

Tutorial 12

Introdução à verossimilhança

BIZ0433 - INFERÊNCIA FILOGENÉTICA: FILOSOFIA, MÉTODO E APLICAÇÕES.

POR DENIS JACOB MACHADO & FERNANDO MARQUES

Conteúdo

Objetivo	198
12.1 Probabilidades	199
12.2 Princípios de estimativas de verossimilhança máxima	200
12.3 Modelos de substituição	203
12.4 Verossimilhança como critério de otimalidade	205
12.4.1 Parâmetros inconvenientes e medidas de máxima verossimilhança . . .	205
12.4.2 Calculando $P_{(D T,\theta)}$	206
12.4.3 Exemplo simples	207
12.4.4 Seleção de modelos e dados lacunares	208
12.5 Referências	211

Objetivo

Este tutorial introduz o conceito de verossimilhança como método de inferência e como ele pode ser aplicado como critério de otimalidade em inferência filogenética. Três tipos de verossimilhança são apresentados como critério de otimalidade. Finalmente, o conceito e a prática da seleção de modelos de substituição é apresentado. Os arquivos associados a este tutorial estão disponíveis no [GitHub](#). Você baixar todos os tutoriais com o seguinte comando:

```
svn checkout https://github.com/fplmarques/cladistica/trunk/tutorials/
```

12.1 Probabilidades

Há várias definições de probabilidades e uma breve consulta a literatura especializada revelará que não há consenso absoluto sobre esse conceito [1]. Independentemente do conceito que você adotar sobre probabilidade, há uma série de axiomas em Teoria de Probabilidades que você deve conhecer – ou relembrar para que nós possamos entender os conceitos de verossimilhança. Neste tutorial iremos adotar a definição “frequencista” de probabilidades. Dentro deste contexto, dado n eventos dentre os quais um conjunto de resultados mutuamente exclusivos (E_i) é observado, à medida em que $n \rightarrow \infty$, dizemos que a probabilidade de E_i (isto é, $P_{(E_i)}$) pode ser expressa por E_i/n . Há alguns axiomas da teoria de probabilidades, ou consequência desses, que você deve ter em mente:

$$0 \leq P_{(E_i)} \leq 1; \quad (12.1)$$

$$P_{(E_i)} = 1 - P_{(\tilde{E}_i)}; \quad (12.2)$$

no qual $P_{(\tilde{E}_i)}$ é a probabilidade de $P_{(E_i)}$ não ocorrer,

$$P_{(E_i \text{ ou } E_j)} = P_{(E_i \cup E_j)} = P_{(E_i)} + P_{(E_j)}; \quad (12.3)$$

assumindo que E_i e E_j sejam eventos disjuntos exclusivos, e

$$P_{(E_i \text{ e } E_j)} = P_{(E_i \cap E_j)} = P_{(E_i)} * P_{(E_j)}; \quad (12.4)$$

no qual E_i e E_j são eventos complementares independentes.

Por fim, probabilidades condicionais são expressas da seguinte forma:

$$P_{(E_i|E_j)} = \frac{P_{(E_i \cap E_j)}}{P_{(E_j)}}; \quad (12.5)$$

Neste caso, $P_{(E_i|E_j)}$ denota a $P_{(E_i)}$ ocorrer dado que $P_{(E_j)}$ já ocorreu.

12.2 Princípios de estimativas de verossimilhança máxima

Considere que você pegou uma moeda e jogou cara ou coroa 20 vezes e obteve o seguinte resultado [o mesmo exemplo é dado utilizando distribuição binomial no Apêndice A de 2]:

Ca Co Co Ca Ca Ca Co Ca Co Co Co Ca Ca Co Ca Co Co Ca Ca Ca

ou seja 11 caras e 9 coroas.

Qual seria a probabilidade desta observação ($P_{(obs)}$)?

Ela seria definida da seguinte forma:

$$P_{(obs)} = P_{(Ca)} * P_{(Co)} * P_{(Co)} * P_{(Ca)} * P_{(Ca)} * P_{(Ca)} * P_{(Co)} * P_{(Ca)} * P_{(Co)} * P_{(Co)} * P_{(Co)} * P_{(Co)} * P_{(Ca)} * P_{(Ca)} * P_{(Co)} * P_{(Ca)} * P_{(Ca)} * P_{(Ca)}$$

ou seja (veja equação 12.4),

$$P_{(obs)} = P_{(Ca)}^{11} * P_{(Co)}^9 \quad (12.6)$$

Segundo a fórmula acima, e assumindo que a moeda não apresenta nenhum vício, você teria:

$$P_{(obs)} = 0.5^{11} * 0.5^9 = 0.00000105415$$

No entanto você assumiu que $P_{(Ca)} = P_{(Co)} = 0.5$. Suponha que você desconheça o valor de $P_{(Ca)}$ ou $P_{(Co)}$. Seria possível estimar esses valores? Estimativas de Verossimilhança Máxima podem ser aplicadas nesse contexto.

Verossimilhança Máxima (L) é definida como:

$$L_{(\theta|obs)} \propto P_{(obs|\theta)} \quad (12.7)$$

onde, a Verossimilhança Máxima (L) de um parâmetro θ dado a observação é proporcional a probabilidade da observação dado um determinado parâmetro. Portanto, Verossimilhança Máxima lhe possibilita estimar o valor de θ que maximize a probabilidade de você observar os dados que observou.

Para aplicar Verossimilhança Máxima em nosso exemplo com caras e coroas, devemos considerar que:

$$P_{(Ca)} + P_{(Co)} = 1$$

portanto,

$$P_{(Co)} = 1 - P_{(Ca)}. \quad (12.8)$$

Se considerarmos que a equação (12.8) e a equação (12.6) teríamos:

$$P_{(obs)} = P_{(Ca)}^{11} * (1 - P_{(Ca)})^9 \quad (12.9)$$

Considere agora que o parâmetro θ que você gostaria de estimar é o valor de $P_{(Ca)}$ que maximizasse L . O valor de $P_{(Ca)}$ que maximizaria a probabilidade de observar seus dados seria definido como:

$$L_{(P_{(Ca)}|obs)} \propto P_{(obs|P_{(Ca)})} \quad (12.10)$$

Se você tem familiaridade com R , você pode usar o *script* abaixo, disponível no diretório deste tutorial¹, para gerar os gráficos apresentados a seguir manipulando os valores de *head* e *tail*, para caras e coroas, respectivamente.

```
head <- 11 # mude aqui o numero de caras
tail <- 9  # mude aqui o numero de coroas
p <- seq(0,1,by=0.001)
for (i in p){L <- c((p^head)*(1-p)^tail)}
maxL <- max(L)
indexL <- match(maxL,L)
best_p <- p[indexL]
plot(p, L, xlab="P(Ca)", ylab="L", col="blue", frame=F, pch=16,cex=0.5)
text(0.8, maxL, paste("L =", maxL, "\n P(Ca) =", best_p))
```

A Figura 12.1 refere-se à nossa observação inicial e exibe os valores de L associados com as possíveis probabilidades de caras (*i.e.*, $P_{(Ca)}$) – que varia de 0 a 1 em intervalos de 0.001).

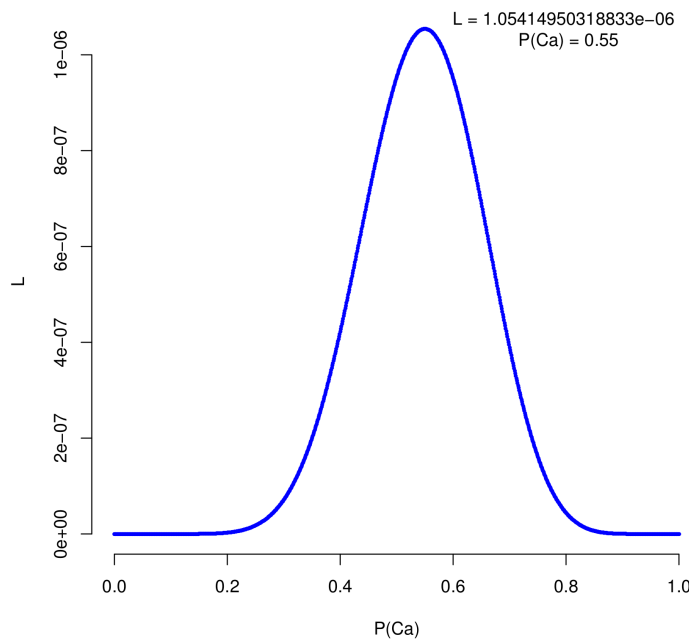


Figura 12.1: Valores de Verossimilhança ($L_{(P_{(Ca)}|obs)}$) em função da probabilidade de obter caras ($P_{(Ca)}$) para 20 eventos dos quais 11 resultaram em caras e 9 em coroas.

Observe que à medida em que $P_{(Ca)}$ se aproxima de 0.5, o valor de L vai aumentando até chegar ao seu máximo. A Verossimilhança Máxima é obtida para $P_{(Ca)} = 0.55$. O que isso significa? De acordo com o conceito de Verossimilhança Máxima, a melhor estimativa de $P_{(Ca)}$ que explica sua

¹ Você pode executar o seguinte comando “Rscript likelihood_coin.r” e o script irá gerar o arquivo “Rplots.pdf” no qual existe um gráfico exemplificando a estimativa de verossimilhança máxima.

observação (*i.e.*, 11 caras e 9 coroas) é 0.55. Desta forma, o modelo probabilístico que melhor explica sua observação é $P_{(Ca)} = 0.55$ e $P_{(Co)} = 0.45$, o que significa que a moeda aparentemente não apresenta nenhum vício, pois essa diferença é muito pequena.

Exercício 13.1

O que aconteceria se a observação fosse outra, como por exemplo, 3 caras e 17 coroas? A Figura 12.2 exibe os valores de L associados com as possíveis probabilidades de caras para esta observação. Note como esse gráfico difere do primeiro. Você considera que o modelo que melhor explica essa observação sugere que a moeda é viciada? Justifique.

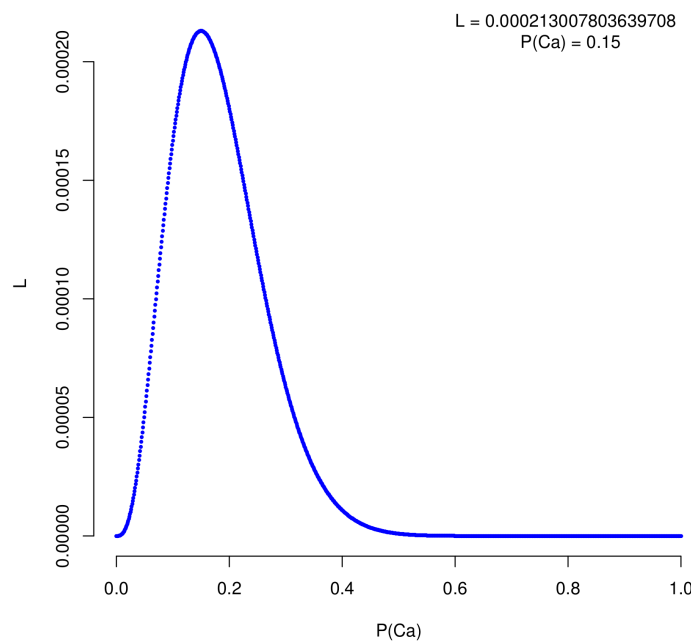


Figura 12.2: Valores de Verossimilhança ($L_{(P_{(Ca)}|obs)}$) em função da probabilidade de obter caras ($P_{(Ca)}$) para 20 eventos dos quais 3 resultaram em caras e 17 em coroas.

Exercício 13.2

O gráfico abaixo representa os valores de Verossimilhança (L) dada a probabilidade de obter caras ($P_{(Ca)}$) em dois ensaios com 20 eventos de cara ou coroa. No primeiro ensaio, uma moeda resultou em 10 caras e 10 coroas (linhas pontilhadas). No segundo ensaio, uma outra moeda resultou em 3 caras e 17 coroas (linhas contínuas). Com base nestas informações, explique como Verossimilhança Máxima é usada para estimar parâmetros.

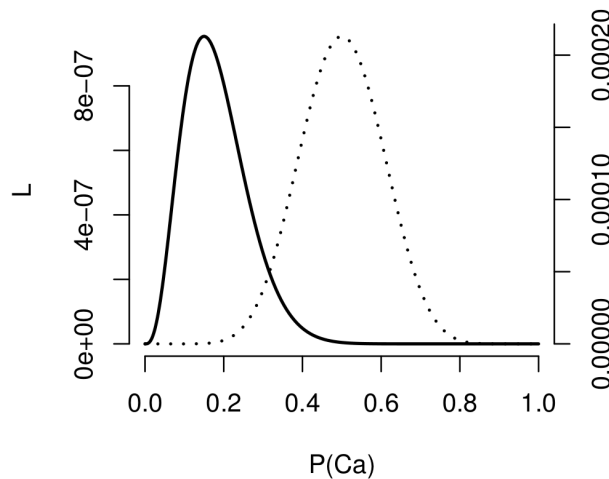


Figura 12.3: Valores de Verossimilhança ($L_{(P_{(Ca)}|obs)}$) em função da probabilidade de obter caras ($P_{(Ca)}$).

12.3 Modelos de substituição

A motivação pela qual Felsenstein [3] propôs o uso de estimativas de verossimilhança como critério de otimalidade em inferência filogenética veio da observação sob dados simulados de que topologias que possuíam comprimentos de ramos desproporcionais não eram recuperadas pelo critério da parcimônia. Desta forma, segundo o autor, a grande virtude do método proposto residia em considerar o comprimento de ramos durante a busca da melhor topologia.

Para considerar o comprimento de ramos é necessário a adoção de um modelo de substituição. De forma geral, o modelo estatístico é a formalização matemática da relação entre variáveis que correspondem a observações potenciais que inclui a descrição das incertezas sobre estas observações devido à variabilidade natural, erros ou informação incompleta. Neste caso em particular, esses modelos descrevem as probabilidades de transformação de caracteres dado sua frequência (π) e o comprimento de ramo (v). O comprimento deste ramo representa o número esperado de substituições por sítio e é definido por $v = 3\alpha * t$; no qual 3α é a taxa total de substituição e t uma unidade de tempo qualquer. Uma vez que taxa e unidade de tempo estão interligados, arbitrariamente convencionou-se que a taxa receberia o valor de **uma substituição** que temos a expectativa que ocorra em **uma unidade de tempo** para cada sítio. Ao fazermos isso, “tempo” (o comprimento de um ramo) é medido em unidades de distância evolutiva (ou ainda, número de substituições esperada por sítio). Essa equivalência é feita assumindo que $\alpha = \frac{1}{3}$, pois teríamos $v = 3(\frac{1}{3})t$, ou seja, $v = t$.

Um dos modelos mais simples de substituição é conhecido como JC69 [4]. Este modelo assume que todas as bases são igualmente frequentes (0.25) e que a taxa de substituição (α) é idêntica

para todas as possíveis substituições cujos eventos obedecem a distribuição de probabilidades de Poisson. Esta distribuição descreve o número de ocorrências de um determinado evento aleatório (estocástico) em um determinado espaço de tempo, assumindo uma taxa média de ocorrência (λ).² Desta forma, de acordo com esse modelo,

$$P_{ij}(t) = \frac{1}{4}(1 - e^{-4v/3}) = \frac{1}{4} - \frac{1}{4}e^{-4v/3} \quad (12.11)$$

e

$$P_{ii}(t) = e^{-4v/3} + \frac{1}{4}(1 - e^{-4v/3}) = \frac{1}{4} + \frac{3}{4}e^{-4v/3} \quad (12.12)$$

onde $P_{ii}(t)$ é a probabilidade de não-mudança do estado de caráter durante o tempo t .

Considere por exemplo, a transformação da sequência ancestral ACGTACGTACGT para a sequência descendente ACGTACGTAAAA dado que v é igual a 0.1. A probabilidade $P_{(ACGTACGTACGT \rightarrow ACGTACGTAAAA | \theta=v=0.1)}$ seria:

$$P_{(ACGTACGTACGT \rightarrow ACGTACGTAAAA | v=0.1)} = \left[\left(\frac{1}{4} - \frac{1}{4}e^{-4(0.1/3)} \right)^3 \right] * \left[\left(\frac{1}{4} + \frac{3}{4}e^{-4(0.1/3)} \right)^9 \right] \quad (12.13)$$

ou seja,

$$P_{(ACGTACGTACGT \rightarrow ACGTACGTAAAA | v=0.1)} = 0.00001254686 \quad (12.14)$$

No entanto, talvez $v = 0.1$ não seja o melhor valor deste parâmetro para explicar a transformação entre estas duas sequências. Usando o conceito de máxima verossimilhança poderíamos estimar o valor de v que maximizaria $P_{(ACGTACGTACGT \rightarrow ACGTACGTAAAA | \theta=v)}$. O script `likelihood_v.r`, disponível no diretório deste tutorial, computa os valores de $L_{(v|ACGTACGTACGT \rightarrow ACGTACGTAAAA)}$ – como ilustrado na Figura 12.4.

Exercício 13.3

Você considera que valor de v utilizado anteriormente é uma boa estimativa para esse parâmetro? Justifique

²Esta distribuição é derivada da distribuição binomial (veja <https://www.khanacademy.org/math/probability/random-variables-topic/poisson-process/v/poisson-process-1>).

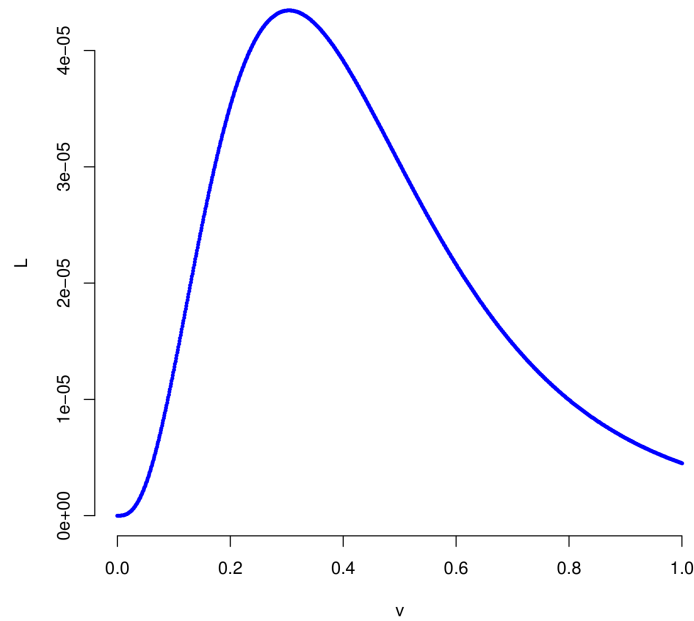


Figura 12.4: Valores de Verossimilhança ($L_{(v|obs)}$) em função do comprimento de ramo v .

12.4 Verossimilhança como critério de otimalidade

Assim como em parcimônia (MP), a busca de árvores sob o critério de máxima verossimilhança (“maximum likelihood”, ML) assinala estados ancestrais (dentro de um contexto absoluto ou de médias) tal que a verossimilhança da árvore (hipótese) é maximizada. Para o conjunto de dados D , o modelo de substituição Θ e a topologia T :

$$L_{(T,\Theta|D)} \propto P_{(D|T,\Theta)} \quad (12.15)$$

Nesta expressão, uma determinada T é selecionada de modo que $P_{(D|T,M)}$ é maximizada.

12.4.1 PARÂMETROS INCONVENIENTES E MEDIDAS DE MÁXIMA VEROSSIMILHANÇA

Os parâmetros inconvenientes são todos os demais parâmetros, fora T e D , necessários para calcular $P_{(D|T)}$. Os três parâmetros inconvenientes mais importantes são: (i) os parâmetros livres do modelo de substituição, (ii) tempo e taxa de transformação nos ramos e (iii) distribuição das taxas entre os caracteres. Estes parâmetros são coletivamente chamados θ . Assumindo uma distribuição para os parâmetros inconvenientes $\Phi(\theta | T)$, pode-se integrar θ (dentro do espaço de parâmetros Θ) para determinar $P_{(D|T)}$:

$$P_{(D|T)} = \int_{(\theta \in \Theta)} P_{(D|T,\theta)} d\Phi(\theta|T) \quad (12.16)$$

O T que maximize $P_{(D|T)}$ desta maneira é chamado máxima verossimilhança integrada

(“maximum integrated likelihood”, MIL) [5]. A MIL é igual à máxima probabilidade posterior (do inglês, “maximum a posteriori”, MAP) quando a distribuição assumidas a priori (“*priors*”) for uniforme (“flat”).

Como θ é composto de muitos parâmetros de distribuição desconhecida, uma abordagem para lidar com estes parâmetros é selecionar θ tal que $P_{(D|T,\theta)}$ seja maximizada. Isto equivale à máxima verossimilhança relativa (do inglês “maximum relative likelihood”, MRL). A MRL é a metodologia mais usada em análises empíricas. Seu cálculo independe de $P_{(T)}$ e $\Phi_{(\theta|T)}$. Os tipos de MRL estão listados abaixo:

- “Maximum average likelihood” [MAL, 6]: soma sobre todos os possíveis estados dos vértices. Esta forma de verossimilhança é a mais comumente empregada em análises empíricas.
- “Most parsimonious likelihood” (MPL, também conhecida como “ancestral maximum likelihood”): valores e parâmetros específicos são atribuídos para cada vértice. Apesar de se parecer em alguns aspectos com uma análise de parcimônia, MPL e parcimônia não convergem pois todas as taxas aplicadas sobre os ramos serão as mesmas para todos os caracteres.
- “Evolutionary path likelihood” [EPL, 7]: toda sequência de estados de caracter intermediários entre os vértices é especificada de tal modo que a verossimilhança de toda árvore é maximizada. A árvore que maximiza EPM é a árvore mais parcimoniosa.

12.4.2 CALCULANDO $P_{(D|T,\theta)}$

Para um único caráter (x) em uma árvore, a verossimilhança do vértice i (L_i) com vértices descendentes j e k será a soma da probabilidade entre x_i e cada um de seus descendentes (dado um comprimento de ramo v) multiplicadas por suas respectivas verossimilhanças e somadas para todos os estados. A verossimilhança dos caracteres é multiplicada sobre todo o conjunto de dados para determinar a verossimilhança.

$$L_i(x) = \sum_i^{\text{estados}} \left[\left(\sum_{x_j} p_{x_i, x_j}(v_j) L_j(x_j) \right) \times \left(\sum_{x_k} p_{x_i, x_k}(L_k)(x_k) \right) \right] \quad (12.17)$$

Usando dados reais, os comprimentos dos ramos são quase sempre desconhecidos e precisam ser estimados. Isto pode ser feito de diversas maneiras mas normalmente depende da probabilidade marginal (mantendo todos os demais parâmetros constantes) de um dado ramo assumindo uma variedade de valores (parâmetro v) e escolhendo um valor ótimo [para mais detalhes, veja 8, Capítulo 11].

O cálculo da MAL de uma árvore é um procedimento heurístico devido ao grande número de parâmetros que devem ser estimados. Assim como em parcimônia, a árvore é obtida assinalando

estados medianos recursivamente. Este procedimento é inicializado assinalando um valor igual a 1 para todos os ramos terminais. A verossimilhança é determinada dos ramos terminais para a raiz, multiplicando-se pelas probabilidades *a priori* dos próprios estados.

$$L_T(x) = \prod_{i=1}^{\text{estados}} \pi_i \prod_{\forall u,v \in E} L_{u,v} \quad (12.18)$$

Dado que estes valores são geralmente muito baixos, é geralmente mais conveniente expressá-los em forma logarítmica.

12.4.3 EXEMPLO SIMPLES

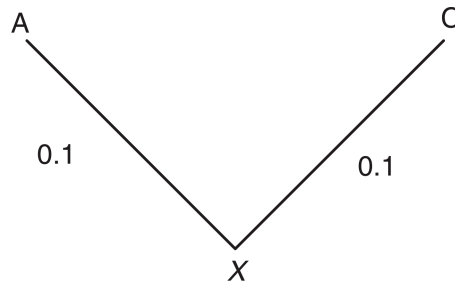
Considere um único nucleotídeo caracterizado sob o model JC69. As probabilidades de ramo serão ficadas em $v = \mu t = 0.1$.

$$f(n) = \begin{cases} \frac{1}{4} + \frac{3}{4}e^{-\mu t} & i = j \\ \frac{1}{4} - \frac{1}{4}e^{-\mu t} & i \neq j \end{cases} \quad (12.19)$$

Deste modo, as probabilidades de ramo são:

$$f(n) = \begin{cases} 0.929 & i = j \\ 0.0238 & i \neq j \end{cases} \quad (12.20)$$

Veja na **Figura 12.5** como seria o cálculo da sub-árvore à seguir com ramos terminais contendo os caracteres A e C e parâmetro de ramo 0.1. A verossimilhança média da árvore é $1,76 \times 10^6$ ou, em $-\log$ (base e), 13,25.



$$L(x = A) = [0.929 \cdot 1.0] \times [0.0238 \cdot 1.0] = 0.0221$$

$$L(x = C) = [0.0238 \cdot 1.0] \times [0.929 \cdot 1.0] = 0.0221$$

$$L(x = G) = [0.0238 \cdot 1.0] \times [0.0238 \cdot 1.0] = 0.000566$$

$$L(x = T) = [0.0238 \cdot 1.0] \times [0.0238 \cdot 1.0] = 0.000566$$

$$\text{Total } L(x) = 0.0453$$

Figura 12.5: Sub-árvore anotada com cálculo de verossimilhança (de Wheeler, 2012: Fig. 11.8).

Exercício 13.4

Qual seria a verossimilhança da sub-árvore da figura acima caso “gaps” (inserções/ deleções, ou “InDels”) fossem tratados como um quinto estado de caráter e o parâmetro dos ramos fosse 0.2?

12.4.4 SELEÇÃO DE MODELOS E DADOS LACUNARES

A adoção do critério de verossimilhança em inferência filogenética requer a escolha de um modelo probabilístico de substituição. Quanto maior é o número de parâmetros livres em um modelo – aqueles são estimados durante a análise –, maior será o ajuste do modelo aos seus dados, no entanto você pode usar parâmetros desnecessários, cuja a inclusão não justifica o ganho no valor de verossimilhança. Desta forma, o procedimento de seleção de modelos, visa adotar o modelo com o menor número de parâmetros livres que de acordo com critérios que visam penalizar a sobre-parametrização de modelos.

A escolha destes modelos deve preceder a análise filogenética e deve ser feita de forma objetiva [veja conceitos e referências em 9]. Neste tutorial nos iremos adotar como critério de escolha de modelos o critério de informação de Akaike corrigido (*Akaike Information Criteria – AIC_c*). Esta métrica não corrigida é computada da seguinte forma:

$$AIC = -2l + 2k \quad (12.21)$$

no qual l é o log da verossimilhança e k é o número de parâmetros livres do modelo – aqueles que são maximizados na função de verossimilhança. Vale ressaltar que o k inclui os parâmetros livres do modelo de substituição [veja Tabela 1 de 9], mais a topologia e seus comprimentos de ramos (*i.e.*, $(t * 2) - 3$, onde t é o número de terminais). Sua correção é necessária quando o número amostral – neste caso número de caracteres n – é pequeno quando comparado com o número de parâmetros livres (*i.e.*, estimados; digamos $\frac{n}{k} < 40$) e é dada pela fórmula:

$$AIC_c = AIC + \frac{(2k(k-1))}{(n-k-1)} \quad (12.22)$$

Em princípio, dados lacunares (“missing data”) não são um problema para análises de verossimilhança. Estados de ramos terminais podem ser definidos com probabilidade de transformação igual a 1.0 para cada estado observado ou implícito. Porém, diferentes implementações podem diferir na forma de tratar estes dados. É evidente que a implementação de dados lacunares irá afetar as análises. Este problema torna-se ainda mais pernóstico quando INDELS (*i.e.*, inserções e deleções) são tratados como dados lacunares. Este tratamento é problemático, porém matematicamente necessário para manter o cálculo da verossimilhança factível dentro dos limites atuais de tempo e recursos computacionais.

Nos exercícios abaixo você deverá selecionar modelos para dois bancos de dados, incluindo três alinhamentos distintos. O objetivo é verificar como alinhamentos modificam os critérios de escolha de modelos e refletir sobre o uso de AIC_c em análises por verossimilhança máxima.

Exercício 13.5

Neste exercício você deverá fazer três alinhamentos distintos para as sequências no arquivo `partition2.fas` (*i.e.*, `partition2aln1.fas`, `partition2aln2.fas` e `partition2aln3.fas`) utilizando MAFFT e os seguintes parâmetros de alinhamento: abertura de gaps (`--op`) 3.06, 1.53 e 0.123 no qual o valor dos gaps de extensão (`--ep`) será fixado em 0.123 (veja Seção 8.2.5 do Tutorial 8 caso tenha dúvidas).

i. Qual valor de “gap opening” resultou em um alinhamento com mais gaps? Porquê?

Existem algumas ferramentas para a seleção de modelos em análises filogenéticas, sendo [jModelTest 2](#) [10, 11] uma das mais tradicionais e a leitura do manual deste programa pode ser

muito instrutiva. No entanto, o programa **IQ-TREE** [12] – um aplicativo relativamente recente – tornou o jModelTest 2 obsoleto na minha opinião. O **IQ-TREE** possibilita a avaliação de um maior número de modelos e é muito mais rápido e robusto que o jModelTest 2.

O **IQ-TREE** utiliza o algoritmo do ModelFinder [13] para a seleção de modelos e maiores detalhes sobre a seleção de modelos em IQ-TREE podem ser encontrados nos tutoriais do programa – na seção “**Choosing the right substitution model**” e/ou na documentação do programa – na seção “**Substitution models**”.

Exercício 13.6

Neste exercício você deverá computar os parâmetros de seleção de modelos para o arquivo `partition1.fas` – que não requer alinhamento – e os três alinhamentos distintos para as sequências no arquivo `partition2.fas` (*i.e.*, `partition2aln1.fas`, `partition2aln2.fas` e `partition2aln3.fas`) que você executou no exercício anterior. Adicionalmente, você fará o mesmo para os bancos de dados concatenados.

O **IQ-TREE** deverá estar instalado em seu computador, caso não esteja, verifique a versão apropriada para seu sistema na [página de download](#) do IQ-TREE. Uma vez instalado, a seleção de modelos no IQ-TREE pode ser obtida pela seguinte linha de comando:

```
$ iqtree2 -s partition1.fas -m TEST
```

Com esse comando, o IQ-TREE irá avaliar 88 modelos de substituição, equivalentes ao que seria examinado pelo jModelTest 2³. Após a execução do comando acima, os dados necessários para completar a Tabela 12.1 estarão no log da execução (*e.g.*, arquivo `partition1.fas.log`).

Tabela 12.1: Seleção de modelos pelo critério de AIC_c .

Dataset	$-\ln L$	k	AIC_c	Model
<code>partition1.fas</code>				
<code>partition2aln1.fas</code>				
<code>partition2aln2.fas</code>				
<code>partition2aln3.fas</code>				
<code>partition1+partition2aln1.nex</code> ⁴				
<code>partition1+partition2aln2.nex</code>				
<code>partition1+partition2aln3.nex</code>				

³ A opção “`-m MFT`” explora modelos adicionais implementados em ModelFinder, consulte a documentação de **IQ-TREE** para maiores detalhes.

⁴ Você deverá utilizar o `sequencematrix` para concatenar os dados (veja seção 7.3.1 do Tutorial 7 e exportar os dados no formato NEXUS para GARLI).

Exercício 13.7

Com base nos exercícios acima responda:

- i. Modelos diferentes podem ser selecionados para diferentes alinhamentos dos mesmos dados?

- ii. Considerando os três alinhamentos para o arquivo `partition2.fas`, você teria algum critério para escolher qual deles deveria ser submetido à análise filogenética?

- iii. Qual modelo de substituição você selecionaria para a análise do arquivo `partition1+partition2aln2.nex`? Justifique⁵.

- iv. Os dados que você analisou são os mesmos da seção 9.2.3 do Tutorial 9. Naquela ocasião, havia um outro conjunto de dados morfológicos (`partition3.tnt`) que se referia aos mesmos táxons. O que seria necessário para incorporá-lo em uma análises simultânea juntamente com os demais dados que você estimou o melhor parâmetro de substituição em uma análise sob o critério de verossimilhança máxima?

12.5 Referências

1. Thacker, N. A. Tutorial: Defining probability for Science. Acesso em: 2 jul. 2014. 2014. <http://tree.bio.ed.ac.uk/>.
2. Anderson, D. R. 2008. Model based inferences in the life sciences: a primer on evidence. Fort Collins, CO: Springer, 2008.
3. Felsenstein, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology* **27**(4): 783–791.
4. Jukes, T. H. & Cantor, C. R. em *Mammalian protein metabolism*. ed. Munro, H. N. New York: Academic Press, 1969.

⁵ Considere o modelo sugerido para cada uma das partições individualmente.

5. Steel, M & Penny, D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular biology and evolution* **17**(6): 839–50.
6. Barry, D & Hartigan, J. 1987. Statistical Analysis of Hominoid Molecular Evolution. *Statistical Science* **2**(2): 191–207.
7. Farris, J. 1973. A Probability Model for Inferring Evolutionary Trees. *Systematic Biology* **22**(3): 250–256.
8. Wheeler, W. 2012. Systematics: A Course of Lectures. First. Oxford: Wiley-Blackwell, 2012, 426. ISBN: 9780470671702.
9. Darriba, D. & Posada, D. jModelTest 2.0 v0.1.1. <http://code.google.com/p/jmodeltest2/>. 2015.
10. Darriba, D.; Taboada, G. L.; Doallo, R & Posada, D. 2003. jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods* **9**: 772.
11. Guidon, S & Gascuel, O. 2012. A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Systematic Biology* **52**: 696–704.
12. Nguyen, L. T.; Schmidt, H. A.; von Haeseler, A & Minh, B. Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood. *Molecular Biology and Evolution* **32**: 268–274.
13. Kalyaanamoorthy, S; Minh, B. Q.; Wong, T. K. F.; von Haeseler, A & Jermiin, L. S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods* **14**: 587–589.