UMA RESPOSTA BAYESIANA AO PARADOXO DE SUZUKI

Elizabeth Belo Hypólito

Instituto de Matemática da Universidade Federal do Rio de Janeiro

Mestrado em Estatística

Orientador: Marco Antonio Rosa Ferreira

Co-orientadora: Maria Alejandra Jaramillo Sierra

Rio de Janeiro

Abril 2005

UMA RESPOSTA BAYESIANA AO PARADOXO DE SUZUKI

Elizabeth Belo Hypólito

Dissertação de Mestrado submetida ao programa de Pós-Graduação em Estatística do Instituto de Matemática da Universidade Federal do Rio de Janeiro-UFRJ, como parte dos requisitos necessários à obtenção do título de Mestre em Estatística.

Aprovada por:

Orientador, Professor Marco Antonio Rosa Ferreira-UFRJ

Co-orientadora, Professora Maria Alejandra Jaramillo Sierra - UFRJ

Professora Alexandra M. Schmidt - UFRJ

Professor Carlos Alberto de Bragança Pereira - USP

Rio de Janeiro

Abril 2005

Ficha Catalográfica

Hypólito, Elizabeth Belo.

Uma resposta bayesiana ao paradoxo de Suzuki.

Rio de Janeiro. Universidade Federal do Rio de Janeiro. 2005.

xii, 71p.

Dissertação de Mestrado em Estatística

1. Inferência bayesiana

2. Modelo de Misturas

3. Seqüências de DNA concatenadas

4. Árvores filogenéticas

I. Universidade Federal do Rio de Janeiro

II. Uma resposta bayesiana ao paradoxo de Suzuki.

Abstract

Since its development in the nineties, the Bayesian phylogenetic analysis has grown in popularity and has contributed to the solution of several important phylogenies. Nevertheless, some researchers are still wary about the application of Bayesian phylogenetic analysis, and papers discussing its performance are common. In one of those papers, Suzuki et al.(2002) use concatenated DNA sequences to prove that Bayesian phylogenetic analysis overestimates the credibility of the branches of the consensus tree. In this work, we show that there is a fundamental shortcoming in the analysis of Suzuki et al. (2002): they simulate the DNA sequences from a mixture of topologies model, and analyse the data with a model that assumes a true unique topology. Because of this shortcoming, their Bayesian analysis frequently chooses one topology over the others, and they claim that is an indication of credibility overestimation. We show mathematically why that happens, and we argue that the correct way to proceed is to analyse their datasets with a mixture of topologies model. Using such a model, implemented in the statistical package BARCE (Husmeier & McGuire 2003), we show that the Bayesian analysis is able to identify that the DNA sequences are not generated by a unique topology.

Resumo

Desde o seu desenvolvimento, em meados dos anos 90, a análise filogenética bayesiana vem ganhando bastante popularidade e contribuindo para a resolução de importantes filogenias. No entanto, esse método ainda é visto com desconfiança por alguns filogeneticistas, o que justifica a freqüente publicação de artigos que discutem sua performance. Em um desses artigos, Suzuki *et al.* (2002) utilizam seqüências de DNA concatenadas com o objetivo de provar que a análise filogenética bayesiana superestima a credibilidade dos ramos da árvore de consenso. Nesta dissertação, iremos mostrar que há uma falha fundamental na análise de Suzuki *et al.* (2002): eles simulam as seqüências de DNA com um modelo de mistura de topologias e analisam esses dados com um modelo que assume uma única topologia como sendo a verdadeira. Devido a essa falha, a análise bayesiana desses autores freqüentemente dá preferência a uma topologia sobre as outras. Suzuki *et al.* (2002) alegam que esse comportamento é indicação de que a credibilidade dos ramos está sendo superestimada. Nós demonstramos matematicamente porque isso acontece e argumentamos que o procedimento correto para analisar esses dados é assumir um modelo de mistura de topologias. Usando tal modelo, implementado no pacote estatístico BARCE (Husmeier & McGuire, 2003), nós mostramos que a análise bayesiana é capaz de identificar que as sequências de DNA não foram geradas de uma única topologia.

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

A evolução pode ser definida como um processo através do qual ocorrem transformações nos seres vivos ao longo do tempo, dando origem a novas espécies. Mecanismos como mutações, variações gênicas, isolamento geográfico e seleção natural fazem com que a vida na Terra esteja em contínua evolução.

A filogenética é um ramo da biologia que tem como objetivos reconstruir a história evolutiva, ou seja, a relação de parentesco entre os organismos, e estimar o tempo de divergência entre ancestrais e seus descendentes. Esses laços genealógicos são ilustrados por uma árvore binária denominada árvore filogenética ou filogenia, a qual é construída com base em características morfológicas ou códigos genéticos dos organismos.

Nos últimos 50 anos, surgiram diversas técnicas de construção de filogenias. Entre elas, podemos destacar a *análise filogenética bayesiana*, introduzida em meados dos anos 90 e revisada por Huelsenbeck *et al.* (2001). Desde o seu desenvolvimento, esse método vem ganhando bastante popularidade e contribuindo para a resolução de importantes filogenias como, por exemplo, a dos mamíferos (Murphy *et al.*, 2001) e a maior árvore já construída para plantas (Karol *et al.*, 2001).

Embora de grande importância, a inferência filogenética bayesiana ainda é vista

com desconfiança por alguns filogeneticistas, o que justifica a freqüente publicação de artigos que discutem sua performance como, por exemplo, Suzuki *et al.* (2002) e Alfaro *et al.* (2003). As principais críticas feitas a essa ferramenta referem-se às prioris assumidas e ao suporte atribuído aos ramos da árvore de consenso.

O artigo de Suzuki, Glazko e Nei, publicado em 2002, discute a utilização da inferência filogenética bayesiana em seqüências de DNA concatenadas. Utilizando a concatenação de dados pertencentes a três diferentes histórias evolutivas, os autores compararam a análise bayesiana com dois métodos clássicos de construção de filogenias, a máxima verossimilhança via bootstrap e a vizinhança conjunta via bootstrap. Eles concluíram que, para dados encadeados, a inferência bayesiana superestima a credibilidade dos ramos da árvore de consenso. Como Murphy *et al.* (2001) e Karol *et al.* (2001) utilizaram seqüências de genes concatenados, Suzuki *et al.* (2002) concluíram também que as filogenias para os mamíferos e para as plantas podem ainda não estar totalmente resolvidas.

A presente dissertação tem por objetivo contestar o resultado obtido por Suzuki *et al.* (2002), mostrando que a principal falha desses autores é a simulação de dados de um modelo de mistura de filogenias e a análise desses dados assumindo um modelo com uma única filogenia.

A estrutura desse trabalho é como descrita a seguir. No capítulo 2 introduzimos conceitos fundamentais da inferência filogenética e apresentamos alguns dos mais importantes modelos de substituição de nucleotídeos. Em seguida, descrevemos as principais técnicas de reconstrução de árvores filogenéticas. Dedicamos um capítulo ao entendimento e estudo do artigo de Suzuki *et al.* (2002). Posteriormente, expomos nossas críticas a esse artigo, explicando matematicamente o paradoxo de Suzuki. Finalmente, apresentamos a proposta de análise dos dados com o modelo de misturas de filogenias em seqüências de DNA de Husmeier & McGuire (2003), implementado no pacote estatístico BARCE.

Capítulo 2

Evolução e árvores filogenéticas

Neste capítulo, faremos uma introdução à análise filogenética, apresentando ao leitor conceitos fundamentais para o entendimento desta dissertação. Começaremos fazendo um resumo das teorias evolutivas ao longo da história. Em seguida, discorreremos a respeito de árvores filogenéticas e seqüências de DNA. Finalizando, exibiremos alguns dos mais importantes modelos de substituição de nucleotídesos da literatura atual.

2.1 Evolução

Há muitos séculos os homens se deparam com questões referentes às suas origens. *Como surgiu a vida na Terra? Como surgiu o primeiro homem?* Para responder a essas perguntas, foram formuladas, ao longo da História, diversas teorias sobre a origem da vida.

Uma das mais antigas teorias, a Teoria da Abiogênese ou Geração Espontânea se fundamentava na afirmação de que a vida não se originou de outra vida preexistente. Na antiga Grécia, Aristóteles afirmava que um princípio ativo, ou força vital, atuando na matéria inanimada, podia dar-lhe vida. Essa teoria foi bem aceita até cerca de 300 anos atrás.

Nem todos os cientistas concordavam com as idéias da abiogênise e muitos deles buscaram combatê-las. Mas foi Louis Pasteur, químico e biólogo francês, quem, em meados do século XIX, provou que a vida não poderia surgir espontaneamente.

Surgia então a Teoria da Biogênese, que assegura que toda vida origina-se de uma vida preexistente. No entanto ainda era necessário responder às perguntas *Como surgiu então a primeira vida, aquela que originou todas as outras? E como ocorreu a evolução?*

No meio científico, a mais famosa teoria da origem da vida é a Teoria dos Coacervados, formulada por Alexander Oparin, em 1936 e comprovada pelas experiências de Lloyd Miller, em 1954 e Sidney Fox, em 1950. De acordo com essa hipótese, a vida teria se originado nos oceanos primitivos da Terra, pela união de grupos de moléculas, os coacervados.

As teorias evolucionistas afirmam que todos os seres vivos originaram-se de um mesmo ancestral. A vida é o resultado de um longo e contínuo processo de evolução em que espécies surgiram e foram extintas, ocorrendo diversas alterações nas formas de vida desde o surgimento da Terra.

O primeiro homem a desenvolver uma teoria evolucionista foi o francês Jean-Baptiste Lamarck. Suas idéias desafiaram o pensamento predominante na época, que Deus criou o mundo com as formas de vida que hoje conhecemos e elas nunca evoluíram.

Lamarck acreditava que as formas de vida mais evoluídas haviam surgido de vidas mais simples. Sua teoria era fundamentada no uso e desuso e na transmissão das características adquiridas. De acordo com o ambiente, um determinado órgão de uma espécie tenderia a ser mais ou menos utilizado. Se fosse pouco utilizado, esse órgão tenderia a desaparecer e se muito utilizado se desenvolveria. Essas modificações seriam transmitidas aos descendentes.

A teoria de Lamarck não é aceita atualmente, pois sabe-se que características

adquiridas não são transmitidas aos descendentes.

Em 1859, surgia outra teoria revolucionária, a do inglês Charles Darwin. Em seu livro, *A origem das espécies*, Darwin defendia a idéia de que os seres vivos apresentam variações em seus caracteres, não sendo, portanto, idênticos entre si. Assim, os organismos que apresentam variações favoráveis às condições do ambiente onde vivem têm maiores chances de sobreviver e deixar descendentes. Esses descendentes também apresentarão variações vantajosas, pois há transmissão de caracteres de pais para filhos. Desta forma, ao longo das gerações, a natureza faz uma seleção natural sobre os indivíduos, mantendo ou melhorando o grau de adaptação destes ao meio.

A atual teoria evolucionista é o Neodarwinismo ou Teoria Sintética, que reafirma as idéias de Darwin e incorpora a elas os conceitos modernos da genética.

Para o Neodarwinismo, os principais fatores evolutivos são a variabilidade genética, a mutação e a seleção natural das espécies.

A *variabilidade genética* ocorre em todos os seres que se reproduzem de forma sexuada. Ela pode ocorrer tanto no genótipo, conjunto de todos os genes do indivíduo, como no fenótipo, conjunto das características observáveis em um organismo. No entanto, somente as variações genéticas que alteram o fenótipo de um organismo são importantes na evolução.

A *mutação gênica* é uma mudança em um determinado gene, durante a duplicação da molécula de DNA. Essa mudança pode ser a perda, a adição ou a substituição de um ou mais nucleotídeos, originando um gene capaz de codificar uma proteína diferente da codificada pelo gene original.

As mutações ocorrem ao acaso, de modo que não é possível prever o gene a ser mutado. Elas variam entre 10 e 80 mutações a cada 1 milhão de gametas.

A grande maioria das mutações são prejudiciais ao organismo, podendo levar à morte. No entanto, se uma mutação proporcionar características adaptativas à espécie, por seleção natural, ela tende a ser mantida. As mutações transmitidas aos

descendentes são as que ocorrem em células germinativas.

A *seleção natural* ocorre quando grupos de organismos competem por uma melhor adequação ao ambiente e apenas os mais aptos sobrevivem, tendo a oportunidade de transmitir seus genes. Dizemos então que esses organismos foram selecionados.

Na próxima seção, faremos uma introduçao à árvore filogenética, mecanismo gráfico que representa a história evolutiva de um conjunto de organismos.

2.2 Árvores filogenéticas

Willi Hennig (1950) desenvolveu um conjunto de idéias para reconstruir a história evolutiva dos organismos, buscando estimar o tempo de divergência entre os ancestrais e os descendentes. A proposta de Henning foi utilizar métodos cladísticos para recontrução de árvores filogenéticas.

Uma *árvore filogenética* é um gráfico composto de nós e ramos, que tem por objetivo ilustrar as relações de parentesco entre organismos. Recebem o nome de *nó* os pontos que representam as entidades biológicas em estudo, também denominadas de *táxon*. Esses táxons podem ser espécies, populações, indivíduos ou genes. Os *ramos* ou *galhos*, fazem a ligação entre os nós e são, geralmente, proporcionais ao tempo ou ao número de mutações ocorridas até o surgimento de um novo táxon.

A relação evolutiva entre um conjunto de espécies é representada por *árvores binárias*, ou seja, árvores para as quais saem exatamente dois galhos de cada nó. A figura 2.1 apresenta duas árvores binárias. À esquerda é ilustrada uma *árvore com raiz* e à direita uma *árvore sem raiz*. Em uma árvore raizada, todos os táxons são descendentes de um mesmo ancestral, o qual é representado pelo *nó raiz*.

Os nós que aparecem nas pontas, 1, 2, 3, 4 e 5 são *nós externos* e representam os táxons que estamos buscando relacionar. Os nós restantes, 6, 7, 8 e 9 (nó raiz) são denominados *nós internos* e representam os ancentrais dos nós externos. Os

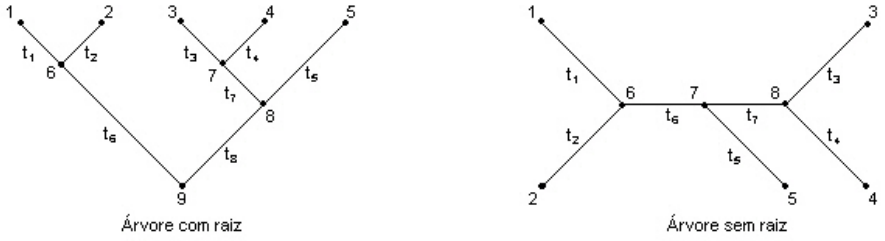

Figura 2.1: Árvores com e sem raiz

segmentos de reta unindo os nós são os galhos. Por exemplo, o segmento que une 1 a 6 é um galho, com comprimento dado por t_6.

A construção de ávores filogenéticas é um proceso complexo, já que para um conjunto de táxons há mais de uma possibilidade de história evolutiva. Chamamos de *topologia*, a árvore que apresenta somente a relação de parentesco entre as espécies, sem considerar os valores dos comprimentos de galhos. Para m táxons, o número de possíveis topologias com raiz é

$$\frac{(2m-3)!}{2^{m-2}(s-2)!} \tag{2.1}$$

e o número de possíveis topologias sem raiz é dado por

$$\frac{(2m-5)!}{2^{m-3}(s-3)!} \tag{2.2}$$

Como exemplo, suponha a existência de 4 espécies, a, b, c e d. As figuras 2.2 e 2.3 mostram todas as possíveis topologias com e sem raiz para essas espécies.

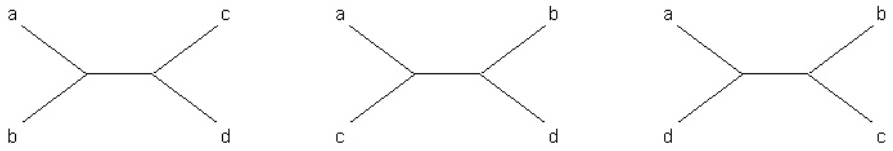

Figura 2.2: Possíveis topologias sem raiz para 4 espécies

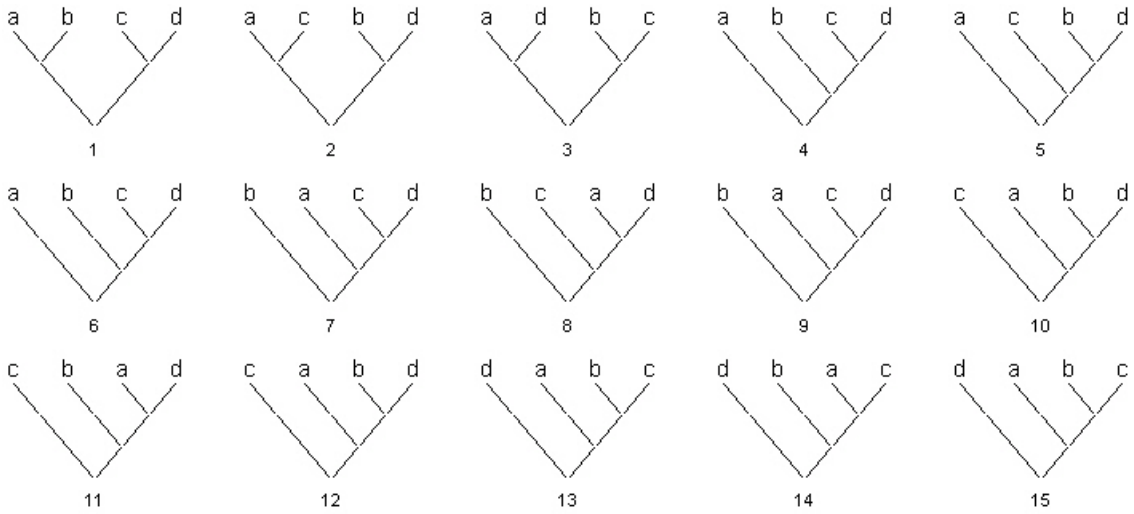

Figura 2.3: Possíveis topologias com raiz para 4 espécies

2.3 Seqüências de DNA

As informações hereditárias da grande maioria dos organismos vivos são transmitidas a seus descendentes por moléculas de *ácido desoxirribonucléico*, mais conhecido como *DNA*.

O modelo de DNA, hoje aceito, foi proposto por Watson & Crick (1953). Sua configuração espacial é a de uma dupla cadeia em forma de hélice, como uma escada em espiral. É composto por milhões ou até bilhões de nucleotídeos. Os *nucleotídeos* são identificados por *bases nitrogenadas*, as quais se dividem em dois grupos:

Purinas: Adenina (A) e Guanina (G);

Pirimidinas: Citosina (C), Timina (T).

Quando o material genético é passado aos descendentes de um indivíduo, a seqüência de DNA pode sofrer mutações. Uma mutação pode ser a substituição, a eliminação ou a inserção de um nucleotídeo.

Se temos indivíduos diversos e queremos verificar se há evidências de que eles

divergiram de um mesmo ancestral por mutação e seleção, nós comparamos suas seqüências de DNA. O procedimento de comparação é feito com o alinhamento.

Dizemos que duas ou mais seqüências estão *alinhadas* quando seus caracteres estão na mesma ordem. As seqüências são colocadas em linhas e seus caracteres similares são colocados na mesma coluna. Na ilustração abaixo temos duas seqüências alinhadas. O símbolo "-" indica os nucleotídeos inseridos ou eliminados.

$$\begin{array}{cccccccccccc} G & G & A & G & C & A & G & T & - & - & - & A \\ | & | & | & | & | & | & | & | & | & | & | & | \\ G & G & A & A & - & - & G & T & T & G & C & C \end{array}$$

O alinhamento pode ser global ou local. Alinhamentos globais usam o maior número de nucleotídeos possível. Alinhamentos locais comparam apenas um gene ou um trecho dele. Um *gene* é um intervalo da molécula de DNA responsável por alguma característica hereditária.

Os livros de Durbin, Eddy, krogh e Mitchison (1998) e de Mount (2001) são boas referências para alinhamento de seqüências de DNA.

2.4 Modelos evolutivos

A evolução pode ser vista como um complicado processo estocástico. Os modelos de substituição de nucleotídeos são simplificações desse processo, sendo, portanto, utilizados na reconstrução de árvores filogenéticas.

Hoje, sabemos que o nucleotídeo de um determinado sítio da seqüência de DNA pode mudar ao longo de um intervalo de tempo. Imagine, por exemplo, a história evolutiva de uma espécie qualquer, digamos E. Comece pelo mais antigo ancestral de E, representado pelo nó raiz da árvore filogenética. Para essa espécie inicial, suponha que o nucleotídeo C foi observado para um particular sítio de uma seqüência de DNA. Considere que, após certo período de tempo, ocorra uma mutação e C seja substituído

por G. Posteriormente, G muta para C. Seguindo a linha de sucessão, chegamos a A, o nucleotídeo da espécie E para esse mesmo sítio. Esse processo pode ser descrito por uma cadeia de Markov, com espaço de estados discreto dado por {A,C,G,T} e tempo contínuo.

2.4.1 Processos de Markov com tempo contínuo

Suponha um processo de Markov homogêneo, com espaço de estados S, discreto, e tempo definido nos reais positivos. Seja X a variável que assume valores em S. A suposição de *markovianidade* implica que, dado um estado $X(t) = i$ em algum tempo t, a probabilidade de que $X(t + h) = j$, em um tempo futuro $t + h$, não depende dos valores de X antes do tempo t. A suposição de homogeneidade significa que a probabilidade condicional

$$P(X(t + h) = j|X(t) = i) \tag{2.3}$$

é independente de t e, portanto, pode ser escrita como $p_{ij}(h)$.

Considere o caso em que a probabilidade de transição $p_{ij}(h)$ assume a forma

$$p_{ij}(h) = q_{ij}h + o(h), \quad i \neq j \tag{2.4}$$

$$p_{ii}(h) = 1 + q_{ii}h + o(h), \tag{2.5}$$

quando $(h \to 0)$. A função $o(h)$ é tal que $\lim_{h\to 0} \frac{o(h)}{h} = 0$.

Pelas equações (2.4) e (2.5), é facil perceber que $q_{ij} \geq 0$ para $i \neq j$ e $q_{ii} \leq 0 \ \forall i$. Além disso, esperamos que $\sum_j p_{ij} = 1$. Assim,

$$1 = \sum_j p_{ij} = 1 + h\sum_j q_{ij} + 2o(h) \simeq 1 + h\sum_j q_{ij}, \tag{2.6}$$

o que nos leva a concluir que

$$\sum_j q_{ij} = 0. \tag{2.7}$$

A quantidade q_{ij} é a taxa instantânea de transição, ou seja, a taxa de mudança do estado i para o estado j . Essa taxa é usualmente definida como

$$q_{ij} = \lim_{h \to 0} \frac{P(X(t+h) = j | X(t) = i)}{h}. \tag{2.8}$$

A probabilidade condicional de $X(t+h) = j$ dado que $X(0) = i$ é dada pela equação de Chapman-Kolmogorov

$$p_{ij}(t+h) = \sum_k p_{ik}(t) p_{kj}(h). \tag{2.9}$$

Pela equação (2.9) e as suposições (2.4) e (2.5), temos que

$$\begin{aligned} p_{ij}(t+h) &= p_{ij}(t)(1 + q_{jj}h + o(h)) + \sum_{k \neq j} p_{ik}(t)(q_{ik}h + o(h)) \\ &= p_{ij}(t) + h \sum_k p_{ik}(t) q_{ki} + o(h) \\ &\simeq p_{ij}(t) + h \sum_k p_{ik}(t) q_{ki}, \end{aligned}$$

de modo que

$$\frac{1}{h}[p_{ij}(t+h) - p_{ij}(t)] \simeq \sum_k p_{ik}(t) q_{kj}. \tag{2.10}$$

Quando $h \to 0$, nós obtemos a equação de Kolmogorov progressiva

$$\frac{d}{dt} p_{ij}(t) = \sum_k p_{ik}(t) q_{kj}, \quad t \geq 0. \tag{2.11}$$

Em geral, nós podemos escrever a equação (2.11) na forma

$$\frac{d}{dt} P(t) = P(t) Q, \tag{2.12}$$

onde $P(t) = \{p_{ij}(t)\}$ é a matriz de probabilidades de transição e Q é a matriz de taxas instantâneas. Considerando que a matriz de transição no tempo inicial seja igual a matiz identidade ($P(0) = I$), temos uma única solução para a equação diferencial (2.12), a qual é dada por

$$P(t) = e^{tQ}. \tag{2.13}$$

Encontrada a matriz de transição $P(t)$, desejamos encontrar a distribuição de equilíbrio dos estados, ou seja, a distribuição estacionária. A distribuição estacionária de um determinado estado, digamos j, é dada por π_j e pode ser entendida como a probabilidade de observarmos o estado j em um ponto aleatório do tempo, quando não temos nenhum conhecimento do estado inicial da cadeia.

Para cadeias de Markov com tempo contínuo, a probabilidade estacionária é definida por $\lim_{t\to 0} p_{ij}(t) = \pi_j$, onde $\sum_j \pi_j = 1$.

As informações a respeito de processos de Markov com estados discretos e tempo contínuo, aqui apresentadas, foram basicamente extraídas de Grimmett & Stirzaker (1992), capítulo 6 e Ewens & Grant (2001), capítulo 10.

2.4.2 O modelo evolutivo de Jukes-Cantor

O modelo Jukes-Cantor (JC69), proposto por Jukes & Cantor (1969), é considerado o mais simples modelo de substituição de nucleotídeos.

Para esse modelo, a taxa instantânea de substituição do nucleotídeo i pelo nucleotídeo j é definida com $q_{ij} = u/3$ para todo $i \neq j$ e $i, j \in \{A, C, T, G\}$. A quantidade u, $u \leq 1$, é a taxa de mudança de um nucleotídeo para um dos outros três (note que $\sum_{j\neq i} q_{ij} = u$). Assim, o modelo de Jukes Cantor tem matriz $Q = \{q_{ij}\}$ dada por

$$Q = \begin{bmatrix} -u & u/3 & u/3 & u/3 \\ u/3 & -u & u/3 & u/3 \\ u/3 & u/3 & -u & u/3 \\ u/3 & u/3 & u/3 & -u \end{bmatrix}.$$

Podemos verificar que, como em (2.7), o somatório dos elementos de cada linha da matriz tem soma zero, ou seja, $\sum_j q_{ij} = 0$.

Utilizando a equação de Kolmogorov progressiva (2.11), temos que

$$\frac{d}{dt}p_{ii}(t) = -up_{ii}(t) + \frac{u}{3}\sum_{k\neq i} p_{ik}(t) = -up_{ii}(t) + \frac{u}{3}(1 - p_{ii}(t)) = \frac{u}{3}(1 - 4p_{ii}(t)). \quad (2.14)$$

A solução para a equação diferencial em (2.14), é dada por

$$p_{ii}(t) = \frac{1}{4} + ce^{-\frac{4}{3}tu}, \quad (2.15)$$

onde c é uma constante que pode ser determinada com a condição $p_{ii}(0) = 1$. Assim, encontramos para c o valor $3/4$.

Suponha que o estado i é dado pelo nucleotídeo C. Há três outros nucleotídeos para os quais C pode mudar, A, G ou T. Logo, podemos calcular p_{Cj}, para $j \neq C$ como sendo

$$p_{Cj}(t) = \frac{(1 - p_{CC})}{3} = \frac{1}{3} - \frac{1}{12} - \frac{1}{4}e^{-\frac{4}{3}tu} = \frac{1}{4} - \frac{1}{4}e^{-\frac{4}{3}tu}. \quad (2.16)$$

Por (2.15) e (2.16), concluímos que a matriz de transição para o modelo de Jukes-Cantor é dada por

$$p_{ij}(t) = \begin{cases} \frac{1}{4}(1 - e^{-\frac{4}{3}ut}) & \text{se } i \neq j \\ \frac{1}{4}(1 + 3e^{-\frac{4}{3}ut}) & \text{se } i = j. \end{cases} \quad (2.17)$$

Calculando $\lim_{t\to\infty} p_{ij}(t)$, encontramos a probabilidade estacionária $\pi_j = 1/4 \ \forall j \in \{A, C, G, T\}$, ou seja, para o modelo JC69,

$$\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}. \quad (2.18)$$

2.4.3 O modelo evolutivo de Kimura

O modelo de Kimura com dois parâmetros (K2P) foi proposto por Kimura (1980). Este é um modelo um pouco mais complexo e realista que o modelo de Jukes-Cantor.

Os dois parâmetros considerados nesse modelo são α, que indica a taxa de *transição* e β, a taxa de *transversão*. Uma *transição* é a substituição de uma purina (A ou G) por outra purina ou de uma pirimidina (T ou C) por outra pirimidina. Uma transversão é a substituição de uma purina por uma pirimidina ou de uma pirimidina por uma purina.

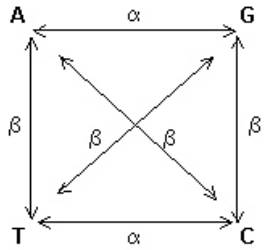

Figura 2.4: Taxas de transição e transversão para o modelo de Kimura

A matriz $Q = \{q_{ij}\}$ de taxas instantânes de substituição de nucleotídeos é dada por

$$Q = \begin{bmatrix} -\alpha - 2\beta & \alpha & \beta & \beta \\ \alpha & -\alpha - 2\beta & \beta & \beta \\ \beta & \beta & -\alpha - 2\beta & \alpha \\ \beta & \beta & \alpha & -\alpha - 2\beta \end{bmatrix}.$$

A matriz de probabilidades de transição, $P = \{p_{ij}\}$, é obtida pela resolução do sistema de equações diferenciais de Kolmogorov progressiva (2.11). Para um comprimento de ramo t e um dado nucleotídeo i, temos que

$$p_{ii}(t) = 0.25 + 0.25e^{-4\beta t} + 0.5e^{-2(\alpha+\beta)t}. \tag{2.19}$$

A expressão para $p_{ij}(t)$, com $i \neq j$, depende das escolhas de i e j. Se o nucleotídeo i é uma purina (respectivamente pirimidina) e j também é uma purina (respectivamente pirimidina), então a probabilidade $p_{ij}(t)$ é dada por

$$p_{ij}(t) = 0.25 + 0.25e^{-4\beta t} - 0.5e^{-2(\alpha+\beta)t}. \tag{2.20}$$

Se o nucleotídeo i é uma purina (respectivamente pirimidina) e j é uma pirimidina (respectivamente purina), temos que

$$p_{ij}(t) = 0.25 - 0.25e^{-4\beta t}. \tag{2.21}$$

Assim como no modelo JC69, a probabilidade estacionária $\pi_j = \lim_{t\to\infty} p_{ij}(t)$ é $1/4$ $\forall i \in \{A, C, G, T\}$. A razão de transição/transversão do modelo é denotada por r e é igual a $\alpha/(2\beta)$. Se α for igual a β e, portanto, $r = 1/2$, teremos o modelo de Jukes-Cantor, que é um caso particular do modelo de Kimura.

2.4.4 O modelo evolutivo de Felsenstein

O modelo de Felsenstein (F81), introduzido por Felsenstein (1981), é uma generalização do modelo de Jukes-Cantor (1969).

A principal suposição do modelo F81 é que a probabilidade de substituição de um nucleotídeo por um dos outros três é proporcional à probabilidade estacionária do nucleotídeo substituto. Essa hipótese define a seguinte matriz de taxas instantânes de substituição de nucleotídeos:

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

onde $k > 0$ é um parâmetro do modelo que representa a taxa total de transversão.

Essa matriz e as suposições (2.4) e (2.5), especificam um sistema equações diferenciais na forma (2.14). A solução para esse sistema é dada pelas seguintes probabilidades de transição:

$$p_{ij}(t) = \begin{cases} \pi_j(1 - e^{-kt}) & \text{se } i \neq j \\ e^{-kt} + \pi_i(1 - e^{-kt}) & \text{se } i = j. \end{cases} \tag{2.22}$$

Capítulo 3

Construção clássica de árvores filogenéticas binárias

O principal interesse da análise filogenética é responder à seguinte pergunta: *Qual é a história evolutiva mais provável para as espécies em análise?* Essa questão é respondida com o auxílo de algum método de inferência para filogenias. Neste capítulo, iremos descrever três dos mais importantes métodos clássicos de construção de árvores filogenéticas: máxima verossimilhança, máxima parcimônia e vizinhança conjunta.

3.1 Máxima Parcimônia

O método da máxima parcimônia é o mais antigo método de inferência filogenética. Sua idéia chave, dada por Edwards & Cavalli-Sforza (1963), consiste em predizer a árvore que minimiza o número de passos necessários para gerar a variação observada nos dados.

Para uma melhor compreensão do método de máxima parcimônia, vejamos um exemplo bem simples.

Considere a existência de 4 espécies, a, b, c e d, para as quais temos o alinhamento

de seqüências de DNA mostrado na tabela 3.1. Alguns sítios desse alinhamento são

Táxon	Sítios						
	1	2	3	4	5	6	7
a	A	A	G	A	G	T	C
b	A	G	C	C	G	T	C
c	A	G	A	G	A	T	C
d	A	G	A	T	A	T	C

Adaptado de David W. Mount(2001)

Tabela 3.1: Alinhamento de 4 seqüências de DNA.

informativos e outros não. Para que um sítio seja classificado como informativo, ele deve conter pelo menos 2 nucleotídeos diferentes. Assim, os sítios 1, 6 e 7 são não informativos e 2, 3, 4 e 5 são informativos.

De acordo com a equação (2.2), são possíveis 3 topologias sem raiz para 4 espécies. Dada uma dessas topologias, devemos verificar o número de mudanças nos nucleotídeos, ocorridas para cada sítio informativo do alinhamento.

A figura 3.1 ilustra as mudanças de nucleotídeos ocorridas nos sítios informativos para a topologia do tipo $((a, b), (c, d))$, denominada topologia A. O total de mudanças, também chamado de custo de substituição, é 1+2+3+1=7. Para uma topologia B, do tipo $((a, d), (b, c))$, o custo de substituição por sítio é mostrado na figura 3.2. O custo total é 1+2+3+2=8. A topologia C, dada por $((a, c), (b, d))$, tem custo total igual a 1+2+3+2=8.

Entre as topologias A, B e C, a que apresenta o menor custo total é a topologia A. Assim, podemos dizer que A explica a seqüência de dados com o menor número de passos possível e, portanto, ela será a árvore de máxima parcimônia.

Na prática, a contagem do número de mudanças de estado em uma filogenia é realizada com o auxílio de algum algoritmo. Entre os algoritmos utilizados, podemos

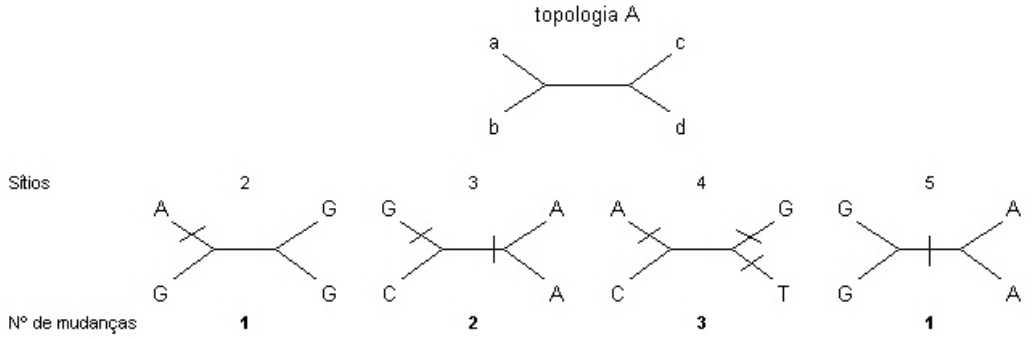

Figura 3.1: Modificações nos nucleotídeos para a topologia A. Os traços que cortam os ramos indicam essas modificações.

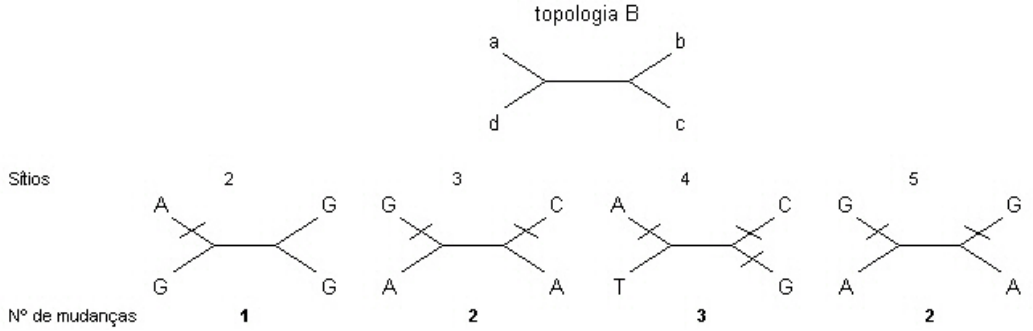

Figura 3.2: Modificações nos nucleotídeos informativos para a topologia B. Os traços que cortam os ramos indicam essas modificações.

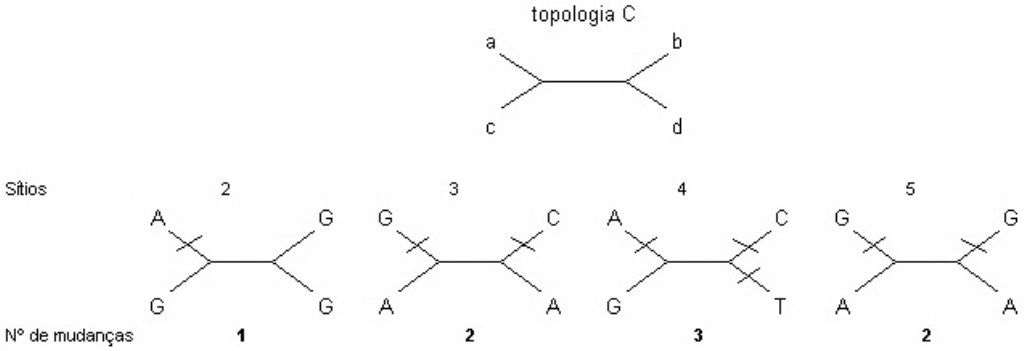

Figura 3.3: Modificações nos nucleotídeos informativos para a topologia C. Os traços que cortam os ramos indicam essas modificações.

citar, por exemplo, os de Fitch (1971), Sankoff (1975) e Sankoff & Rousseau (1975). Em seu livro, Felsenstein (2004) dedica um capítulo à "Contagem de Mudanças Evolucionárias", descrevendo algumas dessas técnicas.

3.2 Vizinhança conjunta

Considere uma matriz D, onde D_{ij} indica a exata distância entre as espécies i e j. Para n espécies, temos

$$D = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} \quad & \quad & \quad & \quad \end{array} \\ \begin{bmatrix} D_{11} & D_{12} & \dots & D_{1n} \\ D_{21} & D_{22} & \dots & D_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ D_{n1} & D_{n2} & \dots & D_{nn} \end{bmatrix} \end{array}.$$

O método de vizinhança conjunta, proposto por Saitou & Nei (1987) e modificado por Studier & Keppler (1988), utiliza essa matriz para reconstruir a árvore filogenética. Os passos desse algoritmo são descritos no capítulo 11 de Felsenstein (2004) como sendo:

1. Para cada linha da matriz D, calcule $u_i = \sum_{j:i\neq j}^{n} D_{ij}/(n-2)$.

2. Escolha as espécies i e j tais que $D_{ij} - u_i - u_j$ seja mínino.

3. Conecte as espécies i e j, formando um nó interno (ij). Calcule o comprimento dos ramos t_i, que liga o nó (ij) à espécie i, e t_j, que liga (ij) à espécie j utilizando as seguintes fórmulas:

$$t_i = \frac{1}{2}D_{ij} + \frac{1}{2}(u_i - u_j)$$
$$t_j = \frac{1}{2}D_{ij} + \frac{1}{2}(u_j - u_i).$$

4. Calcule a distância entre o nó (ij) e as espécies restantes como sendo

$$D_{(ij)k} = \frac{(D_{ik} + D_{jk} + D_{ij})}{2}.$$

5. Delete as espécies i e j da matriz de distâncias e as substitua pelo nó (ij), que será agora tratado como um nó externo, ou seja, um nó que representa uma espécie. Os elementos $D_{(ij)k} = D_{k(ij)}$ são aqueles encontrados no item anterior.

6. Se D ainda for uma matriz maior que 2x2, volte ao primeiro passo. Se D é 2x2, conecte os dois nós externos restantes (digamos, k e l) com comprimento de ramo dado por $D_{kl} = D_{lk}$.

Infelizmente, na prática, a matriz D é desconhecida. Logo, para utilizar o método de vizinhança conjunta, torna-se necessário calcular uma matriz de distâncias estimadas, denotada por $\hat{D}$.

Uma possível solução para encontrar $\hat{D}$ é dada no capítulo 14 de Ewens & Grant (2001). Se temos seqüências de DNA alinhadas para todas as espécies em questão, a distância entre duas delas será proporcional a

$$-\log\left(1 - \frac{4}{3}p\right),$$

sendo p a proporção de nucleotídeos que se diferem nas seqüências dessas duas espécies. Essa estimativa é baseada no modelo evolutivo de Jukes Cantor, o qual foi explicado na seção 2.4.2.

3.3 Máxima Verossimilhança

A utilização da máxima verossimilhança na estimação de árvores filogenéticas foi introduzida por Felsenstein (1981).

Para descrevermos este método, utilizaremos um exemplo específico. Suponha a existência de 4 espécies, a, b, c e d, sendo cada uma delas representada por uma seqüência de DNA com n sítios de um determinado gene.

$$x = \begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & \dots & n \end{array} \\ \begin{bmatrix} A & C & C & G & C & C & A & \dots & T \\ A & C & C & G & T & C & C & \dots & C \\ A & C & C & G & C & C & T & \dots & C \\ A & C & C & G & C & T & A & \dots & G \end{bmatrix} \end{array}$$

Considere uma árvore definida por uma topologia τ e um vetor de comprimentos de ramos denotado por t. Nosso interesse é computar a verossimilhança dessa árvore. O primeiro passo nessa direção é fazer algumas suposições que simplifiquem o cálculo, como por exemplo, supor que o processo de substituição de nucleotídeos em um determinado sítio seja independente dos demais sítios. Essa suposição é de grande importância pois garante que a função de verossimilhança para a matriz de dados x possa ser decomposta no produto das contribuições individuais de cada sítio, ou seja,

$$p(x \mid \tau, t) = \prod_{i=1}^{n} p(x^{(i)} \mid \tau, t), \tag{3.1}$$

onde $x^{(i)}$ é o vetor que contém os nucleotídeos observados para as espécies a, b, c e d para o sítio i.

Considere a árvore para um determinado sítio do alinhamento, digamos i, como sendo a da figura 3.4. O vetor $x^{(i)}$ é dado por (A, C, T, A).

Assuma um modelo de substituição de nucleotídeos estocástico de tempo contínuo. Sob esse modelo, π_i é a probabilidade estacionária do nucleotídeo i e $p_{ij}(t)$ é a probabilidade de substituição do nucleotídeo i pelo nucleotídeo j em um ramo de comprimento t da topologia τ. Assim,

$$p(x^{(i)} \mid \tau, t) = \sum_w \sum_y \sum_z p(A, C, T, A, w, y, z|\tau, t) \tag{3.2}$$

$$= \sum_w \sum_y \sum_z \pi_z p_{zy}(t_6) p_{zw}(t_5) p_{yA}(t_4) p_{yT}(t_3) p_{wC}(t_2) p_{wA}(t_1), \tag{3.3}$$

onde w, y, z são os nucleotídeos para as espécies ancestrais. Esses nucleotídeos são não observáveis, podendo ser do tipo A, C, G ou T.

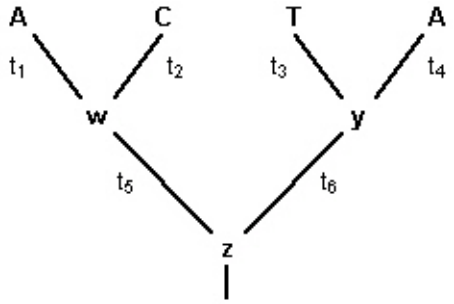

Figura 3.4: Árvore com os dados do i-ésimo sítio

Uma expressão semelhante a (3.3) deve ser calculada para cada um dos sítios do alinhamento e, posteriormente, a função de verossimilhança, dada em (3.1), deve ser calculada. Um algoritmo de maximização pode então ser utilizado para encontrar o estimador de máxima verossimilhança de t.

O procedimento descrito nessa seção deve ser repetido para todas as possíveis topologias de a, b, c e d. A árvore de máxima verossimilhança será aquela que apresentar a maior verossimilhança dentre todas as possíveis topologias.

Para maiores detalhes a respeito da construção de árvores filogenéticas via máxima verossimilhança, recomendamos ver Felsenstein (1981) e Felsentein (2004).

3.4 Medindo incerteza via bootstrap

O método de bootstrap, proposto por Efron (1979), é uma técnica utilizada para medir a incerteza na estimação de parâmetros. Esse método foi introduzido na análise filogenética pelos trabalhos de Mueller & Ayala (1982) e Felsenstein (1985). Para discorrer sobre esse assunto, nos baseamos no livro de Hjorth (1994), no artigo de Felsenstein (1985) e no livro de Felsenstein (2004).

3.4.1 Introdução ao bootstrap

Sejam $x_1, x_2, \ldots, x_n$ n observações independentes e igualmente distribuídas de uma distribuição $F(x)$. Seja θ um parâmetro de interesse, cujo valor é desconhecido. Utilizando uma estatística t baseada na amostra, podemos obter um estimador para θ, dado por

$$\hat{\theta} = t_n(x_1, \ldots, x_n). \tag{3.4}$$

Se a distribuição F é conhecida, podemos calcular a esperança do estimador pela integral

$$E_F(\hat{\theta}) = \int \ldots \int t_n(x_1, \ldots, x_n) f(x_1) \ldots f(x_n) dx_1 \ldots dx_n \tag{3.5}$$

e sua variância por

$$Var(\hat{\theta}) = E_F[(\hat{\theta} - E_F(\hat{\theta}))^2]. \tag{3.6}$$

O desvio padrão de $\hat{\theta}$ pode ser obtido pela raiz quadrada de sua variância. Com essas quantidades, é possível construir um intervalo de confiança para θ.

Quando F é desconhecida, não podemos calcular as expressões (3.5) e (3.6). Uma possível solução para esse problema é substituir a distribuição $F(x)$ pela distribuição empírica dos dados, $F_n(x)$, seguindo a idéia do método de bootstrap.

O primeiro passo do bootstrap é reamostrar n observações dos dados, com reposição. Dessa forma, uma determinada observação dos dados originais pode aparecer mais de uma vez ou mesmo não aparecer na nova amostra, representada por $x_1^*, x_2^*, \ldots, x_n^*$.

Podemos calcular o estimador de θ, dado por $\hat{\theta}_B = t_n(x_1^*, x_2^*, \ldots, x_n^*)$. O índice B do estimador indica que ele está sendo calculado para a amostra de bootstrap.

Considere que o processo descrito no parágrafo anterior seja repetido k vezes, sendo k um número grande. Teremos então k estimativas de bootstrap

$$\begin{aligned} \hat{\theta}_{B_1} &= t_{n_1}(x_{11}^*, x_{12}^*, \ldots, x_{1n}^*) \\ &\vdots \\ \hat{\theta}_{B_k} &= t_{n_k}(x_{k1}^*, x_{k2}^*, \ldots, x_{kn}^*). \end{aligned}$$

Note que $\hat{\theta}_{B_1}, \hat{\theta}_{B_2} \ldots \hat{\theta}_{B_k}$ formam uma amostra de uma distribuição que se aproxima da distribuição de θ. Assim, podemos estimar quantidades como média, viés, variância, entre outras. A média, denotada por $\hat{\theta}_{B(.)}$, é a estimativa de bootstrap para θ.

$$\hat{\theta}_{B(.)} = \frac{\sum_{i=1}^{k} \hat{\theta}_{B_i}}{k} \tag{3.7}$$

A variância da estimativa é dada por

$$Var(\hat{\theta}_{B(.)}) = E_{F_n}[(\hat{\theta}_{B(.)} - \hat{\theta})^2]. \tag{3.8}$$

As estimativas de bootstrap dadas por (3.7) e (3.8) podem ser usadas para construir um intervalo de confiança para o parâmetro θ.

3.4.2 Bootstrap na construção de árvores filogenéticas

O bootstrap também pode ser utilizado na estimação de árvores filogenéticas. Neste caso, o conjunto de dados é geralmente formado por uma matriz do tipo *espécies X caracteres*, para dados morfológicos, ou *espécies X sítios*, para dados moleculares.

Considere, por exemplo, uma matriz com m espécies e n sítios de DNA. Para utilizarmos o método de bootstrap, precisamos realizar uma reamostragem dos nossos dados. No entanto, não sabemos de imediato como aplicar esse procedimento a matrizes. A princípio, poderíamos reamostrar tanto as espécies (linhas) como os sítios (colunas).

Para a correta utilização do método, as observações das amostras de bootstrap devem ser independentes e igualmente distribuídas. Sabemos que as espécies não são independentes, pois compartilham uma história evolutiva. Logo, a melhor opção é fazer uma reamostragem dos sítios encontrados nas colunas da matriz.

Podemos considerar cada sítio como sendo uma variável aleatória independente dos demais sítios, de acordo com um processo estocástico cujos parâmetros são os comprimentos dos ramos e a topologia. Assim, as colunas da matriz de dados, denotadas por $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$, são observações independentes da distribuição de todas as 4^m possíveis configurações dos sítios. Para dados discretos, como é o caso das sequências de DNA, a distribuição dos sítios é dada por uma Multinomial.

Uma amostra de bootstrap é formada pela reamostragem de n sítios, com reposição. Temos então uma matriz do tipo m x n que pode apresentar determinadas colunas da matriz original mais de uma vez e não apresentar outras colunas. A figura 3.5 ilustra a construção de duas amostras de bootstrap.

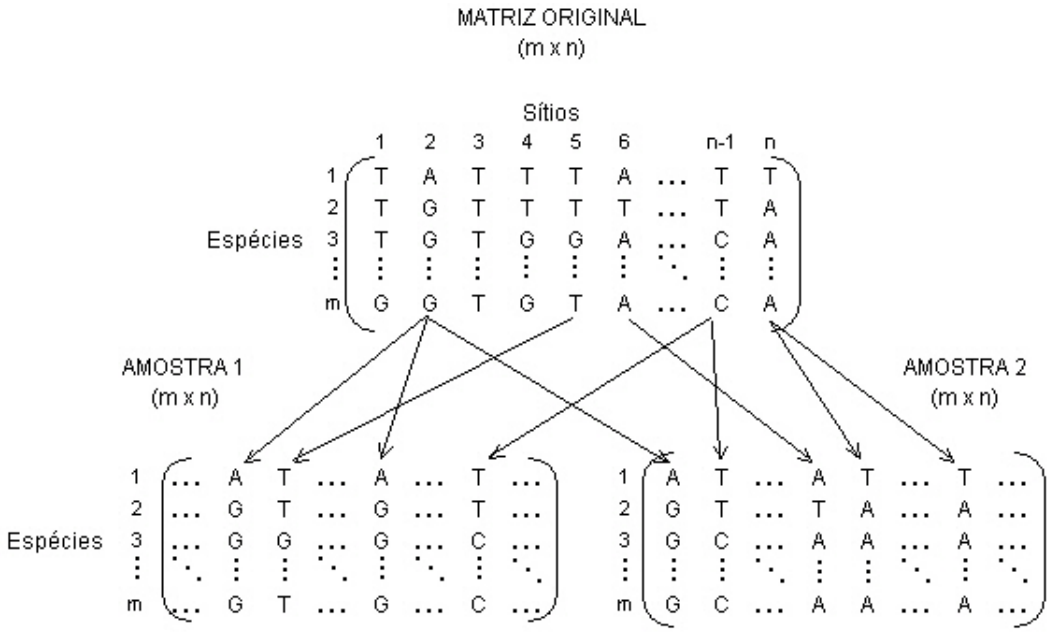

Figura 3.5: Amostras de bootstrap

Considere que são confeccionadas k amostras, sendo k um número suficientemente grande. Para cada uma dessas matrizes podemos inferir a árvore filogenética, utilizando um método como, por exemplo, máxima verossimilhança, parcimônia ou vizinhança conjunta. Como resultado temos k árvores, com as quais construímos a árvore de consenso.

Na seção seguinte, daremos ênfase à construção de uma árvore de consenso.

3.5 Árvores de consenso

Uma árvore de consenso é uma árvore que sintetiza um conjunto de árvores filogenéticas referentes a um determinado grupo de organismos.

É importante ressaltar que a árvore de consenso não é a estimada filogenia para os organismos em estudo. Na verdade, essa árvore é utilizada para que possamos avaliar o quão boa é a filogenia obtida pelo conjunto de dados original. Essa avaliação é feita com base no suporte de bootstrap dado aos grupos inseridos na árvore de consenso.

Existem diversos métodos para calcular uma árvore de consenso. Entre os mais importantes podemos citar a Árvore de Estrito Consenso, de Rohlf (1982), a Árvore de Consenso Adams, de Adams (1972, 1986) e a Árvore de Consenso Majoritária, de Margush & McMorris (1981). Nesse seção, apresentaremos somente o método de Margush & McMorris (1981). Para a obtenção de informações sobre os demais métodos, sugerimos a leitura do trigésimo capítulo de Felsenstein (2004).

3.5.1 Árvore de Consenso Majoritária

A "regra da maioria" é um método de obtenção de árvores de consenso, introduzido por Margush & McMorris (1981). A árvore resultante é composta pelos grupos que aparecem na maioria das árvores que estamos tentando sintetizar. Assim, se um grupo de organismos está presente em mais de 50% das árvores, ele deve ser inserido

na árvore de consenso.

Vejamos um exemplo bem simples, retirado do livro de Felsenstein (2004). Suponha que realizamos uma análise de bootstrap em um conjunto de seis espécies, a, b, c, d, e e f. Considere que foram realizadas cinco reamostragens dos dados e cinco árvores filogenéticas foram inferidas para essas espécies. A figura 3.6 ilustra as 5 árvores em questão. A tabela 3.2 apresenta o número de vezes que cada grupo aparece nessas árvores e a figura 3.7 ilustra a árvore de consenso construída pela regra da maioria.

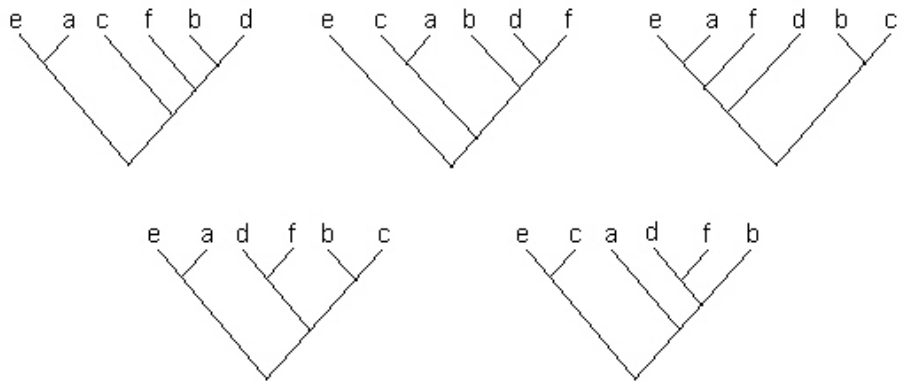

Figura 3.6: As 5 árvores inferidas para as espécies a, b, c, d, e e f

Note que os grupos $\{a, e\}$, $\{d, f\}$, $\{a, c, e\}$, $\{b, d, f\}$, $\{b, c, d, f\}$ e $\{a, b, c, e\}$ aparecem cada um 3 vezes (mais de 50% das vezes) e portanto devem estar presentes na árvore de consenso. O suporte de cada um desses grupos é de 3/5, ou seja, 60%.

grupo	repetições
ae | bcdf	3
ace | bdf	3
acef | bd	1
ac | bdef	1
aef | bcd	1
adef | bc	2
abdf | ec	1
abce | df	3

Tabela 3.2: Número de vezes que cada partição aparece nas árvores inferidas para as espécies a, b, c, d, e e f.

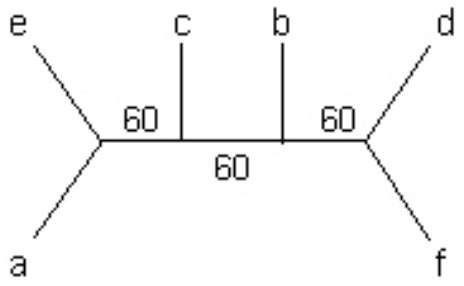

Figura 3.7: Árvore de consenso com os respectivos suportes dos galhos

Capítulo 4

Construção bayesiana de árvores filogenéticas

A inferência bayesiana utiliza-se da função de verossimilhança e de uma distribuição a priori das quantidades de interesse para calcular a probabilidade dos parâmetros condicionada nos dados observados. Na análise filogenética bayesiana, o conhecimento a respeito da topologia, dos comprimentos dos galhos e dos parâmetros de substituição de nucleotídeos é representado com distribuições de probabilidades.

Neste capítulo, faremos uma breve exposição da inferência bayesiana, introduzindo o teorema de Bayes e o uso do método de Monte Carlo via Cadeias de Markov para obtenção de uma amostra da distribuição a posteriori de interesse. As explicações que aqui serão apresentadas foram extraídas de Gamerman (1996), Felsenstein (2004) e Huelsenbeck & Ronquist (2001).

4.1 Inferência bayesiana

Seja θ uma quantidade de interesse desconhecida cujos possíveis valores são pertencentes ao conjunto Θ. O objetivo da inferência bayesiana pode ser a estimação de θ

ou o teste de alguma hipótese envolvendo valores de θ. Um dos principais ingredientes para a realização de inferência bayesiana é a distribuição a posteriori, que representa o conhecimento a respeito de θ após a observação dos dados x.

A distribuição a posteriori é obtida através do teorema de Bayes:

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)}, \tag{4.1}$$

onde $p(x) = \sum_{\theta \in \Theta} p(\theta)p(x|\theta)$. A função de densidade $p(\theta)$ é denominada priori e representa o conhecimento a respeito de θ antes de se observar os dados. A função $p(x|\theta)$ especifica a verossimilhança das observações e $p(x)$ é a probabilidade marginal desses dados. Como a quantidade $p(x)$ não depende de θ, podemos reescrever a equação (4.1) como sendo

$$p(\theta|x) \propto p(x|\theta)p(\theta). \tag{4.2}$$

4.2 Análise filogenética bayesiana

A inferência bayesiana foi introduzida na análise filogenética em meados dos anos 90. Yang & Rannala (1997) propuseram um método de reconstrução de filogenias baseado no Método de Monte Carlo via Cadeias de Markov (MCMC). Mau *et al.* (1999) também utilizaram MCMC para propor um outro método, estendido posteriormente por Larget & Simon (1999). Mais recentemente, Li *et al.* (2000) e Huelsenbeck *et al.* (2001) publicaram importantes artigos sobre esse mesmo tema.

Na análise filogenética bayesiana, o parâmetro θ da equação (4.1) é formado pela topologia τ, pelos parâmetros de substituição de nucleotídeos ϕ e pelo vetor de comprimentos de galhos t. Para dados moleculares, x será a observada matriz de seqüências de DNA/RNA alinhadas. Assim, a probabilidade a posteriori para (τ, ϕ, t) é propo-

rional a:

$$p(\tau, \phi, t|x) \propto p(x|\tau, \phi, t)p(\tau, \phi, t). \tag{4.3}$$

Sejam Φ o conjunto dos possíveis valores de ϕ e T o conjunto dos possíveis valores de t. Em particular, a probabilidade a posteriori de uma determinada topologia τ^* será obtida por

$$p(\tau^*|x) \propto \int_{\Phi} \int_{T} p(x|\tau^*, \phi, t)p(\tau^*, \phi, t)dtd\phi. \tag{4.4}$$

4.2.1 Prioris

Na literatura de construção bayesiana de árvores filogenéticas, em geral assume-se prioris independentes para (τ, ϕ, t), ou seja, assume-se $p(\tau, \phi, t) = p(\tau)p(\phi)p(t)$. Para a topologia τ, uma boa opção de distribuição a priori é a uniforme no espaço de possíveis topologias. Li *et al.* (2000) e Suchard *et al.* (2003) seguem essa abordagem. Já Yang & Rannala (1997) usam um processo de nascimento e morte para definir a priori para a topologia.

Em Husmeier & McGuire (2003), os comprimentos dos ramos são uniformemente distribuídos. O pacote computacional `MrBayes` (Huelsenbeck & Ronquist, 2001) possibilita o uso de prioris uniformes e exponenciais para esse parâmetro.

O vetor paramétrico ϕ varia de acordo com o modelo de substituição de nucleotídeos assumido. Para o modelo de Jukes Cantor, por exemplo, ϕ contém somente o vetor de probabilidades estacionárias $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$. Uma boa priori para π é dada por uma Dirichlet. Para o modelo de Kimura, a taxa de transição/transversão, r, também é um parâmetro a ser estimado. Uma possível priori para r é uma beta (1,1), utilizada como priori padrão no `MrBayes` (Huelsenbeck & Ronquist, 2001).

4.2.2 MCMC

Normalmente, é extremamente difícil ou até mesmo impossível calcular analiticamente a distribuição a posteriori de interesse. No entanto, integrais envolvendo a posteriori e funções dos parâmetros podem ser aproximadas pelo método de Monte Carlo via Cadeias de Markov, mais conhecido pela sigla MCMC.

O MCMC é um método estocástico utilizado para gerar amostras aleatórias de uma distribuição de interesse, que na inferência filogenética é a distribuição a posteriori de (τ, ϕ, t).

O Metropolis-Hastings(Metropolis *et al.*, 1953; Hastings, 1970) é um método de MCMC que tem sido bastante utilizado na análise filogenética. Aqui, daremos ênfase a versão mais utilizada desse algoritmo, que é o Metropolis-Hastings com transição por componentes. Larget & Simon (1999) e Yang & Rannala (1997) usaram essa versão para aproximar a posteriori das árvores. Huelsenbeck & Ronquist (2001) implementaram essa metodologia no software `MrBayes`, um dos mais populares softwares de análise filogenética bayesiana.

O algoritmo de Metropolis-Hastings com transição por componentes se desenvolve da seguinte forma:

1. Defina se a escolha da componente a ser atualizada será feita de forma aleatória ou pré- fixada. Para esse último caso, defina as regras da escolha da componente.

2. Escolha uma árvore inicial $\Psi = \tau_i, t_i, \phi_i$, composta por uma topologia τ_i, um vetor de comprimentos de galhos t_i e por um vetor de parâmetros de substituição de nucleotídeos ϕ_i.

3. Escolha um dos componentes de Ψ, de acordo com o item 1. Proponha uma modificação para esse componente, gerando um novo estado Ψ' da densidade $p(\Psi'|\Psi)$.

4. Calcule a probabilidade de aceitação da proposta, p, dada por :

$$p = mim\left(1, \frac{p(x|\Psi')p(\Psi')p(\Psi|\Psi')}{p(x|\Psi)p(\Psi)p(\Psi'|\Psi)}\right). \tag{4.5}$$

5. Gere uma observação de uma distribuição uniforme no intervalo (0,1). Se p≥u, aceite o estado proposto Ψ' e faça $\Psi = \Psi'$; caso contrário permaneça no estado Ψ.

6. Retorne ao terceiro passo.

O algoritmo deve ser repetido até que haja a convergência da cadeia. A convergência pode ser verificada, por exemplo, investigando se o logaritmo da posteriori conjunta tornou-se estacionário. Essa metodologia foi utilizada por Husmeier & MacGuire (2003).

4.2.3 Suporte Bayesiano

O método de Monte Carlo via Cadeias de Markov gera uma amostra aleatória para a distribuição a posteriori da árvore filogenética. A informação contida nessa amostra pode então ser resumida, dando o suporte bayesiano para os ramos que antecedem as bifurcações na filogenia. Essa sumarização pode ser feita por métodos cladísticos. Assim, podemos determinar a probabilidade a posteriori de um determinado grupo como sendo a proporção desse grupo na amostra. Considere, por exemplo, que na análise bayesiana de quatro espécies, a, b, c e d, o grupo $\{a, c\}$ aparece em 75% das árvores da amostra. A probabilidade a posteriori para esse grupo é então dada por 75%. A inferência filogenética do pacote computacional `MrBayes` (Huelsenbeck & Ronquist, 2001) também utiliza essa metodologia. Outra possibilidade é dada por Li *et al.* (2000), que usam uma medida de distância para encontrar uma árvore central para a amostra.

4.3 Exemplo: Bayes e Bootstrap

Nessa seção, apresentaremos um conjunto fictício de seqüências de DNA alinhadas, que será analisado pelos métodos bayesiano e de bootstrap. O algoritmo escolhido para análise de bootstrap foi o algoritmo de vizinhança conjunta. O conjunto de observações e a escolha dos pacotes computacionais utilizados na geração e na análise filogenética desses dados foram baseados no artigo de Suzuki *et. al* (2002), que será discutido no capítulo 5 dessa dissertação.

Suponha que desejamos investigar a história evolutiva de 4 espécies fictícias, a, b, c e d. Suponha também que a matriz de dados é composta por seqüências de DNA alinhadas, com 5000 nucleotídeos cada.

As seqüências foram geradas com o software SEQGEN (versão 1.25; Rambaut & Grassly, 1997), seguindo a árvore sem raiz dada pela figura 4.1. O modelo de substituição de nucleotídeos utilizado foi o modelo de Jukes-Cantor.

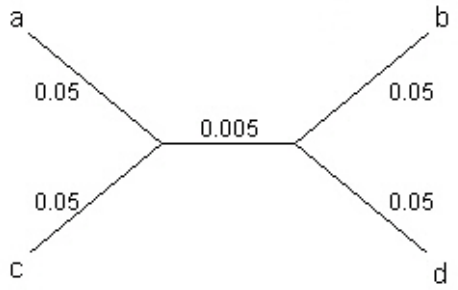

Figura 4.1: Árvore verdadeira. Os números correspondem aos comprimentos dos galhos.

Primeiramente, realizamos a análise de bootstrap pelo método da vizinhança conjunta com o software MEGA (versão 2.01; Kumar *et al.*, 2001). Foram confeccionadas 1000 amostras de bootstrap, com as quais foram estimadas árvores filogenéticas com medida de distância baseada no modelo de Jukes-Cantor. O resultado é mostrado na figura 4.2.

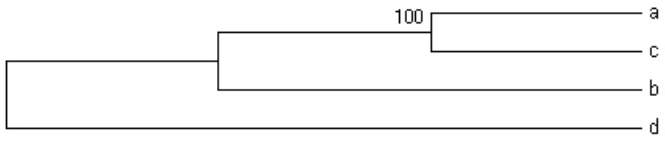

Figura 4.2: Árvore inferida pelo método de Bootstrap, utilizando o algoritmo da Vizinhança Conjunta. Os suportes de todos os ramos são iguais a 100.

Para a inferência bayesiana, utilizamos o software `MrBayes` (versão 2.1; Huelsenbeck & Ronquist, 2001). Foram realizadas 200000 iterações do MCMC, sendo uma árvore coletada a cada 100. Para a análise, foram consideradas apenas as últimas 1000 árvores das 2000 observações geradas. O modelo utilizado foi o modelo padrão do `MrBayes` (Huelsenbeck & Ronquist, 2001), o qual usa a razão de transição/transversão fixa em 0.5 (modelo JC69), priori uniforme para as topologias e priori exponencial com parâmetro igual a 10 para os comprimentos dos ramos. A árvore de consenso está ilustrada na figura 4.3.

Figura 4.3: Árvore inferida pelo método bayesiano. Os suportes dos ramos também são iguais a 100.

Para esse exemplo, bastante simplório, os dois métodos utilizados foram 100% eficientes na estimação da árvore. No entanto, em dados reais, com número maior de espécies, é comum encontrarmos divergências entre essas metodologias, principalmente no que se refere ao suporte dos ramos. Filogeneticistas freqüentemente publicam artigos com o objetivo de responder a seguinte questão: *Qual o método mais eficiente: Bayes ou Bootstrap?* Suzuki *et al.* (2002) e Alfaro *et al.* (2003) são bons exemplos dessa discussão.

Capítulo 5

O Paradoxo de Suzuki

Em dezembro de 2002, foi publicado em *Proceedings of the National Academy of Sciences* (PNAS) o artigo entitulado "Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics", de autoria de Suzuki, Glazko e Nei. Este artigo faz uma comparação entre o suporte bayesiano para a árvore inferida e o suporte de bootstrap. O objetivo dos autores, como o título já indica, é provar que a análise filogenética bayesiana superestima a credibilidade dos galhos da árvore de consenso. Neste capítulo, iremos expor os principais tópicos do artigo de Suzuki *et al.* (2002). Além disso, tentaremos reproduzir os resultados obtidos pelos autores em questão.

5.1 O artigo de Suzuki

Suzuki *et al.* (2002) consideram quatro espécies fictícias a, b, c e d, para as quais são possíveis três topologias sem raiz, A: $((a, b),(c, d))$, B: $((a, c),(b, d))$ e C: $((a, d),(c, d))$. Com o auxílio do pacote computacional SEQGEN (versão 1.25; Rambaut & Grassly, 1997), para cada uma destas árvores foram geradas quatro seqüências de DNA, cada uma com 5000 sítios. Assim, foram obtidos três conjuntos de quatro seqüências (a', b', c', d'), (a", b", c", d") e (a"', b"', c"', d"'). A figura 5.1 ilustra este procedi-

mento.

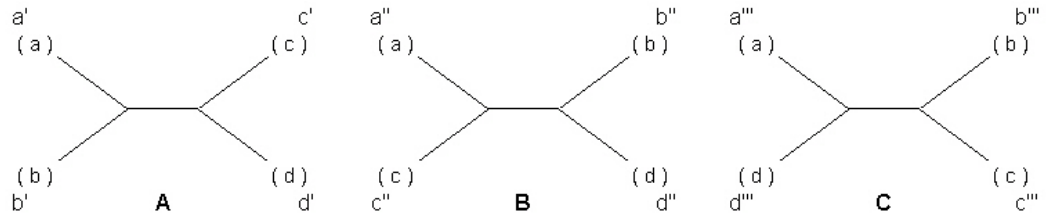

Figura 5.1: Árvores utilizadas para a geração das seqüências concatenadas

Em seguida, as seqüências *a*', *a*" e *a*"' foram concatenadas formando uma única seqüência denominada *a*, com 15000 nucleotídeos. O mesmo procedimento foi repetido para as seqüências *b*', *b*" e *b*"'; *c*', *c*" e *c*"' e *d*', *d*" e *d*"', formando as seqüências *b*, *c* e *d*, respectivamente.

A matriz de dados resultante do alinhamento das seqüências de DNA das espécies *a*, *b*, *c* e *d* foi analisada de três formas diferentes. Utilizando o pacote computacional `MrBayes` (versão 2.1; Huelsenbeck & Ronquist, 2001), foi realizada uma inferência bayesiana dos dados, em que 2000000 de árvores foram geradas, sendo descartadas as primeiras 1000000. Dentre as restantes, uma árvore a cada 100 geradas foi incluída na amostra usada para estimar a árvore de consenso. Foram também realizadas duas análises de bootstrap, uma pelo método da vizinhança conjunta e outra pelo método de máxima verossimilhança. Para a primeira, foi usado o pacote computacional MEGA (versão 2.01; Kumar *et al.*, 2001) e para a segunda o PAUP (versão 4.0b8a; Swofford, 1995). Nos dois casos, foram feitas 1000 reamostragens dos dados.

Suzuki *et al.* (2002) acreditam que utilizando estas seqüências concatenadas, as topologias *A*, *B* e *C* deveriam ser escolhidas com igual probabilidade, e que isto só não ocorreria devido ao erro estocástico da substituição de nucleotídeos. Desta forma, para eles, o suporte bayesiano ou o suporte bootstrap para a árvore inferida não poderia ser muito alto. Mais especificamente, eles classificaram um resultado como falso-positivo quando a probabilidade a posteriori ou a probabilidade bootstrap

superou 95%.

Foram analisados conjuntos de dados gerados com diferentes razões de transição/transversão e comprimentos de galhos internos e externos. Para cada método, bayesiano, máxima verossimilhança e vizinhança conjunta, a inferência foi realizada 50 vezes. A tabela 5.1 apresenta os resultados obtidos no artigo em questão.

			Análise Bayesiana					Vizinhança Conjunta					Máxima Verossimilhança				
b_E	b_I	R	A	B	C	P	M	A	B	C	P	M	A	B	C	P	M
0.05	0.005	0.5(0.5)	8	5	8	21	0.85	1	0	1	2	0.63	0	0	0	0	0.64
			16	14	20	50		17	13	20	50		16	14	20	50	
0.1	0.01	0.5(0.5)	6	7	7	20	0.85	1	0	0	1	0.64	0	0	0	0	0.65
			18	20	12	50		16	19	15	50		18	20	12	50	
0.05	0.005	5(0.5)	14	13	9	36	0.91	0	0	0	0	0.62	0	0	0	0	0.63
			19	19	12	50		17	20	13	50		19	19	12	50	
0.1	0.01	5(0.5)	12	14	11	37	0.95	0	1	1	2	0.68	0	0	1	1	0.69
			18	16	16	50		19	18	13	50		18	16	16	50	
0.05	0	5(0.5)	11	14	6	31	0.89	1	0	1	2	0.68	1	0	1	2	0.66
			18	21	11	50		18	22	10	50		18	21	11	50	
0.1	0	5(0.5)	11	13	15	39	0.95	0	0	1	1	0.69	0	0	1	1	0.68
			18	16	16	50		16	18	16	50		18	16	16	50	

Tabela 5.1: Freqüência de falsos-positivos nas análises bayesiana e de bootstrap. b_E indica o comprimento dos galhos externos; b_I o comprimento dos galhos internos; R a razão de transição/transversão usada na geração da seqüência (o valor de R utilizado na inferência filogenética é dado entre parênteses); P indica a proporção de falsos-positivos e M a média das máximas probabilidades de todas as replicações.

Podemos perceber que a inferência bayesiana produziu um número de falsos-positivos bem superior ao produzido pelas duas análises de bootstrap. Além disso, para todos os conjuntos de dados, a média das máximas probabilidades a posteriori foi de pelo menos 85%, enquanto a média das probabilidades de bootstrap foi inferior a 70%.

Esses resultados levaram Suzuki *et al.* (2002) a concluírem que a análise filogenética bayesiana, quando utilizada em seqüências concatenadas, produz uma pro-

babilidade muito alta para galhos da árvore de consenso. Como a maior árvore já construída para os mamíferos (Murphy *et al.*, 2001) e a maior para as plantas (Karol *et al.*, 2001) foram produzidas através da inferência bayesiana de seqüências de nucleotídeos concatenadas, os autores também concluíram que essas árvores podem ainda não estarem resolvidas. Essas conclusões são bastante polêmicas e, portanto, não são aceitas por toda a comunidade científica.

É importante destacar que, para a construção das árvores dos mamíferos e das plantas, foram concatenadas seqüências de *genes* diferentes, todos com a mesma história evolutiva. O pacote computacional `MrBayes` (Huelsenbeck & Ronquist, 2001) oferece ao usuário a opção de utilizar seqüências concatenadas, desde que "todas as suas partições sejam provenientes da mesma topologia", como é caso dos dados utilizados para construir as árvores dos mamíferos e das plantas. É possível especificar diferentes modelos para diferentes partições, fazendo distinção entre sequências de DNA e RNA, por exemplo. Desta forma, o software acima mencionado oferece um bom suporte para analisar os dados de Murphy *et al.* (2001) e de Karol *et al.* (2001).

Para a construção da árvore evolutiva para as espécies a, b, c e d, foram concatenadas seqüências de 3 *topologias* distintas. Vale ressaltar que estes dados são completamente diferentes dos dados utilizados por Murphy *et al.* (2001). Além disso, a análise filogenética das seqüências de nucleotídeos de Suzuki *et al.* (2002) foi realizada assumindo erroneamente que as mesmas não eram concatenadas e haviam sido geradas de uma única topologia.

5.2 Reproduzindo os resultados de Suzuki

Nesta seção, tentamos reproduzir os resultados da tabela 5.1, referentes ao artigo de Suzuki *et al.* (2002).

Para isso, geramos 4 seqüências de DNA, a, b, c e d, seguindo os mesmos pas-

sos descritos nos dois primeiros parágrafos da seção 5.1. Em seguida, analisamos o conjunto de dados pelos métodos bayesiano e de vizinhança conjunta, sob as mesmas condições de Suzuki *et al.* (2002).

Os resultados que obtivemos foram bem próximos aos resultados de Suzuki *et al.* (2002), como pode ser visto na tabela 5.2. Devido a esse fato e ao enfoque bayesiano dessa dissertação, não consideramos necessária a repetição da análise de máxima verossimilhança dos dados.

			Análise Bayesiana					Vizinhança Conjunta				
b_E	b_I	R	A	B	C	P	M	A	B	C	P	M
0.05	0.005	0.5(0.5)	2	6	7	15	0.79	0	0	0	0	0.58
			13	22	15	50		14	22	14	50	
0.1	0.01	0.5(0.5)	4	5	11	20	0.85	0	1	0	1	0.64
			10	20	20	50		11	19	20	50	
0.05	0.005	5(0.5)	12	12	13	37	0.93	0	0	0	0	0.64
			16	17	17	50		15	15	20	50	
0.1	0.01	5(0.5)	13	16	14	43	0.98	0	0	2	2	0.64
			14	18	18	50		14	19	17	50	
0.05	0	5(0.5)	15	7	4	26	0.89	0	1	0	1	0.63
			21	18	11	50		21	15	14	50	
0.1	0	5(0.5)	9	16	8	33	0.92	0	0	0	0	0.64
			15	21	14	50		16	22	12	50	

Tabela 5.2: Freqüência de falsos-positivos nas análises bayesiana e de bootstrap. b_E indica o comprimento dos galhos externos; b_I o comprimento dos galhos internos; R a razão de transição/transversão usada na geração da seqüência (o valor de R utilizado na inferência filogenética é dado entre parênteses); P indica a proporção de falsos-positivos e M a probabilidade média de todas as replicações.

Capítulo 6

A explicação para o Paradoxo de Suzuki

Neste capítulo, apresentamos uma explicação para o resultado paradoxal encontrado por Suzuki *et al.* (2002).

6.1 Modelo de Mistura

Considere uma distribuição indexada por um parâmetro de interesse θ, com função de densidade $p(x|\theta)$. Sejam duas amostras aleatórias dessa distribuição, cada uma gerada com um particular valor de θ:

$$x_1, \ldots, x_n|\theta = \theta_1 \overset{iid}{\sim} p(x|\theta = \theta_1) \tag{6.1}$$

$$y_1, \ldots, y_n|\theta = \theta_2 \overset{iid}{\sim} p(y|\theta = \theta_2) \tag{6.2}$$

Erroneamente, como feito por Suzuki *et al.* (2002), assuma que $x_1, x_2, \ldots, x_n, y_1, y_2 \ldots, y_n \overset{iid}{\sim} p(x|\theta)$. Desta forma, a função de verossimilhança será dada por:

$$p(\vec{x_n}, \vec{y_n}|\theta) = \prod_{i=1}^{n} p(x_i|\theta)p(y_i|\theta), \tag{6.3}$$

onde $\vec{x_n} = (x_1, \ldots, x_n)$ e $\vec{y_n} = (y_1, \ldots, y_n)$.

Como em Suzuki *et al.* (2002), vamos testar as hipóteses $H_0 : \theta = \theta_1$ versus $H_1 : \theta = \theta_2$, utilizando os dados gerados segundo as equações (6.1) e (6.2), mas com o modelo dado pela equação (6.3). A probabilidade a posteriori da hipótese $H_0 : \theta = \theta_1$ será:

$$p(\theta = \theta_1|\vec{x_n}, \vec{y_n}) = \frac{p(\vec{x_n}, \vec{y_n}|\theta = \theta_1)p(\theta = \theta_1)}{p(\vec{x_n}, \vec{y_n}|\theta = \theta_1)p(\theta = \theta_1) + p(\vec{x_n}, \vec{y_n}|\theta = \theta_2)p(\theta = \theta_2).} \tag{6.4}$$

Assumindo que as duas hipóteses sejam equiprováveis a priori, ou seja, $p(\theta = \theta_1) = p(\theta = \theta_2)$, a razão de chances a posteriori será igual à razão de verossimilhanças:

$$\frac{p(\theta = \theta_1|\vec{x_n}, \vec{y_n})}{p(\theta = \theta_2|\vec{x_n}, \vec{y_n})} = \frac{p(\vec{x_n}, \vec{y_n}|\theta = \theta_1)}{p(\vec{x_n}, \vec{y_n}|\theta = \theta_2)}. \tag{6.5}$$

Podemos reescrever a razão da função de verossimilhanças como:

$$\frac{p(\vec{x_n}, \vec{y_n}|\theta_1)}{p(\vec{x_n}, \vec{y_n}|\theta_2)} = \frac{\prod_{i=1}^{n} p(x_i|\theta_1)p(y_i|\theta_1)}{\prod_{i=1}^{n} p(x_i|\theta_2)p(y_i|\theta_2)} = \frac{\prod_{i=1}^{n} p(x_i|\theta_1)/p(x_i|\theta_2)}{\prod_{i=1}^{n} p(y_i|\theta_2)/p(y_i|\theta_1)} \tag{6.6}$$

O paradoxo de Suzuki está relacionado ao comportamento da expressão (6.6) quando o tamanho da amostra vai para infinito. Neste caso, Suzuki *et al.* (2002) assumem que a razão de verossimilhanças deveria convergir para 1, ou seja, que os dados forneceriam a mesma evidência a favor do modelo H_0 e do modelo H_1. No entanto, mostraremos a seguir que este é um raciocínio equivocado.

Seja l_n o logaritmo da razão de verossimilhança. Então,

$$l_n = \ln\left[\frac{p(\vec{x_n}, \vec{y_n}|\theta_1)}{p(\vec{x_n}, \vec{y_n}|\theta_2)}\right] = \sum_{i=1}^{n}\{[\ln p(x_i|\theta_1) - \ln p(x_i|\theta_2)] - [\ln p(y_i|\theta_2) - \ln p(y_i|\theta_1)]\} \tag{6.7}$$

Suponha que o valor esperado da evidência fornecida por x_i a favor da hipótese H_0 contra o modelo H_1 seja igual ao valor esperado da evidência fornecida por y_i a favor da hipótese H_1 contra o modelo H_0. Suponha também que esta evidência seja medida pela discrepância logarítmica, ou seja,

$$E[\ln p(x_i|\theta_1) - \ln p(x_i|\theta_2)] = E[\ln p(y_i|\theta_2) - \ln p(y_i|\theta_1)]. \tag{6.8}$$

Definindo $g_i = \ln\left[\frac{p(x_i|\theta_1)/p(x_i|\theta_2)}{p(y_i|\theta_2)/p(y_i|\theta_1)}\right]$, podemos escrever l_n como sendo $l_n = l_{n-1} + g_n$, com $E(g_n) = 0$. Isto nos garante que l_n é um passeio aleatório e, portanto, um processo divergente (ver, por exemplo, Fristedt & Gray, 1997). Logo,

$$P[\lim_{n\to\infty} l_n = -\infty] = P[\lim_{n\to\infty} l_n = \infty] = \frac{1}{2}. \tag{6.9}$$

Como $g(z) = e^z$ é uma função contínua, podemos fazer

$$P[\lim_{n\to\infty} e^{l_n} = 0] = P[\lim_{n\to\infty} e^{l_n} = \infty] = \frac{1}{2}, \tag{6.10}$$

e portanto,

$$P(\lim_{n\to\infty} p(\theta = \theta_1|\vec{x_n}, \vec{y_n}) = 0) = P(\lim_{n\to\infty} p(\theta = \theta_2|\vec{x_n}, \vec{y_n}) = 1) = \frac{1}{2}. \tag{6.11}$$

Dessa forma, um dos modelos H_0 ou H_1 será escolhido com probabilidade 1 quando o tamanho da amostra for para infinito. Esse foi exatamente o resultado encontrado por Suzuki *et al.* (2002).

O resultado paradoxal encontrado acima é fruto de uma modelagem inadequada dos dados. Na verdade, como a amostra $x_1, \ldots, x_n, y_1, \ldots, y_n$ é proveniente da mistura das amostras de dois componentes com densidades distintas, o modelo mais apropriado a ser assumido é um modelo de mistura, o qual é dado por

$$x_1, \ldots, x_n, y_1, \ldots, y_n \overset{iid}{\sim} \lambda p(x|\theta_1) + (1-\lambda)p(y|\theta_2) \tag{6.12}$$

Para o exemplo apresentado nesta seção, teríamos $\lambda = \frac{1}{2}$.

O caso de misturas de topologias é análogo aos exemplos acima. Assumido um modelo equivocado, ou seja, um modelo que não leva em conta o fato de termos seqüências concatenadas, longas seqüências tendem a levar à conclusão de que, com alta probabilidade a posteriori, uma particular topologia é a verdadeira. O modelo correto seria um modelo de mistura, como o que será apresentado no próximo capítulo.

Neste momento, uma importante questão pode ser levantada: *Se o modelo utilizado no artigo de Suzuki não é um modelo apropriado, porque então a análise de*

bootstrap funciona? Isso ocorre porque o bootstrap não "acredita" no modelo que está sendo utilizado. A árvore inferida para cada amostra de bootstrap irá variar de acordo com a proporção de sítios amostrados de cada uma das partições dos dados (1-5000, 5001-10000, 10001-15000), que representam três diferentes topologias. Como a reamostragem dos dados é feita aleatoriamente, é pouco provável que uma bifurcação pertencente a uma determinada topologia seja estimada em um número muito grande de amostras de bootstrap, digamos, por exemplo, em 95% dessas amostras. Na verdade, o baixo suporte dado aos ramos pela análise de bootstrap é um indício de que devemos suspeitar da adequação do modelo que está sendo utilizado.

6.1.1 Exemplo com mistura de normais

Vamos agora analisar um exemplo prático, com dados gerados de distribuições $N(\theta, \sigma)$, ou seja, normais com média θ e variância σ.

Seja $x_1, \ldots, x_{5000}$ uma amostra aleatória da distribuição $N(3, 1)$ e $y_1, \ldots, y_{5000}$ uma amostra aleatória da distribuição $N(-3, 1)$. Agrupando as duas amostras, teríamos uma amostra de tamanho 10000, dada por $x_1, \ldots, x_{5000}, y_1, \ldots, y_{5000}$. Vamos denotar essa amostra concatenada por $z_1, \ldots, z_{10000}$. Considere, erroneamente, que essas 10000 observações foram geradas da distribuição $N(\theta, 10)$, onde 10 é o valor aproximado da variância amostral de $z_1, \ldots, z_{10000}$.

Nesse caso, a função de verossimilhança é dada por

$$f(z; \theta) = \left(\frac{1}{\sqrt{20\pi}}\right)^{10000} e^{-\frac{1}{20}\sum_{i=1}^{10000}(z_i-\theta)^2} \tag{6.13}$$

Queremos investigar se a evidência da amostra $z_1, \ldots, z_{10000}$ a favor de $\theta = 3$ é igual à evidência de $\theta = -3$, ou seja, se a razão de verossimilhanças

$$r = \frac{e^{-\frac{1}{20}\sum_{i=1}^{10000}(z_i-3)^2}}{e^{-\frac{1}{20}\sum_{i=1}^{10000}(z_i+3)^2}} = exp\left\{-\frac{1}{20}\sum_{i=1}^{10000}(z_i-3)^2 + \frac{1}{20}\sum_{i=1}^{10000}(z_i+3)^2\right\}, \tag{6.14}$$

converge para 1, e, portanto, seu logaritmo

$$\ln r = -\frac{1}{20}\sum_{i=1}^{10000}(z_i - 3)^2 + \frac{1}{20}\sum_{i=1}^{10000}(z_i + 3)^2 = 0.6\sum_{i=1}^{10000} z_i \tag{6.15}$$

converge para 0. Para averiguar esse comportamento, realizamos um experimento que consiste em gerar computacionalmente a amostra $z_1, \ldots, z_{10000}$ e calcular o valor de $\ln r$. A tabela 6.1 mostra os os valores de $\ln r$ para as análises de 10 conjuntos de dados.

Experimento	1	2	3	4	5	6	7	8	9	10
$\ln r$	-77.74	75.07	-43.87	7.35	67.72	22.14	-70.57	39.22	44.13	-44.87

Tabela 6.1: Valores de $\ln r$ para as análises de 10 conjuntos de dados.

Podemos verificar que, em todos os experimentos apresentados na tabela 6.1, $\ln r$ está distante de zero, o que significa que os dados estão privilegiando $\theta = 3$ sobre $\theta = -3$ ou $\theta = -3$ sobre $\theta = 3$. Para 1000 replicações do experimento, somente em 15 delas observamos $\ln r$ próximo de zero. Se aumentamos o tamanho da amostra, o número relativo de vezes em que $\ln r$ fica próximo de zero tende a ser ainda menor. Para uma amostra de 100000 observações (50000 para $\theta = 3$ e 50000 para $\theta = -3$), verificamos que somente em 2 conjuntos de dados $\ln r$ fica próximo de zero.

Capítulo 7

Solução bayesiana: mistura de topologias

Neste capítulo, discutiremos a implementação de um modelo de mistura para os dados simulados no capítulo 5, seção 5.2.

O pacote computacional utilizado para detectar as 3 topologias das seqüências a, b, c e d foi o BARCE (versão 1.00b; Husmeier & McGuire, 2003). Esse programa faz uso de uma metodologia estatística bayesiana para identificar recombinações, ou seja, mistura de topologias em alinhamentos de DNA.

Na seção seguinte, faremos uma breve exposição do modelo de mistura utilizado no BARCE (Husmeier & McGuire, 2003).

7.1 O modelo estatístico de Husmeier & McGuire

Considere uma matriz de dados x, formada pelo alinhamento de m seqüências de DNA, cada uma com n nucleotídeos. Seja $x^{(i)}$ o vetor de nucleotídeos para o i-ésimo sítio, dado pela i-ésima coluna da matriz. Assim, temos que $x = (x^{(1)} \dots x^{(n)})$. Sejam k o número de possíveis topologias sem raiz para essas m espécies e $\{T_1, \dots, T_k\}$ o

conjunto dessas topologias.

Seja τ_i a topologia para o i-ésimo sítio do alinhamento, com $\tau_i \in \{T_1, \ldots, T_k\}$. O objetivo do modelo de Husmeier & McGuire (2003) é, com base em x, estimar a seqüência "ótima" de topologias ao longo dos sítios. O modelo estatístico assumido é um Modelo de Markov com Estados Escondidos, sendo esses estados as topologias.

Indicamos por t_i o vetor de comprimentos de galhos para τ_i e por θ_i o vetor de parâmetros para um modelo de substituição de nucleotídeos previamente definido. O modelo apresenta também um parâmetro ν, que é a dificuldade de mudança de topologia, isto é, a probabilidade que a topologia do i-ésimo sítio não se modifique quando passamos de i para um dos estados adjacentes, $i-1$ ou $i+1$.

Por conveniência, defina $\tau = (\tau_1, \ldots, \tau_n)$, $t = (t_1, \ldots, t_n)$ e $\theta = (\theta_1, \ldots, \theta_n)$. A predição de τ é baseada na probabilidade a posteriori $p(\tau|x)$, a qual pode ser obtida pela integral

$$p(\tau|x) = \int_t \int_\theta \int_\nu p(\tau, t, \theta, \nu|x) dt d\theta d\nu = \int_t \int_\theta \int_\nu \left[\frac{p(\tau, t, \theta, \nu, x)}{p(x)}\right] dt d\theta d\nu. \tag{7.1}$$

Como a quantidade $p(x)$ não depende dos parâmetros de interesse, ela é considerada uma constante. Assim,

$$p(\tau|x) \propto \int_t \int_\theta \int_\nu p(\tau, t, \theta, \nu, x) dt d\theta d\nu. \tag{7.2}$$

Supondo independência a priori dos parâmetros t, θ e ν, a probabilidade conjunta de todas as quantidades aleatórias pode ser obtida pela seguinte fatorização:

$$p(x, \tau, t, \theta, \nu) = \left[\prod_{i=1}^{n} p(x^{(i)}|\tau_i, t_i, \theta_i)\right] \left[\prod_{i=2}^{n} p(\tau_i|\tau_{i-1}, \nu) p(t_i) p(\theta_i)\right] p(\nu) p(\tau_1). \tag{7.3}$$

A contribuição do i-ésimo sítio para a verossimilhança, dada por $p(x^{(i)}|\tau_i, t_i, \theta_i)$, pode ser calculada pela fórmula dada em (3.3). A quantidade $p(\tau_i|\tau_{i-1}, \nu)$ é a probabilidade de transição entre os estados e é dada por

$$p(\tau_i|\tau_{i-1}, \nu) = \nu \delta(\tau_i, \tau_{i-1}) + \frac{1-\nu}{k-1}[1 - \delta(\tau_i, \tau_{i-1})], \tag{7.4}$$

onde $\delta(\tau_i, \tau_{i-1})$ representa a função delta de Kronecker, que assume o valor 1 quando $\tau_i = \tau_{i-1}$ e 0, caso contrário. Assim,

$$p(\tau_i|\tau_{i-1}, \nu) = \begin{cases} \nu & \text{se } \tau = \tau_{i-1} \\ \frac{1-\nu}{k-1} & \text{se } \tau \neq \tau_{i-1} \end{cases} \tag{7.5}$$

As probabilidades $p(\tau_i)$, $p(t_i)$, $p(\theta_i)$ e $p(\nu)$ contêm as informações a priori a respeito dos parâmetros e serão descritas nos próximos parágrafos.

A priori, a topologia para o primeiro sítio do alinhamento é uniformemente distribuída no conjunto das possíveis topologias, ou seja

$$p(\tau_1 = \tau^*) = \frac{1}{k} \quad \forall \tau^* \in \{T_1, \ldots, T_k\}. \tag{7.6}$$

O vetor de comprimentos dos galhos de τ_i tem como distribuição a priori uma uniforme no intervalo $[0, 1]$.

A distribuição a priori de θ depende do modelo de substituição de nucleotídeos escolhido. Suponha, por exemplo, o modelo de Jukes Cantor. Nesse caso, o vetor de parâmetros θ é dado por

$$\theta = (\pi_A, \pi_C, \pi_G, \pi_T).$$

A priori para as freqüências de nucleotídeos , π_j, com $j = A, C, G, T$, é escolhida para ser uma dirichlet(1,1,1,1), ou seja,

$$p(\pi_A, \pi_C, \pi_G, \pi_T) \propto 1, \qquad \pi_A + \pi_C + \pi_G + \pi_T = 1. \tag{7.7}$$

Para o parâmetro de recombinação, ν, é utilizada uma priori conjugada com distribuição beta, cujos hiperparâmetros são α e β.

$$p(\nu) = B(\nu|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \nu^{\alpha-1}(1 - \nu)^{\beta-1}, \qquad 0 \leq \nu \leq 1. \tag{7.8}$$

Definidas as prioris, podemos retornar ao cálculo da probabilidade a posteriori para o vetor de topologias, dada na equação (7.1). Na prática, essa quantidade não pode

ser tratada analiticamente. Contudo, ela pode ser aproximada pelo método de Monte Carlo via Cadeias de Markov. Para amostrar da distribuição a posteriori conjunta, $p(\tau, t, \theta, \nu|x)$, Husmeier & McGuire (2003) utilizaram os algoritmos de Gibbs-com-Gibbs e de Metropolis-Hastings. Mais especificamente, cada τ_i, $i = 1, \ldots, n$, foi amostrado separadamente, condicional aos demais estados, com o esquema Gibbs-com-Gibbs. Para os parâmetros t e θ, foi utilizado o algoritmo de Metropolis-Hastings. A amostragem de ν é direta de sua distribuição a posteriori, uma beta conjugada com hiperparâmetros $(\sum_{i=1}^{n-1} \delta(\tau_i, \tau_{i+1}) + \alpha)$ e $(n - 1 - \sum_{i=1}^{n-1} \delta(\tau_i, \tau_{i+1}) + \beta)$.

7.1.1 Aplicação do modelo de misturas ao paradoxo de Suzuki

Os dados gerados por Suzuki *et al.* (2002) deveriam ter sido analisados com um modelo de misturas. Fazemos isso nesta seção, utilizando o pacote filogenético BARCE (Husmeier & McGuire, 2003), que implementa uma análise bayesiana para o modelo de misturas para quatro espécies.

Husmeier & McGuire (2003) sugerem que, antes de utilizar o BARCE, devemos analisar as seqüências por outros métodos, como o RECPARS (Hein, 1993) ou o TOPAL (McGuire *et al.*, 1997; McGuire & Wright, 2000). Esses métodos, cujo grau de resolução é mais baixo que o do BARCE (Husmeier & McGuire, 2003), são usados para verificar a suposta existência de recombinações nos dados. A resposta fornecida é uma aproximação da estrutura das seqüências. Essa estrutura é então utilizado para iniciar o BARCE (Husmeier & McGuire, 2003), que visa estimar a "exata" natureza do processo de recombinação dos dados.

O procedimento descrito no parágrafo anterior não é obrigatório. No etanto, quando utilizamos uma estrutura de recombinação aleatória para iniciar a análise no BARCE (Husmeier & McGuire, 2003), o tempo necessário para que a Cadeia de Markov convirja pode ser extremamente longo.

O mais importante resultado fornecido pela análise bayesiana do BARCE (Hus-

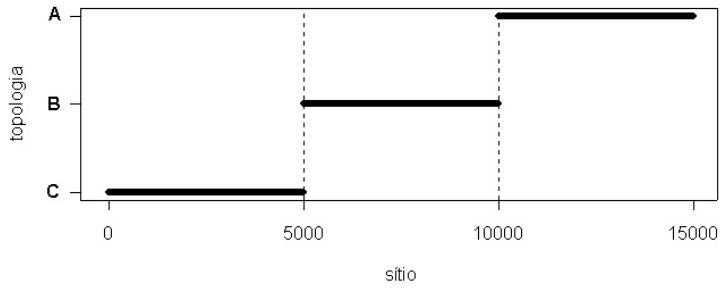

Figura 7.1: Mosaico da real estrutura dos dados de Suzuki *et al.* (2002).

meier & McGuire, 2003) é uma matriz de probabilidades, cujas linhas representam os sítios e as colunas são as 3 possíveis topologias sem raiz para 4 espécies. Para os dados de Suzuki *et al.* (2002), a exata matriz de probabilidades e os gráficos para as suas colunas são ilustrados na figura 7.2. Cabe ressaltar que as topologias A, B e C são as mesmas que estão representadas na figura 5.1, do capítulo 5.

	topologia A	topologia B	topologia C
1	1	0	0
5000	1	0	0
5001	0	1	0
10000	0	1	0
10001	0	0	1
15000	0	0	1

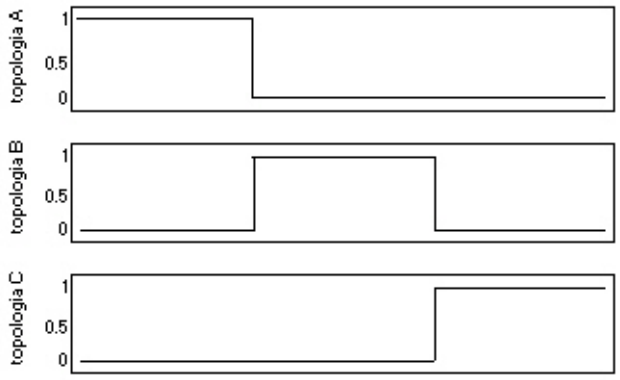

Figura 7.2: Exata matriz de posterioris para os dados de Suzuki *et al.* (2002)

O programa ainda fornece arquivos contendo a evolução para os comprimentos de galhos das topologias, a evolução das probabilidades estacionárias, do raio de transição/transversão, do parâmetro ν e do logaritmo da posteriori.

7.2 Análise dos dados simulados

Como dito anteriormente, as seqüências de DNA utilizadas por Suzuki *et al.* (2002) são resultantes da concatenação de 3 seqüências de 5000 sítios, provenientes das topologias A, B e C, nessa mesma ordem. Cada seqüência possui duas mudanças de topologias, uma de A para B e outra de B para C. Assim, a probabilidade de mudança de topologia ao longo dos 15000 sítios pode ser estimada como sendo 2/15000, ou seja, aproximadamente 0.0001. A probabilidade de que a topologia não se altere quando passamos de um sítio para o seu sucessor, dada pelo parâmetro ν do modelo de Husmeier & McGuire (2003), é estimada como aproximadamente 0.9999. Em nossos estudos, testamos a performance de três prioris para ν: beta com média 0.95, beta com média 0.99 e beta com média 0.999. Como no BARCE (Husmeier & McGuire, 2003) o hiperparâmetro β da equação (7.8) é fixo em 2, as variâncias dessas prioris são calculadas como sendo aproximadamente 115.8536×10^{-5}, 4.9254×10^{-5} e 0.0499×10^{-5}, respectivamente.

7.2.1 Análise do conjunto de dados número 1

O primeiro conjunto de dados escolhido para ser analisado no BARCE (Husmeier & McGuire, 2003) foi um dos alinhamentos de DNA usados para construir a primeira linha da tabela 5.2. Para as 3 topologias, os galhos externos têm comprimentos iguais a 0.05 e os galhos internos 0.005. Os dados foram gerados sob o modelo de Jukes-Cantor. Para a análise de bootstrap com algoritmo de vizinhança conjunta, esses dados produziram a topologia B com suporte de 55% para os ramos. Na análise filogenética bayesiana realizada pelo `MrBayes` (Huelsenbeck & Ronquist, 2001), a topologia estimada também foi B com suporte de 94% para os ramos.

Primeiramente, realizamos uma análise dos dados com mosaico aleatório para a estrutura de recombinação das seqüências. O modelo de substituição de nucleotídeos

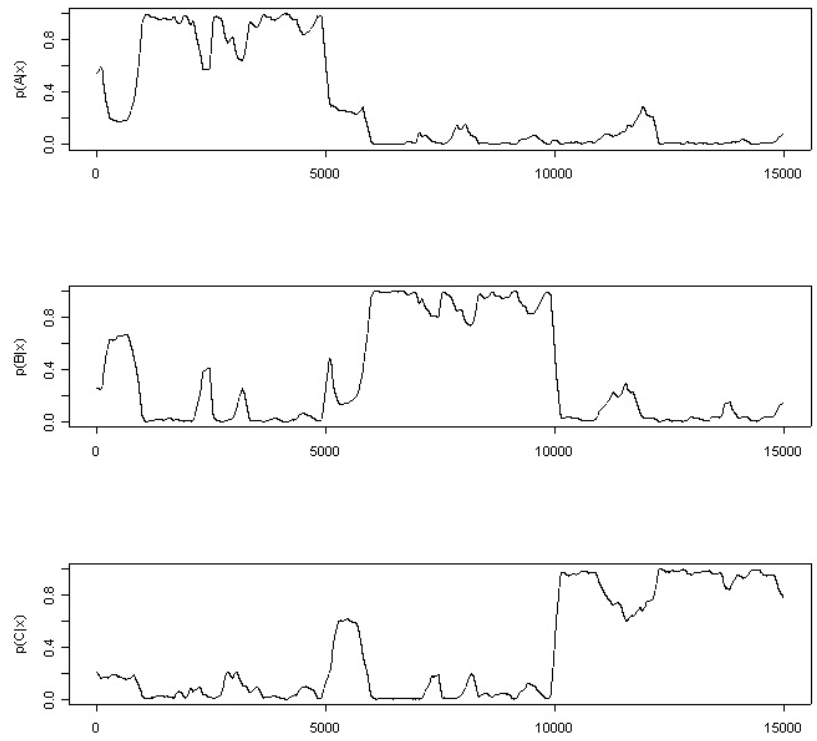

Figura 7.3: O gráfico no topo da figura representa $p(A|x)$ ao longo dos sítios. O gráfico do meio refere-se a $p(B|x)$ e o da parte inferior a $p(C|x)$. A análise dos dados foi realizada com mosaico aleatório e média(ν)=0.99

usado foi o de Jukes-Cantor. Foram realizadas 2000000 iterações do MCMC sendo as primeiras 1000000 delas descartadas. Os parâmetros foram amostrados a cada 100 iterações, gerando uma amostra de tamanho 10000. O tempo computacional foi de aproximadamente 24h para cada 1000000 de iterações do MCMC em um Athlon 2600.

A priori para o parâmetro de dificuldade de mudança de árvore (ν) foi escolhida para ter média 0.99.

Estamos interessados em investigar as probabilidades a posteriori para as topologias A, B, e C em cada sítio do alinhamento. Sejam $p(A^{(i)}|x)$, $p(B^{(i)}|x)$ e $p(C^{(i)}|x)$

essas probabilidades, onde x é a matriz de seqüências de DNA alinhadas e i é o i-ésimo sítio do alinhamento. A figura 7.3 apresenta as evoluções de $p(\tau^{(i)}|x)$, $\tau = A$, B, ou C, ao longo dos 15000 sítios.

Podemos verificar que os valores das posterioris oscilam consideravelmente nas três topologias, fazendo com que a figura 7.3 seja bem diferente da figura 7.2, que ilustra a realidade dos dados. Para uma análise mais detalhada desses resultados, criamos uma variável T, tal que

$$T^{(i)} = \begin{cases} 1 & \text{se } p(A^{(i)}|x) = max\{p(A^{(i)}|x), p(B^{(i)}|x), p(C^{(i)}|x)\} \\ 2 & \text{se } p(B^{(i)}|x) = max\{p(A^{(i)}|x), p(B^{(i)}|x), p(C^{(i)}|x)\} \\ 3 & \text{se } p(C^{(i)}|x) = max\{p(A^{(i)}|x), p(B^{(i)}|x), p(C^{(i)}|x)\}. \end{cases}$$

Essa variável tem a função de classificar cada sítio como pertencente à topologia A (se $T = 1$), B (se $T = 2$) ou C (se $T = 3$). A figura 7.4 ilustra os valores de T obtidos com as probabilidades apresentadas na figura 7.3. Os sítios que foram classificados erroneamente são representados em vermelho. Ao todo, 1410 (9.40%) sítios foram classificados de forma equivocada.

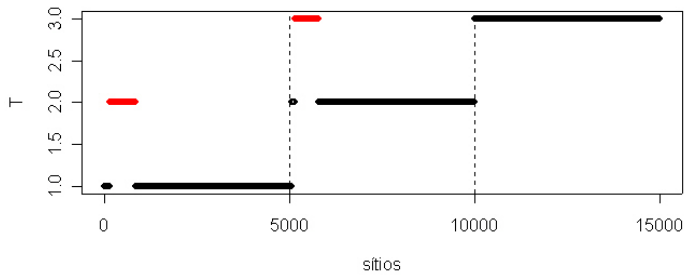

Figura 7.4: Valores da variável T ao longo dos 15000 nucleotídeos para a análise com mosaico aleatório e priori para ν com média 0.99. Os pontos vermelhos indicam os sítios classificados de forma equivocada.

Uma segunda análise dos dados foi realizada seguindo a sugestão de Husmeier & McGuire (2003) de usar um mosaico com uma pré-classificação dos sítios, ou seja, com

uma estrutura de recombinação. Optamos por utilizar o RECPARS (versão 1.00b; Hein, 1993), que estima a seqüência de topologias $\tau = (\tau_1, \ldots, \tau_{15000})$ pelo método da máxima parcimônia. A estrutura de recombinação predita pelo RECPARS (Hein, 1993) é mostrada na figura 7.5.

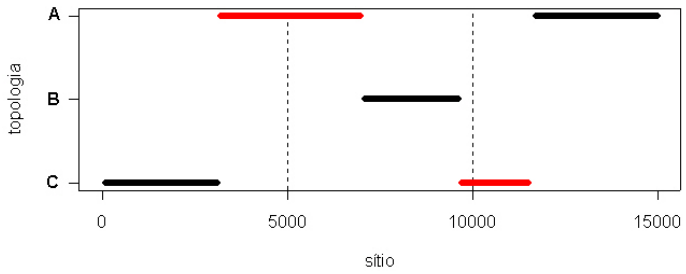

Figura 7.5: Estrutura de recombinação dos dados estimada pelo RECPARS. Os pontos vermelhos indicam os sítios classificados de forma equivocada (37.32%). 3.95% dos sítios não possuem classificação.

O BARCE (Husmeier & McGuire, 2003) foi então utilizado com o mesmo número de iterações e o mesmo tamanho de amostra utilizados anteriormente. Já a média para a priori de ν foi modificada para 0.95. O gráfico para T é apresentado na figura 7.6.

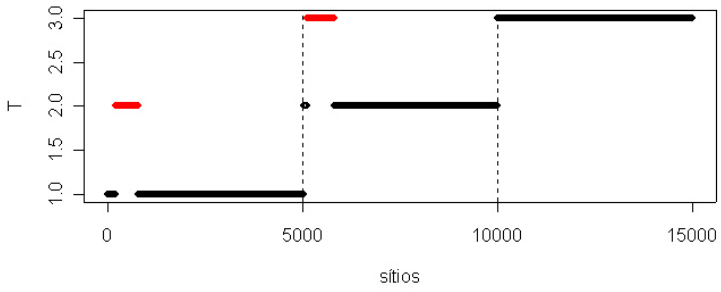

Figura 7.6: Valores da variável T para a análise iniciada com RECPARS e priori para ν com média 0.95. Os pontos vermelhos indicam erros de classificação de sítios.

Esta segunda análise dos dados apresentou 8.67% de erros de classificação de sítios, sendo essa proporção inferior a do primeiro estudo, realizado com mosaico aleatório.

Realizamos uma terceira análise desses dados, a qual se difere da segunda apenas pelo parâmetro ν, que passou a ter média igual a 0.99. Os resultados de T são ilustrados na figura 7.7. A porcentagem de erros de classificação para os sítios do alinhamento foi de 8.26%.

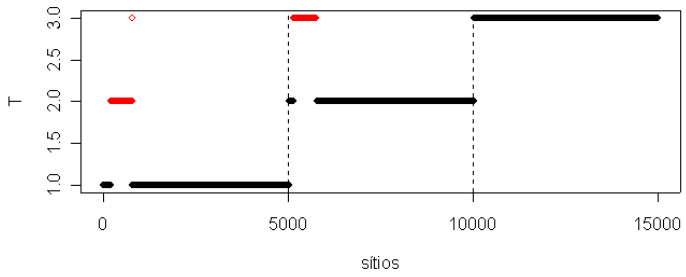

Figura 7.7: Valores de T para a análise iniciada com RECPARS e priori para ν com média 0.99. Os pontos vermelhos indicam erros de classificação de sítios.

A quarta análise dos dados foi realizada com mosaico inicial para a estrutura de recombinação construído pelo RECPARS (Hein, 1993) e priori para ν com média 0.999. O número de iterações e o processo de amostragem permaneceu inalterado. O resultado obtido para a variável T é mostrado na figura 7.8. A porcentagem de sítios classificados como pertencentes a uma topologia errada foi reduzida a 5.60%.

Note que, em todas as análises realizadas, fica claro que os dados não foram gerados de uma única topologia, e sim de uma mistura de topologias. Além disso, o modelo de mistura bayesiano utilizado estima de forma satisfatória a topologia a que pertence cada sítio do alinhamento.

A figura 7.9 ilustra o logaritmo da posteriori conjunta para as análises até então realizadas. Estamos interessados em examinar o intervalo entre a 10000-ésima e a 20000-ésima observação, o qual contém os 10000 elementos da amostra. O logaritmo

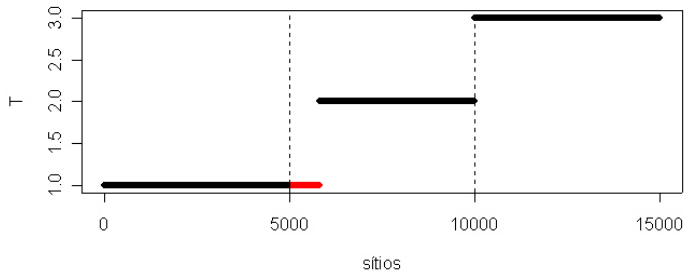

Figura 7.8: Valores de T para a análise iniciada com RECPARS e priori para ν com média 0.999. Os pontos vermelhos indicam erros de classificação de sítios.

da posteriori para esse intervalo é ilustrado no lado direito da gravura 7.9. Podemos verificar que o gráfico mais estável é referente à análise com iniciada com mosaico estimado pelo RECPARS (Hein, 1993) e priori de ν com média igual a 0.999. Cabe relembrar que essa análise foi a que apresentou o menor número erros na classificação dos sítios.

Buscando melhorar os resultados obtidos, decidimos implementar o BARCE (Husmeier & McGuire, 2003) com 3000000 de iterações, utilizando apenas as últimas 1000000 para coletar uma amostra de tamanho 10000. Para a priori de ν, escolhemos a beta com média 0.99. Infelizmente, os resultados obtidos não foram muito diferentes daquelas obtidos para 2000000 de iterações. O número de erros de classificação e o gráfico para o logaritmo da posteriori não sofreram alterações significativas. No entanto, o tempo computacional aumentou em aproximadamente 24h.

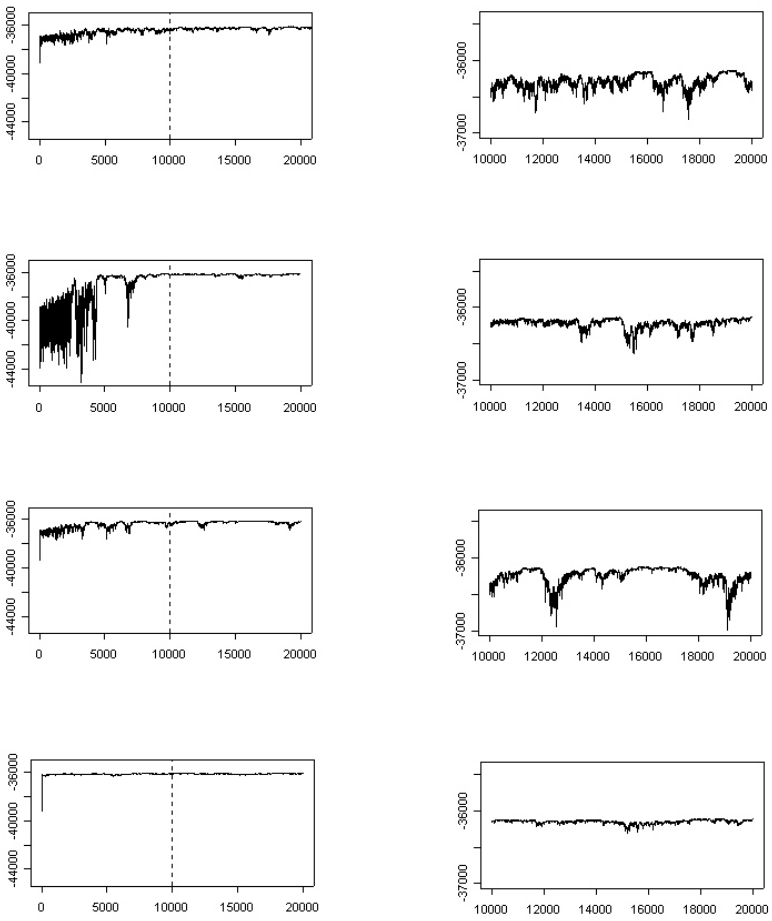

Figura 7.9: À esquerda: logaritmo da posteriori para as 20000 observações de cada análise. À direita: logaritmo da posteriori para as últimas 10000 observações (amostra). De cima para baixo: gráfico referente a análise com mosaico aleatório e priori para ν com média 0.99; gráfico referente a análise iniciada com RECPARS e priori para ν com média 0.95; gráfico referente a análise iniciada com RECPARS e priori para ν com média 0.99; gráfico referente a análise iniciada com RECPARS e priori para ν com média 0.999.

7.2.2 Análise do conjunto de dados número 2

Um novo conjunto de dados foi escolhido. Novamente os galhos externos têm comprimentos iguais a 0.05 e os galhos internos 0.005. O modelo sob o qual os dados foram gerados também foi o de Jukes-Cantor. Para esses dados, a análise de bootstrap com algoritmo de vizinhança conjunta estimou a topologia B com suporte de 53% para os ramos. A análise filogenética realizada pelo `MrBayes` (Huelsenbeck & Ronquist, 2001) produziu a topologia B com suporte bayesiano de 65% para os ramos.

Decidimos aplicar a esses dados todas as análises realizadas para o conjunto de dados número 1.

Utilizando um mosaico aleatório para a estrutura de recombinação dos dados e priori para ν com média igual a 0.99, obtivemos as estimativas dadas na figura 7.10 para as probabilidades a posteriori das três topologias e os valores de T apresentados na figura 7.11.

Para essa primeira análise, a porcentagem de sítios erroneamente classificados foi 14.15%. Para tentar reduzir esse número, introduzimos o mosaico de recombinação estimado pelo RECPARS (Hein, 1993) na análise bayesiana do BARCE (Husmeier & McGuire, 2003). Esse mosaico é ilustrado na figura 7.12.

Note que o RECPARS (Hein, 1993) estima uma estrutura de recombinação muito distante da realidade dos dados (figura 7.1), sendo apenas 20.8% dos sítios associados a correta topologia. Iremos perceber que esse resultado influencia significativamente na estimação das probabilidades a posteriori para a topologia de cada sítio.

Utilizando o mosaico da figura 7.12 para iniciar o BARCE (Husmeier & McGuire, 2003), o número de erros de classificação dos sítios tornou-se maior que o da primeira análise desses dados, realizada com mosaico aleatório. A figura 7.13 contém o gráfico dos valores de T para o estudo feito com priori para ν com média igual a 0.95. A porcentagem de sítios erroneamente classificados foi de 30.52%. Para a análise realizada com priori para ν com média igual a 0.99, a porcentagem de erros foi de

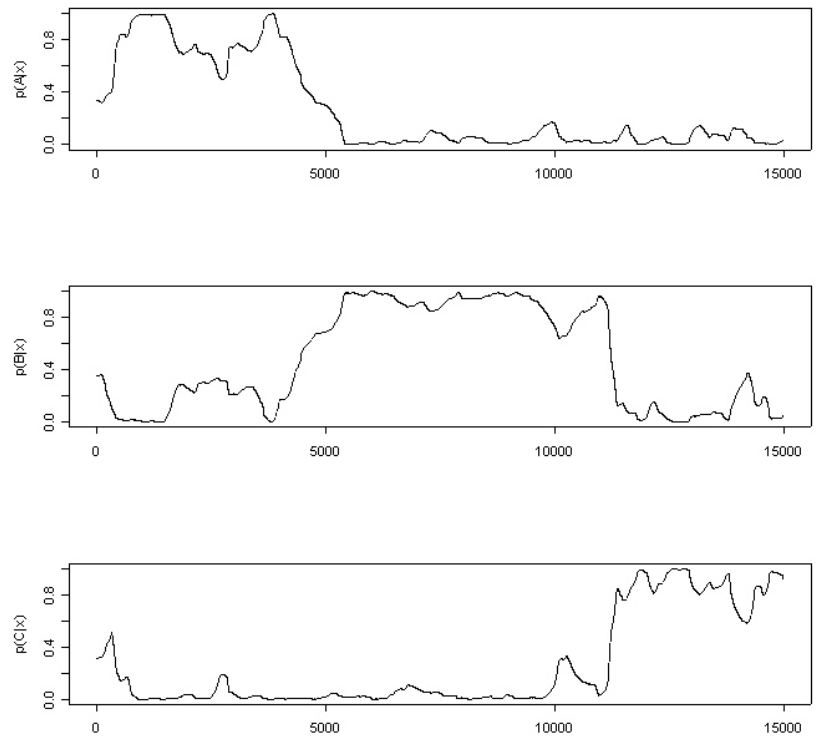

Figura 7.10: Probabilidades a posteriori para as topologias A, B e C para a análise dos dados iniciada com estrutura de recombinação aleatória e priori para ν com média 0.99.

23.68%. Os valores de T referentes a esse último estudo são ilustrados na figura 7.14.

Finalmente realizamos a análise com mosaico inicial estimado pelo RECPARS (Hein, 1993) e com priori para ν com média igual a 0.999. Obtivemos 25.19% dos sítios classificados de forma errada. A figura 7.15 mostra os valores de T para esse estudo.

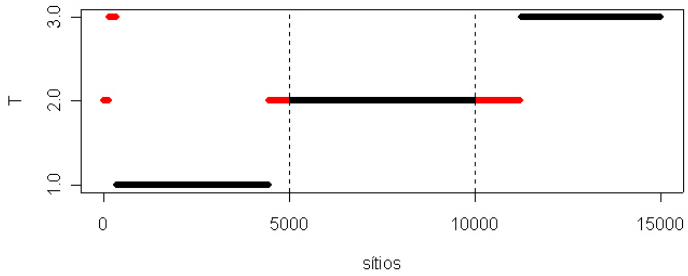

Figura 7.11: Valores de T para a análise dos dados iniciada com estrutura de recombinação aleatória e priori para ν com média 0.99.

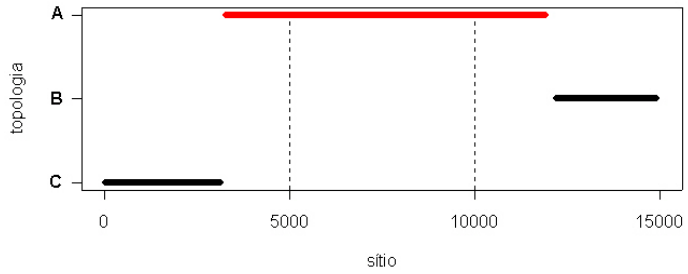

Figura 7.12: Estrutura de recombinação dos dados estimada pelo RECPARS. Os pontos vermelhos indicam os sítios classificados de forma equivocada (75.43%). 3.77% dos sítios não possuem classificação.

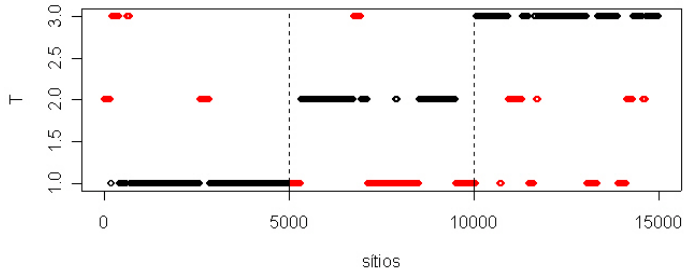

Figura 7.13: Valores de T para a análise iniciada com RECPARS e priori para ν com média 0.95. Os pontos vermelhos indicam erros de classificação de sítios.

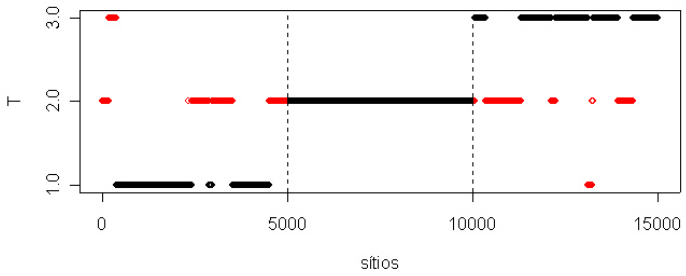

Figura 7.14: Valores de T para a análise iniciada com RECPARS e priori para ν com média 0.99. Os pontos vermelhos indicam erros de classificação de sítios.

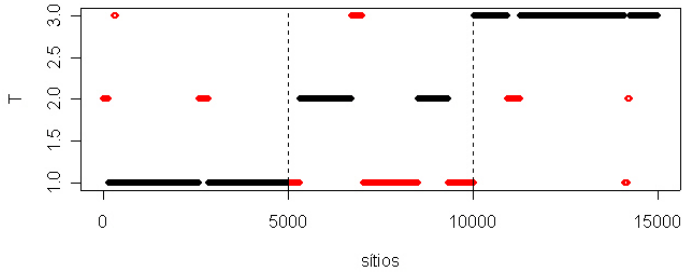

Figura 7.15: Valores de T para a análise iniciada com RECPARS e priori para ν com média 0.999. Os pontos vermelhos indicam erros de classificação de sítios.

O logaritmo da posteriori conjunta dos estudos realizados com o conjunto de dados número 2 é mostrado na figura 7.16. Assim como ocorre com conjunto de dados número 1, o gráfico referente à análise iniciada com mosaico do RECPARS (Hein, 1993) e priori para ν com média 0.999 apresenta a menor oscilação. No entanto, para o segundo conjunto de dados, a análise com o menor número de erros de classificação de sítios é a análise realizada com o mosaico aleatório gerado pelo BARCE (Husmeier & McGuire, 2003).

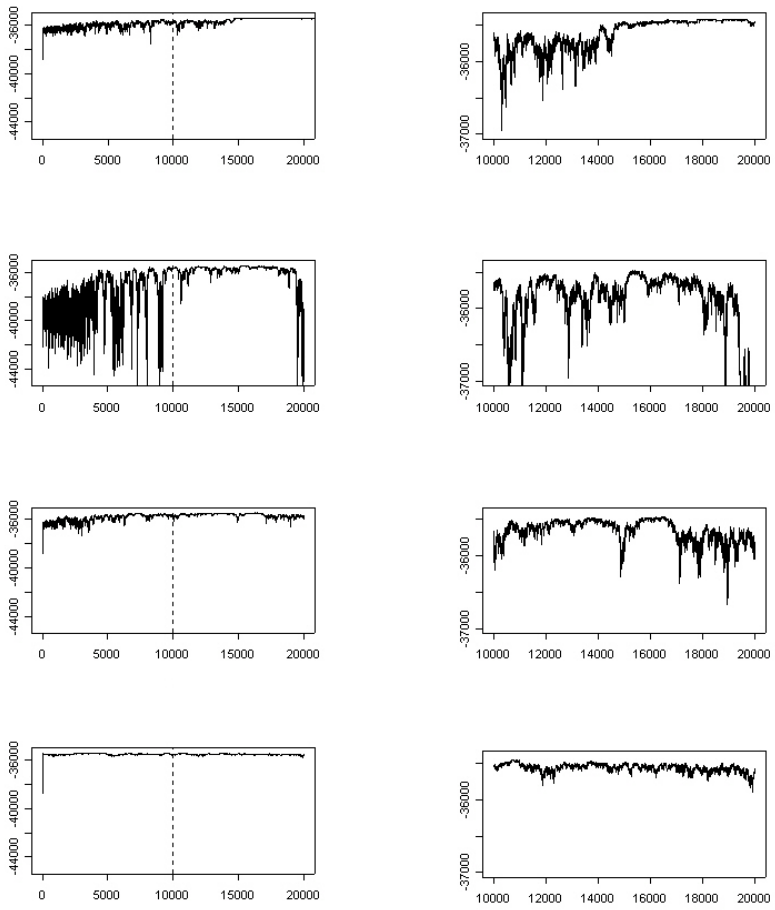

Figura 7.16: À esquerda: logaritmo da posteriori para as 20000 observações de cada análise. À direita: logaritmo da posteriori para as 10000 observações de cada amostra. De cima para baixo: gráfico referente a análise com mosaico aleatório e priori para ν com média 0.99; gráfico referente a análise iniciada com RECPARS e priori para ν com média 0.95; gráfico referente a análise iniciada com RECPARS e priori para ν com média 0.99; gráfico referente a análise iniciada com RECPARS e priori para ν com média 0.999.

Note que para as duas seqüências de DNA estudadas neste capítulo, todas as análises realizadas classificaram mais de 60% dos sítios do alinhamento como pertencentes à topologia correta . Além disso, todos os estudos revelam a mistura das topologias A, B e C nos dados.

Capítulo 8

Conclusão

Neste trabalho, estudamos o resultado paradoxal encontrado por Suzuki *et al.* (2002). Demonstramos que o modelo bayesiano utilizado por esses autores é inadequado para dados concatenados pois, como vimos no capítulo 6, quando analisamos seqüências de DNA com mistura de topologias sob um modelo que supõe topologia única, os dados tendem a privilegiar uma das topologias sobre as demais. Esse comportamento justifica a alta probabilidade à posteriori encontrada para os ramos da árvore de consenso pela análise realizada por Suzuki *et al.* (2002) no `Mrbayes` (Huelsenbeck & Ronquist, 2001).

Sugerimos a análise filogenética de dados concatenados com um modelo bayesiano de misturas de topologias. No capítulo 7, implementamos o modelo de Husmeier & McGuire (2003) para identificar as diferentes topologias que compõem as seqüências de DNA geradas sob as mesmas condições dos dados de Suzuki *et al.* (2002). Realizamos análises com diferentes iniciações para a estrutura de recombinação e para distribuição a priori para a dificuldade de mudança de topologia ao longo dos sítios (ν).

Escolhemos dois conjuntos de dados que quando analisados pelo `MrBayes` (Huelsenbeck & Ronquist, 2001), haviam produzido uma única topologia, a do tipo B, com suporte bayesiano de 94% e 65%. Pudemos verificar que, em todas as análises real-

izadas no BARCE (Husmeier & McGuire, 2003), fica evidente que esses dados não foram gerados de uma única topologia, e sim de uma mistura das topologias A, B e C.

Vimos que o mosaico de estrutura dos dados e a priori para ν escolhidos para iniciar a análise influenciam significativamente na probabilidade a posteriori para a topologia de cada sítio. De qualquer forma, em todos os estudos realizados, mais de 60% dos sítios são associados a correta topologia.

Com os resultados que obtivemos, mostramos que, ao contrário do que dizem Suzuki *et al.*(2002), a inferência filogenética bayesiana pode ser usada em seqüências de DNA concatenadas. No entanto, se temos partições da seqüência de DNA com diferentes histórias evolutivas, um modelo de mistura de topologias deve ser assumido.

Referências Bibliográficas

[1] Adams, E. N.,III. 1972. Consensus techniques and the comparison of taxonomic trees. *Systematic Zoology* **21**: 390-397.

[2] Adams, E. N.,III. 1986. N-trees as nestings: complexity, similarity, and consensus. *Journal of Classification* **3**: 299-317.

[3] Alfaro, M., S. Zoller and F. Lutzoni. 2003. Bayes or Bootstrap? A Simulation Study Comparing the Performance of Bayesian Markov Chain Monte Carlo Sampling and Bootstrapping in Assessing Phylogenetic Confidence. *Molecular Biology and Evolution* **20**: 255-266.

[4] Bininda-Emonds, O., J. Gittleman and M. Steel. 2002.The (Super)Tree of Life: Procedures, Problems, and Prospects. *Annual Review of Ecology Systematics* **33**: 265-289.

[5] Cheida, L. E. 2003. Biologia integrada: volume único. Coleção Delta, São Paulo.

[6] Darwin, C. 1859. On The Origin of Species. John Murray, London.

[7] Durbin, R., S. Eddy, A. Krogh and G. Mitchison. 1998. Probabilistic models of proteins and nucleic acids. The United Kingdom at the University Press, Cambridge.

[8] Edwards, A. W. F. and L. L. Cavalli-Sforza. 1936. The reconstruction of evolution. *Annals of Human Genetics* **27**:105-106. *Heredity* **18**: 553.

[9] Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**: 1-26.

[10] Ewens, W. J. and G.R.Grant. 2001. Statistical Methods in Bioinformatics-An Introduction. Springer-Verlag New York, Inc.

[11] Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution***17**: 368-376.

[12] Felsenstein, J. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution* **39**: 783-791.

[13] Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Inc. Publishers Sunderland, Massachusetts.

[14] Fitch, W. M. 1971. Toward defining the course of evolution: Minimum change for a specified tree topology. *Systematic Zoology* **20**: 406-416

[15] Fristedt, B. and L. Gray. 1997. A Modern Approach to Probability Theory. Birkhäuser, Boston.

[16] Gamerman, D. 1996. Simulação Estocástica via Cadeias de Markov. ABE-Associação Brasileira de Estatística.

[17] Graves, J. A. M. 2003. The Tree of life: View from a Twig. *Science* **300**: 1621.

[18] Grimment, G. R. and D. R. Stirzaker. 1992. Probability and Random Processes. Oxford University Press Inc., New York.

[19] Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their aplications. *Biometrika* **57**: 97-109.

[20] Hein, J. 1993. A heuristic method to reconstruct the history of sequences subject to recombination. *Journal of Molecular Evolution* **36**: 396-405.

[21] Hennig, W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin.

[22] Hjorth, J. S. U. 1994. Computer Intensive Satistical Methods: Validation model selection and bootstrap. Chapman & Hall, London.

[23] Huelsenbeck, J. P. and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogeny. *Bioinformatics* **17**: 754-755.

[24] Huelsenbeck, J. P., F. Ronquist, R. Nielsen and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**: 2310-2314.

[25] Husmeier, D. and G.McGuire. 2003. Detecting Recombination in 4-Taxa DNA Sequence Alignments with Bayesian Hidden Markov Models and Markov Chain Monte Carlo. *Molecular Biology and Evolution* **20**: 315-337.

[26] Jukes, T. H. and C.R.Cantor. 1969. Mammalian Protein Metabolist. Vol. III. M. N. Munro. Academic Press, New York.

[27] Kimura, M. 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotides sequences. *Journal of Molecular Evolution* **16**: 111-120.

[28] Kumar, S., K. Tamura, I. B. Jakobsen and M. Nei. 2001. MEGA. *Bioinformatics* **17**: 1244-1245.

[29] Lamark, J. 1809. Philosophie Zoologique. France.

[30] Larget, B. and D. L. Simon. 1999. Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees.*Molecular Biology and Evolution* **16**: 750-759.

[31] Li, S., D. Pearl and H. Doss. 2000. Phylogenetic Tree Construction Using Markov Chain Monte Carlo. *Journal of the American Statistical Association* **95**: 493-508.

[32] Lopes, S. 1994. Bio: volume único. Editora Saraiva, São Paulo.

[33] Mau, B. and M. A. Newton. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics* **6**: 122-131.

[34] Mau, B., M. A. Newton and B. Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics* **5**: 1-12.

[35] McGuire, G. and F. Wright. 2000. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. *Bioinformatics* **16**: 130-134.

[36] McGuire, G. and F. Wright and m. J. Prentice. 1997. A graphical method for detecting recombination in phylogenetic data sets. *Molecular Biology and Evolution* **14**: 1125-1131.

[37] Margush, T. and F. R. McMorris. 1981. Consensus n-trees. *Bulletin of Mathematical Biology* **43**:239-244.

[38] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller. 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21**: 1087-1092.

[39] Misawa, K. and M. Nei. 2003. Reanalysis of Murphy *et al.*'s Data Gives Various Mammalian Phylogenies and Suggests Overcredibility of Bayesian Trees. *Journal of Molecular Evolution* **57**: S290-S296.

[40] Mount, D. W. 2001. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, New York.

[41] Mueller, L. D. and F. J. Ayala. 1982. Estimation and interpretation of genetic distance in empirical studies. *Genetical Research* **40**: 127-137.

[42] Oparin, A. 1936. Origin of Life. Dover, New York.

[43] Pagel, M. and A. Meade. 2004. A Phylogenetic Mixture Model for Detecting Pattern-Heterogeneity in Gene Sequence or Character-State Data. *Systematic Biology* **53**: 571-581.

[44] Rambaut, A. and N. C. Grassly. 1997. SEQGEN. *Computer Applications in the Biosciences.* **12**: 291-295.

[45] Rohlf, F. J. 1982. Consensus indices for comparing classifications. *Mathematical Biosciences* **59**: 131-144.

[46] Saitou, N. and M. Nei. 1987. The neighbor-joining method: A new method for reconstruction phylogenetic trees. *Molecular Biology and Evolution* **4**: 406-425.

[47] Sanderson, M., A. Purvis and C. Henze. 1998. Phylogenetic supertrees: assembling the trees of life. *Trends in Ecology and Evolution* **13**: 105-109.

[48] Sankoff, D. and P. Rousseau. 1975. Locating the vertices of a Steiner tree in arbitrary space. *Mathematical Programming* **9**: 240-246.

[49] Studier, J .A. and K. J. Keppler. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution* **5**: 729-731.

[50] Suchard, M. A., C. M. R. Kitchen, J .S. Sinsheimer and R. E. Weiss. 2003. Hierarchical Phylogenetic Models for Analyzing Multipartite Sequence Data. *Systematic Biology* **52**: 649-664.

[51] Suzuki, Y., G. Glazko and M. Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *PNASProceedings of the National Academy of Sciences* **99**: 16138-16143.

[52] Swofford, D. L.(1995). PAUP*. Phylogenetic Analysis Using Parsimony(*and other Methods). Sinauer Associates, Sunderland, Massachusets.

[53] Watson, J. and F. Crick. 1953. A Structure for Deoxyribose Nucleic Acid. *Nature* **171**: 737-738.

[54] Wilcox, T. P., D. J. Zwickl, T. A. Heath, and D . M. Hillis. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Molecular Biology and Evolution* **25**: 361-371

[55] Yang, Z. and B. Rannala. 1997. Bayesian Phylogenetic Inference Using DNA Sequences: A Markov Chain Monte Carlo Method. *Molecular Biology and Evolution* **14**: 717-724.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo