guidon e Gascuel, 2003).

Por exemplo, no pacote PHYML (Guidon e Gascuel, 2003), utilizado nesse trabalho, inicialmente a árvore filogenética é construída através do método "Neighbor-Joining", baseado em distâncias entre as seqüências de DNA. Este método é rápido mas menos preciso do que o método da máxima verossimilhança. Em seguida, é calculada a função de verossimilhança da árvore, e os parâmetros do modelo de substituição de bases são otimizados utilizando-se um método de otimização numérica. Após isso, são feitas sucessivas etapas de modificações da topologia da árvore e comprimento dos ramos, seguido de novo cálculo da função de verossimilhança e otimização dos parâmetros do modelo. Sempre que a nova árvore apresenta função de verossimilhança inferior à antiga ela é descartada. Caso contrário, ela é mantida e a velha árvore descartada. Esse procedimento segue até que a função de verossimilhança convirja, e que não hajam mais modificações na árvore capazes de aumentá-la.

Dessa forma, estima-se a filogenia que melhor relaciona as seqüências e os parâmetros do modelo de substituição de bases escolhido pelo método da máxima verossimilhança. Assim também se obtém o máximo da função de verossimilhança para o modelo de substituição escolhido que é necessário para a realização do teste que compara os modelos.

3.2 Teste da Razão de Verossimilhança

Nesta seção, utilizamos o teste da razão de verossimilhança para escolher, entre dois modelos diferentes, aquele que melhor descreve a evolução da seqüência de DNA. O objetivo é testar entre os modelos apresentados na Seção 2.3. Note que todos eles têm os seguintes pressupostos em comum:

A1. As seqüências são relacionadas por uma estrutura filogenética;

A2. As mutações de cada sítio são independentes e identicamente distribuídas;

A3. As mutações são regidas por uma cadeia de Markov, com matriz de taxas infinitesimais $\mathbf{Q}$ e probabilidades iniciais $\mathbf{p}_0$.

A distinção entre estes modelos é feita unicamente pela matriz $\mathbf{Q}$ e pelo vetor $\mathbf{p}_0$. Dessa forma, sejam

$$\begin{aligned} &H_0\text{: Pressupostos A1, A2, A3 e } \mathbf{Q} = \mathbf{Q}_{\mathrm{M}_0}, \quad \mathbf{p}_0 = \mathbf{p}_{0,\mathrm{M}_0}; \\ &H_1\text{: Pressupostos A1, A2, A3 e } \mathbf{Q} = \mathbf{Q}_{\mathrm{M}_1}, \quad \mathbf{p}_0 = \mathbf{p}_{0,\mathrm{M}_1}, \end{aligned} \tag{3.12}$$

em que M_0 e M_1 pertencem à classe dos modelos de substituição de bases, dados na Seção 2.3, $\mathcal{M} = \{\text{JC69, K80, K81, F81, TN93, HKY85, F84, GTR}\}$.

Definição 3.2. *Considere o teste de hipóteses dado em (3.12). A função do teste da razão de verossimilhança para testar H_0 versus H_1 é dada por*

$$-2\Delta(\mathbf{X}) = -2\left(\log\left(\hat{L}_0(\mathbf{X})\right) - \log\left(\hat{L}_1(\mathbf{X})\right)\right), \tag{3.13}$$

em que $\hat{L}_0(\mathbf{X})$ é o máximo da função de verossimilhança sob H_0 e $\hat{L}_1(\mathbf{X})$ é o máximo da função de verossimilhança sob H_1.

Sob H_0 consideramos sempre o modelo que possui o menor número de parâmetros (é o modelo mais simples), pois dessa forma ele só será rejeitado se o modelo mais genérico tiver uma diferença de desempenho significativa.

Observe que, como os modelos que consideramos neste trabalho são todos casos particulares uns dos outros, teremos sempre que $-2\Delta(\mathbf{X}) \geq 0$, pois a função de máxima verossimilhança sob H_0 é no máximo igual à função de máxima verossimilhança sob H_1.

Vamos considerar primeiro um exemplo em que analisamos se o modelo Jukes-Cantor é plausível para o conjunto de dados $\mathbf{X}$ considerado, ou se corremos um alto risco de erro devido a diferença entre taxas de *transições* e *transversões*. Para isso, utilizamos o teste da razão de verossimilhança com a hipótese nula H_0: A1, A2, A3 e JC69 (a cadeia se comporta de acordo com o modelo JC69), contra a hipótese alternativa, H_1: A1, A2, A3 e K80 (a cadeia se comporta de acordo com o modelo K80). Nesse caso, sabemos que o modelo Kimura-2 Parâmetros, que permite diferentes taxas de *transições* e *transversões*, terá um melhor desempenho, mas a pergunta é se a diferença é estatisticamente significativa.

Assumindo como conhecida a árvore F que relaciona $\mathbf{X}$, calculamos a função do teste $-2\Delta(\mathbf{X})$ de acordo com a expressão (3.13), obtendo $\hat{L}_0(\mathbf{X})$ e $\hat{L}_1(\mathbf{X})$ como na Seção 3.1. Comparamos a função do teste $-2\Delta(\mathbf{X})$ com a sua distribuição sob H_0 para obter o p-valor, e com base nele decidir se rejeitamos ou não a hipótese nula. Observe que, nesse teste, estamos apenas comparando o aspecto em que diferem os dois modelos (taxas de *transições* e *transversões*), mas assumimos como verdadeiros todos os pressupostos comuns a ambos. Se algum dos pressupostos comuns aos dois modelos não estiver correto, isso pode afetar o resultado do teste.

Definição 3.3. *O p-valor é a estatística definida como o menor nível de significância λ para o qual um experimentador utilizando a estatística $T(\mathbf{X})$ do teste rejeitaria H_0 com base na amostra observada.*

Dessa forma, o p-valor corresponde à probabilidade de, sob a distribuição de H_0, encontrarmos valores mais extremos para a estatística do teste do que o valor observado. Assim, calculamos o p-valor do teste da razão de verossimilhança que compara dois modelos de substituição de bases como

$$\text{p-valor} = \mathbb{P}(-2\Delta(\mathbf{X}) < Y), \tag{3.14}$$

onde Y é uma variável aleatória com a distribuição de $-2\Delta(\mathbf{X})$ sob H_0.

Teorema 3.1. *Sejam $F_{\mathbf{X}}(\cdot)$ função de distribuição do vetor de variáveis aleatórias $\mathbf{X}$ e $\boldsymbol{\theta} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_s)$ o vetor de parâmetros, tais que estejam satisfeitas as seguintes condições*

$$\left\| \mathbb{E}\left(\frac{\partial}{\partial \theta_u} \log(L) \frac{\partial}{\partial \theta_w} \log(L) \right) \right\| < \infty, \tag{3.15}$$

$$\mathbb{E}\left(\frac{\partial}{\partial \theta_u} \log(L) \frac{\partial}{\partial \theta_w} \log(L) \right) + \mathbb{E}\left(\frac{\partial^2}{\partial \theta_u \partial \theta_w} \log(L) \right) = \frac{\partial^2}{\partial \theta_u \partial \theta_w} \int L = 0, \tag{3.16}$$

para todo $\theta_u, \theta_w \in \boldsymbol{\theta}$. Considere a hipótese nula $H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}$ e a alternativa $H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r_0}$ a serem testadas. Então, sob H_0, temos que

$$-2\Delta(\mathbf{X}) \xrightarrow{d} \chi^2_r, \quad \textit{quando } S \to \infty, \tag{3.17}$$

onde $-2\Delta(\mathbf{X})$ é a estatística do teste (3.12), dada na expressão (3.13), χ^2_r é a distribuição qui-quadrado central com r graus de liberdade, e $\xrightarrow{d}$ significa convergência em distribuição.

Demonstração: Ver Wilks (1962). □

Observação 3.1. Para o teste JC69 versus K80 comentado anteriormente, temos $r = 2 - 1 = 1$.

Observação 3.2. Se X é uma variável aleatória com distribuição qui-quadrado central com r graus de liberdade ($X \sim \chi^2_r$), então, a função densidade de probabilidade de X é dada por

$$f_X(x) = \begin{cases} \frac{1}{\Gamma(r/2)2^{r/2}} e^{-\frac{x}{2}} x^{\frac{r}{2}-1}, & 0 < x < \infty, \\ 0, & x \leq 0, \end{cases}$$

onde a função gama $\Gamma(\cdot)$ é definida por $\Gamma(a) = \int_0^\infty t^{a-1}e^{-t}dt$. Observe que $\mathbb{E}(X) = r$ e $\mathrm{Var}(X) = 2r$.

Lema 3.1. *Considere os modelos de substituição de bases apresentados na Seção 2.3, no caso em que a filogenia F é conhecida. Então, as condições de regularidade apresentadas em (3.15) e (3.16) do Teorema 3.1 estão satisfeitas.*

Demonstração: Para conferir se a propriedade (3.16) é válida, notamos que as derivadas de primeira e segunda ordem de $\log(L)$, em relação aos parâmetros $\boldsymbol{\theta}$ existem, contanto que os parâmetros estejam no interior do espaço paramétrico $\boldsymbol{\Theta}$. Isso pode ser facilmente visto, uma vez que $\log(L(\boldsymbol{\theta}|\mathbf{X})) = \sum_{u=1}^{S} \log\left(\mathbb{P}(X_u^1, \cdots, X_u^{2N-1}|\boldsymbol{\theta})\right)$, em que $\mathbb{P}(X_u^1, \cdots, X_u^{2N-1}|\boldsymbol{\theta}) > 0$ é constituído de soma e produto de funções do tipo (2.16), (2.19) e (2.41).

Por outro lado, como $\mathbf{X}$ pode assumir apenas um conjunto discreto e limitado de valores, temos que

$$\begin{aligned} & \mathbb{E}\left(\frac{\partial}{\partial\theta_u}\log(L)\frac{\partial}{\partial\theta_w}\log(L)\right) + \mathbb{E}\left(\frac{\partial^2}{\partial\theta_u\partial\theta_w}\log(L)\right) \\ = & \sum_{\mathbf{X}}\left(\frac{\partial}{\partial\theta_u}\log(L)\frac{\partial}{\partial\theta_w}\log(L) + \frac{\partial^2}{\partial\theta_u\partial\theta_w}\log(L)\right) f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}). \end{aligned}$$

Ainda, como $L \equiv L(\boldsymbol{\theta}|\mathbf{X}) = f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, podemos desenvolver as derivadas, de modo que temos

$$\begin{aligned} & \mathbb{E}\left(\frac{\partial}{\partial\theta_u}\log(L)\frac{\partial}{\partial\theta_w}\log(L)\right) + \mathbb{E}\left(\frac{\partial^2}{\partial\theta_u\partial\theta_w}\log(L)\right) \\ = & \sum_{\mathbf{X}}\left(\frac{\partial}{\partial\theta_u}\log(L)\frac{\partial}{\partial\theta_w}\log(L) + \frac{\partial^2}{\partial\theta_u\partial\theta_w}\log(L)\right) L \\ = & \sum_{\mathbf{X}}\left(\left(\frac{1}{L}\frac{\partial L}{\partial\theta_u}\right)\left(\frac{1}{L}\frac{\partial L}{\partial\theta_w}\right) + \frac{1}{L}\frac{\partial^2 L}{\partial\theta_u\partial\theta_w} - \frac{1}{L^2}\frac{\partial L}{\partial\theta_u}\frac{\partial L}{\partial\theta_w}\right) L \\ = & \sum_{\mathbf{X}}\frac{\partial^2 L}{\partial\theta_u\partial\theta_w} = \frac{\partial^2}{\partial\theta_u\partial\theta_w}\sum_{\mathbf{X}} L = \frac{\partial^2}{\partial\theta_u\partial\theta_w}1 = 0, \end{aligned} \tag{3.18}$$

em que a primeira igualdade da última linha da expressão (3.18) é válida porque o somatório em $\mathbf{X}$ é finito.

Para conferir se a propriedade (3.15) é válida, notamos que

$$\left\|\mathbb{E}\left(\frac{\partial}{\partial\theta_u}\log(L)\frac{\partial}{\partial\theta_w}\log(L)\right)\right\| = \left\|\sum_{\mathbf{X}}\left(\frac{\partial}{\partial\theta_u}\log(L)\frac{\partial}{\partial\theta_w}\log(L)\right)f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\right\|$$
$$= \left\|\sum_{\mathbf{X}}\left(\frac{1}{L}\frac{\partial^2 L}{\partial\theta_u\partial\theta_w}\right)f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})\right\| \quad (3.19)$$

é finito se $\left\|\frac{\partial}{\partial\theta_u}\log(L)\right\| < \infty$. Por outro lado, temos que

$$\frac{\partial}{\partial\theta_u}\log(L) = \frac{1}{L}\frac{\partial L}{\partial\theta_u},$$

onde $L > 0$ para todo $\mathbf{X}$, desde que os parâmetros do modelo de substituição de bases e os tempos $\bar{\tau}$ não sejam nulos. Além disso, como L é composto de soma e produto de expressões do tipo (2.16), (2.19) e (2.41), se os parâmetros do modelo de substituição de bases e os tempos $\bar{\tau}$ (comprimentos dos ramos da filogenia) não são nulos nem infinitos, então a derivada $\frac{\partial L}{\partial\theta_u}$ é finita.

Assim, como temos $\frac{1}{L} < \infty$ e $\frac{\partial L}{\partial\theta_u} < \infty$ para todo $\theta_u \in \Theta$, então $\left\|\frac{\partial}{\partial\theta_u}\log(L)\right\| < \infty$. Assim, a soma finita de (3.19) é finita, de forma que temos

$$\left\|\mathbb{E}\left(\frac{\partial}{\partial\theta_u}\log(L)\frac{\partial}{\partial\theta_w}\log(L)\right)\right\| < \infty.$$

Deste modo, mostramos que a condição (3.15) também é válida.

□

Goldman (1993), um dos primeiros a utilizar o teste da razão de verossimilhança para comparar modelos de substituição de bases, ressalta alguns problemas na utilização da distribuição χ^2_r para esses testes. Ele trata a matriz de dados $\mathbf{X}$ $(N \times S)$ como S sucessivas observações *i.i.d.* de um processo com distribuição multinomial. Assim, a combinação de bases que aparece em um sítio do DNA é uma observação, e cada possível observação tem probabilidade determinada pela filogenia F e o modelo de substituição de bases. Para que a aproximação qui-quadrado seja válida, é necessário que cada categoria (possível combinação de bases em um sítio da seqüência) apareça um número mínimo de vezes na amostra. Para um dado número N de seqüências consideradas, o número de categorias possíveis é 4^N. Assim, temos que, com algumas poucas seqüências, o número de categorias é maior do que os comprimentos de seqüência usuais. Além disso, na maioria das seqüências de DNA as categorias com todas as bases iguais são muito mais frequentes que as demais categorias. Assim, algumas das categorias nunca apareceriam na amostra. Desta forma, para utilizar devidamente a aproximação χ^2 precisaríamos de seqüências muito longas para um número moderado de organismos.

Outra crítica de Goldman (1993) ao uso da aproximação χ^2_r é que exemplos como o explorado acima, em que a árvore filogenética F, que relaciona os elementos de $\mathbf{X}$, é

conhecida, são muito raros. Na maioria dos casos, é com os dados de $\mathbf{X}$ que se fazem inferências sobre a árvore F e principalmente sobre os tempos de coalescência $\bar{\tau}$, e esses resultados podem variar de acordo com o modelo de substituição de bases adotado. Nesses casos, os elementos da filogenia fazem parte dos parâmetros do modelo sob os quais a função de verossimilhança deve ser maximizada e deveriam ser levados em consideração na determinação dos graus de liberdade da distribuição χ^2_r. Como a parametrização da estimativa de árvores filogenéticas, pelo método da máxima verossimilhança, não é bem compreendida, é difícil determinar os graus de liberdade r.

Em função dessas críticas, Goldman (1993) sugere o método apresentado a seguir de reamostragem paramétrica para obter a distribuição da $-2\Delta(\mathbf{X})$ sob H_0.

3.2.1 Método para o Cálculo do p-Valor

Como alternativa à distribuição χ^2_r, Goldman (1993) propõe uma variação do teste de Cox para obter a distribuição de $-2\Delta(\mathbf{X})$ sob H_0 e o p-valor, dado pela expressão (3.14). O teste de Cox compara modelos não necessariamente hierárquicos, em que a hipótese nula pode ser composta, ou seja não há um único valor especificado para os parâmetros sob H_0. Como sob H_0 a melhor explicação para os dados (máximo da função de verossimilhança) é justamente o modelo com os parâmetros estimados pelo método da máxima verossimilhança, utilizam-se essas estimativas para os valores dos parâmetros sob H_0. Se a hipótese nula com estes parâmetros for rejeitada, então ela seria rejeitada com qualquer outra escolha de parâmetros permitidos por H_0, pois a sua função de verossimilhança seria menor.

Com base no mesmo princípio, utilizam-se as estimativas de máxima verossimilhança dos parâmetros do modelo para fazer uma simulação de Monte Carlo dos dados e obter a distribuição de $-2\Delta(\mathbf{X})$ sob H_0. No caso de seqüências de DNA, os parâmetros a serem estimados incluem a árvore filogenética F, o comprimento dos ramos $\bar{\tau}$ e os parâmetros livres do modelo de substituição de bases adotado.

Para obter uma distribuição através de uma simulação de Monte Carlo, utiliza-se o modelo sob H_0 para gerar m conjuntos de dados, e com estes dados constrói-se um histograma. Quando m aumenta, o histograma deve se aproximar da distribuição que estamos tentando obter.

Apresentamos o *método baseado em simulações de Monte Carlo* (também conhecido como *bootstrap paramétrico*) para a obtenção do p-valor.

1. Utilizar os dados observados para estimar, pelo método da máxima verossimilhança, os parâmetros do modelo de substituição de bases, a árvore filogenética F e os tamanhos dos ramos $\bar{\tau}$, utilizando o modelo M_0 sob H_0.

2. Gerar, aleatoriamente, uma seqüência para a raiz da filogenia segundo uma distribuição multinomial com vetor de probabilidades $\mathbf{p}_0$ designada pelo modelo M_0 sob H_0.

3. Utilizando os parâmetros estimados no passo **1**, simular as seqüências filhas da árvore. Para isso, inicia-se pela raiz da filogenia, e geram-se seqüências para os nós adjacentes partindo da seqüência da raiz e utilizando as probabilidades de mutação $\mathbb{P}(i|j,\tau)$. Essas probabilidades são determinadas pelos parâmetros estimados para o modelo de substituição de bases e os $\bar{\tau}$ estimados. Em seguida, repete-se o procedimento para os nós descendentes daqueles recém gerados, seguindo o desenho da filogenia estimada até que se obtenham seqüências para os N nós externos.

4. Calcular $-2\Delta(\cdot)$ dado pela expressão (3.13) para os dados gerados no passo **3**.

5. Repetir os passos **2**, **3** e **4**, m vezes [2].

6. Calcular $-2\Delta(\cdot)$ para os valores da amostra.

7. Construir um histograma com os valores de $-2\Delta(\cdot)$ simulados.

8. Verificar a posição de $-2\Delta(\mathbf{X})$ produzido pela amostra $\mathbf{X}$ relativo ao histograma do passo **7** para obter o p-valor.

Uma alternativa para o procedimento descrito no passo **3** para gerar N seqüências de acordo com o modelo M_0 sob H_0 é atribuir à amostra uma distribuição multinomial em que cada uma das possíveis combinações de bases em um sítio tem uma probabilidade. A probabilidade de cada uma das 4^N possíveis categorias pode ser calculada utilizando a árvore filogenética, os comprimentos dos ramos e os parâmetros do modelo de substituição de bases estimados. Entretanto, como o número de categorias cresce rapidamente com o número de seqüências, para muitos casos é mais rápido fazer a simulação da forma descrita no passo **3** acima.

3.2.2 Poder do Teste

Em testes estatísticos, erros do tipo I ocorrem quando a hipótese nula é rejeitada sendo ela verdadeira, e o erro do tipo II ocorre quando H_0 não é rejeitada mas ela é falsa, ou seja

$$\begin{aligned} \mathbb{P}(\text{erro tipo I}) &= \mathbb{P}(H_0 \text{ é rejeitada} | H_0 \text{ é verdadeira}) \\ \mathbb{P}(\text{erro tipo II}) &= \mathbb{P}(H_0 \text{ é aceita} | H_0 \text{ é falsa}). \end{aligned} \tag{3.20}$$

Definição 3.4. *O poder do teste é definido como* $1 - \mathbb{P}(\text{erro do tipo II})$.

Assim, o poder do teste equivale à probabilidade de rejeitar H_0 dado que a hipótese alternativa H_1 é verdadeira.

[2] Goldman (1993) comenta que, para esse tipo de teste com 95% de confiança, é aceito que $m \geq 100$ apresenta bons resultados.

Teorema 3.2. *Seja $\boldsymbol{\theta} = (\boldsymbol{\theta}_r, \boldsymbol{\theta}_s) \in \boldsymbol{\Theta}$ o vetor de parâmetros do modelo submetido ao teste da razão de verossimilhança. Considere a hipótese nula $H_0 : \boldsymbol{\theta}_r = \boldsymbol{\theta}_{r_0}$ e a alternativa $H_1 : \boldsymbol{\theta}_r \neq \boldsymbol{\theta}_{r_0}$ a serem testadas. Então, sob as mesmas condições do Teorema 3.1, temos que, para todo $\epsilon > 0$,*

$$\lim_{S\to\infty} \mathbb{P}\left(|-2\Delta(\mathbf{X}) - \chi^2_r(D)| > \epsilon\right) = 0 \tag{3.21}$$

onde $-2\Delta(\mathbf{X})$ é a estatística do teste (3.13), $\chi^2_r(D)$ é a distribuição qui-quadrado com r graus de liberdade e parâmetro de não centralidade D, e S é comprimento das seqüências. Temos ainda que, sob H_0, $D = 0$ e sob H_1, $D = (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})'$. Nessa expressão, $\mathbf{I} = (I_{u,w})_{u,w}$ é a matriz de Informação de Fisher, cujos termos são dados por

$$I_{u,w} = -\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\theta_u \partial\theta_w}\right), \tag{3.22}$$

onde L é a função de verossimilhança e $\theta_u, \theta_w \in \boldsymbol{\theta}_r$.

Demonstração: Ver Kendall e Stuart (1973). □

Observação 3.3. Se X é uma variável aleatória com distribuição qui-quadrado não-central com r graus de liberdade e parâmetro de não centralidade D, $(X \sim \chi^2_r(D))$, então, a função densidade de probabilidade de X é dada por

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{\pi}2^{r/2}} e^{-\frac{x+D}{2}} x^{\frac{r}{2}-1} \sum_{j=0}^{\infty} \frac{(xD)^j \Gamma(j+\frac{1}{2})}{(2j)!\Gamma((j+\frac{1}{2r})}, & 0 < x < \infty, \\ 0, & x \leq 0. \end{cases} \tag{3.23}$$

Observe que $\mathbb{E}(X) = r + D$ e $\mathrm{Var}(X) = 2r + 4D$.

Com base no Teorema 3.2 temos que o poder do teste da razão de verossimilhança ao nível de significância λ é dado por

$$\text{Poder} = \mathbb{P}(\chi^2_r(D) > C_\lambda), \tag{3.24}$$

em que C_λ é o valor crítico do teste, tal que $\mathbb{P}(-2\Delta(\mathbf{X}) > C_\lambda) = \lambda$ se H_0 é verdadeira. Como estamos considerando o resultado assintótico, $C_\lambda = \chi^2_{r,\lambda}$ é o valor da distribuição qui-quadrado central com r graus de liberdade χ^2_r tal que $\mathbb{P}(\chi^2_r > \chi^2_{r,\lambda}) = \lambda$.

Apresentamos a seguir um resultado assintótico para a distribuição da estatística do teste sob H_1. Para tanto, precisamos do seguinte lema.

Lema 3.2. *Seja X uma variável aleatória tal que $X \sim \chi^2_r(D)$. Se*

(a) *D é constante, então $\frac{X-(r+D)}{\sqrt{2r+4D}} \xrightarrow{d} \mathcal{Z}$, quando $r \to \infty$;*

(b) *r é constante, então $\frac{X-(r+D)}{\sqrt{2r+4D}} \xrightarrow{d} \mathcal{Z}$, quando $D \to \infty$;*

onde $\mathcal{Z} \sim N(0,1)$ é a distribuição normal padrão com média 0 e variância 1.

Demonstração: Ver Johnson e Kotz (1970). □

O Lema 3.2 pode ser utilizado para estabelecer o seguinte resultado sobre a distribuição assintótica em S de $-2\Delta(\mathbf{X})$ sob H_1.

Teorema 3.3. *Considere o teste da razão de verossimilhança que compara dois modelos de substituição de bases, dado em (3.12). Então, a distribuição assintótica da estatística do teste $-2\Delta(\mathbf{X})$, dada em (3.13), sob H_1, é tal que*

$$\frac{-2\Delta(\mathbf{X}) - (r+D)}{\sqrt{2r+4D}} \xrightarrow{d} \mathcal{Z}, \quad \text{quando } S \to \infty, \tag{3.25}$$

onde r e D estão definidos no Teorema 3.2, e $\mathcal{Z} \sim N(0,1)$.

Demonstração: Pelo Lema 3.1, sabemos que os modelos de substituição de bases satisfazem as condições de regularidade do Teorema 3.1. Observe que, pela expressão (3.2), a função de verossimilhança para qualquer modelo de substituição de bases, apresentados na Seção 2.3, é tal que

$$L(F, \bar{\tau}|\mathbf{X}) = L(F, \bar{\tau}|X^1, \cdots, X^{2N-1}) = \prod_{u=1}^{S} \mathbb{P}(X_u^1, \cdots, X_u^{2N-1}|F, \bar{\tau}).$$

Note ainda que existem 4^N possíveis combinações de bases para uma posição de todas as seqüências $(X_1^u, \cdots, X_N^1)$. Se definimos p_i como a probabilidade da combinação i, e s_i como o número de vezes que a combinação i aparece na amostra, temos que

$$\log\left(L(F, \bar{\tau}|\mathbf{X})\right) = \sum_{i=1}^{N^4} s_i \log(p_i).$$

A função acima ainda pode ser escrita em função das estatísticas $\bar{s}_i = \frac{s_i}{S}$, $i \in \{1, \cdots, S^4\}$, de modo que $\bar{s}_i$ representa a proporção de sítios da seqüência que apresenta a combinação i. Assim, temos que

$$\log\left(L(F, \bar{\tau}|\mathbf{X})\right) = S \sum_{i=1}^{N^4} \bar{s}_i \log(p_i).$$

Para calcular o parâmetro de não centralidade D, descrito no Teorema 3.2, precisamos da matriz de Informação de Fisher $\mathbf{I} = (I_{u,w})_{u,w}$, cujos termos são dados por

$$I_{u,w} = -\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\theta_u \partial\theta_w}\right) = -S\mathbb{E}\left(\frac{\partial^2}{\partial\theta_u \partial\theta_w}\sum_{i=1}^{N^4} \bar{s}_i \log(p_i)\right). \tag{3.26}$$

Assim, temos que

$$\begin{aligned} D &= (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})' \\ &= (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\left[-S\mathbb{E}\left(\frac{\partial^2}{\partial\theta_u \partial\theta_w}\sum_{i=1}^{N^4} \bar{s}_i \log(p_i)\right)\right]_{\theta_u,\theta_w \in \boldsymbol{\theta_r}} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})' \\ &= S\left((\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\left[-\mathbb{E}\left(\frac{\partial^2}{\partial\theta_u \partial\theta_w}\sum_{i=1}^{N^4} \bar{s}_i \log(p_i)\right)\right]_{\theta_u,\theta_w \in \boldsymbol{\theta_r}} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})'\right). \end{aligned} \tag{3.27}$$

A probabilidade p_i da combinação i é uma função da filogenia que relaciona as seqüências, os tempos $\bar{\tau}$, e os parâmetros do modelo de substituição de bases. Entretanto, tal probabilidade não depende de S, de forma que a expressão dentro do parênteses na igualdade (3.27) não depende de S. Dessa forma, temos que $\lim_{S\to\infty} D = \infty$.

Assim, lembrando que convergência em probabilidade implica convergência em distribuição, pelo Lema 3.2 temos que

$$\frac{-2\Delta(\mathbf{X}) - (r + D)}{\sqrt{2r + 4D}} \xrightarrow{d} \mathcal{Z}, \quad \text{quando } S \to \infty, \tag{3.28}$$

onde $\mathcal{Z} \sim N(0, 1)$.

□

Os resultados enunciados acima são assintóticos em S e válidos para os testes da razão de verossimilhança, desde que satisfeitas as condições do Teorema 3.1. Entretanto, se considerarmos os argumentos de Goldman (1993) quanto à convergência da distribuição e às particularidades da estimação de filogenias, somos levados a buscar uma forma alternativa para encontrar o poder do teste.

3.2.3 Método para o Cálculo do Poder do Teste

Utilizando uma abordagem semelhante à sugerida por Goldman (1993) para a obtenção da distribuição da estatística do teste $-2\Delta(\cdot)$, sob H_0, criamos o seguinte *algoritmo para o cálculo do poder do teste.*

1. Utilizar os dados observados para estimar, pelo método da máxima verossimilhança, os parâmetros do modelo de substituição de bases, a árvore filogenética F e os tamanhos dos ramos $\bar{\tau}$, utilizando o modelo M_0 sob H_0.

2. Gerar, aleatoriamente, uma seqüência para a raiz da filogenia segundo uma distribuição multinomial com vetor de probabilidades $\mathbf{p}_0$ designada pelo modelo M_0 sob H_0.

3. Utilizando os parâmetros estimados no passo **1**, simular as seqüências filhas da árvore. Para isso, inicia-se pela raiz da filogenia, e geram-se seqüências para os nós adjacentes partindo da seqüência da raiz e utilizando as probabilidades de mutação $\mathbb{P}(i \mid j, \tau)$. Essas probabilidades são determinadas pelos parâmetros estimados para o modelo de substituição de bases e os tempos $\bar{\tau}$ estimados. Em seguida, repete-se o procedimento para os nós descendentes daqueles recém gerados, seguindo o desenho da filogenia estimada até que se obtenham seqüências para os N nós externos.

4. Calcular $-2\Delta(\cdot)$ dado pela expressão (3.13) para os dados gerados no passo **3**.

5. Repetir os passos **2**, **3** e **4**, m vezes.

6. Construir um histograma com os valores de $-2\Delta(\cdot)$ simulados para descobrir qual o valor crítico C_λ.

7. Utilizando os dados observados, estimar, pelo método da máxima verossimilhança os parâmetros do modelo de substituição de bases, a árvore filogenética F e os tamanhos dos ramos $\bar{\tau}$ utilizando o modelo M_1 sob H_1.

8. Gerar, aleatoriamente, uma seqüência para a raiz da filogenia segundo uma distribuição multinomial com vetor de probabilidades $\mathbf{p}_0$ designada pelo modelo M_1 sob H_1.

9. Utilizando os parâmetros estimados no passo **7**, simular as seqüências filhas da árvore como feito no passo **3**.

10. Calcular $-2\Delta(\cdot)$ dado pela expressão (3.13) para os dados gerados no passo **9**.

11. Repetir os passos **8**, **9** e **10**, m vezes.

12. Construir um histograma com os valores de $-2\Delta(\cdot)$ simulados a partir do modelo sob H_1 e verificar a posição do valor crítico C_λ.

13. O poder do teste é dado pela proporção de elementos do passo **11** maiores do que C_λ.

3.2.4 Estimador $\hat{D}$ para o Parâmetro de Não-Centralidade da Distribuição Qui-Quadrado

A obtenção do poder do teste segundo o algoritmo apresentado na Seção 3.2.3 requer um esforço computacional considerável. Por exemplo, para conjuntos de 15 seqüências

de comprimento $S = 1000$, nos computadores pessoais comuns, essa análise leva em torno de duas horas e meia para ser concluída em linguagem C. Em geral, quando um pesquisador aplica o teste da razão de verossimilhança para comparar modelos de evolução do DNA, seu objetivo é escolher o melhor modelo para análises posteriores, ou responder a questões sobre a evolução das seqüências. O conhecimento do poder desses testes é um dado complementar que pode ter valor para avaliar a certeza da decisão do teste. Caso o pesquisador entenda que o tempo computacional do algoritmo utilizado para obter o poder do teste (dado na Seção 3.2.3) é muito grande, ele poderá, alternativamente, utilizar o método descrito a seguir.

Desejamos obter uma forma alternativa de estimar o poder do teste, que envolva menor tempo computacional. Para tanto, recorremos ao Teorema 3.2, que afirma que, sob H_1, a distribuição assintótica da estatística do teste $-2\Delta(\mathbf{X})$ é uma qui-quadrado não central com r graus de liberdade. Os graus de liberdade r desta distribuição são facilmente determinados para os testes que comparam os modelos apresentados na Seção 2.3. Já o parâmetro de não centralidade D desta distribuição é mais difícil de ser determinado.

Lembramos que, segundo o Teorema 3.2, temos que $D = (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})'$. Nessa expressão, $\mathbf{I} = (I_{u,w})_{u,w}$ é a matriz de Informação de Fisher, cujos termos são dados por

$$I_{u,w} = -\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\theta_u \partial\theta_w}\right), \tag{3.29}$$

onde L é a função de verossimilhança e $\theta_u, \theta_w \in \boldsymbol{\theta}_r$.

Primeiramente, notamos que, em situações reais, não conhecemos o verdadeiro valor de $\boldsymbol{\theta}_r$, o que dificulta o cálculo de $(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})$ e a obtenção da esperança na expressão (3.29). Além disso, como o formato da função de verossimilhança varia dependendo da filogenia que relaciona as seqüências, não temos uma expressão geral para as derivadas na expressão (3.29).

Para contornar estas questões, recorremos à Informação de Fisher Observada como estimador da matriz de Informação de Fisher.

Definição 3.5. A matriz da Informação de Fisher Observada $\hat{\mathbf{I}} = (\hat{I}_{u,w})_{u,w}$ é dada por

$$\hat{I}_{u,w} = -\left.\frac{\partial^2 \log(L)}{\partial\theta_u \partial\theta_w}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -\left.\frac{1}{L}\frac{\partial^2 (L)}{\partial\theta_u \partial\theta_w}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \tag{3.30}$$

para todo $\theta_u, \theta_w \in \boldsymbol{\theta}$, onde L é a função de verossimilhança, e $\widehat{\boldsymbol{\theta}}$ o estimador de máxima verossimilhança para $\boldsymbol{\theta}$.

Grande parte dos pacotes computacionais, para a análise de seqüências de DNA, que utilizam o método da máxima verossimilhança já obtém a Informação de Fisher Observada como meio para estimar a variância dos estimadores de máxima verossimilhança

dos parâmetros do modelo de substituição de bases. Desse modo, a utilização desta estatística para estimar o poder do teste não acrescenta tempo computacional à análise das seqüências.

Assim, propomos o seguinte estimador para o parâmetro de não-centralidade D e o poder do teste:

Lema 3.3. *Considere a distribuição assintótica qui-quadrado não-central da estatística do teste* $-2\Delta(\mathbf{X})$*, sob* H_1*, dada no Teorema 3.2. O estimador* $\hat{D}$ *para o parâmetro de não-centralidade* D *desta distribuição assintótica é dado por*

$$\hat{D} = (\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r_0})\hat{\mathbf{I}}(\widehat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r_0})', \tag{3.31}$$

em que $\widehat{\boldsymbol{\theta}}_r$ *é o estimador de máxima verossimilhança para* $\boldsymbol{\theta}_r$*, e* $\hat{\mathbf{I}}$ *é a matriz da Informação de Fisher Observada. Deste modo, o poder do teste pode ser estimado por*

$$\widehat{Poder} = \mathbb{P}\left(\chi_r^2 < \chi_r^2(\hat{D})\right), \tag{3.32}$$

em que os graus de liberdade r *são determinados como no Teorema 3.1,* χ_r^2 *representa a distribuição qui-quadrado central com* r *graus de liberdade, e* $\chi_r^2(\hat{D})$ *a distribuição qui-quadrado não central com parâmetro de não centralidade* $\hat{D}$*.*

No teorema apresentado a seguir estabelecemos a consistência do estimador $\hat{D}$. Note que um estimador $\hat{\theta}$ é dito consistente para o parâmetro θ se $\hat{\theta} \xrightarrow{p} \theta$ (ou seja, $\lim_{n\to\infty} \mathbb{P}(|\hat{\theta} - \theta| < \epsilon) = 1$, para todo $\epsilon > 0$).

Teorema 3.4. *Considere o teste da razão de verossimilhança que compara dois modelos de substituição de bases i.i.d., dado na expressão (3.12), e o estimador* $\hat{D}$ *para o parâmetro de não centralidade* D *da distribuição assintótica qui-quadrado de* $-2\Delta(\mathbf{X})$*, apresentado no Lema 3.3. Então,* $\hat{D}$ *é um estimador consistente para* D*.*

Demonstração: Queremos mostrar que $\hat{D} \xrightarrow{p} D$. Primeiramente, observe que, sob as condições do Teorema 3.1, temos que o estimador pelo método da máxima verossimilhança é consistente (ver Shao, 2003). Já o Lema 3.1 estabelece que, para os modelos i.i.d. considerados na Seção 2.3, as condições do Teorema 3.1 estão satisfeitas. Assim, temos que $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$.

Por outro lado, note que, se tomamos

$$\mathbf{I}_S = \left(\mathbb{E}\left[\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_1, \cdots, \mathbf{X}_S))}{\partial\theta_u \partial\theta_w}\right]\right)_{\theta_u, \theta_w \in \boldsymbol{\theta}_r} \qquad \text{e} \qquad \mathbf{I}_1 = \left(\mathbb{E}\left[\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_1))}{\partial\theta_u \partial\theta_w}\right]\right)_{\theta_u, \theta_w \in \boldsymbol{\theta}_r},$$

e consideramos que o processo que designa a combinação de bases de cada sítio é i.i.d., então, temos a seguinte relação

$$\begin{aligned}\mathbf{I}_S &= \left(\mathbb{E}\left[\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_1,\cdots,\mathbf{X}_S))}{\partial\theta_u\partial\theta_w}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} = \left(\mathbb{E}\left[\frac{\partial^2 \log\left(\prod_{i=1}^{S} L(\boldsymbol{\theta}|\mathbf{X}_i)\right)}{\partial\theta_u\partial\theta_w}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} \\ &= \left(\sum_{i=1}^{S}\mathbb{E}\left[\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u\partial\theta_w}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} = S\mathbf{I}_1. \qquad (3.33)\end{aligned}$$

De forma semelhante, temos que

$$\hat{\mathbf{I}} = \left(-\left.\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_1,\cdots,\mathbf{X}_S))}{\partial\theta_u\partial\theta_w}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} = \left(-\sum_{i=1}^{S}\left.\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u\partial\theta_w}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r}.$$

Ainda, pela Lei Fraca dos Grandes Números (ver Shao, 2003), temos que

$$\frac{\hat{\mathbf{I}}}{S} = \frac{1}{S}\left(-\sum_{i=1}^{S}\left.\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u\partial\theta_w}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} \xrightarrow{p} \left(-\mathbb{E}\left[\left.\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u\partial\theta_w}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r}. \qquad (3.34)$$

Note que a esperança da expressão (3.34) é uma função contínua de $\hat{\boldsymbol{\theta}}$. Assim, como temos que $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, podemos aplicar o Teorema de Slutzky (ver Shao, 2003), de forma que obtemos

$$\left(-\mathbb{E}\left[\left.\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u\partial\theta_w}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} \xrightarrow{p} \left(-\mathbb{E}\left[\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}_i))}{\partial\theta_u\partial\theta_w}\right]\right)_{\theta_u,\theta_w\in\boldsymbol{\theta}_r} = \mathbf{I}_1,$$

Deste modo, concluímos que a Informação de Fisher Observada é um estimador consistente para a informação de Fisher, ou seja

$$\hat{\mathbf{I}} \xrightarrow{p} S\mathbf{I}_1 = \mathbf{I}_S.$$

Como temos que $\hat{\mathbf{I}} \xrightarrow{p} \mathbf{I}$ e $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$, utilizamos novamente o Teorema de Slutzky, e obtemos

$$\hat{D} = (\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r_0})\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}}_r - \boldsymbol{\theta}_{r_0}) \xrightarrow{p} (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0}) = D,$$

ou seja, $\hat{D}$ é um estimador consistente para D. □

Observação 3.4. A relação apresentada na expressão (3.33) nos dá uma forma alternativa para a demonstração do Teorema 3.3, que garante que

$$\frac{-2\Delta(\mathbf{X}) - (r+D)}{\sqrt{2r+4D}} \xrightarrow{d} \mathcal{Z}, \quad \text{quando S} \to \infty,$$

onde r e D estão definidos no Teorema 3.2, e $\mathcal{Z} \sim N(0,1)$.

Observe que, como $S\mathbf{I}_1 = \mathbf{I}_S$, temos que

$$D = (\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}_S(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0}) = S(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0})\mathbf{I}_1(\boldsymbol{\theta}_r - \boldsymbol{\theta}_{r_0}).$$

Dessa forma, estabelece-se facilmente que $\lim_{S\to\infty} D = \infty$, e podemos aplicar o Lema 3.2 para concluir a demonstração.

Para uma comparação entre o poder do teste estimado utilizando $\hat{D}$ e o método do bootstrap paramétrico, ver Seção 6.2.

3.3 Aplicação

Nesta seção serão apresentadas aplicações da teoria deste capítulo a seqüências reais. O teste da razão de verossimilhança foi utilizado para comparar os modelos da Seção 2.3 em dois conjuntos de dados.

Ressaltamos que um dos principais motivos para comparar modelos de substituição de bases é compreender melhor o processo de evolução molecular das seqüências de DNA. Isso porque a maioria dos modelos são baseados em pressupostos com significado biológico em relação a tal processo. Deste modo, a escolha de um determinado modelo é um indício de quais fatores tem maior importância no processo de evolução da seqüência. Por exemplo, em um teste que compara os modelos JC69 e K80, a escolha do modelo K80 indica que taxas de *transições* e *transversões* distintas são importantes na evolução das seqüências de DNA em questão.

Outra importante razão pela qual pode-se aplicar esses testes é escolher o melhor modelo para realizar análises posteriores utilizando a função de verossimilhança. Diversas análises moleculares, como a escolha da árvore que relaciona a seqüência, a datação de ancestrais comuns e a inferência de caracteres ancestrais dependem da função de verossimilhança. Deste modo, para realizar essas análises, deve-se utilizar um modelo de substituição de bases. Se por um lado, a utilização de modelos equivocados pode afetar essas análises, por outro, a utilização de modelos muito complexos (menos parcimoniosos) exige muito esforço computacional. Desta forma, utilizam-se testes estatísticos para averiguar se o desempenho de um modelo mais complexo é significativamente superior ao do modelo mais simples.

Todos os cálculos da função de verossimilhança foram realizados utilizando o pacote computacional PAML distribuído por Ziheng Yang e disponibilizado *on-line* em

http:abacus.gene.ucl.ac.uk/software/paml.html. Para a aplicação do bootstrap paramétrico, foi utilizada uma adaptação do *software* "baseml", parte do pacote PAML, e algumas rotinas criadas para "R-project".

3.3.1 Exemplo

Apresentamos um exemplo detalhado de aplicação da teoria deste capítulo a um conjunto de seqüências hipotéticas no Apêndice A. Tal exemplo tem como propósito ilustrar o cálculo da função de verossimilhança utilizando tanto o algoritmo da Seção 3.1.1, quanto diretamente a expressão (3.1). Além disso, este exemplo mostra a aplicação do teste da razão de verossimilhança e a obtenção do seu poder.

Para simplificar o exemplo, supomos que temos uma amostra constituída de três seqüências de DNA alinhadas, com S=1000, e relacionadas por uma filogenia conhecida. Assim, o objetivo da análise é comparar o desempenho dos modelos JC69 e K80 para essa amostra fictícia.

No exemplo, mostramos passo a passo como é feito o cálculo da função de verossimilhança, através das probabilidades das combinações de bases nos nós externos da filogenia. Para o cálculo de tais probabilidades, consideramos as seqüências desconhecidas dos nós internos da filogenia. A comparação entre o procedimento utilizado para os modelo JC69 e K80 ilustra os cálculos adicionais necessários quando se adiciona um parâmetro ao modelo.

Utilizamos, ainda, o bootstrap paramétrico para obter a distribuição da estatística do teste $-2\Delta(\mathbf{X})$, sob H_0, e a comparamos com a distribuição assintótica teórica dada no Teorema 3.1. O poder do teste também é obtido pelo bootstrap paramétrico, conforme o algoritmo da Seção 3.2.3.

3.3.2 Felinos

Apresentamos agora uma aplicação da teoria deste capítulo a um conjunto de seqüências reais. O conjunto de dados consiste de 37 seqüências de DNA de felinos, obtidas do Genbank. Após a etapa de alinhamento, feita com o *software*, ClustalX (ver Seção 2.1.3), todas as seqüências tem comprimento de 382 nucleotídeos. A filogenia utilizada na análise foi estimada com o pacote PAML (disponível *on line* em *http:abacus.gene.ucl.ac.uk/software/paml.html*), utilizando o modelo GTR, e está apresentada na Figura 3.2.

Foram aplicados 5 testes comparando os modelos JC69, K80, F81, HKY85 e GTR. Os parâmetros dos modelos de substituição de bases estimados pelo método da máxima verossimilhança para cada um dos modelos estão apresentados na Tabela 3.1; nos modelos K80 e HKY85, em vez dos valores para β e γ foi estimada a taxa de *transições/transversões* $K = \frac{\beta}{\gamma}$, como no Apêndice A. A Tabela 3.2 apresenta os resultados obtidos para os 5 testes, JC69 $\times$ K80, F81 $\times$ HKY85, JC69 $\times$ F81, K80 $\times$ HKY85 e HKY85 $\times$ GTR, segundo as distribuições assintóticas teóricas (ver Teorema 3.1) e segundo simulações de Monte

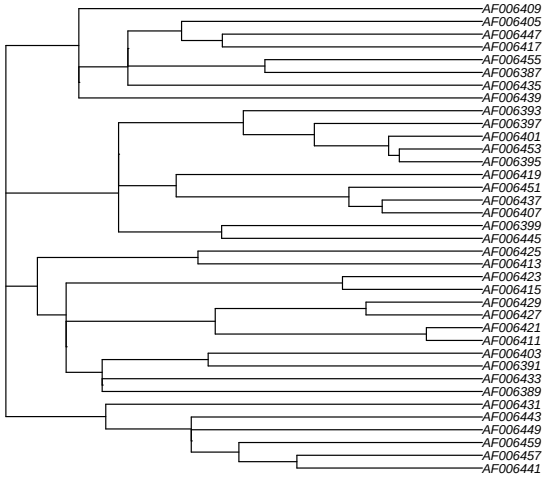


Figura 3.2: Filogenia estimada pelo modelo GTR para as 37 seqüências de felinos, utilizada na Seção de Aplicação.

Carlo (ver Seções 3.2.1 e 3.2.3). A Figura 3.3 apresenta os histogramas obtidos pelas simulações de Monte Carlo em todos os testes para as distribuições de $-2\Delta(\mathbf{X})$, sob H_0 e H_1. Nos histogramas sob H_1 a linha vertical tracejada representa o valor crítico à 99% de confiança. A Figura 3.4 apresenta os Q-Q plot comparando os quantís das distribuições de $-2\Delta(\mathbf{X})$ obtidas sob H_0 com os da distribuição χ^2 e os quantís de $-2\Delta(\mathbf{X})$ sob H_1 com os quantís da distribuição normal.

O primeiro teste aplicado foi JC69 $\times$ K80, e o valor obtido para $-2\Delta(\mathbf{X})$ foi 229.8426. O valor crítico 99% encontrado por meio de simulações de Monte Carlo foi 7.9013, já o valor crítico sugerido pela distribuição assintótica ($-2\Delta(\mathbf{X}) \sim \chi^2_1$) é 6.64, de forma

Tabela 3.1: Estimadores para os Modelos das Seqüências dos Felinos.

Modelo	Estimadores
K80	$\hat{K} = 9.4280$,
F81	$\hat{\mathbf{p}}_0 = (0.3420, 0.2037, 0.2243, 0.2300)$
HKY85	$\hat{\mathbf{p}}_0 = (0.3420, 0.2037, 0.2243, 0.2300)$, $\hat{K}$= 9.8142
GTR	$\hat{\mathbf{p}}_0 = (0.3420, 0.2037, 0.2243, 0.2300)$,
	$\hat{\alpha} = 1$, $\hat{\eta} = 2.4983$, $\hat{\beta} = 0.2072$, $\hat{\delta} = 0.1026$, $\hat{\gamma} = 0.2722$ e $\hat{\epsilon} = 0.0001$

Tabela 3.2: Resultado do Teste para as Seqüências dos Felinos.

Teste	$-2\Delta(\mathbf{X})$	Distribuição Teórica		Simulação de Monte Carlo			Decisão
		r	p-valor	p-valor	Valor Crítico 99%	Poder	
JC69 × K80	229.8426	1	6.4503×10^{-52}	< 0.0010	7.9013	1	rejeita
F81 × HKY85	248.2718	1	6.1825×10^{-56}	< 0.0010	7.4679	1	rejeita
JC69 × F81	1.4760	3	6.8851×10^{-1}	0.6930	10.0948	0.9860	aceita
K80 × HKY85	19.9052	3	1.7780×10^{-4}	< 0.0010	11.4302	0.7500	rejeita
HKY85 × GTR	44.0106	4	6.3833×10^{-9}	< 0.0010	12.2034	1	rejeita

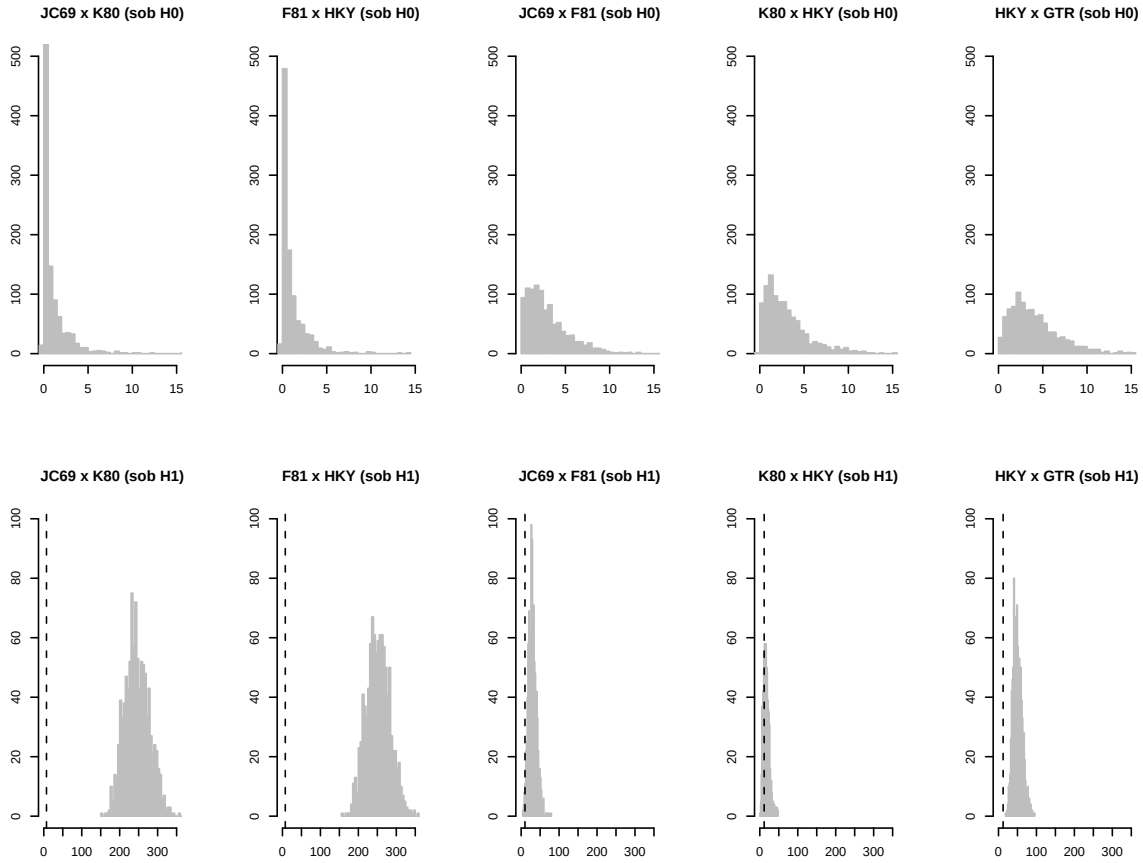


Figura 3.3: Histogramas de $-2\Delta(\mathbf{X})$ obtidos por simulações de Monte Carlo para os testes JC69 × K80, JC69 × F81, K80 × HKY85, F81 × HKY85 e HKY85 × GTR, sob H_0 e H_1, para as seqüências dos Felinos.

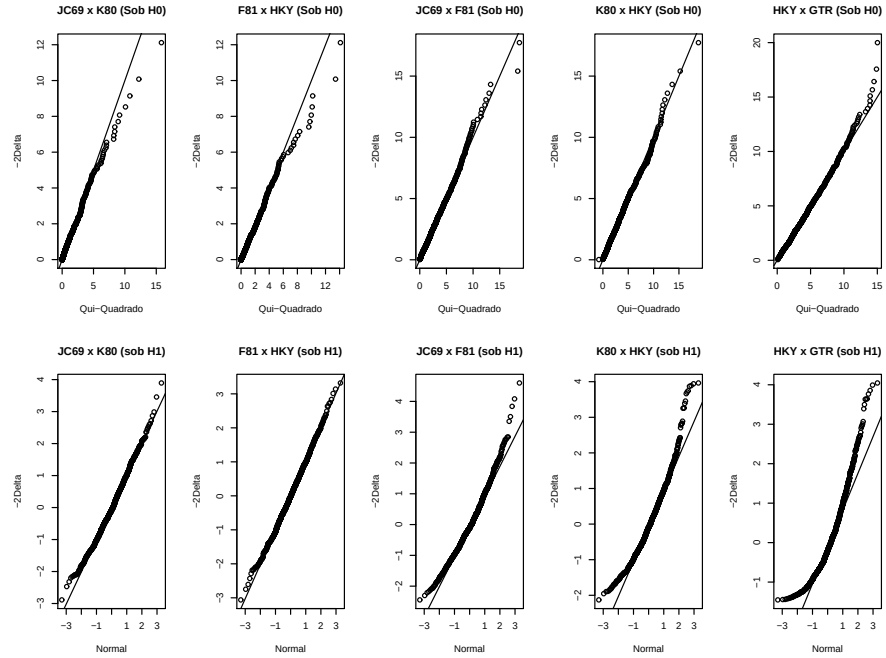


Figura 3.4: Q-Q plots de $-2\Delta(\mathbf{X})$ obtidos por simulações de Monte Carlo para os testes JC69 × K80, JC69 × F81, K80 × HKY85, F81 × HKY85 e HKY85 × GTR, sob H_0 e H_1, para as seqüências dos Felinos.

que rejeitamos H_0 independente de qual distribuição utilizamos. Notamos que o p-valor obtido por ambas as distribuições consideradas é < 0.001. O segundo teste que aplicamos foi F81 × HKY85, e o valor obtido para a estatística do teste foi 248.2718. O valor crítico assintótico é o mesmo do primeiro teste, enquanto o valor obtido das simulações foi 7.4679, de forma que, para esse teste, também rejeitamos H_0 com um p-valor < 0.001. Notamos que ambos os testes rejeitam modelos que não permitem taxas diferentes para *transições* e *transversões*. O poder de ambos os testes foi 1.

Também aplicamos os testes JC69 × F81 e K80 × HKY85, ambos visando avaliar o efeito ao permitir diferentes freqüências entre as bases. No teste JC69 × F81 obtivemos $-2\Delta(\mathbf{X}) = 1.4760$, enquanto o valor crítico da simulação de Monte Carlo foi 10.0948, com um p-valor de 0.6930, de forma que não rejeitamos H_0, ao nível de 1% de significância. O valor crítico segundo a distribuição assintótica ($-2\Delta(\mathbf{X}) \sim \chi^2_3$) é 11.350 e o poder obtido para este teste foi 0.9860. Já para o teste F81 × HKY85 a estatística do teste foi 19.9052, enquanto o valor crítico obtido foi de 11.4302, de forma que rejeitamos H_0 com um p-valor < 0.001. A distribuição assintótica para esse teste é a mesma de JC69 × F81, e o poder obtido por meio das simulações foi de 0.7500. Assim, obtivemos resultados conflitantes em relação à distribuição das bases ser diferente da homogênea, mas, como os testes anteriores trazem forte evidência a favor de modelos que diferenciam taxas de *transições* e *transversões*, favorecemos o resultado do teste F81 × HKY85.

Tabela 3.3: Estimadores para os Modelos das Seqüências dos Primatas.

Modelo	Estimadores
K80	$\hat{K} = 2.0939$,
F81	$\hat{\mathbf{p}}_0 = (0.2800, 0.2281, 0.2572, 0.2547)$
HKY85	$\hat{\mathbf{p}}_0 = (0.2862, 0.2073, 0.2540, 0.2525)$, $\hat{K}$= 2.12470
GTR	$\hat{\mathbf{p}}_0 = (0.2647, 0.1942, 0.2571, 0.2840)$,
	$\hat{\alpha} = 1$, $\hat{\gamma} = 0.4601$, $\hat{\beta} = 0.2126$, $\hat{\delta} = 0.2443$, $\hat{\eta} = 0.5061$ e $\hat{\epsilon} = 0.3359$

O último teste aplicado foi HKY85 × GTR. Obtivemos $-2\Delta(\mathbf{X}) = 44.0106$ e um valor crítico obtido pelo bootstrap paramétrico de 12.2034. Desta forma, rejeitamos H_0 com um p-valor $< 0,001$. O valor crítico segundo a distribuição assintótica ($-2\Delta(\mathbf{X}) \sim \chi^2_4$) é 13.28, e o poder obtido para esse teste foi 1.

Assim, após a seqüência de testes, escolhemos o modelo GTR para descrever o processo de evolução deste gene nesta filogenia.

Observamos que os mesmos testes de hipóteses realizados utilizando filogenias estimadas com os outros modelos para o cálculo da função de verossimilhança tiveram resultados semelhantes.

Pela análise da Figura 3.4 percebemos que a distribuição de $-2\Delta(\mathbf{X})$ sob H_0 obtida pelas simulações de Monte Carlo coincide com a distribuição χ^2 teórica para os quantís baixos, entretanto os Q-Q plots demonstram uma divergência dessa distribuição na cauda direita. Para os testes JC69 × K80 e F81 × HKY85 a distribuição de $-2\Delta(\mathbf{X})$ sob H_1 se aproxima de uma distribuição normal. Entretanto, para os outros testes a distribuição é bem distante da normal.

3.3.3 Primatas

Estudamos um conjunto de 17 seqüências de genes da família ECP-EDN dos primatas com $S = 483$. A família de genes ECP-EDN é composta de duas ribonucleasas (ECP e EDN) que têm função na resposta imunológica inespecífica desses animais. Esses genes, que atualmente têm funções distintas, sofreram um evento de duplicação em um ancestral dos primatas, de forma que cada primata carrega uma cópia para o gene ECP e outra para o gene EDN. As seqüências, já alinhadas, foram obtidas do conjunto de exemplos que acompanham o pacote PAML. Utilizamos, para as análises, a topologia da filogenia obtida por Bielawski e Yang (2003). Neste trabalho, que apresenta modelos para eventos de duplicação gênica, os autores estudam o mesmo conjunto de dados, e concluem que a topologia em questão é a que melhor descreve os eventos ocorridos nessa família gênica. A Figura 3.5 apresenta esta filogenia, em que o círculo próximo à raiz da árvore representa o evento de duplicação gênica.

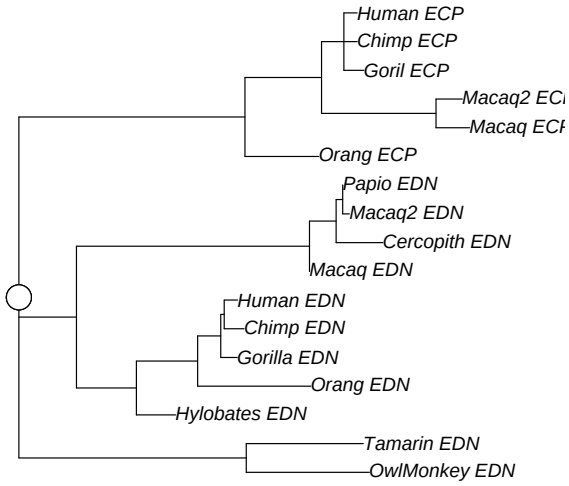


Figura 3.5: Filogenia Utilizada para a Aplicação dos Primatas.

Tabela 3.4: Resultado do Teste para as Seqüências dos Primatas.

Teste	$-2\Delta(\mathbf{X})$	Distribuição Assintótica			Simulação de Monte Carlo			Decisão
		r	p-valor	Poder	p-valor	Valor Crítico 99%	Poder	
JC69 × K80	32.0992	1	1.4650×10^{-8}	0.9343	< 0.0010	6.7860	1	rejeita
F81 × HKY85	33.1716	1	8.4374×10^{-9}	0.9412	< 0.0010	6.2904	0.999	rejeita
JC69 × F81	8.2774	3	4.0613×10^{-2}	0.4707	0.0440	11.6883	0.468	aceita
K80 × HKY85	9.3498	3	2.4984×10^{-2}	0.5340	0.0270	11.6545	0.536	aceita
HKY85 × GTR	29.4102	4	6.4520×10^{-6}	-	< 0.0010	13.4503	0.977	rejeita

Assim como na aplicação dos Felinos, consideramos os modelos JC69, K80, F81, HKY85 e GTR. Os parâmetros dos modelos estimados pelo método da máxima verossimilhança estão apresentados na Tabela 3.3. Notamos que a parametrização utilizada para os modelos K80 e HKY85 difere daquela dada na Seção 2.3, e está apresentada no exemplo do Apêndice A. Os modelos utilizados foram comparados nos seguintes testes: JC69 $\times$ K80, F81 $\times$ HKY85, JC69 $\times$ F81, K80 $\times$ HKY85 e HKY85 $\times$ GTR. Os resultados desses testes estão expostos na Tabela 3.4.

As distribuições da estatística do teste $-2\Delta(\cdot)$, sob H_0 e sob H_1, foram obtidas por simulação de Monte Carlo. A Figura 3.6 apresenta os histogramas das distribuições obtidas para todos os testes de hipóteses estudados. Nos histogramas sob H_0, a linha vertical pontilhada representa o valor da estatística do teste calculada para os dados $\mathbf{X}$. Já nos histogramas sob H_1, a linha pontilhada representa o valor crítico à 99% de confiança, obtido nas simulações de Monte Carlo. A Figura 3.4 apresenta os Q-Q plot comparando os quantís das distribuições de $-2\Delta(\mathbf{X})$, obtidas sob H_0, com os da distribuição χ^2, e os quantís de $-2\Delta(\mathbf{X})$, sob H_1, com os quantís da distribuição normal.

O poder do teste também foi obtido utilizando o estimador $\hat{D}$, para os quatro primeiros testes. Observamos que os valores obtidos para o poder do teste por este método foram muito próximos daqueles obtidos com as Simulações de Monte Carlo.

Observamos que nos testes de hipóteses JC69 $\times$ K80 e F81 $\times$ HKY85, que testam para o uso de diferentes taxas para *transições* e *transversões*, a hipótese nula foi sempre rejeitada. Deste modo, concluímos que, para a evolução desta família de genes nos primatas, diferentes taxas de *transições* e *transversões* tiveram um papel importante. O poder obtido para ambos os testes foi muito próximo de 1.

Já em relação aos testes que comparam modelos com e sem o pressuposto de homogeneidade nas freqüências das bases (JC69 $\times$ F81 e K80 $\times$ HKY85), a hipótese nula foi aceita em ambos os testes. Assim, notamos que as freqüências das bases ao longo dessa família gênica são suficientemente próximas das freqüências homogêneas. Portanto, para posteriores análises envolvendo a função de verossimilhança, a utilização de três parâmetros adicionais que representam as freqüências das bases aparentemente não se justifica. Ou seja, o desempenho dos modelos que não assumem que a distribuição das taxas é homogênea não é significativamente superior, de forma que, por parcimônia, optamos pelo modelo mais simples. Entretanto, observamos que o poder em ambos os testes foi próximo de 50% (ver Tabela 3.4). Portanto, existe uma probabilidade de que a hipótese alternativa esteja correta, e o teste não tem poder suficiente para rejeitar a hipótese H_0. Assim, esse resultado deve ser utilizado com cuidado.

Na comparação dos modelos JC69, K80, F81 e HKY85, com base nos testes aqui apresentados, somos levados a escolher como melhor descrição do processo de evolução dessas seqüências o modelo K80, uma vez que este modelo permite diferentes taxas de *transições* e *transversões* e assume que as freqüências das bases são homogêneas.

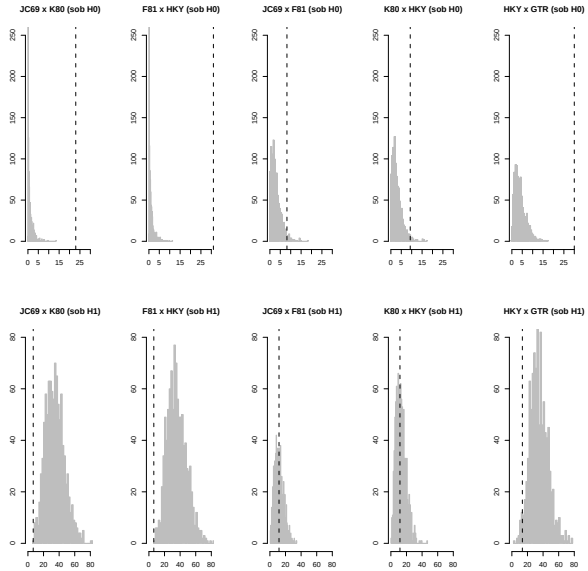


Figura 3.6: Histogramas de $-2\Delta(\mathbf{X})$ obtidos por simulações de Monte Carlo para os testes JC69 $\times$ K80, JC69 $\times$ F81, K80 $\times$ HKY85, F81 $\times$ HKY85 e HKY85 $\times$ GTR sob H_0 e H_1 para as seqüências de Primatas.

O último teste realizado foi HKY85 $\times$ GTR. Nesse teste, a hipótese nula foi rejeitada com um p-valor < 0.0010, de modo que somos levados a escolher o modelo GTR. Se compararmos o modelo K80, selecionado pelos testes anteriores, temos que a estatística do teste é $-2\Delta(\mathbf{X}) = 38.76$, e sua distribuição assintótica teórica é χ^2_7. Assim, com um p-valor de 2.1717×10^{-6} rejeitamos também o modelo K80.

Ressaltamos que o modelo mais amplamente utilizado na literatura é HKY85, seguido de modelos mais simples (ver Bielawski e Yang, 2003). Isso se deve ao fato de que, além de apresentar mais parâmetros que aumentam os tempos computacionais, modelos mais complexos que mantém o pressuposto de sítios i.i.d. são de difícil interpretação biológica. Assim, uma forma de interpretar a escolha do modelo GTR em detrimento de modelos mais simples, é que diferentes taxas de *transições* e *transversões* e variação nas freqüências das bases não são suficientes para descrever a evolução das seqüências de DNA. E, portanto, deve haver outros fatores envolvidos. O Capítulo 4 apresenta modelos que tratam de alguns desses possíveis fatores adicionais.

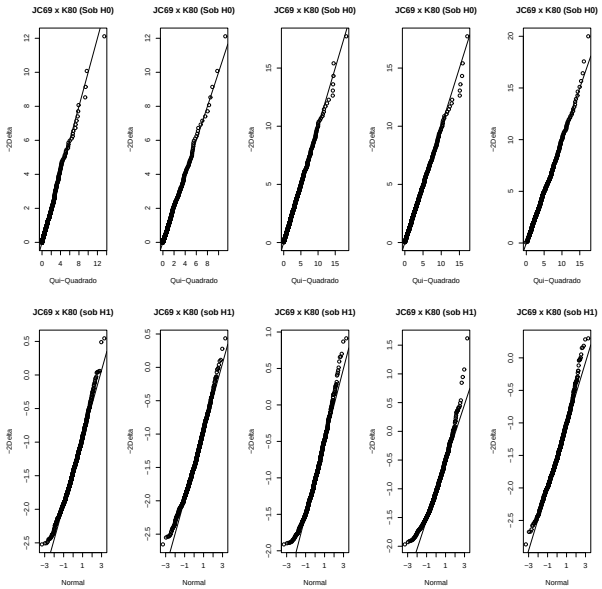


Figura 3.7: Q-Q plots de $-2\Delta(\mathbf{X})$ obtidos por simulações de Monte Carlo para os testes JC69 $\times$ K80, JC69 $\times$ F81, K80 $\times$ HKY85, F81 $\times$ HKY85 e HKY85 $\times$ GTR sob H_0 e H_1 para as seqüências de Primatas.

Capítulo 4

Variação nas Taxas de Mutação e Dependência entre Sítios

Os modelos de substituição de bases apresentados na Seção 2.3 têm em comum diversos pressupostos, tais como: a estrutura filogenética que relaciona seqüências de DNA; a evolução temporal de forma Markoviana dessas seqüências; e a evolução dos sítios à taxas constantes segundo uma mesma matriz de transição, em um processo cujas variáveis aleatórias são i.i.d.. Entretanto, essas hipóteses são muito restritas, de forma que os modelos deixam de captar diversos aspectos do complexo processo de evolução molecular.

Nesse capítulo serão apresentados modelos mais complexos para a evolução molecular de seqüências de DNA. Esses modelos são mais realistas, no sentido de permitir fenômenos já verificados em seqüências reais, como variação entre as taxas de mutação nos diferentes sítios da seqüência ou mesmo permitir dependência entre os sítios no processo de mutação. À luz desses modelos, a teoria apresentada no Capítulo 3 deve ser revisada, para que sejam feitas as devidas alterações relacionadas às peculiaridades de cada modelo.

A comparação entre os modelos apresentados nesse capítulo e aqueles da Seção 2.3 resulta em novos testes estatísticos que avaliam a importância dessas novas hipóteses (como dependência entre sítios ou variação nas taxas de mutação) para descrever o processo de evolução molecular.

4.1 Modelos com Variação da Taxa de Mutação entre os Sítios

Nessa seção serão considerados alguns modelos que relaxam a hipótese de que todos os sítios evoluem segundo uma mesma taxa. Essa hipótese não tem fundamento biológico, já que os diferentes sítios de uma seqüência de DNA podem estar submetidos a diferentes pressões evolutivas e de mutação, resultado da estrutura secundária da molécula de DNA ou de funções dos aminoácidos correspondentes em proteínas.

Estudos evolutivos realizados nos últimos 30 anos revelaram que variação nas taxas de mutação entre os sítios existe em praticamente todos os genes e proteínas, com a possível exceção de alguns pseudogenes e regiões intergênicas (Yang, 1996).

Yang et all. (1994) verificaram, por meio de simulações, que o uso de modelos que não permitem heterogeneidade entre os sítios levam a vício na estimação de parâmetros das filogenias, como a subestimação dos comprimento dos ramos.

Note que esses modelos trazem informação sobre como são distribuídas as diferentes taxas de mutação entre os sítios, mas não informam sobre o processo de evolução da seqüência em si. Desta forma, faz-se necessário aliar ao modelo uma matriz $\mathbf{Q}$ com as probabilidades de mutação entre bases e um vetor $\mathbf{p}_0$ de probabilidades iniciais, representados pelos modelos de substituição de bases apresentados na Seção 2.3.

4.1.1 Distribuição Discreta para as Taxas de Mutação

Utilizamos a notação +disc sempre que assumimos uma distribuição discreta, nas condições apresentadas nesta sub-seção, para as taxas de mutação nos diferentes sítios. Assim, o modelo JC69 com distribuição discreta para as taxas de mutação nos sítios é denotado por JC69+disc.

Nesse modelo, os sítios da seqüência de DNA são divididos em C categorias, cada uma com taxa de mutação distinta. A probabilidade de um determinado sítio pertencer à categoria l com taxa de mutação μ_l é dada por q_l. Essas probabilidades estão sujeitas à restrição $\sum_{l=1}^{C} q_l = 1$, e ainda $\sum_{l=1}^{C} \mu_l q_l = 1$. A segunda restrição fixa a média das taxas de mutação da seqüência, e equivale às restrições feitas à taxa de mutação geral dos modelos de substituição de bases do Capítulo 2, para que seja possível a estimação dos comprimentos dos ramos. Com essas restrições, o modelo com C categorias de taxas de mutação possui $2C - 1$ parâmetros além daqueles da matriz $\mathbf{Q}$ de transição. Para cada sítio, o processo de evolução da base segue o formato da filogenia, e as taxas μ_l alongam ou encurtam os ramos para cada sítio. Assim, os sítios com taxa de mutação geral μ_l têm taxas de mutação de uma base para outra dada por $\mu_l\mathbf{Q}$.

Em seqüências reais, não se conhece a taxa de mutação de cada sítio. Portanto, o modelo assume que todos os sítios têm a mesma probabilidade de pertencer a cada categoria. Dessa forma, percebe-se que apesar dos sítios evoluírem a taxas distintas o processo é i.i.d..

Assim, para um determinado sítio, a probabilidade de encontrarmos a combinação de bases $\mathbf{X}_u$ é dada por uma soma ponderada das probabilidade de encontrarmos $\mathbf{X}_u$ segundo cada uma das diferentes taxas μ_l. Temos então o modelo de misturas dado por

$$\mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}) = \sum_{l=1}^{C} q_l \, \mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, \mu_l). \tag{4.1}$$

Como a expressão é a mesma para cada sítio, a função de verossimilhança da seqüência inteira para esse modelo é dada por

$$L(F, \bar{\tau} | \mathbf{X}) = \prod_{u=1}^{S} \left(\sum_{l=1}^{C} q_l \, \mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, \mu_l) \right). \tag{4.2}$$

Note que a probabilidade $\mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, \mu_l)$ depende da matriz de taxas $\mathbf{Q}$ do modelo de substituição escolhido, e pode ser determinada pelo algoritmo de Felsenstein (1996) apresentado na Seção 3.1.1.

As taxas μ_i podem ser definidas previamente ou estimadas a partir dos dados. A definição prévia das taxas é um tanto complicada por depender apenas de sensibilidade e experiência de quem faz a análise, tendo pouco suporte nos dados. Se optarmos por estimar as taxas a partir dos dados é necessário definir previamente o número C de categorias. Note entratanto que, como esse é um modelo de mistura, em que são misturados C grupos com taxa de mutação μ_l a proporções q_l, para $l \in \{1, \cdots, C\}$, o número C de categorias para que haja identificabilidade é limitado. Yang (1995) sugere a escolha de $C \in \{3, 4\}$.

Modelos com Proporção de Sítios Invariantes

Um dos modelos com distribuição discreta para as taxas de mutação mais utilizados, proposto por Hasegawa et al. (1985), é o modelo de sítios invariantes. Esse modelo, denotado por +I (por exemplo, JC69+I, HKY85 +I, GTR +I, dependendo da matriz de transição escolhida), divide os sítios da seqüência em 2 grupos: um que segue o processo determinado por sua matriz de transição com taxa de mutação geral μ_1 e outro invariante, com taxa de mutação $\mu_0 = 0$. Assim como no caso geral, a média da taxa de mutação sobre todos os sítios é 1, ou seja, $q_1\mu_1 + q_0\mu_0 = q_1\mu_1 = 1$. Dessa forma, $\mu_1 = q_1^{-1} = (1 - q_0)^{-1}$.

Observa-se que, se um sítio não é constante, ou seja, tem bases diferentes nas diferentes seqüências, então 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. Portanto, temos

$$\mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}) = \begin{cases} q_0 + q_1 \mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, \mu_1), & \text{se o sítio é constante} \\ q_1 \mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, \mu_1), & \text{se o sítio não é constante.} \end{cases} \tag{4.3}$$

Existem diversos trechos das seqüências de DNA que são altamente conservados, de forma que, na prática funcionam como sítios invariantes. Usualmente esse fato decorre de alguma função específica desempenhada pela seqüência, que seria perdida caso a base fosse alterada. Exemplo disso são os motivos de reconhecimento (porções do DNA que dependem de uma seqüência específica para que possam ser reconhecidos por proteínas) que são imprescindíveis para o processamento da informação contida nas seqüências.

4.1.2 Distribuição Gama para as Taxas de Mutação

Outra forma de considerar variação nas taxas de mutação entre os sítios é assumir que as taxas seguem uma distribuição contínua. Várias distribuições ja foram utilizadas, como por exemplo a log-normal (ver Waddell et al., 1997) mas a distribuição mais amplamente utilizada para esse propósito é a gama. Os modelos com distribuição gama para as taxas de mutação recebem o sufixo $+\Gamma$ (por exemplo GTR+Γ, HKY85+Γ e K80+Γ). Não existe nenhuma razão biológica para a escolha desta distribuição. No entanto, o uso da gama decorre da sua versatilidade (ver Yang, 2007).

Observação 4.1. Se X é uma variável aleatória tal que $X \sim \Gamma(a, b)$ então a função densidade de probabilidade de X é dada por

$$f_X(x) = \begin{cases} \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}, & x \geq 0 \\ 0, & \text{caso contrário,} \end{cases} \tag{4.4}$$

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

$$\Gamma(a) = \int_0^\infty t^{a-1} e^{-t} dt. \tag{4.5}$$

Observe que $\mathbb{E}(X) = \frac{a}{b}$ e $\text{Var}(X) = \frac{a}{b^2}$.

Assim como no caso em que taxas têm distribuição discreta, não sabemos qual a taxa de mutação de cada sítio. Então, para calcular a probabilidade dos dados devemos fazer uma média sobre todas as possíveis taxas. Como a distribuição é contínua, isso é obtido por integração. Desta forma, temos que

$$\begin{aligned} \mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, a, b) &= \int_0^\infty f_X(\mu) \mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, \mu) d\mu \\ &= \int_0^\infty \frac{b^a \mu^{a-1} e^{-b\mu}}{\Gamma(a)} \mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, \mu) d\mu. \end{aligned} \tag{4.6}$$

A função de verossimilhança para este modelo, dada na expressão (4.6), depende dos parâmetros a e b da distribuição gama. Entretanto, assim como feito quando temos taxas

de mutação distintas e discretas, impomos que a taxa de mutação geral média nos sítios seja 1. Como a esperança da distribuição gama é dada por $\frac{a}{b}$ temos que essa condição é satisfeita se tomarmos $a = b$. Com esta restrição, tomamos $\mathbb{P}(\mathbf{X}|F, \bar{\tau}, a, b) = \mathbb{P}(\mathbf{X}|F, \bar{\tau}, a)$. Assim, além dos parâmetros do modelo de substituição de bases e dos comprimentos de ramos, o parâmetro a deve ser estimado a partir dos dados. Valores pequenos de a sugerem grande variação nas taxas de mutação, enquanto que valores grandes de a indicam que as taxas estão todas próximas a 1.

Cálculo da Função de Verossimilhança

Apresentamos a seguir o algorítmo para o cálculo da função de verossimilhança quando as taxas de mutação têm distribuição gama (ver Yang, 1993). Note que este é o mesmo algorítmo da Seção 3.1.1, com algumas alterações.

1. Para cada sítio $u \in \{1, \cdots, S\}$ das seqüências:

(a) Para cada uma das seqüências da amostra ($k \in \{1, \cdots, N\}$) e para cada $i \in E = \{A, G, C, T\}$ calcular:

$$\mathbb{P}(L_k|i, \mu)_u = I(X_u^k = i),$$

em que $I(A)$ representa a função indicadora do conjunto A.

(b) Para cada nó interno ($k \in \{N+1, \cdots, 2N-1\}$) e para cada $i \in E = \{A, G, C, T\}$ calcular:

$$\mathbb{P}(L_k|i, \mu)_u = \sum_{j,l \in E} \mathbb{P}(j|i, \tau_n, \mu)\mathbb{P}(L_n|j, \mu)_u \mathbb{P}(l|i, \tau_m, \mu)\mathbb{P}(L_m|l, \mu)_u,$$

em que n e m são os nós descendentes do nó k.

(c) Calcular a função de verossimilhança no sítio u como

$$L(F, \bar{\tau}|X_u) = \sum_{i \in E} \pi_i \mathbb{P}(L_{2N-1}|i)_u = \sum_{i \in E} \pi_i \mathbb{E}(\mathbb{P}(L_{2N-1}|i, \mu)_u).$$

2. Calcular a função de verossimilhança da árvore como $L(F, \bar{\tau}|\mathbf{X}) = \prod_{u=1}^{S} L(F, \bar{\tau}|X_u)$.

Nos passos **1a** e **1b** do algorítmo de Yang (1993), para $k \in \{1, \cdots, N\}$, é calculada a probabilidade da sub-árvore L_k que tem como raiz o nó k, dado que a base i está na raiz dessa sub-árvore e a taxa de mutação para o sítio u é μ. Essas probabilidades são dependentes de μ.

No passo **1c**, para obter a probabilidade de toda a árvore (não condicionada à taxa de mutação), calcula-se a esperança de $\mathbb{P}(L_{2N-1}|i, \mu)_u$ em relação à variável aleatória μ. Note que $\mathbb{E}(\mathbb{P}(L_{2N-1}|i, \mu)_u)$ corresponde à integral da expressão (4.6). Na prática, dependendo do modelo de substituição de bases utilizado e da filogenia em questão, pode se explicitar tal esperança.

Teste da Razão de Verossimilhança

Para verificar se um modelo de substituição de bases com distribuição gama para as taxas de mutação nos sítios (+Γ) apresenta melhora significativa na descrição dos dados em relação ao mesmo modelo com taxa de mutação constante entre os sítios, pode-se utilizar o teste da razão de verossimilhança. Dessa forma, consideramos o teste de hipóteses

H_0 : Pressupostos A1, A2, A3 e modelo M_0;

H_1 : Pressupostos A1, A2, A3 e modelo $\mathrm{M}_1 = \mathrm{M}_0 + \Gamma$,

em que M_0 pertence à classe dos modelos de substituição de bases, dados na Seção 2.3 $\mathcal{M} = \{$JC69, K80, K81, F81, TN93, HKY85, F84, GTR$\}$, e os pressupostos acima são

A1. As seqüências são relacionadas por uma estrutura filogenética;

A2. As mutações de cada sítio são independentes;

A3. As mutações são regidas por uma cadeia de Markov, com matriz $\mathbf{Q}$ de taxas infinitesimais e probabilidades iniciais $\mathbf{p}_0$.

A função do teste da razão de verossimilhança para testar H_0 versus H_1 é dada por (3.13), ou seja

$$-2\Delta(\mathbf{X}) = -2\left(\log\left(\hat{L}_0(\mathbf{X})\right) - \log\left(\hat{L}_1(\mathbf{X})\right)\right),$$

em que $\hat{L}_0(\mathbf{X})$ é a função de máxima verossimilhança do modelo sob H_0 e $\hat{L}_1(\mathbf{X})$ é a do modelo sob H_1 calculada como indicado acima.

Por meio de simulações, Whelan e Goldman (1999) encontraram que, para o teste da razão de verossimilhança que avalia o uso da distribuição gama para as taxas de mutação, a distribuição de $-2\Delta(\mathbf{X})$ é diferente da χ^2. A justificativa apresentada por esses autores para tal desvio é o fato de o parâmetro a sob H_0 encontrar-se na fronteira do espaço de parâmetros, contrariando a condição do Teorema 3.1 que garante que a distribuição da estatística do teste da razão de verossimilhança é χ^2. Note que, quando o parâmetro a da distribuição gama aumenta, a variação entre as taxas de mutação diminui; no limite, para que todos os sítios tenham a mesma taxa de mutação ($\mathrm{M}_0 + \Gamma = \mathrm{M}_0$) é necessário que a seja ∞. Nesse caso, sob H_0, a estaria na fronteira do espaço de parâmetros, pois, sob H_1, $a \in [0, \infty)$. O teorema a seguir apresenta a distribuição da estatística do teste em tais casos.

Teorema 4.1. *Suponha que a função de verossimilhança $L(\boldsymbol{\theta}|\mathbf{X})$ possui as três primeiras derivadas parciais em relação a θ_i, para todo $\theta_i \in \boldsymbol{\theta}$ (ou derivadas laterais se $\boldsymbol{\theta}$ está na fronteira do espaço de parâmetros), e que $\left|\frac{1}{S}\frac{\partial^3 L(\boldsymbol{\theta}|\mathbf{X})}{\partial\theta_i^3}\right|$ é limitado por uma função cuja esperança existe, para todo $\theta_i \in \boldsymbol{\theta}$. Além disso, suponha que a matriz de Informação de Fisher $\mathbf{I}(\boldsymbol{\theta})$ é positiva definida. Então, para o teste da razão de verossimilhança que compara $H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ versus $H_1 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_1$, tal que $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}_1$ e apenas um dos componentes de $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)$ está na fronteira do espaço de parâmetros sob H_0, temos que*

$$-2\Delta(\mathbf{X}) \sim 0.5\chi^2_{r-1} + 0.5\chi^2_r, \tag{4.7}$$

em que r é a diferença de parâmetros livres entre os modelo sob H_0 e H_1.

Demonstração: Ver Self e Liang (1987). □

Observação 4.2. No Teorema 4.1 a expressão $0.5\chi^2_{r-1} + 0.5\chi^2_r$ representa uma mistura de duas distribuições qui-quadrado independentes. Assim, se $X \sim 0.5\chi^2_{r-1} + 0.5\chi^2_r$, temos que a função densidade de X é dada por $f_X = 0.5f_Y + 0.5f_Z$, em que f_Y é a função densidade de uma variável aleatória Y, tal que $Y \sim \chi^2_{r-1}$, e f_Z é a função densidade de Z, tal que $Z \sim \chi^2_r$.

Um exemplo de tal teste seria comparar H_0 : K80 versus H_1 : F84 + Γ, nesse caso teríamos $r = 4$ pois os parâmetros livres seriam π_A, π_C, π_G e a (lembrando que $\pi_A + \pi_C + \pi_G + \pi_T = 1$), e a distribuição da estatística do teste seria $0.5\chi^2_3 + 0.5\chi^2_4$.

Já no caso do teste em que comparamos um mesmo modelo na ausência e presença da distribuição gama, por exemplo H_0: JC69 versus H_1: JC69+Γ, temos que $-2\Delta(\mathbf{X}) \sim 0.5\chi^2_0 + 0.5\chi^2_1$. Nesse caso, consideramos a distribuição χ^2_0 como aquela que assume o valor 0 com probabilidade um (degenerada em 0).

Para obter o p-valor do teste para o caso destas misturas utilizamos o seguinte resultado.

Lema 4.1. *Suponha que a função de verossimilhança $L(\boldsymbol{\theta}|\mathbf{X})$ possui as três primeiras derivadas parciais em relação a θ_i, para todo $\theta_i \in \boldsymbol{\theta}$ (ou derivadas laterais se $\boldsymbol{\theta}$ está na fronteira do espaço de parâmetros), e que $\left|\frac{1}{S}\frac{\partial^3 L(\boldsymbol{\theta}|\mathbf{X})}{\partial\theta_i^3}\right|$ é limitado por uma função cuja esperança existe, para todo $\theta_i \in \boldsymbol{\theta}$. Além disso, suponha que a matriz de Informação de Fisher $\mathbf{I}(\boldsymbol{\theta})$ é positiva definida. O p-valor do teste da razão de verossimilhança que compara diferentes modelos, com um dos parâmetros fixos na fronteira do espaço de parâmetros sob H_0, é dado por*

$$\text{p-valor}(\mathbf{X}) = \mathbb{P}\left(\frac{1}{2}\chi^2_{r-1} + \frac{1}{2}\chi^2_r < -2\Delta(\mathbf{X})\right) = \frac{\mathbb{P}\left(\chi^2_{r-1} < -2\Delta(\mathbf{X})\right) + \mathbb{P}\left(\chi^2_r < -2\Delta(\mathbf{X})\right)}{2}, \tag{4.8}$$

onde r é a diferença de parâmetros livres entre H_1 e H_0. Além disso, sob H_0, temos que $\mathbb{E}(-2\Delta(\mathbf{X})) = r - \frac{1}{2}$ *e* $\text{Var}(-2\Delta(\mathbf{X})) = r - \frac{1}{2}$.

Demonstração: Como as condições do Teorema 4.1 estão satisfeitas, temos que $-2\Delta(\mathbf{X}) \sim \frac{1}{2}\chi^2_{r-1} + \frac{1}{2}\chi^2_r$. Assim, temos que $f_{-2\Delta(\mathbf{X})}(\cdot) = \frac{1}{2}f_Y(\cdot) + \frac{1}{2}f_Z(\cdot)$, em que $Y \sim \chi^2_{r-1}$ e $Z \sim \chi^2_r$. Portanto,

$$\begin{aligned}
\text{p-valor}(\mathbf{x}) &= \mathbb{P}\left(\frac{1}{2}\chi^2_{r-1} + \frac{1}{2}\chi^2_r < -2\Delta(\mathbf{x})\right) = \int_{-\infty}^{-2\Delta(\mathbf{x})} \left(\frac{1}{2}f_Y(t) + \frac{1}{2}f_Z(t)\right) dt \\
&= \frac{1}{2}\int_{-\infty}^{-2\Delta(\mathbf{x})} f_Y(t)dt + \frac{1}{2}\int_{-\infty}^{-2\Delta(\mathbf{x})} f_Z(t)dt \\
&= \frac{\mathbb{P}\left(\chi^2_{r-1} < -2\Delta(\mathbf{x})\right) + \mathbb{P}\left(\chi^2_r < -2\Delta(\mathbf{x})\right)}{2}.
\end{aligned}$$

Além disso, como $\mathbb{E}(\chi^2_r) = r$ e $\text{Var}(\chi^2_r) = 2r$,

$$\mathbb{E}(-2\Delta(\mathbf{X})) = \mathbb{E}\left(\frac{1}{2}\chi^2_{r-1} + \frac{1}{2}\chi^2_r\right) = \frac{1}{2}\mathbb{E}(\chi^2_{r-1}) + \frac{1}{2}\mathbb{E}(\chi^2_r) = \frac{1}{2}(r-1) + \frac{1}{2}r = r - \frac{1}{2},$$

e

$$\text{Var}(-2\Delta(\mathbf{X})) = \text{Var}\left(\frac{1}{2}\chi^2_{r-1} + \frac{1}{2}\chi^2_r\right) = \frac{1}{4}\text{Var}(\chi^2_{r-1}) + \frac{1}{4}\text{Var}(\chi^2_r) = \frac{1}{4}(2r-2) + \frac{1}{4}2r = r - \frac{1}{2},$$

onde $Y \sim \chi^2_{r-1}$ e $Z \sim \chi^2_r$ são variáveis aleatórias independentes. □

Goldman e Whelman (2000) realizaram simulações de testes da razão de verossimilhança para o uso da distribuição gama, comparando um mesmo modelo com e sem a distribuição gama conforme apresentado nessa seção. Os autores também realizaram testes hierárquicos em que o modelo sob H_1 tem distribuição gama para as taxas de mutação, como por exemplo JC69 × K80+Γ. Em ambos os casos, verificaram que as misturas de distribuições χ^2, sugeridas por Self e Liang (1987), são compatíveis com os resultados das simulações.

4.1.3 Comparação entre as Distribuições Discreta e Gama para as Taxas de Mutação

Yang (1996) faz um estudo comparativo entre as vantagens e desvantagens do uso da distribuição gama e de uma distribuição discreta para as taxas de mutação dos diferentes sítios da seqüência.

Como vantagem da distribuição discreta, ele destaca que os cálculos envolvidos nas análises da função de verossimilhança são relativamente simples e rápidos. Por outro

lado, as estimativas das taxas são sensíveis à escolha do número de organismos, de forma que a interpretação do modelo se torna mais difícil. Além disso, não se pode fazer uma comparação dos resultados obtidos de seqüências 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 requerem a estimação de muitos parâmetros.

Yang (1996) destaca que o modelo que designa uma distribuição gama para as taxas de mutação possui a vantagem de explicar, através de apenas um parâmetro, a variação das taxas, além de ser de simples interpretação e possuir maior apelo biológico, por ser contínuo. Entretanto, o tempo computacional envolvido é tão grande que a utilização do modelo só é viável para amostras de até 6 seqüências.

O modelo da distribuição gama discretizada (ver Yang, 1994) é uma boa alternativa a estes dois modelos. Este modelo apresenta, além da fácil interpretabilidade e da boa aderência do modelo com distribuição gama, um tempo computacional compatível com o das taxas discretas.

Distribuição Gama Discretizada para as Taxas de Mutação

O modelo da distribuição gama discretizada, denotado por $+\Gamma_d$, utiliza C categorias para aproximar a distribuição gama (ver Yang, 1994). Assumimos que todas as categorias tem a mesma probabilidade $1/C$.

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

$$\mu_i = \frac{\int_A^B x f_X(x)dx}{\int_A^B f_X(x)dx} = \frac{\int_A^B x f_X(x)dx}{1/C}, \tag{4.9}$$

em que A é o percentíl $(i-1)/C$ da distribuição gama, B é o percentíl i/C da mesma distribuição, e $f_X(\cdot)$ é a função densidade da distribuição gama, dada na Observação 4.1. Lembramos que, assim como com o modelo apresentado na Seção 4.1.2, queremos que a taxa de mutação média seja 1. Portanto, os parâmetros da distribuição gama devem atender à condição $a = b$, de forma que os únicos parâmetros a serem estimados neste modelo são a e C. Entretanto, na maioria das aplicações o pesquisador deve escolher com quantas categorias deseja trabalhar, de forma que o único parâmetro a ser estimado é a.

Cálculo da Função de Verossimilhança

O cálculo da função de verossimilhança para este modelo se dá de forma idêntica ao caso com as taxas discretas. A única diferença é que os valores das taxas de mutação μ_i estão atrelados à distribuição gama, na forma da expressão (4.9), e que todas as categorias tem igual probabilidade. Assim, temos que

$$L(F, \bar{\tau}|\mathbf{X}) = \prod_{u=1}^{S} \left(\sum_{l=1}^{C} \frac{1}{C} \, \mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}, \mu_l) \right). \tag{4.10}$$

O custo computacional deste modelo é aproximadamente C vezes aquele do modelo em que todos os sítios evoluem às mesmas taxas.

4.1.4 Taxas de Mutação Determinadas através de uma Cadeia de Markov Oculta (HMM)

Outra forma de considerar taxas variadas de mutação nos diferentes sítios da seqüência de DNA é atribuir uma cadeia de Markov oculta ao processo de designação de taxas de mutação nos sítios. Este modelo (HMM), apresentado em Felsenstein e Churchill (1996), tem como objetivo incorporar as seguintes propriedades:

- permitir que as taxas de mutação variem entre os sítios;
- não assumir que conhecemos as taxas de mutação de cada sítio, mas sim inferí-las a partir dos dados $\mathbf{X}$;
- permitir algum tipo de correlação entre as taxas dos sítios adjacentes. (4.11)

Essas são as propriedades que os autores consideram que qualquer tratamento minimamente realista de heterogeneidade entre os sítios deve possuir.

Observe que os modelos apresentados nas Seções 4.1.1 e 4.1.2 possuem as primeiras duas propriedades. Entretanto, ambos não permitem correlação entre taxas de sítios adjacentes, uma vez que a taxa de mutação de cada sítio é retirada de uma população com determinada distribuição de taxas em um processo independente.

O modelo de Felsenstein e Churchill (1996) assume que existe um conjunto discreto de C possíveis taxas de mutação μ_u. Além disso, o modelo exige que determinemos, à priori, as probabilidades q_u de cada taxa μ_u.

As taxas de mutação são determinadas por um processo markoviano que percorre o comprimento da seqüência, designado a taxa de cada sítio. As taxas são escolhidas de $\{\mu_1, \cdots, \mu_C\}$, e assume-se que o processo markoviano é estacionário e irredutível,

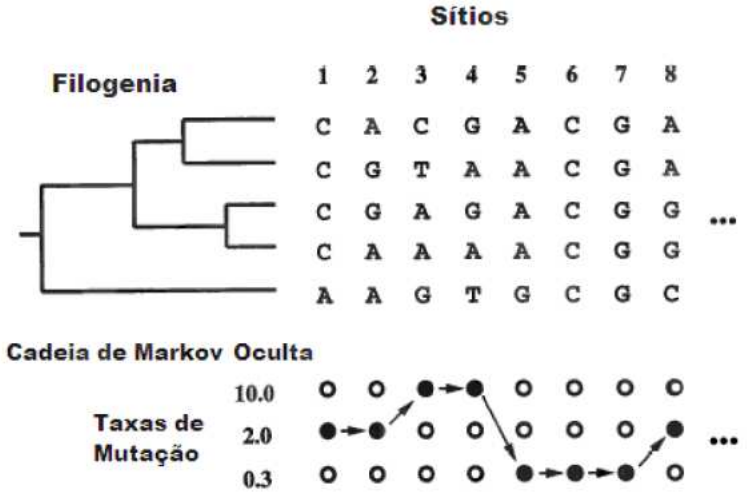


Figura 4.1: Representação Esquemática das Mudanças de Estado do Modelo HMM de Felsenstein e Churchill (1996).

existindo, desta forma, as probabilidades de equilíbrio das taxas. Assume-se que as probabilidades de transição desse processo são conhecidas. Este processo é oculto, de forma que não podemos observar as taxas de mutação de cada sítio.

Uma vez que as taxas de mutação dos sítios sejam designadas, cada sítio evolui de forma independente ao longo da filogenia F de acordo com um modelo de substituição de bases. Dessa forma, assim como com aqueles apresentados nas Seções 4.1.1 e 4.1.2, este modelo deve ser aliado a um dos modelos apresentados na Seção 2.3. Além disso, assume-se que toda correlação entre os sítios é resultado do agrupamento de taxas altas ou baixas em sítios adjacentes.

A Figura 4.1.4, retirada de Felsenstein e Churchill (1996), representa esquematicamente o processo que designa as taxas μ_u ao longo dos sítios.

Cálculo da Função de Verossimilhança

Assim como para os modelos da Seção 4.1.1, a função de verossimilhança para uma determinada filogenia F é dada pela soma, sob todas as possíveis combinações de taxas nos sítios, da probabilidade dos dados $\mathbf{X}$ condicionada à combinação de taxas multiplicada pela probabilidade à priori da combinação de taxas. Assim, se c_u denota a categoria que uma dada combinação de taxas designa para o sítio u, de forma que a taxa de mutação do sítio u é μ_{c_u}, a função de verossimilhança pode ser escrita como

$$L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{c_1}\sum_{c_2}\cdots\sum_{c_S}\mathbb{P}(c_1, c_2, \cdots, c_S) \times \mathbb{P}(\mathbf{X}|\bar{\tau}, \mu_{c_1}, \mu_{c_2}, \cdots, \mu_{c_S}). \tag{4.12}$$

O pressuposto de que, uma vez designadas as taxas de mutação, cada sítio evolui de maneira independente permite expressar a probabilidade $\mathbb{P}(\mathbf{X}|\bar{\tau}, \mu_{c_1}, \mu_{c_2}, \cdots, \mu_{c_S})$ em forma de um produto, de forma que a expressão (4.12) é reescrita por

$$L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{c_1}\sum_{c_2}\cdots\sum_{c_S}\mathbb{P}(c_1, c_2, \cdots, c_S) \times \prod_{u=1}^{S}\mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u}). \tag{4.13}$$

Como as taxas dos sítios são determinadas por uma cadeia de Markov oculta, temos que as probabilidades $\mathbb{P}(c_1, c_2, \cdots, c_S)$ podem ser escritas como

$$\mathbb{P}(c_1, c_2, \cdots, c_S) = q_{c_1}\mathbb{P}(c_2|c_1)\mathbb{P}(c_3|c_2) \times \cdots \times \mathbb{P}(c_S|c_{S-1}), \tag{4.14}$$

onde q_{c_1} é a probabilidade à priori da categoria c_1, e as probabilidades $\mathbb{P}(c_{u+1}|c_u)$ são determinadas pela matriz de transição da cadeia de Markov oculta, para todo $u \in \{1, \cdots, S-1\}$.

Observe que, para 3 categorias e seqüências de comprimento $S = 1000$, teríamos $3^{1000} \approx 10^{447}$ termos na soma de (4.13). Entretanto, os cálculos podem ser reduzidos utilizando um algoritmo semelhante ao descrito na Seção 3.1.1 para a obtenção da função de verossimilhança.

Note que substituindo (4.14) na expressão (4.13), temos que

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{X}) &= \sum_{c_1}\sum_{c_2}\cdots\sum_{c_S} q_{c_1}\mathbb{P}(c_2|c_1)\mathbb{P}(c_3|c_2) \times \cdots \times \mathbb{P}(c_S|c_{S-1}) \times \prod_{u=1}^{S}\mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u}) \\ &= \sum_{c_1} q_{c_1}\left(\sum_{c_2}\cdots\sum_{c_S}\mathbb{P}(c_2, c_3, \cdots, c_S|c_1) \times \prod_{u=1}^{S}\mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u})\right), \end{aligned} \tag{4.15}$$

pois $\mathbb{P}(c_2, c_3, \cdots, c_S|c_1) = \mathbb{P}(c_2|c_1)\mathbb{P}(c_3|c_2) \times \cdots \times \mathbb{P}(c_S|c_{S-1})$. Observe que o termo em parênteses na expressão (4.15) corresponde à função de verossimilhança para os dados $\mathbf{X}$, dado que a categoria do sítio 1 é c_1.

Definindo $L_u^{(c_u)} = \mathbb{P}(\mathbf{X}_u, \mathbf{X}_{u+1}, \cdots, \mathbf{X}_S|\bar{\tau}, c_u)$, temos que

$$L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{c_1} q_{c_1} L_1^{(c_1)} \tag{4.16}$$

onde

$$L_1^{c_1} = \mathbb{P}(\mathbf{X}_1, \cdots, \mathbf{X}_S|\bar{\tau}, c_1) = \sum_{c_2} \cdots \sum_{c_S} \mathbb{P}(c_2, c_3, \cdots, c_S|c_1) \times \prod_{u=1}^{S} \mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u}). \quad (4.17)$$

Note ainda que, utilizando as expressões (4.13) e (4.14), podemos reescrever a expressão (4.17), de forma que temos

$$\begin{aligned} L_1^{(c_1)} &= \mathbb{P}(\mathbf{X}_1|\bar{\tau}, \mu_{c_1}) \sum_{c_2} \cdots \sum_{c_S} \mathbb{P}(c_2, c_3, \cdots, c_S|c_1) \times \prod_{u=2}^{S} \mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u}) \\ &= \mathbb{P}(\mathbf{X}_1|\bar{\tau}, \mu_{c_1}) \sum_{c_2} \mathbb{P}(c_2|c_1) \sum_{c_3} \cdots \sum_{c_S} \mathbb{P}(c_3, c_4, \cdots, c_S|c_2) \times \prod_{u=2}^{S} \mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u}) \\ &= \mathbb{P}(\mathbf{X}_1|\bar{\tau}, \mu_{c_1}) \sum_{c_2} \mathbb{P}(c_2|c_1) L_2^{(c_2)}. \end{aligned} \quad (4.18)$$

Repetindo o processo realizado em (4.18), chegamos à seguinte regra recursiva

$$L_u^{(c_u)} = \mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u}) \sum_{c_{u+1}} \mathbb{P}(c_{u+1}|c_u) L_{u+1}^{(c_{u+1})}, \quad (4.19)$$

válida para $u \in \{1, \cdots, S-1\}$. Para $u = S$, define-se $L_S^{(c_S)} = \mathbb{P}(\mathbf{X}_S|\bar{\tau}, \mu_{c_S})$.

Desta forma, o seguinte algoritmo, proposto por Felsenstein e Churchill (1996), pode ser usado para calcular a função de verossimilhança deste modelo, através dos passos:

1. Para cada categoria $c \in \{1, \cdots, C\}$ de taxas de mutação, calcular $L_S^{(c_S)} = \mathbb{P}(\mathbf{X}_S|\bar{\tau}, \mu_{c_S})$;

2. Para cada sítio $i \in \{S-1, S-2, \cdots, 1\}$ e para cada categoria $c \in \{1, \cdots, C\}$ de taxas de mutação, calcular
$L_u^{(c_u)} = \mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u}) \sum_{c_{u+1}} \mathbb{P}(c_{u+1}|c_u) L_{u+1}^{(c_{u+1})}$;

3. Calcular a função de verossimilhança de $\mathbf{X}$ como $L(\boldsymbol{\theta}|\mathbf{X}) = \sum_{c_1} q_{c_1} L_1^{(c_1)}$, onde c_1 representa a categoria do sítio 1.

Note que as probabilidades $\mathbb{P}(\mathbf{X}_u|\bar{\tau}, \mu_{c_u})$ são calculadas de acordo com o modelo de substituição de bases escolhido, utilizando o algoritmo da Seção 3.1.1.

Uma questão interessante sobre o processo é qual a combinação de taxas mais provável. Como a função de verossimilhança foi obtida por uma soma da função de verossimilhança de cada combinação de taxas, podemos obter a combinação de taxas mais provável como

aquela que teve a maior contribuição no valor da função de verossimilhança, ou seja, $\max_{c_1,c_2,\cdots,c_S} \mathbb{P}(c_1, c_2, \cdots, c_S)\mathbb{P}(\mathbf{X}|\bar{\tau}, \mu_1, \cdots, \mu_S)$. Um algoritmo semelhante ao utilizado para o cálculo da função de verossimilhança deste modelo foi apresentado por Felsenstein e Churchill (1996) para a determinação de tal combinação de taxas.

Caso Particular de Taxas para o Modelo HMM

Diversas cadeias de Markov podem ser utilizadas para o processo oculto que designa as taxas de mutação dos sítios da molécula de DNA no modelo HMM. Apresentamos aqui, como exemplo, o modelo implementado nos pacotes DNAML e DNAMLK distribuídos por Joseph Felsenstein através do site *http://evolution.genetics.washington.edu/phylip.html*.

O programa permite que o usuário especifique o número C de diferentes categorias, as taxas de mutação μ_u de cada categoria, e as freqüências q_u de equilíbrio de cada categoria. Note que $\sum_{u=1}^{C} q_u = 1$. Além disso, deve-se especificar o parâmetro de autocorrelação λ, que representa a probabilidade de que a taxa de mutação de um determinado sítio seja a mesma do sítio anterior. Com probabilidade $1 - \lambda$ o processo de Markov oculto escolhe uma nova categoria para o sítio, retirando-a de uma urna que contém todas as categorias nas freqüências q_u.

As probabilidades de transição da cadeia de Markov oculta são dadas por

$$\mathbb{P}(c_u|c_v) = \lambda\delta_{u,v} + (1 - \lambda)q_u, \tag{4.20}$$

em que $\delta_{u,v}$ representa a função delta de Kronecker, que assume valor 1, se $u = v$, e 0, caso contrário. Note que, se o sítio i pertence à categoria c_u, com probabilidade $\lambda + (1 - \lambda)q_u$ o sítio $i + 1$ pertencerá à mesma categoria. A distribuição de equilíbrio para esse modelo é composta, justamente, das freqüências q_u determinadas para as taxas.

No caso de $C = 3$, por exemplo, a matriz de transição para a cadeia oculta desse modelo é dada por

$$\begin{array}{c} \\ \mathbf{P}_{\text{HMM}} = \\ \\ \end{array} \begin{array}{c} \begin{array}{ccc} c_1 & c_2 & c_3 \end{array} \\ \begin{pmatrix} \lambda + (1-\lambda)q_1 & (1-\lambda)q_2 & (1-\lambda)q_3 \\ (1-\lambda)q_1 & \lambda + (1-\lambda)q_2 & (1-\lambda)q_3 \\ (1-\lambda)q_1 & (1-\lambda)q_2 & \lambda + (1-\lambda)q_3 \end{pmatrix} \end{array}. \tag{4.21}$$

Felsenstein e Churchill (1996) destacam que as taxas μ_u, as probabilidades q_u e o parâmetro de autocorrelação λ podem ser estimados utilizando o algoritmo EM de Baum et al. (1970). Entretanto, a implementação de tal algoritmo para estimar as taxas μ_u traria um aumento significativo no esforço computacional. Os autores comentam que, na prática, é melhor analisar os dados com um conjunto de diferentes taxas e valores para λ, e escolher aqueles que geram maior valor para a função de verossimilhança.

Conforme comentado anteriormente, este modelo para a determinação das taxas de

mutação pode ser aliado a qualquer um dos modelos de substituição de bases apresentados na Seção 2.3. Nos pacotes DNAML e DNAMLK, o modelo utilizado é o Felsenstein (1984) denotado por F84 na Seção 2.3.

As diferentes taxas de mutação nos sítios são aplicadas nos ramos das filogenias. Desta forma, todos os sítios evoluem de acordo com a mesma matriz de transição $\mathbf{P}_{\mathrm{F84}}$ do processo de substituição de bases. Entretanto, para todos os sítios de uma determinada categoria, os ramos da filogenia são multiplicados pela taxa de mutação μ. Assim, para sítios com taxa de mutação maior, a filogenia possui ramos mais longos, já para os sítios com taxa de mutação menor, a filogenia com mesma topologia possui ramos mais curtos.

Felsenstein e Churchill (1996) destacam que o tempo computacional necessário para este modelo é C vezes maior do que o tempo necessário para a mesma análise de máxima verossimilhança quando temos apenas uma taxa. Assim, se tivermos $C = 4$, o tempo computacional envolvido na estimação de filogenias, por exemplo, é quatro vezes maior do que a mesma estimação realizada quando a taxa de mutação é única.

4.2 Modelos com Dependência entre Sítios

A maioria dos modelos probabilísticos para a evolução de seqüências de DNA assume que os sítios evoluem independentemente, ou seja, a mutação ocorrida em um sítio não afeta as probabilidades de mutação dos demais sítios da seqüência. Esta é uma hipótese importante para a utilização de tais modelos na análise molecular por máxima verossimilhança, uma vez que simplifica os cálculos e reduz significativamente o tempo computacional envolvido (ver Yang, 2005). Entretanto, em muitas situações reais, essa hipótese é claramente violada. Dessa forma, são úteis modelos que consideram diferentes formas de dependência entre os sítios da seqüência.

4.2.1 Modelos Dupla Fita

Algumas Informações sobre o RNA

Seqüências de DNA que codificam moléculas de RNA ribossômico (rRNA) são freqüentemente utilizadas em análises filogenéticas que relacionam grupos de organismos muito distintos (com alto grau de divergência). Isso se deve ao fato do rRNA ser altamente conservado entre todos os seres vivos, de forma que é possível alinhar seqüências de rRNA de espécies muito distantes.

As moléculas de RNA são seqüências lineares de nucleotídeos (assim como no DNA). Entretanto, em vez de serem nucleotídeos de DNA, são de RNA. Os nucleotídeos de RNA possuem três das mesmas bases do DNA (A, C e G) além da base uracila (U). Há uma

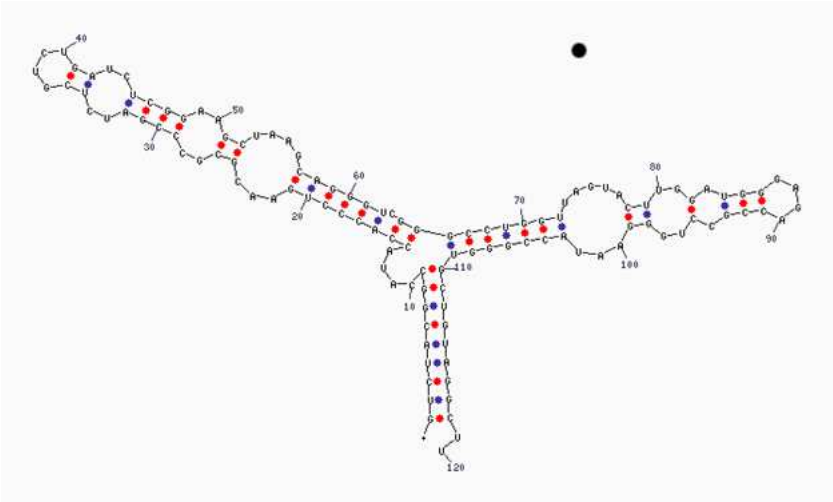


Figura 4.2: Estrutura secundária da porção 5S do ribossomo.

correspondência entre as bases do DNA e as de seu respectivo RNA. Assim, se a seqüência de DNA for A-A-G-T-C, a seqüência de RNA correspondente será A-A-G-U-C[1].

Além disso, a molécula de rRNA desempenha uma importante função na maquinaria celular, resultado da sua conformação tri-dimensional (estrutura secundária). Esta estrutura é determinada principalmente pela seqüência das bases, especialmente por pareamento dos nucleotídeos de RNA de porções diferentes da seqüência, causando um dobramento da molécula. A Figura 4.2 apresenta a estrutura secundária da porção 5S do ribossomo, e ilustra o pareamento de diferentes sítios da molécula de rRNA.

Como o pareamento é importante para a função da molécula de rRNA, uma mutação em uma das bases do par gera uma grande pressão seletiva para que no outro sítio do par ocorra uma mutação compensatória para restabelecer o pareamento. Por exemplo, se temos um par U = A e ocorrer uma mutação na primeira base para G, de forma que temos G = A, o pareamento poderá ser desfeito e a molécula poderá perder sua função. Assim, por pressão seletiva, há um aumento na probabilidade de verificarmos uma mutação para C no segundo sítio do par, de forma a restabelecer o pareamento, agora com G = C.

Portanto, as moléculas de rRNA claramente contrariam a hipótese de independência entre os sítios. Entretanto, pelo fato da estrutura do rRNA ser altamente conservada e bastante estudada, conhecem-se os sítios em que os pareamentos ocorrem. Com isso em vista, Tillier e Collins (1995) propuseram um modelo para tratar os casos de pareamento nas moléculas de RNA, e o chamaram de modelo fita dupla. Esse modelo aplica-se apenas aos sítios em que há pareamento, devendo os outros sítios serem tratados com os modelos de substituição de bases convencionais.

[1]Note que a correspondente da base T do DNA é a base U do RNA. O RNA não possui T e o DNA não possui U. Em seqüências de DNA os pares complementares são A = T e C = G, já no RNA os pares complementares são A = U e C = G.

Apresentação do Modelo Dupla Fita

Para o modelo fita dupla, a unidade de evolução que está sujeita a mutações com probabilidades determinadas pela matriz de transição é o par de bases. Dessa forma, temos 16 possíveis estados para o processo ($\mathrm{A}=\mathrm{A}, \mathrm{A}=\mathrm{C}, \mathrm{A}=\mathrm{G}, \mathrm{A}=\mathrm{U}, \mathrm{G}=\mathrm{A}, \cdots, \mathrm{U}=\mathrm{U}$), e a matriz de transição desse processo tem dimensão 16×16. O modelo assume que as mutações são independentes em sítios não pareados entre si, assim recaímos facilmente em um modelo semelhante aos modelos de substituição de bases apresentados na Seção 2.3, com a diferença que o espaço de estados E tem cardinalidade 16 em vez de 4. Desta forma, no caso mais geral que mantém a reversibilidade no tempo, temos 119 parâmetros livres para serem estimados, além das freqüências das bases e dos comprimentos dos ramos da filogenia. Entretanto, se forem consideradas algumas peculiaridades do pareamento de bases no RNA, o número de parâmetros pode ser reduzido.

Por estarmos tratando de bases pareadas, na grande maioria das vezes que se considera um desses sítios encontra-se um dos pares A = U, U = A, C = G e G = C, pois são essas as combinações que formam ligações químicas estáveis entre as bases. O par G = U (ou U = G), apesar de não ser complementar como os anteriores, as vezes é encontrado nas seqüências formando uma ligação instável. Muitas vezes esse par serve de passo intermediário para uma mutação entre pares complementares distintos. Todos os outros pares de bases raramente são encontrados no rRNA. Assim, o modelo de dupla fita designa a mesma probabilidade para todos eles, tratando-os como um só estado do processo. Definimos os estado $O = T$ do processo como o conjunto de todos os pares de bases não complementares e diferentes de $U = G$ e $G = U$. Deste modo, os estados do processo são os pares de bases A = U, G = U, G = C, U = A, U = G, C = G e O = T. Suas respectivas freqüências são dadas por π_1, π_2, π_3, π_4, π_5, π_6 e π_7.

A Figura 4.3, obtida de Tillier e Collins (1994), representa esquematicamente as mutações do processo dupla fita. Nela, os pares de bases são divididos em três caixas dependendo de a primeira base do par ser *purina* ou *pirimidina*, ou o par pertencer ao estado O = T. Dentro das duas primeiras caixas, uma transição simples para formar o par G = U ocorre com taxa inifitesimal α_s, e dupla transição para sair de uma par de bases complementar para outro ocorre com taxa α_d. As mutações entre esses dois grupos ocorrem com taxa β, e as mutações de um desses grupos para o estado O = T, com taxa γ.

A matriz de taxas infinitesimais deste processo é dada por

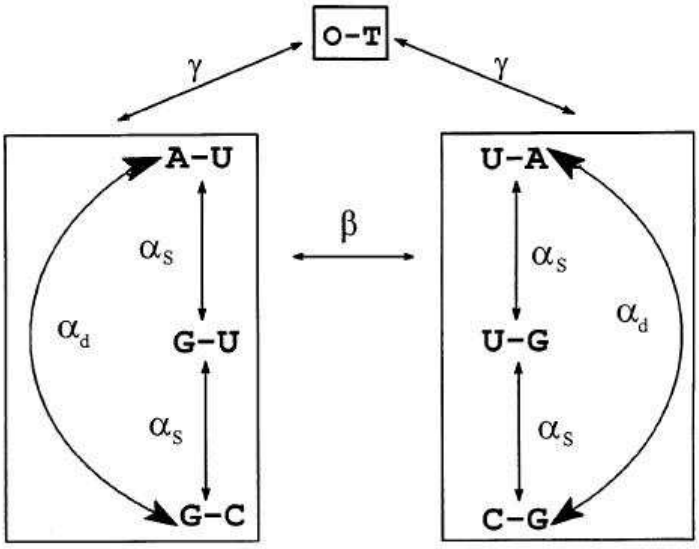


Figura 4.3: Representação esquemática das mudanças de estado do modelo de Tillier e Collins (1994).

$$\mathbf{Q}_{\text{df}} = \begin{array}{c} \begin{array}{ccccccc} \text{A}=\text{U} & \text{G}=\text{U} & \text{G}=\text{C} & \text{U}=\text{A} & \text{U}=\text{G} & \text{C}=\text{G} & \text{O}=\text{T} \end{array} \\ \begin{pmatrix} m_1 & \alpha_s\pi_2 & \alpha_d\pi_3 & \beta\pi_4 & \beta\pi_5 & \beta\pi_6 & \gamma\pi_7 \\ \alpha_s\pi_1 & m_2 & \alpha_s\pi_3 & \beta\pi_4 & \beta\pi_5 & \beta\pi_6 & \gamma\pi_7 \\ \alpha_d\pi_1 & \alpha_s\pi_2 & m_3 & \beta\pi_4 & \beta\pi_5 & \beta\pi_6 & \gamma\pi_7 \\ \beta\pi_1 & \beta\pi_2 & \beta\pi_3 & m_4 & \alpha_s\pi_5 & \alpha_d\pi_6 & \gamma\pi_7 \\ \beta\pi_1 & \beta\pi_2 & \beta\pi_3 & \alpha_s\pi_4 & m_5 & \alpha_s\pi_6 & \gamma\pi_7 \\ \beta\pi_1 & \beta\pi_2 & \beta\pi_3 & \alpha_d\pi_4 & \alpha_s\pi_5 & m_6 & \gamma\pi_7 \\ \gamma\pi_1 & \gamma\pi_2 & \gamma\pi_3 & \gamma\pi_4 & \gamma\pi_5 & \gamma\pi_6 & m_7 \end{pmatrix} \end{array}, \tag{4.22}$$

em que os parâmetros α_s, α_d, β, γ e π_i , $i \in \{1, 2, \cdots, 7\}$ são os mesmos que aparecem na Figura 4.3, e m_i é tal que a soma da linha i é 0, para $i \in \{1, \cdots, 7\}$.

Assim, pode-se obter a matriz das probabilidades de mutação $\mathbf{P}_{\text{df}}$, dada por

$$\mathbf{P}_{\text{df}} = \begin{pmatrix} m_\tau^1 & \alpha_{s,\tau}^2 & \alpha_{d,\tau}^3 & \beta_\tau^4 & \beta_\tau^5 & \beta_\tau^6 & \gamma_\tau^7 \\ \alpha_{s,\tau}^1 & m_\tau^2 & \alpha_{s,\tau}^3 & \beta_\tau^4 & \beta_\tau^5 & \beta_\tau^6 & \gamma_\tau^7 \\ \alpha_{d,\tau}^1 & \alpha_{s,\tau}^2 & m_\tau^3 & \beta_\tau^4 & \beta_\tau^5 & \beta_\tau^6 & \gamma_\tau^7 \\ \beta_\tau^1 & \beta_\tau^2 & \beta_\tau^3 & m_\tau^4 & \alpha_{s,\tau}^5 & \alpha_{d,\tau}^6 & \gamma_\tau^7 \\ \beta_\tau^1 & \beta_\tau^2 & \beta_\tau^3 & \alpha_{s,\tau}^4 & m_\tau^5 & \alpha_{s,\tau}^6 & \gamma_\tau^7 \\ \beta_\tau^1 & \beta_\tau^2 & \beta_\tau^3 & \alpha_{d,\tau}^4 & \alpha_{s,\tau}^5 & m_\tau^6 & \gamma_\tau^7 \\ \gamma_\tau^1 & \gamma_\tau^2 & \gamma_\tau^3 & \gamma_\tau^4 & \gamma_\tau^5 & \gamma_\tau^6 & m_\tau^7 \end{pmatrix}, \tag{4.23}$$

em que a probabilidade de mutação de uma caixa para outra (ou seja, de um par de bases

em que há uma *purina* na primeira posição para outro em que há uma *pirimidina* na primeira posição) é dada por

$$\beta_\tau^i = \pi_i \left(1 + f_1 \frac{\pi_7}{1-\pi_7} - f_2\right), \tag{4.24}$$

com f_1 e f_2 definidos em (4.29). Já a probabilidade de mutação de um par de bases complementar para os intermediários G = U ou U = G ou vice versa é dado por

$$\alpha_{s,\tau}^i = \pi_i \left(1 + f_1 \frac{\pi_7}{1-\pi_7} + f_2 - f_3\right), \tag{4.25}$$

para todo $i \in \{1, \cdots, 6\}$, com f_1, f_2 e f_3 definidos em (4.29). A probabilidade de mutar de um par complementar para o outro da mesma caixa é

$$\alpha_{d,\tau}^i = \pi_i \left(1 + f_1 \frac{\pi_7}{1-\pi_7} + f_2 + \frac{\pi_2}{\pi_1+\pi_3} f_3 - \pi_4\right), \tag{4.26}$$

se $i \in \{1, 3\}$ e

$$\alpha_{d,\tau}^i = \pi_i \left(1 + f_1 \frac{\pi_7}{1-\pi_7} + f_2 + \frac{\pi_5}{\pi_4+\pi_6} f_3 - \pi_4\right), \tag{4.27}$$

se $i \in \{4, 6\}$, com f_1, f_2 e f_3 definidos em (4.29). E a probabilidade de mutação de um par de bases instável, da categoria O = T, para um par de bases mais estável (representados pelos demais estados do processo) é dada por

$$\gamma_\tau^i = \pi_i f_1, \tag{4.28}$$

para todo $i \in \{1, \cdots, 7\}$, e os elementos m_τ^i são tais que a soma da linha i é 1. Ainda, temos que (ver Tillier e Collins, 1995)

$$\begin{aligned} f_1 &= e^{-\gamma\tau} \\ f_2 &= \frac{1}{1-\pi_7} e^{-(\gamma\pi_7 + (1-\pi_7)\beta)\tau} \\ f_3 &= \frac{2}{1-\pi_7} e^{-(\gamma\pi_7 + \frac{1}{2}(1-\pi_7)(\beta+\alpha_s))\tau}. \end{aligned} \tag{4.29}$$

Teste da Razão de Verossimilhança

O modelo dupla fita, apresentado nesta sub-seção, tem como pressuposto a possibilidade de que ocorram duas mutações em tempo infinitesimal. Esta é uma característica controversa, uma vez que usualmente assume-se que, em tempos infinitesimais, apenas uma mutação deve ocorrer. O argumento de Tillier e Collins (1994) é que, para que a integridade da molécula seja mantida, é necessário que sítios pareados tenham bases complementares ou $G = U$ (ou seja, um par que não pertença à categoria $O = T$). Assim, se ocorresse uma primeira mutação que desmanchasse o par complementar, uma segunda mutação no outro sítio do par que reestabelecesse a complementariedade seria altamente favorecida (essas mutações são conhecidas como mutações compensatórias). Por outro lado, se a segunda mutação não ocorresse, o par seria rapidamente eliminado pela seleção natural, e assim, não o encontraríamos na natureza. Esse é um aspecto importante do modelo e, por ser controverso, é interessante ser testado.

Queremos utilizar o teste da razão de verossimilhança para avaliar o modelo dupla fita. Para tanto, devemos primeiro obter a função de verossimilhança para este modelo. Lembramos que a unidade que sofre evolução nesse modelo não é um sítio, mas sim dois sítios pareados. Ainda, notamos que cada par de sítios evolui independentemente dos outros pares e segundo as mesmas taxas. Desse modo, para obter a função de verossimilhança desse modelo, podemos utilizar o algoritmo apresentado na Seção 3.1.1, sem alterações. Note que as expressões (4.24) a (4.28) nos dão as probabilidades de mutação de um par de bases para outro, necessárias ao algoritmo.

Temos que as hipóteses do teste da razão de verossimilhança que avalia a possibilidade de ocorrerem mutações compensatórias em tempos infinitesimais são

$$H_0\text{: Pressupostos A1, A2, A3 e } \beta = 0;$$
$$H_1\text{: Pressupostos A1, A2, A3,}$$

em que

A1. As seqüências são relacionadas por uma estrutura filogenética;

A2. As mutações de cada par de sítio são independentes e identicamente distribuídas;

A3. As mutações são regidas por uma cadeia de Markov, com matriz de taxas infinitesimais $\mathbf{Q}_{\text{df}}$.

Note que o parâmetro β representa as mutações compensatórias, como pode ser visto na Figura 4.3. Temos que, quando $\beta = 0$, não ocorrem mutações compensatórias em tempos infinitesimais. Mas $\beta = 0$ representa a fronteira do espaço de parâmetros, uma vez que $0 \leq \beta < \infty$. Assim, recaímos no caso apresentado no Teorema 4.1, em que a distribuição da estatística do teste $-2\Delta(\mathbf{X})$, dada em (3.13), é obtida por uma mistura $0.5\chi^2_0 + 0.5\chi^2_1$.

4.2.2 Correlação entre Pares de Sítios

O modelo dupla fita trata de dependência entre as mutações ocorridas em sítios distintos, no caso específico do RNA ribossômico. Entretanto, em diversas outras condições, seqüências de DNA podem apresentar dependência entre os sítios. Deste modo, gostaríamos de um modelo mais geral que considerasse a dependência entre os sítios.

Schöniger e van Haesler (1994) sugerem que a ferramenta matemática adequada para tratar de dependência entre sítios, no caso geral, deve ser as cadeias de Markov de ordem superior. Assim, o estado (base) do sítio i depende da seqüência dos v sítios anteriores (e em geral, poderia depender das bases de sítios seguintes também), em que v representa a ordem da cadeia da Markov. Tal modelo ainda deveria preservar ordem da cadeia de Markov ao longo da seqüência, e considerar correlação entre os sítios vizinhos.

Tavaré e Giddings (1989), utilizam a teoria da informação para estimar o valor de v, obtendo $v = 1$ ou 2, dependendo da região da seqüência, para o DNA do bacteriófago λ. Note que, quando $v = 0$, os processos de mutação nos diferentes sítios são independentes, o que corresponde aos modelos apresentados nas Seções 2.3 e 4.1.

Schöniger e van Haesler (1994) argumentam que este modelo não é matematicamente, ou computacionalmente, viável, e assim apresentam uma simplificação, o modelo com correlação entre pares de sítios. Este modelo considera a existência de correlação entre pares de sítios que não se sobrepõem.

Um exemplo importante em que o modelo com correlação entre pares de sítios se aplica é o de seqüências codificantes, em que se considera apenas as primeiras duas bases de cada códon. Neste caso, os pares trariam informações importantes sobre os aminoácidos da seqüência codificada, uma vez que as duas primeiras posições do códon são as mais importantes na determinação do aminoácido (ver Seção 2.1). Pela mesma razão, é razoável assumir que a probabilidade de uma mutação em um dos sítios da dupla é afetada pela base do outro sítio.

Note que o modelo dupla fita também é um caso particular do modelo com correlação entre pares de sítios. Neste caso, os pares correlacionados não são de sítios adjacentes, mas sim os sítios que pareiam na estrutura tri-dimensional do rRNA.

O modelo com correlação entre pares de sítios considera uma seqüência de DNA de comprimento S, constituída de $S/2$ pares de bases. Assim, temos que a base do sítio $2i$ é dependente da base do sítio $2i - 1$, e essas duas bases são independentes do restante da molécula. A matriz de transição da cadeia de Markov, à tempo discreto, que determina a base do sítio $2i$, a partir da base do sítio $2i - 1$, é chamada de matriz de autocorrelação $\boldsymbol{\rho}$. Ela pode ser determinada pelas freqüências de cada par de bases, de forma que temos

$$\boldsymbol{\rho} = \begin{pmatrix} \pi_{AA} & \pi_{AG} & \pi_{AC} & \pi_{AT} \\ \pi_{GA} & \pi_{GG} & \pi_{GC} & \pi_{GT} \\ \pi_{CA} & \pi_{CG} & \pi_{CC} & \pi_{CT} \\ \pi_{TA} & \pi_{TG} & \pi_{TC} & \pi_{TT} \end{pmatrix}, \tag{4.30}$$

em que π_{ij} representa a freqüência do par de bases (i, j) nas primeiras duas posições do códon.

Com base na matriz $\boldsymbol{\rho}$, o processo evolutivo é definido como aquele que preserva as freqüências dos pares. Assumimos, ainda, que não ocorrem mutações nas duas bases do par em um tempo infinitesimal, e que o processo atinge o estado estacionário. Deste modo, podemos descrever a evolução dos pares independentes da seqüência através de um processo markoviano com 16 estados, semelhante ao apresentado para o modelo dupla fita. Os estados desse processo serão os 16 possíveis pares de bases $M = \{AA, AG, AC, \cdots, TT\}$, e a matriz $\mathbf{Q}_{\text{cp}} = (\vartheta_{\nu,\upsilon})_{\nu,\upsilon \in M}$ de taxas infinitesimais que governa as mutações de um par de bases para outro tem entradas dadas por

$$\vartheta_{\nu,\upsilon} = \begin{cases} \pi_{\upsilon}, & \text{se } \nu \text{ e } \upsilon \text{ diferem por apenas uma base,} \\ -\sum_{\upsilon \in M(\nu \neq \upsilon)} \vartheta_{\nu,\upsilon}, & \text{se } \nu = \upsilon, \\ 0, & \text{caso contrário.} \end{cases}$$

Note que este modelo assemelha-se ao modelo F81 apresentado na Seção 2.3, pois ele mantém as freqüências dos estados (pares de bases) e consequentemente mantém as freqüências das bases, mas não assume que estas freqüências sejam homogêneas. Além disso, ele não permite probabilidades distintas para transições e transversões e possui a propriedade de reversibilidade no tempo.

O modelo poderia ser alterado para incorporar diferentes probabilidades de transições e transversões, de forma semelhante aos modelos HKY85 e F84. Para tanto, as entradas da matriz de taxas infinitesimal $\mathbf{Q}_{\text{cp}} = (\vartheta_{\nu,\upsilon})_{\nu,\upsilon \in M}$, deveriam ser dadas por

$$\vartheta_{\nu,\upsilon} = \begin{cases} k\pi_{\upsilon}, & \text{se } \nu \text{ e } \upsilon \text{ diferem por apenas uma } \textit{transição}, \\ \pi_{\upsilon}, & \text{se } \nu \text{ e } \upsilon \text{ diferem por apenas uma } \textit{transversão}, \\ -\sum_{\upsilon \in M(\nu \neq \upsilon)} \vartheta_{\nu,\upsilon}, & \text{se } \nu = \upsilon, \\ 0, & \text{caso contrário.} \end{cases}$$

Teste para Dependência entre Sítios (Caso uma Seqüência)

Schöniger e von Haesler (1994) sugerem a utilização do teste qui-quadrado para verificar a dependência entre os sítios no caso de uma única seqüência. Para tanto, eles sugerem que as entradas da matriz $\boldsymbol{\rho}$ sejam estimadas a partir dos dados, utilizando os estimadores de máxima verossimilhança. No caso de uma única seqüência estes estimadores são dados por

$$\hat{\pi}_\nu = \frac{\text{número de pares } \nu}{S/2}.$$

Note que, se vale a independência entre os sítios, então temos que $\pi_\nu = \pi_{ij} = \pi_i^{(1)}\pi_j^{(2)}$, em que π_{ij} representa a freqüência do par (i,j), $\pi_i^{(1)}$ representa a freqüência da base i na primeira posição dos códons, e $\pi_j^{(2)}$ representa a freqüência da base j na segunda posição dos códons. Desta forma, $\pi_i^{(1)}$ pode ser estimado como

$$\hat{\pi}_i^{(1)} = \frac{\text{número de bases } i \text{ nos primeiros sítios dos códons}}{S/2},$$

e $\pi_i^{(2)}$ pode ser estimado de maneira análoga. Desse modo, definimos o seguinte teste de hipóteses:

$$\begin{aligned} &H_0\text{: Todos os sítios são independentes, ou seja, } \boldsymbol{\rho} = (\pi_{ij})_{i,j} = (\pi_i^{(1)}\pi_j^{(2)})_{i,j};\\ &H_1\text{: Há dependência entre o primeiro e segundo sítio de cada códon, ou}\\ &\quad\text{seja, } \boldsymbol{\rho} = (\pi_{ij})_{i,j} \neq (\pi_i^{(1)}\pi_j^{(2)})_{i,j}. \end{aligned} \tag{4.31}$$

A estatística do teste qui-quadrado para essas hipóteses é dada por

$$\mathcal{T} = \frac{S}{2}\sum_{i\in E}\sum_{j\in E}\frac{\left(\hat{\pi}_i^{(1)}\hat{\pi}_j^{(2)} - \hat{\pi}_{ij}\right)^2}{\hat{\pi}_i^{(1)}\hat{\pi}_j^{(2)}}, \tag{4.32}$$

onde $\mathcal{T} \sim \chi^2_r$, com $r = 9$.

Note que, assim como para o teste da razão de verossimilhança, para o teste qui-quadrado, os graus de liberdade são determinados pela diferença entre parâmetros livres dos dois modelos. Sob H_1, os parâmetros do modelo são $\pi_{AA}, \pi_{AG}, \pi_{AC}, \cdots, \pi_{TT}$, mas eles estão sujeitos à restrição de que $\sum_{\nu\in M}\pi_\nu = 1$, de forma que temos 15 parâmetros livres. Sob H_0, os parâmetros do modelo são $\pi_A^{(1)}, \pi_G^{(1)}, \pi_C^{(1)}, \pi_T^{(1)}$ e $\pi_A^{(2)}, \pi_G^{(2)}, \pi_C^{(2)}, \pi_T^{(2)}$, sujeitos às restrições $\sum_{i\in E}\pi_i^{(1)} = 1 = \sum_{i\in E}\pi_i^{(2)}$, de forma que ficamos com 6 parâmetros livres. Desta forma, temos que $r = 15 - 6 = 9$, e assim, obtemos que $\mathcal{T}$ tem distribuição assintótica χ^2_9. Assim, como o quantil 99% da distribuição χ^2_9 é dado por 21.96, temos

que, se $\mathcal{T} > 21.96$, rejeitamos H_0 ao nível de 99% de confiança, e concluímos que há dependência entre os dois primeiros sítios do códon.

Schöniger e von Haesler (1994) ainda apresentam outro teste qui-quadrado que utiliza duas seqüências homólogas para testar a homogeneidade do processo. Tal teste é realizado comparando as matrizes $\boldsymbol{\rho}$ obtidas em cada seqüência.

Teste para Dependência entre Sítios (Caso Várias Seqüências)

O teste sugerido por Schöniger e von Haesler (1994) não considera a estrutura filogenética, de forma que não deve ser utilizado para mais de duas seqüências. Com isso em vista, e na linha do trabalho realizado com os modelos apresentados até aqui, sugerimos a utilização do teste da razão de verossimilhança para testar a dependência entre os sítios. Não só o teste da razão de verossimilhança pode ser utilizado para qualquer número de seqüências maior do que 1, mas ele também tem a vantagem de avaliar o processo de evolução das seqüências considerando a filogenia que as relaciona. Este teste ainda avalia a adequação do modelo como um todo, aceitando-o apenas se ele tiver desempenho significativamente melhor.

Para utilizar o teste da razão de verossimilhança, devemos primeiro obter a função de verossimilhança para o modelo com correlação entre pares de sítios. Para tanto, observamos que temos uma seqüência de comprimento $S/2$ de pares de sítios evoluindo independentemente de acordo com a matriz $\mathbf{Q}_{\text{cp}} = (\vartheta_{\nu,\upsilon})_{\nu,\upsilon \in M}$. Assim, podemos utilizar exatamente o mesmo algoritmo apresentado na Seção 3.1.1, observando que o processo possui 16 estados no lugar de 4. Para tanto, necessitamos apenas da matriz de transição $\mathbf{P}_{\text{cp}}(\tau)$, que pode ser obtida computacionalmente através de $\mathbf{P}_{\text{cp}}(\tau) = \exp(\tau \mathbf{Q}_{\text{cp}})$. Apartir da função de verossimilhança, podemos optimizá-la computacionalmente para obter a estatística do teste $-2\Delta(\mathbf{X})$, dada em (3.13).

Note que, embora o objetivo tanto deste teste quanto daquele apresentado para apenas uma seqüência seja verificar se há dependência entre os sítios na forma do modelo apresentado nessa seção, as hipóteses desse teste são diferentes daquelas apresentadas em (4.31). São elas:

$$H_0\text{: Pressupostos A1, A2, A3 e } \boldsymbol{\rho} = (\pi_{ij})_{i,j \in E} = (\pi_i^{(1)} \pi_j^{(2)})_{i,j \in E};$$
$$H_1\text{: Pressupostos A1, A2, A3 e } \boldsymbol{\rho} = (\pi_{ij})_{i,j \in E} \neq (\pi_i^{(1)} \pi_j^{(2)})_{i,j \in E},$$

em que

A1. As seqüências são relacionadas por uma estrutura filogenética;

A2. As mutações de cada par de sítio são independentes e identicamente distribuídas;

A3. As mutações são regidas por uma cadeia de Markov, com matriz de taxas infinitesimais $\mathbf{Q}_{\text{cp}}$.

Observe que, como valem esses pressupostos, recaímos em um caso similar ao dos modelos apresentados na Seção 2.3, e assim, o Teorema 3.1 é válido para esse teste. Portanto, a estatística do teste $-2\Delta(\mathbf{X})$ tem distribuição assintótica qui-quadrado com 9 graus de liberdade.

4.3 Aplicações

Nesta seção, são apresentadas aplicações dos modelos deste capítulo a conjuntos de seqüências reais. Inicialmente, utilizamos o teste da razão de verossimilhança para comparar o desempenho dos modelos das Seções 2.3 e 4.1 nas seqüências analisadas na aplicação do final do Capítulo 3.

Por questões computacionais, o modelo com distribuição gama contínua, apresentado na Seção 4.1.2, não está implementado no pacote PAML para conjuntos de dados com mais de oito seqüências. Assim, como as duas aplicações apresentadas na Seção 3.3 possuem mais de oito seqüências, utilizaremos apenas o modelo da distribuição gama discretizada, apresentado na Seção 4.1.3 , nestas duas aplicações.

Todos os modelos utilizados nesse capítulo tem como objetivo tratar diferentes formas de heterogeneidade nas seqüências. Essa heterogeneidade é justificada por vários fatores, como pressão seletiva diferenciada em trechos da seqüência e restrições de conformação tridimensional da molécula (seja no próprio DNA, no RNA ou na proteína codificada). É por meio de testes de hipóteses que comparam esses modelos que avaliamos a importância desses fatores na evolução do DNA.

Deste modo, como ressaltado na Seção 3.3, temos dois interesses principais em aplicar o teste da razão de verossimilhança a este modelos. O primeiro é conhecer melhor o processo de evolução das seqüências, destacando que fatores tem maior influência. Assim, diferentes taxas de mutação entre os sítios e correlação das taxas de mutação em sítios vizinhos são pressupostos que têm bom fundamentos biológicos como justificativa. Mas será que eles realmente são determinantes na evolução de um determinado gene?

Nosso segundo interesse na aplicação destes testes é a escolha de um modelo para a realização de análises posteriores envolvendo a função de verossimilhança (como inferência de filogenia, determinação de tempo de divergência entre espécies e datação de ancestral comum). Note que estas análises podem ter resultados diferentes para modelos diferentes. Por outro lado, a escolha de modelos muito complexos aumenta bastante o tempo computacional (principalmente quando se considera distribuição contínua para as taxas ou dependência entre os sítios). Assim, utilizam-se modelos mais complexos apenas quando eles apresentam desempenho significativamente superior. Além disso, deve-se verificar se o tempo computacional envolvido não torna a análise inviável.

4.3.1 Felinos

Consideramos, agora, o mesmo conjunto de dados utilizado na Seção 3.3.2, composto de 37 seqüências com $S = 382$ sítios de comprimento, proveniente de um gene dos felinos. Sempre que possível, dentro do que há disponível, o pacote PAML foi utilizado para a análise das seqüências. Entretanto, o pacote não conta com rotinas para simular seqüências segundo os modelos com taxas de mutação determinadas por distribuição discreta e por cadeias de Markov ocultas. Como tal simulação é necessária para o bootstrap paramétrico, foram desenvolvidas rotinas através do "R-project".

Os modelos apresentados neste capítulo devem ser sempre utilizados junto com um dos modelos de substituição de bases apresentados na Seção 2.3. Os modelos da Seção 2.3 considerados foram JC69, K80, F81, HKY85 e GTR. Já modelos deste capítulo considerados na análise foram os que atribuem às taxas de mutação nos sítios uma distribuição discreta como na Seção 4.1.1 (denotado por +disc), a distribuição gama discretizada (denotado por $+\Gamma_d$), e uma cadeia de Markov oculta (denotado por +HMM). Além disso, os modelos da Seção 2.3 foram considerados na sua versão original, em que todos os sítios da seqüências evoluem de maneira independente e segundo as mesmas taxas (denotado por i.i.d.). Como todas as combinações entre os modelos foram consideradas, ao todo utilizamos 20 modelos.

Utilizamos nesta seção a notação i.i.d. para designar aqueles modelos nos quais os sítios evoluem de maneira independente e todos de acordo com as mesmas taxas, como utilizado no Capítulo 3. Notamos, entretanto que os modelos $+\Gamma_d$ e +disc também representam processos estocásticos i.i.d., conforme a Definição 2.4.

Para os modelos +HMM, $+\Gamma_d$ e +disc utilizamos o número de diferentes taxas $C = 3$. A filogenia utilizada para as análises foi a mesma da Seção 3.3.2, e está apresentada na Figura 3.2.

Dentro de cada modelo deste capítulo (+HMM, $+\Gamma_d$ e +disc) os modelos de substituição de bases do Capítulo 2 foram comparados entre si segundo os mesmos testes utilizados na Seção 3.3. O objetivo desses testes é verificar o efeito que os diferentes pressupostos em comum tem na decisão do teste. Os resultados estão apresentados na Tabela 4.1, que também apresenta o valor da estatística do teste $-2\Delta(\mathbf{X})$ para cada comparação, e o p-valor encontrado de acordo com a distribuição assintótica qui-quadrado.

Notamos que, entre os testes de hipóteses apresentados na Tabela 4.1, apenas o teste JC69 $+\Gamma_d$ × F81 $+\Gamma_d$ teve resultado diferente de seu equivalente com taxas constantes (JC69 (i.i.d.) × F81 (i.i.d.)). Assim, percebemos que, neste caso, a utilização de modelos para diferentes taxas de mutação nos sítios pouco afeta a escolha do modelo de substituição de bases que descreve a evolução de um sítio. Além disso, destacamos que, em todos os modelos para diferentes taxas de mutação nos sítios ($+\Gamma_d$, +disc, +HMM, i.i.d.) o modelo de substituição de bases da Seção 2.3 que melhor descreve a evolução temporal desse conjunto de dados é o GTR.

Tabela 4.1: Resultado do Teste para as Seqüências dos Felinos.

Pressupostos Comuns	Teste	$-2\Delta(\mathbf{X})$	Distribuição Teórica r	p-valor	Decisão
i.i.d.	JC69 × K80	229.8426	1	6.4503×10^{-52}	rejeita
	F81 × HKY85	248.2718	1	6.1825×10^{-56}	rejeita
	JC69 × F81	1.4760	3	6.8851×10^{-1}	aceita
	K80 × HKY85	19.9052	3	1.7780×10^{-4}	rejeita
	HKY85× GTR	44.0106	4	6.3833×10^{-9}	rejeita
+ Γ_d	JC69 × K80	234.7156	1	5.5846×10^{-53}	rejeita
	F81 × HKY85	248.5566	1	5.3594×10^{-56}	rejeita
	JC69 × F81	14.0022	3	2.9021×10^{-3}	rejeita
	K80 × HKY85	27.8432	3	3.9179×10^{-6}	rejeita
	HKY85× GTR	34.1664	4	6.8886×10^{-7}	rejeita
+ disc	JC69 × K80	228.6578	1	1.1696×10^{-51}	rejeita
	F81 × HKY85	261.0316	1	1.0224×10^{-58}	rejeita
	JC69 × F81	1.0436	3	7.9070×10^{-1}	aceita
	K80 × HKY85	19.9536	3	1.7354×10^{-4}	rejeita
	HKY85× GTR	28.2796	4	1.0946×10^{-5}	rejeita
+ HMM	JC69 × K80	240.2154	1	3.5297×10^{-54}	rejeita
	F81 × HKY85	262.8502	1	4.1044×10^{-59}	rejeita
	JC69 × F81	1.0442	3	7.9055×10^{-1}	aceita
	K80 × HKY85	23.679	3	2.9146×10^{-5}	rejeita
	HKY85× GTR	28.873	4	8.2957×10^{-6}	rejeita

Queremos, agora, avaliar o desempenho dos modelo $+\Gamma_d$, +disc e +HMM em relação ao i.i.d., para verificar se sua utilização se justifica. Para tanto, utilizamos o teste da razão de verossimilhança para comparar os três modelos com o i.i.d.. Como o modelo GTR foi escolhido como aquele que melhor descreve as seqüências, nós o utilizamos em todas as comparações. A Tabela 4.2 apresenta os valores dos parâmetros estimados para cada modelo. Notamos que as estimativas dos parâmetros do modelo GTR variam, dependendo do modelo que o acompanha.

O resultado destes testes está apresentado na Tabela 4.3. Os valores obtidos para a estatística do teste foram comparados com a distribuição dada pelo bootstrap paramétrico, como apresentado na Seção 3.2.3. A Figura 4.4 apresenta os histogramas obtidos para $-2\Delta(\mathbf{X})$, sob H_0 e sob H_1, para os quatro testes considerados.

Notamos que, nos três primeiros testes, a hipótese nula é rejeitada, de forma que temos fortes indícios a favor de modelos que considerem diferentes taxas de mutação. Assim, concluímos que a evolução mais rápida de alguns sítios da seqüência é um fator importante na evolução desse gene.

A Tabela 4.3 também apresenta o teste GTR+disc × GTR+HMM, que tem como objetivo avaliar o efeito que a correlação de taxas entre sítios adjacentes tem sobre essa amostra. A tabela indica que, segundo a distribuição obtida pelo bootstrap paramétrico, a hipótese nula deve ser rejeitada ao nível de 99% de confiança. Além disso, temos que o

Tabela 4.2: Estimadores para os Modelos das Seqüências dos Felinos.

Modelo	Estimadores
GTR	$\hat{\mathbf{p}}_0 = (0.3420,\ 0.2037,\ 0.2243,\ 0.2300)$, $\hat{\alpha} = 1$, $\hat{\gamma} = 0.2722$, $\hat{\beta} = 0.2072$, $\hat{\delta} = 0.1026$, $\hat{\eta} = 2.4983$ e $\hat{\epsilon} = 0.0001$
GTR $+\Gamma_d$	$\hat{\mathbf{p}}_0 = (0.3318, 0.2237, 0.2468, 0.1977)$, $\hat{\alpha} = 1$, $\hat{\gamma} = 0.2785$, $\hat{\beta} = 0.1912$, $\hat{\delta} = 0.1057$, $\hat{\eta} = 2.8458$ e $\hat{\epsilon} = 0.0001$ $\hat{a} = 1.7778$
GTR + disc	$\hat{\mathbf{p}}_0 = (0.3420, 0.2037, 0.2300, 0.2243)$, $\hat{\alpha} = 1$, $\hat{\gamma} = 0.2089$, $\hat{\beta} = 0.1777$, $\hat{\delta} = 0.1149$, $\hat{\eta} = 2.5467$ e $\hat{\epsilon} = 0.0001$, $\hat{\mu}_1 = 0.0001$, $\hat{q}_1 = 0.6104$, $\hat{\mu}_2 = 1.3223$, $\hat{q}_2 = 0.3141$, $\hat{\mu}_3 = 7.7507$, $\hat{q}_3 = 0.0754$
GTR + HMM	$\hat{\mathbf{p}}_0 = (0.3420, 0.2037, 0.2300, 0.2243)$, $\hat{\alpha} = 1$, $\hat{\gamma} = 0.1995$, $\hat{\beta} = 0.1768$, $\hat{\delta} = 0.1278$, $\hat{\eta} = 2.6001$ e $\hat{\epsilon} = 0.0001$, $\hat{\mu}_1 = 0.0967$, $\hat{q}_1 = 0.7381$, $\hat{\mu}_2 = 2.0048$, $\hat{q}_2 = 0.1926$, $\hat{\mu}_3 = 7.8273$, $\hat{q}_3 = 0.0693$ $\mathbf{P}_{\text{HMM}} = \begin{pmatrix} 0.8510 & 0.1063 & 0.0427 \\ 0.5238 & 0.3276 & 0.1486 \\ 0.1308 & 0.7368 & 0.1324 \end{pmatrix}$

Tabela 4.3: Resultado do Teste para as Seqüências dos Felinos.

Teste	$-2\Delta(\mathbf{X})$	Simulação de Monte Carlo			Decisão
		p-valor	Valor Crítico 99%	Poder	
GTR (i.i.d.) $\times$ GTR $+\Gamma_d$	292.0344	< 0.001	6.2749	1.0000	rejeita
GTR (i.i.d.) $\times$ GTR + disc	328.7630	< 0.001	5.5257	1.0000	rejeita
GTR (i.i.d.) $\times$ GTR + HMM	359.5958	< 0.001	11.4482	1.0000	rejeita
GTR + disc $\times$ GTR + HMM	30.8328	< 0.001	13.6605	0.3580	rejeita

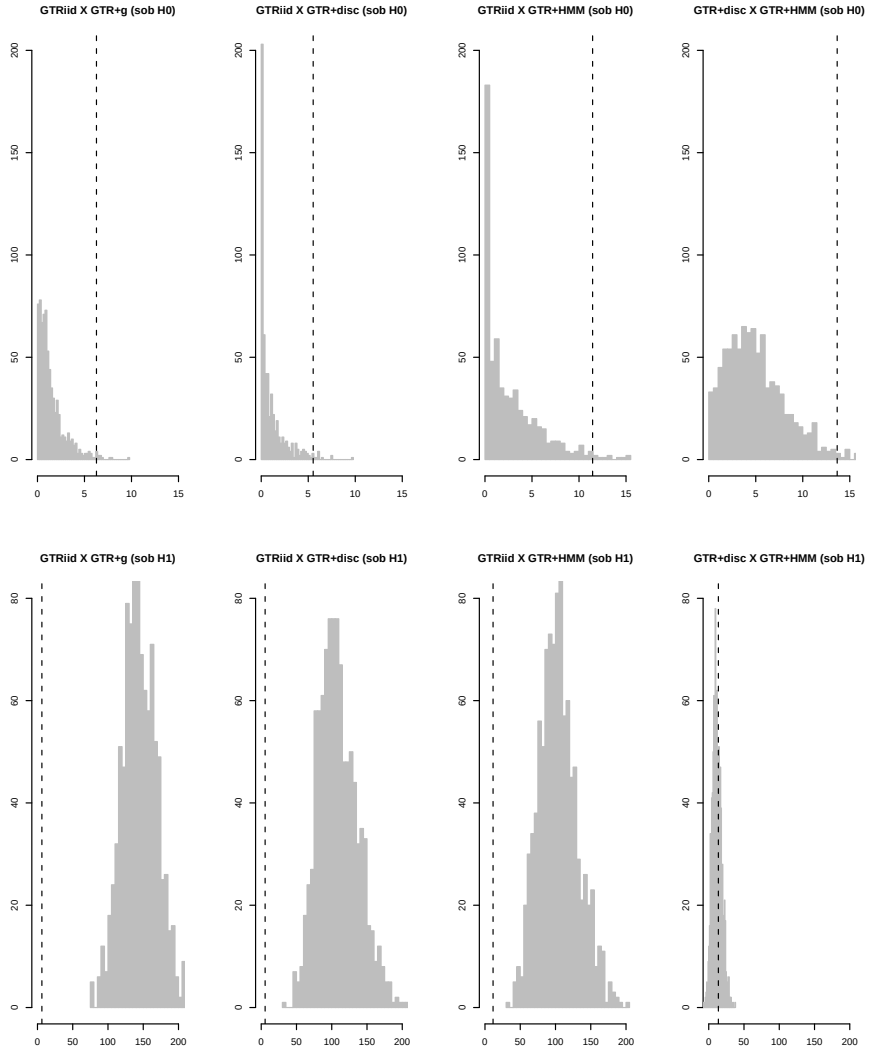


Figura 4.4: Histogramas de $-2\Delta(\mathbf{X})$ obtidos por Simulações de Monte Carlo para os Testes GTR (i.i.d.) $\times$ GTR+Γ_d, GTR(i.i.d.) $\times$ GTR+disc, GTR(i.i.d.) $\times$ GTR+HMM e GTR+disc $\times$ GTR+HMM sob H_0 e H_1 para as Seqüências de Felinos.

modelo com distribuição discreta para as taxas de mutação é um caso particular do modelo HMM, em que há restrições à matriz de transição da cadeia oculta. Assim, temos que a estatística deste teste $-2\Delta(\mathbf{X})$ tem distribuição assintótica χ^2_6. Segundo esta distribuição, a hipótese nula deve ser rejeitada ao nível 99% de confiança (p-valor= 2.7315×10^{-5}). Assim, ambas as distribuições concordam que, nestas seqüências, sítios adjacentes tendem a evoluir a taxas semelhantes.

De acordo com o resultado destes testes, somos levados a escolher o modelo GTR+HMM para descrever este conjunto de dados. Notamos, entretanto, que, dependendo da análise para a qual se deseja utilizá-lo, o tempo computacional envolvido pode ser muito grande devido ao grande número de seqüências. Assim, por questões práticas, pode se fazer a escolha de um modelo mais simples.

Observamos que o poder do teste, obtido pelo bootstrap paramétrico, foi 1 para os três primeiros testes. Entretanto, para o quarto teste, que avalia a presença de correlação entre as taxas de mutação de sítios adjacente, o poder foi bastante baixo (Poder= 0.3580).

4.3.2 Primatas

O segundo conjunto de dados que analisamos consiste de 17 seqüências, com $S = 483$ sítios de comprimento, da família gênica ECP-EDN em primatas. Essas são as mesmas seqüências utilizadas na Seção 3.3.3. Para dar continuidade àquilo que foi trabalhado naquela seção, utilizamos a mesma árvore filogenética, apresentada na Figura 3.5. Para a aplicação dos testes, foram utilizadas as mesmas rotinas computacionais da Seção 4.3.1.

Os modelos e testes considerados foram os mesmos da aplicação anterior dos felinos, apresentada na Seção 4.3.1. Assim, consideramos todas as combinações dos modelos JC69, K80, F81, HKY85 e GTR com +HMM, $+\Gamma_d$, +disc e i.i.d.. Inicialmente, comparamos os modelos de substituição de bases da Seção 2.3 entre si, mantendo constantes os pressupostos em relação às taxas de mutação nos diferentes sítios. Os resultados destes testes de hipóteses estão apresentados na Tabela 4.4.

Como pode ser visto, independente dos pressupostos comuns em relação à taxa de mutação nos sítios, o resultado de todos os testes coincidem com aqueles encontrados na Seção 3.3.3, apresentados na Tabela 4.4 como i.i.d.. Assim, confirmamos as conclusões então obtidas de que, para esse conjunto de dados, parâmetros adicionais que representem diferentes proporções das bases na seqüência não trazem melhoras significativas no desempenho do modelo. Lembramos que o que diferencia os modelos JC69 e K80 dos modelos F81 e HKY85 é exatamente a existência de tais parâmetros. Além disso, destacamos que o melhor modelo de substituição de bases, entre os da Seção 2.3, para esses dados é o GTR.

Utilizando sempre o modelo GTR, prosseguimos para avaliar a utilização de diferentes distribuições para as taxas dos sítios. Assim, comparamos os modelos GTR+Γ_d,

Tabela 4.4: Resultado do Teste para as Seqüências dos Primatas.

Pressupostos Comuns	Teste	$-2\Delta(\mathbf{X})$	Distribuição Teórica		Decisão
			r	p-valor	
i.i.d.	JC69 × K80	32.0992	1	1.4650×10^{-8}	rejeita
	F81 × HKY85	33.1716	1	8.4374×10^{-9}	rejeita
	JC69 × F81	8.2774	3	4.0613×10^{-2}	aceita
	K80 × HKY85	9.3498	3	2.4984×10^{-2}	aceita
	HKY85× GTR	29.4102	4	6.4520×10^{-6}	rejeita
$+\Gamma_d$	JC69 × K80	33.5126	1	7.0803×10^{-9}	rejeita
	F81 × HKY85	35.3276	1	2.7865×10^{-9}	rejeita
	JC69 × F81	8.7361	3	3.3013×10^{-2}	aceita
	K80 × HKY85	10.5510	3	1.4418×10^{-2}	aceita
	HKY85× GTR	25.6118	4	3.7890×10^{-5}	rejeita
+disc	JC69 × K80	34.2200	1	4.9220×10^{-9}	rejeita
	F81 × HKY85	34.3584	1	4.5841×10^{-9}	rejeita
	JC69 × F81	6.3892	3	9.4136×10^{-2}	aceita
	K80 × HKY85	6.5276	3	8.8580×10^{-2}	aceita
	HKY85× GTR	29.9072	4	5.1120×10^{-6}	rejeita
+HMM	JC69 × K80	33.7966	1	6.1185×10^{-9}	rejeita
	F81 × HKY85	36.6028	1	1.4482×10^{-9}	rejeita
	JC69 × F81	5.4344	3	1.4261×10^{-1}	aceita
	K80 × HKY85	8.2406	3	4.1292×10^{-2}	aceita
	HKY85× GTR	24.0342	4	7.8623×10^{-5}	rejeita

GTR+disc e GTR+HMM com o modelo GTR (i.i.d.). Os estimadores de máxima verossimilhança para os parâmetros destes modelos estão apresentados na Tabela 4.5, e o resultado dos testes de hipóteses estão na Tabela 4.6. Note que, para a decisão dos testes utiliza-se a distribuição obtida no bootstrap paramétrico, realizado de acordo com o algoritmo apresentado na Seção 3.2.1. Além disso, a Tabela 4.6 apresenta o poder do teste, obtido de acordo com o algoritmo da Seção 3.2.3, com as devidas alterações para acomodar o fato de que estes modelos têm taxas de mutação diferentes nos sítios. A Figura 4.5 apresenta os histogramas de $-2\Delta(\mathbf{X})$, sob H_0 e sob H_1, obtidos pelo bootstrap paramétrico.

O resultado dos três primeiros testes apresentados na Tabela 4.6 indica que, para este conjunto de dados, modelos que pressupõem diferentes taxas de mutação para os diferentes sítios são preferidos. Além disso, o resultado do teste GTR+disc × GTR+HMM indica que a correlação entre as taxas de mutação de sítios adjacentes foi importante na evolução destas seqüências.

Como resultado dos testes de hipóteses aplicados, concluímos que, na família gênica ECP-EDN de primatas, o modelo que melhor descreve a evolução das seqüências é o GTR+HMM.

Destacamos que o poder do teste, obtido pelo bootstrap paramétrico, foi baixo para os testes GTR (i.i.d.) × GTR+disc, GTR (i.i.d.) × GTR+HMM e GTR+disc × GTR+HMM (ver Tabela 4.6). Como a hipótese nula H_0 foi rejeitada em todos os casos,

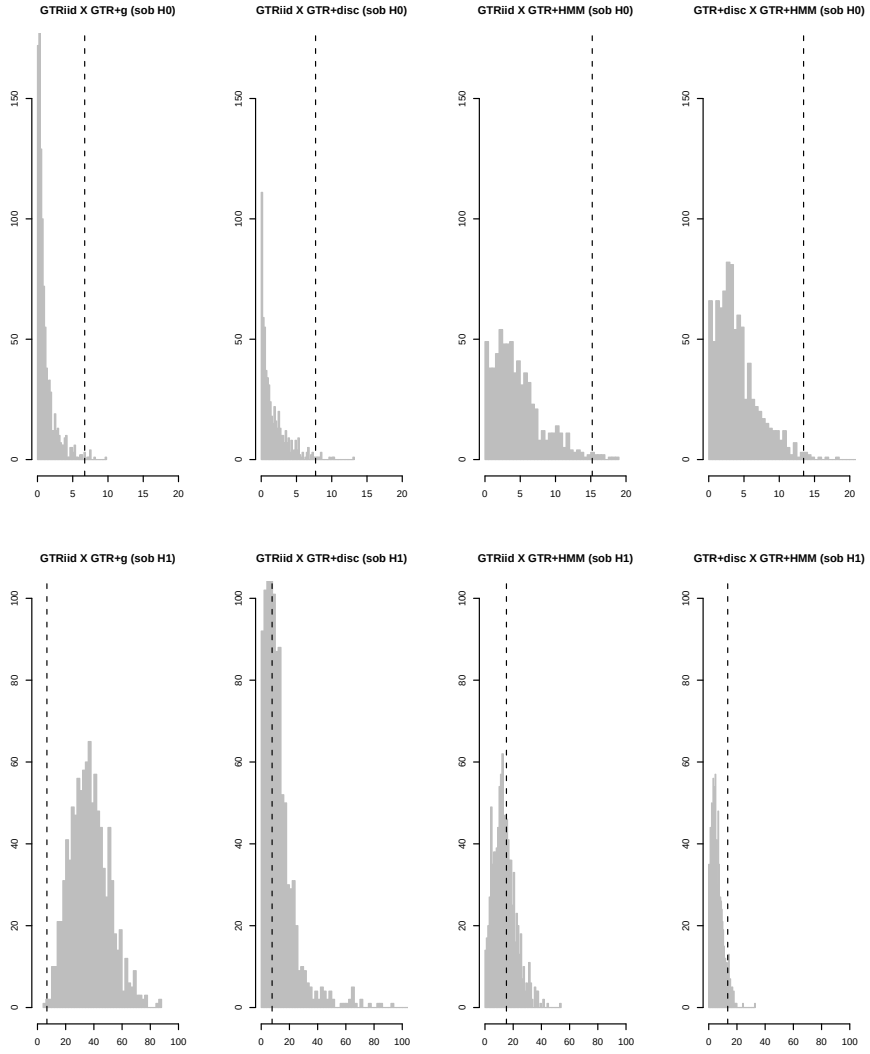


Figura 4.5: Histogramas de $-2\Delta(\mathbf{X})$ obtidos por Simulações de Monte Carlo para os Testes GTR (i.i.d.) $\times$ GTR+Γ_d, GTR (i.i.d.) $\times$ GTR+disc, GTR (i.i.d.) $\times$ GTR+HMM e GTR+disc $\times$ GTR+HMM sob H_0 e H_1 para as Seqüências de Primatas.

Tabela 4.5: Estimadores para os Modelos das Seqüências dos Primatas.

Modelo	Estimadores
GTR	$\hat{\mathbf{p}}_0 = (0.2647, 0.1942, 0.2571, 0.2840)$, $\hat{\alpha} = 1$, $\hat{\gamma} = 0.4601$, $\hat{\beta} = 0.2126$, $\hat{\delta} = 0.2443$, $\hat{\eta} = 0.5061$ e $\hat{\epsilon} = 0.3359$
GTR+Γ_d	$\hat{\mathbf{p}}_0 = (0.2655, 0.1905, 0.2813, 0.2627)$, $\hat{\alpha} = 1$, $\hat{\gamma} = 0.4580$, $\hat{\beta} = 0.1962$, $\hat{\delta} = 0.2341$, $\hat{\eta} = 0.4206$ e $\hat{\epsilon} = 0.3003$ $\hat{a} = 0.7260$
GTR+disc	$\hat{\mathbf{p}}_0 = (0.2755, 0.1957, 0.2729, 0.2559)$, $\hat{\alpha} = 1$, $\hat{\gamma} = 0.4742$, $\hat{\beta} = 0.1967$, $\hat{\delta} = 0.2355$, $\hat{\eta} = 0.4531$ e $\hat{\epsilon} = 0.3091$, $\hat{\mu}_1 = 0.3878$, $\hat{q}_1 = 0.7145$, $\hat{\mu}_2 = 2.3520$, $\hat{q}_2 = 0.2830$, $\hat{\mu}_3 = 23.5863$, $\hat{q}_3 = 0.0024$
GTR+HMM	$\hat{\mathbf{p}}_0 = (0.2729, 0.2559, 0.2755, 0.1957)$, $\hat{\alpha} = 1$, $\hat{\gamma} = 0.4817$, $\hat{\beta} = 0.2002$, $\hat{\delta} = 0.2590$, $\hat{\eta} = 0.4919$ e $\hat{\epsilon} = 0.3234$, $\hat{\mu}_1 = 0.3411$, $\hat{q}_1 = 0.5942$, $\hat{\mu}_2 = 1.4707$, $\hat{q}_2 = 0.3469$, $\hat{\mu}_3 = 4.8748$, $\hat{q}_3 = 0.0589$ $\mathbf{P}_{\text{HMM}} = \begin{pmatrix} 0.9016 & 0.0001 & 0.0983 \\ 0.1685 & 0.8313 & 0.0002 \\ 0.0001 & 0.9929 & 0.0070 \end{pmatrix}$

Tabela 4.6: Resultado do Teste para as Seqüências dos Primatas.

Teste	$-2\Delta(\mathbf{X})$	Simulação de Monte Carlo			Decisão
		p-valor	Valor Crítico 99%	Poder	
GTR (i.i.d.) × GTR+Γ_d	37.2386	< 0.001	6.6985	0.9980	rejeita
GTR (i.i.d.) × GTR+disc	41.0562	< 0.001	7.7079	0.5850	rejeita
GTR (i.i.d.) × GTR+HMM	57.3578	< 0.001	15.2119	0.3670	rejeita
GTR+disc × GTR+HMM	16.3016	0.006	13.2119	0.0590	rejeita

não precisamos nos preocupar com isso. Entretanto, caso H_0 fosse aceita para algum destes testes, teríamos que analisar este resultado com cuidado, uma vez que o poder do teste acusou valor baixo.

Destacamos novamente que esses modelos, especialmente o +HMM, apresentam alto custo computacional. Assim, embora ele tenha sido escolhido pelo teste, ele pode não ser adequado para certas análises por as tornar inviáveis.

4.3.3 Dependência entre Sítios

Nesta sub-seção, realizamos uma aplicação do teste para dependência entre sítios apresentado na Seção 4.2.2. Note, entretanto, que os modelos que consideram dependência entre os processos evolutivos que ocorrem nos sítios não são tão amplamente difundidos quanto os demais modelos apresentados nesse trabalho. Assim, os aplicativos que realizam análises de máxima verossimilhança em conjuntos de seqüências de DNA, como o PAML, utilizado neste trabalho, em geral não possuem estes modelos implementados. Deste modo, aplicamos o teste para dependência entre sítios apenas no caso de uma seqüência, pois assim, não precisamos considerar a filogenia.

Ressaltamos que, por considerar apenas uma seqüência, este teste não considera a história evolutiva do gene. Assim, embora possamos concluir a respeito da dependência nesta seqüência, seria apressado concluir se devemos utilizar o modelo de correlação entre

Tabela 4.7: Estimadores para as Freqüências das Bases e dos Pares de Bases no Modelo de Correlação entre Sítios.

Parâmetro	Estimadores
$\pi_i^{(1)}$	$\hat{\pi}_A^{(1)} = 0.3018$, $\hat{\pi}_C^{(1)} = 0.23509$, $\hat{\pi}_G^{(1)} = 0.1754$, $\hat{\pi}_T^{(1)} = 0.2877$
$\pi_i^{(2)}$	$\hat{\pi}_A^{(2)} = 0.0772$, $\hat{\pi}_C^{(2)} = 0.3965$, $\hat{\pi}_G^{(2)} = 0.3263$, $\hat{\pi}_T^{(2)} = 0.2000$
$\pi_{i,j}$	$\hat{\pi}_{AA} = 0.0386$, $\hat{\pi}_{AC} = 0.1333$, $\hat{\pi}_{AG} = 0.0912$, $\hat{\pi}_{AT} = 0.0386$ $\hat{\pi}_{CA} = 0.0140$, $\hat{\pi}_{CC} = 0.1193$, $\hat{\pi}_{CG} = 0.0246$, $\hat{\pi}_{CT} = 0.0772$ $\hat{\pi}_{GA} = 0.0140$, $\hat{\pi}_{GC} = 0.0807$, $\hat{\pi}_{GG} = 0.0386$, $\hat{\pi}_{GT} = 0.0421$ $\hat{\pi}_{TA} = 0.0105$, $\hat{\pi}_{TC} = 0.0632$, $\hat{\pi}_{TG} = 0.1719$, $\hat{\pi}_{TT} = 0.0421$
$\pi_i^{(1)} \times \pi_j^{(2)}$	$\hat{\pi}_A^{(1)}\hat{\pi}_A^{(2)} = 0.0233$, $\hat{\pi}_A^{(1)}\hat{\pi}_C^{(2)} = 0.1196$, $\hat{\pi}_A^{(1)}\hat{\pi}_G^{(2)} = 0.0985$, $\hat{\pi}_A^{(1)}\hat{\pi}_T^{(2)} = 0.0604$ $\hat{\pi}_C^{(1)}\hat{\pi}_A^{(2)} = 0.0181$, $\hat{\pi}_C^{(1)}\hat{\pi}_C^{(2)} = 0.0932$, $\hat{\pi}_C^{(1)}\hat{\pi}_G^{(2)} = 0.0767$, $\hat{\pi}_C^{(1)}\hat{\pi}_T^{(2)} = 0.0470$ $\hat{\pi}_G^{(1)}\hat{\pi}_A^{(2)} = 0.0135$, $\hat{\pi}_G^{(1)}\hat{\pi}_C^{(2)} = 0.0696$, $\hat{\pi}_G^{(1)}\hat{\pi}_G^{(2)} = 0.0572$, $\hat{\pi}_G^{(1)}\hat{\pi}_T^{(2)} = 0.0351$ $\hat{\pi}_T^{(1)}\hat{\pi}_A^{(2)} = 0.0222$, $\hat{\pi}_T^{(1)}\hat{\pi}_C^{(2)} = 0.1141$, $\hat{\pi}_T^{(1)}\hat{\pi}_G^{(2)} = 0.0939$, $\hat{\pi}_T^{(1)}\hat{\pi}_T^{(2)} = 0.0575$

sítios para a evolução deste gene. Para tal conclusão, deveríamos utilizar o teste para dependência entre sítios (caso várias seqüências), que considera uma filogenia e a história evolutiva do gene.

Para esta análise, utilizamos a seqüência do gene de humanos da proteína abglobina, disponível no conjunto de exemplos distribuído junto com o aplicativo PAML. Uma característica importante desta seqüência é que ela é composta apenas da região codante do gene e está na fase de leitura correta (ou seja, sabemos exatamente quais são os códons da seqüência). O conhecimento dos códons da seqüência é fundamental para a aplicação do modelo de correlação entre pares de sítios, pois este assume que há correlação entre as mutações que ocorrem nos dois primeiros sítios de cada códon.

Inicialmente, preparamos a seqüência, composta de 855 nucleotídeos, removendo a terceira posição de cada códon da seqüência, pois estes não estão contemplados no modelo. Assim, obtivemos uma seqüência com $S = 570$ sítios que contém apenas os dois primeiros sítios de cada códon. Em seguida, calculamos as freqüências das bases na primeira e segunda posição de cada códon, $\hat{\pi}_i^{(1)}$ e $\hat{\pi}_i^{(2)}$, além das freqüências dos pares de bases nos códons $\hat{\pi}_{i,j}$, com $i, j \in E$. A Tabela 4.7 apresenta os valores obtidos para as freqüências das bases e dos pares de bases. Além disso, essa tabela fornece os valores de $\hat{\pi}_i^{(1)} \times \hat{\pi}_j^{(2)}$, com $i, j \in E$, para fins de comparação, uma vez que se $\hat{\pi}_i^{(1)} \times \hat{\pi}_j^{(2)} = \hat{\pi}_{i,j}$ temos independência entre os sítios.

Observe, por exemplo, que a freqüência esperada do par TA, sob a hipótese da independência seria 0.0222. Entretanto, a freqüência observada deste par é menor, $\hat{\pi}_{TA} = 0.0105$. Por outro lado, o par TG apresenta uma freqüência de 0.1719, que é consideravelmente maior do que a freqüência esperada, sob a hipótese de independência, $\hat{\pi}_T^{(1)} \times \hat{\pi}_G^{(2)} = 0.0939$. Assim, devemos verificar se esses desvios das proporções esperadas,

sob a hipótese de independência, são significativos.

O teste que consideramos tem as seguintes hipóteses:

H_0: Todos os sítios são independentes, ou seja, $\boldsymbol{\rho} = (\pi_{ij})_{i,j} = (\pi_i^{(1)} \times \pi_j^{(2)})_{i,j}$;
H_1: Há dependência entre o primeiro e segundo sítio de cada códon, ou seja, $\boldsymbol{\rho} = (\pi_{ij})_{i,j} \neq (\pi_i^{(1)} \times \pi_j^{(2)})_{i,j}$.

A estatística do teste, dada pela expressão (4.32), é tal que $\mathcal{T} = 54.2126$. Lembramos que $\mathcal{T} \sim \chi^2_9$, assim, temos que o p-valor deste teste é de 1.7200×10^{-08}.

Deste modo, rejeitamos a hipótese nula, e concluímos que nesta seqüência há dependência entre o primeiro e segundo sítio de cada códon. Ou seja, existe uma correlação entre as bases das duas primeiras posições de cada códon, que são as posições fundamentais na determinação do aminoácido codificado.

Capítulo 5

Estimação

Os testes estatísticos apresentados nos Capítulos 3 e 4 têm por objetivo responder a perguntas sobre o processo de evolução molecular de seqüências de DNA. Tais perguntas são traduzidas em termos dos diferentes modelos para esse processo. Cada parâmetro dos modelos de substituição de bases representa um aspecto do processo biológico. Desta forma, através da estimativa de cada parâmetro, pode-se avaliar a importância relativa dos aspectos do processo representado por cada parâmetro.

Além disso, para que o teste da razão de verossimilhança possa ser aplicado, são necessários os estimadores de máxima verossimilhança para os parâmetros dos modelos envolvidos.

Neste capítulo, serão estudados estimadores para os parâmetros de diferentes modelos de evolução molecular. Primeiro consideramos o caso em que temos apenas duas seqüências; nessas condições, a estimativa dos parâmetros se confunde com a da distância evolutiva entre as seqüências. Nesse caso particular, podemos obter informações sobre a distribuição assintótica dos estimadores. Em seguida, consideramos um conjunto de hipóteses restritivas que nos permitem obter informações sobre a variância dos estimadores de máxima verossimilhança dos parâmetros dos modelos de substituição de bases.

Além disso, apresentamos um método baseado em parcimônia para a estimação do parâmetro a da distribuição gama, conforme modelo apresentado na Seção 4.1.2, e sugerimos a utilização do Jackknife para a correção do vício dos estimadores de máxima verossimilhança.

5.1 Distâncias entre Duas Seqüências

Iniciamos considerando o caso em que temos apenas duas seqüências. Assim, como elas têm um ancestral comum, a relação entre elas está representada na Figura 5.1. Portanto, a probabilidade de encontrarmos a base i no sítio u da seqüência 1 e a base j no mesmo sítio da seqüência 2 é dada por

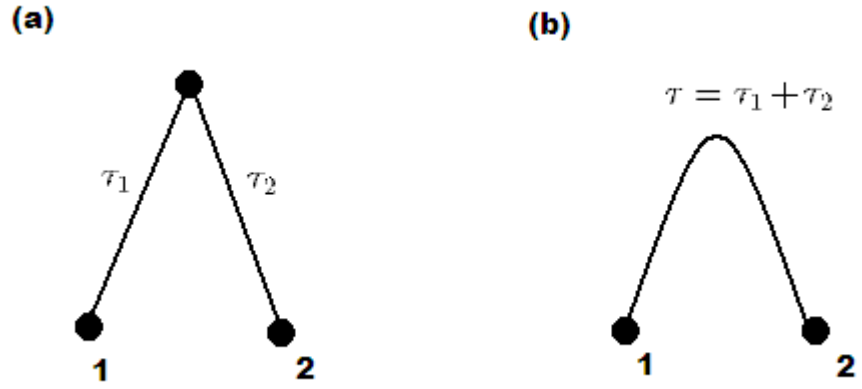


Figura 5.1: Representação da relação entre duas seqüências: (a) sem e (b) com reversibilidade no tempo.

$$\mathbb{P}(X_u^1 = i, X_u^2 = j) = \sum_{l \in E} \pi_l \mathbb{P}(i|l, \tau_1)\mathbb{P}(j|l, \tau_2), \tag{5.1}$$

em que l representa a base encontrada na posição u da seqüência do ancestral comum entre 1 e 2. Nessa expressão são consideradas, para o cálculo da probabilidade em questão, todas as possíveis bases na seqüência ancestral.

Observe que a variável aleatória X_u^v representa a base que ocupa o u-ésimo sítio da v-ésima seqüência da amostra. De forma análoga, temos que $\mathbf{X}^v$ representa toda a v-ésima seqüência da amostra e $\mathbf{X}_u$ é o vetor com as bases do u-ésimo sítio de todas as seqüências da amostra.

Note que todos os modelos que consideramos têm a propriedade de reversibilidade no tempo, ou seja

$$\pi_i \mathbb{P}(j|i, \tau) = \pi_j \mathbb{P}(i|j, \tau).$$

Dessa forma, podemos reescrever a expressão (5.1) como

$$\mathbb{P}(X_u^1 = i, X_u^2 = j) = \sum_{l \in E} \pi_j \mathbb{P}(l|j, \tau_2)\mathbb{P}(i|l, \tau_1) = \mathbb{P}(i|j, \tau_1 + \tau_2),$$

em que a segunda igualdade é conseqüência das equações de Chapman-Kolmogorov (ver Karlin e Taylor, 1975). Com base nesse resultado, vemos que, para modelos em que a propriedade de reversibilidade no tempo é satisfeita, não podemos distinguir os tempos τ_1 e τ_2, pois obtemos apenas informações sobre a soma desses tempos. Assim, para fins de

estimativas, podemos considerar apenas $\tau = \tau_1 + \tau_2$, a soma dos comprimentos dos ramos que separam as duas seqüências (ver Figura 5.1).

Nessa seção estudamos os estimadores de máxima verossimilhança para os parâmetros dos modelos de substituição de bases no caso de apenas duas seqüências. Para tal análise, são fundamentais as propriedades assintóticas dos estimadores de máxima verossimilhança apresentadas no teorema a seguir.

Teorema 5.1. *Seja $\hat{\boldsymbol{\theta}}_n$ o estimador de máxima verossimilhança para $\boldsymbol{\theta}$, vetor de parâmetros da distribuição $f_X(\cdot)$. Sob as hipóteses do Teorema 3.1, temos que*

$$\sqrt{S}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{Z}, \quad \textit{quando } S \to \infty, \tag{5.2}$$

onde $\mathcal{Z} \sim \mathbf{N}(0, \boldsymbol{\Sigma})$ é um vetor aleatório com distribuição normal multivariada com média 0 e matriz de variâncias-covariâncias $\boldsymbol{\Sigma} = [I(\boldsymbol{\theta})]^{-1}$, e $\mathbf{I}(\boldsymbol{\theta})$ a matriz de informação de Fisher.

Demonstração: Ver Rohatgi (1976). □

Observação 5.1. Note que, no Lema 3.1, demonstramos que as condições de regularidade do Teorema 3.1 são válidas para os modelos de substituição de bases apresentados na Seção 2.3.

Além disso, precisamos obter a distribuição de funções de variáveis aleatórias, e para tanto, recorremos ao método delta.

Teorema 5.2. (Método Delta). *Seja $\mathbf{T}_n = (T_1, \cdots, T_n)'$ vetor de estimadores com distribuição assintótica normal com média $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)'$ e matriz de variâncias-covariâncias $\boldsymbol{\Sigma}$. Suponha que a função $\mathbf{g}(t_1, \cdots, t_n) = (g_1(t_1, \cdots, t_n), \cdots, g_n(t_1, \cdots, t_n))$ tem matriz de derivadas $\mathbf{J}$ em $\mathbf{t} = \boldsymbol{\theta}$. Então,*

$$\sqrt{n}(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{Z}, \quad \textit{quando } n \to \infty,$$

onde $\mathcal{Z} \sim \mathbf{N}(0, \mathbf{J}\ \boldsymbol{\Sigma}\ \mathbf{J}')$. Assim, $\mathbf{g}(\mathbf{T}_n)$ tem variância assintótica normal multivariada com vetor de médias $\mathbf{g}(\boldsymbol{\theta})$ e matriz de variâncias-covariâncias $\mathbf{J}\ \boldsymbol{\Sigma}\ \mathbf{J}'$.

Demonstração: Ver Apêndice B. □

Modelo Jukes Cantor

Iniciando pelo caso mais simples, consideramos o modelo JC69. Assim, por (2.12) temos que

$$\mathbb{P}(X_u^1 = i, X_u^2 = j) = \alpha_\tau = \frac{1}{4}(1 - e^{-4\alpha\tau}),$$

se $i \neq j$, e

$$\mathbb{P}(X_u^1 = i, X_u^2 = j) = 1 - 3\alpha_\tau = \frac{1}{4}(1 + 3e^{-4\alpha\tau}),$$

se $i = j$. Seja s o número de sítios em que as duas seqüências diferem e S o comprimento total da seqüência. Assim, a função de verossimilhança dessas seqüências segundo o modelo JC69 é dada por

$$\begin{aligned} L(\tau|\mathbf{X}) &= \prod_{u=1}^{S} \mathbb{P}(X_u^1, X_u^2) = \alpha_\tau^s (1 - 3\alpha_\tau)^{S-s} \\ &= \left(\frac{1}{4} - \frac{1}{4}e^{-4\alpha\tau}\right)^s \left(\frac{1}{4} + \frac{3}{4}e^{-4\alpha\tau}\right)^{S-s}. \end{aligned} \tag{5.3}$$

Na expressão acima, o parâmetro do modelo de substituição de bases α e o tempo de divergência entre as espécies τ aparecem sempre juntos. Dessa forma, sem informação específica sobre um desses parâmetros só podemos obter dos dados $\mathbf{X}$ a estimativa do produto $\alpha\tau$.

Note que, para o modelo JC69, 3α é a taxa de mutação geral, ou seja o número médio do mutações por tempo. Seja d a distância entre duas seqüências definida como o produto entre a taxa de mutação e o tempo de divergência ($d = 3\alpha\tau$, para o modelo JC69). Dessa forma, d é uma medida da quantidade de evolução ocorrida entre as duas seqüências. Como não podemos obter separadamente α e τ, reescrevemos a expressão (5.3) em função de d,

$$L(d|\mathbf{X}) = \left(\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}d}\right)^s \left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right)^{S-s}.$$

Com base nesta expressão, podemos encontrar o estimador de máxima verossimilhança para d.

Lema 5.1. *Considere o modelo Jukes Cantor, cuja matriz* $\mathbf{Q}_{\text{JC69}}$ *de taxas infinitesimais é dada por (2.9) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.10). O estimador de máxima verossimilhança para a distância d entre duas seqüências é dado por*

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{s}{S}\right). \tag{5.4}$$

Demonstração: Queremos estimar d pelo método da máxima verossimilhança, e para tanto devemos maximizar $L(d|\mathbf{X})$. Note, entretanto, que como o logaritmo é uma função crescente, o valor de d em que $L(d|\mathbf{X})$ atinge seu máximo é o mesmo em que $\log(L(d|\mathbf{X}))$ atinge seu máximo. Além disso, como para seqüências de comprimento moderado os valores da função de verossimilhança são muito baixos é mais natural considerar seu logaritmo. Assim,

$$\log(L(d|\mathbf{X})) = s\log\left(\frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}d}\right) + (S-s)\log\left(\frac{1}{4} + \frac{3}{4}e^{-\frac{4}{3}d}\right). \tag{5.5}$$

Para encontrar os pontos críticos da função acima resolvemos a equação

$$\frac{\partial \log(L(d|\mathbf{X}))}{\partial d} = 0.$$

Mas note que

$$\frac{\partial \log(L(d|\mathbf{X}))}{\partial d} = \frac{4}{3}\frac{s\ e^{-\frac{4}{3}d}}{(1-e^{-\frac{4}{3}d})} - 4\frac{(S-s)e^{-\frac{4}{3}d}}{(1+3e^{-\frac{4}{3}d})},$$

de forma que temos

$$0 = \frac{s}{3(1-e^{-\frac{4}{3}\hat{d}})} - \frac{(S-s)}{(1+3e^{-\frac{4}{3}\hat{d}})},$$

e ainda

$$0 = 4s - 3S + 3Se^{-\frac{4}{3}\hat{d}}.$$

Assim, o estimador de máxima verossimilhança para d é dado por

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{s}{S}\right),$$

que só está definido se $\frac{s}{S} < \frac{3}{4}$. □

Observação 5.2. Note que quando $\frac{s}{S} \to \frac{3}{4}$ na expressão (5.4), temos que $\hat{d} \to \infty$. Dessa forma, quando $\frac{s}{S} \geq \frac{3}{4}$ define-se $\hat{d} = \infty$. Assim, temos

$$\hat{d} = \begin{cases} -\frac{3}{4}\log\left(1 - \frac{4}{3}\frac{s}{S}\right), & \text{se } \frac{s}{S} < \frac{3}{4} \\ \infty, & \text{caso contrário.} \end{cases} \tag{5.6}$$

O estimador $\hat{d}$ definido em (5.6) é viciado.

Para encontrar propriedades do estimador $\hat{d}$, recorremos a uma variável auxiliar. Note que se p for definido como a probabilidade de um sítio ser heteromorfo (apresentar bases diferentes nas duas seqüências) e W uma variável aleatória definida como o número de sítios heteromorfos na amostra, então W tem distribuição binomial ($W \sim \mathcal{B}(S, p)$).

Note que se W é uma variável aleatória com distribuição binomial, $W \sim \mathcal{B}(S, p)$, então $\mathbb{E}(W) = Sp$ e $\mathrm{Var}(W) = Sp(1-p)$.

Podemos rearranjar a expressão (5.4) da seguinte forma

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{p}\right), \tag{5.7}$$

e assim obtemos $\hat{d}$ como função de $\hat{p}$. A vantagem dessa abordagem é que a distribuição de W é conhecida.

Desejamos utilizar a expressão (5.7) para encontrar a distribuição assintótica de $\hat{p}$. Entretanto, como a variância não é função linear de d recorremos ao método delta para obter $\mathrm{Var}(\hat{d})$.

Lema 5.2. *Considere o modelo Jukes Cantor, cuja matriz* $\mathbf{Q}_{\mathrm{JC69}}$ *de taxas infinitesimais é dada por (2.9) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.10). Considere* $p, \hat{p} \in \left[0, \frac{3}{4}\right)$. *A distribuição assintótica do estimador de máxima verossimilhança* $\hat{d}$ *para a distância d entre duas seqüências, dado em (5.4), é dada por*

$$\sqrt{S}(\hat{d} - d) \xrightarrow{d} \mathcal{Z}, \quad \textit{quando } S \to \infty,$$

onde $\mathcal{Z} \sim N(0, \mathrm{Var}(\hat{d}))$, *com*

$$\mathrm{Var}(\hat{d}) = \frac{p\,(1-p)}{\left(1 - \frac{4}{3}p\right)^2}, \tag{5.8}$$

onde $\mathrm{Var}(\hat{d})$ *é a variância assintótica de* $\hat{d}$. *Assim, o estimador* $\hat{d}$ *é consistente, pois sua variância, dada por* $Var(\hat{d})/S$, *tende a zero, quando o número de sítios S tende a infinito, e é assintoticamente não viciado.*

Demonstração: Lembramos que $\hat{p}$ é estimador de máxima verossimilhança para p, e pelo Teorema 5.1 temos que $\sqrt{S}(\hat{p} - p) \xrightarrow{d} \mathcal{Z}$, quando $S \to \infty$, onde $\mathcal{Z} \sim N(0, \mathrm{Var}(\hat{p}))$ e $\mathrm{Var}(\hat{p}) = (p(1-p))$ é a variância assintótica de $\hat{d}$. Ainda observamos que $\hat{d} = g(\hat{p})$, onde $g(p) = -\frac{3}{4}\log\left(1 - \frac{4}{3}p\right) = d$, e que

$$\left.\frac{\partial g(w)}{\partial w}\right|_{w=p} = \frac{1}{(1 - \frac{4}{3}p)}.$$

Portanto, utilizando o Método Delta, temos que

$$\begin{aligned}\mathrm{Var}(\hat{d}) &= \mathrm{Var}(\hat{p})\left(\left.\frac{\partial g(\hat{p})}{\partial \hat{p}}\right|_{\hat{p}=p}\right)^2 = p(1-p)\times\frac{1}{(1-\frac{4}{3}p)^2}\\ &= \frac{p\,(1-p)}{\left(1-\frac{4}{3}p\right)^2}.\end{aligned} \tag{5.9}$$

Ou seja,

$$\sqrt{S}(\hat{d}-d)\xrightarrow{d}\mathcal{Z},\ \text{ quando } S\to\infty,$$

onde $\mathcal{Z}\sim N(0,\mathrm{Var}(\hat{d}))$.

□

Lema 5.3. *Considere o modelo Jukes Cantor, cuja matriz* $\mathbf{Q}_{\mathrm{JC69}}$ *de taxas infinitesimais é dada por (2.9) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.10). Considere o estimador de máxima verossimilhança* $\hat{d}$ *para a distância* d *entre duas seqüências, dado em (5.4). A variância de* $\hat{d}$ *obtida utilizando a informação de Fisher é dada por*

$$\mathrm{Var}(\hat{d}) = -\left(\frac{\partial^2(\log L(d|\mathbf{X}))}{\partial d^2}\right)^{-1} = \frac{p\,(1-p)}{\left(1-\frac{4}{3}p\right)^2 S}, \tag{5.10}$$

e coincide com a aproximação obtida pelo Método Delta.

Demonstração: Note que $L(d|\mathbf{X})$, dado em (5.3), pode ser reescrito em função da probabilidade $p=3\alpha_\tau$ da proporção de sítios heteromorfos na seqüência. Desta forma, temos que

$$L(d|\mathbf{X}) = (\alpha_\tau)^s\,(1-3\alpha_\tau)^{S-s} = \left(\frac{p}{3}\right)^s(1-p)^{S-s}.$$

Assim, temos que

$$\log(L(d|\mathbf{X})) = s\log(p) - s\log(3) + (S-s)\log(1-p).$$

Desta forma, temos que

$$\frac{\partial^2\log(L(d|\mathbf{X}))}{\partial d^2} = s\left(\frac{p''p-(p')^2}{p^2}\right) - (S-s)\left(\frac{p''-p''p+(p')^2}{(1-p)^2}\right),$$

em que $p'=\frac{\partial p}{\partial d}$ e $p''=\frac{\partial^2 p}{\partial d^2}$. Note que $\mathbb{E}\left(\frac{s}{S}\right)=p$. Assim, escrevemos

$$\begin{aligned}\mathbb{E}\left(\frac{\partial^2 \log(L(d|\mathbf{X}))}{\partial d^2}\right) &= S\left(\frac{p''p-(p')^2}{p}\right) - S\left(\frac{p''-p''p+(p')^2}{(1-p)}\right) \\ &= \frac{-S(p')^2}{p(1-p)}.\end{aligned}$$

Como $p = 3\alpha_\tau = \frac{3}{4}(1-e^{-\frac{4}{3}d})$, temos $p' = \exp(-\frac{4}{3}d)$, e ainda $d = -\frac{3}{4}\log\left(1-\frac{4}{3}p\right)$. Assim, temos que

$$\mathbb{E}\left(\frac{\partial^2 \log(L(d|\mathbf{X}))}{\partial d^2}\right) = \frac{-S\left(1-\frac{4}{3}p\right)^2}{p(1-p)}.$$

Desta forma, obtemos a variância assintótica de $\hat{d}$, que é dada por

$$\mathrm{Var}(\hat{d}) = \left[-\mathbb{E}\left(\frac{\partial^2 \log(L(d|\mathbf{X}))}{\partial d^2}\right)\right]^{-1} = \frac{p\,(1-p)}{\left(1-\frac{4}{3}p\right)^2 S}.$$

□

Modelo Kimura-2 Parâmetros

Para compreender como se comportam os estimadores de máxima verossimilhança em modelos com mais parâmetros, estudaremos o caso do modelo Kimura-2 Parâmetros. Para este modelo, são importantes duas estatísticas das seqüências que estão sendo comparadas. Sejam s_1 o número de sítios em que se verificam *transições* (ou seja, as bases das seqüências 1 e 2 pertencem à mesma categoria, mas não são iguais), e s_2 o número de sítios em que se verificam *transversões* (ou seja, as seqüências tem bases de grupos diferentes). Lembramos ainda que γ_τ representa a probabilidade de verificarmos uma transversão, β_τ a probabilidade de verificarmos uma transição, e que essas quantidades são dadas, respectivamente, pelas seguintes expressões

$$\begin{aligned}\gamma_\tau &= \frac{1}{4}(1-e^{-4\gamma\tau}), \\ \beta_\tau &= \frac{1}{4}(1+e^{-4\gamma\tau}-2e^{2(\beta+\gamma)\tau}),\end{aligned} \tag{5.11}$$

em que γ e β são os parâmetros do processo K80 apresentados na matriz (2.14). Assim, temos que a função de verossimilhança para as duas seqüências segundo o modelo K80 é dada por

$$
\begin{aligned}
L(\tau,\beta,\gamma|\mathbf{X}) &= \prod_{u=1}^{S}\mathbb{P}(X_u^1,X_u^2)=\beta_\tau^{s_1}\gamma_\tau^{s_2}(1-\beta_\tau-2\gamma)^{S-s_1-s_2}\\
&= \left(\frac{1}{4}+\frac{1}{4}e^{-4\gamma\tau}-\frac{1}{2}e^{2(\beta+\gamma)\tau}\right)^{s_1}\left(\frac{1}{4}-\frac{1}{4}e^{-4\gamma\tau}\right)^{s_2}\\
&\times \left(\frac{1}{4}-\frac{1}{4}e^{-4\gamma\tau}+\frac{1}{2}e^{2(\beta+\gamma)\tau}\right)^{S-s_1-s_2}.
\end{aligned}
$$

Substituindo $\beta\tau$ por d_1 a distância por *transições* (dada pelo produto da taxa de *transições* pelo tempo τ) e $2\gamma\tau$ por d_2, a distância por *transversões* (definida de maneira análoga) e aplicando o logaritmo, obtemos a seguinte expressão

$$
\begin{aligned}
\log(L(d_1,d_2|\mathbf{X})) &= s_1\log\left(\frac{1}{4}+\frac{1}{4}e^{-2d_2}-\frac{1}{2}e^{2d_1+d_2}\right)+s_2\log\left(\frac{1}{4}-\frac{1}{4}e^{-2d_2}\right)\\
&+ (S-s_1-s_2)\log\left(\frac{1}{4}-\frac{1}{4}e^{-2d_2}+\frac{1}{2}e^{2d_1+d_2}\right).
\end{aligned}
$$

Resolvendo o sistema de equações

$$
\begin{aligned}
\frac{\partial\log(L(d_1,d_2|\mathbf{X}))}{\partial d_1}&=0\\
\frac{\partial\log(L(d_1,d_2|\mathbf{X}))}{\partial d_2}&=0,
\end{aligned}
$$

obtemos os estimadores de máxima verossimilhança $\hat{d_1}$ e $\hat{d_2}$, dados por

$$
\begin{aligned}
\hat{d_1} &= -\frac{1}{2}\log(1-2\hat{p}_1-\hat{p}_2)+\frac{1}{4}\log(1-2\hat{p}_2), \quad \text{se } 1-2\hat{p}_2>0 \text{ e } 1-2\hat{p}_1-\hat{p}_2>0,\\
\hat{d_2} &= -\frac{1}{2}\log(1-2\hat{p}_2), \qquad\qquad\qquad\qquad\qquad\qquad \text{se } 1-2\hat{p}_2>0,
\end{aligned}
\tag{5.12}
$$

em que $\hat{p}_1=s_1/S$ e $\hat{p}_2=s_2/S$. Note que esses estimadores só estão definidos se $1-2\hat{p}_1-\hat{p}_2>0$ e $1-2\hat{p}_2>0$. Assim como feito no modelo JC69, estendem-se esses estimadores para todos os valores de $\hat{p}_1$ e $\hat{p}_2$, de forma que temos

$$
\hat{d_1}=\begin{cases}
-\frac{1}{2}\log(1-2\hat{p}_1-\hat{p}_2)+\frac{1}{4}\log(1-2\hat{p}_2), & \text{se } 1-2\hat{p}_1-\hat{p}_2>0 \text{ e } 1-2\hat{p}_2>0,\\
\infty, & \text{caso contrário,}
\end{cases}
\tag{5.13}
$$

e

$$\hat{d}_2 = \begin{cases} -\frac{1}{2}\log(1-2\hat{p}_2), & \text{se } 1-2\hat{p}_2 > 0, \\ \infty, & \text{caso contrário.} \end{cases} \tag{5.14}$$

Observe ainda que a distância d entre as duas seqüências está definida como $d = (\beta + 2\gamma)\tau$ para o modelo K80. Assim, seu estimador de máxima verossimilhança é dado por

$$\hat{d} = \hat{d}_1 + \hat{d}_2 = \begin{cases} -\frac{1}{2}\log(1-2\hat{p}_2-\hat{p}_2) - \frac{1}{4}\log(1-2\hat{p}_2), & \text{se } 1-2\hat{p}_1-\hat{p}_2 > 0 \,\text{e}\, 1-2\hat{p}_2 > 0, \\ \infty, & \text{caso contrário,} \end{cases} \tag{5.15}$$

No modelo K80 temos três parâmetros de interesse: τ, γ e β. Entretanto, temos apenas como estimar $d_2 = 2\gamma\tau$ e $d_1 = \beta\tau$. A razão para isso é que, assim como para o modelo JC69, o tempo de divergência τ entre as espécies aparece sempre multiplicando os parâmetros do modelo de substituição de bases nas expressões das probabilidades de mutação. Dessa forma, a menos que se fixe o valor de algum parâmetro, não podemos obter as estimativas dos parâmetros individualmente.

A variável aleatória W que apresenta três resultados, transição com probabilidade p_1, transversão com probabilidade p_2 e nenhuma mutação com probabilidade $1-p_1-p_2$ tem distribuição multinomial.

Observação 5.3. Seja $\mathbf{W} = (W_1, W_2, W_3)$ um vetor de variáveis aleatórias com distribuição multinomial, $\mathbf{W} \sim \mathcal{M}(S, p_1, p_2, p_3)$, tal que $p_1 + p_2 + p_3 = 1$ e $0 \leq p_i \leq 1$, para $i \in \{1, 2, 3\}$. Então, W_i assume valores em $\{0, \cdots, S\}$, para $i \in \{1, 2, 3\}$, tais que $w_1 + w_2 + w_3 = S$. A função massa de probabilidade $f_{\mathbf{W}}(\cdot)$ de W é dada por

$$f_{\mathbf{W}}(\mathbf{w}) = \begin{cases} \frac{S!}{w_1!w_2!w_3!} p_1^{w_1} p_2^{w_2} p_3^{w_3}, & \text{se } w_1 + w_2 + w_3 = S, \\ 0, & \text{caso contrário.} \end{cases}$$

Temos que $\mathbb{E}(W_i) = Sp_i$, $\text{Var}(W_i) = Sp_i(1-p_i)$, $\text{Cov}(W_i, W_j) = -Sp_ip_j$, com $i, j \in \{1, 2, 3\}$ e $i \neq j$. Ainda, $\hat{p}_i = \frac{w_i}{S}$ é o estimador de máxima verossimilhança para p_i, para $i \in \{1, 2, 3\}$.

Para o cálculo da distribuição assintótica dos estimadores $\hat{d}_1$ e $\hat{d}_2$, recorremos ao Método Delta no caso multivariado, considerando $\hat{d}_1 = g_1(\hat{p}_1, \hat{p}_2)$ e $\hat{d}_2 = g_2(\hat{p}_1, \hat{p}_2)$. Assim, temos o seguinte resultado.

Lema 5.4. *Considere o modelo Kimura-2 Parâmetros, cuja matriz* $\mathbf{Q}_{\text{K80}}$ *de taxas infinitesimais é dada por (2.14) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.10).*

Sejam $\hat{p}_1$ e $\hat{p}_2$ tais que $1-2\hat{p}_1-\hat{p}_2>0$ e $1-2\hat{p}_2>0$. O estimador de máxima verossimilhança das distâncias por transições e por transversões $\mathcal{D}=(d_1,d_2)$, dado por (5.12), tem distribuição assintótica dada por

$$\sqrt{S}(\hat{\mathcal{D}}-\mathcal{D}) \xrightarrow{d} \mathcal{Z}, \quad \textit{quando } S\to\infty,$$

onde $\mathcal{Z}\sim N(0,\boldsymbol{\Sigma}_{\mathcal{D}})$, com

$$\boldsymbol{\Sigma}_{\mathcal{D}}=\begin{pmatrix} \mathrm{Var}(\hat{d}_1) & \mathrm{Cov}(\hat{d}_1,\hat{d}_2) \\ \mathrm{Cov}(\hat{d}_1,\hat{d}_2) & \mathrm{Var}(\hat{d}_2) \end{pmatrix},$$

onde

$$\mathrm{Var}(\hat{d}_1)=a\,p_1(1-p_1)-b\,p_1p_2+c\,p_2(1-p_2),$$

$$\mathrm{Var}(\hat{d}_2)=\frac{p_2(1-p_2)}{(1-2p_2)^2},$$

$$\mathrm{Cov}(\hat{d}_1,\hat{d}_2)=\frac{p_1p_2}{(1-2p_1-p_2)(1-2p_2)}+\frac{p_2(1-p_2)(1+2p_1-3p_2)}{2(1-2p_1-p_2)(1-2p_2)^2},$$

com

$$\begin{aligned} a &= \frac{1}{(1-2p_1-p_2)^2} \\ b &= \frac{1+2p_1-3p_2}{(1-2p_1-p_2)^2(1-2p_2)} \\ c &= \frac{(1+2p_1-3p_2)^2}{4(1-2p_1-p_2)^2(1-2p_2)^2}. \end{aligned}$$

Assim, $\hat{d}_1$ e $\hat{d}_2$ são assintoticamente não viciados e consistentes, uma vez que suas variâncias, dadas, respectivamente, por $Var(\hat{d}_1)/S$ e $Var(\hat{d}_2)/S$, tendem a zero, quando $S\to\infty$.

Demonstração: Sabemos que $\hat{\mathcal{P}}=(\hat{p}_1,\hat{p}_2)$ é estimador de máxima verossimilhança para $\mathcal{P}=(p_1,p_2)$. Assim, nas condições do Teorema 5.1, temos que $\sqrt{S}(\hat{\mathcal{P}}-\mathcal{P}) \xrightarrow{d} \mathcal{Z}$, quando $S\to\infty$, onde $\mathcal{Z}\sim\mathbf{N}(0,\boldsymbol{\Sigma}_{\mathcal{P}})$. Ainda, pela Observação 5.3, temos que a matriz de variâncias-covariâncias $\boldsymbol{\Sigma}_{\mathcal{P}}$ de $\hat{p}_1$ e $\hat{p}_2$ é tal que

$$\boldsymbol{\Sigma}_{\mathcal{P}}=\begin{pmatrix} \hat{p}_1(1-\hat{p}_1) & \hat{p}_1\hat{p}_2 \\ \hat{p}_1\hat{p}_2 & \hat{p}_2(1-\hat{p}_2) \end{pmatrix}. \tag{5.16}$$

Tomando $\mathcal{D}=\mathbf{g}(\mathcal{P})=(d_1,d_2)=(g_1(p_1,p_2),g_2(p_1,p_2))$, temos, pelo Método Delta, que $\sqrt{S}(\mathbf{g}(\hat{\mathcal{P}})-\mathbf{g}(\mathcal{P})) \xrightarrow{d} \mathcal{Z}$, quando $S\to\infty$, em que $\mathcal{Z}\sim\mathbf{N}(0,\boldsymbol{\Sigma}_{\mathcal{D}})$. Ainda

$$\boldsymbol{\Sigma}_{\mathcal{D}} = \begin{pmatrix} \text{Var}(\hat{d}_1) & \text{Cov}(\hat{d}_1, \hat{d}_2) \\ \text{Cov}(\hat{d}_1, \hat{d}_2) & \text{Var}(\hat{d}_2) \end{pmatrix} = \mathbf{J} \ \boldsymbol{\Sigma}_{\mathcal{P}} \ \mathbf{J}', \tag{5.17}$$

em que $\boldsymbol{\Sigma}_{\mathcal{P}}$ é dada pela expressão (5.16), e $\mathbf{J}$ é a matriz Jacobiana, dada por

$$\mathbf{J} = \left(\begin{array}{cc} \frac{\partial \hat{d}_1}{\partial p_1} & \frac{\partial \hat{d}_1}{\partial p_2} \\ \frac{\partial \hat{d}_2}{\partial p_1} & \frac{\partial \hat{d}_2}{\partial p_2} \end{array} \right) \Bigg|_{(p_1,p_2)=(\hat{p}_1,\hat{p}_2)}.$$

Com base na expressão (5.12), encontramos as derivadas parciais de $\hat{d}_1$ e $\hat{d}_2$, de forma que

$$\mathbf{J} = \begin{pmatrix} \frac{1}{1-2\hat{p}_1-\hat{p}_2} & \frac{1}{1-2\hat{p}_1-\hat{p}_2} - \frac{1}{2(1-2\hat{p}_2)} \\ 0 & \frac{1}{1-2\hat{p}_2} \end{pmatrix}. \tag{5.18}$$

Note ainda que pela igualdade (5.17), a variância de $\hat{d}_1$ é dada por

$$\text{Var}(\hat{d}_1) = \left(\frac{\partial \hat{d}_1}{\partial \hat{p}_1}\right)^2 \text{Var}(\hat{p}_1) + 2\left(\frac{\partial \hat{d}_1}{\partial \hat{p}_1}\right)\left(\frac{\partial \hat{d}_1}{\partial \hat{p}_2}\right)\text{Cov}(\hat{p}_1, \hat{p}_2) + \left(\frac{\partial \hat{d}_1}{\partial \hat{p}_2}\right)^2 \text{Var}(\hat{p}_2). \tag{5.19}$$

Substituindo (5.16) e (5.18) em (5.19) obtemos a variância assintótica de d_1, dada por

$$\text{Var}(\hat{d}_1) = a\, p_1(1-p_1) - b\, p_1 p_2 + c\, p_2(1-p_2),$$

onde

$$\begin{aligned} a &= \frac{1}{(1-2p_1-p_2)^2} \\ b &= \frac{1+2p_1-3p_2}{(1-2p_1-p_2)^2(1-2p_2)} \\ c &= \frac{(1+2p_1-3p_2)^2}{4(1-2p_1-p_2)^2(1-2p_2)^2}. \end{aligned}$$

A variância assintótica de $\hat{d}_2$ é obtida substituindo (5.16) e (5.18) em forma análoga a (5.19). Assim, temos que

$$\text{Var}(\hat{d}_2) = \frac{p_2(1-p_2)}{(1-2p_2)^2}.$$

Além disso, da expressão (5.17) encontramos a seguinte expressão para a covariância de $\hat{d}_1$ e $\hat{d}_2$

$$\begin{aligned}\mathrm{Cov}(\hat{d}_1,\hat{d}_2) &= \left(\frac{\partial\hat{d}_1}{\partial\hat{p}_1}\right)\left(\frac{\partial\hat{d}_2}{\partial\hat{p}_1}\right)\mathrm{Var}(\hat{p}_1)+\left(\frac{\partial\hat{d}_2}{\partial\hat{p}_2}\right)\left(\frac{\partial\hat{d}_1}{\partial\hat{p}_2}\right)\mathrm{Var}(\hat{p}_2)\\ &+ \mathrm{Cov}(\hat{p}_1,\hat{p}_2)\left[\left(\frac{\partial\hat{d}_1}{\partial\hat{p}_1}\right)\left(\frac{\partial\hat{d}_2}{\partial\hat{p}_2}\right)+\left(\frac{\partial\hat{d}_1}{\partial\hat{p}_2}\right)\left(\frac{\partial\hat{d}_2}{\partial\hat{p}_1}\right)\right]. \qquad (5.20)\end{aligned}$$

Substituindo (5.16) e (5.18) em (5.20) obtemos a covariância assintótica, dada por

$$\mathrm{Cov}(\hat{d}_1,\hat{d}_2)=\frac{p_1p_2}{(1-2p_1-p_2)(1-2p_2)}+\frac{p_2(1-p_2)(1+2p_1-3p_2)}{2(1-2p_1-p_2)(1-2p_2)^2}.$$

□

Lema 5.5. *Considere o modelo Kimura-2 Parâmetros, cuja matriz* $\mathbf{Q}_{\mathrm{K80}}$ *de taxas infinitesimais é dada por (2.14) e cujo vetor* $\mathbf{p}_0$ *de probabilidades iniciais é dado por (2.10). Seja* $d=(\beta+2\gamma)\tau$ *a distância entre duas seqüências para o modelo* K80*, cujo estimador de máxima verossimilhança* $\hat{d}$ *é dado por (5.15). Então, sua variância assintótica é dada por*

$$Var(\hat{\mathrm{d}}) = \mathrm{a}'^2\mathrm{d}_1+\mathrm{b}'^2\mathrm{d}_2-(\mathrm{a}'\mathrm{d}_1+\mathrm{b}'\mathrm{d}_2)^2$$

onde

$$a'=\frac{1}{1-2p_1-p_2} \quad e \quad b'=\frac{a'}{2}+\frac{1}{2(1-2p_2)}.$$

Além disso, o estimador $\hat{d}$ *é consistente, pois sua variância, dada por* $Var(\hat{\mathrm{d}})/\mathrm{S}$*, tende a zero quando o número de sítios* S *tende a infinito.*

Demonstração: Ver Yang (2007).

Quando consideramos apenas duas seqüências, a estimação dos parâmetros dos modelos de substituição de bases e do tempo de divergência entre as seqüências é relativamente simples, e propriedades desses estimadores podem ser obtidas. Entretanto, quando consideramos mais de duas seqüências, a filogenia que as relaciona deve ser considerada, pois ela altera a expressão da função de verossimilhança. Isso dificulta a obtenção de expressões para os estimadores dos parâmetros, que geralmente são obtidos por otimização numérica. Entretanto, se considerarmos um caso particular do problema, em que são conhecidas todas as seqüências ancestrais, encontramos resultados interessantes sobre a variância dos estimadores dos parâmetros dos modelos.

5.2 Cota Inferior para a Variância dos Estimadores

Queremos encontrar uma cota mínima para a variância dos estimadores dos parâmetros de modelos de substituição de bases. Para tanto utilizamos o limite inferior de Cramér-Rao (também conhecido como limite da Informação de Fisher), apresentado a seguir.

Teorema 5.3. *Seja* $T(\mathbf{X})$ *uma estatística tal que* $\mathrm{Var}(T(\mathbf{X})) < \infty$, $\forall\theta \in \Theta$. *Denote* $\mathbb{E}(T(\mathbf{X})) \equiv \psi(\theta)$, *e suponha que valem*

- *o conjunto* $A = \{\mathbf{X}; L(\theta|\mathbf{X}) > 0\}$ *não depende de* θ, *e* $\forall\mathbf{X} \in A$, $\forall\theta \in \Theta$ *existe* $\frac{\partial}{\partial\theta}(\log(L(\theta|\mathbf{X})))$ *e é finita;*
- *se* $T(\mathbf{X})$ *é uma estatística tal que* $\mathbb{E}(T(\mathbf{X})) < \infty$, $\forall\theta \in \Theta$, *então as operações de integração e diferenciação com relação a* θ *podem ser trocadas entre si em*

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} T(\mathbf{X})\log(L(\theta|\mathbf{X}))d\mathbf{X} \tag{5.21}$$

- *e* $0 < \mathbf{I}(\theta) < \infty$,

onde $\mathbf{I}(\theta) = \frac{\partial^2}{\partial\theta^2}(\log(L(\theta|\mathbf{X})))$. *Então,* $\psi(\mathbf{X})$ *é diferenciável e*

$$Var(T(\mathbf{X})) \leq \frac{[\psi'(\theta)]^2}{\mathbf{I}(\theta)}. \tag{5.22}$$

Demonstração: Ver Kendall e Stuart (1973). □

Corolário 5.1. *Nas condições do Teorema 5.3, se* $T(\mathbf{X})$ *é um estimador não viciado de* θ, *então* $Var(T(\mathbf{X})) \leq \frac{1}{\mathbf{I}(\theta)}$.

Note que para utilizar esse resultado, devemos primeiro garantir que as condições de regularidade estejam satisfeitas.

Lema 5.6. *As condições do Teorema 5.3 estão satisfeitas para os modelos de substituição de bases i.i.d. com taxas de mutação constantes.*

Demonstração: Ver demonstração do Lema 3.1. □

Observe, ainda, que embora os estimadores de máxima verossimilhança possam ser viciados, sabemos que eles são assintoticamente não viciados. Assim, o Lema 5.6 pode ser utilizado.

Cota Inferior para a Variância do Estimador de Máxima Verossimilhança para o Parâmetro α do Modelo JC69 (Caso Particular)

Queremos determinar uma cota inferior para a variância do estimador do parâmetro α do modelo Jukes-Cantor. Para tanto, consideramos o caso em que temos a maior informação possível sobre o processo. Assim, considere a hipótese

$$\begin{aligned}\mathbf{H}: \quad &\bullet \text{ conhecemos a estrutura da filogenia } F \text{ que relaciona as seqüências;} \\ &\bullet \text{ conhecemos os tempos de coalescência } \bar{\tau} = \{\tau_1, \cdots, \tau_{2N-2}\}; \\ &\bullet \text{ conhecemos as seqüências dos nós internos da filogenia } X_{N+1}, \cdots, X_{2N-1} \\ &\quad \text{além das seqüências da amostra } X_1, \cdots, X_N. \end{aligned} \tag{5.23}$$

Vamos começar considerando o caso de uma árvore com raiz e $N = 3$. Desta forma, temos o seguinte lema.

Lema 5.7. *Considere α o parâmetro do modelo JC69 cuja matriz Q_{JC69} de taxas infinitesimais é dada na expressão (2.9). Considere o caso em que $N = 3$ e a hipótese* **H**, *dada em (5.23), está satisfeita. Então, o estimador de máxima verossimilhança $\hat{\alpha}_{\mathbf{H}}$ para α é dado pela solução da equação*

$$0 = \sum_{i=1}^{4} \left(\frac{s_i^+ \tau_i e^{-4\alpha\tau_i}}{\frac{1}{4}(1 - e^{-4\alpha\tau_i})} - \frac{3 s_i^- \tau_i e^{-4\alpha\tau_i}}{\frac{1}{4}(1 + 3e^{-4\alpha\tau_i})} \right), \tag{5.24}$$

onde $s_i^+ = \sum_{j=1}^{S} I(X_i^j \neq X_{h(i)}^j)$ e $s_i^- = S - s_i^+$. Além disso, uma cota inferior para a variância de $\hat{\alpha}_{\mathbf{H}}$ é dada por

$$\text{Var}(\hat{\alpha}_{\mathbf{H}}) = \left(48S \sum_{i=1}^{4} \tau_i^2 e^{-4\alpha\tau_i} \left(\frac{\tau_i}{(1 - e^{-4\alpha\tau_i})^2} - \frac{(1 - 3\tau_i)}{(1 + 3e^{-4\alpha\tau_i})^2} \right) \right)^{-1}. \tag{5.25}$$

Demonstração: Note que, para $N = 3$, a estrutura da filogenia que relaciona as seqüências está representada na Figura 5.2, a menos da troca de ordem das seqüências.

Existem $4^{2N-1} = 4^5$ possíveis combinações de bases nas seqüências $X_1, \cdots, X_{2N-1}$. Entretanto, devido às peculiaridades do modelo Jukes-Cantor, os padrões podem ser reunidos em 16 grupos de probabilidades distintas. O primeiro grupo, por exemplo, reúne aqueles padrões em que todas as seqüências possuem a mesma base. A probabilidade do modelo Jukes-Cantor gerar uma das combinações do padrão 1, p_1, é dada por

$$p_1 = \frac{1}{4}(1 - 3\alpha_{\tau_1})(1 - 3\alpha_{\tau_2})(1 - 3\alpha_{\tau_3})(1 - 3\alpha_{\tau_4}),$$

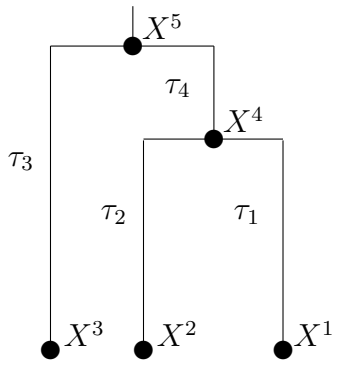


Figura 5.2: Filogenia F.

em que α_{τ_i}, dado pela expressão (2.13), é a probabilidade de uma mutação específica em um tempo τ_i segundo o modelo Jukes-Cantor. As probabilidades dos demais padrões podem ser obtidas de maneira semelhante, de forma que temos

$$\begin{aligned}
p_2 &= \frac{1}{4}\alpha_{\tau_1}(1-3\alpha_{\tau_2})(1-3\alpha_{\tau_3})(1-3\alpha_{\tau_4}) \\
p_3 &= \frac{1}{4}(1-3\alpha_{\tau_1})\alpha_{\tau_2}(1-3\alpha_{\tau_3})(1-3\alpha_{\tau_4}) \\
p_3 &= \frac{1}{4}(1-3\alpha_{\tau_1})(1-3\alpha_{\tau_2})(1-3\alpha_{\tau_3})(1-3\alpha_{\tau_4}) \\
&\vdots \\
p_{15} &= \frac{1}{4}\alpha_{\tau_1}\alpha_{\tau_2}\alpha_{\tau_3}(1-3\alpha_{\tau_4}) \\
p_{16} &= \frac{1}{4}\alpha_{\tau_1}\alpha_{\tau_2}\alpha_{\tau_3}\alpha_{\tau_4}.
\end{aligned}$$

Seja s_i o número de ocorrências de padrões do grupo i na amostra. Então, pela expressão (3.2), a função de verossimilhança é dada por

$$L(\alpha|\mathbf{X}) = \prod_{j=1}^{S} \mathbb{P}(\mathbf{X}|\alpha) = \prod_{i=1}^{16} p_i^{s_i}. \tag{5.26}$$

Assim, temos que

$$\log(L(\alpha|\mathbf{X})) = \sum_{i=1}^{16} s_i \log(p_i).$$

Reagrupando as probabilidades, obtemos

$$\begin{aligned}\log(L(\alpha|\mathbf{X})) &= S\log(0.25) + s_1^+\log(\alpha_{\tau_1}) + s_1^-\log(1-3\alpha_{\tau_1}) \\ &+ s_2^+\log(\alpha_{\tau_2}) + s_2^-\log(1-3\alpha_{\tau_2}) + s_3^+\log(\alpha_{\tau_3}) \\ &+ s_3^-\log(1-3\alpha_{\tau_3}) + s_4^+\log(\alpha_{\tau_4}) + s_4^-\log(1-3\alpha_{\tau_4}),\end{aligned}$$

em que

$$\begin{aligned}s_1^+ &= \sum_{j=1}^{S} I(X_1^j = X_4^j) \qquad s_1^- = S - s_1^+ \\ s_2^+ &= \sum_{j=1}^{S} I(X_2^j = X_4^j) \qquad s_2^- = S - s_2^+ \\ s_3^+ &= \sum_{j=1}^{S} I(X_3^j = X_5^j) \qquad s_3^- = S - s_3^+ \\ s_4^+ &= \sum_{j=1}^{S} I(X_4^j = X_5^j) \qquad s_4^- = S - s_4^+.\end{aligned}$$

Assim, S_i^+ conta o número de vezes em que foi verificada mutação entre as seqüências ligadas pelo ramo de comprimento τ_i, e S_i^- conta o número de vezes em que não se verificou mudança na base.

Derivando a log-verossimilhança obtemos

$$0 = \frac{\partial \log(L(\alpha|\mathbf{X}))}{\partial\alpha} = \sum_{i=1}^{4}\left(\frac{s_i^+}{\alpha_{\tau_i}}\frac{\partial\alpha_{\tau_i}}{\partial\alpha} + \frac{s_i^-}{(1-3\alpha_{\tau_i})}\frac{\partial(1-3\alpha_{\tau_i})}{\partial\alpha}\right).$$

Mas, lembrando que, pela expressão (2.13),

$$\begin{aligned}\alpha_{\tau_i} &= \frac{1}{4}(1-e^{-4\alpha\tau_i}) \\ 1-3\alpha_{\tau_i} &= \frac{1}{4}(1+3e^{-4\alpha\tau_i}),\end{aligned} \tag{5.27}$$

temos que

$$0 = \frac{\partial \log(L(\alpha|\mathbf{X}))}{\partial\alpha} = \sum_{i=1}^{4}\left(\frac{s_i^+\tau_i e^{-4\alpha\tau_i}}{\frac{1}{4}(1-e^{-4\alpha\tau_i})} - \frac{3s_i^-\tau_i e^{-4\alpha\tau_i}}{\frac{1}{4}(1+3e^{-4\alpha\tau_i})}\right).$$

A solução dessa equação fornece o estimador de máxima verossimilhança $\hat{\alpha}_{\mathbf{H}}$ para o parâmetro α, dado em (5.24).

Queremos prosseguir para encontrar a variância de $\hat{\alpha}_{\mathbf{H}}$, e para isso lembramos o Corolário 5.1, que garante que o estimador de máxima verossimilhança é assintoticamente normal, não viciado e com variância dada por

$$\mathrm{Var}(\hat{\alpha}_{\mathbf{H}}) = \frac{1}{\mathbb{E}\left(-\frac{\partial^2 \log(L(\alpha|\mathbf{X}))}{\partial \alpha^2}\right)}.$$

A segunda derivada de $\log(L(\alpha|\mathbf{X}))$ é dada por

$$\frac{\partial^2 \log(L(\alpha|\mathbf{X}))}{\partial \alpha^2} = \sum_{i=1}^{4} \left(\frac{-16 s_i^+ \tau_i^2 e^{-4\alpha\tau_i}}{(1-e^{-4\alpha\tau_i})^2} + \frac{48 s_i^- \tau_i^2 e^{-4\alpha\tau_i}}{(1+3e^{-4\alpha\tau_i})^2} \right).$$

Para obter a variância do estimador, devemos ainda lembrar que

$$\begin{aligned} \mathbb{E}(s_i^+) &= 3\tau_i S \\ \mathbb{E}(s_i^-) &= \mathbb{E}(S - s_i^+) = S(1-3\tau_i). \end{aligned}$$

Assim, temos que

$$\mathbb{E}\left(\frac{\partial^2 \log(L(\alpha|\mathbf{X}))}{\partial \alpha^2}\right) = 48S \sum_{i=1}^{4} \tau_i^2 e^{-4\alpha\tau_i} \left(\frac{-\tau_i}{(1-e^{-4\alpha\tau_i})^2} + \frac{(1-3\tau_i)}{(1+3e^{-4\alpha\tau_i})^2} \right). \qquad (5.28)$$

Pela expressão (5.28) e pelo Corolário 5.1 concluímos o lema. □

Observação 5.4. O número total de nós de uma filogenia é $2N-1$, pois uma filogenia com N nós externos possui exatamente $N-1$ nós internos. Isso pode ser visto se considerarmos o número de linhagens (ramos) existentes à altura de cada nó interno da filogenia. Na altura dos nós externos temos exatamente N linhagens, já na altura do primeiro nó interno (por exemplo X^4 na Figura 5.2) temos a união de duas linhagens, de forma que ficamos com exatamente $N-1$ ramos. Esse procedimento se segue, de forma que na altura do k-ésimo nó interno restam $N-k$ ramos. Assim, a filogenia fica reduzida a apenas um ramo no $N-1$-ésimo nó interno.

Cota Inferior para a Variância do Estimador de Máxima Verossimilhança para o Parâmetro α do Modelo JC69 (Caso Geral)

O Lema 5.7 pode ser facilmente extendido para o caso de uma filogenia genérica, relacionando N seqüências filhas, com raiz conhecida. Nesse caso, assim como antes, consideramos as hipóteses **H**, dadas em (5.23), ou seja, temos acesso às seqüências $X_1, \cdots, X_{2N-1}$ de todos os nós da filogenia F e aos comprimentos dos ramos $\tau_1, \cdots, \tau_{2N-1}$.

Lema 5.8. *Considere α o parâmetro do modelo JC69 cuja matriz Q_{JC69} de taxas infinitesimais é dada na expressão (2.9). Considere, que a hipótese* **H**, *dada em (5.23), está satisfeita. Então, o estimador de máxima verossimilhança $\hat{\alpha}_{\mathbf{H}}$ para α é dado pela solução da equação*

$$\sum_{i=1}^{2N-2} \left(\frac{s_i^+ \tau_i e^{-4\alpha\tau_i}}{\frac{1}{4}(1 - e^{-4\alpha\tau_i})} - \frac{3 s_i^- \tau_i e^{-4\alpha\tau_i}}{\frac{1}{4}(1 + 3e^{-4\alpha\tau_i})} \right) = 0, \tag{5.29}$$

onde $s_i^+ = \sum_{j=1}^{S} I(X_i^j \neq X_{h(i)}^j)$ e $s_i^- = S - s_i^+$. Além disso, uma cota inferior para a variância de $\hat{\alpha}_{\mathbf{H}}$ é dada por

$$\text{Var}(\hat{\alpha}_{\mathbf{H}}) = \left(48S \sum_{i=1}^{2N-2} \tau_i^2 e^{-4\alpha\tau_i} \left(\frac{\tau_i}{(1 - e^{-4\alpha\tau_i})^2} - \frac{(1 - 3\tau_i)}{(1 + 3e^{-4\alpha\tau_i})^2} \right) \right)^{-1}. \tag{5.30}$$

Demonstração: Para uma árvore filofenética F com N seqüências, temos 4^{2N-1} possíveis combinações de bases nas $2N-1$ seqüências. Entretanto, como estamos considerando o modelo Jukes-Cantor, essas combinações podem ser reunidas em 2^{2N-2} combinações com probabilidades distintas. Denotamos por p_i a probabilidade da combinação do grupo i e s_i o número de vezes que combinações desse grupo aparecem na amostra. A função de verossimilhança para esse modelo é dada por

$$L(\alpha|\mathbf{X}) = \prod_{j=1}^{S} \mathbb{P}(\mathbf{X}|\alpha) = \prod_{i=1}^{2^{2N-2}} p_1^{s_i}.$$

Note que o ramo de comprimento τ_i é aquele entre a seqüência X_i e seu ancestral imediato na filogenia, denotado por $X_{h(i)}$ (ver Figura 5.2). Reagrupando as probabilidades como feito anteriormente e escrevendo

$$s_i^+ = \sum_{j=1}^{S} I(X_i^j \neq X_{h(i)}^j) \quad \text{e} \quad s_i^- = S - s_i^+, \tag{5.31}$$

para todo $i \in \{1, \cdots, 2^{2N-2}\}$, é facil ver que

$$\log(L(\alpha|\mathbf{X})) = S\log\left(\frac{1}{4}\right) + \sum_{i=1}^{2N-2}\left(s_i^+\log(\alpha_{\tau_i}) + s_i^-\log(1-3\alpha_{\tau_i})\right).$$

Considerando a expressão (5.27) temos que o estimador de máxima verossimilhança $\hat{\alpha}_{\mathbf{H}}$ para o parâmetro α é dado pela solução da equação

$$0 = \frac{\partial\log(L(\alpha|\mathbf{X}))}{\partial\alpha} = \sum_{i=1}^{2N-2}\left(\frac{s_i^+\tau_i e^{-4\alpha\tau_i}}{\frac{1}{4}(1-e^{-4\alpha\tau_i})} - \frac{3s_i^-\tau_i e^{-4\alpha\tau_i}}{\frac{1}{4}(1+3e^{-4\alpha\tau_i})}\right), \tag{5.32}$$

onde s_i^+ e s_i^-, para todo $i \in \{1, \cdots, 2N-2\}$, estão definidos na expressão (5.31).

Procedendo de forma análoga ao caso com três seqüências, derivamos (5.32) para obter a segunda derivada de $\log(L(\alpha|\mathbf{X}))$, dada por

$$\frac{\partial^2\log(L(\alpha|\mathbf{X}))}{\partial\alpha^2} = \sum_{i=1}^{2N-2}\left(\frac{-16s_i^+\tau_i^2 e^{-4\alpha\tau_i}}{(1-e^{-4\alpha\tau_i})^2} + \frac{48s_i^-\tau_i^2 e^{-4\alpha\tau_i}}{(1+3e^{-4\alpha\tau_i})^2}\right).$$

Para obter a variância do estimador, devemos ainda lembrar que

$$\begin{aligned}\mathbb{E}(s_i^+) &= 3\tau_i S\\ \mathbb{E}(s_i^-) &= \mathbb{E}(S - s_i^+) = S(1-3\tau_i).\end{aligned}$$

Assim, do Corolário 5.1 obtemos a variância assintótica do estimador, dada por

$$\begin{aligned}\mathrm{Var}(\hat{\alpha}_{\mathbf{H}}) &= \left(\mathbb{E}\left(\frac{\partial^2}{\partial\alpha^2}\log(L(\alpha|\mathbf{X})\right)\right)^{-1}\\ &= \left(48S\sum_{i=1}^{2N-2}\tau_i^2 e^{-4\alpha\tau_i}\left(\frac{\tau_i}{(1-e^{-4\alpha\tau_i})^2} - \frac{(1-3\tau_i)}{(1+3e^{-4\alpha\tau_i})^2}\right)\right)^{-1}.\end{aligned} \tag{5.33}$$

□

Cota Inferior para a Variância do Estimador de Máxima Verossimilhança para o Parâmetro K do Modelo K80

Considere as mesmas hipóteses **H** utilizadas para $\hat{\alpha}_{\mathbf{H}}$, dadas por (5.23). Considere, agora, o modelo K80, sob a restrição de que a taxa de mutação geral, que é dada por $\beta + 2\gamma$, é 1.

Lema 5.9. *Considere o modelo K80 cuja matriz Q_{K80} de taxas infinitesimais é dada por (2.14), e seja $K = \frac{\beta}{\gamma}$ a taxa de transições e transversões deste modelo. Suponha, ainda, que a hipótese* **H**, *dada em (5.23), está satisfeita. Então, o estimador de máxima verossimilhança $\hat{K}_{\mathbf{H}}$ para K é dado pela solução da equação*

$$\begin{aligned}\sum_{i=1}^{2N-2} \frac{\tau_i}{(1+K)^2} \Bigg[& \frac{s_i'}{\beta_{\tau_i}} \left(e^{-\frac{4\tau_i}{K+2}} + e^{-\frac{2\tau_i(K+1)}{K+2}} \right) \\ + \quad & \frac{s_i^0}{1-\beta_{\tau_i}-2\gamma_{\tau_i}} \left(e^{-\frac{4\tau_i}{K+2}} - e^{-\frac{2\tau_i(K+1)}{K+2}} \right) - \frac{s_i''}{\gamma_{\tau_i}} e^{-\frac{4\tau_i}{K+2}} \Bigg] = 0, \end{aligned} \tag{5.34}$$

onde s_i'' é o número de vezes que se observa uma transversão no ramo de comprimento τ_i, s_i' é o número de vezes que se observa uma transição no mesmo ramo, e s_i^0 é o numero de vezes em que não há mutação aparente neste ramo. Além disso, uma cota inferior para a variância de $\hat{K}_{\mathbf{H}}$ é dada por

$$\mathrm{Var}(\hat{K_{\mathbf{H}}}) = \left(- \sum_{i=1}^{2N-2} \left[\frac{S\tau_i}{K+2} a(\tau_i) + S(1-\tau_i) b(\tau_i) + \frac{SK\tau_i}{K+2} c(\tau_i) \right] \right)^{-1},$$

onde

$$\begin{aligned} a(\tau_i) &= \frac{\left(\frac{4\tau_i}{K+2} - 2\right) \tau_i e^{-\frac{4\tau_i}{K+2}} - \left(\frac{2\tau_i}{K+2} + 2\right) \tau_i e^{-\frac{2\tau_i(K+1)}{K+2}}}{(K+2)^3 \beta_{\tau_i}} - \frac{\left(\tau_i e^{-\frac{4\tau_i}{K+2}} + \tau_i e^{-\frac{2\tau_i(K+1)}{K+2}} \right)^2}{(K+2)^4 \beta_{\tau_i}^2} \\ b(\tau_i) &= \frac{\left(\frac{4\tau_i}{K+2} - 2\right) \tau_i e^{-\frac{4\tau_i}{K+2}} + \left(\frac{2\tau_i}{K+2} + 2\right) \tau e^{-\frac{2\tau_i(K+1)}{K+2}}}{(K+2)^3 \beta_{\tau_i}} - \frac{\left(\tau_i e^{-\frac{4\tau_i}{K+2}} - \tau_i e^{-\frac{2\tau_i(K+1)}{K+2}} \right)^2}{(K+2)^4 \beta_{\tau_i}^2} \\ c(\tau_i) &= -\frac{\left(\frac{4\tau_i}{K+2} - 2\right) \tau_i e^{-\frac{4\tau_i}{K+2}}}{(K+2)^3 \gamma_{\tau_i}} - \frac{\left(\tau_i e^{-\frac{4\tau_i}{K+2}} \right)^2}{(K+2)^4 \gamma_{\tau_i}^2}, \end{aligned}$$

e β_{τ_i} e γ_{τ_i} são as probabilidades de transição do modelo K80, dadas em (2.16).

Demonstração: Sob a restrição de que a taxa de mutação geral satisfaz $\beta + 2\gamma = 1$, as probabilidades de mutação podem ser reescritas em função do parâmetro $K = \frac{\beta}{\gamma}$, que representa a taxa de *transições* e *transversões*. Assim, as probabilidades de mutação para esse modelo, dadas na expressão (2.16), podem ser reescritas como

$$\beta_\tau = \frac{1}{4}\left(1 + e^{-\frac{4\tau}{K+2}} - 2e^{-\frac{2\tau(K+1)}{K+2}}\right) \tag{5.35}$$

$$\gamma_\tau = \frac{1}{4}\left(1 - e^{-\frac{4\tau}{K+2}}\right) \tag{5.36}$$

$$\mathbb{P}(i|i,\tau) = 1 - \beta_\tau - 2\gamma_\tau = \frac{1}{4}\left(1 + e^{-\frac{4\tau}{K+2}} + 2e^{-\frac{2\tau(K+1)}{K+2}}\right), \tag{5.37}$$

para todo $i \in E$. Deste modo, a função de verossimilhança, dada por

$$L(\boldsymbol{\theta}|\mathbf{X}) = \prod_{l=1}^{2^{2N-1}} p_l^{s_l} = \exp\left(\sum_{l=1}^{4^{2N-1}} s_l \log p_l\right),$$

pode ser facilmente rearranjada, da mesma forma como foi feito para o modelo JC69, em função das mutações que ocorrem em cada ramo da filogenia. Note que, para este modelo existem três probabilidades de mutação distintas para um dado τ. Assim, a log-verossimilhança para esse modelo pode ser escrita como

$$\log(L(\boldsymbol{\theta}|\mathbf{X})) = \sum_{i=1}^{2N-2} S\log\frac{1}{4} + s_i''\log\gamma_{\tau_i} + s_i'\log\beta_{\tau_i} + s_i^0\log(1-\beta_{\tau_i}-2\gamma_{\tau_i}), \tag{5.38}$$

onde s_i'' é o número de vezes que se observa uma transversão no ramo de comprimento τ_i (duas bases de categorias químicas distintas nos extremos do ramo), s_i' é o número de vezes que se observa uma transição no mesmo ramo, e s_i^0 é o numero de vezes em que não há mutação aparente neste ramo.

Derivando a expressão (5.38), podemos encontrar o estimador de máxima verossimilhança $\hat{K}_{\mathbf{H}}$ para o parâmetro K, dado pela solução da seguinte equação

$$\begin{aligned} 0 = \frac{\partial\log(L(K|\mathbf{X}))}{\partial K} = \sum_{i=1}^{2N-2} \frac{\tau_i}{(1+K)^2}\Bigg[\frac{s_i'}{\beta_{\tau_i}}\left(e^{-\frac{4\tau_i}{K+2}} + e^{-\frac{2\tau_i(K+1)}{K+2}}\right) \\ + \frac{s_i^0}{1-\beta_{\tau_i}-2\gamma_{\tau_i}}\left(e^{-\frac{4\tau_i}{K+2}} - e^{-\frac{2\tau_i(K+1)}{K+2}}\right) - \frac{s_i''}{\gamma_{\tau_i}}e^{-\frac{4\tau_i}{K+2}}\Bigg]. \end{aligned} \tag{5.39}$$

Para encontrar a variância desse estimador, derivamos duas vezes o logaritmo da função de verossimilhança, obtendo

$$\frac{\partial^2\log(L(K|\mathbf{X}))}{\partial K^2} = \sum_{i=1}^{2N-2} s_i'a(\tau_i) + s_i^0 b(\tau_i) + s_i''c(\tau_i),$$

em que

$$
\begin{aligned}
a(\tau_i) &= \frac{\left(\frac{4\tau_i}{K+2}-2\right)\tau_i e^{-\frac{4\tau_i}{K+2}} - \left(\frac{2\tau_i}{K+2}+2\right)\tau_i e^{-\frac{2\tau_i(K+1)}{K+2}}}{(K+2)^3\beta_{\tau_i}} - \frac{\left(\tau_i e^{-\frac{4\tau_i}{K+2}} + \tau_i e^{-\frac{2\tau_i(K+1)}{K+2}}\right)^2}{(K+2)^4\beta_{\tau_i}^2} \\
b(\tau_i) &= \frac{\left(\frac{4\tau_i}{K+2}-2\right)\tau_i e^{-\frac{4\tau_i}{K+2}} + \left(\frac{2\tau_i}{K+2}+2\right)\tau e^{-\frac{2\tau_i(K+1)}{K+2}}}{(K+2)^3\beta_{\tau_i}} - \frac{\left(\tau_i e^{-\frac{4\tau_i}{K+2}} - \tau_i e^{-\frac{2\tau_i(K+1)}{K+2}}\right)^2}{(K+2)^4\beta_{\tau_i}^2} \\
c(\tau_i) &= -\frac{\left(\frac{4\tau_i}{K+2}-2\right)\tau_i e^{-\frac{4\tau_i}{K+2}}}{(K+2)^3\gamma_{\tau_i}} - \frac{\left(\tau_i e^{-\frac{4\tau_i}{K+2}}\right)^2}{(K+2)^4\gamma_{\tau_i}^2},
\end{aligned} \tag{5.40}
$$

onde β_{τ_i} e γ_{τ_i} são dados por (5.35) e (5.36), respectivamente, para todo $i \in E$.

Lembramos que, sob a restrição $1 = \beta + 2\gamma$ temos $\beta = \frac{K}{K+2}$, e $\gamma = \frac{1}{K+2}$. Assim, as esperanças das estatísticas s_i'', s_i' e s_i^0 são dadas por

$$
\begin{aligned}
\mathbb{E}(s_i'') &= S2\gamma\tau_i = \frac{S\tau_i}{K+2} \\
\mathbb{E}(s_i') &= S\beta\tau_i = \frac{SK\tau_i}{K+2} \\
\mathbb{E}(s_i^0) &= \mathbb{E}(S - s_i'' - s_i') = S(1-\tau_i).
\end{aligned}
$$

Dessa forma, obtemos o limite inferior para a variância de $\hat{K}_{\mathbf{H}}$, dado por

$$
\frac{1}{\mathbb{E}\left(-\frac{\partial^2 \log(L(K|\mathbf{X}))}{\partial K^2}\right)} = \left(-\sum_{i=1}^{2N-2}\left[\frac{S\tau_i}{K+2}a(\tau_i) + S(1-\tau_i)b(\tau_i) + \frac{SK\tau_i}{K+2}c(\tau_i)\right]\right)^{-1},
$$

onde $a(\tau_i)$, $b(\tau_i)$ e $c(\tau_i)$ são dados em (5.40). □

Cota Inferior para as Variâncias dos Estimadores de Máxima Verossimilhança para os Parâmetros β e γ do Modelo K80

Considere as mesmas hipóteses **H** utilizadas nos casos anteriores, dadas por (5.23).

Analisamos agora uma cota inferior para a variância dos parâmetros do modelo $K80$ quando não estão sujeitos à restrição de que a taxa de mutação geral do processo seja 1. Nesse caso, o processo possui dois parâmetros. Escolhemos a parametrização desse modelo apresentada no Capítulo 1, em função dos parâmetros γ e β.

Lema 5.10. *Considere o modelo K80 cuja matriz* $\mathbf{Q}_{\text{K80}}$ *de taxas infinitesimais é dada por (2.14). Suponha, ainda, que a hipótese* $\mathbf{H}$*, dada em (5.23), está satisfeita. Então, os estimadores de máxima verossimilhança* $\hat{\beta}_{\mathbf{H}}$ *e* $\hat{\gamma}_{\mathbf{H}}$ *dos parâmetros* β *e* γ*, respectivamente, são dados pela solução do sistema de equações*

$$\begin{aligned}
0 = \frac{\partial \log(L(\boldsymbol{\theta}|\mathbf{X}))}{\partial \beta} &= \sum_{i=1}^{2N-2} \frac{4\tau_i s_i' e^{-2(\beta+\gamma)\tau_i}}{(1+e^{-4\gamma\tau_i}-2e^{-2(\gamma+\beta)\tau_i})} - \frac{4\tau_i s_i^0 e^{-2(\beta+\gamma)\tau_i}}{(1+e^{-4\gamma\tau_i}+2e^{-2(\gamma+\beta)\tau_i})} \\
0 = \frac{\partial \log(L(\boldsymbol{\theta}|\mathbf{X}))}{\partial \gamma} &= \sum_{i=1}^{2N-2} \frac{4\tau_i s_i' (e^{-2(\beta+\gamma)\tau_i}-e^{-4\gamma\tau})}{(1+e^{-4\gamma\tau_i}-2e^{-2(\gamma+\beta)\tau_i})} - \frac{4\tau_i s_i^0 (e^{-2(\beta+\gamma)\tau_i}+e^{-4\gamma\tau})}{(1+e^{-4\gamma\tau_i}+2e^{-2(\gamma+\beta)\tau_i})} \\
&+ \frac{4 s_i'' \tau e^{-4\gamma\tau}}{(1-e^{-4\gamma\tau})}. \qquad (5.41)
\end{aligned}$$

onde s_i'' *é o número de vezes que se observa uma transversão no ramo de comprimento* τ_i*,* s_i' *é o número de vezes que se observa uma transição no mesmo ramo, e* s_i^0 *é o numero de vezes em que não há mutação aparente neste ramo. Uma cota inferior para a variância de* $\hat{\beta}_{\mathbf{H}}$ *é dada por*

$$\begin{aligned}
\text{Var}(\hat{\beta}_{\mathbf{H}}) &= \frac{-2S}{\det(\mathbf{I}(\beta,\gamma))} \left(\sum_{i=1}^{2N-2} \tau_i^3 \gamma_{\tau_i} e^{-2(\beta+\gamma)\tau_i} \left[\frac{(1-\beta-2\gamma)}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{\beta}{\beta_{\tau_i}^2} \right] - S\tau_i^3 e^{-4\gamma\tau_i} \right. \\
&\left. \times \left[\frac{(1-\beta-2\gamma)(1+e^{-2(\beta+\gamma)\tau_i})}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} + \frac{\beta(1-e^{-2(\beta+\gamma)\tau_i})}{\beta_{\tau_i}^2} + \frac{\gamma}{\gamma_{\tau_i}^2} \right] \right), \qquad (5.42)
\end{aligned}$$

em que β_{τ_i} *e* γ_{τ_i} *são as probabilidades de transição do modelo K80, dadas por (2.16), e* $\det(\mathbf{I}(\beta,\gamma))$ *é o determinante da matriz de informação de Fisher, dado por*

$$\det(\mathbf{I}(\beta,\gamma)) = \mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial \beta^2}\right) \mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial \gamma^2}\right) - \left(\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial \beta \partial \gamma}\right)\right)^2.$$

De forma análoga, uma cota inferior para a variância de $\hat{\gamma}_{\mathbf{H}}$ *e dada por*

$$\text{Var}(\hat{\gamma}_{\mathbf{H}}) = \frac{-S}{\det(\mathbf{I}(\beta,\gamma))} \sum_{i=1}^{2N-2} \tau_i^3 \left(\frac{1}{2} + \frac{e^{-4\gamma\tau_i}}{2}\right) \left[\frac{1-\beta-2\gamma}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{\beta}{\beta_{\tau_i}^2} \right]. \qquad (5.43)$$

Alem disso, a covariância assintótica entre esses dois estimadores é dada por

$$\mathrm{Cov}(\hat{\beta}_{\mathbf{H}}, \hat{\gamma}_{\mathbf{H}}) = \frac{2S}{\det(\mathbf{I}(\beta,\gamma))} \sum_{i=1}^{2N-2} \tau_i^3 \gamma_{\tau_i} e^{-2(\beta+\gamma)\tau_i} \left[\frac{(1-\beta-2\gamma)}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{\beta}{\beta_{\tau_i}^2} \right]. \tag{5.44}$$

Demonstração: De acordo com a parametrização do modelo K80 apresentada na Seção 2.3, temos que

$$\begin{aligned}
\beta_\tau &= \frac{1}{4}\left(1 + e^{-4\gamma\tau} - 2e^{-2(\gamma+\beta)\tau}\right) && (5.45)\\
\gamma_\tau &= \frac{1}{4}\left(1 - e^{-4\gamma\tau}\right) && (5.46)\\
\mathbb{P}(i|i,\tau) &= 1 - \beta_\tau - 2\gamma_\tau = \frac{1}{4}\left(1 + e^{-4\gamma\tau} + 2e^{-2(\gamma+\beta)\tau}\right). && (5.47)
\end{aligned}$$

O logarítmo da função de verossimilhança desse modelo está dado na expressão (5.38). Assim, derivando-o em relação aos parâmetros β e γ, obtemos o seguinte sistema

$$\begin{aligned}
0 = \frac{\partial \log(L(\boldsymbol{\theta}|\mathbf{X}))}{\partial\beta} &= \sum_{i=1}^{2N-2} \frac{4\tau_i s_i' e^{-2(\beta+\gamma)\tau_i}}{(1 + e^{-4\gamma\tau_i} - 2e^{-2(\gamma+\beta)\tau_i})} - \frac{4\tau_i s_i^0 e^{-2(\beta+\gamma)\tau_i}}{(1 + e^{-4\gamma\tau_i} + 2e^{-2(\gamma+\beta)\tau_i})}\\
0 = \frac{\partial \log(L(\boldsymbol{\theta}|\mathbf{X}))}{\partial\gamma} &= \sum_{i=1}^{2N-2} \frac{4\tau_i s_i' (e^{-2(\beta+\gamma)\tau_i} - e^{-4\gamma\tau})}{(1 + e^{-4\gamma\tau_i} - 2e^{-2(\gamma+\beta)\tau_i})} - \frac{4\tau_i s_i^0 (e^{-2(\beta+\gamma)\tau_i} + e^{-4\gamma\tau})}{(1 + e^{-4\gamma\tau_i} + 2e^{-2(\gamma+\beta)\tau_i})}\\
&+ \frac{4 s_i'' \tau e^{-4\gamma\tau}}{(1 - e^{-4\gamma\tau})}. && (5.48)
\end{aligned}$$

A solução do sistema (5.48) fornece os estimadores de máxima verossimilhança dos parâmetros β e γ sujeitos às hipóteses de que F, $\bar{\tau}$ e $(\mathbf{X}^1, \cdots, \mathbf{X}^{2N-1})$ são conhecidos.

Como estamos tratando com um modelo a dois parâmetros, necessitamos da versão matricial do limite inferior de Cramér-Rao. Assim,

$$\boldsymbol{\Sigma}_{(\hat{\beta}_{\mathbf{H}}, \hat{\gamma}_{\mathbf{H}})} = (\mathbf{I}(\beta,\gamma))^{-1}, \tag{5.49}$$

em que $\boldsymbol{\Sigma}_{(\hat{\beta}_{\mathbf{H}}, \hat{\gamma}_{\mathbf{H}})}$ é a matriz de variâncias-covariâncias dos estimadores de máxima verossimilhança $\hat{\beta}_{\mathbf{H}}$ e $\hat{\gamma}_{\mathbf{H}}$, e $[\mathbf{I}(\beta,\gamma)]^{-1}$ é a inversa da matriz de informação de Fisher, dada por

$$\mathbf{I}(\beta,\gamma) = \left(-\mathbb{E}\left[\frac{\partial^2 \log(L)}{\partial\theta_l \partial\theta_j} \right] \right)_{\theta_l, \theta_j \in \{\beta,\gamma\}} = \begin{pmatrix} -\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta^2}\right) & -\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta\partial\gamma}\right) \\ -\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta\partial\gamma}\right) & -\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\gamma^2}\right) \end{pmatrix}. \tag{5.50}$$

Note que as derivadas de segunda ordem de $\log(L(\boldsymbol{\theta}|\mathbf{X})$ são dadas por

$$\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}))}{\partial^2 \beta} = \sum_{i=1}^{2N-2} \tau_i^2 \left(\frac{1}{2} + \frac{e^{-4\gamma\tau_i}}{2}\right) \left[\frac{s_i^0}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{s_i'}{\beta_{\tau_i}^2}\right] \tag{5.51}$$

$$\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}))}{\partial\beta\partial\gamma} = \sum_{i=1}^{2N-2} 2\tau_i^2 \gamma_{\tau_i} e^{-2(\beta+\gamma)\tau_i} \left[\frac{s_i^0}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{s_i'}{\beta_{\tau_i}^2}\right] \tag{5.52}$$

$$\begin{aligned}\frac{\partial^2 \log(L(\boldsymbol{\theta}|\mathbf{X}))}{\partial\gamma^2} &= \sum_{i=1}^{2N-2} 2\tau_i^2 \gamma_{\tau_i} e^{-2(\beta+\gamma)\tau_i} \left[\frac{s_i^0}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{s_i'}{\beta_{\tau_i}^2}\right] - \tau_i^2 e^{-4\gamma\tau_i} \\ &\quad \times \left[\frac{s_i^0(1+e^{-2(\beta+\gamma)\tau_i})}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} + \frac{s_i'(1-e^{-2(\beta+\gamma)\tau_i})}{\beta_{\tau_i}^2} + \frac{s_i''}{\gamma_{\tau_i}^2}\right].\end{aligned} \tag{5.53}$$

Lembrando que temos $\mathbb{E}(s_i'') = S2\gamma\tau_i$, $\mathbb{E}(s_i') = S\beta\tau_i$, e $\mathbb{E}(s_i^0) = S - \mathbb{E}(s_i'') - \mathbb{E}(s_i') = (1-\beta-2\gamma)S\tau_i$, podemos encontrar a matriz de variâncias-covariâncias $\boldsymbol{\Sigma}_{(\hat{\beta}_{\mathbf{H}},\hat{\gamma}_{\mathbf{H}})}$.

Com isso em vista, calculamos a variância assintótica do estimador $\hat{\gamma}_{\mathbf{H}}$ como

$$\begin{aligned}\mathrm{Var}(\hat{\gamma}_{\mathbf{H}}) &= -\frac{\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta^2}\right)}{\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta^2}\right)\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\gamma^2}\right) - \mathbb{E}\left[\left(\frac{\partial^2 \log(L)}{\partial\beta\partial\gamma}\right)\right]^2} \\ &= \frac{-S}{\det(\mathbf{I}(\beta,\gamma))} \sum_{i=1}^{2N-2} \tau_i^3 \left(\frac{1}{2} + \frac{e^{-4\gamma\tau_i}}{2}\right) \left[\frac{1-\beta-2\gamma}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{\beta}{\beta_{\tau_i}^2}\right].\end{aligned}$$

A variância de $\hat{\beta}_{\mathbf{H}}$ é obtida de forma análoga, de modo que temos

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_{\mathbf{H}}) &= -\frac{\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\gamma^2}\right)}{\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta^2}\right)\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\gamma^2}\right) - \mathbb{E}\left[\left(\frac{\partial^2 \log(L)}{\partial\beta\partial\gamma}\right)\right]^2} \\ &= \frac{-S}{\det(\mathbf{I}(\beta,\gamma))} \sum_{i=1}^{2N-2} 2\tau_i^2 \gamma_{\tau_i} e^{-2(\beta+\gamma)\tau_i} \left[\frac{(1-\beta-2\gamma)}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{\beta}{\beta_{\tau_i}^2}\right] - \tau_i^2 e^{-4\gamma\tau_i} \\ &\quad \times \left[\frac{(1-\beta-2\gamma)(1+e^{-2(\beta+\gamma)\tau_i})}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} + \frac{\beta(1-e^{-2(\beta+\gamma)\tau_i})}{\beta_{\tau_i}^2} + \frac{\gamma}{\gamma_{\tau_i}^2}\right].\end{aligned} \tag{5.54}$$

Da mesma forma, a covariância é dada por

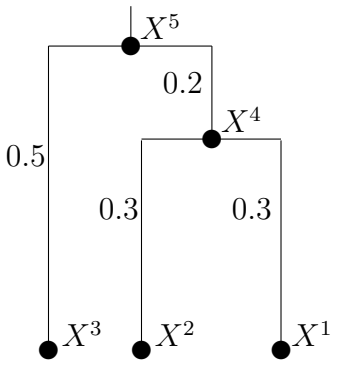


Figura 5.3: Filogenia da Seção 5.2.1.

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_{\mathbf{H}}, \hat{\gamma}_{\mathbf{H}}) &= \frac{\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta\partial\gamma}\right)}{\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta^2}\right)\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\gamma^2}\right) - \mathbb{E}\left[\left(\frac{\partial^2 \log(L)}{\partial\beta\partial\gamma}\right)\right]^2} \\ &= \frac{-S}{\det(\mathbf{I}(\beta,\gamma))}\sum_{i=1}^{2N-2} 2\tau_i^2\gamma_{\tau_i}e^{-2(\beta+\gamma)\tau_i}\left[\frac{(1-\beta-2\gamma)}{(1-\beta_{\tau_i}-2\gamma_{\tau_i})^2} - \frac{\beta}{\beta_{\tau_i}^2}\right],\end{aligned}$$

onde

$$\det(\mathbf{I}(\beta,\gamma)) = \mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\beta^2}\right)\mathbb{E}\left(\frac{\partial^2 \log(L)}{\partial\gamma^2}\right) - \mathbb{E}\left[\left(\frac{\partial^2 \log(L)}{\partial\beta\partial\gamma}\right)\right]^2.$$

□

5.2.1 Avaliação das Variâncias de $\hat{\alpha}$ e $\hat{\alpha}_{\mathbf{H}}$

Para avaliar a relação entre a variância do estimador de máxima verossimilhança e a cota inferior para a variância apresentada na Seção 5.2, realizamos uma aplicação computacional. Utilizando como exemplo o parâmetro α do modelo JC69, implementamos uma rotina no pacote "R-project" para estimar as variâncias dos estimadores $\hat{\alpha}$, $\hat{\alpha}_{\mathbf{H}}$, e as comparamos com a cota inferior apresentada no Lema 5.8.

Neste estudo, utilizamos a filogenia apresentada na Figura 5.3 para simular conjuntos de seqüências com $\alpha \in \{0.2, 0.3, 0.4, 0.6, 0.8, 1\}$. Para cada valor de α, simulamos 40 seqüências com S variando entre 100 e 2000. Para cada seqüência simulada, assumindo a

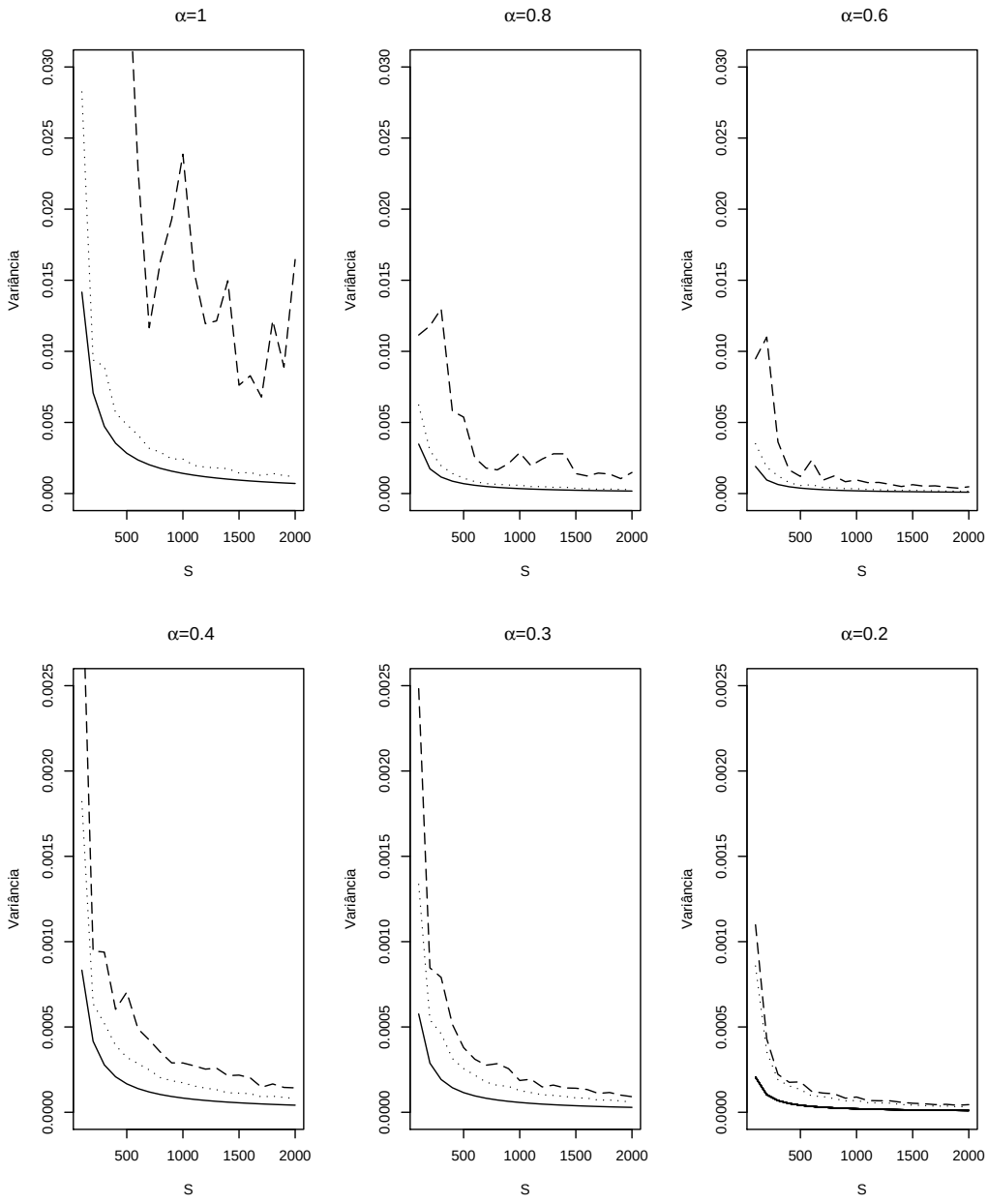


Figura 5.4: Comparação entre as Variâncias de $\hat{\alpha}$ (linha tracejada), $\hat{\alpha}_{\mathbf{H}}$ (linha pontilhada), e a Cota Inferior para a Variância de $\hat{\alpha}_{\mathbf{H}}$ (linha cheia).

topologia da filogenia e os tempos $\bar{\tau}$ conhecidos, estimamos as variâncias dos estimadores $\hat{\alpha}$ e $\hat{\alpha}_{\mathbf{H}}$ utilizando a informação observada de Fisher.

A Figura 5.4 apresenta a comparação entre as variâncias de $\hat{\alpha}$, $\hat{\alpha}_{\mathbf{H}}$, e a cota inferior para a variância de $\hat{\alpha}_{\mathbf{H}}$ para todos os valores de α analisados. Note que a escala dos gráficos da segunda linha da Figura 5.4 é diferente daquela da primeira linha para permitir melhor visualização da relação entre os três estimadores da variância.

Note que, como as seqüências foram simuladas, dispomos da informação das seqüências nos nós ancestrais $X_{N+1}, \cdots, X_{2N-1}$. Essa informação permite a utilização do estimador $\hat{\alpha}_{\mathbf{H}}$. Em casos reais, não dispomos de tais seqüências, de forma que a utilização de $\hat{\alpha}_H$ para estimar α não seria possível.

Observe que a única diferença entre os estimadores $\hat{\alpha}$ e $\hat{\alpha}_{\mathbf{H}}$ é a hipótese adicional de que as seqüências ancestrais $X_{N+1}, \cdots, X_{2N-1}$ são conhecidas. No caso de $\hat{\alpha}$, contorna-se o fato de não conhecer essas seqüências com a soma de todas as possíveis combinações de bases nos nós ancestrais, como apresentado na Seção 3.1. Assim, se $\hat{\alpha}$ é o estimador semelhante a $\hat{\alpha}_{\mathbf{H}}$ na ausência da informação de $X_{N+1}, \cdots, X_{2N-1}$, espera-se que sua variância seja maior do que a variância de $\hat{\alpha}_{\mathbf{H}}$. Isso é de fato verificado nas simulações, como pode ser visto na Figura 5.4. Desta forma, a cota inferior para a variância de $\hat{\alpha}_{\mathbf{H}}$ é também uma cota inferior para a variância de $\hat{\alpha}$.

A Figura 5.4 confirma que $\mathrm{Var}(\hat{\alpha}_{\mathbf{H}})$, dado na expressão (5.30), serve como uma cota inferior para as variâncias de $\hat{\alpha}$ e $\hat{\alpha}_{\mathbf{H}}$, para todos os valores de α e todos os comprimentos de seqüência testados. Além disso, esta figura evidencia que, com o aumento de S, a variância de $\hat{\alpha}_{\mathbf{H}}$ converge para a cota inferior (esse resultado é esperado, uma vez que a cota foi calculada com base na distribuição assintótica do estimador de máxima verossimilhança). A variância de $\hat{\alpha}$ também se aproxima da cota inferior, sempre que S aumenta.

Além disso, notamos que com α pequeno as variâncias de $\hat{\alpha}$ e $\hat{\alpha}_{\mathbf{H}}$ são próximas. Este fato pode ser interpretado da seguinte forma: com α pequeno, o número de mutações que ocorrem na amostra também é pequeno. Com poucas mutações, a informação adicional contida nas seqüências dos nós internos também é pequena. Desta forma, o estimador $\hat{\alpha}_{\mathbf{H}}$, que utiliza essa informação, não apresenta uma grande vantagem em relação à $\hat{\alpha}$, o que é evidenciado pela semelhança das suas respectivas variâncias.

5.3 Correção de Vício por Jackknife

Conforme o Teorema 5.1, sob certas condições de regularidade, estimadores de máxima verossimilhança são assintoticamente não viciados. Entretanto, dispomos apenas de resultados assintóticos para a esperança destes estimadores; e, de fato, freqüentemente, eles são viciados. Para determinar o vício de um estimador de máxima verossimilhança, é preciso calcular sua esperança. Entretanto, no caso dos modelos de substituição de bases, quando se consideram diversas seqüências e a filogenia que as relaciona, tal esperança é

difícil de ser determinada.

Uma estimativa não paramétrica para o vício de um estimador pode ser obtida através do Jackknife.

Lema 5.11. *Considere uma amostra de variáveis aleatórias i.i.d.* $x_1, \cdots, x_n$*, e o estimador* $\hat{\theta}$ *para o parâmetro* θ *da população. Seja* $\hat{\theta}_{(-i)}$ *o estimador* $\hat{\theta}$ *calculado sobre a amostra* $x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_n$*, quando o elemento* x_i *é retirado da amostra. Então, a estimativa do Jackknife para o vício de* $\hat{\theta}$ *é dada por*

$$\widehat{\text{vício}(\hat{\theta})} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}) \tag{5.55}$$

onde $\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)}$*. Assim, temos que o estimador corrigido por Jackknife é dado por*

$$\hat{\theta}_J = \hat{\theta} - \widehat{\text{vício}(\hat{\theta})} = n\hat{\theta} - (n-1)\hat{\theta}_{(\cdot)} \tag{5.56}$$

Demonstração: Ver Efron e Tibshirani (1993).

No caso dos modelos de substituição de bases, esta técnica pode ser aplicada tanto para os parâmetros dos modelos quanto para os comprimentos dos ramos das filogenias. Para tanto, removemos o primeiro sítio da amostra, ficando com seqüências de comprimento $S-1$, e então, estimamos os parâmetros do modelo por máxima verossimilhança, para obter $\hat{\boldsymbol{\theta}}_{(-1)}$, em que $\hat{\boldsymbol{\theta}}$ contém os comprimentos de ramos $\bar{\tau}$ e os parâmetros do modelo de substituição de bases. Esse procedimento é repetido para os S sítios da seqüência, e então $\hat{\boldsymbol{\theta}}_J$ é calculado conforme a expressão (5.56).

Note que o estimador $\hat{\boldsymbol{\theta}}_J$ tem um custo computacional aproximadamente $S+1$ vezes maior do que $\hat{\boldsymbol{\theta}}$. Assim, a decisão de utilizá-lo deve ser feita comparando o custo computacional e o ganho em precisão da estimativa dado pelo Jackknife.

A correção do vício por Jackknife pode ser utilizada para todos os modelos apresentados nesse trabalho que possuem a propriedade de que os sítios são i.i.d.. Assim, o único modelo que apresentamos para o qual esta correção não pode ser aplicada é o modelo que designa uma cadeia de Markov oculta para as taxas de mutação dos sítios, dado na Seção 4.1.4. Note que, para utilizar a correção do vício para os modelos que assumem dependência entre os sítios, apresentados na Seção 4.2, não devemos considerar os sítios isolados, mas sim, o par de sítios, pois estes são i.i.d..

Na Seção 6.3 são apresentadas simulações computacionais que visam avaliar o desempenho da correção do vício por Jackknife para modelos de substituição de bases.

5.3.1 Nova Versão do Algoritmo para a Obtenção do Poder do Teste

Utilizando a correção do vício dos estimadores por Jackknife, podemos obter uma nova versão, mais precisa, do algoritmo apresentado na Seção 3.2.3 para obter o poder do teste da razão de verossimilhança que compara dois modelos de substituição de bases. Para tanto, em vez de utilizar os estimadores de máxima verossimilhança para gerar as replicações das seqüências, utilizamos os estimadores com a correção do vício por Jackknife. Assim, temos o *algoritmo para o cálculo do poder do teste* apresentado a seguir.

1. Utilizar os dados observados para estimar, pelo método da máxima verossimilhança, os parâmetros do modelo de substituição de bases, a árvore filogenética F e os tamanhos dos ramos $\bar{\tau}$, utilizando o modelo M_0 sob H_0.

2. Aplicar o Jackknife aos dados, e realizar a correção do vício por Jackknife aos parâmetros estimados no passo **1**.

3. Gerar, aleatoriamente, uma seqüência para a raiz da filogenia com a distribuição de bases $\mathbf{p}_0$ designada pelo modelo M_0 sob H_0.

4. Utilizando os parâmetros estimados no passo **2**, simular as seqüências filhas da árvore. Para isso, inicia-se pela raiz da filogenia, e geram-se seqüências para os nós adjacentes partindo da seqüência da raiz e utilizando as probabilidades de mutação $\mathbb{P}(i \mid j, \tau)$. Essas probabilidades são determinadas pelos parâmetros estimados para o modelo de substituição de bases e os tempos $\bar{\tau}$ estimados. Em seguida, repete-se o procedimento para os nós descendentes daqueles recém gerados, seguindo o desenho da filogenia estimada até que se obtenham seqüências para os N nós externos.

5. Calcular a estatística do teste $-2\Delta(\cdot)$, dada pela expressão (3.13), para os dados gerados no passo **4**.

6. Repetir os passos **3**, **4** e **5**, m vezes.

7. Construir um histograma com os valores de $-2\Delta(\cdot)$ simulados para descobrir qual o valor crítico C_λ.

8. Utilizando os dados observados, estimar, pelo método da máxima verossimilhança os parâmetros do modelo de substituição de bases, a árvore filogenética F e os tamanhos dos ramos $\bar{\tau}$ utilizando o modelo M_1 sob H_1.

9. Aplicar o Jackknife aos dados, e realizar a correção do vício por Jackknife aos parâmetros estimados no passo **8**.

10. Gerar, aleatoriamente, uma seqüência para a raiz da filogenia com a distribuição de bases $\mathbf{p}_0$ designada pelo modelo M_1 sob H_1.

11. Utilizando os parâmetros estimados no passo **9**, simular as seqüências filhas da árvore como feito no passo **4**.

12. Calcular a estatística do teste $-2\Delta(\cdot)$, dada pela expressão (3.13), para os dados gerados no passo **11**.

13. Repetir os passos **10**, **11** e **12**, m vezes.

14. Construir um histograma com os valores de $-2\Delta(\cdot)$ simulados a partir de H_1 e verificar a posição do valor crítico C_λ.

15. O poder do teste é dado pela proporção de elementos do passo **13** maiores do que C_λ.

5.4 Estimação do Parâmetro a no Modelo com Distribuição Gama para as Taxas de Mutação

No modelo que designa uma distribuição gama para as taxas de mutação nos sítios da seqüência (ver Seção 4.1.2), o parâmetro a está inversamente relacionado à variabilidade entre os sítios. A distribuição gama com $a > 1$ assume forma de um sino, assim temos o caso em que a maioria dos sítios tem taxas de mutação intermediárias, e alguns poucos tem taxas muito altas ou muito baixas. No limite, quando $a = \infty$, temos o caso em que todos os sítios evoluem à mesma taxa de mutação. Para valores de $a < 1$, a distribuição gama assume forma de L, neste caso, a maioria dos sítios teria taxa de mutação pequena, e alguns poucos sítios teriam taxa de mutação muito grande (os chamados "hot spots" mutacionais).

Deste modo, observamos que o parâmetro a tem informações preciosas sobre as características mutacionais do gene. Por esta razão, estamos interessados em estimá-lo. Yang e Kumar (1996) sugerem um método baseado no princípio da parcimônia para estimar a.

Assumindo que cada sítio da seqüência evolui conforme o modelo JC69 (ver Seção 2.3), temos que

$$\mathbb{P}(i|j,\tau) = \frac{1}{4}(1 - e^{-4\alpha\tau})$$
$$\mathbb{P}(i|i,\tau) = \frac{1}{4}(1 + 3e^{-4\alpha\tau}),$$

onde $i, j \in E$. Além disso, a probabilidade condicional de se observar um sítio no qual ocorra um padrão específico[1] em que se observa k mutações, dado a taxa de mutação μ é

[1]Para um exemplo dos padrões de bases da amostra $\mathbf{X}$ e do cálculo de suas probabilidades, ver Apêndice A.

dada por

$$\mathbb{P}(k|\mu) = \mathbb{P}(i|i,\mu\tau)^{(2N-2-k)}\mathbb{P}(i|j,\mu\tau)^{k},$$

onde $2N-2$ representa o número de ramos da filogenia sem raiz. Assim, como μ tem distribuição gama, temos que

$$\mathbb{P}(k) = \int_0^\infty \mathbb{P}(k|\mu) f_X(\mu),$$

onde $f_X(\mu)$ representa a função densidade da distribuição gama, dada na Observação 4.1. Desta forma, temos que a função de log-verossimilhança, que é calculada como o logaritmo da probabilidade de encontrar N_k sítios pertencentes ao padrão de bases k, com k pertencente ao conjunto de possíveis padrões de bases, é dado por

$$\begin{aligned}\log(L) &= \sum_k N_k \log(\mathbb{P}(k)) \\ &= \sum_k N_k \left(\int_0^\infty \left(\frac{1}{4} - \frac{e^{-4\alpha\tau}}{4}\right)^{(2N-2-k)} f_X(\mu) \left(\frac{1}{4} + \frac{3e^{-4\alpha\tau}}{4}\right)^{k} f_X(\mu) d\mu \right).\end{aligned} \tag{5.57}$$

Para estimar a otimizamos numericamente o logaritmo da função de verossimilhança, dada na expressão (5.57).

Em Yang e Kumar (1996) o tempo τ que aparece na expressão (5.57) é obtido como a razão entre o número médio de mutações por sítio e o número de ramos da filogenia $2N-2$. Note, entretanto, que não dispomos do exato número de mutações ocorrido em cada sítio, mas apenas das seqüências dos nós externos da filogenia. Assim, o número de mutações ocorrido em cada sítio deve ser estimado por parcimônia, como o menor número de mutações necessário para gerar o padrão de bases encontrado no sítio. Esse número mínimo de mutações pode ser obtido pelo algoritmo de Hartigan (1973).

Note que o algoritmo da parcimônia claramente sub-estima o número de mutações dos sítios, e consequentemente o número de mutações médio nos ramos τ, uma vez que não considera a possibilidade de mutações silenciosas. Yang e Kumar (1996) comentam que a também é sub-estimado por este método, ainda que menos do que outros métodos de estimação de a. Eles ainda destacam que a utilização de valores maiores para τ não parece melhorar as estimativas de a.

Observamos que a ainda pode ser estimado por máxima verossimilhança, utilizando a expressão (4.6) para a função de verossimilhança, e otimização numérica, conforme feito para todos os parâmetros tratados até aqui neste trabalho. Destacamos, entretanto, que

o custo computacional deste procedimento é muito maior do que quando utilizamos a expressão (5.57).

Capítulo 6

Simulações

Neste capítulo são apresentadas simulações que visam avaliar o comportamento de diversas estatísticas estudadas nesse trabalho.

Inicialmente estudamos o comportamento dos testes da razão de verossimilhança que comparam dois modelos de substituição de bases. Tal estudo é feito através da estatística do teste $-2\Delta(\mathbf{X})$, dada pela expressão (3.13). O efeito de diversos fatores, como o número de seqüências, comprimento das seqüências, características das filogenias e parâmetros são avaliados para os modelos apresentados na Seção 2.3. Além disso, apresentamos um estudo do comportamento da correção do vício dos estimadores dos parâmetros dos modelos pelo método do Jackknife.

6.1 Caso i.i.d. com Taxas de Mutação Constantes

Nessa seção, são analisados, por meio de simulações, os testes da razão de verossimilhança que comparam os modelos de substituição de bases apresentados na Seção 2.3. Estes modelos são i.i.d. e apresentam taxas de mutação constantes entre os sítios. Foram avaliadas as funções de distribuição da estatística do teste $-2\Delta(\mathbf{X})$, dada pela expressão (3.13), tanto sob H_0 quanto sob H_1. O objetivo desta análise é avaliar o efeito da variação do comprimento S da seqüência, do número N de espécies e dos parâmetros dos modelos na distribuição da estatística do teste e no seu poder.

A metodologia utilizada é aquela apresentada no Capítulo 3 para a obtenção do valor crítico e do poder do teste por meio de simulações de Monte Carlo. Foram utilizadas adaptações de rotinas do pacote PAML, disponibilizado *on-line* em *http:abacus.gene.ucl.ac.uk/software/paml.html*.

Consideramos os testes de hipóteses que comparam os modelos JC69 $\times$ K80, JC69 $\times$ F81, K80 $\times$ HKY85, F81 $\times$ HKY85 e HKY85 $\times$ GTR, para três filogenias diferentes. As filogenias, estimadas originalmente de seqüências reais, foram retiradas dos exemplos que acompanham o pacote PAML e estão apresentadas na Figura 6.1. A

Tabela 6.1: Parâmetros Utilizados na Simulação.

Número de Seqüências	$N \in \{4, 13, 19\}$
Comprimento da Seqüência	$S \in \{100, 500, 1000, 2000\}$
Número de Repetições	$Re = 1000$

Tabela 6.2: Parâmetros dos Modelos Utilizados na Simulação.

Modelo	Parâmetros
JC69	$K = 1$ e $\mathbf{p}_0 = (0.25, 0.25, 0.25, 0.25)$
K80	$K \in \{3, 5\}$ e $\mathbf{p}_0 = (0.25, 0.25, 0.25, 0.25)$
F81	$K = 1$ e $\mathbf{p}_0 = (0.2, 0.3, 0.3, 0.2)$
HKY85	$K = 3$ e $\mathbf{p}_0 = (0.2, 0.3, 0.3, 0.2)$
GTR	$\alpha = 1$, $\beta = 0.2$, $\gamma = 0.3$, $\delta = 0.2$, $\epsilon = 0.1$, $\eta = 2$ e $\mathbf{p}_0 = (0.2, 0.3, 0.3, 0.2)$

árvore 1 apresenta $N = 4$, a árvore 2 apresenta $N = 13$, e a árvore 3 tem $N = 19$. Para cada árvore filogenética, foram simuladas seqüências filhas com comprimentos de $S \in \{100, 500, 1000, 2000\}$ segundo cada um dos 5 modelos envolvidos nos testes. Para o modelo K80 foram utilizados valores de $K \in \{3, 5\}$ (taxa de *transições* e *transversões*); para o modelo HKY85, $K = 3$; e para o modelo GTR, conforme a notação da matriz (2.43), foi utilizado $\alpha = 1$, $\beta = 0.2$, $\gamma = 0.3$, $\delta = 0.2$, $\epsilon = 0.1$ e $\eta = 2$. Para os modelos F81, HKY85 e GTR utilizamos $\mathbf{p}_0 = (\pi_A, \pi_G, \pi_C, \pi_T,) = (0.2, 0.3, 0.3, 0.2)$. Para todas as simulações consideramos $Re = 1000$ replicações. Os parâmetros utilizados nas simulações estão sintetizados nas tabelas 6.1 e 6.2.

A escolha dos valores para os parâmetros foi feita com base em resultados de análise de seqüências reais, para que eles façam sentido do ponto de vista biológico. Os parâmetros do modelo GTR ainda foram escolhidos de forma que as seqüências simuladas com esse modelo, quando analisadas sob outros modelos, apresentem estimativas de K próximo a 3 (valor de K utilizado para K80 e HKY85).

Valor Crítico da Função de Distribuição de $-2\Delta(\mathbf{X})$, sob H_0

As Tabelas 6.3 e 6.4 apresentam o valor crítico à 99% obtido para cada um dos testes de hipóteses estudados, nos comprimentos $S \in \{100, 500, 1000, 2000\}$ para todas as árvores, com $\mathbf{p}_0$ estimado, respectivamente, pelos métodos da máxima verossimilhança (denotado por $\hat{\mathbf{p}}_0$) e dos momentos (denotado por $\tilde{\mathbf{p}}_0$). Além disso, para fins de comparação, a última coluna das tabelas apresenta o valor crítico sugerido pela distribuição assintótica χ^2 (ver Teorema 3.1).

Note que, para ambos os testes de hipóteses JC69 $\times$ K80 e F81 $\times$ HKY85, a função de distribuição teórica de $-2\Delta(\mathbf{X})$ é χ^2_1. Observe que, tanto para o teste JC69 $\times$ K80 quanto para F81 $\times$ HKY85, os valores críticos simulados variam em torno do valor crítico

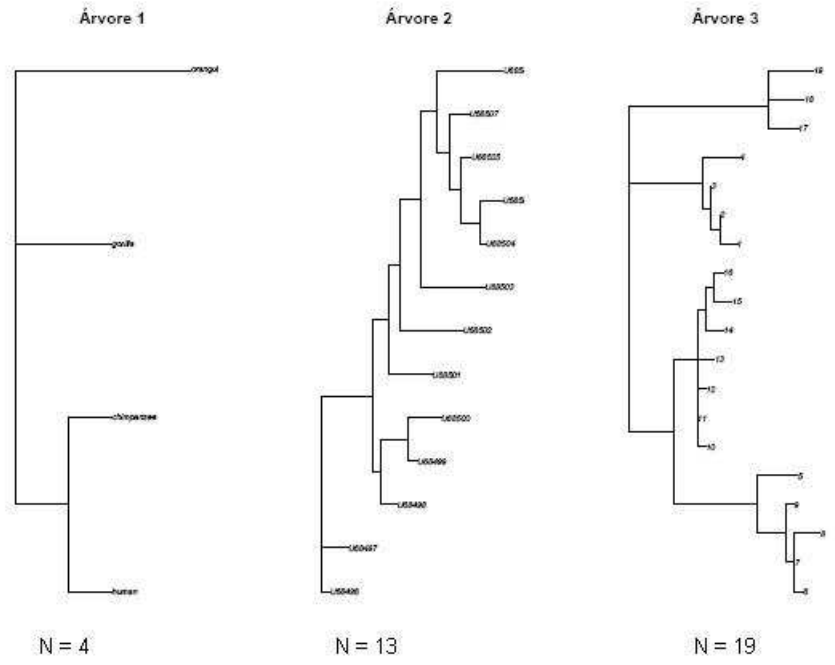


Figura 6.1: Árvores Utilizadas nas Simulações.

teórico.

Já para os testes da razão de verossimilhança que comparam JC69 × F81 e K80 × HKY85 a função de distribuição teórica de $-2\Delta(\mathbf{X})$ é χ^2_3. Os valores que encontramos para ambos os testes, nas simulações em que $\mathbf{p}_0$ é estimado pelo método da máxima verossimilhança, estão próximos do valor crítico teórico.

Finalmente, para o teste de hipóteses HKY85×GTR, a estatística do teste tem distribuição assintótica teórica χ^2_4. De forma geral, pode-se dizer que os valores encontrados na simulação com $\mathbf{p}_0$ estimado pelo método da máxima verossimilhança estão próximos ao valor crítico tabelado.

Nas simulações em que $\mathbf{p}_0$ é estimado pelo método dos momentos, o valor crítico simulado tende a ser inferior ao valor crítico tabelado para o teste de hipóteses que comparam JC69 × F81 e K80 × HKY85. Já para HKY85×GTR, com $\mathbf{p}_0$ estimado pelo método dos momentos, o valor crítico simulado foi sempre maior do que o tabelado, como pode ser visto na Tabela 6.4.

Notamos, através das Tabelas 6.3 e 6.4 que, apesar dos valores críticos simulados estarem próximos dos valores críticos tabelados da distribuição assintótica qui-quadrado, há bastante variabilidade nesses resultados. Desta forma, sugere-se novo estudo com maior número de replicações para verificar se esta variação é realmente apenas devida ao

Tabela 6.3: Comparação entre Valor Crítico Simulado e Teórico Utilizando $\hat{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$	Valor Crítico Teórico
JC69 × K80	1	7.2678	6.5562	6.8426	6.8322	6.64
	2	7.6886	6.5456	6.6568	6.6048	6.64
	3	6.5702	7.0386	5.6932	5.9630	6.64
JC69 × F81	1	11.8492	11.4164	12.6830	11.6228	11.36
	2	11.3966	12.0164	11.0836	12.6642	11.36
	3	11.7898	12.3754	11.4188	11.3166	11.36
K80 × HKY85	1	10.5066	11.5928	12.1312	10.4880	11.36
	2	12.2736	11.7202	10.0436	12.7340	11.36
	3	11.9502	9.9860	12.0482	10.3848	11.36
F81 × HKY85	1	7.6480	6.1212	7.1450	6.6690	6.64
	2	6,7022	6.3468	6.4874	6.2496	6.64
	3	6.1860	6.9846	6.4940	6.0468	6.64
HKY85 × GTR	1	12.3402	13.1654	12.3796	14.0130	13.28
	2	13.5666	13.9564	12.7916	13.5186	13.28
	3	130438	12.6556	12.7004	13.6898	13.28

Nota: Os testes foram realizados ao nível de 99% de confiança.

efeito aleatório.

As Figuras 6.2 a 6.6 apresentam histogramas das distribuições obtidas nas simulações para $-2\Delta(\mathbf{X})$ utilizando o estimador $\hat{\mathbf{p}}_0$ na Árvore 3. As Figuras 6.7 a 6.10 apresentam histogramas das distribuições obtidas utilizando o estimador $\tilde{\mathbf{p}}_0$ para a mesma árvore. Nos histogramas em que aparece a linha vertical tracejada, ela representa o valor crítico à 99% de confiança obtido para aquele teste (tal linha não está presente nos histogramas sob H_0). Já as Figuras 6.14 a 6.16 apresentam os Q-Q plots obtidos para a Árvore 3, em que os quantís da distribuição de $-2\Delta(\mathbf{X})$ sob H_0 são comparados com os quantís da respectiva distribuição χ^2, e os quantís da distribuição de $-2\Delta(\mathbf{X})$ sob H_1 são comparados com aqueles da distribuição normal.

Além disso, foram realizados testes de hipóteses que avaliam a adequação da distribuição obtida pelas simulações para $-2\Delta(\mathbf{X})$ com a distribuição teórica. Foi utilizado o teste qui-quadrado para verificar se $-2\Delta(\mathbf{X})$, sob H_0, é de fato uma qui-quadrado, e o teste de normalidade Shapiro-Wilks para verificar se $-2\Delta(\mathbf{X})$, sob H_1, possui distribuição normal. As decisões desses testes então apresentadas nas Tabelas D.1 a D.4, no Apêndice D.

Observa-se que a distribuição assintótica teórica da estatística do teste sob H_0, quando S cresce, é qui-quadrado. Nas simulações para os testes de hipóteses JC69 × K80, HKY85 × F84 e HKY85 × GTR a função de distribuição de $-2\Delta(\mathbf{X})$ se aproxima de uma qui-

Tabela 6.4: Comparação entre Valor Crítico Simulado e Teórico Utilizando $\tilde{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$	Valor Crítico Teórico
JC69 × F81	1	11.4614	11.2738	12.5556	9.4124	11.36
	2	10.3364	10.1318	9.5160	10.8180	11.36
	3	10.1400	10.5134	10.7418	9.9574	11.36
K80 × HKY85	1	9.5134	10.8692	11.9666	10.3136	11.36
	2	11.2508	10.1412	8.6866	11.0700	11.36
	3	10.9548	9.1172	11.8954	10.0656	11.36
F81 × HKY85	1	7.4444	6.2936	7.2370	6.6206	6.64
	2	7.0422	6.4498	6.8994	6.5502	6.64
	3	6.2488	7.0320	4.5482	4.7254	6.64
HKY85 × GTR	1	14.0524	14.3326	13.8066	15.9540	13.28
	2	17.4364	17.8594	15.8970	16.4914	13.28
	3	15.3268	15.0882	14.5502	14.9148	13.28

Nota: Os testes foram realizados ao nível de 99% de confiança.

quadrado para todos os comprimentos de seqüência e todas as árvores, independente do método de estimação utilizado para $\mathbf{p}_0$. Os testes qui-quadrado confirmam esses resultados, com poucas exceções pontuais, como pode ser visto nas Tabelas D.1 e D.2. Assim, notamos que, para o uso da distribuição assintótica χ^2, seqüências de comprimento 100 já assumem o comportamento assintótico, como pode ser visto nas Figuras 6.2 à 6.16. Além disso, o número de seqüências analisadas aparentemente não afeta a forma dessa distribuição.

Whelan e Goldman (1999) realizaram análises semelhantes, comparando os modelos JC69 × K80 e F81 × HKY85, e obtiveram conclusões semelhantes em relação ao comprimento das seqüências. Também verificaram que a filogenia e o número de seqüências têm pouco efeito na distribuição da estatística do teste da razão de verossimilhança sob H_0, e que as funções de distribuição de ambos os testes são compatíveis com uma χ^2_1.

Nas simulações dos testes de hipóteses JC69 × F81 e K80 × F84 com $\tilde{\mathbf{p}}_0$ notamos que a função de distribuição da estatística do teste não é qui-quadrado, o que pode ser facilmente evidenciado pela presença de valores negativos para $-2\Delta(\mathbf{X})$. Nos testes qui-quadrado, a hipótese nula (de que $-2\Delta(\mathbf{X})$, sob H_0, tem distribuição χ^2_3) foi rejeitada à nível de confiança 99% para todos os testes, como pode ser visto na Tabela D.2. Note que, como esses testes são hierárquicos, isto é $\hat{L}_0(\mathbf{X}) \leq \hat{L}_1(\mathbf{X})$, onde $\hat{L}_0(\mathbf{X})$ é o máximo da função de verossimilhança sob H_0 e $\hat{L}_1(\mathbf{X})$ é o mesmo máximo sob H_1. Dessa forma valores negativos de $-2\Delta(\mathbf{X})$ não são possíveis para o teste da razão de verossimilhança. Eles aparecem nas simulações devido ao fato de que o parâmetro $\mathbf{p}_0$ é estimado pelo método dos momentos e não pelo método da máxima verossimilhança. Desta forma, o teste que está sendo realizado, no caso de JC69 × F81, não é $H_0 : \mathbf{p}_0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4},)$

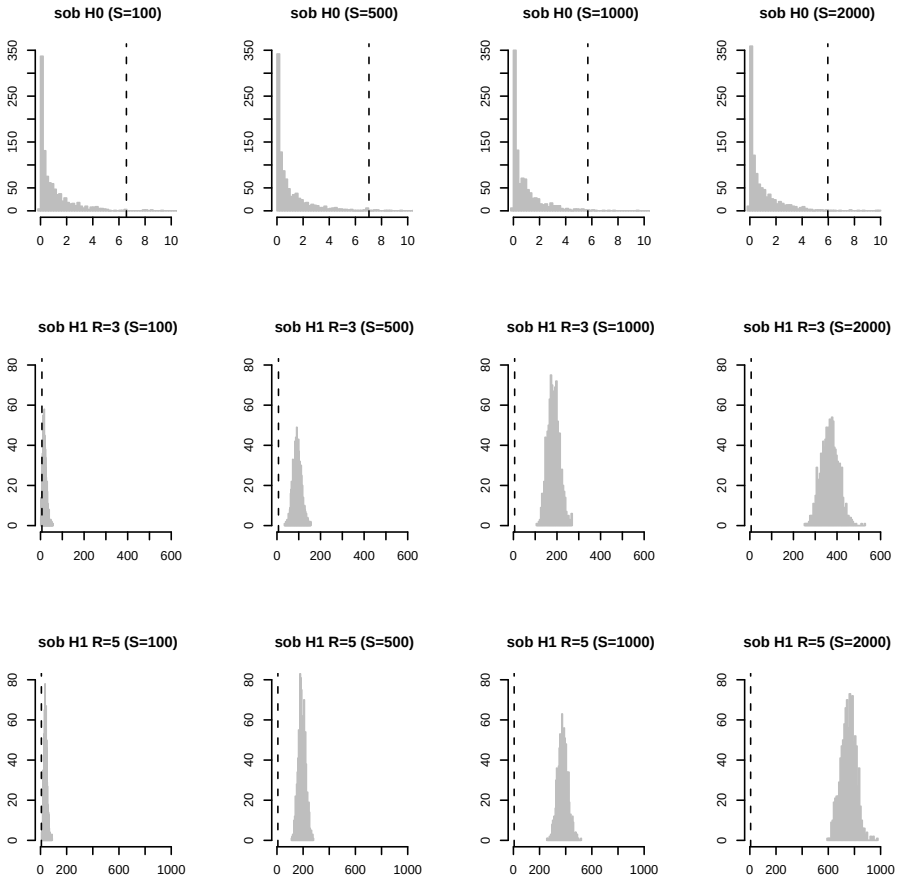


Figura 6.2: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ K80, sob H_0 e sob H_1, para a Árvore 3.

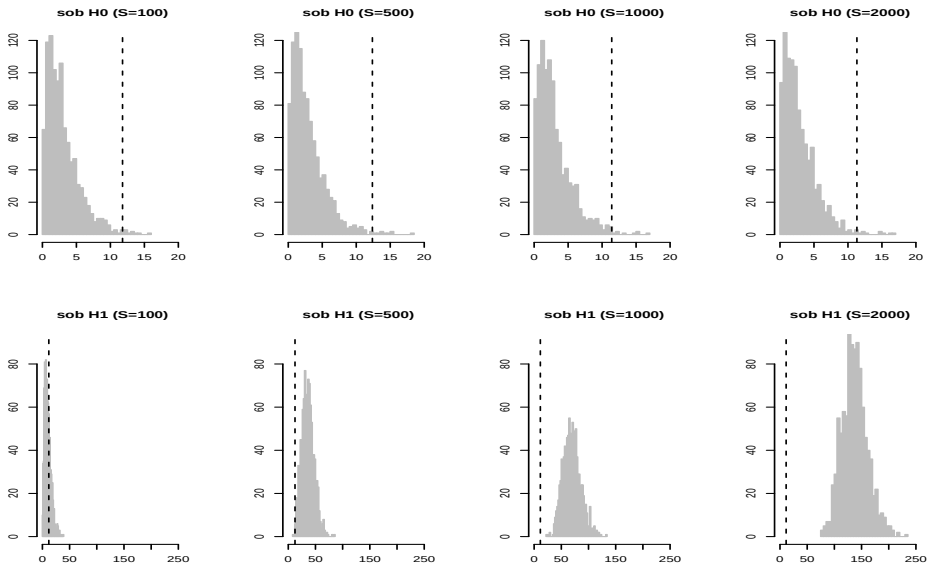


Figura 6.3: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ F81, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

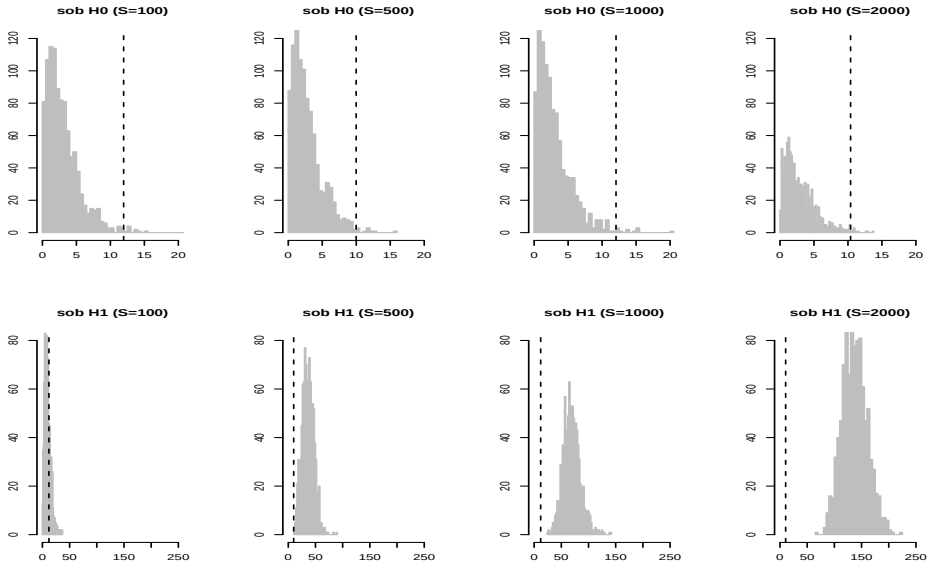


Figura 6.4: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 $\times$ HKY85, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

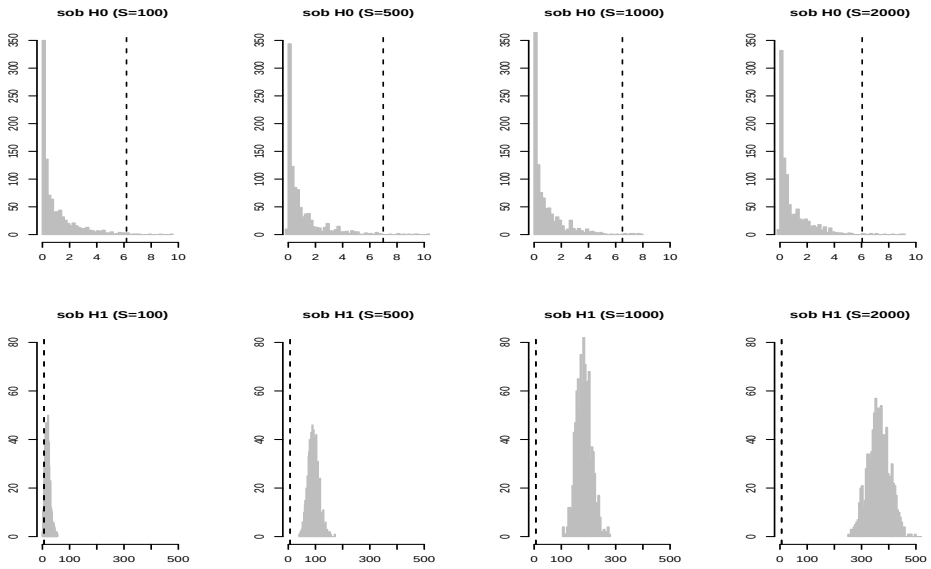


Figura 6.5: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F81 $\times$ HKY85, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

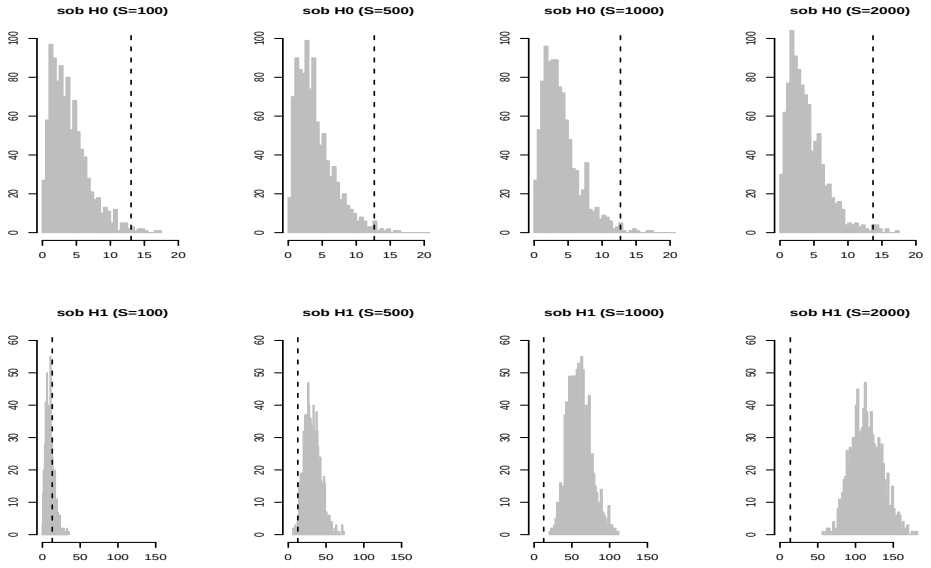


Figura 6.6: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses HKY85 $\times$ GTR, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

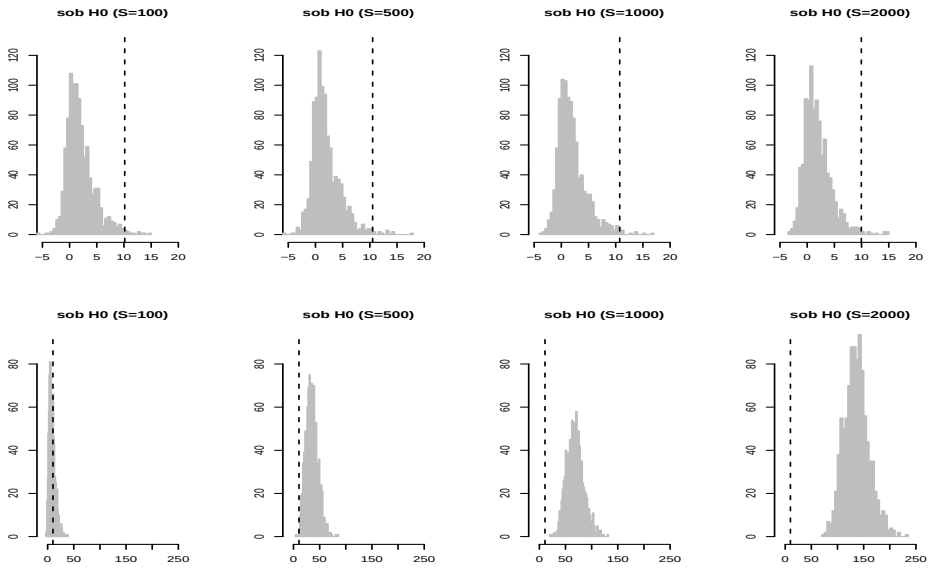


Figura 6.7: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ F81, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

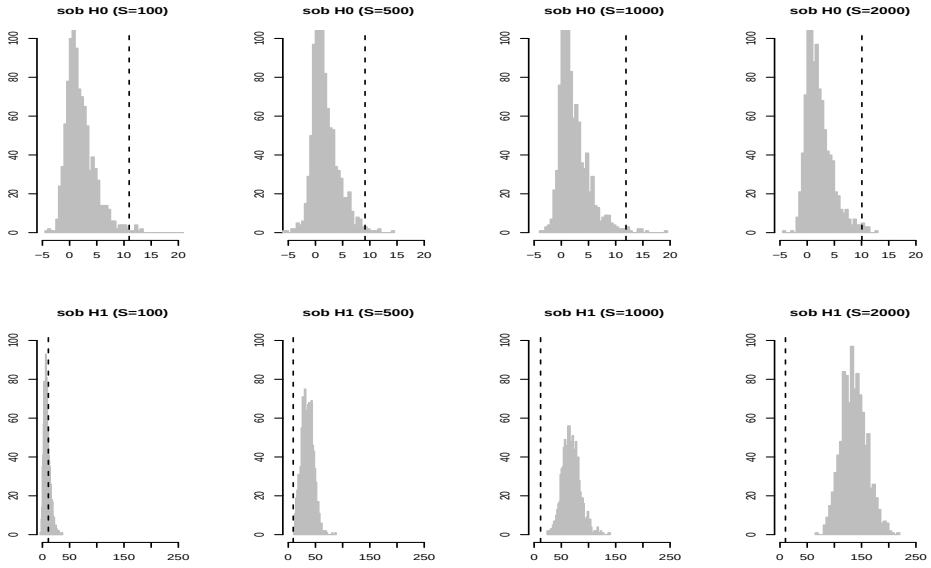


Figura 6.8: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 $\times$ HKY85, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 3.

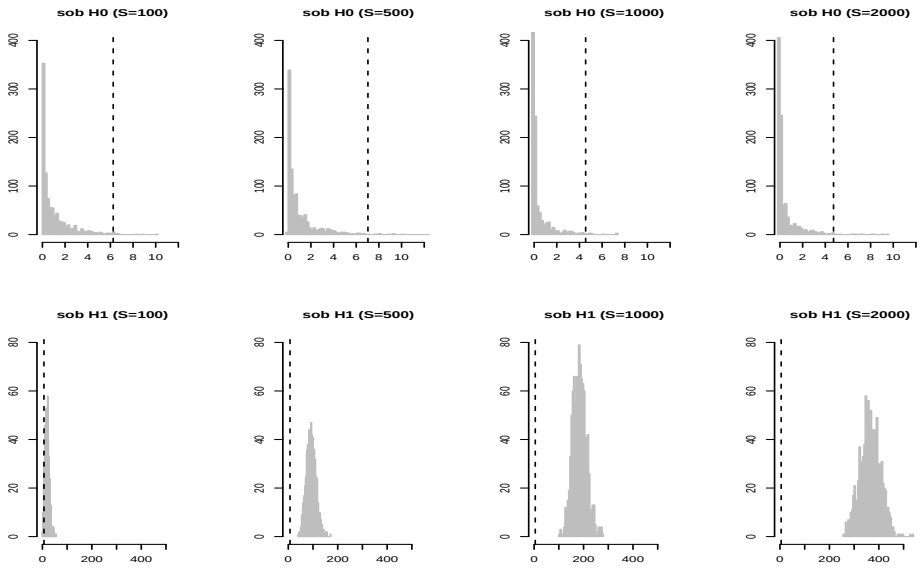


Figura 6.9: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F81 $\times$ HKY85 sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 3.

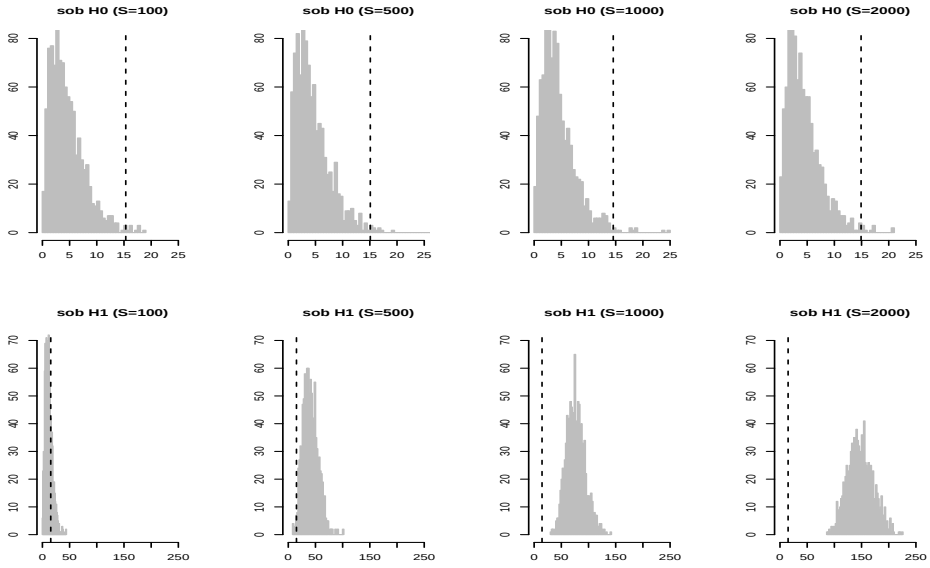


Figura 6.10: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses HKY85 $\times$ GTR, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 3.

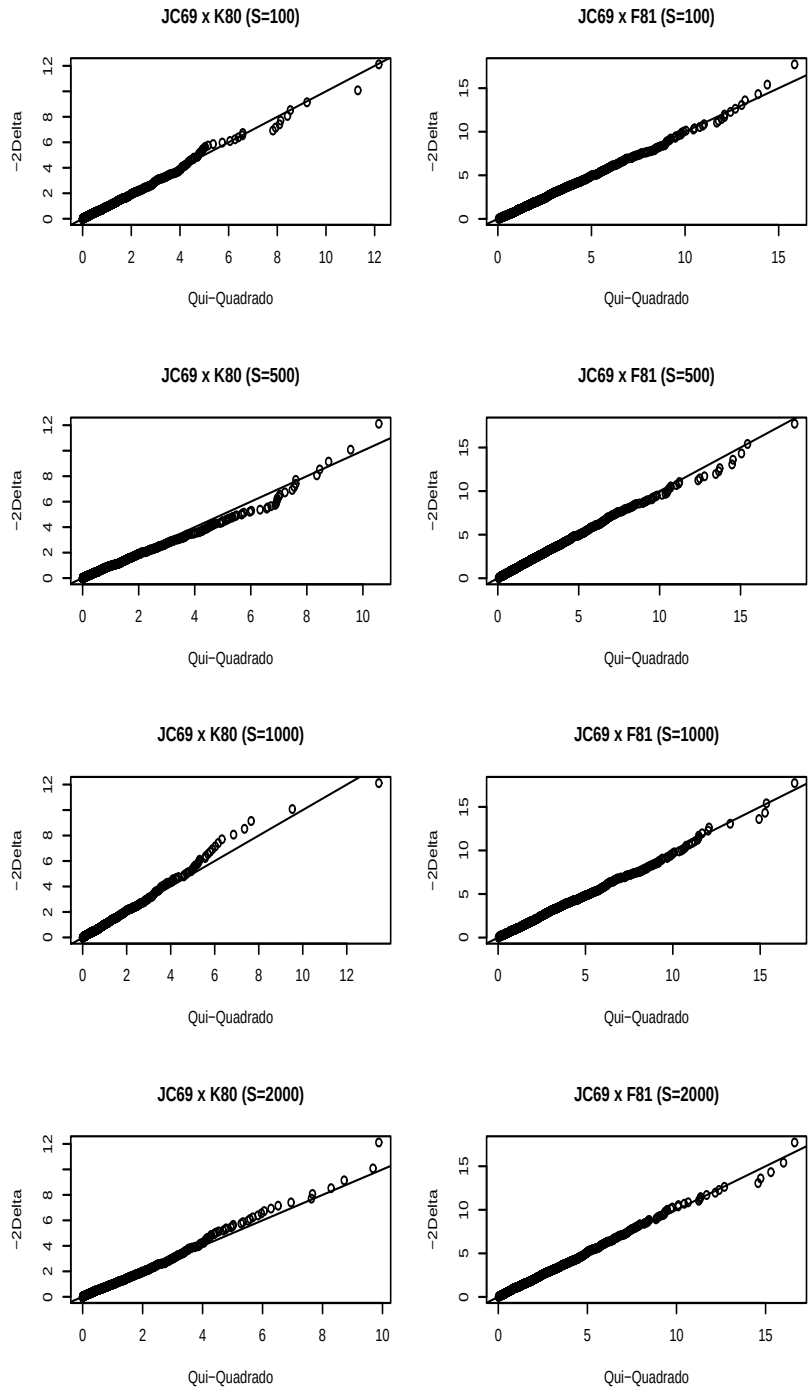


Figura 6.11: Q-Q plot sob H_0 para os Testes de Hipóteses JC69 × K80 e JC69 × F81, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

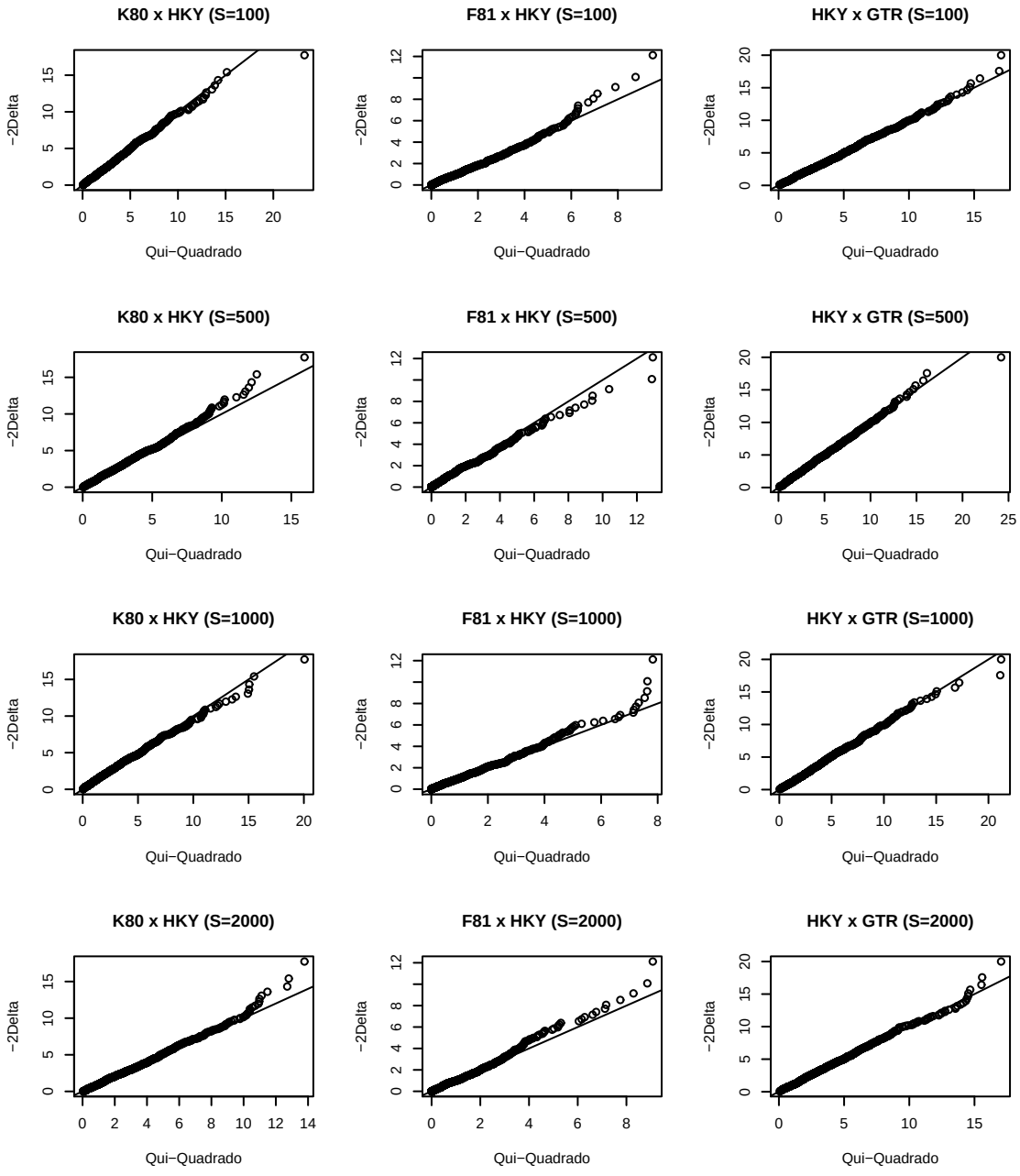


Figura 6.12: Q-Q plot sob H_0 para os Testes de Hipóteses K80 × HKY85, F81 × HKY85 e HKY85 × GTR, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

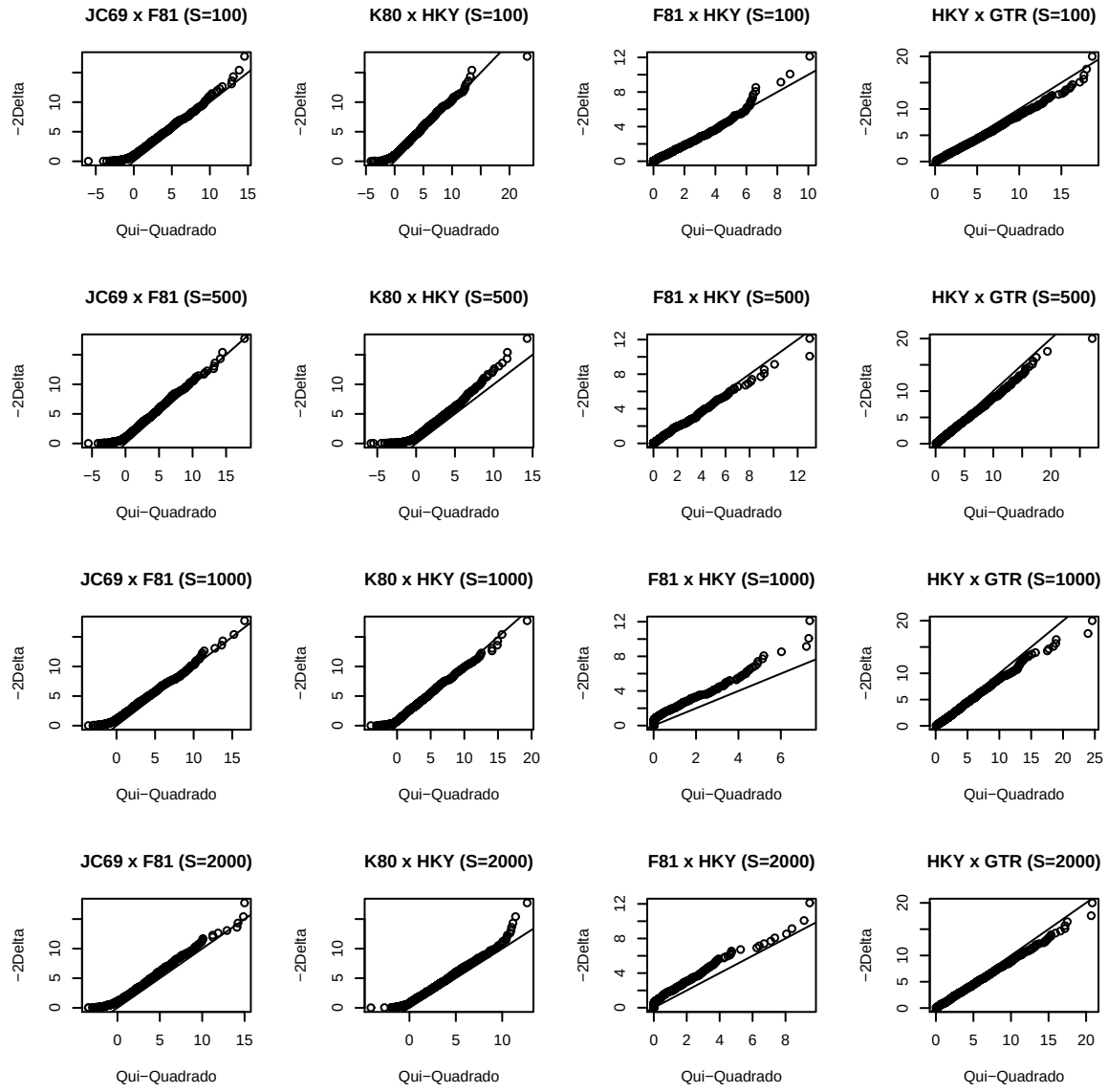


Figura 6.13: Q-Q plot sob H_0 para os Testes de Hipóteses JC69 $\times$ F81, K80 $\times$ HKY85, F81 $\times$ HKY85 e HKY85 $\times$ GTR, com $\tilde{\mathbf{p}}_0$, para a Árvore 3.

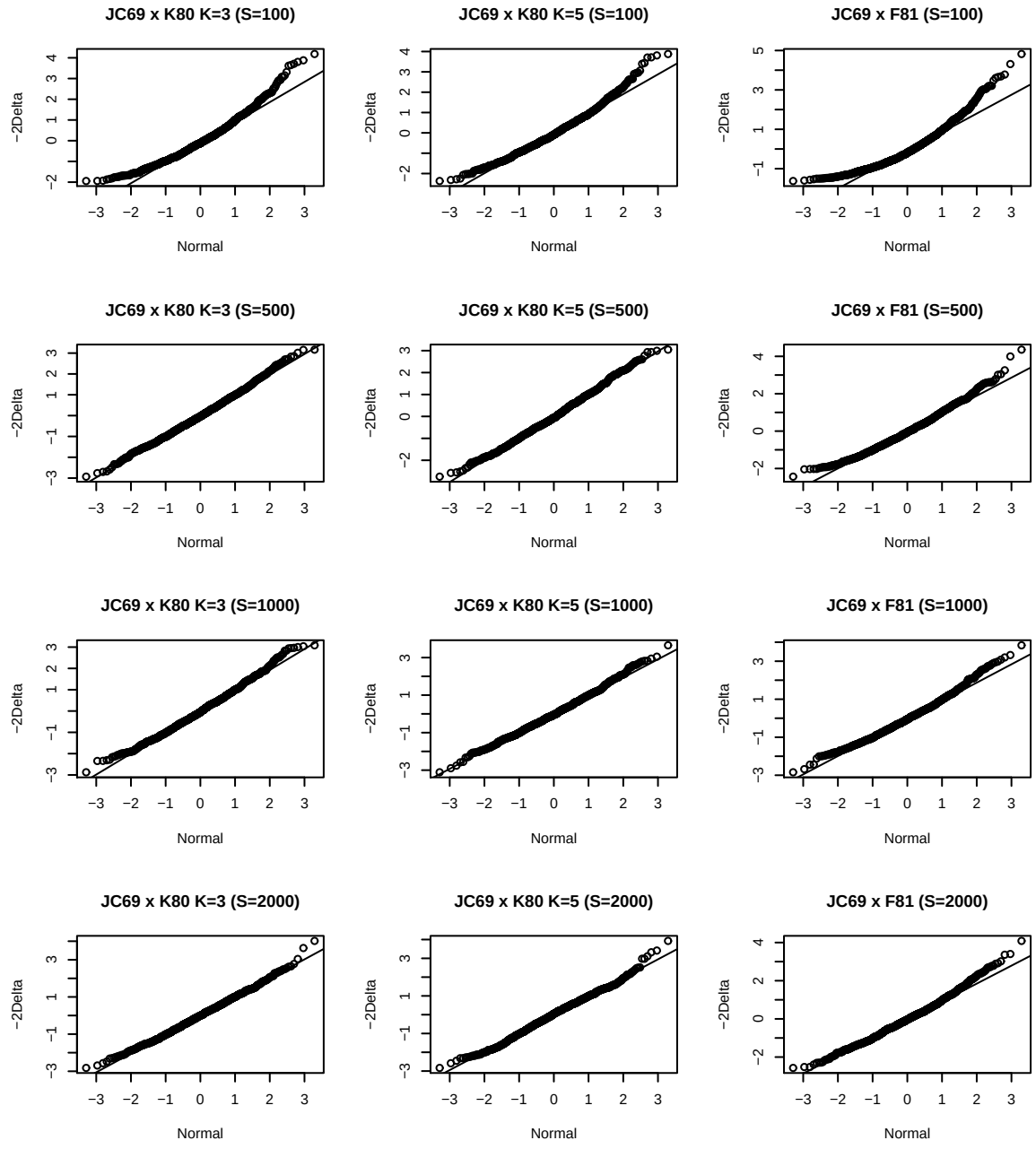


Figura 6.14: Q-Q plot sob H_1 para os Testes de Hipóteses JC69 × K80 e JC69 × F81, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

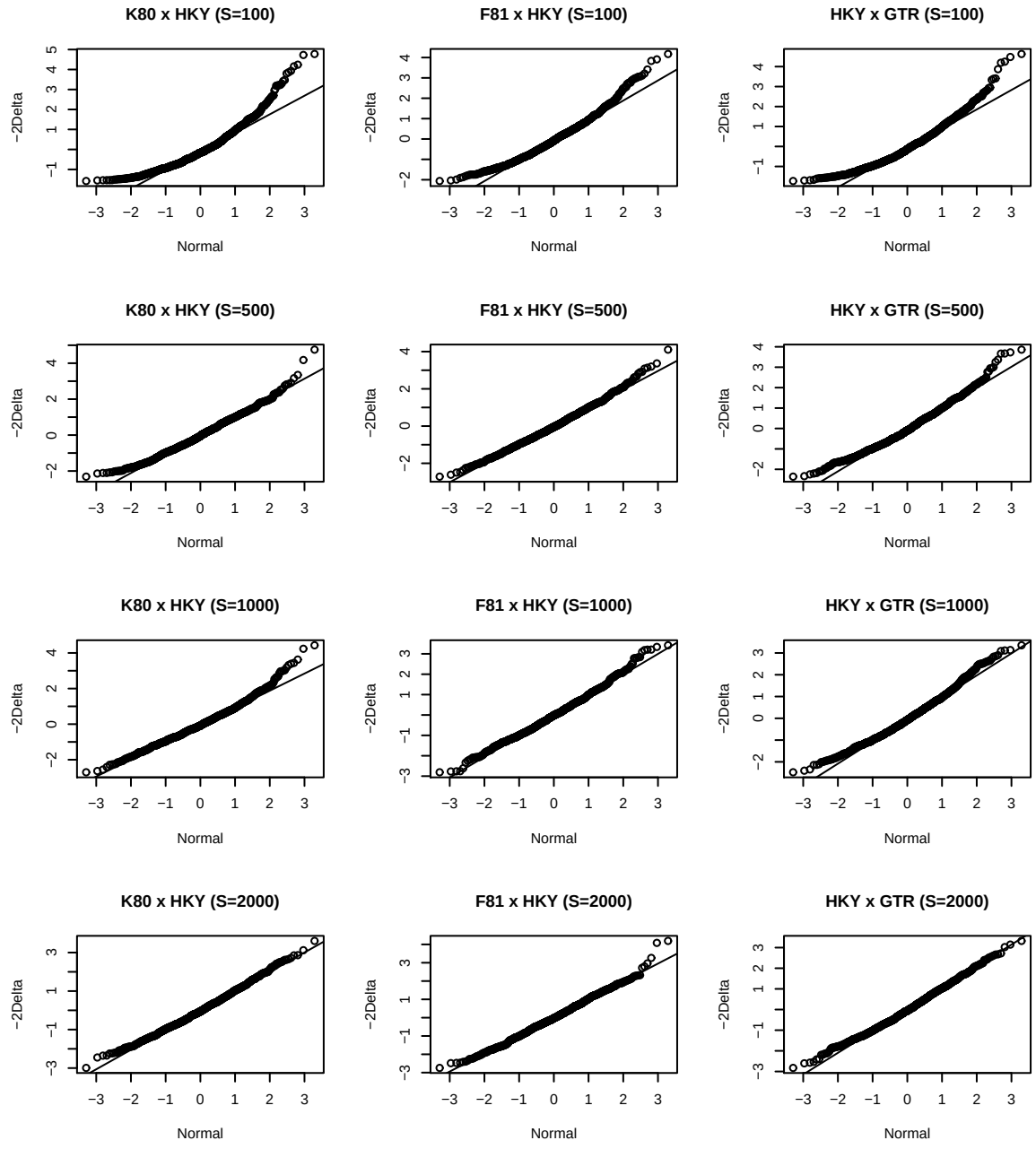


Figura 6.15: Q-Q plot sob H_1 para os Testes de Hipóteses K80 × HKY85, F81 × HKY85 e HKY85 × GTR, com $\hat{\mathbf{p}}_0$, para a Árvore 3.

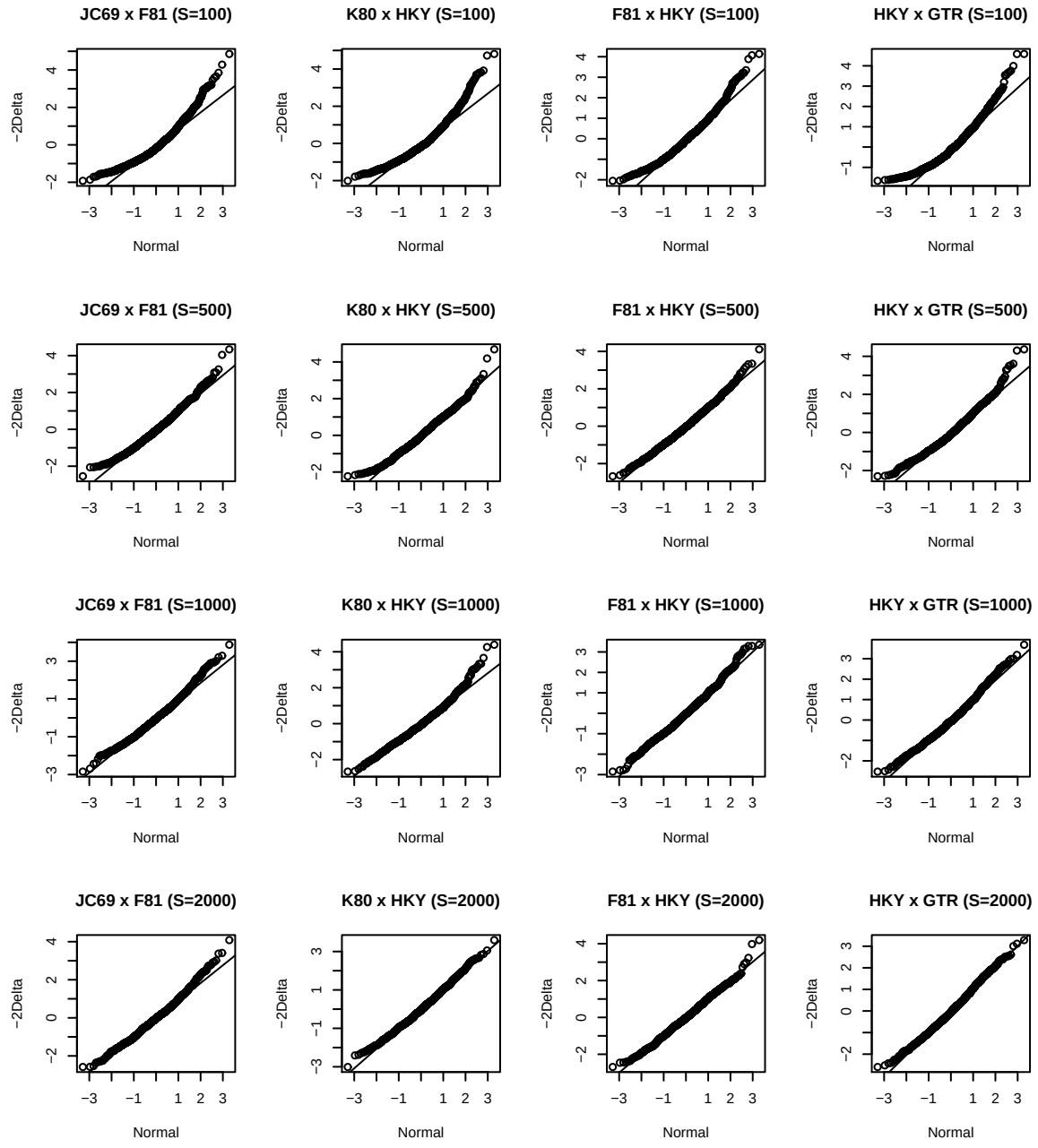


Figura 6.16: Q-Q plot sob H_1 para os Testes de Hipóteses JC69 × F81, K80 × HKY85, F81 × HKY85 e HKY85 × GTR, com $\tilde{\mathbf{p}}_0$, para a Árvore 3.

versus $H_1 : \mathbf{p}_0 = (\pi_A, \pi_G, \pi_C, \pi_T)$ tais que $0 \leq \pi_i \leq 1$ e $\pi_A + \pi_G + \pi_C + \pi_T = 1$, mas sim $H_0 : \mathbf{p}_0 = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4},)$ versus $H_1 : \mathbf{p}_0 = (\hat{\pi}_A, \hat{\pi}_G, \hat{\pi}_C, \hat{\pi}_T)$, em que $\hat{\pi}_i$ é a freqüência observada da base i na amostra $\mathbf{X}$. Assim, temos um teste de hipóteses não hierárquico e não podemos esperar a distribuição qui-quadrado.

Por outro lado, muitos dos aplicativos computacionais disponíveis para análises de máxima verossimilhança em seqüências de DNA por meio de filogenias utilizam o método dos momentos para estimar $\mathbf{p}_0$, devido à redução no custo computacional, de forma que é interessante analisar o comportamento desse teste. Destacamos que o uso de $\tilde{\mathbf{p}}_0$ pode acarretar vários problemas estatísticos. Observamos, ainda, que as distribuições encontradas nas simulações sob H_0 não foram afetadas por variações em S ou N, assim como acontece para os testes que utilizam $\hat{\mathbf{p}}_0$.

Já nas simulações dos testes de hipóteses JC69 $\times$ F81 e K80 $\times$ F84 com $\mathbf{p}_0$ estimado pelo método da máxima verossimilhança, tanto os testes qui-quadrado, quanto os Q-Q plots indicam que, sob H_0, $-2\Delta(\mathbf{X}) \sim \chi^2_3$.

Whelan e Goldman (1999) realizaram análises semelhantes, estudando os testes de hipóteses JC69 $\times$ F81 e K80 $\times$ HKY85 utilizando $\hat{\mathbf{p}}_0$ e $\tilde{\mathbf{p}}_0$. Os autores encontraram, de forma compatível com os nossos resultados, que ambos os testes utilizando o método dos momentos têm distribuição diferente da χ^2_3. Além disso, quando utilizaram os estimadores de máxima verossimilhança para $\mathbf{p}_0$, obtiveram distribuições próximas à χ^2_3.

Poder do Teste e Função de Distribuição de $-2\Delta(\mathbf{X})$ sob H_1

As Tabelas 6.5 e 6.6 apresentam o poder do teste obtido por meio de simulações para $S \in \{100, 500\}$, para as árvores 1, 2 e 3, com $\hat{\mathbf{p}}_0$ e $\tilde{\mathbf{p}}_0$, respectivamente. Para todos os testes estudados e todas as árvores o poder do teste quando $S \in \{1000, 2000\}$ foi 1. Como as tabelas e os histogramas obtidos nas simulações (histogramas das árvores 1 e 2 estão no Apêndice C) evidenciam, o poder do teste aumenta à medida que o tamanho da seqüência aumenta para todas as árvores e em todos os testes. Além disso, notamos que, para as três árvores, o poder do teste que compara os modelos JC69 e K80 aumenta quando utilizamos o parâmetro $K = 5$ em vez de $K = 3$. Esses resultados já eram esperados e evidenciam o efeito que um aumento no tamanho da amostra e no valor do parâmetro que distingue os modelos causam no poder do teste.

Por outro lado, não encontramos um padrão no efeito causado pelo aumento do número de seqüências N no poder do teste. Entretanto, percebemos que os valores para o poder do teste simulado para os testes na árvore 2 foram sempre maiores do que o mesmos valores simulados para os testes nas árvores 1 e 3. Como esse efeito não pode ser relacionado ao número de seqüências das árvores, já que a árvore 2 apresenta valor de N maior do que da árvore 1 e menor do que da árvore 3, devemos olhar para outras características da filogenia para buscar explicação para esse comportamento.

Anisimova et al. (2001) avaliaram, por meio de simulações de Monte Carlo, o poder de

Tabela 6.5: Poder do Teste Utilizando $\hat{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$
JC69 × K80 $K = 2.5$	1	0.8750	1.0000
	2	1.0000	1.0000
	3	0.9440	1.0000
JC69 × K80 $K = 1.5$	1	0.9990	1.0000
	2	1.0000	1.0000
	3	1.0000	1.0000
JC69 × F81	1	0.3290	1.0000
	2	0.5310	1.0000
	3	0.3080	0.9990
K80 × HKY85	1	0.4120	0.9980
	2	0.5410	1.0000
	3	0.2680	1.0000
F81 × HKY85	1	0.8390	1.0000
	2	1.0000	1.0000
	3	0.9520	1.0000
HKY85 × GTR	1	0.2320	0.9370
	2	0.4840	1.0000
	3	0.2360	0.9860

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela 6.6: Poder do Teste Utilizando $\tilde{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$
JC69 × F81	1	0.3250	1.0000
	2	0.4710	1.0000
	3	0.3520	0.9990
K80 × HKY85	1	0.4500	0.9980
	2	0.4790	1.0000
	3	0.2710	1.0000
F81 × HKY85	1	0.8480	1.0000
	2	1.0000	1.0000
	3	0.9500	1.0000
HKY85 × GTR	1	0.2370	0.9500
	2	0.5460	1.0000
	3	0.2680	0.9880

Nota: Os testes foram realizados ao nível de 99% de confiança.

testes que comparam modelos de mutações sinônimas e não sinônimas em seqüências de DNA. Em suas simulações, encontraram que o poder do teste é afetado pelos comprimentos das seqüências e pelos parâmetros do modelo. Além disso, avaliaram o efeito que a divergência entre as seqüências tem sobre o poder do teste, percebendo que o poder do teste é pequeno para árvores com pequena divergência entre as seqüências (div = 0.11), aumentando para valores intermediários (div = 1.1), e voltando a diminuir para grandes valores de divergência (div = 11). Ou seja, testes realizados em árvores com valores intermediários de divergência entre as seqüências apresentam maior poder do teste. Eles explicam tal fato da seguinte forma: em árvores com div pequeno houve pequena quantidade de evolução, de forma a dificultar a detecção de qual modelo é mais adequado. Assim, o aumento de div acarreta em um aumento do poder do teste. Entretanto, para valores muito grandes de div, o número de mutações ocorridas na filogenia é tão grande que várias mutações provavelmente ocorreram no mesmo ramo, mascarando a história do processo, e reduzindo a nossa capacidade de obter informações sobre as seqüências.

Observamos que nas simulações aqui realizadas, as árvores 1 e 3, que apresentam poder do teste mais baixo, têm respectivamente $\text{div}_1 = 0.9400$ e $\text{div}_3 = 0.6990$. Já a árvore 2, na qual obtivemos o maior poder do teste para todos os testes de hipóteses analisados, tem $\text{div}_2 = 1.7600$. Assim, esses resultados são compatíveis com aqueles apresentados em Anisimova et al. (2001).

Para comparar o poder dos diferentes testes devemos considerar a influência da escolha dos parâmetros para a simulação. Assim, não podemos comparar o poder de testes não relacionados. Entretanto, devido à utilização dos mesmos parâmetros, podemos comparar testes que avaliam o mesmo efeito na evolução das seqüências.

Por exemplo, nos testes de hipóteses JC69 × K80 e F81 × HKY85, o que distingue os dois modelos testados é a taxa de *transições* e *transversões*, representada pelo parâmetro K. Nos dois testes de hipóteses, sob H_0, temos que $K = 1$, enquanto que, sob H_1, K é um parâmetro a ser estimado. Portanto, ambos os testes de hipóteses estão avaliando o efeito da taxa de *transições* e *transversões* no desempenho do modelo. A diferença é que o primeiro teste assume que a distribuição das bases é homogênea, enquanto que no segundo a distribuição das bases deve ser estimada a partir dos dados. Observando as Tabelas 6.5 e 6.6 não temos resultados conclusivos sobre qual teste teria o maior poder.

Já na comparação entre os testes de hipóteses JC69 × F81 e K80 × HKY85, que testam a homogeneidade na distribuição das bases, não encontramos padrão distintivo. Enquanto que, para as árvores 1 e 2, o teste K80 × HKY85 tem maior poder do teste, para a árvore 3 o maior poder do teste ocorre no JC69 × F81 (tanto utilizando $\hat{\mathbf{p}}_0$ quanto utilizando $\tilde{\mathbf{p}}_0$).

De modo semelhante, ao compararmos os testes de hipóteses com $\mathbf{p}_0$ estimado tanto pelo método da máxima verossimilhança quanto pelo dos momentos não percebemos distinção clara no poder do teste. Em geral, um mesmo teste utilizando os diferentes estimadores para $\mathbf{p}_0$ pode apresentar poder do teste semelhante. Portanto, para aqueles testes de hipóteses que não testam para diferenças de $\mathbf{p_0}$, em que a distribuição de $-2\Delta(\mathbf{X})$, sob H_0, é qui-quadrado, devido à redução no esforço computacional, a utilização de $\tilde{\mathbf{p}}_0$ pode ser vantajosa. Já para os testes de hipóteses que testam para diferenças de $\mathbf{p_0}$, é necessária uma forma precisa de determinar a distribuição de $-2\Delta(\mathbf{X})$, sob H_0, para que o estimador $\tilde{\mathbf{p}}_0$ possa ser amplamente usado.

Por fim, notamos que nas funções de distribuição obtidas para $-2\Delta(\mathbf{X})$ sob H_1, não só os valores da estatística do teste aumentam com o aumento de S, mas também ocorre uma modificação no formato da função de distribuição. Para muitos dos testes de hipóteses, nos valores menores de S, encontramos distribuições assimétricas e, à medida que S aumenta, a distribuição se torna mais simétrica. Os testes de normalidade, cujos resultados estão apresentados nas Tabelas D.3 e D.4, não apresentaram padrão conclusivo. Percebe-se, entretanto, que há uma tendência para que a hipótese de normalidade seja aceita para valores maiores de S. Por outro lado, para alguns testes, a hipótese de normalidade é aceita para algum valor baixo de S e depois rejeitada para os valores maiores de S.

6.1.1 Efeito do Comprimento das Seqüências sobre o Poder do Teste

Comprimento Ótimo da Seqüência

Nessa seção analisamos o efeito do comprimento da seqüência sobre o poder do teste. Conforme observado na Seção 6.1, o poder do teste aumenta com o aumento de

Tabela 6.7: Valor Crítico do Teste de Hipóteses JC69×K80, variando S.

S	Árvore 1	Árvore 2	Árvore 3
50	6.9406	7.0134	6.0831
100	6.9406	7.6887	6.5701
200	6.5581	7.8732	6.1061
300	6.1416	7.3673	6.3780
400	6.6519	7.0430	6.3829
500	6.5562	6.5456	7.0386
600	6.0699	7.4848	6.1244
700	7.1687	7.3865	5.8399
800	6.4658	6.1187	6.2562
900	6.3819	6.2504	6.3979
1000	6.8426	6.6568	5.6946

Nota: Os testes foram realizados ao nível de 99% de confiança.

S. Entretanto, a utilização de seqüências demasiadamente longas acarreta em aumento de custos para o seqüenciamento e aumento do tempo computacional. Assim, o objetivo desse estudo é prescrever um comprimento de seqüência adequado para que o teste da razão de verossimilhança tenha boa capacidade de discriminar os modelos.

Por meio de simulações de Monte Carlo, estudamos o teste de hipóteses JC69×K80. Pelos resultados da Seção 6.1, percebemos que para esse teste de hipóteses, o poder do teste é sempre 1, já para $S = 1000$. Assim, para este conjunto de simulações refinamos o intervalo, utilizando $S \in \{50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000\}$. Além disso, na Seção 6.1, verificamos que o valor de K e a filogenia F também afetam o poder do teste. Desta forma, nestas simulações utilizamos as mesmas filogenias da Figura 6.1, e valores de $K \in \{1.5, 2, 3\}$. Todas as simulações foram realizadas com $Re = 1000$ replicações.

A Tabela 6.7 apresenta o valor crítico simulado ao nível de 99% de confiança, para as três árvores utilizadas. Já as Tabelas 6.8, 6.9 e 6.10 apresentam o poder do teste simulado nos casos em que $K = 1.5$, $K = 2$ e $K = 3$, respectivamente.

Notamos que, quando $K = 2$, obtemos um poder do teste superior a 95% para todas as três árvores consideradas quando $S = 300$. Já, quando $K = 3$, $S = 200$ é suficiente para que o poder do teste seja maior do que 95%. No caso de $K = 1.5$ apenas para $S = 900$ que o poder de 95% é atingido. Além disso, percebemos que aumentos posteriores em S têm pouco efeito sobre o poder do teste.

Entretanto, notamos que o poder do teste varia tanto com a filogenia utilizada, quanto com o valor do parâmetro K. Aparentemente, a influência do valor de K sobre o poder do teste é maior do que a da filogenia. Assim, para valores $K \geq 2$, podemos dizer que $S = 300$ é um bom comprimento de seqüência para o teste de hipóteses JC69×K80. Para

Tabela 6.8: Poder do Teste de Hipóteses JC69×K80, variando S e com $K = 1.5$.

S	Árvore 1	Árvore 2	Árvore 3
50	0.0490	0.1870	0.0850
100	0.1160	0.3090	0.1490
200	0.2720	0.6660	0.4050
300	0.4270	0.8790	0.5590
400	0.5450	0.9610	0.7150
500	0.6700	0.9920	0.7950
600	0.7850	0.9970	0.8970
700	0.8190	0.9990	0.9390
800	0.8940	1.0000	0.9680
900	0.9620	1.0000	0.9790
1000	0.9420	1.0000	0.9950

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela 6.9: Poder do Teste de Hipóteses JC69×K80, variando S e com $K = 2$.

S	Árvore 1	Árvore 2	Árvore 3
50	0.1840	0.5660	0.2810
100	0.4110	0.8620	0.5360
200	0.8000	0.9950	0.8920
300	0.9560	1.0000	0.9790
400	0.9860	1.0000	0.9990
500	0.9980	1.0000	0.9990

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela 6.10: Poder do Teste de Hipóteses JC69×K80, variando S e com $K = 3$.

S	Árvore 1	Árvore 2	Árvore 3
50	0.5460	0.9620	0.6800
100	0.8840	1.0000	0.9440
200	0.9990	1.0000	1.0000
300	1.0000	1.0000	1.0000
400	1.0000	1.0000	1.0000
500	1.0000	1.0000	1.0000

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela 6.11: Decisão do Teste de Normalidade Shapiro-Wilks para $-2\Delta(\mathbf{X})$, sob H_1.

Teste	Árvore	$S = 4000$	$S = 6000$	$S = 8000$	$S = 10000$
JC69 × K80	1	-	aceita	aceita	aceita
	2	-	aceita	aceita	aceita
	3	-	aceita	aceita	aceita
JC69 × F81	1	-	aceita	aceita	aceita
	2	aceita	aceita	aceita	aceita
	3	aceita	aceita	aceita	aceita
K80 × HKY85	1	-	aceita	aceita	aceita
	2	aceita	aceita	aceita	aceita
	3	aceita	aceita	aceita	aceita
F81 × HKY85	1	aceita	aceita	aceita	aceita
	2	aceita	aceita	aceita	aceita
	3	aceita	aceita	aceita	aceita
HKY85 × GTR	1	aceita	aceita	aceita	aceita
	2	aceita	aceita	-	aceita
	3	aceita	aceita	aceita	aceita

Nota: Os testes foram realizados ao nível de 99% de confiança.

seqüências com $K < 2$, são necessários comprimentos maiores para S.

Estudo da Distribuição Assintótica de $2\Delta(\mathbf{X})$ sob H_1.

No Teorema 3.3 obtivemos o resultado analítico de que a distribuição assintótica da estatística do teste sob H_1 é regida por uma variável aleatória normal. Entretanto, nas simulações apresentadas na primeira parte dessa seção, encontramos apenas indícios fracos de que, com o aumento de S, a distribuição de $-2\Delta(\mathbf{X})$, sob H_1, se aproxima da normal. Deste modo, com o objetivo de corroborar o resultado do Teorema 3.3, realizamos novas simulações de bootstrap paramétrico utilizando os mesmos parâmetros das Tabelas 6.1 e 6.2, mas com valores de $S \in \{4000, 6000, 8000, 10000\}$. As distribuições obtidas para $-2\Delta(\mathbf{X})$, sob H_1, foram submetidas ao Teste de Normalidade Shapiro-Wilks, e o resultado desse teste está apresentado na Tabela 6.11. Na Tabela 6.11, para os testes em que não aparece a palavra "aceita", a hipótese nula de normalidade foi rejeitada.

Os resultados apresentados na Tabela 6.11 indicam que, para valores grandes de S, a distribuição da estatística do teste, sob H_1, é normal, conforme o Teorema 3.3. Além disso, o resultado das simulações quando $S \in \{100, 500, 1000, 2000\}$ é condizente com uma distribuição qui-quadrado não-central, onde o parâmetro de não-centralidade aumenta à medida que S aumenta, como a Figura 6.6 evidencia. Este é o argumento utilizado na demonstração do Teorema 3.3.

Tabela 6.12: Poder do Teste de Hipóteses JC69×K80, variando N e com $K = 2$.

N	$S = 100$	$S = 300$
2	0.1030	0.3870
5	0.5370	0.9810
10	0.7080	0.9940
20	0.7790	0.9970
30	0.6900	0.9990
40	0.7780	0.9990
50	0.7900	0.9990
100	0.8150	1.0000

Nota: Os testes foram realizados ao nível de 99% de confiança.

É importante notar, que apesar de interessante, esse resultado não pode ser utilizado para obter aproximações para o poder do teste. Isso se deve ao fato de que, à medida que S aumenta, o poder do teste também aumenta. E como visto nas simulações, em geral, a normalidade assintótica só é atingida quando o poder do teste já é muito próximo de 1.

6.1.2 Efeito do Número de Seqüências sobre o Poder do Teste

Nesta sub-seção, estudamos, por meio de simulações de Monte Carlo, o efeito que o número de seqüências N tem sobre o poder do teste JC69×K80. Para tanto, foram geradas 8 filogenias com $N \in \{2, 5, 10, 20, 30, 40, 50, 100\}$. Para isolar o efeito de N do efeito da divergência div $= \sum \tau_i$, os comprimentos dos ramos das árvores geradas foram padronizados, de forma que div $= 1$. Para cada uma das 8 filogenias geradas, foram simuladas seqüências com $S \in \{100, 300\}$. Para o modelo K80, foram utilizados valores de $K \in \{2, 3\}$. Todas as simulações foram realizadas com $Re = 1000$ replicações.

As Tabelas 6.12 e 6.13 apresentam o poder do teste obtido para as simulações quando $K = 2$ e $K = 3$, respectivamente. Percebemos que o aumento de N leva a um aumento no poder do teste. Entretanto, para $N \geq 5$ este aumento é muito lento, se comparado ao efeito de incrementar S, como pode ser visto nas Tabelas 6.9 e 6.10. [1]

Aplicamos o teste χ^2 para verificar se a distribuição da estatística do teste $-2\Delta(\mathbf{X})$, sob H_0, é de fato χ^2_1, conforme o Teorema 3.1. Para todas as árvores utilizadas e nos dois comprimentos de seqüência utilizados, a hipótese nula, de que $-2\Delta(\mathbf{X})$ tem distribuição qui-quadrado com um grau de liberdade foi aceita. Aplicamos, também, o teste de normalidade Shapiro Wilks para a distribuição de $-2\Delta(\mathbf{X})$, sob H_1. Em todos as simulações realizadas, a hipótese nula de normalidade foi rejeitada ao nível de significância 95%. Entretanto, a análise dos histogramas de $-2\Delta(\mathbf{X})$, sob H_1, indica que, com o aumento

[1] Além disso, notamos que o aumento no tempo computacional é linearmente proporcional ao aumento de S. Já um aumento em N afeta mais rapidamente o tempo computacional, devido à natureza do algoritmo do cálculo e otimização da função de verossimilhança (ver Capítulo 3).

Tabela 6.13: Poder do Teste de Hipóteses JC69×K80, variando N e com $K = 3$.

N	$S = 100$	$S = 300$
2	0.3750	0.9260
5	0.9670	1.0000
10	0.9800	1.0000
20	0.9960	1.0000
30	0.9910	1.0000
40	0.9980	1.0000
50	0.9990	1.0000
100	0.9990	1.0000

Nota: Os testes foram realizados ao nível de 99% de confiança.

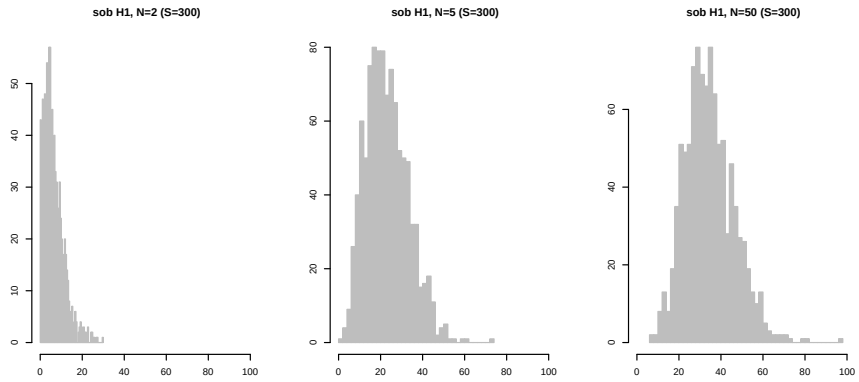


Figura 6.17: Histogramas da Distribuição de $-2\Delta(\mathbf{X})$, sob H_1, com $N \in \{2, 5, 50\}$.

de N, a distribuição se torna progressivamente mais simétrica, como pode ser visto na Figura 6.17.

6.1.3 Efeito do Índice de Divergência entre as Seqüências sobre o Poder do Teste

Nesta sub-seção estudamos, através de simulações de Monte Carlo, o efeito que a divergência div entre as seqüências tem sobre o poder do teste JC69×K80. A divergência entre as seqüências é uma característica da filogenia F, e é dada pela soma dos comprimentos dos ramos, div $= \sum \tau_i$.

Para as simulações, geramos uma árvore filogenética com $N = 10$ e, posteriormente, multiplicamos seus comprimentos de ramos por constantes para obter os valores de div desejados. Assim, a mesma estrutura filogenética gerou as 14 árvores utilizadas nesse estudo, com div $\in \{0.25, 0.5, 0.75, 1, 1.5, 1.75, 2, 3, 4, 5, 6, 8, 10, 12\}$. Foram geradas

Tabela 6.14: Poder do Teste de Hipóteses JC69×K80, variando div.

div	0.25	0.5	0.75	1	1.5	1.75	2
Poder do Teste	0.3840	0.5550	0.6400	0.8030	0.8230	0.8590	0.9180
div	3	4	5	6	8	10	12
Poder do Teste	0.9320	0.9360	0.9060	0.8930	0.7630	0.6230	0.4050

Nota: Os testes foram realizados ao nível de 99% de confiança.

seqüências com $S = 100$. Sob o modelo K80, o valor utilizado para o parâmetro K foi 2. Em todas as simulações utilizamos $Re = 1000$ replicações.

A Tabela 6.14 apresenta os valores obtidos para o poder do teste. Observamos que, para valores muito pequenos ou grandes de div o poder do teste é baixo. Mas para valores intermediários de div, por exemplo entre 2 e 5, o poder do teste é alto. Esses resultados são condizentes com o que foi comentado na Seção 6.1, referente às simulações das árvores 1, 2, e 3. Além disso, nossos resultados estão de acordo com o reportado por Anisimova et al. (2001) em modelos de mutações sinônimas e não sinônimas em seqüências de DNA.

6.2 Comparação entre o Poder do Teste Obtido com o Estimador $\hat{D}$ e com o Bootstrap Paramétrico

Nesta seção, avaliamos a comparação entre a estimativa do poder do teste utilizando o estimador $\hat{D}$, conforme apresentado na Seção 3.2.4, e utilizando o bootstrap paramétrico.

Inicialmente, estudamos o teste de hipóteses JC69 ×K80. Para tanto, estimamos o parâmetro de não centralidade D para seqüências simuladas com o modelo K80 com $S \in \{50, 100, 200, 300, 400, 500\}$. Nas simulações foram utilizadas as árvores 1, 2 e 3 apresentadas na Figura 6.1. Todas as seqüências utilizadas apresentaram $\hat{K} = 1.5$. Os valores obtidos para o poder do teste utilizando o estimador $\hat{D}$ foram então comparados com os valores obtidos para o poder do teste utilizando o bootstrap paramétrico.[2] Os resultados estão apresentados na Tabela 6.15.

Notamos que, para todas as seqüências analisadas, o poder do teste obtido pelo bootstrap paramétrico foi superior à estimativa do poder do teste obtida utilizando o estimador $\hat{D}$. Notamos, ainda, que com, o aumento de S, o poder do teste estimado por ambos os métodos aumenta. Além disso, também por ambos os métodos de estimação, o poder do teste obtido para um determinado comprimento de seqüência foi sempre maior para a árvore 2.

O poder obtido pelo bootstrap paramétrico nessas simulações foi sempre maior para a árvore 3 do que para a árvore 1, embora essa diferença não seja tão grande quanto para

[2]Note que os resultados do bootstrap paramétrico já foram utilizados na Seção 6.1.1 para outra análise.

Tabela 6.15: Comparação do Poder do Teste JC69 × K80 obtido por Bootstrap Paramétrico e com $\hat{D}$, para $K = 1.5$.

Árvore	S	Estimativa com $\hat{D}$		Bootstrap
		$\hat{D}$	$\widehat{\text{Poder}}$	Poder
1	50	0.6414	0.0337	0.0490
	100	1.3167	0.0686	0.1160
	200	2.8386	0.1905	0.2720
	300	4.0466	0.3204	0.4270
	400	5.5242	0.4095	0.5450
	500	6.9036	0.5267	0.6700
2	50	1.8329	0.1107	0.1870
	100	3.5768	0.2465	0.3090
	200	7.7119	0.5794	0.6660
	300	10.7818	0.7601	0.8790
	400	14.3642	0.8875	0.9610
	500	18.2805	0.9553	0.9920
3	50	0.8532	0.0494	0.0850
	100	1.0824	0.0624	0.1490
	200	2.9369	0.1941	0.4050
	300	4.8021	0.3500	0.5590
	400	6.7524	0.5087	0.7150
	500	8.2342	0.6130	0.7950

Nota: Os testes foram realizados ao nível de 99% de confiança.

a árvore 2. Essa mesma relação entre as árvores foi encontrada para as estimativas do poder utilizando $\hat{D}$ para todos os comprimentos de seqüência, com exceção de $S = 100$. Nas seqüências de comprimento 100, o poder do teste obtido utilizando o estimador $\hat{D}$ foi muito semelhante para as árvores 1 e 3, embora tenha sido um pouco maior para a árvore 1 do que para a árvore 3.

Assim, desses resultados concluímos que a estimativa do poder do teste utilizando $\hat{D}$ para JC69 $\times$ K80 subestima o poder quando comparado àquele obtido pelo método do bootstrap paramétrico. Entretanto, em geral, o poder do teste estimado acompanha os padrões obtidos com o bootstrap paramétrico, ou seja, se o teste A tem maior poder do que o teste B segundo o bootstrap paramétrico então a estimativa do poder utilizando $\hat{D}$ tenderá a apresentar essa mesma distinção.

Ressaltamos novamente que a principal vantagem da utilização do estimador $\hat{D}$ para obter o poder do teste é o baixo custo computacional. Enquanto o método do bootstrap paramétrico requer um número alto de replicações, o estimador $\hat{D}$ pode ser obtido diretamente a partir da análise da amostra, sem a necessidade de cálculos adicionais além daqueles que usualmente são realizados para esses testes de hipóteses.

6.3 Avaliação da Correção de Vício por Jackknife

Nesta seção apresentamos os resultados de simulações realizadas para avaliar o uso da correção do vício de estimadores de máxima verossimilhança por Jackknife, conforme Seção 5.3. Para tais simulações, utilizou-se o modelo HKY85 (i.i.d.), e a filogenia 1 da Figura 6.1. Foram utilizadas 5 combinações de parâmetros para o modelo HKY85, que estão apresentadas na Tabela 6.16. Para cada conjunto de parâmetros, foram estimados K, π_A, π_G, π_C (parâmetros do modelo HKY85), além dos comprimentos dos ramos da filogenia, $\tau_1, \cdots, \tau_5$. Note que o parâmetros π_T não precisa ser estimado independentemente, uma vez que $\pi_T = 1 - \pi_A - \pi_C - \pi_G$. Os conjuntos de parâmetros 4 e 5 são tais que os valores de K e p_0 sejam próximos daqueles estimados para as aplicações de felinos e primatas das Seções 3.3.2 e 3.3.3, respectivamente.

Para cada conjunto de parâmetros da Tabela 6.16, simulamos um conjunto de dados

Tabela 6.16: Parâmetros Utilizados nas Simulações para Avaliar o Jackknife.

P	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
P1	0.12	0.1	0.1	0.22	0.4	2	0.25	0.25	0.25
P2	0.12	0.1	0.1	0.22	0.4	3	0.20	0.20	0.25
P3	0.12	0.1	0.1	0.22	0.4	5	0.25	0.25	0.25
P4	0.12	0.1	0.1	0.22	0.4	9	0.32	0.20	0.23
P5	0.12	0.1	0.1	0.22	0.4	2	0.29	0.21	0.25

da mesma forma como feito nos passos **2** e **3** do algoritmo da Seção 3.2.1. Os parâmetros foram estimados por máxima verossimilhança, utilizando o pacote PAML. Em seguida, aplicamos o Jackknife, e realizamos a correção do vício dos estimadores, de acordo com o procedimento apresentado na Seção 5.3. Foram realizadas 200 replicações deste procedimento para cada conjunto de parâmetros, nos comprimentos de seqüência $S = 80, 140, 200$.

Nesta seção apresentamos apenas os dados referentes ao conjunto de parâmetros P1. Os demais resultados encontram-se no Apêndice F.

Para avaliar o efeito que o aumento no comprimento das seqüências têm sobre a estimativa do vício pelo Jackknife, a Tabela 6.17 e as Tabelas F.2, F.5, F.8 e F.11, do Apêndice F, trazem a média sobre as 200 replicações da estimativa do vício para cada parâmetro, respectivamente para os conjuntos de parâmetros de P1 a P5. Ao observá-las, notamos que, em geral, o vício estimado pelo Jackknife diminui com o aumento de S. Este resultado é esperado, uma vez que estamos tratando com estimadores de máxima verossimilhança, que são assintoticamente não viciados.

Deste modo, observamos que a utilização do Jackknife para correção do vício parece se justificar mais quando aplicada a seqüências curtas, uma vez que nas seqüências longas, além do vício ser muito pequeno, o custo computacional do Jackknife é muito grande.

Para avaliar qual o real efeito que a correção do vício por Jackknife tem sobre as estimativas, comparamos os respectivos estimadores com os valores reais dos parâmetros. Para tanto, utilizamos a seguinte estatística

$$g = |\theta - \hat{\theta}| - |\theta - \hat{\theta}_J|, \tag{6.1}$$

em que θ é o parâmetro de interesse, $\hat{\theta}$ seu estimador de máxima verossimilhança e $\hat{\theta}_J$ o estimador com a correção de vício por Jackknife, conforme apresentado na Seção 5.3. Note que, se a correção do vício por Jackknife melhorar a estimativa de θ ($|\theta - \hat{\theta}| > |\theta - \hat{\theta}_J|$), então g é positivo; já, se a estimativa de $\hat{\theta}_J$ for pior, teremos g negativo. Desse modo, g mede o desempenho da correção do vício por Jackknife, isto é, quanto maior for g, melhor terá sido o desempenho da correção.

Note que a utilização da estatística g só é possível porque estamos tratando com simulações, em que conhecemos o verdadeiro valor de θ. Para seqüências reais, ela não poderá ser utilizada.

A Tabela 6.18 e as Tabelas F.3, F.6, F.9 e F.12, do Apêndice F, apresentam os valores médios de g nas 200 replicações, respectivamente, para os conjuntos de parâmetros P1 a P5. Notamos que, na maioria das estimativas, assim como acontece com a estimativa do vício, o ganho obtido com a correção por Jackknife é reduzido com o aumento de S na maioria das vezes. Como $\hat{\theta}_J$ é calculado com base na estimativa do vício, é esperado que o módulo de g diminua com a redução do vício. Entretanto, em um número considerável das estimativas obtivemos g maior quando $S = 140$, e não quando $S = 80$.

Tabela 6.17: Vício Médio para P1.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00320	0.00234	0.00088	0.00612	0.01363	0.13249	-0.00007	-0.00005	0.00010
140	0.00237	0.00091	0.00061	0.00246	0.00696	0.06213	0.00002	-0.00003	-0.00002
200	0.00083	0.00059	0.00058	0.00224	0.00509	0.04310	0.00008	0.00006	-0.00005

Tabela 6.18: Ganho Médio com $\hat{\theta}_J$ para P1.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00129	0.00065	0.00062	0.00252	0.00221	0.05652	-0.00007	0.00002	-0.00002
140	0.00035	0.00027	0.00018	0.00044	0.00100	0.01441	0.00003	0.00003	-0.00001
200	0.00070	0.00019	0.00022	0.00068	0.00142	0.00829	0.00003	-0.00003	-0.00005

As tabelas de ganho ainda indicam que os parâmetros das freqüências das bases (π_A, π_G e π_C) são os que menos são afetados pela correção do vício pelo Jackknife. Assim, para estes parâmetros isoladamente, a correção pode não se justificar.

A Tabela 6.19 e as Tabelas F.4, F.7, F.10 e F.13, do Apêndice F, apresentam o percentual das replicações em que $\hat{\theta}_J$ representa uma melhora na estimativa de θ, em relação a $\hat{\theta}$, para, respectivamente, os conjuntos de parâmetros de P1 a P5. Notamos que os diferentes parâmetros tem comportamentos distintos em relação a esta estatística. Em alguns casos, temos proporções que se aproximam dos 80%, que é um resultado muito bom. No entanto, em outros temos proporções próximas de 50%. Notamos, entretanto, que, diferentemente do que ocorre com a estimativa do vício e com g, não identificamos tendência distinta desta proporção com o aumento de S.

A Figura 6.18 apresenta uma representação gráfica do vício médio, ganho médio e percentual de melhora das estimativas para cada parâmetro de P1, dados respectivamente nas Tabelas 6.17, 6.18 e 6.19. Nela fica evidenciado que tanto o vício estimado quanto o ganho para o parâmetro K é muito maior que nos outros parâmetros do modelo. Entretanto, essa relação não se repete no percentual de melhoras com a correção do vício por Jackknife. Já a Figura 6.19 apresenta o gráfico do vício médio e do ganho médio para cada parâmetro de P1, com exceção do parâmetro K. Nela podemos ver como a correção do vício com o Jackknife pouco afeta as estimativas das freqüências das bases.

Tabela 6.19: Percentual das Replicações com g Positivo para P1.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	64.5	62.5	66.5	54.0	45.0	52.0	46.0	54.5	51.5
140	69.0	56.0	59.5	50.5	46.5	49.5	53.5	49.5	47.0
200	71.0	56.0	63.0	50.5	54.5	48.0	53.5	46.5	46.0

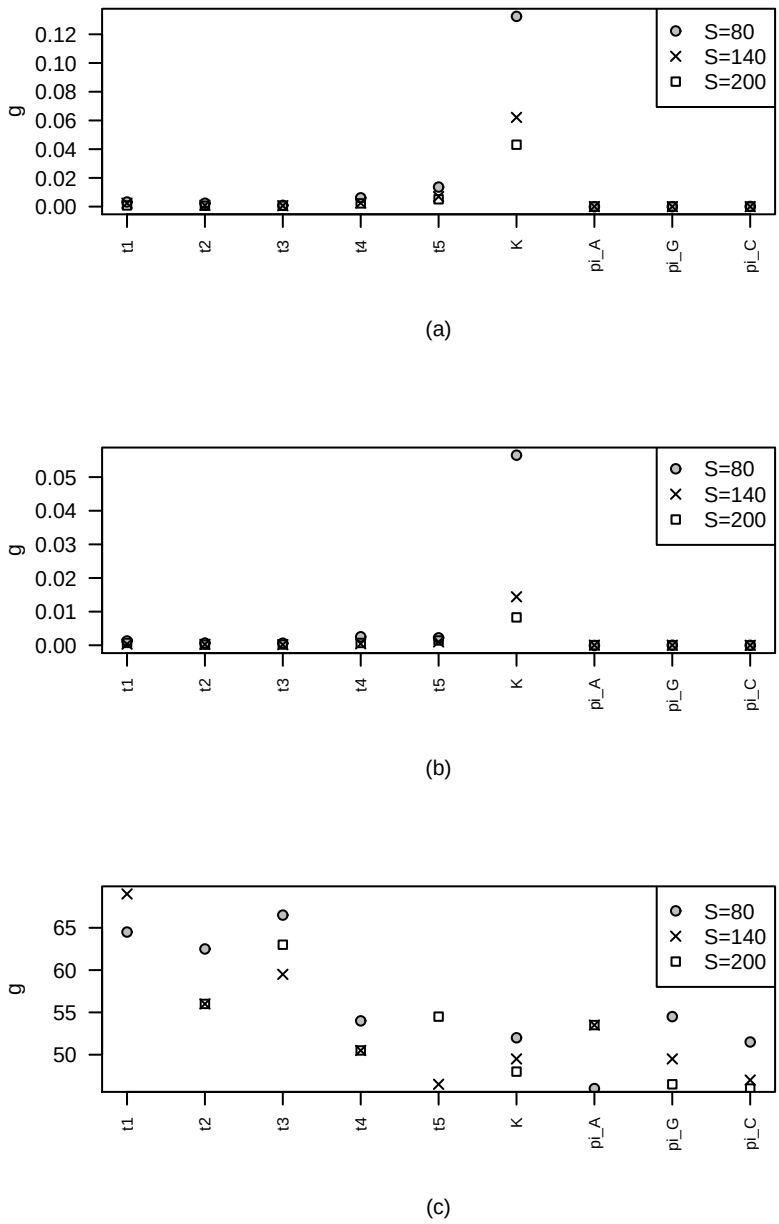
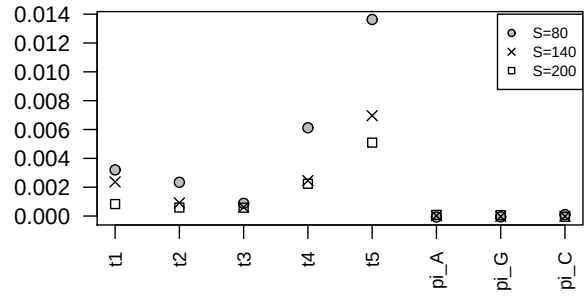
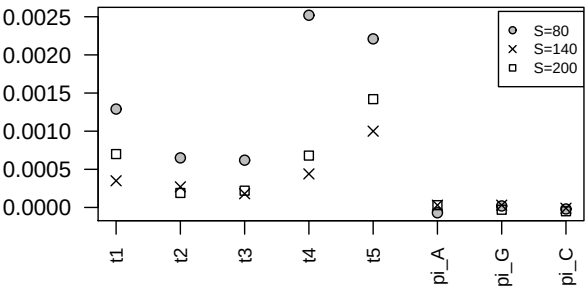


Figura 6.18: Gráficos de (a) Vício Médio, (b) Ganho Médio, e (c) Percentual de Replicações com g Positivo para P1.



(a)



(b)

Figura 6.19: Gráficos de (a) Vício Médio e (b) Ganho Médio para P1, para Todos os Parâmetros Exceto K.

Tabela 6.20: Ganho Médio e Probabilidade de Melhorar as Estimativas para cada Conjunto de Parâmetros.

S	Parâmetros	pm	$\bar{g}$
80	P1	0.52500	0.01069
140	P1	0.55110	0.00423
200	P1	0.53500	0.00204
80	P2	0.55833	0.00762
140	P2	0.53444	0.00185
200	P2	0.54333	0.00127
80	P3	0.54556	0.02107
140	P3	0.55277	0.00815
200	P3	0.56111	0.00639
80	P4	0.54056	0.08351
140	P4	0.56778	0.03150
200	P4	0.52944	0.00963
80	P5	0.55166	0.00708
140	P5	0.53778	0.00249
200	P5	0.54278	0.00128

Finalmente, apresentamos um resumo destas estatísticas para cada conjunto de parâmetros. Na Tabela 6.20 constam o valor de g médio ($\bar{g}$) e a probabilidade de melhora (pm) das estimativas, em cada conjunto de parâmetros utilizado. A análise da tabela confirma uma redução do ganho médio com o aumento de S. Já a probabilidade de haver uma melhora na estimativa com a correção do vício, calculada como a proporção das replicações em que a estimativa foi melhorada, não sofre influência de S.

Nas simulações apresentadas, obtivemos uma probabilidade de melhora da estimativa média próxima de 0.55. A questão importante é se este ganho é suficiente para justificar a utilização do Jackknife (e seu custo computacional). Esta resposta depende de diversos fatores como o comprimento das seqüências, a precisão desejada, o objetivo das estimativas, entre outros. Uma forma de abordar esta questão, sugerida na página 8 de Efron (1982), é através da comparação entre as estimativas de vício e variância. O autor comenta que, se a razão

$$\text{razão} = \frac{\widehat{\text{vício}(\hat{\theta})}}{\sqrt{\widehat{\text{Var}(\hat{\theta})}}}, \tag{6.2}$$

em que $\widehat{\text{vício}(\hat{\theta})}$ e $\widehat{\text{Var}(\hat{\theta})}$ são estimados pelo Jackknife, for muito pequena, então o vício tem importância pequena em relação à variância. Assim, ele sugere que a correção de vício por Jackknife seja utilizada apenas quando esta razão for superior a 1/4.

Tabela 6.21: Percentual das Replicações em que razão $> 1/4$ para P1.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	92.5	95.0	92.0	97.5	99.5	100.0	77.5	74.0	72.5
140	94.5	93.0	89.5	99.0	100.0	100.0	90.0	86.0	87.5
200	96.0	97.0	94.0	99.5	100.0	100.0	96.0	95.0	95.5

Calculamos a razão dada pela expressão (6.2) para todas as replicações do conjunto de parâmetros número 1. A Tabela 6.21 apresenta o percentual das replicações em que a razão entre vício e desvio padrão, estimados por Jackknife, foi superior a 1/4. A variância dos parâmetros foi estimada utilizando a seguinte expressão:

$$\widehat{\mathrm{Var}(\hat{\theta})} = \frac{S-1}{S} \sum_{i=1}^{S} \left(\hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)} \right)^2, \tag{6.3}$$

onde $\hat{\theta}_{(-i)}$ é a estimativa do parâmetro θ para a amostra quando se retira a observação i, e $\hat{\theta}_{(\cdot)} = \frac{1}{S} \sum_{i=1}^{S} \hat{\theta}_{(-i)}$, conforme dado no Lema 5.11 (ver Efron e Tibshirani, 1993).

Notamos que, tanto para os comprimentos de ramos quanto para o parâmetro K, obtivemos valores superiores a 1/4 para a razão entre vício e desvio padrão em cerca de 90% das replicações. E, até mesmo para os parâmetros de freqüência das bases, na grande maioria das replicações a condição foi satisfeita. Assim, segundo o critério sugerido por Efron (1982), na grande maioria das vezes a utilização da correção do vício por Jackknife é recomendada para o conjunto P1 de parâmetros.

Entretanto, conforme comentado anteriormente, a estimativa do vício diminui com o aumento da amostra. Então, ainda que o critério anterior esteja satisfeito, é possível que para comprimentos grandes de S, o uso do Jackknife não se justifique. A determinação para quais valores de S o uso do Jackknife não seria recomendado é uma questão complexa. Diversos fatores como o número de seqüências na amostra, o modelo utilizado e o grau de precisão desejado nas estimativas estão envolvidos. Assim, esta é uma questão que requer uma análise mais detalhada, e pode ser objeto de futuros trabalhos.

Capítulo 7

Conclusões e Futuros Trabalhos

Neste trabalho estudamos diversos aspectos dos testes da razão de verossimilhança que comparam modelos de substituição de bases em seqüências de DNA. Percebemos, tanto por vias analíticas quanto por meio das simulações de Monte Carlo, que a distribuição assintótica qui-quadrado é realmente uma boa aproximação para a estatística do teste $-2\Delta(\mathbf{X})$ com S grande, quando consideramos apenas modelos em que os sítios são independentes e suas taxas de mutação são idênticas. Além disso, as simulações mostram que a assintoticidade é atingida até para comprimentos de seqüência muito pequenos, como $S = 100$.

Utilizamos a mesma teoria assintótica para estudar o poder dos testes da razão de verossimilhança, através da distribuição da estatística do teste sob H_1. Percebemos que, diferentemente do que acontece com o valor crítico, o poder do teste é influenciado pelo comprimento das seqüências. Mostramos não só que o poder do teste aumenta com o aumento de S, mas também que a distribuição da estatística do teste, sob H_1, converge para uma distribuição normal, quando as seqüências ficam mais longas. As simulações de Monte Carlo corroboram esses resultados, e indicam que a distribuição assintótica normal é atingida apenas para valores maiores de S, como $S = 4000$.

Por meio das simulações do Capítulo 6, notamos também que o número N de seqüências da amostra tem influência muito pequena sobre o poder do teste. Já o índice de divergência entre as seqüências apresenta uma relação interessante com o poder: o poder do teste é maior para valores intermediários de div. Essa relação é bastante forte, e deverá ser alvo de futuros trabalhos para verificar se existe um valor de div ótimo para todas a seqüências, ou se esse valor varia dependendo de outros parâmetros.

Propomos um novo estimador $\hat{D}$ para o parâmetro de não-centralidade da distribuição qui-quadrado, que pode ser utilizado, em conjunto com a teoria assintótica, para obter o poder do teste. Além disso, demonstramos a consistência de $\hat{D}$. Tal estimador tem a vantagem de não exigir esforço computacional adicional, além daquele necessário para o cálculo da estatística do teste. Como estudos do poder destes testes são praticamente inexistentes na literatura, não pudemos comparar o desempenho deste estimador com

outros trabalhos. Assim, para avaliar a estimativa do poder do teste utilizando $\hat{D}$, a comparamos com as estimativas para o poder obtidas pelas simulações de Monte Carlo. Notamos que a variação de parâmetros e características da amostra parecem influenciar as duas estimativas do poder do teste de forma semelhante. Entretanto, o estimador do poder que utiliza $\hat{D}$ tende a sub-estimar o poder do teste, quando comparado com as simulações de Monte Carlo.

Em trabalhos futuros, pretendemos seguir investigando as propriedades estatísticas do estimador $\hat{D}$. Além disso, pretendemos avaliar como ele se comporta quando aplicado a outros testes e analisar melhor a relação que existe entre as estimativas do poder do teste por simulações de Monte Carlo e pelo $\hat{D}$. Tudo indica que este estimador é um forte candidato a ser implementado nos aplicativos de análise de máxima verossimilhança para usuários, uma vez que ele fornece o poder do teste praticamente sem custo computacional adicional. Ressaltamos que, até o momento, o poder do teste dificilmente é considerado quando se utilizam testes de hipóteses para escolher entre modelos de evolução do DNA.

No Capítulo 5, através de hipóteses restritivas, obtivemos cotas inferiores para a variância dos estimadores de máxima verossimilhança dos parâmetros de alguns modelos de substituição de bases. Notamos ainda que, ao menos no caso do modelo JC69, com o aumento de S, a variância de $\hat{\alpha}$ se aproxima de sua cota inferior. Além disso, a variância de $\hat{\alpha}$ se torna mais próxima de sua cota inferior à medida que a taxa de mutação α diminui.

Destacamos que, em trabalhos futuros, esta mesma idéia pode ser aplicada aos demais modelos, para a obtenção de cotas para seus parâmetros. Acreditamos que a aproximação da variância do estimador da sua cota inferior, quando S aumenta, seja um padrão que se repete nos demais modelos.

Sugerimos, ainda no Capítulo 5, a utilização da técnica de reamostragem Jackknife para a correção do vício dos estimadores de máxima verossimilhança utilizados nesse trabalho. Realizamos diversos estudos de simulação para avaliar quais os reais benefícios de seu uso. Observamos que o ganho obtido com o uso deste método é reduzido à medida que as seqüências se tornam maiores, ainda que geralmente seja positivo. Por outro lado, a comparação entre estimativas de vício e variância indica que, na grande maioria das vezes, a importância da estimativa do vício é grande em relação à variância. Assim, temos bons argumentos que favorecem o uso desta técnica para melhorar as estimativas dos parâmetros, ao menos no caso de S pequeno. Ainda assim, faz-se necessário um estudo que avalie esse método em casos mais gerais e identifique para quais valores de S ele é recomendado.

As aplicações da teoria deste trabalho aos genes de primatas e felinos ilustram bem a interpretação dos resultados dos testes de hipótese em seqüências reais e o cuidado que se deve ter com os pressupostos do modelo, que podem levar a conclusões equivocadas. Por meio delas, exemplificamos como se aplicam os testes de hipóteses e quais conclusões podem ser obtidas. As aplicações também foram importantes para confirmar o que freqüentemente é dito na literatura: são necessários modelos que considerem

as diferentes pressões que existem no DNA e que expliquem a variabilidade gerada por elas. Os modelos devem ainda observar as restrições impostas pelo tempo computacional. Apresentar alguns destes modelos foi o objetivo do Capítulo 4.

O estudo de modelos que expliquem a evolução temporal das seqüências de DNA é muito importante para a biologia evolutiva. Através da estimação de parâmetros dos modelos e comparação entre modelos diferentes obtém-se informações sobre o processo biológico. Além disso, a utilização de modelos inadequados nas análises que utilizam a função de verossimilhança pode levar a conclusões equivocadas. Neste contexto, destacamos a importância dos estudos realizados sobre as características do teste da razão de verossimilhança, um dos critérios estatísticos mais amplamente utilizados para escolher entre os modelos.

Referências

[1] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. e Walter, P. (2004). *Molecular Biology of the Cell*. Garland Science & Taylor and Francis Books.

[2] Anisimova, S., Bielawski, J. e Yang, Z. (2001). "Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution". *Molecular Biology and Evolution,* Vol. **18(8)**, 1585-1592.

[3] Baum, E., Petrie, T., Soules, G. e Weiss, N. (1970). "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains". *Annals of Mathematical Statistics,* Vol. **41**, 164-171.

[4] Bielawski, P. e Yang, Z. (2003). "Maximum likelihood methods for detecting adaptive evolution after gene duplication". *Journal of Structural and Functional Genomics,* Vol. **3**, 201-212.

[5] Durbin, R., Krogh, A., Eddy, S. e Mitchison, G. (2004). *Biological Sequence Analysis.* Cambridge: Cambridge University Press.

[6] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: CBMS-NSF Regional Conference Series in Applied Mathematics.

[7] Efron, B. e Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

[8] Felsenstein, J. (1981). "Evolutionary Trees From DNA Sequences: A Maximum Likelihood Approach". *Journal of Molecular Evolution,* Vol. **17**, 368-376.

[9] Felsenstein, J. (2004). *Inferring Phylogenies.* Sunderland: Sinauer Associates.

[10] Felsenstein, J. e Churchill, G.A. (1996). "A Hidden Markov model approach to variation among sites in rate of evolution". *Molecular Biology and Evolution*, Vol. **13**, 93-104.

[11] Goldman, N. (1993). "Statistical tests of models of DNA substitution". *Journal of Molecular Evolution*, Vol. **36**, 182-198.

[12] Goldman, N. e Whelan, S. (2000). "Statistical Tests of Gamma-Distributed Rate Heterogeneity in Models of Sequence Evolution in Phylogenetics". *Molecular Biology and Evolution*, Vol. **17**, 975-978.

[13] Guidon, S. e Gascuel, O. (2003). "A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood". *Systematic Biology*, Vol. **52(5)**, 696-704.

[14] Hartigan, J. (1973). "Minimum evolution fits to a given tree". *Biometrics*, Vol. **29**, 53-65.

[15] Hasegawa, M., Kishino, H. e Yano, T. (1985). "Dating of human-ape splitting by a molecular clock of mitochondrial DNA". *Journal of Molecular Evolution*, Vol. **22**, 160-174.

[16] Johnson, N. e Kotz, S. (1970). *Continuous Univariate Distributions*, Vol. **2**. New York: John Wiley.

[17] Jukes, T. e Cantor, C. (1969). *Evolution of Protein Molecules*, 21-132. New York: Academic Press.

[18] Karlin, S. e Taylor, H. (1975). *A First Course in Stochastic Processes*. San Diego: Academic Press.

[19] Kendall, M. e Stuart, A. (1973). *The Advanced Theory of Statistics*, Vol. **2**. London: Griffin.

[20] Kimura, M. (1980). "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences". *Journal of Molecular Evolution*, Vol. **16**, 111-120.

[21] ——— (1981). "Estimation of evolutionary distances between homologous nucleotide sequences". *Proceedings of the National Academy of Sciences USA*, Vol. **78**, 454-458.

[22] Rohatgi,V. (1976). *An Introduction to Probability Theory and Mathematical Statistics*. New York: John Willey.

[23] Saitou, N. e Nei, M. (1987). "The Neighbour-Joining Method: a New Method for Reconstructing Phylogenetic Trees". *Molecular Biology and Evoltion*, Vol. **4**, 406-425.

[24] Schöniger, M. e von Haesler, A. (1994). "A Stochastic Model of Autocorrelated DNA Sequences". *Molecular Phylogenetics and Evolution*, Vol. **3(3)**, 240-247.

[25] Self, S. e Liang, K. (1987). "Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions". *J. Am. Stat. Assoc.,* Vol. **82**, 605-610.

[26] Shao, J. (2003). *Mathematical Statistics.* New York: Springer-Verlag.

[27] Sokal, R. e Sneath, P. (1963) *Numerical Taxonomy.* San Francisco: W. H. Freeman.

[28] Tamura, K. e Nei, M. (1993). "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees". *Molecular Biology and Evolution*, Vol. **10**, 512-526.

[29] Tavaré, S. e Giddings, B. (1989). "Some statistical aspects of the primary structure of nucleotide sequences". In *Mathematical Methods for DNA Sequences*, 117-132. Boca Raton: CRC Press.

[30] Tillier, E. e Collins, R. (1995). "Neighbor Joining and Maximum Likelihood with RNA Sequences: Addressing the Interdependence of Sites". *Molecular Biology and Evolution,* Vol. **12**, 7-15.

[31] Waddell, P., Penny, D. e Moore, T. (1997). "Hadamard conjugations and modeling sequence evolution with unequal rates across sites". *Molecular Phylogenetic Evolution,* Vol. **8**, 33-50.

[32] Whelan, S. e Goldman, N. (1999). "Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics". *Molecular Biology and Evolution,* Vol. **16**, 1292-1299.

[33] Wilks, S. (1962) *Mathematical Statistics.* New York: John Wiley.

[34] Yang, Z. (1993). "Maximum-Likelihood Estimation of Phylogeny from DNA Sequences when Substitution Rates Differ over Sites". *Molecular Biology and Evolution* , Vol. **10**, 1396-1401.

[35] ——— (1994). "Maximum Likelihood Phylogenetic Estimation from DNA Sequances with Variable Rates over Sites: Approximate Methods". *Journal of Molecular Evolution,* Vol. **39**, 306-314.

[36] ——— (1995). "A Space-Time Process Model for the Evolution of DNA Sequences". *Genetics,* Vol. **139**, 993-1005.

[37] ——— (1996). "Among-Site Rate Variation and its Impact on Phylogenetic Analyses". *TREE,* Vol. **11(9)**, 367-372.

[38] ——— (2007). *Computational Molecular Evolution.* New York: Oxford University Press.

[39] Yang, Z. e Kumar, S. (1996). "Approximate Methods for Estimating the Pattern of Nucleotide Substitution and the Variation of Substitution Rates Among Sites". *Molecular Biology and Evolution,* Vol. **13(5)**, 650-659.

[40] Yang, M., Goldman, N. e Friday, A. (1994). "Comparison of Models for Nucleotide Substitution Used in Maximum-Likelihood Phylogenetic Estimation". *Molecular Biology and Evolution,* Vol. **11(2)**, 316-324.

Apêndice A

Nessa apêndice será apresentado um exemplo de aplicação do teste da razão de verossimilhança. Neste exemplo comparamos a eficiência dos modelos Jukes Cantor (JC69) e Kimura 2-Parâmetros (K80) para uma amostra fictícia. O objetivo é demonstrar todas as etapas envolvidas no cálculo da função de verossimilhança e na aplicação do teste.

Suponha que temos uma amostra constituída de três seqüências de DNA alinhadas, cada uma de comprimento $S = 1000$. Suponha ainda, por simplicidade, que a filogenia F que relaciona as seqüências é conhecida e dada na Figura A.1.

Como a função de verossimilhança é calculada como a probabilidade do modelo condicionada aos dados, queremos calcular qual a probabilidade de que um determinado modelo gere as seqüências da amostra. Como tanto o modelo JC69 quanto o K80 assumem que os sítios são $i.i.d$, podemos considerar cada sítio isoladamente. Assim, inicialmente, queremos encontrar a probabilidade de cada uma das possíveis combinações de bases na posição u das seqüências da amostra, mas para tanto, a filogenia deve ser considerada.

Note que a amostra $\mathbf{X}$ é constituída de três seqüências com 1000 bases em cada. Dessa forma, no exemplo em questão, a amostra, que é dada por uma matriz $N \times S$, tem dimensão 3×1000. Entretanto, para a análise desejada devemos considerar as seqüências desconhecidas dos nós internos. Dessa forma, extendemos a matriz $\mathbf{X}$ para incluir também os vetores de variáveis aleatórias $\mathbf{X}^4$ e $\mathbf{X}^5$ que representam respectivamente as seqüências desconhecidas dos nós internos 4 e 5. Portanto, $\mathbf{X} = (\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}^3, \mathbf{X}^4, \mathbf{X}^5)'$.

Função de Verossimilhança para JC69

Começaremos calculando a probabilidade $\mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A)$ de encontrarmos a base A na posição u das três seqüências da amostra, segundo o modelo JC69. Para tanto consideramos a árvore apresentada na Figura A.1.

Como os estados dos nós 4 e 5 são desconhecidos, para calcular a probabilidade desejada devemos considerar todas as possíveis combinações de bases nos nós ancestrais.

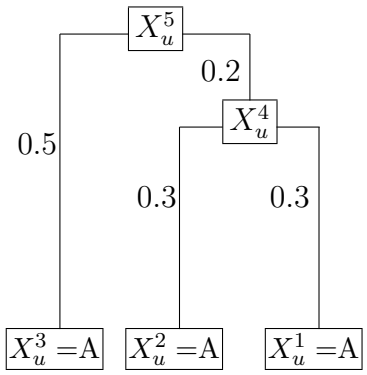


Figura A.1: Filogenia F na posição hipotética u.

Desta forma, a probabilidade desejada é dada por

$$\begin{aligned}\mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A) &= \sum_{i\in E} \pi_i \ \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A \mid X_u^5 = i) \\ &= \frac{1}{4}\sum_{i\in E} \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A \mid X_u^5 = i), \qquad \text{(A.1)}\end{aligned}$$

pois, para o modelo JC69, $\pi_i = \frac{1}{4}$ para todas as bases. Ainda temos que

$$\begin{aligned}\frac{1}{4}\sum_{i\in E} \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A \mid X_u^5 = i) &= \frac{1}{4}\sum_{i\in E} \mathbb{P}(X_u^3 = A \mid X_u^5 = i) \\ &\times \sum_{j\in E} \mathbb{P}(X_u^4 = j \mid X_u^5 = i)\ \mathbb{P}(X_u^1 = A, X_u^2 = A, \mid X_u^4 = j). \qquad \text{(A.2)}\end{aligned}$$

Note que, como estamos considerando o modelo JC69, temos

$$\mathbb{P}(Y_{\tau+\tau_0} = i \mid Y_{\tau_0} = j) = \mathbb{P}(Y_{\tau+\tau_0} = i \mid Y_{\tau_0} = k) \qquad \text{(A.3)}$$

para quaisquer bases i, j e k, tais que $i \neq j$ e $i \neq k$. Na expressão (A.3) Y representa a variável aleatória de uma cadeia de Markov regida pela matriz (2.9) de taxas infinitesimais do processo JC69.

A expressão (A.3) significa que as mutações para bases diferentes são equiprováveis. Dessa forma, o somatório em (A.2) pode ser discriminado e eventos de igual probabilidade reunidos de forma que temos

$$
\begin{aligned}
& \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A) \\
= \; & \frac{1}{4}(\mathbb{P}(X_u^3 = A|X_u^5 = A)\mathbb{P}(X_u^4 = A|X_u^5 = A)\mathbb{P}(X_u^2 = A|X_u^4 = A)\mathbb{P}(X_u^1 = A|X_u^4 = A) \\
+ \; & 3\mathbb{P}(X_u^3 = A|X_u^5 = A)\mathbb{P}(X_u^4 = C|X_u^5 = A)\mathbb{P}(X_u^2 = A|X_u^4 = C)\mathbb{P}(X_u^1 = A|X_u^4 = C) \\
+ \; & 3\mathbb{P}(X_u^3 = A|X_u^5 = C)\mathbb{P}(X_u^4 = A|X_u^5 = C)\mathbb{P}(X_u^2 = A|X_u^4 = A)\mathbb{P}(X_u^1 = A|X_u^4 = A) \\
+ \; & 3\mathbb{P}(X_u^3 = A|X_u^5 = C)\mathbb{P}(X_u^4 = C|X_u^5 = C)\mathbb{P}(X_u^2 = A|X_u^4 = C)\mathbb{P}(X_u^1 = A|X_u^4 = C) \\
+ \; & 6\mathbb{P}(X_u^3 = A|X_u^5 = G)\mathbb{P}(X_u^4 = C|X_u^5 = G)\mathbb{P}(X_u^2 = A|X_u^4 = C)\mathbb{P}(X_u^1 = A|X_u^4 = C)).
\end{aligned} \tag{A.4}
$$

Cada uma das probabilidades da expressão (A.4) deve ser calculada individualmente, levando em conta os comprimentos de ramos da árvore F. Assim temos que

$$
\begin{aligned}
& \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A) \\
= \; & \frac{1}{4}\left(\mathbb{P}(Y_{0.5} = A|Y_0 = A)\mathbb{P}(Y_{0.2} = A|Y_0 = A)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = A)\,\mathbb{P}(Y_{0.5} = A|Y_{0.2} = A)\right. \\
+ \; & 3\mathbb{P}(Y_{0.5} = A|Y_0 = A)\mathbb{P}(Y_{0.2} = C|Y_0 = A)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = C)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = C) \\
+ \; & 3\mathbb{P}(Y_{0.5} = A|Y_0 = C)\mathbb{P}(Y_{0.2} = A|Y_0 = C)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = A)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = A) \\
+ \; & 3\mathbb{P}(Y_{0.5} = A|Y_0 = C)\mathbb{P}(Y_{0.2} = C|Y_0 = C)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = C)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = C) \\
+ \; & \left. 6\mathbb{P}(Y_{0.5} = A|Y_0 = G)\mathbb{P}(Y_{0.2} = C|Y_0 = G)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = C)\mathbb{P}(Y_{0.5} = A|Y_{0.2} = C)\right).
\end{aligned} \tag{A.5}
$$

De acordo com as probabilidades de transição do processo dadas pela matriz (2.12) podemos reescrever a expressão (A.5) como

$$
\begin{aligned}
& \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A) \\
= \; & \frac{1}{4}((1-3\alpha_{0.5})(1-3\alpha_{0.2})(1-3\alpha_{0.3})(1-3\alpha_{0.3}) + 3(1-3\alpha_{0.5})\alpha_{0.2}\;\alpha_{0.3}\;\alpha_{0.3} \\
+ \; & 3\alpha_{0.5}(1-3\alpha_{0.2})\alpha_{0.3}\;\alpha_{0.3} \quad + 3\alpha_{0.5}\alpha_{0.2}(1-3\alpha_{0.3})(1-3\alpha_{0.3}) + 6\,\alpha_{0.5}\;\alpha_{0.2}\;\alpha_{0.3}\;\alpha_{0.3}).
\end{aligned} \tag{A.6}
$$

Utilizando a expressão (2.13) obtemos $\mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A)$ em função do parâmetro α do modelo JC69. Entretanto, é comum que os métodos de estimativa de filogenias utilizem uma versão padronizada dos modelos de substituição de bases em que a taxa de mutação geral é 1 (ver Felsenstein, 2004). Dessa forma, são os comprimentos dos ramos que indicam a quantidade de evolução ocorrida naquela linhagem. Assumindo essa padronização e considerando que a taxa de mutação do processo JC69 é 3α temos

Tabela 1: Probabilidades das possíveis combinações de bases em um sítio da amostra segundo o modelo JC69 e número de combinações em cada grupo de igual probabilidade.

Tipo de Evento	Nº de Combinações	Probabilidade
$X_u^1 = A, X_u^2 = A, X_u^3 = A$	4	0.0780
$X_u^1 = A, X_u^2 = A, X_u^3 = C$	12	0.0229
$X_u^1 = A, X_u^2 = C, X_u^3 = A$	24	0.0113
$X_u^1 = A, X_u^2 = C, X_u^3 = G$	24	0.0059

que $\alpha = 1/3$, e com isso podemos encontrar α_τ, dado pela expressão (2.13), para cada tempo τ dado pela Figura A.1. Assim, temos que

$$\alpha_{0.5} = \frac{1}{4}(1 - e^{-4\times\frac{1}{3}\times 0.5}) \approx 0.1217 \tag{A.7}$$

$$\alpha_{0.3} = \frac{1}{4}(1 - e^{-4\times\frac{1}{3}\times 0.3}) \approx 0.0824 \tag{A.8}$$

$$\alpha_{0.2} = \frac{1}{4}(1 - e^{-4\times\frac{1}{3}\times 0.2}) \approx 0.0585. \tag{A.9}$$

Dessa forma, obtemos a probabilidade desejada, dada por

$$\mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A) \approx 0.0780. \tag{A.10}$$

Observe que realizamos exatamente o procedimento descrito na Seção 3.1.1 deste trabalho para o cálculo de $\mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau})$. Como conhecemos a filogenia F e os comprimentos dos ramos $\bar{\tau}$, recaímos na expressão (3.1) apresentada na Seção 3.1.1, dada por

$$\mathbb{P}(X_u^1, \cdots, X_u^N | F, \bar{\tau}) = \sum_{i^{N+1}, \cdots, i^{2N-1}} \pi_{i^{2N-1}} \prod_{k=N+1}^{2N-2} \mathbb{P}(i^k | i^{h(k)}, \tau_k) \prod_{l=1}^{N} \mathbb{P}(X_u^l | i^{h(l)}, \tau_l).$$

Note que, para o modelo JC69, $\mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A) = \mathbb{P}(X_u^1 = i, X_u^2 = i, X_u^3 = i)$, para qualquer $i \in E$.

Ainda nos resta calcular a probabilidade das outras combinações. Este cálculo é feito utilizando o mesmo procedimento apresentado para a combinação $\{X_u^1 = A, X_u^2 =$

$A, X_u^3 = A\}$, e seus resultados estão apresentados na Tabela 1.

Ao todo são possíveis 64 combinações de bases em X_u. Entretanto, devido a peculiaridades da filogenia F e ao fato do modelo JC69 não distinguir entre as probabilidades de mutação, essas combinações podem ser reunidas em apenas quatro grupos com probabilidades distintas. Se considerarmos outros modelos ou filogenias com número de nós externos maior, o número de probabilidades a serem calculadas aumentará.

Através das probabilidades acima, podemos finalmente calcular a verossimilhança das seqüências segundo o modelo JC69. Para isso utilizamos a expressão (3.2), dada por

$$L(F, \bar{\tau}|\mathbf{X}) = L(F, \bar{\tau}|X^1, \cdots, X^N) = \prod_{u=1}^{S} \mathbb{P}(X_u^1, \cdots, X_u^N|F, \bar{\tau}). \tag{A.11}$$

Para simplificar a notação da função de verossimilhança definimos a estatística S_{AAA} que conta o número de sítios da amostra em que aparece a combinação $\{X_u^1 = A, X_u^2 = A, X_u^3 = A\}$, ou outras de igual probabilidade. Da mesma forma, sejam S_{ACA}, S_{AAC} e S_{ACG} o número de sítios da amostra em que aparecem combinações de base que apresentam, respectivamente, as mesmas probabilidades de $\{X_u^1 = A, X_u^2 = C, X_u^3 = A\}$, $\{X_u^1 = A, X_u^2 = A, X_u^3 = C\}$ e $\{X_u^1 = A, X_u^2 = C, X_u^3 = G\}$. Assim podemos escrever (3.2) como

$$\begin{aligned} L(F, \bar{\tau}|\mathbf{X}) &= \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A)^{S_{AAA}} \mathbb{P}(X_u^1 = A, X_u^2 = C, X_u^3 = A)^{S_{ACA}} \\ &\times \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = C)^{S_{AAC}} \mathbb{P}(X_u^1 = A, X_u^2 = C, X_u^3 = G)^{S_{ACG}}. \end{aligned} \tag{A.12}$$

Suponhamos que na amostra $\mathbf{X}$ as combinações estejam distribuídas nos sítios de tal forma que

$$S_{AAA} = 354; \qquad S_{ACA} = 310; \qquad S_{AAC} = 268; \qquad S_{ACG} = 68.$$

Assim, podemos calcular o logaritmo da função de verossimilhança para estes dados segundo o modelo JC69, que é dado por

$$\log\left(L(F, \bar{\tau}|\mathbf{X})\right) \approx -3653.348. \tag{A.13}$$

Função de Verossimilhança para K80

Agora vamos fazer o mesmo procedimento utilizando o modelo K80, para obter o valor da função de verossimilhança das seqüências de DNA sob este modelo. Entretanto, para o cálculo de $\mathbb{P}(X_u^1, X_u^2, X_u^3)$, utilizaremos o algoritmo apresentado na Seção 3.1.1.

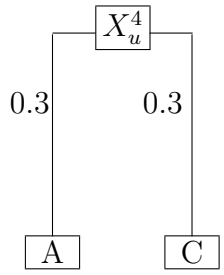


Figura A.2: Sub-árvore formada pelos nós externos 1 e 2 e pelo nó interno 4, na posição hipotética u.

Vamos primeiro calcular $\mathbb{P}(X_u^1 = A, X_u^2 = C, X_u^3 = A)$. Inicialmente, explicitamos os vetores $(\mathbb{P}(L_k|i)_u)_{i\in E}$ para os nós externos, ou seja $k = 1, 2, 3$. Como temos que,

$$\mathbb{P}(L_k|i)_u = I(X_u^k = i), \tag{A.14}$$

em que $I(A)$ representa a função indicadora do conjunto A, notamos que,

$$(\mathbb{P}(L_1|i)_u)_{i\in E} = (1, 0, 0, 0) \tag{A.15}$$

$$(\mathbb{P}(L_2|i)_u)_{i\in E} = (0, 0, 1, 0) \tag{A.16}$$

$$(\mathbb{P}(L_3|i)_u)_{i\in E} = (1, 0, 0, 0). \tag{A.17}$$

Em seguida, consideramos a sub-árvore formada pelos nós externos 1 e 2 e pelo nó interno 4, representada na Figura A.2, para calcular $\mathbb{P}(L_4|i)_u$.

Para calcular $\mathbb{P}(L_4|A)_u$ notamos que

$$\mathbb{P}(L_k|i)_u = \sum_{j,l\in E} \mathbb{P}(j|i, \tau_n)\mathbb{P}(L_n|j)_u\mathbb{P}(l|i, \tau_m)\mathbb{P}(L_m|l)_u, \tag{A.18}$$

em que n e m são os nós descendentes do nó k. Assim, temos que

$$\mathbb{P}(L_4|A)_u = \sum_{j,l\in E} \mathbb{P}(j|A, 0.3)\mathbb{P}(L_1|j)_u\mathbb{P}(l|C, 0.3)\mathbb{P}(L_2|l)_u, \tag{A.19}$$

Como 1 e 2 são nós externos temos $\mathbb{P}(L_1|j)_u = 0$, para todo $j \neq A$, e $\mathbb{P}(L_2|l)_u = 0$, para todo $l \neq C$, a expressão acima se reduz a

$$\mathbb{P}(L_4|A)_u = \mathbb{P}(A|A,0.3)\mathbb{P}(L_1|A)_u\mathbb{P}(A|C,0.3)\mathbb{P}(L_2|C)_u = \mathbb{P}(A|A,0.3)\mathbb{P}(C|A,0.3). \quad \text{(A.20)}$$

Observando a matriz de probabilidades de transição do processo K80 dada pela expressão (2.15), podemos explicitar as probabilidades da expressão (A.20). Assim, obtemos a seguinte expressão,

$$\begin{aligned}\mathbb{P}(L_4|A)_u &= (1-2\gamma_{0.3}-\beta_{0.3})(\gamma_{0.3}) \\ &= \left(1-\frac{1}{2}(1-e^{-4\gamma\times 0.3})-\frac{1}{4}(1+e^{2(\beta+\gamma)\times 0.3})\right)\left(\frac{1}{4}(1-e^{-4\gamma\times 0.3})\right) \quad \text{(A.21)}\end{aligned}$$

em que γ e β são os parâmetros do modelo K80, $\gamma_{0.3}$ é a probabilidade de que uma determinada *transversão* ocorra em $\tau = 0.3$, e $\beta_{0.3}$ é a probabilidade de uma *transição* em $\tau = 0.3$.

Utilizamos a mesma padronização realizada para o modelo JC69, de forma que a taxa geral de mutação seja 1, ou seja $2\gamma + \beta = 1$. Dessa forma, podemos reescrever os parâmetros γ e β do modelo K80 em função da razão de *transições* e *transversões* $K = \frac{\beta}{\gamma}$, obtendo

$$\gamma = \frac{1}{K+2} \quad \text{e} \quad \beta = \frac{K}{K+2}. \quad \text{(A.22)}$$

O valor de K deve ser estimado a partir dos dados. Supondo que o valor estimado de K seja 10, com base nas expressões de (A.22) podemos calcular a probabilidade na expressão (A.21), obtendo

$$\mathbb{P}(L_4|A)_u \approx 0.0182. \quad \text{(A.23)}$$

Fazendo o mesmo procedimento para as outras três bases obtemos o vetor

$$(\mathbb{P}(L_4|i)_u)_{i\in E} \approx (0.0182,\ \ 0.0045,\ \ 0.0182,\ \ 0.0045). \quad \text{(A.24)}$$

Agora seguiremos para o nó imediatamente anterior, ou seja o nó 5. Para calcular $\mathbb{P}(L_5|A)_u$ devemos considerar os vetores (A.17) e (A.24) e a sub-árvore representada na Figura A.3.

Diferentemente de $(\mathbb{P}(L_k|i)_u)_{i\in E}$ para $k = 1, 2, 3$, $(\mathbb{P}(L_4|i)_u)_{i\in E}$ é um vetor com todas as entradas não nulas. Dessa forma, o cálculo de $\mathbb{P}(L_5|i)_u$ não pode ser simplificado da forma como feito com $\mathbb{P}(L_4|i)_u$. Assim, temos que

$$\begin{aligned}\mathbb{P}(L_5|A)_u &= \sum_{j,l\in E}\mathbb{P}(j|A,0.5)\mathbb{P}(L_3|j)_u\mathbb{P}(l|A,0.2)\mathbb{P}(L_4|l)_u \\ &= \mathbb{P}(A|A,0.5)\sum_{l\in E}\mathbb{P}(l|A,0.2)\mathbb{P}(L_4|l)_u\end{aligned}$$

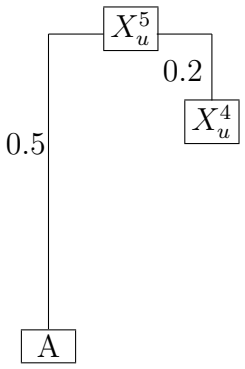


Figura A.3: Sub-árvore formada pelo nó externo 3 e pelos nós internos 4 e 5, na posição hipotética u.

e ainda,

$$\begin{aligned}\mathbb{P}(L_5|A)_u &= (1-2\gamma_{0.5}-\beta_{0.5})(1-2\gamma_{0.2}-\beta_{0.2})\mathbb{P}(L_4|A)_u + \beta_{0.2}\mathbb{P}(L_4|G)_u \\ &+ \gamma_{0.2}\mathbb{P}(L_4|C)_u + \gamma_{0.2}\mathbb{P}(L_4|T)_u) \approx 0.0106,\end{aligned}$$

considerando $K = 10$. Utilizando o mesmo procedimento com as outras três bases, temos que

$$(\mathbb{P}(L_5|i)_u)_{i\in E} \approx (0.0107,\ \ 0.0017,\ \ 0.0006,\ \ 0.0003). \tag{A.25}$$

Assim, podemos calcular o valor da função de verossimilhança da combinação $(X_u^1 = A, X_u^2 = C, X_u^3 = A)$, que é dado por

$$L(F, \bar{\tau}|X_u^1 = A, X_u^2 = C, X_u^3 = A) = \sum_{i\in E} \mathbb{P}(L_5|i)_u \pi_i = \frac{1}{4}\sum_{i\in E} \mathbb{P}(L_5|i)_u, \tag{A.26}$$

já que o modelo K80 assume distribuição homogênea de bases. Portanto, temos que

$$L(F, \bar{\tau}|X_u^1 = A, X_u^2 = C, X_u^3 = A) \approx 0.0033. \tag{A.27}$$

Este mesmo procedimento deve ser repetido para todas as outras possíveis combinações de bases em $(X_u^1, X_u^2, X_u^3,)$. Os resultados desses cálculos estão apresentados na Tabela 2.

Agora queremos escrever a função de verossimilhança da amostra $\mathbf{X}$. Sejam S_{AAA}, S_{AAC}, S_{AAG}, S_{ACA}, S_{AGA}, S_{AGC} e S_{ACG} os números de sítios com combinações de base que apresentam, respectivamente, as mesmas probabilidades de $\{X_u^1 = A, X_u^2 = A, X_u^3 = A\}$,

Tabela 2: Probabilidade das possíveis combinações de bases em um sítio da amostra, segundo o modelo K80, e número de combinações em cada grupo de igual probabilidade.

Tipo de Evento	Nº de Combinações	Probabilidade
$X_u^1 = A, X_u^2 = A, X_u^3 = A$	4	0.0885
$X_u^1 = A, X_u^2 = A, X_u^3 = C$	8	0.0082
$X_u^1 = A, X_u^2 = A, X_u^3 = G$	4	0.0504
$X_u^1 = A, X_u^2 = C, X_u^3 = A$	16	0.0033
$X_u^1 = A, X_u^2 = G, X_u^3 = A$	8	0.0322
$X_u^1 = A, X_u^2 = G, X_u^3 = C$	8	0.0039
$X_u^1 = A, X_u^2 = C, X_u^3 = G$	16	0.0024

$\{X_u^1 = A, X_u^2 = A, X_u^3 = C\}$, $\{X_u^1 = A, X_u^2 = A, X_u^3 = G\}$, $\{X_u^1 = A, X_u^2 = C, X_u^3 = A\}$, $\{X_u^1 = A, X_u^2 = G, X_u^3 = A\}$, $\{X_u^1 = A, X_u^2 = G, X_u^3 = C\}$ e $\{X_u^1 = A, X_u^2 = C, X_u^3 = G\}$. Assim podemos escrever (3.2) como

$$\begin{aligned} L(F, \bar{\tau}|\mathbf{X}) &= \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = A)^{S_{AAA}} \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = C)^{S_{AAC}} \\ &\times \mathbb{P}(X_u^1 = A, X_u^2 = A, X_u^3 = G)^{S_{AAG}} \mathbb{P}(X_u^1 = A, X_u^2 = C, X_u^3 = A)^{S_{ACA}} \\ &\times \mathbb{P}(X_u^1 = A, X_u^2 = G, X_u^3 = A)^{S_{AGA}} \mathbb{P}(X_u^1 = A, X_u^2 = G, X_u^3 = C)^{S_{AGC}} \\ &\times \mathbb{P}(X_u^1 = A, X_u^2 = C, X_u^3 = G)^{S_{ACG}}. \end{aligned} \tag{A.28}$$

Note que o modelo K80 apresenta um número maior de grupos de combinações de bases com probabilidades distintas do que o modelo JC69. Isso é uma conseqüência natural do fato de o modelo K80 designar probabilidades diferentes para *transições* e *transversões*, enquanto o modelo JC69 trata todas as mutações de forma igual.

Suponhamos que na amostra $\mathbf{X}$ as combinações estejam distribuídas nos sítios de forma que

$$S_{AAA} = 354; \quad S_{AAG} = 202; \quad S_{AAC} = 66; \quad S_{ACA} = 53;$$

$$S_{AGA} = 257; \quad S_{AGC} = 31; \quad S_{ACG} = 37.$$

Assim, podemos calcular o logaritmo da função de verossimilhança para os dados, segundo o modelo K80, que é dado por

$$\log\left(L(F, \bar{\tau}|\mathbf{X})\right) \approx -3361.126. \tag{A.29}$$

Aplicação do Teste JC69 $\times$ K80

Na análise acima mostramos apenas como se obtém o valor da função de verossimilhança da amostra $\mathbf{X}$ para uma filogenia com topologia e comprimentos de ramos conhecidos. Essa é uma situação hipotética, na maioria dos casos tais informações (especialmente os tempos $\bar{\tau}$) dificilmente estão disponíveis ou são confiáveis. Entretanto, ainda dentro do exemplo, desejamos aplicar o teste de razão de verossimilhança, ver expressão (3.13), para descobrir qual o melhor modelo para as seqüências $\mathbf{X}$.

Para tanto, é necessário maximizar as funções de verossimilhança sob os modelos JC69 e K80. Neste exemplo, conhecemos todos os parâmetros da filogenia. Então, as funções de verossimilhança devem apenas ser otimizadas em relação aos parâmetros livres dos modelos de substituição de bases. Para o modelo JC69, devido à restrição de que a taxa geral de mutação é 1, não restam parâmetros livres para que se faça uma otimização, e assim o máximo da função de verossimilhança é o mesmo valor calculado em (A.13). Já no caso do modelo K80, a função de verossimilhança deve ser maximizada em relação ao parâmetro K (taxa de *transições* e *transversões*). Para o cálculo de tal função, neste exemplo, tivemos que supor que o valor de K foi estimado. Se utilizarmos o valor de K estimado pelo método da máxima verossimilhança, então, por definição, a função atingirá seu máximo neste ponto. Desta forma, temos que

H_0 : Pressupostos A1, A2, A3 e $\mathbf{Q} = \mathbf{Q}_{\text{JC69}}$, $\mathbf{p}_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$;

H_1 : Pressupostos A1, A2, A3 e $\mathbf{Q} = \mathbf{Q}_{\text{K80}}$, $\mathbf{p}_0 = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$,

onde os pressupostos A1, A2 e A3 são os mesmos apresentados na Seção 3.2. Assim, utilizando as expressões (A.13) e (A.29), temos que

$$\Delta(\mathbf{X}) = \log\left(\hat{L}_{\text{K80}}(\mathbf{X})\right) - \log\left(\hat{L}_{\text{JC69}}(\mathbf{X})\right) \approx -3653.348 + 3361.126 \approx -292.222, \quad \text{(A.30)}$$

onde $\hat{L}_{\text{JC69}}(\mathbf{X})$ é o valor da função de verossimilhança para o modelo JC69, e $\hat{L}_{\text{K80}}(\mathbf{X})$ é o valor da função de verossimilhança para o modelo K80, calculado no valor de K, estimado pelo método da máxima verossimilhança. Dessa forma, a estatística do teste $-2\Delta(\mathbf{X})$ assume valor 584.444.

Para calcular o p-valor devemos comparar o valor de $-2\Delta(\mathbf{X})$ obtido com a distribuição de $-2\Delta(\mathbf{X})$ sob H_0 (ver Yang, 2007). Conforme apresentado na Seção 3.2, a distribuição assintótica teórica de $-2\Delta(\mathbf{X})$ é χ^2_r, em que r representa os graus de liberdade e é dado pela diferença de parâmetros livres dos modelos de H_0 e H_1. Assim $r = 2 - 1 = 1$. Note que da comparação de $-2\Delta(\mathbf{X})$ com a distribuição χ^2_1 temos que

$$\text{p-valor} = \mathbb{P}(-2\Delta(\mathbf{X}) < \chi^2_1) < 0.0001. \quad \text{(A.31)}$$

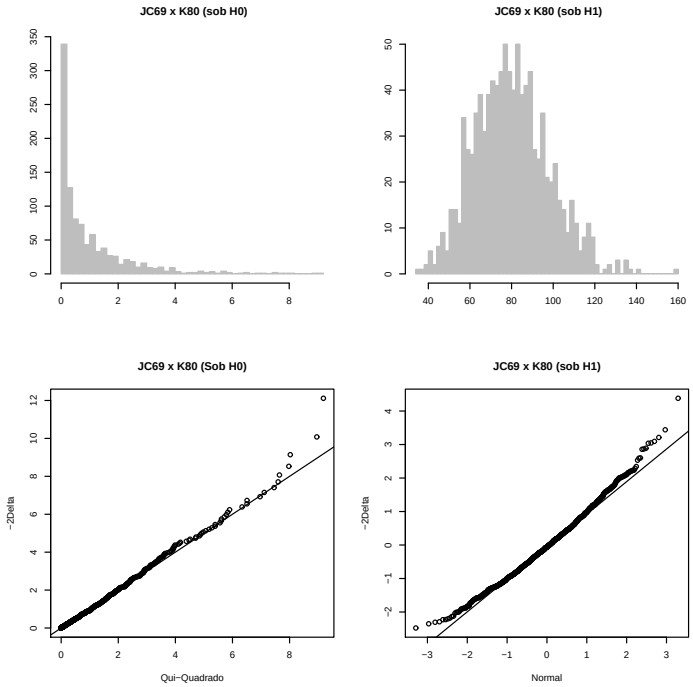


Figura A.4: Histograma e Q-Q plots dos valores de $-2\Delta(\mathbf{X})$ obtidos por simulações de Monte Carlo sob H_0 e H_1.

Desta forma, temos forte evidência de que o modelo K80 descreve melhor a evolução da amostra $\mathbf{X}$, e rejeitamos H_0. Note que, se quisermos realizar um teste de nível de confiança $\lambda = 0.01$ (isto é, $\mathbb{P}(H_1 \text{ é aceita } | H_0 \text{ é verdadeira}) = 0.01$), com base na distribuição assintótica teórica, devemos comparar $-2\Delta(\mathbf{X}) = 584.444$ com o valor 6.64, obtido da distribuição χ^2_1. Neste caso também temos forte evidência de que o modelo K80 é mais adequado ao nível de 99% de confiança.

Por outro lado, podemos utilizar o bootstrap paramétrico, descrito na Seção 3.2.1, para obter a distribuição da estatística do teste sob H_0. E, neste caso, utilizando 1000 repetições para as simulações de Monte Carlo com filogenia fixa F, temos que o valor crítico para o nível de confiança 99% é de 6.5070. O p-valor obtido é também menor do que 0.001, de forma que, com base nesse método, também temos fortes indícios para rejeitar H_0.

Utilizando o método proposto na Seção 3.2.3 podemos obter o poder do teste da razão de verossimilhança que compara os modelo JC69 e K80 para a amostra $\mathbf{X}$ com base na distribuição obtida pelas simulações. Utilizamos a filogenia F e o valor estimado da taxa de *transições* e *transversões* $K = 10$, com 1000 repetições. Assim temos que o poder do teste é 1. A Figura A.4 apresenta histogramas e Q-Q plots da distribuição de $-2\Delta(\mathbf{X})$ obtida pelas simulações sob H_0 e H_1. Nos Q-Q plots, a distribuição de $-2\Delta(\mathbf{X})$, sob H_0, é comparada com a χ^2_1 e a distribuição de $-2\Delta(\mathbf{X})$, sob H_1, é comparada com a distribuição

Tabela 3: Resultado do Teste do Exemplo.

Teste	$-2\Delta(\mathbf{X})$	Distribuição Teórica		Simulação de Monte Carlo			Decisão
		r	p-valor	p-valor	Valor Crítico 99%	Poder	
JC69 × K80	584.444	1	4.0497×10^{-129}	< 0.0001	6.5070	1	rejeita

normal.

Os resultados obtidos para o teste de hipóteses estão apresentados na Tabela 3.

Apêndice B

Apresentamos aqui uma demonstração para o Teorema 5.2 dado na Seção 5.1.

Teorema B.1. (Método Delta). *Seja* $\mathbf{T}_n = (T_1, \cdots, T_n)'$ *vetor de estimadores com distribuição assintótica normal multivariada com média* $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_n)'$ *e matriz de variâncias-covariâncias* $\boldsymbol{\Sigma}$. *Suponha que a função* $\mathbf{g}(t_1, \cdots, t_n) = (g_1(t_1, \cdots, t_n), \cdots, g_n(t_1, \cdots, t_n))$ *tem matriz de derivadas* $\mathbf{J}$ *em* $\mathbf{t} = \boldsymbol{\theta}$. *Então,*

$$\sqrt{n}(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{Z}, \quad \textit{quando } n \to \infty, \tag{B.1}$$

onde $\mathcal{Z} \sim \mathbf{N}(0, \mathbf{J}\ \boldsymbol{\Sigma}\ \mathbf{J}')$. *Assim,* $\mathbf{g}(\mathbf{T}_n)$ *tem variância assintótica normal multivariada com vetor de médias* $\mathbf{g}(\boldsymbol{\theta})$ *e matriz de variâncias-covariâncias* $\mathbf{J}\ \boldsymbol{\Sigma}\ \mathbf{J}'$.

Demonstração: A expansão em série de Taylor de $\mathbf{g}(\mathbf{T}_n)$ em torno de $\boldsymbol{\theta}$ é dada por

$$\mathbf{g}(\mathbf{T}_n) = \mathbf{g}(\boldsymbol{\theta}) + (\mathbf{T_n} - \boldsymbol{\theta})'\mathbf{J} + O(\|\mathbf{T_n} - \boldsymbol{\theta}\|^2) \tag{B.2}$$

em que $\|\mathbf{z}\| = \sqrt{\sum z_i^2}$, e $O(z)$ é uma função tal que $\lim_{z\to 0} \frac{O(z)}{z}$ é uma constante e $\lim_{z\to 0} O(z) = 0$. Assim, rearranjando a expressão acima, temos

$$\sqrt{n}\,(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})) = \sqrt{n}(\mathbf{T_n} - \boldsymbol{\theta})'\mathbf{J} + \sqrt{n}O(\|\mathbf{T_n} - \boldsymbol{\theta}\|^2). \tag{B.3}$$

Mas note que por hipótese

$$\sqrt{n}[\mathbf{T_n} - \boldsymbol{\theta}] \to \mathbf{N}(0, \boldsymbol{\Sigma}). \tag{B.4}$$

Deste modo, temos que

$$\mathbb{E}\left(\sqrt{n}\,(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta}))\right) \approx \mathbb{E}\left(\sqrt{n}(\mathbf{T_n} - \boldsymbol{\theta})'\mathbf{J}\right) \to 0 \tag{B.5}$$

e

$$\text{Var}\left(\sqrt{n}\left(\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})\right)\right) \approx \text{Var}\left(\sqrt{n}(\mathbf{T_n} - \boldsymbol{\theta})'\mathbf{J}\right) \to \mathbf{J}\ \boldsymbol{\Sigma}\ \mathbf{J}'. \tag{B.6}$$

Assim, pelo Teorema de Slutzky, temos que, quando $n \to \infty$,

$$\sqrt{n}[\mathbf{g}(\mathbf{T}_n) - \mathbf{g}(\boldsymbol{\theta})] \to \mathbf{N}(0, \mathbf{J}\ \boldsymbol{\Sigma}\ \mathbf{J}'). \tag{B.7}$$

Apêndice C

Apresentamos aqui os histogramas omitidos da Seção 6.1. Os histogramas apresentam a distribuição obtida para a estatística do teste $-2\Delta(\mathbf{X})$ apresentada em (3.13), sob H_0 e sob H_1, definidas em (3.12). Estes histogramas são para os testes de hipóteses JC69×K80, JC69×F81, K80×HKY85, F81×HKY85 e HKY85×GTR. Os quatro últimos testes de hipóteses foram estudados no caso em que $\mathbf{p}_0$ é estimado tanto pelo método da máxima verossimilhança quanto pelo método dos momentos. Os parâmetros utilizados nas simulações estão apresentados no texto da Seção 6.1 e os histogramas a seguir são correspondentes às árvores 1 e 2 apresentadas na Figura 6.1.

Os histogramas apresentados neste apêndice ratificam as observações feitas na Seção 6.1. Destacam-se as observações sobre a função de distribuição de $-2\Delta(\mathbf{X})$ sob H_0 para os testes de hipóteses JC69×F81 e K80×HKY85, quando utilizamos o estimador $\tilde{\mathbf{p}}_0$, que claramente não é uma qui-quadrado. Isso pode ser facilmente evidenciado pela presença de valores negativos para a estatística do teste. Outra observação evidenciada pelos histogramas desse apêndice é o fato de o poder do teste aumentar quando o número de sítios S de cada seqüência aumenta. Com o aumento de S, percebe-se que os valores de $-2\Delta(\mathbf{X})$, sob H_1, aumentam.

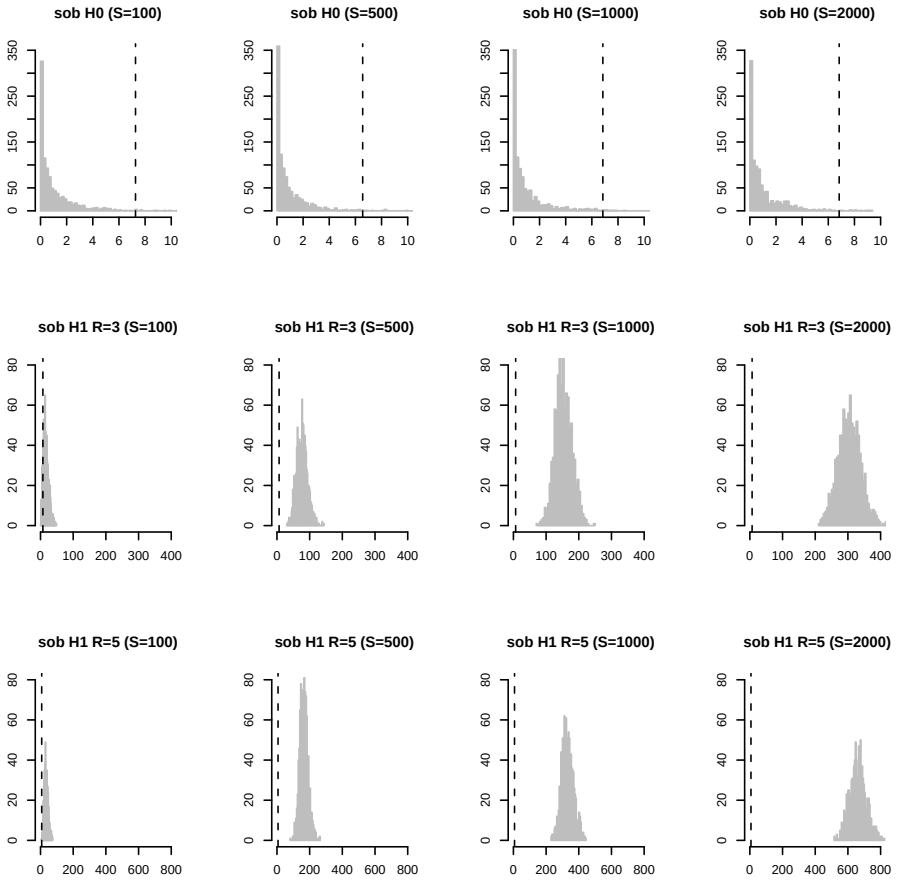


Figura C.1: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ K80, sob H_0 e sob H_1, para a Árvore 1.

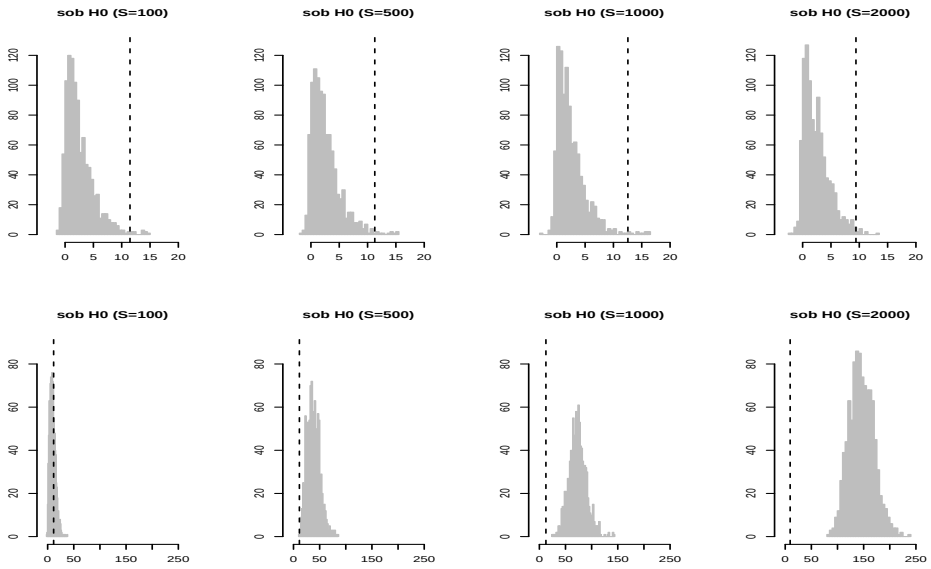


Figura C.2: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ F81, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 1.

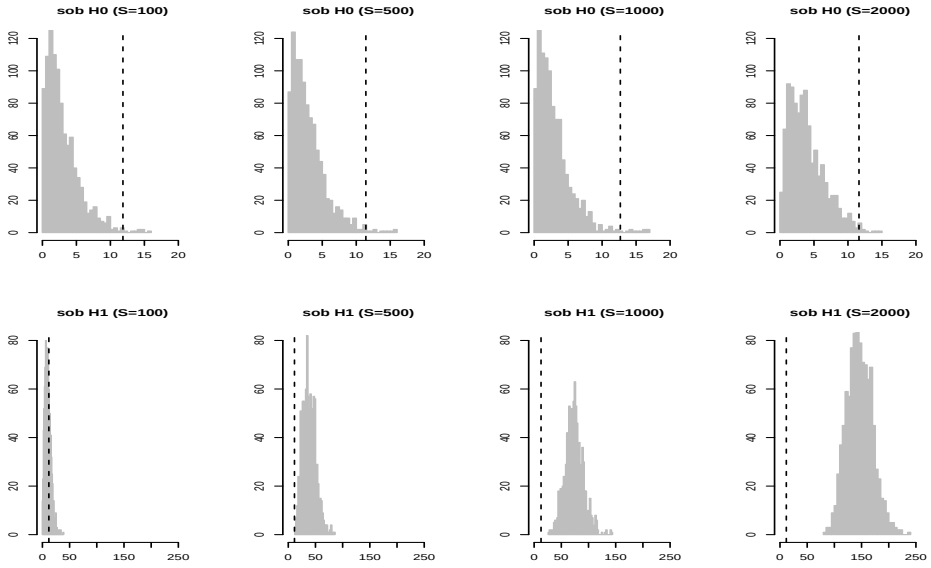


Figura C.3: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ F81, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 1.

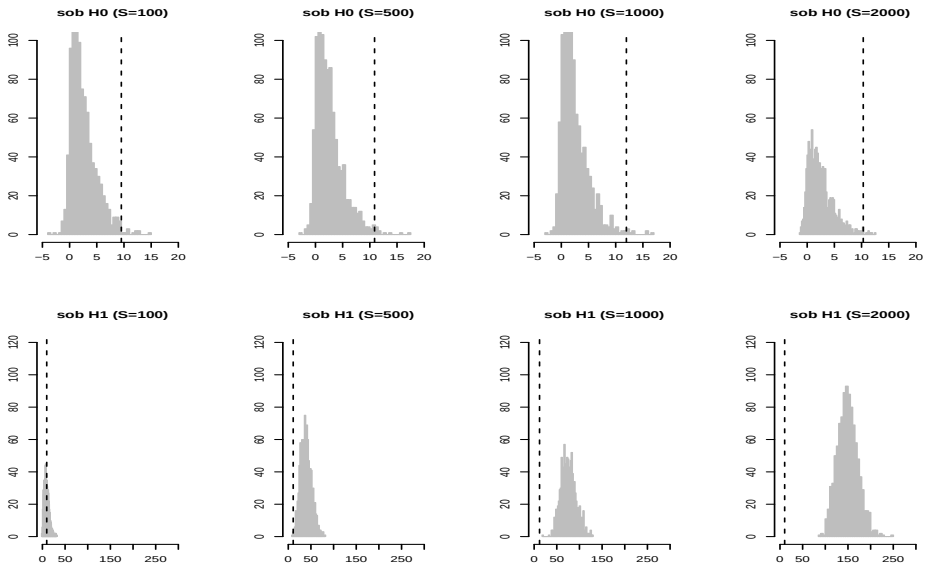


Figura C.4: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 $\times$ HKY85, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 1.

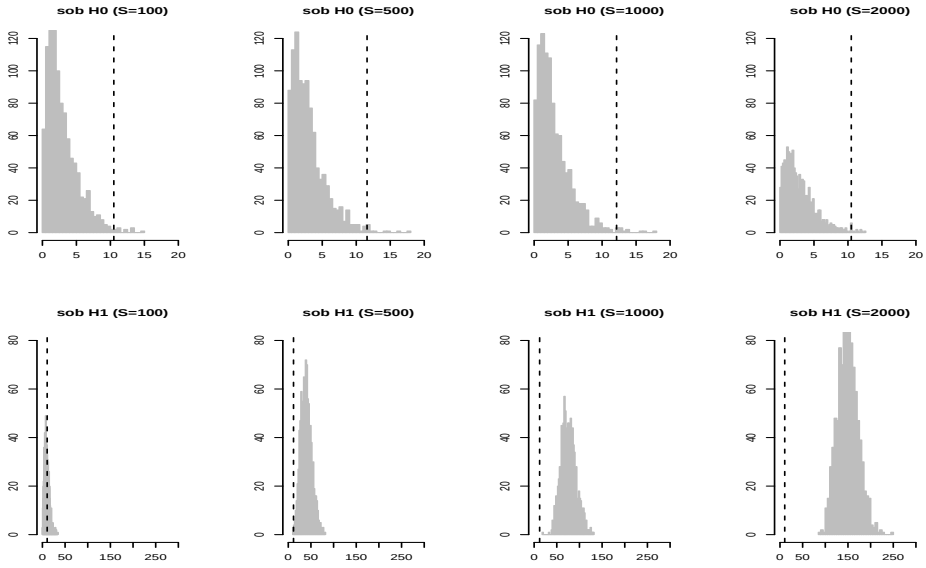


Figura C.5: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 $\times$ HKY85, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 1.

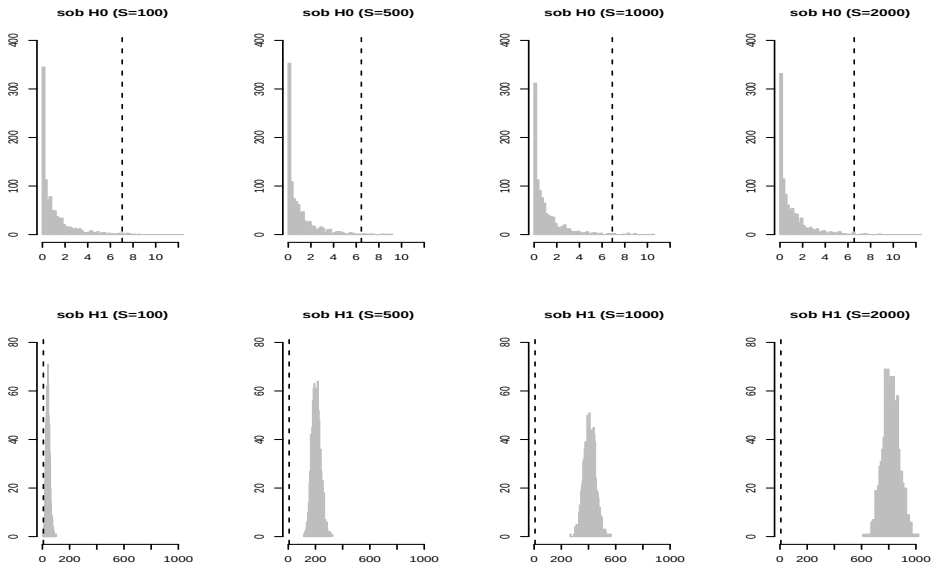


Figura C.6: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F81 $\times$ HKY85, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 1.

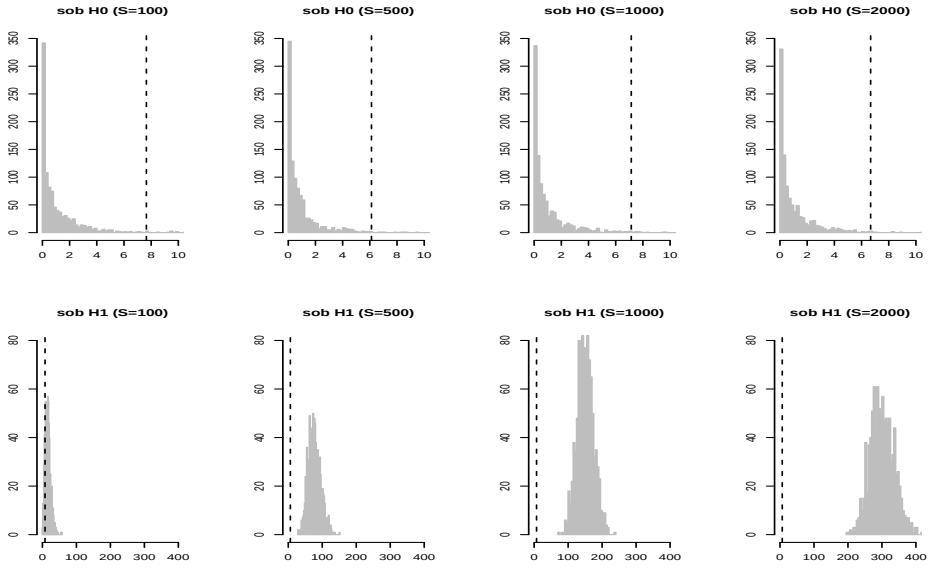


Figura C.7: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F81 $\times$ HKY85, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 1.

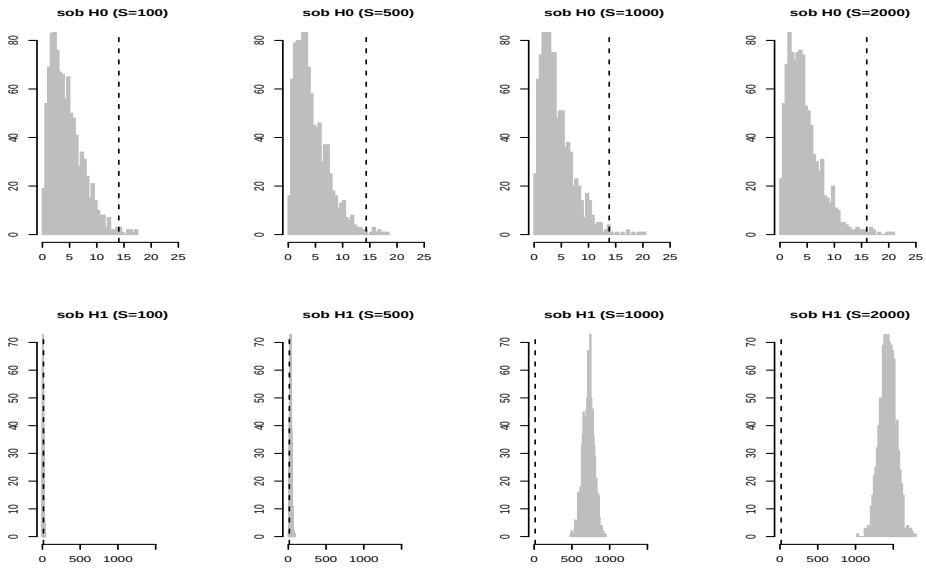


Figura C.8: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses HKY85 $\times$ GTR, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 1.

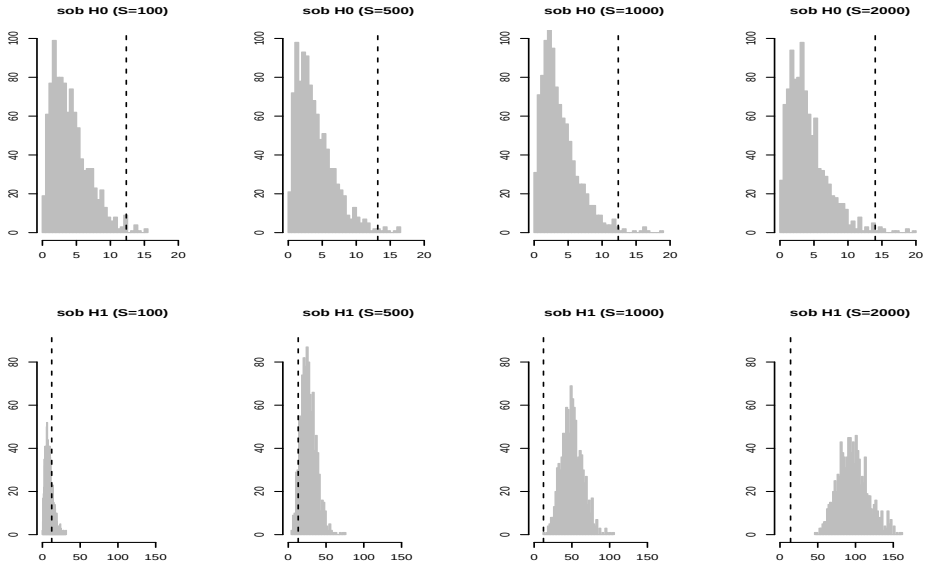


Figura C.9: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses HKY85 $\times$ GTR, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 1.

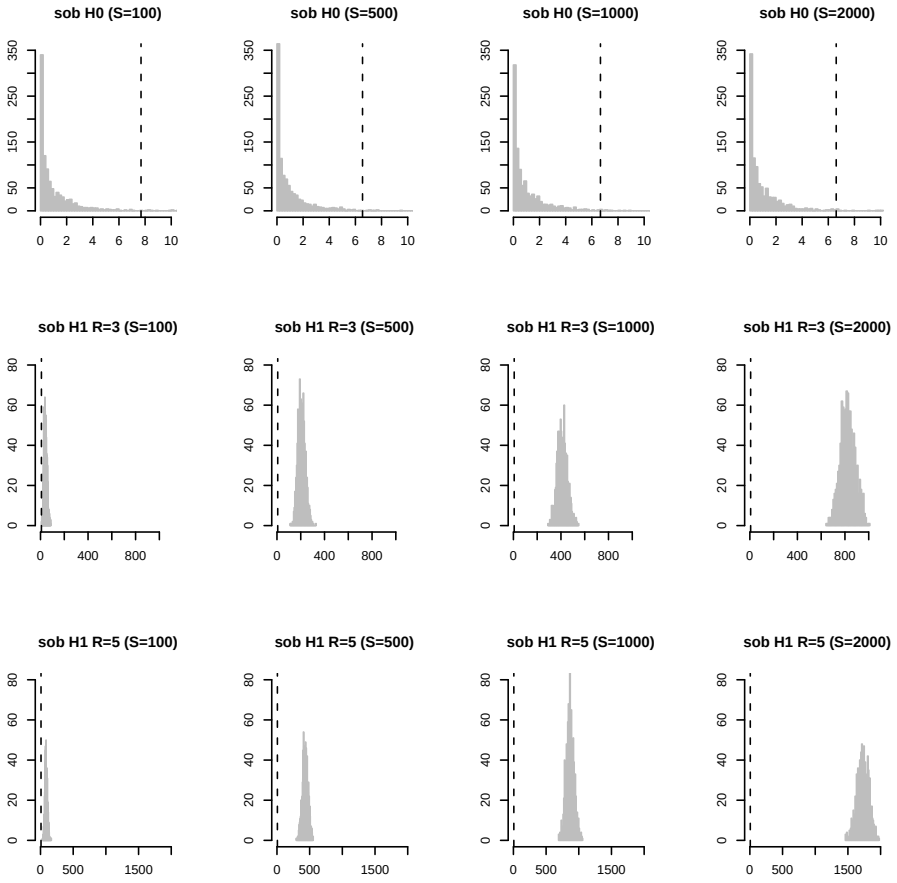


Figura C.10: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ K80, sob H_0 e sob H_1, para a Árvore 2.

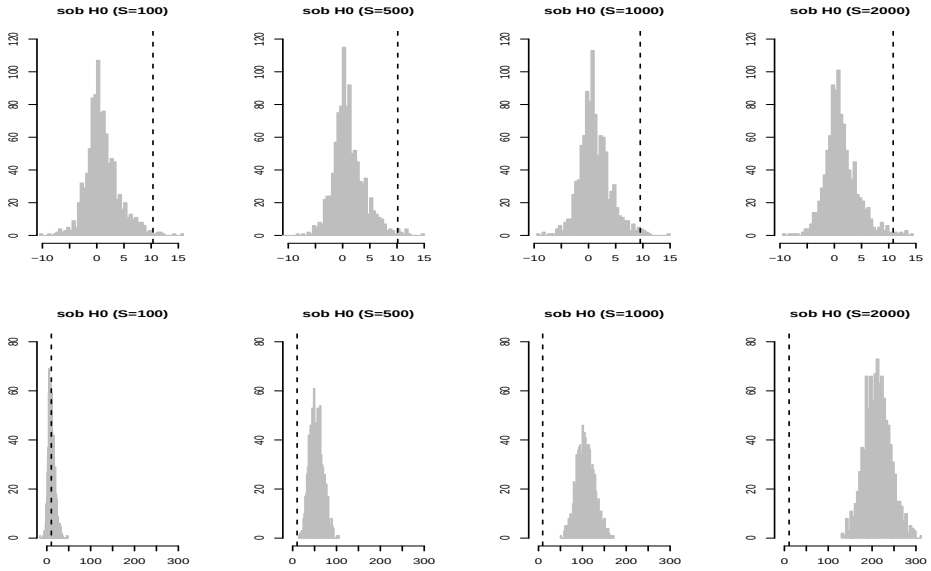


Figura C.11: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ F81, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 2.

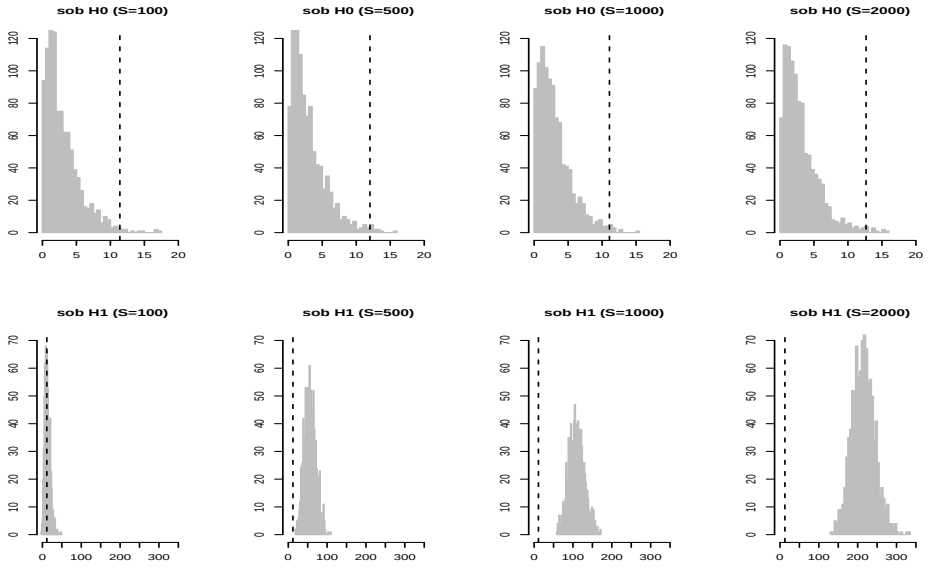


Figura C.12: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses JC69 $\times$ F81, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 2.

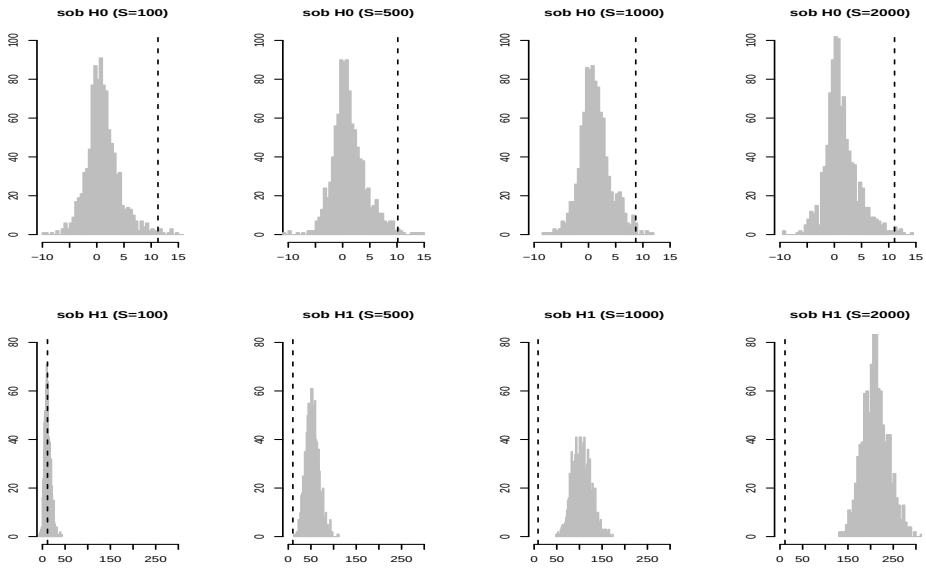


Figura C.13: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 $\times$ HKY85, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 2.

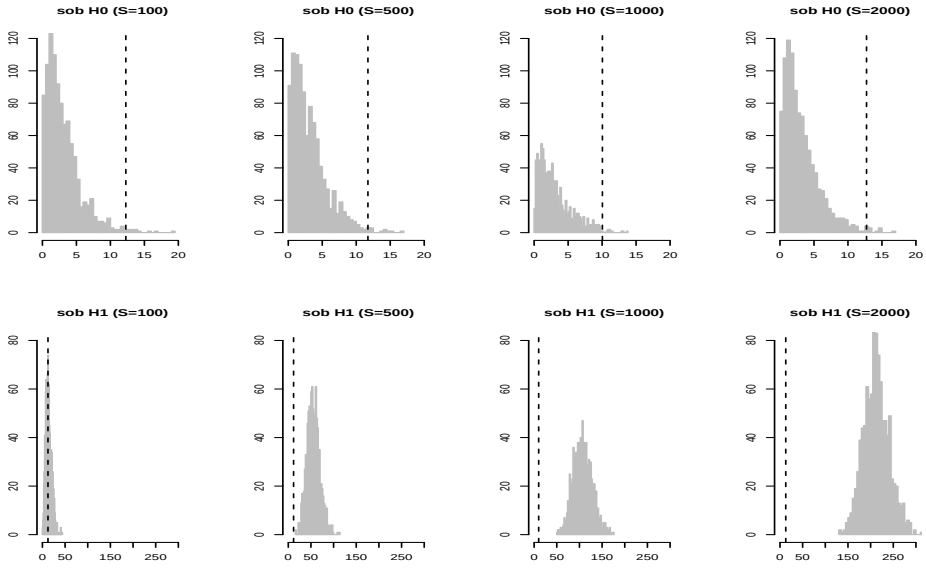


Figura C.14: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 $\times$ HKY85, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 2.

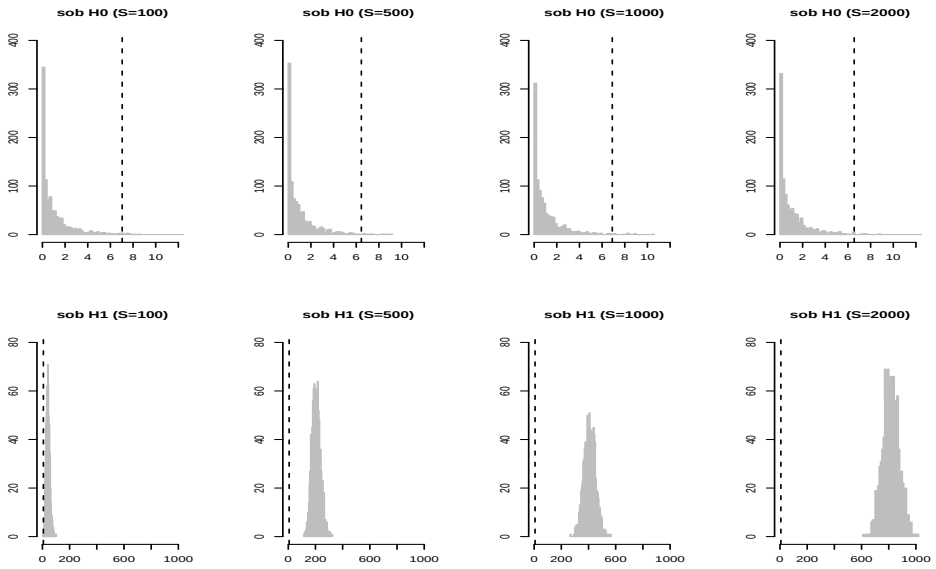


Figura C.15: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F81 $\times$ HKY85, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 2.

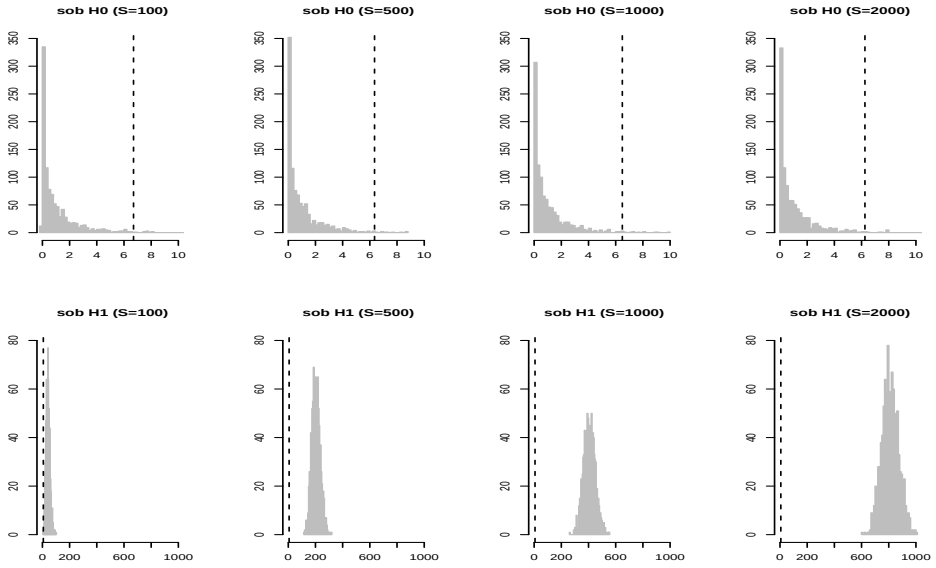


Figura C.16: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F81 $\times$ HKY85, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 2.

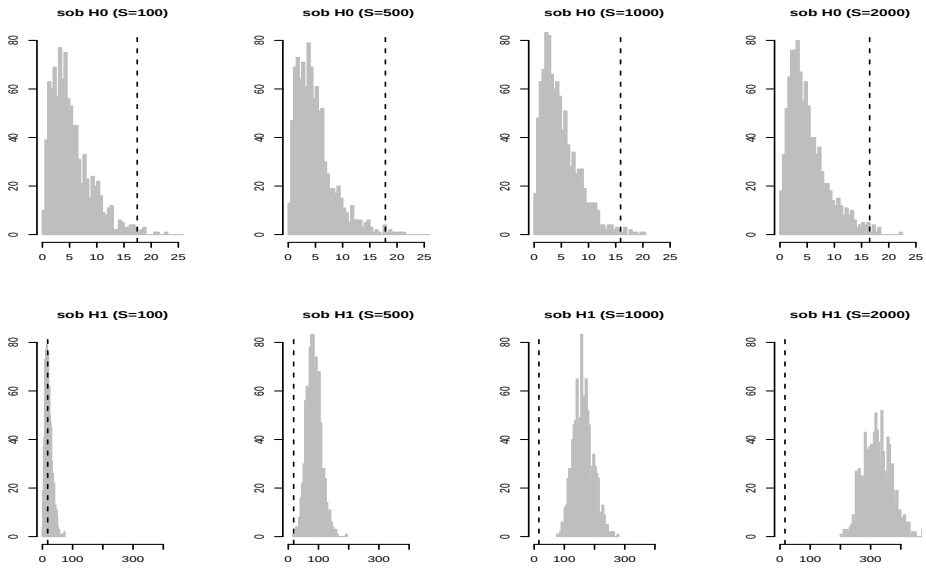


Figura C.17: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses HKY85 $\times$ GTR, sob H_0 e sob H_1, com $\tilde{\mathbf{p}}_0$, para a Árvore 2.

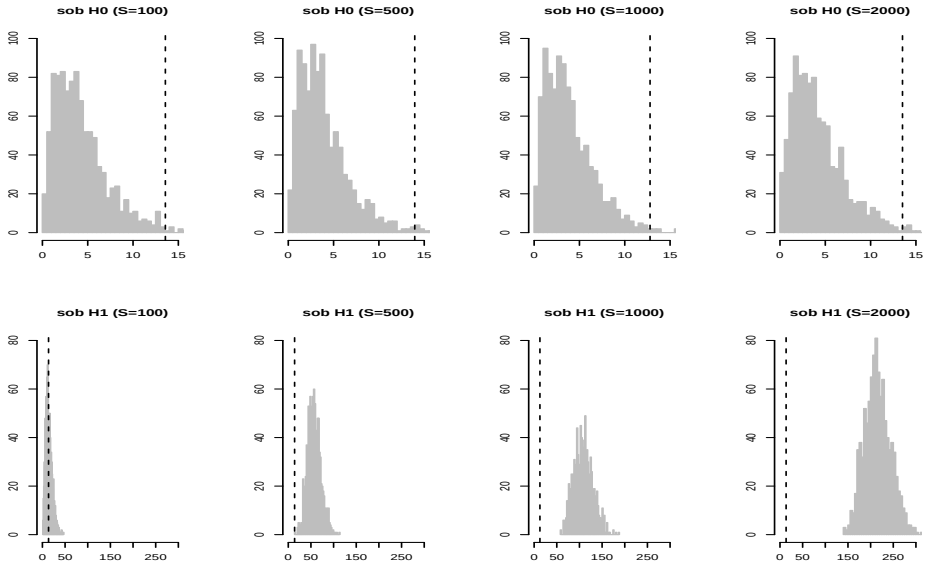


Figura C.18: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses HKY85 $\times$ GTR, sob H_0 e sob H_1, com $\hat{\mathbf{p}}_0$, para a Árvore 2.

Apêndice D

Apresentamos aqui os Q-Q plots e resultados dos testes de aderância omitidos da Seção 6.1, para os testes de hipóteses JC69×K80, JC69×F81, K80×HKY85, F81×HKY85 e HKY85×GTR. Os quatro últimos testes foram estudados no caso em que $\mathbf{p}_0$ é estimado tanto pelo método da máxima verossimilhança quanto pelo método dos momentos. Os parâmetros utilizados nas simulações estão apresentados no texto da Seção 6.1.

Os gráficos apresentam comparações de quantís entre as distribuições obtidas para a estatística do teste $-2\Delta(\mathbf{X})$, definida em (3.13), sob H_0, e as respectivas distribuições teóricas χ^2. Apresentam ainda as comparações de quantís entre as distribuições de $-2\Delta(\mathbf{X})$, obtidas sob H_1, e a distribuição normal. As figuras de D.1 a D.6 referem-se às simulações realizadas na árvore 1, com $\hat{\mathbf{p}}_0$ e $\tilde{\mathbf{p}}_0$. Já as figuras D.7 a D.12 referem-se às mesmas simulações realizadas na árvore 2

As Tabelas D.1 e D.2 apresentam a decisão dos testes de aderência qui-quadrado que tem como hipótese nula: $-2\Delta(\mathbf{X})$, sob H_0, tem distribuição χ^2 determinada pelo Teorema 3.1. Já as Tabelas D.3 e D.4 apresentam a decisão dos testes de normalidade Shapiro-Wilks, que tem como hipótese nula: $-2\Delta(\mathbf{X})$, sob H_1, tem distribuição normal. Em todas as tabelas, quando não aparece "aceito"na célula correspondente a um teste, significa que H_0 foi rejeitada.

Os gráficos neste apêndice ratificam o que foi mencionado na Seção 6.1 em relação à função de distribuição de $-2\Delta(\mathbf{X})$, sob H_0 e sob H_1. Notamos que a função de distribuição de $-2\Delta(\mathbf{X})$, sob H_0, é próxima à distribuição qui-quadrado prevista pelo Teorema 3.1 para todos os testes de hipóteses, com exceção de JC69×F81 e K80×HKY85 quando utilizamos $\tilde{\mathbf{p}}$. Nesses testes de hipóteses percebe-se um claro desvio da distribuição qui-quadrado na cauda esquerda. Já em relação à função de distribuição de $-2\Delta(\mathbf{X})$, sob H_1, nota-se que para $S \in \{100, 500, 1000\}$ a distribuição se aproxima da distribuição normal com o aumento de S. Para $S = 100$, a função de distribuição é distante da normal para a maioria dos testes de hipóteses analisados. Entretanto, nem sempre se verifica uma aproximação da distribuição normal quando S passa de 1000 para 2000.

O resultado dos testes de aderência reforçam o que foi comentado anteriormente. Para os testes que utilizam o estimador $\hat{\mathbf{p}}_0$, a estatística do teste sob H_0 tem distribuição qui-quadrado. E, entre os testes que utilizam o estimador $\tilde{\mathbf{p}}_0$, apenas aqueles que não testam para $\mathbf{p}_0$ apresentam distribuição qui-quadrado (embora não de maneira tão consistente quanto com $\hat{\mathbf{p}}_0$). Sob H_1, a estatística do teste apresenta distribuição normal apenas para alguns testes, e em geral, para valores grandes de S.

Tabela D.1: Decisão do Teste Qui-quadrado para $-2\Delta(\mathbf{X})$, sob H_0, Utilizando $\hat{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$	Distribuição Teórica
JC69 × K80	1	aceita	aceita	aceita	aceita	χ^2_1
	2	aceita	aceita	aceita	aceita	
	3	aceita	aceita	aceita	aceita	
JC69 × F81	1	aceita	aceita	aceita	-	χ^2_3
	2	aceita	aceita	aceita	aceita	
	3	aceita	aceita	aceita	aceita	
K80 × HKY85	1	aceita	aceita	aceita	aceita	χ^2_3
	2	aceita	aceita	aceita	aceita	
	3	aceita	aceita	aceita	aceita	
F81 × HKY85	1	aceita	-	aceita	aceita	χ^2_1
	2	aceita	aceita	aceita	aceita	
	3	aceita	aceita	aceita	aceita	
HKY85 × GTR	1	aceita	aceita	-	aceita	χ^2_4
	2	aceita	aceita	aceita	aceita	
	3	aceita	aceita	aceita	aceita	

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela D.2: Decisão do Teste Qui-quadrado para $-2\Delta(\mathbf{X})$, sob H_0, Utilizando $\tilde{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$	Distribuição Teórica
JC69 × F81	1	-	-	-	-	χ^2_3
	2	-	-	-	-	
	3	-	-	-	-	
K80 × HKY85	1	-	-	-	-	χ^2_3
	2	-	-	-	-	
	3	-	-	-	-	
F81 × HKY85	1	aceita	aceita	aceita	aceita	χ^2_1
	2	aceita	aceita	aceita	aceita	
	3	aceita	aceita	-	-	
HKY85 × GTR	1	aceita	aceita	aceita	aceita	χ^2_4
	2	-	-	-	-	
	3	aceita	aceita	-	aceita	

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela D.3: Decisão do Teste de Normalidade Shapiro-Wilks para $-2\Delta(\mathbf{X})$, sob H_1, Utilizando $\hat{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$
JC69 × K80 $K = 3$	1	-	-	aceita	aceita
	2	-	aceita	aceita	aceita
	3	-	aceita	-	aceita
JC69 × K80 $K = 5$	1	-	-	-	-
	2	-	aceita	aceita	aceita
	3	-	aceita	aceita	-
JC69 × F81	1	-	-	-	-
	2	-	-	-	-
	3	-	-	-	-
K80 × HKY85	1	-	-	-	-
	2	-	-	-	-
	3	-	-	-	-
F81 × HKY85	1	-	-	aceita	-
	2	-	-	aceita	aceita
	3	-	-	-	-
HKY85 × GTR	1	-	-	-	-
	2	-	-	-	-
	3	-	-	-	-

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela D.4: Decisão do Teste de Normalidade Shapiro-Wilks para $-2\Delta(\mathbf{X})$, sob H_1, Utilizando $\tilde{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$
JC69 × F81	1	-	-	-	-
	2	-	-	-	-
	3	-	-	-	-
K80 × HKY85	1	-	-	-	-
	2	-	-	-	-
	3	-	-	-	-
F81 × HKY85	1	-	-	aceita	-
	2	-	-	aceita	aceita
	3	-	-	-	-
HKY85 × GTR	1	-	-	aceita	aceita
	2	aceita	-	-	-
	3	aceita	-	-	-

Nota: Os testes foram realizados ao nível de 99% de confiança.

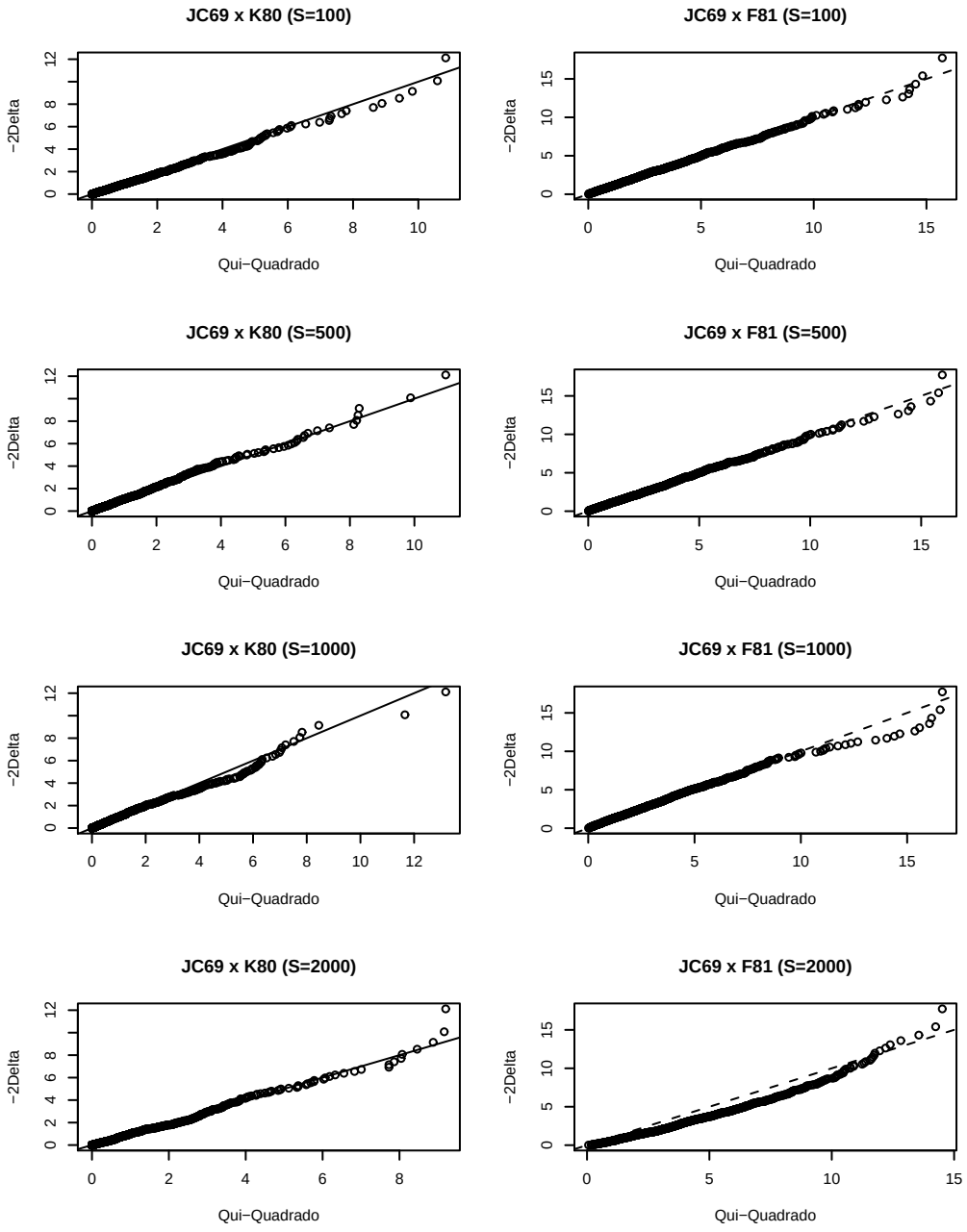


Figura D.1: Q-Q plot, sob H_0, para os Testes de Hipóteses JC69 $\times$ K80 e JC69 $\times$ F81 para a Árvore 1, Utilizando $\hat{\mathbf{p}}_0$.

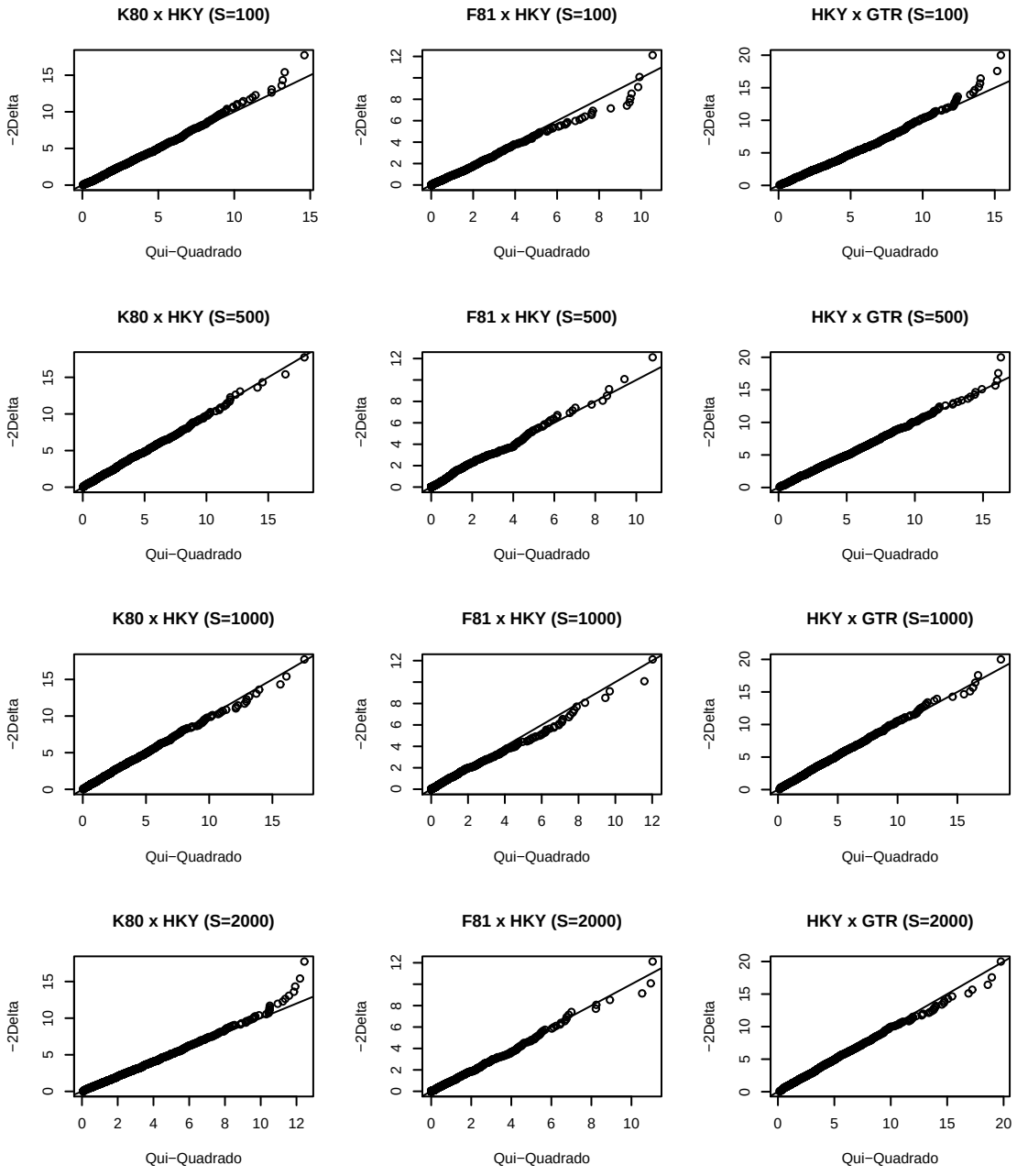


Figura D.2: Q-Q plot, sob H_0, para os Testes de Hipóteses K80 × HKY85, F81 × HKY85 e HKY85 × GTR para a Árvore 1, Utilizando $\hat{\mathbf{p}}_0$.

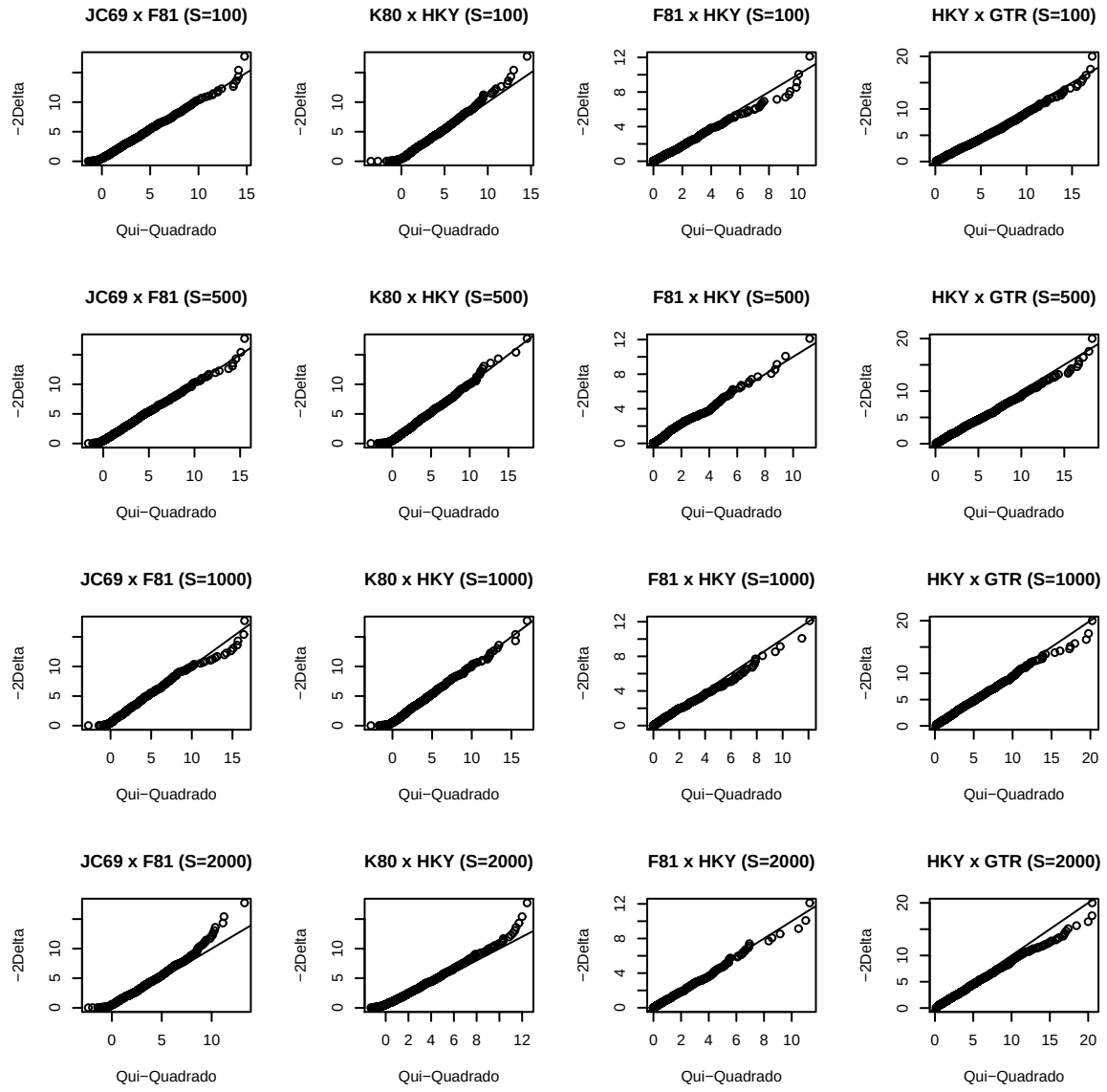


Figura D.3: Q-Q plot, sob H_0, para os Testes de Hipóteses JC69 × F81, K80 × HKY85, F81 × HKY85 e HKY85 × GTR para a Árvore 1, Utilizando $\tilde{\mathbf{p}}_0$.

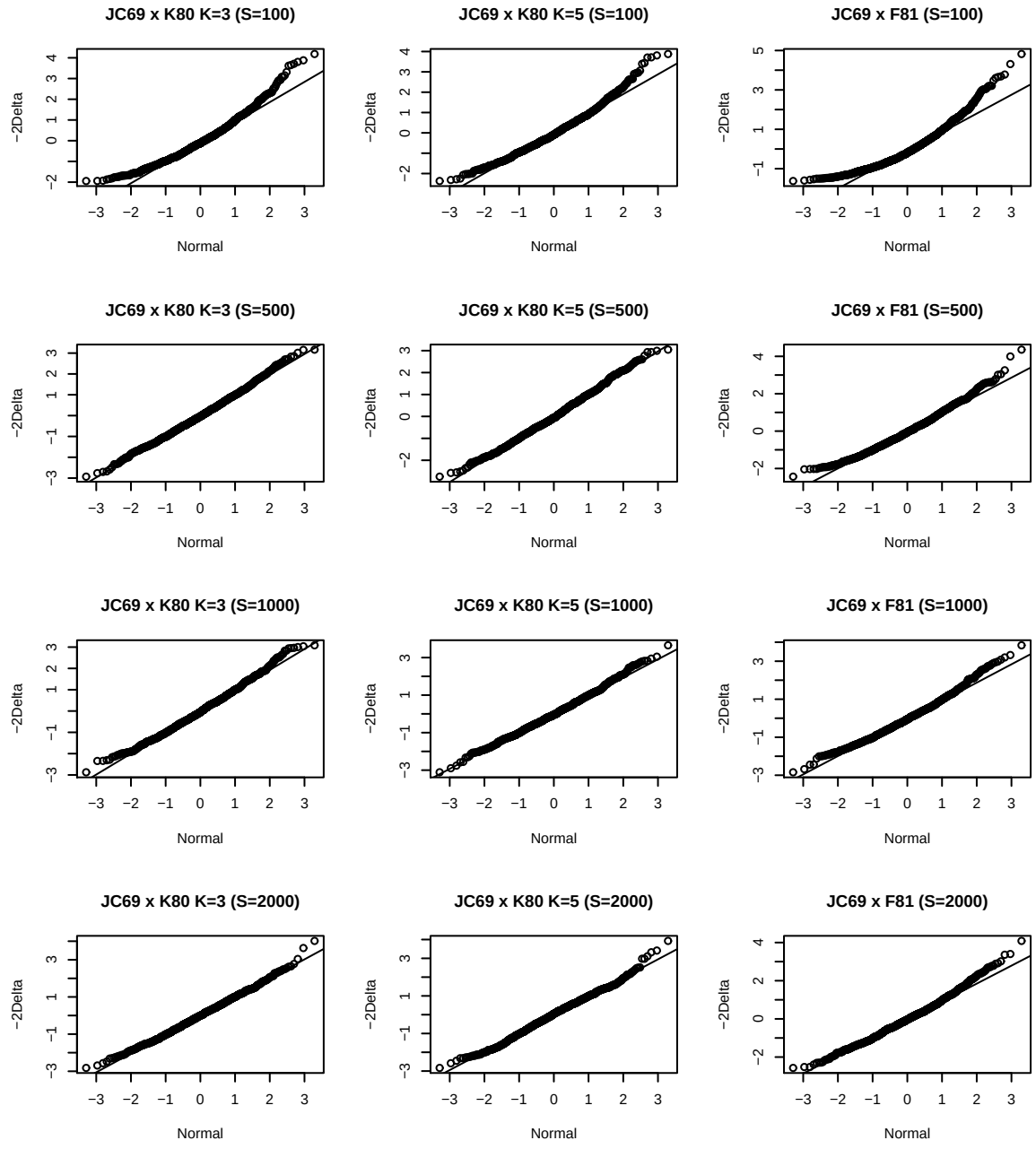


Figura D.4: Q-Q plot, sob H_1, para os Testes de Hipóteses JC69 × K80 e JC69 × F81 para a Árvore 1, Utilizando $\hat{\mathbf{p}}_0$.

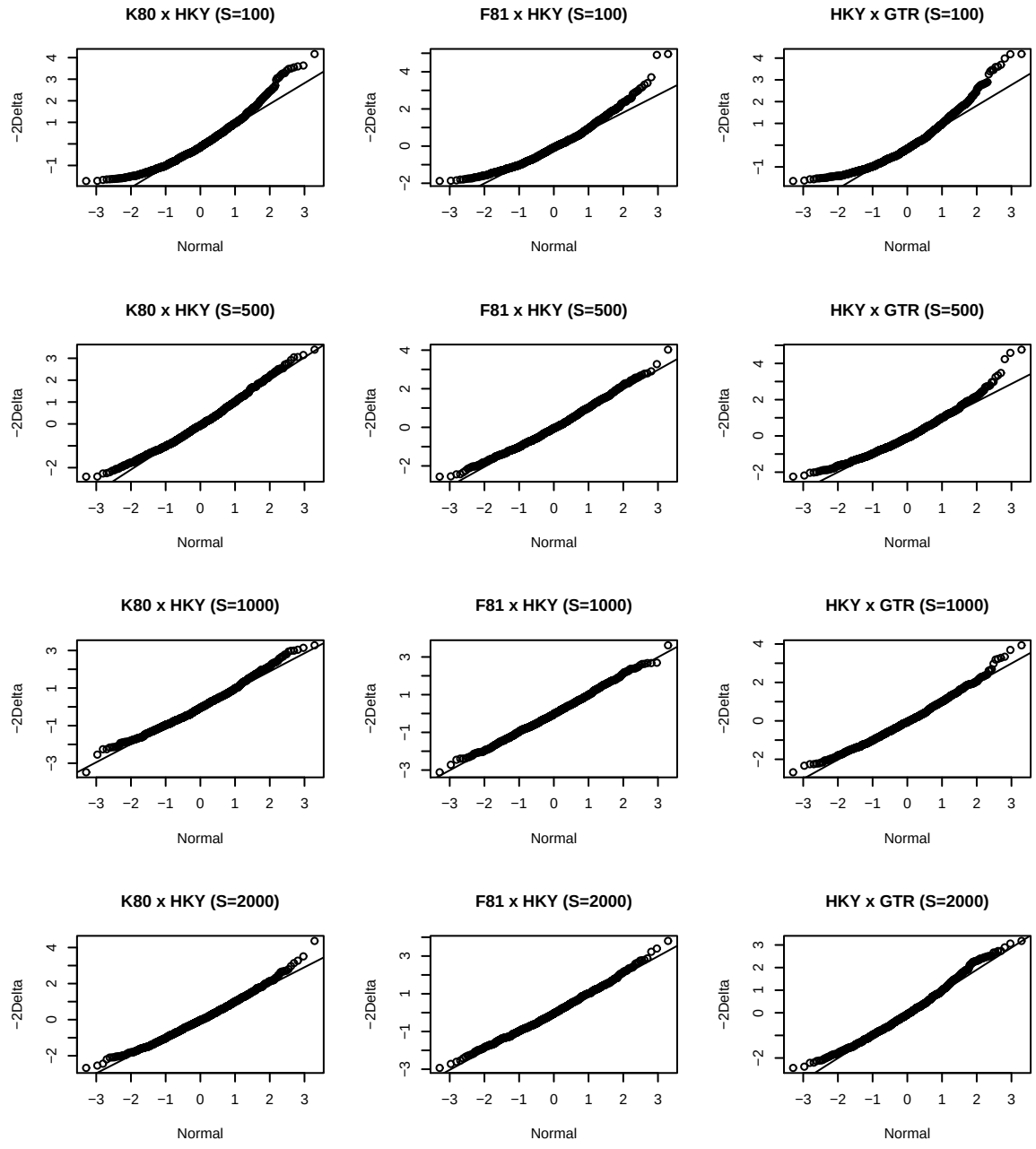


Figura D.5: Q-Q plot, sob H_1, para os Testes de Hipóteses K80 × HKY85, F81 × HKY85 e HKY85 × GTR para a Árvore 1, Utilizando $\hat{\mathbf{p}}_0$.

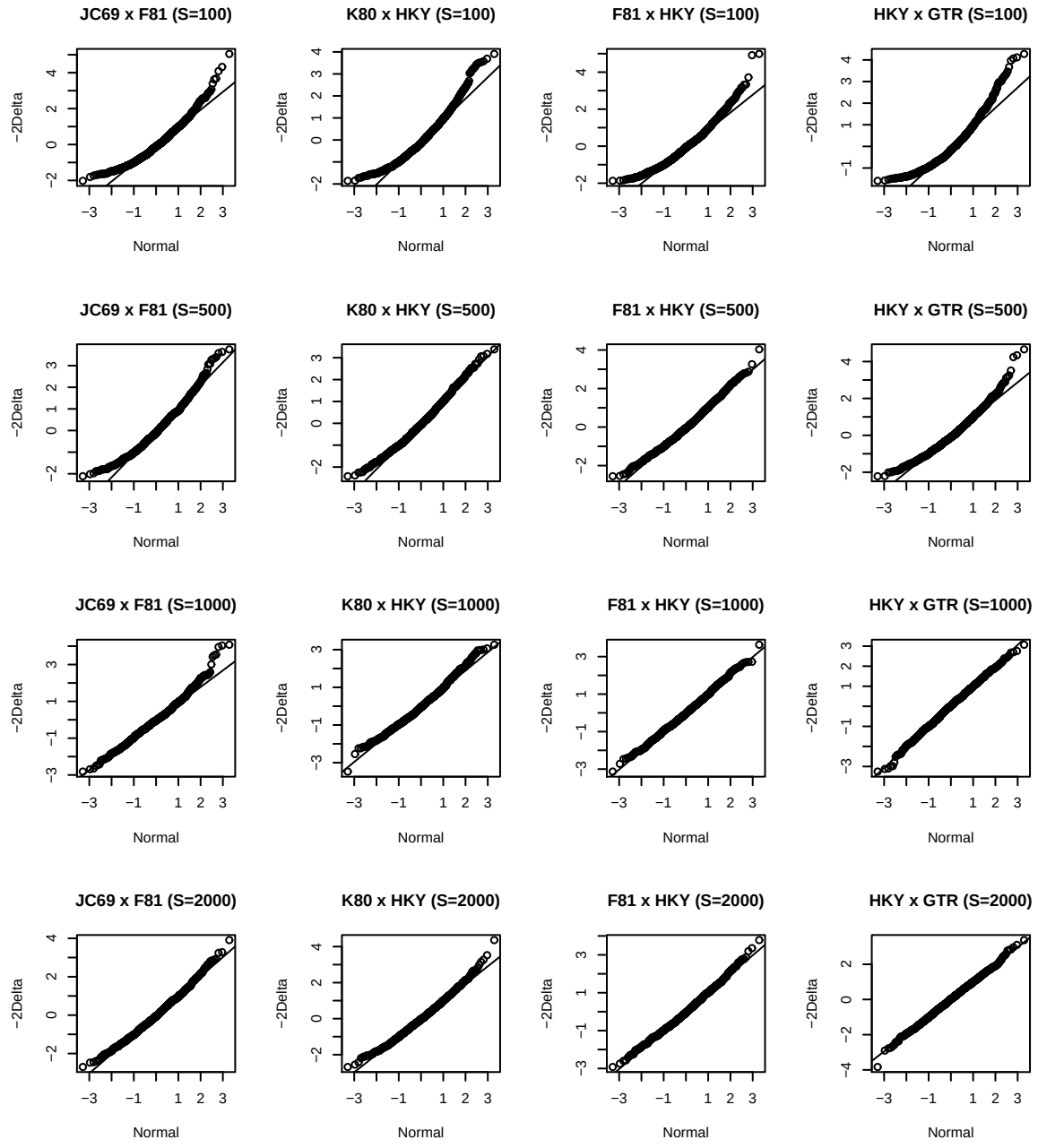


Figura D.6: Q-Q plot, sob H_1, para os Testes de Hipóteses JC69 × F81, K80 × HKY85, F81 × HKY85 e HKY85 × GTR para a Árvore 1, Utilizando $\tilde{\mathbf{p}}_0$.

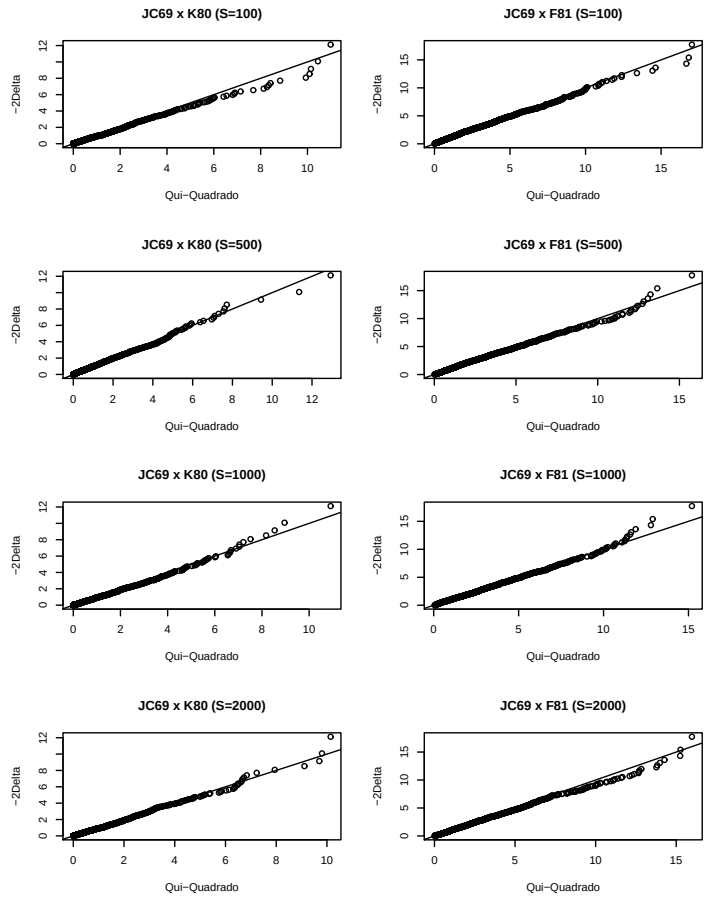


Figura D.7: Q-Q plot, sob H_0, para os Testes de Hipóteses JC69 $\times$ K80 e JC69 $\times$ F81 para a Árvore 2, Utilizando $\hat{\mathbf{p}}_0$.

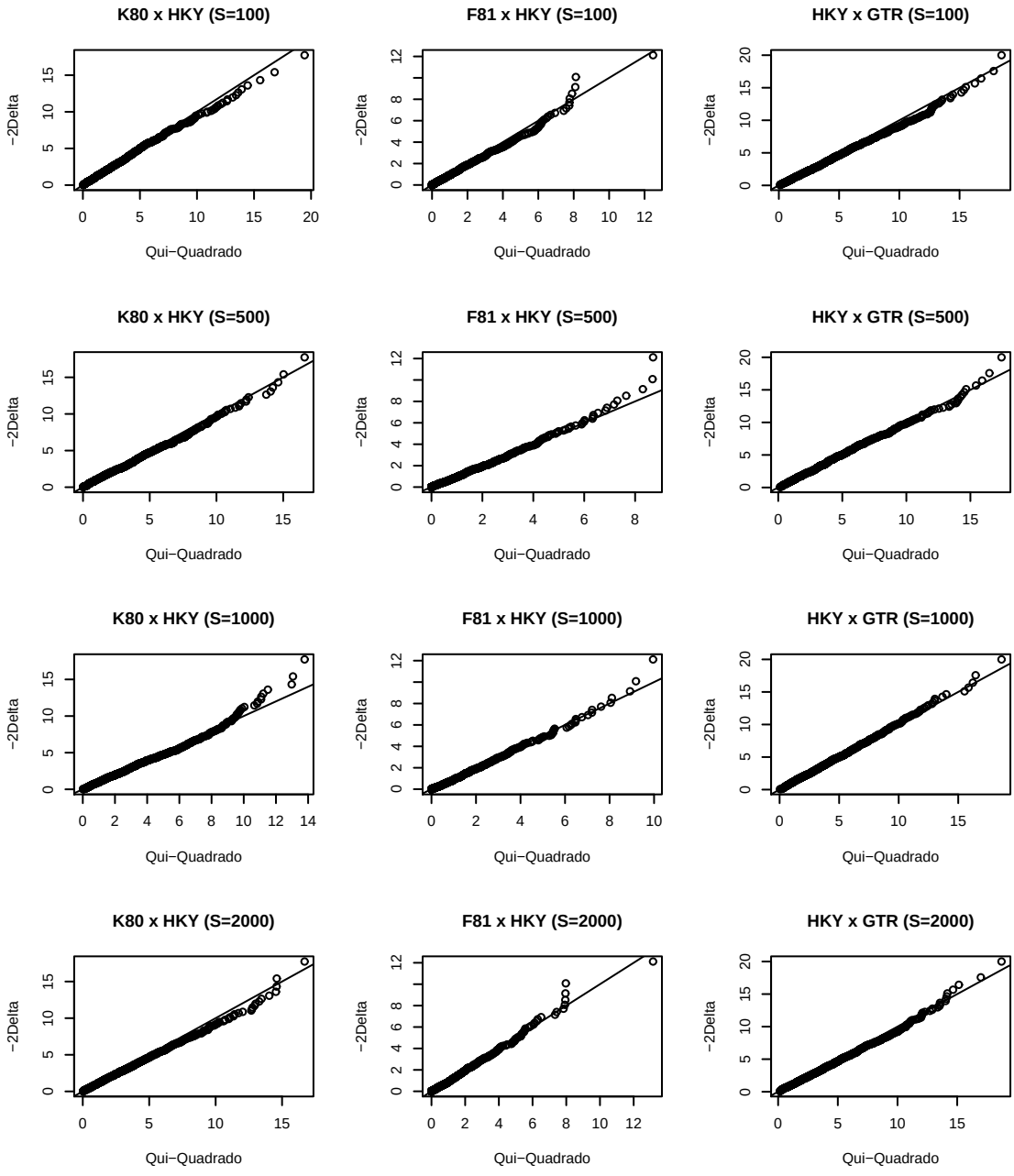


Figura D.8: Q-Q plot, sob H_0, para os Testes de Hipóteses K80 × HKY85, F81 × HKY85 e HKY85 × GTR para a Árvore 2, Utilizando $\hat{\mathbf{p}}_0$.

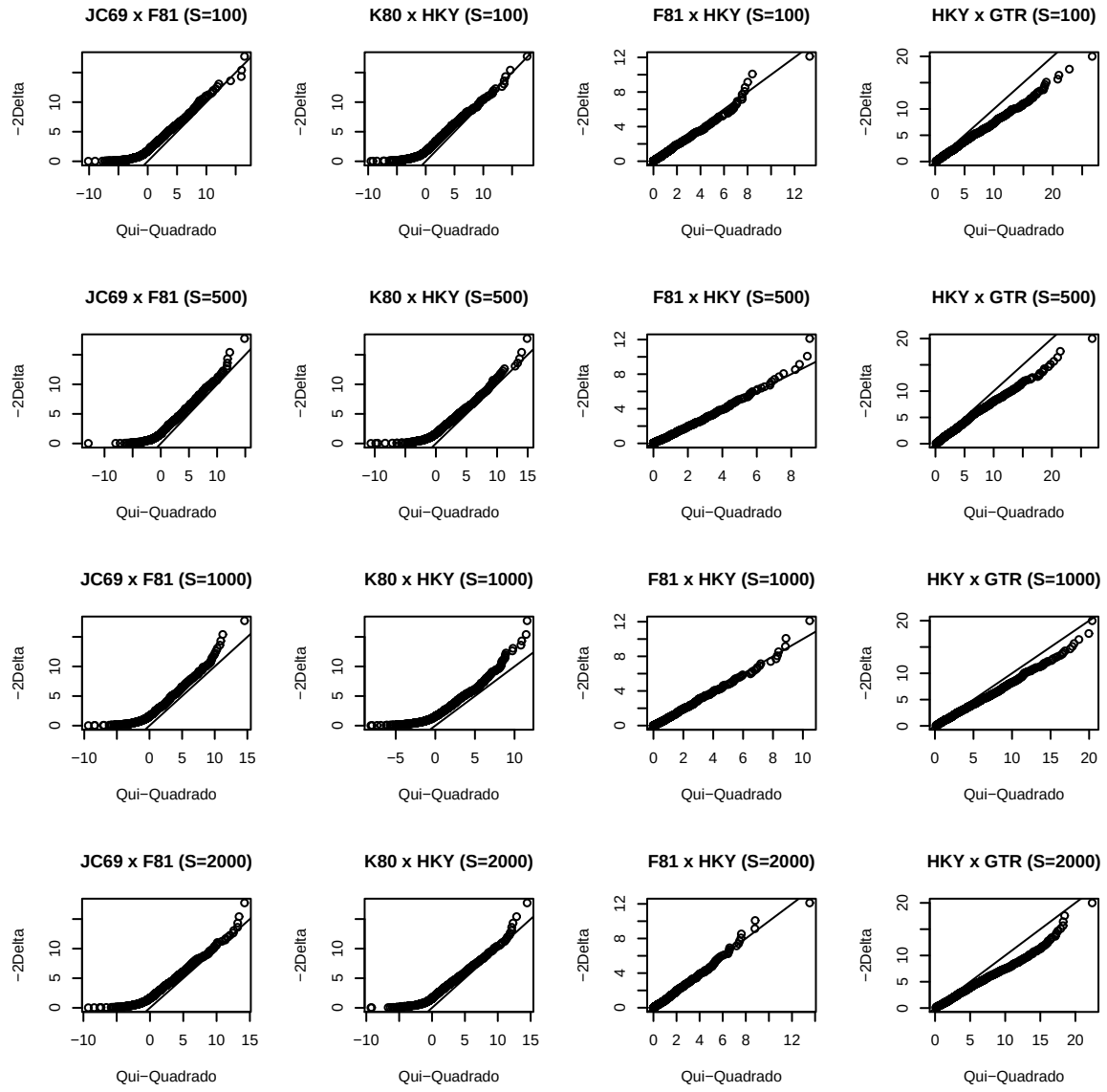


Figura D.9: Q-Q plot, sob H_0, para os Testes de Hipóteses JC69 × F81, K80 × HKY85, F81 × HKY85 e HKY85 × GTR para a Árvore 2, Utilizando $\tilde{\mathbf{p}}_0$.

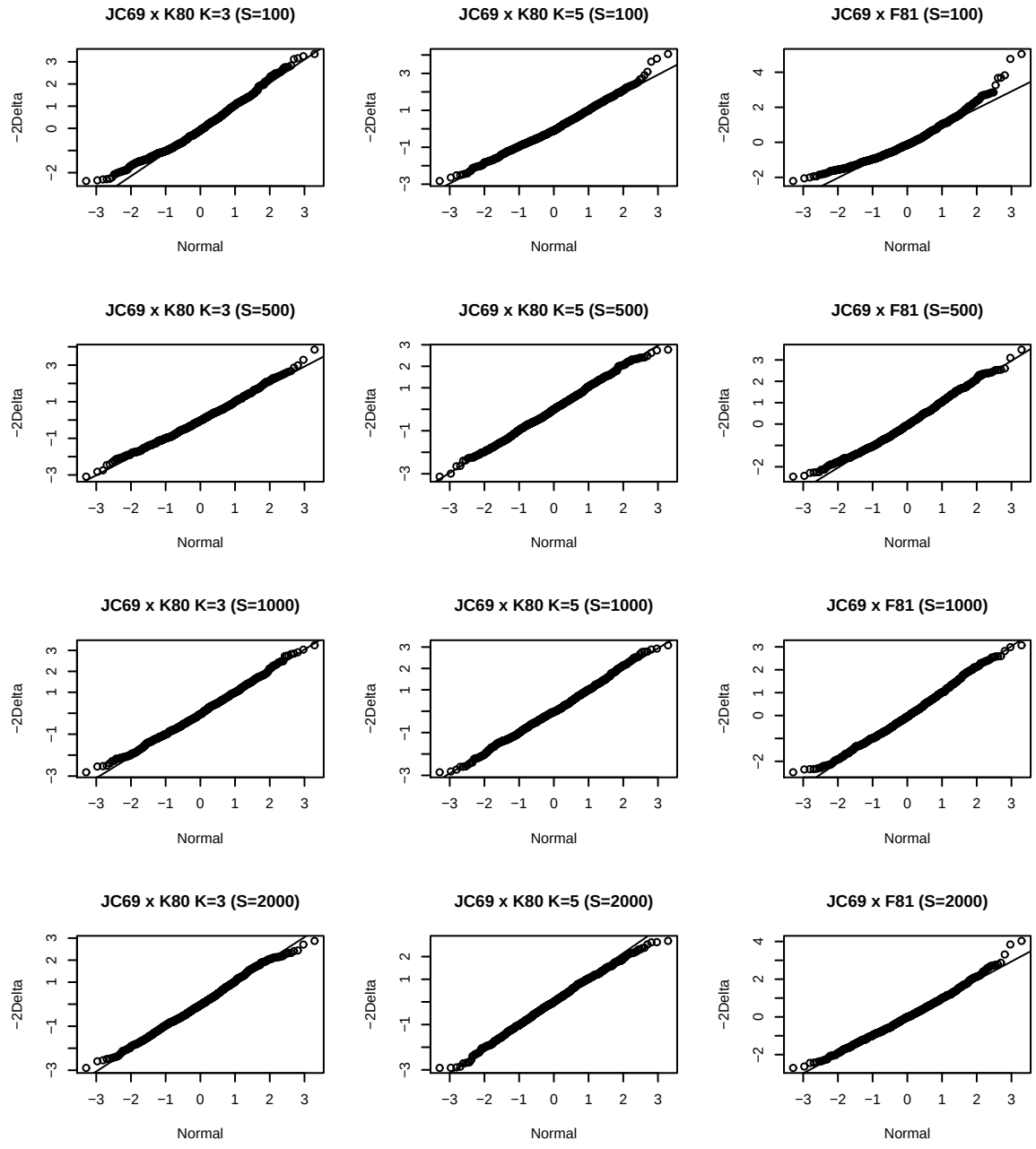


Figura D.10: Q-Q plot, sob H_1, para os Testes de Hipóteses JC69 × K80 e JC69 × F81 para a Árvore 2, Utilizando $\hat{\mathbf{p}}_0$.

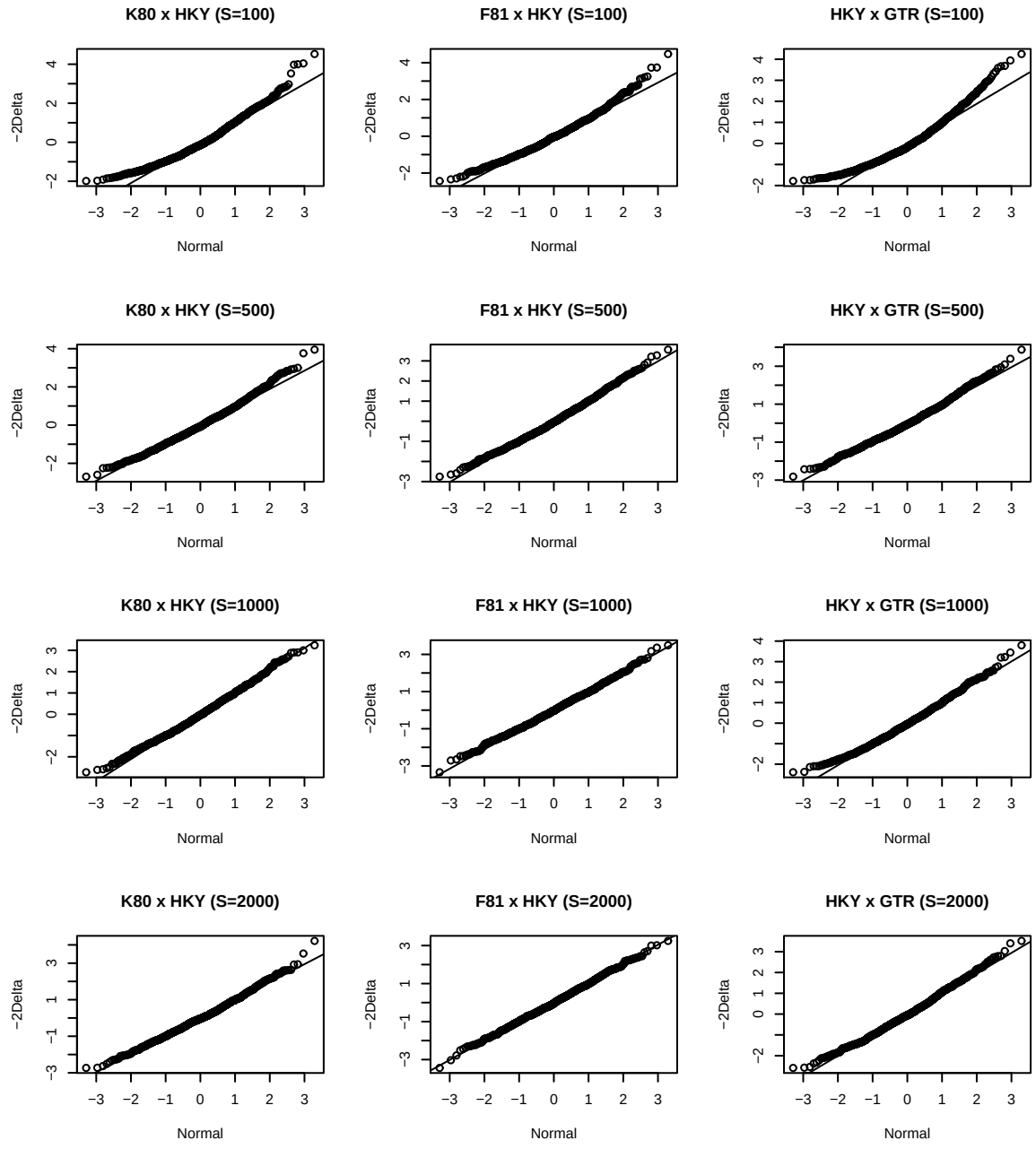


Figura D.11: Q-Q plot, sob H_1, para os Testes de Hipóteses K80 × HKY85, F81 × HKY85 e HKY85 × GTR para a Árvore 2, Utilizando $\hat{\mathbf{p}}_0$.

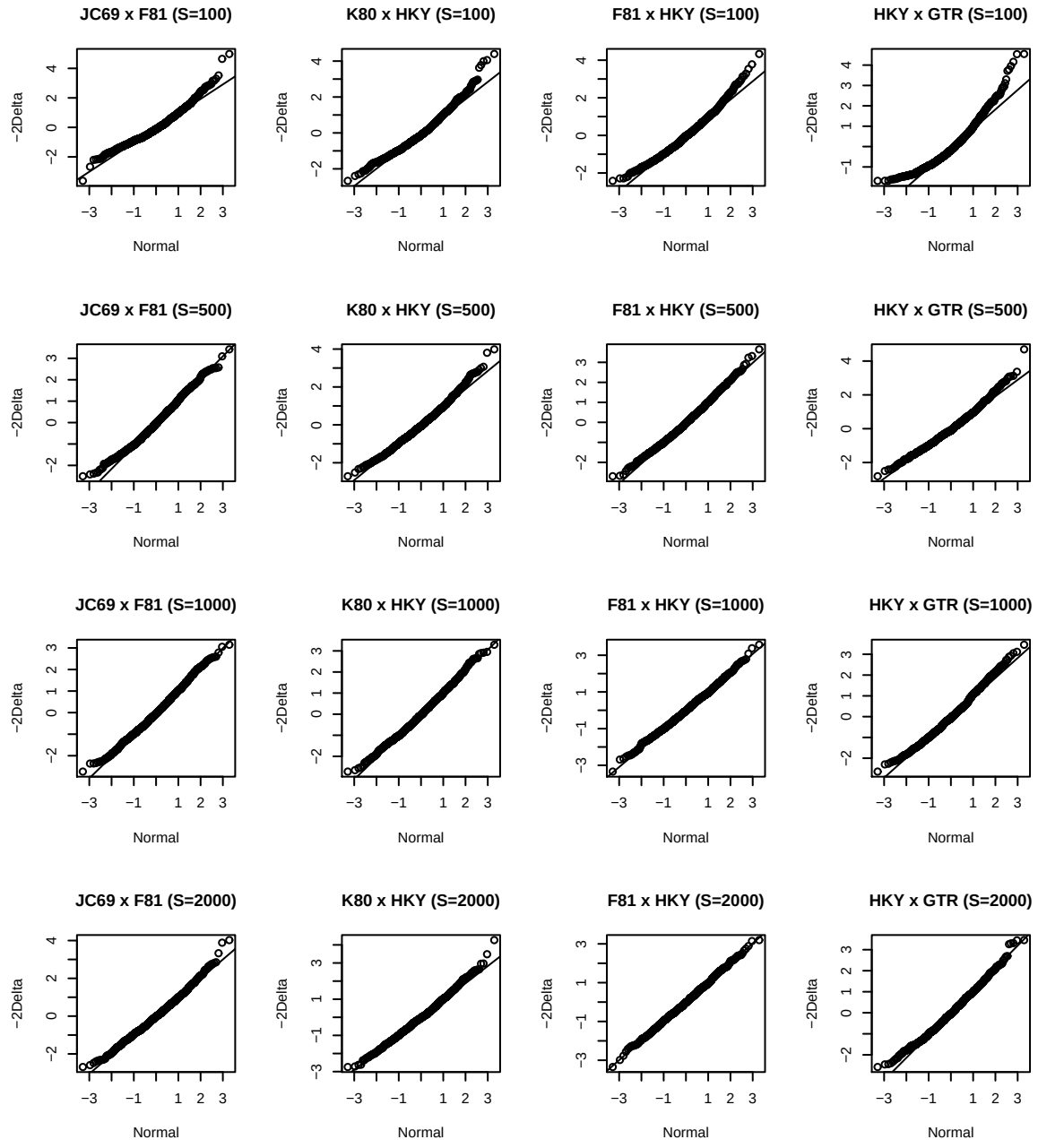


Figura D.12: Q-Q plot, sob H_1, para os Testes de Hipóteses JC69 × F81, K80 × HKY85, F81 × HKY85 e HKY85 × GTR para a Árvore 2, Utilizando $\tilde{\mathbf{p}}_0$.

Apêndice E

Apresentamos aqui os resultados de simulações feitas com as árvores da Figura 6.1 para os testes de hipóteses K80×F84 e F84×GTR. Os parâmetros utilizados para os modelos K80 e GTR foram os mesmos da Seção 6.1. Para o modelo F84 foi utilizado $K = 3$ e $\mathbf{p}_0 = (0.2, 0.3, 0.3, 0.2)$, e os comprimentos das seqüências utilizados foram $S \in \{100, 500, 1000, 2000\}$. O parâmetro $\mathbf{p}_0$ foi estimado utilizando $\tilde{\mathbf{p}}_0$. As Tabelas E.1 e E.2 apresentam, respectivamente, o valor crítico à 99% de confiança e o poder do teste para todas as simulações. A Tabela E.3 apresenta a decisão do teste qui-quadrado que tem como hipótese nula: $-2\Delta(\mathbf{X})$, sob H_0, tem distribuição χ^2 determinada pelo Teorema 3.1. Já a Tabela E.4 apresenta a decisão do teste de normalidade Shapiro-Wilks, que tem como hipótese nula: $-2\Delta(\mathbf{X})$, sob H_1, tem distribuição normal. Em ambas as Tabelas E.3 e E.4, quando não aparece aceito na célula correspondente a um teste, significa que H_0 foi rejeitada. As Figuras E.1 a E.6 apresentam os histogramas obtidos para a distribuição da estatística do teste $-2\Delta(\mathbf{X})$, definida em (3.13), sob as hipóteses H_0 e H_1, dadas em (3.12).

Notamos que os valores críticos simulados para o teste de hipóteses K80×F84 foram sempre inferiores ao valor crítico da distribuição teórica χ^2_3 (11.36). Por outro lado, para o teste de hipóteses F84×GTR, obtivemos valores críticos simulados em geral maiores do que o valor crítico teórico, que é de 13.28. Observamos também que o poder do teste aumenta, com o aumento de S, e para valores de $S \geq 500$, o poder do teste já é muito próximo de 1. Esses resultados são compatíveis com aqueles encontrados para outros testes de hipóteses na Seção 6.1.

Os resultados dos testes qui-quadrado são condizentes com o apresentado na Seção 6.1. No teste K80 × F84, que testa para $\mathbf{p}_0$, a hipótese de que $-2\Delta(\mathbf{X}) \sim \chi^2$ foi sempre rejeitada. E, com poucas exceções, a hipótese de que $-2\Delta(\mathbf{X}) \sim \chi^2$ foi aceita no teste F84 × GTR, que não testa para a freqüência das bases. Notamos ainda que, sob H_1, a hipótese de normalidade foi aceita em alguns testes, e para valores grandes de S.

Tabela E.1: Comparação entre Valor Crítico Simulado e Teórico.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$	Valor Crítico Teórico
K80 × F84	1	10.2536	11.1868	11.9346	10.2632	11.36
	2	11.0078	10.7020	8.9476	11.6952	11.36
	3	11.0894	9.2974	11.8948	10.0656	11.36
F84 × GTR	1	13.6606	13.0518	13.6120	15.0006	13.28
	2	14.5004	14.4504	15.4392	14.1852	13.28
	3	13.7744	15.1752	13.5312	13.2064	13.28

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela E.2: Poder do Teste Utilizando $\tilde{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$
K80 × F84	1	0.3910	0.9990	1.0000	1.0000
	2	0.5410	1.0000	1.0000	1.0000
	3	0.2880	0.9990	1.0000	1.0000
F84 × GTR	1	0.2250	0.9800	1.0000	1.0000
	2	0.6690	1.0000	1.0000	1.0000
	3	0.3120	0.9920	1.0000	1.0000

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela E.3: Decisão do Teste Qui-quadrado para $-2\Delta(\mathbf{X})$, sob H_0, Utilizando $\tilde{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$	Distribuição Teórica
K80 × F84	1	-	-	-	-	χ^2_3
	2	-	-	-	-	
	3	-	-	-	-	
F84 × GTR	1	aceita	-	aceita	aceita	χ^2_4
	2	-	aceita	aceita	aceita	
	3	aceita	aceita	aceita	aceita	

Nota: Os testes foram realizados ao nível de 99% de confiança.

Tabela E.4: Decisão do Teste de Normalidade Shapiro-Wilks para $-2\Delta(\mathbf{X})$, sob H_1, Utilizando $\tilde{\mathbf{p}}_0$.

Teste	Árvore	$S = 100$	$S = 500$	$S = 1000$	$S = 2000$
K80 × F84	1	-	-	-	-
	2	-	-	aceita	aceita
	3	-	-	-	aceita
F84 × GTR	1	-	-	-	-
	2	-	-	-	aceita
	3	-	-	-	-

Nota: Os testes foram realizados ao nível de 99% de confiança.

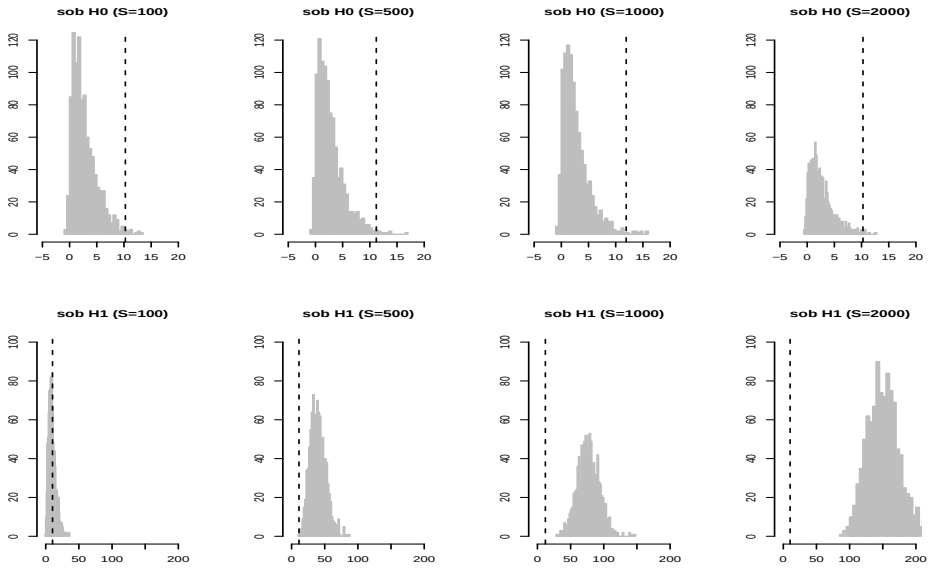


Figura E.1: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 × F84, sob H_0 e sob H_1, Utilizando $\tilde{\mathbf{p}}_0$, para a Árvore 1.

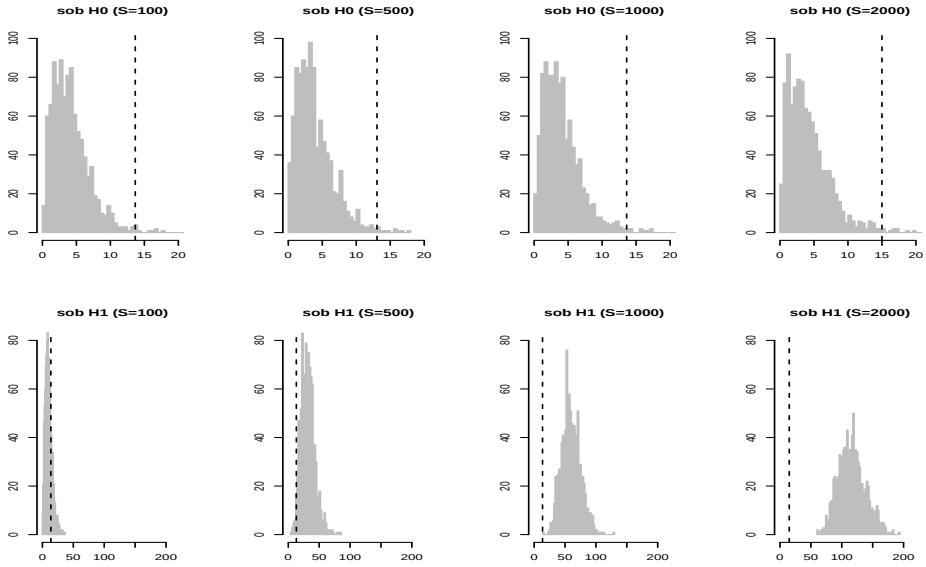


Figura E.2: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F84 $\times$ GTR, sob H_0 e sob H_1, Utilizando $\tilde{\mathbf{p}}_0$, para a Árvore 1.

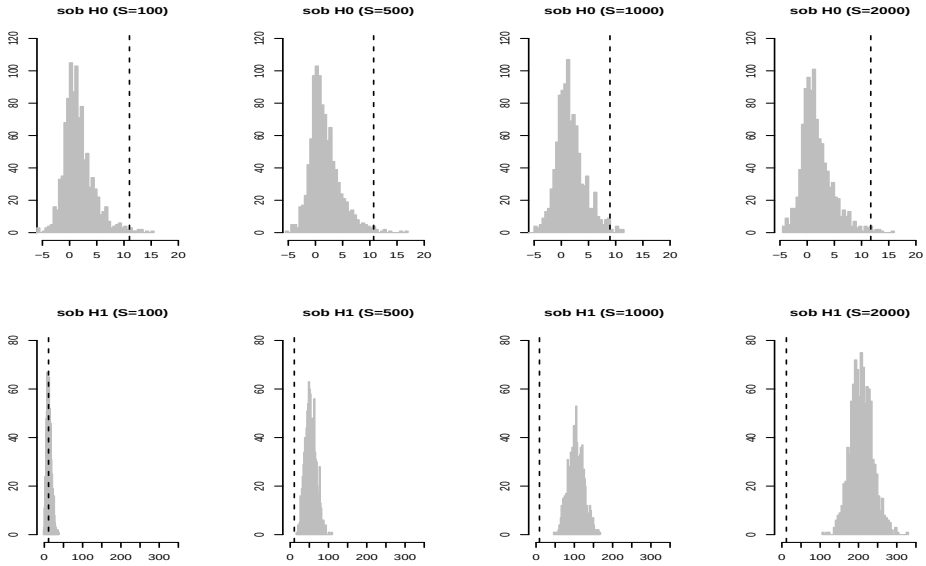


Figura E.3: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 $\times$ F84, sob H_0 e sob H_1, Utilizando $\tilde{\mathbf{p}}_0$, para a Árvore 2.

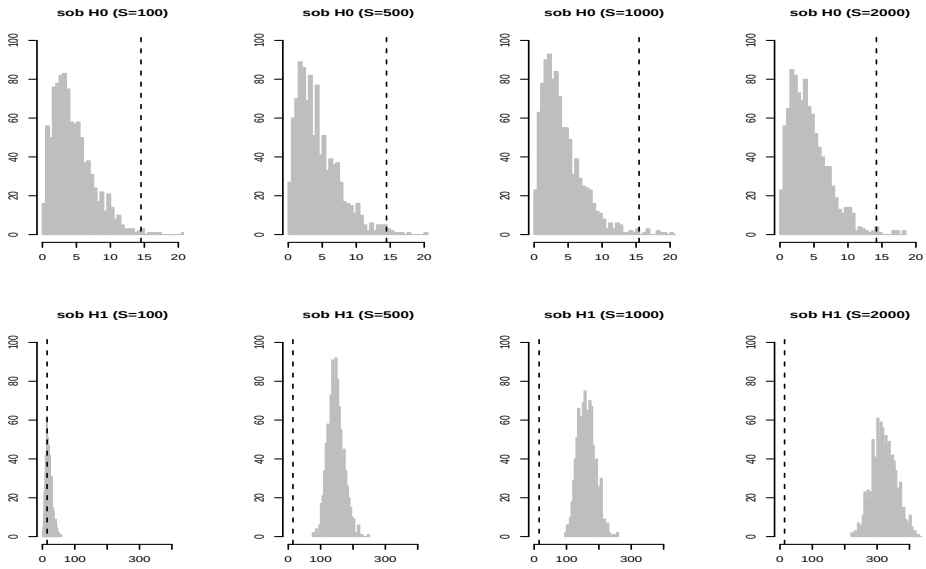


Figura E.4: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F84 $\times$ GTR, sob H_0 e sob H_1, Utilizando $\tilde{\mathbf{p}}_0$, para a Árvore 2.

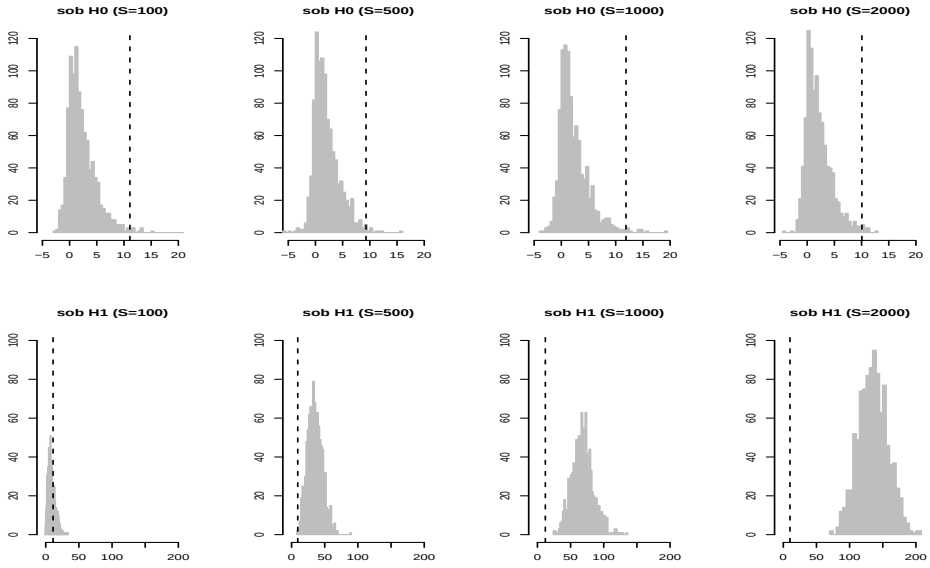


Figura E.5: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses K80 $\times$ F84, sob H_0 e sob H_1, Utilizando $\tilde{\mathbf{p}}_0$, para a Árvore 3.

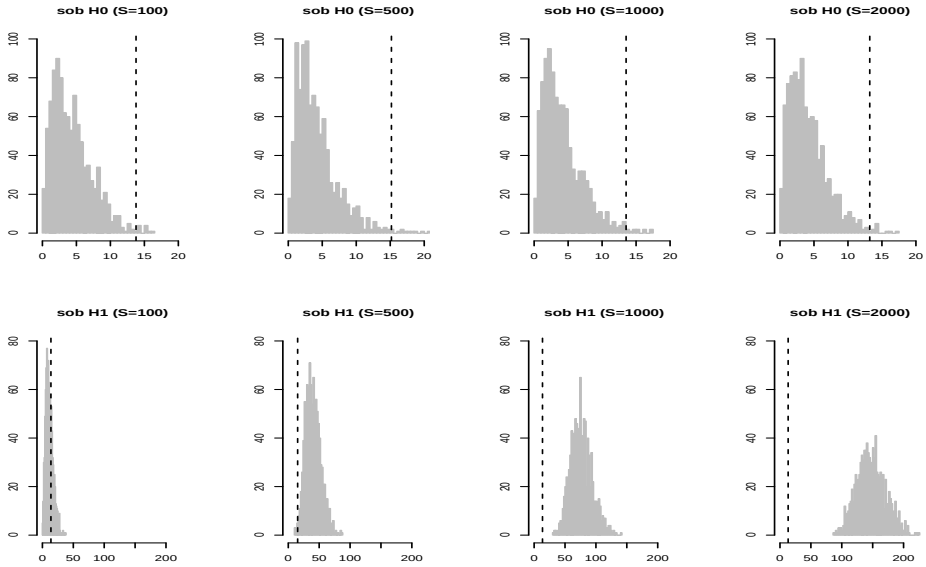


Figura E.6: Histogramas de $-2\Delta(\mathbf{X})$ para o Teste de Hipóteses F84 $\times$ GTR, sob H_0 e sob H_1, Utilizando $\tilde{\mathbf{p}}_0$, para a Árvore 3.

Apêndice F

Neste apêndice estão apresentadas as tabelas omitidas na Seção 6.3, referentes ao estudo da correção por Jackknife do vício dos estimadores obtidos pelo método da máxima verossimilhança, e seus respectivos gráficos. As Tabelas F.2, F.5, F.8 e F.11 trazem a média sobre as 200 replicações da estimativa do vício para cada parâmetro, respectivamente para os conjuntos de parâmetros de P2 a P5, apresentados na Tabela F.1. Já as Tabelas F.3, F.6, F.9 e F.12 apresentam os valores médios da estatística g, dada pela expressão (6.1), nas 200 replicações, respectivamente, para os conjuntos de parâmetros P2 a P5. E as Tabelas F.4, F.7, F.10 e F.13 apresentam o percentual das replicações em que $\hat{\theta}_J$ representa uma melhora na estimativa de θ, em relação a $\hat{\theta}$, para os mesmos conjuntos de parâmetros.

As Figuras de F.1 a F.4 apresentam as representações gráficas das estatísticas utilizadas para avaliar a correção do vício por Jackknife para, respectivamente, os conjuntos de parâmetros P2 a P5.

Fica evidenciado neste apêndice que não há uniformidade nos resultados obtidos pela correção do vício por Jackknife. Freqüentemente, o ganho obtido em um parâmetro é compensado por um resultado ruim em outro parâmetro. Entretanto, em geral, observa-se que o vício estimado tende a diminuir com o aumento de S, e, em grande parte das estimativas, o ganho com a correção também diminui. Verifica-se também que o maior benefício da utilização da correção do vício por jackknife ocorreu no parâmetro K. No entanto, esse benefício foi pequeno para os parâmetros π_A, π_G e π_C.

Tabela F.1: Parâmetros Utilizados nas Simulações para Avaliar o Jackknife.

P	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
P1	0.12	0.1	0.1	0.22	0.4	2	0.25	0.25	0.25
P2	0.12	0.1	0.1	0.22	0.4	3	0.20	0.20	0.25
P3	0.12	0.1	0.1	0.22	0.4	5	0.25	0.25	0.25
P4	0.12	0.1	0.1	0.22	0.4	9	0.32	0.20	0.23
P5	0.12	0.1	0.1	0.22	0.4	2	0.29	0.21	0.25

Tabela F.2: Vício Médio para P2.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00554	0.00106	0.00244	0.00634	0.01567	0.26063	-0.00024	-0.00027	0.00025
140	0.00190	0.00115	0.00073	0.00372	0.00923	0.12285	-0.00011	-0.00007	0.00008
200	0.00119	0.00057	0.00071	0.00263	0.00610	0.08365	-0.00013	-0.00002	0.00006

Tabela F.3: Ganho Médio com $\hat{\theta}_J$ para P2.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00008	0.00032	0.00126	0.00209	0.00565	0.08685	0.00000	0.00002	-0.00007
140	0.00170	0.00012	0.00039	0.00124	0.00230	0.03219	0.00000	0.00004	0.00006
200	0.00072	0.00023	0.00033	0.00068	0.00112	0.01525	0.00003	-0.00005	0.00006

Tabela F.4: Percentual das Replicações g Positivo para P2.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	60	59	66	55	48.5	47	54.5	49	42.5
140	70.5	56	63.5	49.5	47.5	51.5	48.5	53	56
200	62.5	59.5	64.5	47.5	48.5	49	52.5	44.5	53

Tabela F.5: Vício Médio para P3.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00569	0.00143	0.00227	0.00723	0.01832	0.54675	-0.00003	-0.00005	0.00003
140	0.00232	0.00106	0.00097	0.00426	0.01059	0.26646	0.00003	-0.00004	-0.00002
200	0.00123	0.00061	0.00077	0.00301	0.00705	0.18317	-0.00001	0.00007	0.00002

Tabela F.6: Ganho Médio com $\hat{\theta}_J$ para P3.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	-0.00156	0.00032	0.00018	0.00023	0.00338	0.18745	-0.00024	-0.00006	0.00003
140	0.00178	0.00062	0.00054	0.00176	0.00334	0.06522	-0.00003	0.00007	0.00000
200	0.00098	0.00035	0.00035	0.00123	0.00142	0.05306	0.00005	0.00002	0.00004

Tabela F.7: Percentual das Replicações com g Positivo para P3.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	62.5	63.5	62.0	55.0	44.5	49.0	48.5	54.0	52.0
140	72.0	64.5	60.5	52.0	51.0	49.5	45.5	53.0	49.5
200	66.0	64.0	57.5	55.5	48.0	52.5	56.0	52.5	53.0

Tabela F.8: Vício Médio para P4.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00920	0.00108	0.00363	0.00897	0.03022	1.74506	0.00022	0.00012	-0.00014
140	0.00377	0.00129	0.00117	0.00541	0.01670	0.79466	0.00011	0.00003	-0.00007
200	0.00151	0.00078	0.00089	0.00414	0.00972	0.48154	0.00002	0.00002	0.00000

Tabela F.9: Ganho Médio com $\hat{\theta}_J$ para P4.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00282	0.00051	0.00117	0.00359	0.00939	0.73387	0.00005	0.00003	0.00010
140	0.00279	0.00107	0.00115	0.00345	0.00657	0.26833	0.00009	-0.00002	0.00008
200	0.00156	0.00044	0.00057	0.00181	0.00284	0.07950	-0.00001	-0.00001	0.00001

Tabela F.10: Percentual das Replicações com g Positivo para P4.

S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	55.0	59.0	62.0	52.0	49.5	48.0	53.5	52.5	55.0
140	70.0	61.5	66.0	56.5	50.5	46.5	55.0	49.0	56.0
200	66.0	60.5	60.5	53.0	49.5	44.0	45.0	49.0	49.0

Tabela F.11: Vício Médio para P5.

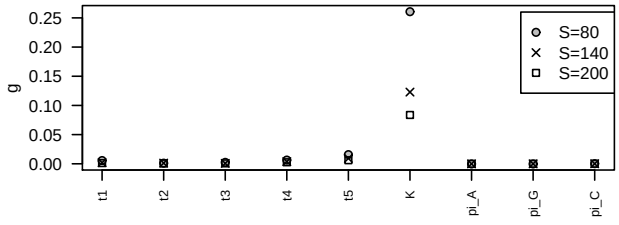
S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00554	0.00106	0.00244	0.00634	0.01567	0.26063	-0.00024	-0.00027	0.00025
140	0.00190	0.00115	0.00073	0.00372	0.00923	0.12285	-0.00011	-0.00007	0.00008
200	0.00119	0.00057	0.00071	0.00263	0.00610	0.08365	-0.00013	-0.00002	0.00006

Tabela F.12: Ganho Médio com $\hat{\theta}_J$ para P5.

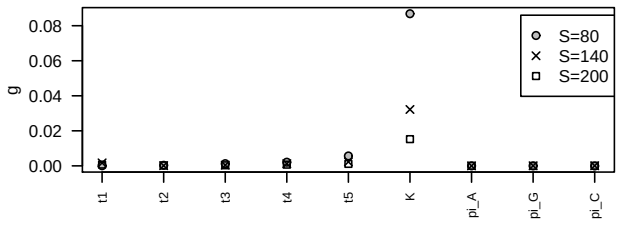
S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	0.00008	0.00032	0.00126	0.00209	0.00565	0.08685	0.00000	0.00002	-0.00007
140	0.00170	0.00012	0.00039	0.00124	0.00230	0.03219	0.00000	0.00004	0.00006
200	0.00072	0.00023	0.00033	0.00068	0.00112	0.01525	0.00003	-0.00005	0.00006

Tabela F.13: Percentual das Replicações com g Positivo para P5.

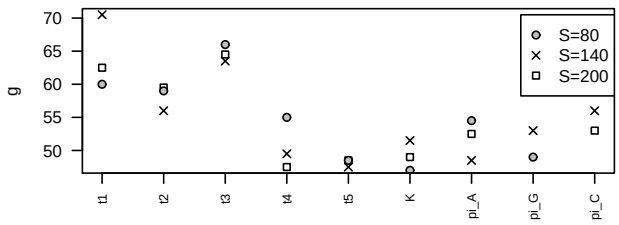
S	τ_1	τ_2	τ_3	τ_4	τ_5	K	π_A	π_G	π_C
80	60	59	66	55	48.5	47	54.5	49	42.5
140	70.5	56	63.5	49.5	47.5	51.5	48.5	53	56
200	62.5	59.5	64.5	47.5	48.5	49	52.5	44.5	53



(a)



(b)



(c)

Figura F.1: Gráficos de (a) Vício Médio, (b) Ganho Médio, e (c) Percentual de Replicações com g Positivo para P2.

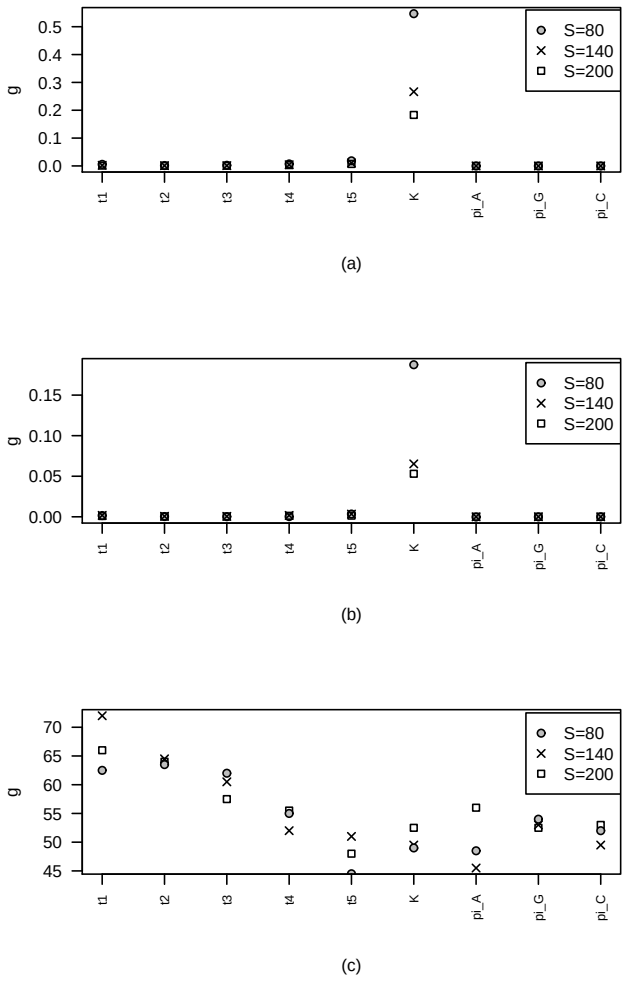


Figura F.2: Gráficos de (a) Vício Médio, (b) Ganho Médio, e (c) Percentual de Replicações com g Positivo para P3.

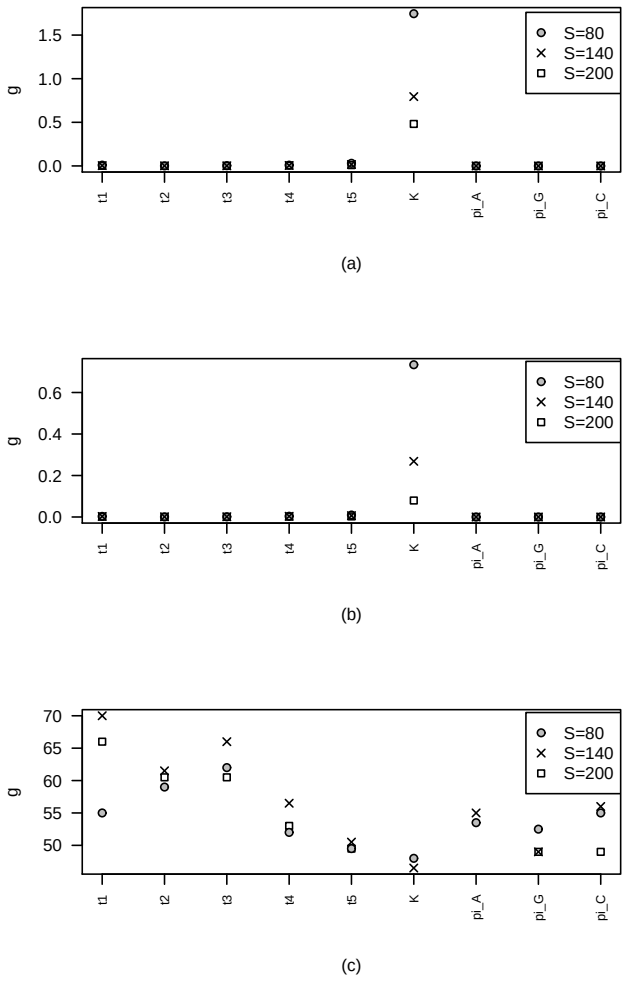


Figura F.3: Gráficos de (a) Vício Médio, (b) Ganho Médio, e (c) Percentual de Replicações com g Positivo para P4.

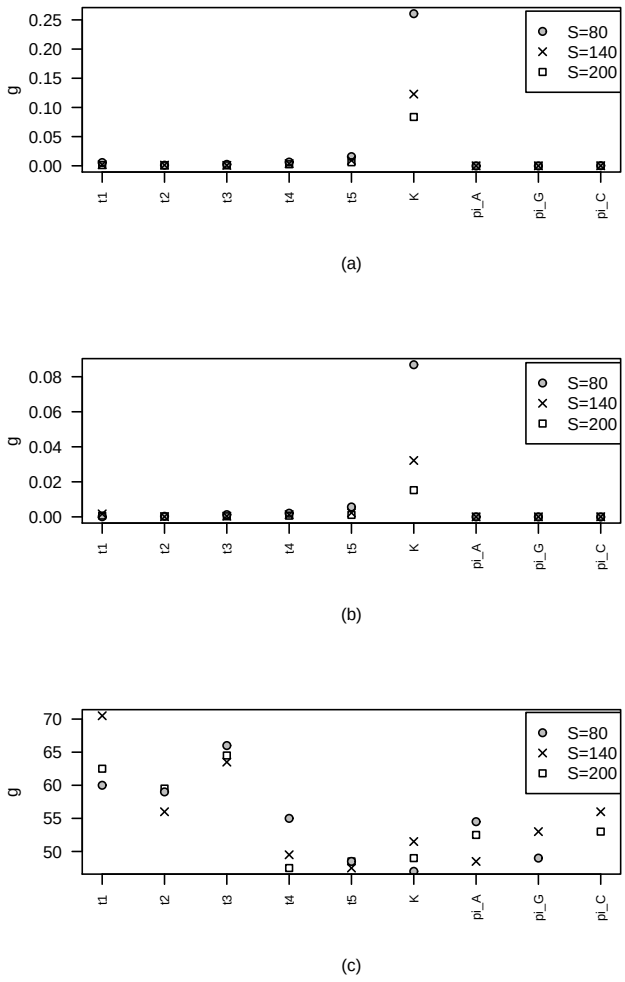


Figura F.4: Gráficos de (a) Vício Médio, (b) Ganho Médio, e (c) Percentual de Replicações com g Positivo para P5.