Chapter 16
Maximum Likelihood in Molecular Phylogenetics

1 Introduction

Both MP and ML methods operate on a set of aligned sequences typically represented as an $N \times L$ character matrix, where N is the number of sequences and L is the aligned sequence length. The core algorithms of the two methods take the character matrix and a topology as input, assign a value to each of the L sites, and sum up the values as a criterion for choosing the best topology. The site-specific value is the substitution cost (c_i) in the MP method and log-likelihood ($\ln L_i$) in the ML method. The best topology is one with the smallest Σc_i in MP or the largest $\Sigma \ln L_i$ in ML, where $i = 1, 2, \ldots, L$.

While the key algorithm in for the MP method is the Fitch (1971) and the Sankoff (1975) algorithm, the key algorithm in the ML approach is the pruning algorithm for computing the likelihood based on equilibrium frequencies and transition probabilities in phylogenetic reconstruction. The pruning algorithm not only speeds up computation but also offers a natural and statistically valid way to handle missing data.

This chapter deals only with tree reconstruction by the ML method based on sequence data. For the ML method for computing pairwise evolutionary distances, please review the chapter on substitution models and the chapter on distance-based phylogenetic methods. For comparative methods using the likelihood approach based on Brownian motion model, with a post hoc modification for characterizing directional changes, please read the chapter on comparative methods. For the likelihood method and likelihood ratio test used to characterize association between discrete characters, please read the chapter on large-scale characterization on association between discrete characters.

We will first present simple examples of likelihood-based estimation to illustrate the rationale of the maximum likelihood approach and then detail the likelihood calculation given a topology and a set of aligned sequences. The basic knowledge needed for this chapter is tree traversal and transition probabilities derived from

X. Xia, *Bioinformatics and the Cell*, https://doi.org/10.1007/978-3-319-90684-3_16

substitution models. Readers who do not know transition probabilities should review a previous chapter on substitution models. We will illustrate the likelihood method with the simplest scenario with four sequences labeled S1 to S4 and three possible unrooted trees. We need to compute the log-likelihood (lnL) for each tree and choose the tree with the maximum likelihood as the ML tree. We will first take a brute-force approach to computing the lnL of the tree and then introduce the pruning algorithm (Felsenstein 1973, 1981, 2004, pp. 253–255). The last section illustrates a bias in the likelihood method when missing data is associated with rate heterogeneity among sites (Xia 2014). Likelihood-based phylogenetic methods, as well as the associated statistics for alternative tree topologies, are implemented in DAMBE (Xia 2013, 2017d).

2 The Rationale of Maximum Likelihood Approach

Maximum likelihood (ML) is a criterion in model selection and in statistical estimation of model parameters, i.e., the best model and the best parameter values given a model are those that maximize the likelihood. Recall that we always need to have a criterion whenever we need to choose one from several alternatives. Two sequences could have different pairwise alignments, and our criterion is the alignment score given a scoring scheme (specified by gap penalties and a match/mismatch matrix). Alignment A is better than Alignment B if the former has higher alignment score than the latter. In the chapter on Gibbs sampler, we could have different sets of putative motifs, and our criterion is the Kullback-Leibler information (F). Any set of putative motifs that gives us the largest F is the best set. In the chapters on substitution models and distance-based phylogenetic methods, we have used the least-squares criterion (both the ordinary and the weighted least-squares). A set of branch lengths that minimize the residual sum of squares is the best branch length estimates. For choosing the best tree among different alternatives, we have used the minimum evolution criterion, i.e., whichever tree has the shortest tree length is the best tree. Maximum likelihood is just one of these criteria for making a choice among alternatives.

Let us illustrate the ML approach with a few examples. Suppose we wish to estimate the proportion of males (p) of a fish population in a large lake. A random sample of N fish contains M males. With the binomial distribution, the likelihood, which is the probability mass function for discrete variables, is

$$L = \frac{N!}{M!(N-M)!} p^M (1-p)^{N-M}. \tag{16.1}$$

The maximum criterion states that the best p should maximize the likelihood value given the observation. This maximization process is simplified by maximizing the natural logarithm of L instead:

$$\begin{aligned} \ln L &= A + M \ln(p) + (N - M) \ln(1 - p) \\ \frac{\partial \ln L}{\partial p} &= \frac{M}{p} - \frac{N - M}{1 - p} = 0 \\ p &= \frac{M}{N}. \end{aligned} \tag{16.2}$$

The likelihood estimate of the variance of p is the negative reciprocal of the second derivative,

$$\mathrm{Var}(p) = -\frac{1}{\frac{\partial^2 \ln(L)}{\partial p^2}} = -\frac{1}{-\frac{M}{p^2} - \frac{N-M}{(1-p)^2}} = \frac{p(1-p)}{N}. \tag{16.3}$$

Note that the likelihood method needs a model (binomial distribution in our example) to formulate the likelihood function, which is in Eq. (16.1) for our example. For this reason, the likelihood method is always model-based.

Let us have just one more example to illustrate the likelihood approach in model selection. Suppose we wish to know whether body height differs between male and female students. We randomly sampled seven male and six female students and measured body height (shown in the first two columns in Table 16.1). We consider three hypotheses (or models). The first (M1) assumes that male and female students do not differ in body height and all 13 values represent sample values from the same normal distribution specified by mean μ and standard deviation σ. The likelihood for continuous variables is the probability density function. For example, the likelihood for body height of 170:

$$L(170|M1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{170-\mu}{\sigma}\right)^2} \tag{16.4}$$

The log-likelihood function for the 13 sample values is

Table 16.1 Body height (in cm) measured from seven male (M) and six female (F) students

Height		M1 (μ,σ)		M2 (μ_M,σ_M,μ_F,σ_F)		M3 (μ_M,μ_F,σ)	
M	F	lnL_M	lnL_F	lnLM	lnLF	lnLM	lnLF
170	170	−2.92876	−2.92876	−3.92540	−2.49898	−3.92540	−2.58489
175	165	−2.86487	−3.54639	−2.63108	−2.49893	−2.63108	−2.58485
175	172	−2.86487	−2.83676	−2.63108	−2.93880	−2.63108	−2.92317
180	170	−3.35471	−2.92876	−2.54483	−2.49898	−2.54483	−2.58489
181	168	−3.51913	−3.10936	−2.67254	−2.31048	−2.67254	−2.43991
179	160	−3.21244	−4.71774	−2.46543	−4.06956	−2.46543	−3.79288
185		−4.39828		−3.66665		−3.66665	

Three models (M1, M2, and M3) are evaluated, assuming normal distribution. M1: male and female students do not differ in body height, and all 13 values represent sample values from the same population. M2: the seven male and six female sample values are from two populations differing in both mean and standard deviation. M3: the seven male and six female sample values are from two populations differing in mean but having the same standard deviation

$$\ln L_{M1} = \ln [L(170|M1)] + \ln [L(175|M1)] + \cdots + \ln [L(160|M1)] \tag{16.5}$$

which is a function of two unknowns (μ and σ). The likelihood criterion states that the best μ and σ should maximize ln$L$. We thus take partial derivatives of lnL_{M1} with respect to μ and σ, set the partial derivatives to zero, and solve the two simultaneous equations. This gives us $\mu = 173.0769$ and $\sigma = 6.7192$ and $\ln L_{\mathrm{M1}} = -43.2108$. ln$L$ for individual observations are shown in Table 16.1, in the two columns under the heading of M1. This M1 is equivalent to the null hypothesis of no difference in body height between male and female students.

Suppose we now have a second hypothesis (M2) that sample values from males and females belong to two different normal distributions, with the male population specified by μ_{M} and σ_{M}, and the female population specified by μ_{F} and σ_{F}. Now the same sample value (e.g., 170) for a male and a female student will have different likelihoods, denoted $L_{\mathrm{M}.170}$ and $L_{\mathrm{F}.170}$ below

$$L(170, \text{male}|M2) = \frac{1}{\sigma_M\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{170-\mu_M}{\sigma_M}\right)^2}; \quad L(170, \text{female}|2)\frac{1}{\sigma_F\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{170-\mu_F}{\sigma_F}\right)^2} \tag{16.6}$$

Designating body height for males and females as MH and FH, respectively, the log-likelihood (ln$L$) for the seven male and six female sample values is

$$\ln L_{M2} = \sum_{i=1}^{7} \ln L(\mathrm{MH}_i|M2) + \sum_{i=1}^{6} \ln L(\mathrm{FH}_i|M2) \tag{16.7}$$

Now the function lnL_{M2} has four parameters. The maximum likelihood estimates of these parameters (i.e., the parameters that maximize lnL_{M2}) are $\mu_{\mathrm{M}} = 177.8570$, $\sigma_{\mathrm{M}} = 4.5491$, $\mu_{\mathrm{F}} = 167.4998$, and $\sigma_{\mathrm{F}} = 3.9896$, with the resulting $\ln L_{\mathrm{M2}} = -37.3527$. The log-likelihood of these individual sample values are shown in Table 16.1 in the two columns under M2.

Which model (M1 or M2) is the more preferable? M1 is simpler, with only two parameters, but is it too simple as to fail to describe nature adequately? What criterion should we use to discriminate between these two models? Just as R^2 is a measure of how well a model fits the data in a least-squares context, ln$L$ is a measure of how well a model fits the data in a likelihood context. However, just as R^2 is not a good criterion for model selection, so is ln$L$. A model will increase its fit to the data if we keep adding parameters. If we use R^2 or ln$L$ as a criterion for model selection, then complicated models will be favored against simple models.

A series of models are termed nested models if the simpler model is a special case of the more complicated model. For example, the following are nested models because the second can be reduced to the first if $b = 0$ and the third can be reduced to the second if $b_2 = 0$:

$$\begin{aligned} y &= a \\ y &= a + bx \\ y &= a + b_1 x + b_2 x^2 \end{aligned} \tag{16.8}$$

In our case, M1 and M2 are nested model because M2 is reduced to M1 if $\mu_M = \mu_F$ and $\sigma_M = \sigma_F$.

Nested models can be tested by the likelihood ratio test (LRT). Any statistical significance test will have two essential quantities: a statistic that measures the difference between the two models (two hypotheses) and a known distribution of the statistic. The statistic in LRT is $2\Delta \ln L$ (twice the difference in $\ln L$ between the two nested models, which is often referred to as the likelihood ratio chi-square) which follows approximately the χ^2 distribution with the degree of freedom being Δp (the difference in the number of parameters between the two nested models). In our example, $2\Delta \ln L = 2^*(\ln L_{M2} - \ln L_{M1}) = 11.7162$. With two degrees of freedom, $p = 0.0029$. Note that the null hypothesis being tested in LRT is that M1 and M2 are equally good, and this null hypothesis is rejected at $p = 0.0029$, so we adopt M2 and conclude that male and female students differ highly significantly in body height.

Now suppose we have a third hypothesis (M3) stating that the male and female populations differ in means but not in standard deviation, i.e., $\mu_M \neq \mu_F$ but $\sigma_M = \sigma_F = \sigma$. Is this M3 as good as M2? We can do the same calculation to obtain $\mu_M = 177.8570$, $\mu_F = 167.4998$, and $\sigma = 4.5491$, with $\ln L_{M3} = -37.4476$. The $\ln L$ for individual sample values are also shown in Table 16.1 in the two columns under M3. The likelihood ratio chi-square $2\Delta \ln L = 0.1897$. With one degree of freedom, $p = 0.6631$, so we cannot reject the null hypothesis that M3 and M2 are equally good.

3 Likelihood for a Phylogenetic Tree

Students often find it a steep learning curve to proceed from the sections above to the sections below. The first difficulty is the tree traversal, the second is the pruning algorithm, and the third is that they have forgotten about substitution models when the classes come to this point. It is important that students should review the substitution models. We will first go slowly with tree traversal with a brute-force approach without the pruning algorithm, which is followed by a detailed numerical illustration of the pruning algorithm, together with missing data handling.

3.1 The Brute-Force Approach

Given a site in a set of aligned sequences and a topology (Fig. 16.1), we have two internal nodes with unknown states, leading to 16 possible nucleotide configurations

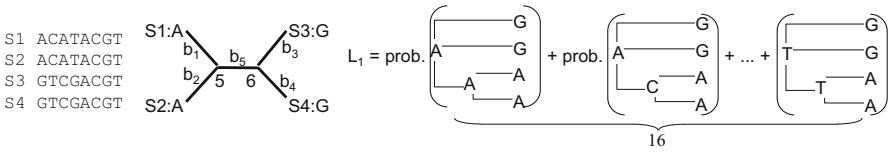

Fig. 16.1 Computing likelihood with a brute-force approach involving one site (site 1) in a set of four aligned sequences. Nodes 5 and 6 have unknown states that could be A, C, G, or T, generating 16 different combinations. We need to compute likelihood for each site designated L_1 to L_{16}. The log-likelihood of a tree is $\ln L = \sum \ln(L_i)$, and the maximum likelihood tree is the one with the highest $\ln L$ among all possible topologies

with the same topology. At first sight, it seems that we need to sum up 16 terms to obtain the likelihood of each site (Fig. 16.1). The likelihood for site 1 is

$$\begin{aligned} \mathrm{L}_1 = {} & \pi_{\mathrm{A}} P_{\mathrm{AA}}(b_1) P_{\mathrm{AA}}(b_2) P_{\mathrm{AA}}(b_5) P_{\mathrm{AG}}(b_3) P_{\mathrm{AG}}(b_4) \\ & + \pi_{\mathrm{C}} P_{\mathrm{CA}}(b_1) P_{\mathrm{CA}}(b_2) P_{\mathrm{CA}}(b_5) P_{\mathrm{AG}}(b_3) P_{\mathrm{AG}}(b_4) \\ & + \cdots + \pi_{\mathrm{T}} P_{\mathrm{TA}}(b_1) P_{\mathrm{TA}}(b_2) P_{\mathrm{TT}}(b_5) P_{\mathrm{TG}}(b_3) P_{\mathrm{TG}}(b_4) \end{aligned} \tag{16.9}$$

where π_{A}, π_{C}, π_{C}, and π_{T} are equilibrium frequencies, b_1 to b_5 are branch lengths, and P_{AA}, P_{AG}, …, P_{TT} are transition probabilities derived from a given substitution model. One can express the likelihood for other sites in the same way. For example, the likelihood for site 5 is

$$\begin{aligned} \mathrm{L}_5 = {} & \pi_{\mathrm{A}} P_{\mathrm{AA}}(b_1) P_{\mathrm{AA}}(b_2) P_{\mathrm{AA}}(b_5) P_{\mathrm{AA}}(b_3) P_{\mathrm{AA}}(b_4) \\ & + \pi_{\mathrm{C}} P_{\mathrm{CA}}(b_1) P_{\mathrm{CA}}(b_2) P_{\mathrm{CA}}(b_5) P_{\mathrm{AA}}(b_3) P_{\mathrm{AA}}(b_4) \\ & + \cdots + \pi_{T} P_{\mathrm{TA}}(b_1) P_{\mathrm{TA}}(b_2) P_{\mathrm{TT}}(b_5) P_{\mathrm{TA}}(b_3) P_{\mathrm{TA}}(b_4) \end{aligned} \tag{16.10}$$

For illustration, we may take equal nucleotide frequencies and the simplest JC69 model with only two distinct transition probabilities:

$$P_{ii}(b) = \frac{1}{4} + \frac{3}{4} e^{-4b/3}, P_{ij}(b) = \frac{1}{4} - \frac{1}{4} e^{-4b/3} \tag{16.11}$$

where b is branch length between neighboring nodes.

Because the JC69 model assumes that the four nucleotides replace each other with equal rate, the four sequences in Fig. 16.1 have only two site patterns, one shared among sites 1–4 and the other shared among sites 5–8. Given the simple JC60 model, the 16 terms for the first site pattern in Eq. (16.9) can be reduced to seven terms because many terms are identical. For example, seven terms are identical and equal to $P_{ij}(b_1)$*$P_{ij}(b_2)$* $P_{ij}(b_3)$* $P_{ij}(b_4)$* $P_{ij}(b_5)$*0.25 when internal nodes 5 and 6 are occupied by nucleotide pairs (C, A), (C, T), (G, A), (G, C), (G, T), (T, A), and (T, C), where the first nucleotide within parenthesis is at node 5 and second at node 6. Similarly, the 16 terms for the second pattern in Eq. (16.10) can be reduced to five terms. The tree lnL given the four sequences in Fig. 16.1 is

$$\ln L = 4 \ln L_1 + 4 \ln L_5 \tag{16.12}$$

Maximizing ln$L$ results in $b_1 = b_2 = b_3 = b_4 = 0$ and $b_5 = 0.823959217$, with ln$L = -21.029981488111$. That $b_1 = b_2 = b_3 = b_4 = 0$ is as expected because S1 = S2 and S3 = S4 (Fig. 16.1), so there are really just two sequences instead of four. Note that if we do not use lnL but use L instead, then $L = 7.358598123*10^{-10}$, which is already small. A larger tree with many OTUs will result in an L so small that computers will not distinguish it from zero.

We have evaluated just one topology in Fig. 16.1 for the four OTUs. There are two other unrooted topologies for four OTUs, one with OTUs S1 and S3 clustered together and the other with OTUs S1 and S4 clustered together. L_5 in Eq. (16.10) is the same for all three topologies, but we need to recompute L_1. The resulting ln$L$, given the JC69 model, for these two topologies are the same and equal to -30.96960809. Thus, the topology in Fig. 16.1, with S1 and S2 clustered together and $b_1 = b_2 = b_3 = b_4 = 0$ and $b_5 = 0.823959217$, is the best tree of the three, because it has the largest ln$L$ ($= -21.029981488111$).

3.2 *The Pruning Algorithm*

The brute-force approach for computing ln$L$ is unnecessary and is not used in practice other than classroom or textbook illustration. The pruning algorithm (Felsenstein 1973, 1981, 2004, pp. 253–255), illustrated below, economizes the computation substantially.

As in the maximum parsimony method, we only need to illustrate the application of the pruning algorithm for a single site because all sites are computed the same way. For any given topology, e.g., the four-species topology in Fig. 16.1, we first define a likelihood vector (L_j) for each of the nodes including the leaf nodes. The vector contains four elements for nucleotide sequences, 20 for amino acid sequences or the number of sense codons for codon sequences with codon-based models. We will use nucleotide sequences for illustration, but the computation is the same for amino acid or codon sequences.

We have four sequences with the first site being A, G, C, and T for species 1, 2, 3, and 4, respectively (Fig. 16.2). Our task is to compute the lnL_1 for this first site with the pruning algorithm. The computation is the same for all other sites.

For a leaf node j with a nucleotide s (where s is either A, C, G, or T), $L_j(\mathrm{s}) = 1$, and $L_j(\bar{\mathrm{s}}) = 0$ (Fig. 16.2). For example, for the first sequence with nucleotide A, $L_1(\mathrm{A}) = 1$, and $L_1(\mathrm{C}) = L_1(\mathrm{G}) = L_1(\mathrm{T}) = 0$. For an internal node j with two offspring (o_1 and o_2), L_j is recursively defined as

$$L_j(s) = \left[\sum_{k=0}^{3} P_{sk}\left(b_{j,o_1}\right) L_{o_1}(k)\right]\left[\sum_{k=0}^{3} P_{sk}\left(b_{j,o_2}\right) L_{o_2}(k)\right] \tag{16.13}$$

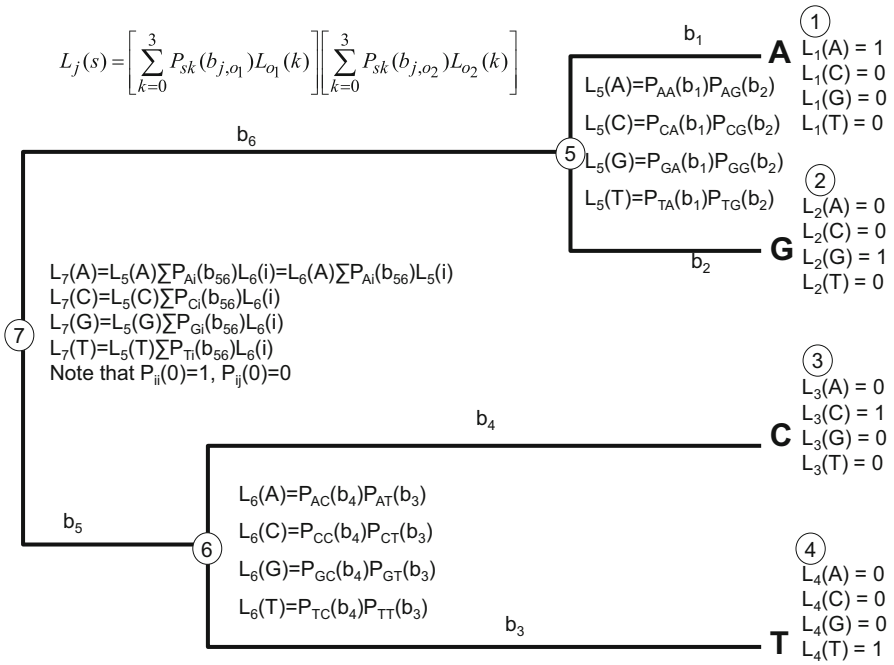

Fig. 16.2 Likelihood computation with the pruning algorithm on a four-species tree. Node j is represented by a vector (L_j) of four elements for nucleotide sequences or 20 for amino acid sequences. L_j is computed according to equation on the upper-left, i.e., Eq. (16.13). b_5 and b_6 cannot be estimated separately without assuming a molecular clock, and only their summation (b_{56}) is used in likelihood calculation

where s is either A, C, G, or T; k takes value of 0, 1, 2, and 3 corresponding to nucleotides A, C, G, and T; $b_{j,o1}$ is the branch length between internal node j and its offspring o_1; and P_{sk} is the transition probability from state s to state k (where s and k are A, C, G, or T for nucleotide sequences). Transition probabilities P_{sk} and its derivation from various substitution models have been detailed in the chapter on substitution models.

The application of Eq. (16.13) is straightforward. Take, for example, internal node 5 with its two offspring nodes 1 and 2,

$$L_5(\mathrm{A}) = \left[\sum_{k=0}^{3} P_{Ak}(b_1) L_1(k)\right]\left[\sum_{k=0}^{3} P_{Ak}(b_2) L_2(k)\right] \tag{16.14}$$

Because $L_1(\mathrm{A}) = 1$, $L_1(\mathrm{C}) = L_1(\mathrm{G}) = L_1(\mathrm{T}) = 0$, $L_2(\mathrm{G}) = 1$, $L_2(\mathrm{A}) = L_2(\mathrm{C}) = L_2(\mathrm{T}) = 0$, $L_5(\mathrm{A})$ becomes

$$L_5(\mathrm{A}) = P_{\mathrm{AA}}(b_1) P_{\mathrm{AG}}(b_2) \tag{16.15}$$

The other three elements, as well as L_j vectors for other internal nodes, are listed in Fig. 16.2.

Internal node 7 is special in that we cannot estimate b_5 and b_6 separately because the substitution model is time-reversible and the resulting tree is consequently unrooted. We simply move node 7 to the location of node 6 (or node 5), so that either b_5 or b_6 is 0 and the other is then equal to $(b_5 + b_6)$ represented as b_{56} in L_7 in Fig. 16.2. If b_5 is 0, then $P_{ii}(b_5) = 1$ and $P_{ij}(b_5) = 0$, i.e., no time for anything to change. This leads to the simplified equations for computing $L_7(i)$ in Fig. 16.2. The final likelihood for the tree is

$$L = \sum_{i=0}^{3} \pi_i L_7(i) \tag{16.16}$$

where π_i is the equilibrium frequency of nucleotide i and reflects the assumption that sufficient time has elapsed for the frequencies to reach equilibrium. Note that, for the JC69 model, $\pi_i = 1/4$.

The application of the pruning algorithm to the aligned sequences and the topology in Fig. 16.1 with the JC69 model will also result in $b_1 = b_2 = b_3 = b_4 = 0$ and $b_5 = 0.823959217$, with $\ln L = -21.029981488111$, just as we have obtained with the brute-force approach. The benefit of the pruning algorithm is the reduction of repeated calculation.

4 Calculating Likelihood by Imposing a Molecular Clock

Recall that molecular phylogenetics has two main objectives: (1) defining the branching pattern and (2) dating evolutionary events such as speciation events or gene duplication events. A good molecular clock is important for dating. To test the molecular clock hypothesis, we need to compute ln$L$ by imposing a molecular clock and lnL without a molecular clock and then use the likelihood ratio test to see if the two fit the data equally well. If we have long sequence alignment but the molecular clock hypothesis is not rejected, then we can calibrate the clocked tree to perform dating.

This section illustrates the calculation of likelihood given the sequence alignment and topology in Fig. 16.3. The pruning algorithm is the same as before except that we constrain $b_1 = b_2$, $b_3 = b_4$, and $(b_1 + b_5) = (b_3 + b_6)$. For simplicity, we will again use the JC69 model, which allows us to reduce the eight aligned sites (Fig. 16.3) to three site patterns: pattern 1 for sites 1–2, pattern 2 for sites 3–4, and pattern 3 for sites 5–8. Thus, we only need to compute the likelihood for sites 1, 3, and 5.

For site 1 (site pattern 1), the likelihood vector L for the internal nodes 5, 6 and 7 are

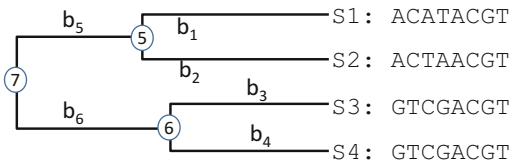

Fig. 16.3 Sequence alignment and topology for illustrating likelihood calculation with a molecular clock. Imposing a molecular clock implies $b_1 = b_2$, $b_3 = b_4$, and $(b_1 + b_5) = (b_3 + b_6)$. So there are only three branch lengths to estimate. Without a clock we would need to estimate five branch lengths: b_1, b_2, b_3, b_4, and b_{56}

$$\begin{aligned} L_{5.1}(\mathrm{A}) &= P_{\mathrm{AA}}(b_1)P_{\mathrm{AA}}(b_2) = P_{ii}^2(b_1) \\ L_{5.1}(\mathrm{C}) &= P_{\mathrm{CA}}(b_1)P_{\mathrm{CA}}(b_2) = P_{ij}^2(b_1) = L_{5.1}(\mathrm{G}) = L_{5.1}(\mathrm{T}) \end{aligned} \tag{16.17}$$

where $L_{5.1}$ denotes the L vector for internal node 5 with site pattern 1.

$$\begin{aligned} L_{6.1}(\mathrm{A}) &= P_{\mathrm{AG}}(b_3)P_{\mathrm{AG}}(b_4) = P_{ij}^2(b_3) = L_{6.1}(\mathrm{C}) = L_{6.1}(\mathrm{T}) \\ L_{6.1}(\mathrm{G}) &= P_{\mathrm{GG}}(b_3)P_{\mathrm{GG}}(b_4) = P_{ii}^2(b_3) \end{aligned} \tag{16.18}$$

$$\begin{aligned} L_{7.1}(\mathrm{A}) &= \left[\sum_{k=0}^{3} P_{\mathrm{A}k}(b_5)L_{5.1}(k)\right]\left[\sum_{k=0}^{3} P_{\mathrm{A}k}(b_6)L_{6.1}(k)\right] \\ L_{7.1}(\mathrm{C}) &= \left[\sum_{k=0}^{3} P_{\mathrm{C}k}(b_5)L_{5.1}(k)\right]\left[\sum_{k=0}^{3} P_{\mathrm{C}k}(b_6)L_{6.1}(k)\right] \\ L_{7.1}(\mathrm{G}) &= \left[\sum_{k=0}^{3} P_{\mathrm{G}k}(b_5)L_{5.1}(k)\right]\left[\sum_{k=0}^{3} P_{\mathrm{G}k}(b_6)L_{6.1}(k)\right] \\ L_{7.1}(\mathrm{T}) &= \left[\sum_{k=0}^{3} P_{\mathrm{T}k}(b_5)L_{5.1}(k)\right]\left[\sum_{k=0}^{3} P_{\mathrm{T}k}(b_6)L_{6.1}(k)\right] \end{aligned} \tag{16.19}$$

For site 3 (site pattern 2), the L vectors for internal nodes 5 and 6 are specified below, and $L_{7.1}$ is specified the way as that in Eq. (16.19) except that $L_{5.1}$ and $L_{6.1}$ are, respectively, replaced by $L_{5.2}$ and $L_{6.2}$ specified below:

$$\begin{aligned} L_{5.2}(\mathrm{A}) &= L_{5.2}(\mathrm{T}) = P_{ii}(b_1)P_{ij}(b_1) \\ L_{5.2}(\mathrm{C}) &= L_{5.2}(\mathrm{G}) = P_{ij}^2(b_1) \end{aligned} \tag{16.20}$$

$$\begin{aligned} L_{6.2}(\mathrm{A}) &= L_{6.2}(\mathrm{G}) = L_{6.2}(\mathrm{T}) = P_{ij}^2(b_3) \\ L_{6.2}(\mathrm{C}) &= P_{ii}^2(b_3) \end{aligned} \tag{16.21}$$

For site 5 (site pattern 3), the L vectors for internal nodes 5 and 6 are shown below, and L_7 is specified the way as that in Eq. (16.19) except that $L_{5.1}$ and $L_{6.1}$ are, respectively, replaced by $L_{5.3}$ and $L_{6.3}$ specified below:

$$\begin{aligned} L_{5.3}(\mathrm{A}) &= P_{ii}^2(b_1) \\ L_{5.3}(\mathrm{C}) &= L_{5.3}(\mathrm{G}) = L_{5.3}(\mathrm{T}) = P_{ij}^2(b_1) \end{aligned} \tag{16.22}$$

$$\begin{aligned} L_{6.3}(\mathrm{A}) &= P_{ii}^2(b_3) \\ L_{6.3}(\mathrm{C}) &= L_{6.3}(\mathrm{G}) = L_{6.3}(\mathrm{T}) = P_{ij}^2(b_3) \end{aligned} \tag{16.23}$$

The likelihood for the three site patterns, designated as L_1, L_2, and L_3, are

$$L_1 = \frac{1}{4}\sum_{i=0}^{3} L_{7.1}(i);\;\; L_2 = \frac{1}{4}\sum_{i=0}^{3} L_{7.2}(i);\;\; L_3 = \frac{1}{4}\sum_{i=0}^{3} L_{7.3}(i) \tag{16.24}$$

where 1/4 is the equilibrium frequencies (π_i) for the JC69 model. The log-likelihood ($\ln L$) given the topology and the sequence alignment in Fig. 16.3 is

$$\ln L = 2\ln(L_1) + 2\ln(L_2) + 4\ln(L_3) \tag{16.25}$$

which has three unknown branch lengths (b_1, b_3, and b_5). The b_1, b_3, and b_5 values that maximize $\ln L$, subject to the constraints that branch lengths cannot be negative, are $b_1 = 0.1597105$, $b_3 = 0$, and $b_5 = 0.3011361$. Given the clock constraint that $(b_1 + b_5) = (b_3 + b_6)$, we have $b_6 = b_1 + b_5 - b_3 = 0.460846$. The L-BFGS-B algorithm (Zhu et al. 1997) is often used for such constrained optimization, e.g., with the lower bounds for branch lengths set to zero. The resulting $\ln L = -27.63046$, with $L_1 = 0.02970894$, $L_2 = 0.003665145$, and $L_3 = 0.09584556$. Note that variable sites in general will have smaller likelihood than conserved sites.

Before one calibrate the clocked tree with dated fossils, it is customary to test the clock hypothesis by likelihood rate test (LRT). Note that a tree without a clock is more general than a tree with a clock. With the former we need to estimate $(N-2)$ more branch lengths than with the latter, where N is the number of OTUs. With four OTUs, a tree without clock will have five branch lengths to estimate, in contrast to only three in a tree without a clock. The procedure of LRT is to estimate $\ln L$ for the tree with and without the clock, designated $\ln L_{\text{clock}}$ and $\ln L_{\text{no.clock}}$, respectively. The likelihood ratio chi-square is $2(\ln L_{\text{no.clock}} - \ln L_{\text{clock}})$ with $(N-2)$ degrees of freedom. To discriminate between the two models, one typically needs to have much longer sequences than those in Fig. 16.3, for two reasons. First, there would be little statistical power to reject the null hypothesis if we have little data, even if the null hypothesis is false. Second, the likelihood ratio chi-square may not follow chi-square distribution when there is little data, so that resulting p value will not be accurate.

The sequences in Fig. 16.3 cannot be used to illustrate the test of molecular clock because S3 and S4 are identical and S1 and S2 diverge equally from S3/S4, so we know a priori that the sequences conform to the clock hypothesis. In other words, the clock model and the non-clock model will have the same $\ln L$. If we change the third site of S3 from C to T, then, again applying the JC69 model and the pruning algorithm as illustrated before for the clock and no-clock hypotheses, we will find lnL equal to -31.2924 with clock and $-$ 30.4874 without clock. This gives us the likelihood ratio chi-square of $2\Delta\ln L$ equal 1.6099. With two degrees of freedom,

$p = 0.4471$, so we do not reject the molecular clock hypothesis. Keep in mind that this illustration has two problems because of the short sequences (sequence length of 8). First, there is little power to reject the null hypothesis. Second, approximate of $2\Delta\ln L$ distribution by χ^2 distribution may not be accurate with small samples. Fortunately, we always have more data in real research.

5 Handling of Missing Data with the Pruning Algorithm (and the Potential Bias)

The pruning algorithm facilitates the handling of missing data. The key requirement for handling missing data is that what is missing will not contribute anything to the choice of the best tree. The maximum likelihood method, when implemented properly, is not biased with missing data. However, phylogenetic bias may be induced by missing data in conjunction with rate heterogeneity over sites. I will illustrate the handling of missing data by the pruning algorithm and the potential phylogenetic bias based on a previous publication (Xia 2014).

Suppose we have data in Fig. 16.4a and need to evaluate the three possible unrooted trees (T_1, T_2, and T_3 in Fig. 16.4). It is obvious that the only information we have for the data set is the distance between S1 and S2, so we should not be able to discriminate among the three topologies. We also note that, given the JC69 model, the sequences in Fig. 16.4a have two site patterns, with the first four sites sharing one site pattern (i.e., with the same site-specific likelihood) and the last four sites sharing the other site pattern. We therefore need to compute the log-likelihood for only the first site ($\ln L_1$) and the fifth site ($\ln L_5$) and multiply them by 4 to get $\ln L$ for the entire alignment.

Fig. 16.4 Aligned sequences with missing data (**a**-**b**) for illustrating missing data handling by the likelihood method and the potential bias induced by missing data in conjunction with rate heterogeneity

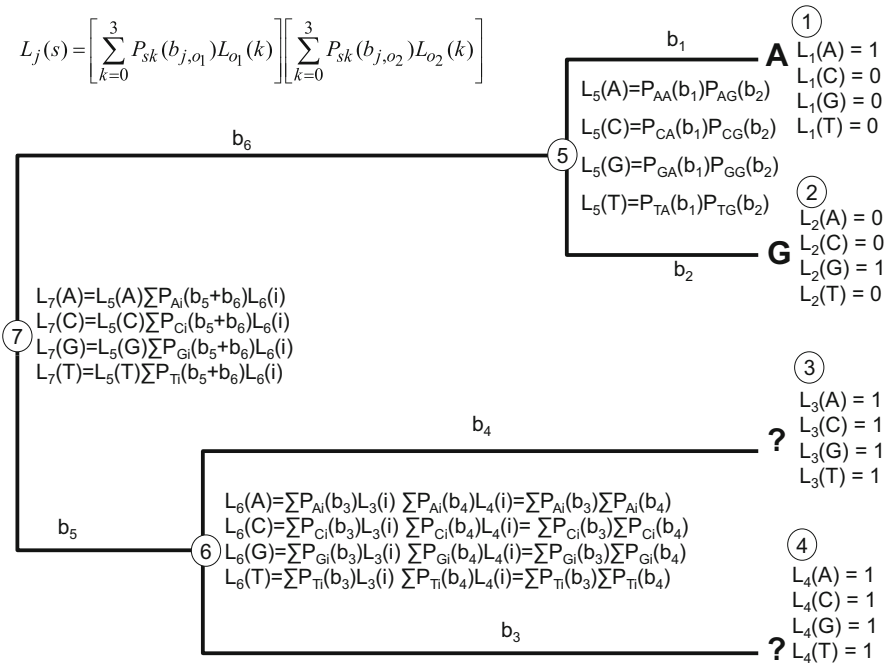

Fig. 16.5 Likelihood computation with the pruning algorithm and missing data

The computation of $\ln L_1$, given Topology T_1, is illustrated in Fig. 16.5. For an unknown or missing nucleotide, we simply set $L_j(\mathrm{A}) = L_j(\mathrm{C}) = L_j(\mathrm{G}) = L_j(\mathrm{T}) = 1$. The likelihood calculation them proceeds exactly as before. The log-likelihood ($\ln L$) for all eight sites, given topology T_1 in Fig. 16.4, is

$$\ln L = 4 \ln L_1 + 4 \ln L_5 \tag{16.26}$$

which, upon maximization, leads to $b_1 + b_2 = 0.8239592165$ and $\ln L = -21.02998149$. Terms containing b_3, b_4, and $b_5 + b_6$ all cancel out, i.e., the sequences in Fig. 16.4a have no information for estimating b_3, b_4, and $b_5 + b_6$, which again is what we would have expected. The resulting distance between S1 and S2 ($= b_1 + b_2$) is the same if we just use the distance formula for two sequences.

If we perform the computation again with topology T_2 in Fig. 16.4, we will have exactly the same $\ln L$, but $b_5 + b_6$ will be 0 and $b_1 + b_3 = 0.8239592165$ (i.e., the distance between OTUs S1 and S2 is 0.8239592165 as before). This again is perfectly consistent with our common sense. Topology T_3 in Fig. 16.4 will lead to the same $\ln L$ and the same conclusion with distance between S1 and S2 being 0.8239592165.

Our happy feeling with the likelihood method, however, does not last forever. Suppose now we have sequence data in Fig. 16.4b, with Gene1 being variable but Gene2, which is missing in S3 and S4, is so conservative as to be invariant. In

practice, Gene1 and Gene2 could be different segments within the same gene, e.g., the conserved and variable domains in ribosomal RNAs with no clear boundary between them. Note that the three variable sites at the 5′-end could be scattered over different sites in the data instead of clumping together to be as easily recognizable as in Fig. 16.4a.

The sequences are intentionally made not to favor any one of the three possible topologies in Fig. 16.4). For Gene1, the four OTUs are exactly equally divergent from each other given the JC69 or more complicated models, i.e., each pair of sequences differ in exactly one transition and two transversions so that no particular topology is favored over the other two. Gene2 is extremely conservative and no substitution has been observed, so it also should not favor any topology over the other two. If Gene2 is not missing in S3 and S4, then all three topologies will be equally supported.

With the sequence data in Fig. 16.4b and topology T_1 in Fig. 16.4, we can apply the pruning algorithm and the JC69 model to compute the likelihood. There are only three different site patterns with the JC69 model, i.e., sites 1–3 sharing the first site pattern, sites 4–7 sharing the second, and sites 8–39 sharing the third. Maximizing the likelihood leads to $\ln L = -83.56464029$ which is reached when $b_1 = b_2 = 0.04153005797$, $b_3 = b_4 = 0.3787544804$, and $(b_5 + b_6) = 0.3511004094$.

The maximum $\ln L$ value for topology T_2 in Fig. 16.4 is -83.96663731, reached when $b_1 = b_3 = 0.04184900$, $b_2 = b_4 = 0.60765526$, and $(b_5 + b_6) = 0.000947018$. The maximum $\ln L$ value for topology T_3 is the same as that for T_2 and both are significantly smaller ($p < 0.001$) than that for T_1 (Fig. 16.4) based on either the Kishino-Hasegawa test or RELL test (Kishino and Hasegawa 1989) or Shimodaira and Hasegawa test (Shimodaira and Hasegawa 1999).

We have previously mentioned that the most fundamental criterion for missing data handling methods is that the missing data should not contribute phylogenetically relevant information. The demonstration above shows clearly that missing data do contribute such information, even with the likelihood approach. If the data are not missing, then the three topologies will be equally supported. So the bias in favor of topology T_1 in the presence of missing data can only be attributed to the presence of missing data.

While the phylogenetic bias in the sequence configuration in Fig. 16.4b favors the grouping S_1 and S_2 together, one can easily envision scenarios in which S_1 and S_2 would repulse each other, e.g., when the last 32 sites in Fig. 16.4b are far more variable than the first seven sites. Thus, the direction of the bias cannot be predicted before data analysis.

Postscript

We have covered the maximum likelihood framework in molecular phylogenetics in depth, but this book does not cover the Bayesian approach which extended the likelihood framework to incorporate prior knowledge. The Bayesian framework can not only help us with molecular phylogenetics but also reduce our tendency to develop prejudice and social bias.

Suppose we live in a multiracial society and need to decide whom our family should interact with. We implicitly would want to estimate the proportion of good people (P_{good}) in a race (or an ethnic group), with "good people" defined as those whom we have pleasant experience interacting with. Naturally one wants to interact with people in a race whose P_{good} is high and avoid people in a race whose P_{good} is low.

Now suppose we have interacted with a small number of people, say three, in one race and our experiences are all bad. A likelihood estimate of P_{good} is then 0 because it is based on data only. If we take this estimated P_{good} seriously in spite of the small sample size of three, then we become a racist.

With the Bayesian approach, we would first conceive a prior for P_{good} before any interaction with people of different races. If we are fair-minded, our prior of P_{good} will be the same for all races to start with. If we are unfortunate to have a bad experience with a member of one race, we would reduce P_{good} for that race a bit. If our second encounter with people of this race is also bad, then we reduce P_{good} still further for that race. Eventually these different P_{good} values for different races constitute our private model of racial differences, and the model, correct or wrong, will affect our behavior.

The model of racial differences thus developed in our mind may be quite different from models in other people's mind, because different people often interact with different samples from different races. Because few of us could claim to have a representative sample of people to interact with, P_{good} is almost always biased. However, it may not be as biased as what one gets from a likelihood framework.

In this context of unrepresentative samples from differences, racism, as well as other kinds of prejudices, is almost inevitable. What is important to keep in mind is that much of the differences in P_{good} among races or ethnic groups are due to historical differences in racial environment. If a little boy is driven by poverty to steal a loaf of bread for his sick and hungry mother, then it is the ruler of the society, not the boy, who is bad. May the joint effort of mankind lead to a monotonic increase in P_{good} in all races.