Let $\hat{\kappa}_{ab}$, $\hat{\kappa}_{bc}$, and $\hat{\kappa}_{ca}$ be the estimates of the transition/transversion rate ratio κ in the three comparisons. Considering that the three sequences are related by a phylogenetic tree, we see that we estimated κ for the branch leading to sequence a as $\hat{\kappa}_{ab}$ in one comparison but as $\hat{\kappa}_{ca}$ in another. This inconsistency is problematic when complex models involving unknown parameters are used, and when information about model parameters is visible only when one compares multiple sequences simultaneously. An example is the variation of evolutionary rates among sites. With only two sequences, it is virtually impossible to decide whether a site has a difference because the rate at the site is high or because the overall divergence between the two sequences is high. Even if the parameters in the rate distribution (such as the shape parameter α of the gamma distribution) are fixed, the pairwise approach does not guarantee that a high-rate site in one comparison is also a high-rate site in another.

A second limitation is important in analysis of highly divergent sequences, in which substitutions have nearly reached *saturation*. The distance between two sequences is the sum of branch lengths on the phylogeny along the path linking the two sequences. By adding branch lengths along the tree, the pairwise distance can become large even if all branch lengths on the tree are small or moderate. As discussed above, large distances involve large sampling errors in the estimates or even cause the distance formulae to be inapplicable. By summing up branch lengths, the pairwise approach exacerbates the problem of saturation and may be expected to be less tolerant of high sequence divergences than likelihood or Bayesian methods, which compare all sequences simultaneously.

1.7 Exercises

1.1 Use the transition probabilities under the JC69 model (equation 1.3) to confirm the Chapman–Kolmogorov theorem (equation 1.4). It is sufficient to consider two cases: (a) $i = \mathrm{T}, j = \mathrm{T}$; and (b) $i = \mathrm{T}, j = \mathrm{C}$. For example, in case (a), confirm that $p_{\mathrm{TT}}(t_1 + t_2) = p_{\mathrm{TT}}(t_1)p_{\mathrm{TT}}(t_2) + p_{\mathrm{TC}}(t_1)p_{\mathrm{CT}}(t_2) + p_{\mathrm{TA}}(t_1)p_{\mathrm{AT}}(t_2) + p_{\mathrm{TG}}(t_1)p_{\mathrm{GT}}(t_2)$.

1.2 Derive the transition-probability matrix $P(t) = \mathrm{e}^{Qt}$ for the JC69 model (Jukes and Cantor 1969). Set $\pi_{\mathrm{T}} = \pi_{\mathrm{C}} = \pi_{\mathrm{A}} = \pi_{\mathrm{G}} = 1/4$ and $\alpha_1 = \alpha_2 = \beta$ in the rate matrix (1.15) for the TN93 model to obtain the eigenvalues and eigenvectors of Q under JC69, using results of Subsection 1.2.3. Alternatively you can derive the eigenvalues and eigenvectors from equation (1.1) directly. Then apply equation (1.17).

1.3 Derive the transition-probability matrix $P(t)$ for the Markov chain with two states 0 and 1 and generator matrix $Q = \begin{pmatrix} -u & u \\ v & -v \end{pmatrix}$. Confirm that the spectral decomposition of Q is given as

$$Q = U\Lambda U^{-1} = \begin{pmatrix} 1 & -u \\ 1 & v \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & -u-v \end{pmatrix} \begin{pmatrix} v/(u+v) & u/(u+v) \\ -1/(u+v) & 1/(u+v) \end{pmatrix}, \quad (1.70)$$

so that

$$P(t) = \mathrm{e}^{Qt} = \frac{1}{u+v}\begin{pmatrix} v + u\mathrm{e}^{-(u+v)t} & u - u\mathrm{e}^{-(u+v)t} \\ v - v\mathrm{e}^{-(u+v)t} & u + v\mathrm{e}^{-(u+v)t} \end{pmatrix}. \tag{1.71}$$

Note that the stationary distribution of the chain is given by the first row of U^{-1}, as $[v/(u+v), u/(u+v)]$, which can also be obtained from $P(t)$ by letting $t \to \infty$. A special case is $u = v = 1$, when we have

$$P(t) = \begin{pmatrix} \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2t} & \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-2t} \\ \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-2t} & \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2t} \end{pmatrix}. \tag{1.72}$$

This is the binary equivalent of the JC69 model.

1.4 Confirm that the two likelihood functions for the JC69 model, equations (1.42) and (1.43), are proportional and the proportionality factor is a function of n and x but not of d. Confirm that the likelihood equation, $\mathrm{d}\ell/\mathrm{d}d = \mathrm{d}\log\{L(d)\}/\mathrm{d}d = 0$, is the same whichever of the two likelihood functions is used.

***1.5** Suppose $x = 9$ heads and $r = 3$ tails are observed in $n = 12$ independent tosses of a coin. Derive the MLE of the probability of heads (θ). Consider two mechanisms by which the data are generated.

(a) *Binomial*. The number $n = 12$ tosses was fixed beforehand. In $n = 12$ tosses, $x = 9$ heads were observed. Then the number of heads x has a binomial distribution, with probability

$$f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{1.73}$$

(b) *Negative binomial*. The number of tails $r = 3$ was fixed beforehand, and the coin was tossed until $r = 3$ tails were observed, at which point it was noted that $x = 9$ heads were observed. Then x has a negative binomial distribution, with probability

$$f(x|\theta) = \binom{r+x-1}{x}\theta^x(1-\theta)^{n-x}. \tag{1.74}$$

Confirm that under both models, the MLE of θ is x/n.

2.7 Exercises

2.1 Obtain two sequences from GenBank, align the sequences and then apply the methods discussed in this chapter to estimate d_S and d_N and discuss their differences. One way of aligning protein-coding DNA sequences is to use CLUSTAL (Thompson *et al.* 1994) to align the protein sequences first and then construct the DNA alignment based on the protein alignment, using, for example, MEGA3.1 (Kumar *et al.* 2005*a*) or BAMBE (Xia and Xie 2001), followed by manual adjustments.

***2.2** Are there really three nucleotide sites in a codon? How many synonymous and nonsynonymous sites are in the codon TAT (use the universal code)?

2.3 Behaviour of LWL85 and related methods under the *two-fold and four-fold mixture regular code*. Imagine a genetic code in which a proportion γ of codons are four-fold degenerate while all other codons are two-fold degenerate. (If $\gamma = 48/64$, the code would encode exactly 20 amino acids.) Suppose that neutral mutations occur according to the K80 model, with transition rate α and transversion rate β, with $\alpha/\beta = \kappa$. The proportion of nonsynonymous mutations that are neutral is ω. The numbers of nondegenerate, two-fold, and four-fold degenerate sites in a codon are $L_0 = 2$, $L_2 = 1 - \gamma$, and $L_4 = \gamma$. Over time interval t, the numbers of transitional and transversional substitutions at the three degeneracy classes are thus $A_0 = \alpha t\omega$, $B_0 = 2\beta t\omega$, $A_2 = \alpha t$, $B_2 = 2\beta t\omega$, $A_4 = \alpha t$, $B_4 = 2\beta t$. (a) Show that the LWL85 method (equation 2.10) gives

$$\begin{aligned} d_S &= \frac{3(\kappa + 2\gamma)\beta t}{1 + 2\gamma}, \\ d_N &= \frac{3(\kappa + 3 - \gamma)\beta t\omega}{4 - \gamma}, \end{aligned} \tag{2.29}$$

with the ratio $d_N/d_S = \omega[(\kappa + 3 - \gamma)(1 + 2\gamma)]/[(4 - \gamma)(\kappa + 2\gamma)]$, which becomes $\omega(\kappa + 3)/(4\kappa)$ if $\gamma = 0$ (so that the code is the two-fold regular code) and ω if $\gamma = 1$ (so that the code is the four-fold regular code). (b) Show that both LPB93 (equation 2.11) and LWL85m (equation 2.12) give $d_S = (\alpha + 2\beta)t$ and $d_N = d_S\omega$. (Comment: under this model, LWL85 gives d_S^* and d_N^*, distances using the physical-site definition.)

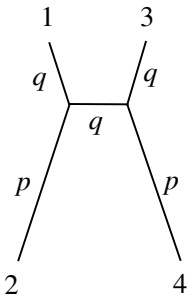


Fig. 4.13 A tree of four species with two branch lengths p and q, defined as the probability that any site is different at the two ends of the branch. For a binary character, this probability is $p = (1-e^{-2t})/2$, where t is the expected number of character changes per site (see Exercise 1.3 in Chapter 1).

4.8 Exercises

***4.1** Collapsing site patterns for likelihood calculation under the JC69 model. Under JC69, the probability of data at a site depends on whether the nucleotides are different in different species, but not on what the nucleotides are. For example, sites with data TTTC, TTTA, AAAG all have the same probability of occurrence. Show that if such sites are collapsed into patterns, there is a maximum of $(4^{s-1} + 3 \times 2^{s-1} + 2)/6$ site patterns for s sequences (Saitou and Nei 1986).

***4.2** Try to estimate the single branch length under the JC69 model for the star tree of three sequences under the molecular clock (see Saitou (1988) and Yang (1994c, 2000a), for discussions of likelihood tree reconstruction under this model). The tree is shown in Fig. 4.8, where t is the only parameter to be estimated. Note that there are only three site patterns, with one, two, or three distinct nucleotides, respectively. The data are the observed numbers of sites with such patterns: n_0, n_1, and n_2, with the sum to be n. Let the proportions be $f_i = n_i/n$. The log likelihood is $\ell = n \sum_{i=0}^{2} f_i \log(p_i)$, with p_i to be the probability of observing site pattern i. Derive p_i by using the transition probabilities under the JC69 model, given in equation (1.3). You can calculate $p_0 = \Pr(\text{TTT})$, $p_1 = \Pr(\text{TTC})$, and $p_2 = \Pr(\text{TCA})$. Then set $d\ell/dt = 0$. Show that the transformed parameter $z = e^{-4/3t}$ is a solution to the following quintic equation:

$$36z^5 + 12(6 - 3f_0 - f_1)z^4 + (45 - 54f_0 - 42f_1)z^3 + (33 - 60f_0 - 36f_1)z^2 + (3 - 30f_0 - 2f_1)z + (3 - 12f_0 - 4f_1) \equiv 0. \quad (4.30)$$

4.3 Calculate the probabilities of sites with data *xxyy*, *xyyx*, and *xyxy* in four species for the unrooted tree of Fig. 4.13, using two branch lengths p and q under a symmetrical substitution model for binary characters (Exercise 1.3). Here it is more convenient to define the branch length as the proportion of different sites at the two ends of the branch. Show that $\Pr(xxyy) < \Pr(xyxy)$ if and only if $q(1 - q) < p^2$. With such branch lengths, parsimony for tree reconstruction is inconsistent (Felsenstein 1978a).

5.8 Exercises

5.1 (a) In the example of testing for infection in Subsection 5.1.2, suppose that a person tested negative. What is the probability that he has the infection (b) Suppose a person was tested twice and found to be positive both times. What is the probability that he has the infection?

5.2 *Criticism of unbiasedness*. Both likelihood and Bayesian proponents point out that strict adherence to unbiasedness may be unreasonable. For the example of Subsection 5.1.2, consult any statistics textbook to confirm that the expectation of the sample frequency x/n is θ under the binomial model and $\theta(n-1)/n$ under the negative binominal model. Thus the unbiased estimator of θ is $x/n = 9/12$ under the binomial and $x/(n-1) = 9/11$ under the negative binomial. Unbiasedness thus violates the likelihood principle. Another criticism of unbiased estimators is that they are not invariant to reparametrization; if $\hat{\theta}$ is an unbiased estimator of θ, $h(\hat{\theta})$ will not be an unbiased estimator of $h(\theta)$ if h is not a linear function of θ.

***5.3** Suppose the target density is $N(\theta, 1)$, and the MCMC uses the sliding-window proposal with normal proposals, with the jump kernel $x^* \sim N(x, \sigma^2)$. Show that the acceptance proportion (the proportion at which the proposals are accepted) is (Gelman *et al.* 1996)

$$P_{\text{jump}} = \frac{2}{\pi} \tan^{-1}\left(\frac{2}{\sigma}\right). \tag{5.43}$$

5.4 Write a program to implement the MCMC algorithm of Subsection 5.3.2 to estimate the distance between the human and orangutan 12s rRNA genes under the JC69 model. Use any programming language of your choice, such as BASIC, Fortran, C/C++, Java, or Mathematica. Investigate how the acceptance proportion changes with the window size w. Also implement the proposal of equation (5.34). (Hint: use the logarithms of the likelihood and prior in the algorithm to avoid numerical problems.)

5.5 Modify the program above to estimate two parameters under the JC69 model: the substation rate $\mu = 3\lambda$ and the time of species divergence T, instead of the distance $\theta = 3\lambda \times 2T$. Consider one time unit as 100 million years, and assign an exponential prior $f(T) = (1/m)e^{-T/m}$ for T with mean $m = 0.15$ (15 million year for human–orangutan divergence) and another exponential prior with mean 1.0 for rate μ (corresponding to a prior mean rate of about 1 substitution per 100 million years). Use two proposal steps, one updating T and another updating μ. Change the prior to examine the sensitivity of the posterior to the prior.

5.6 Modify the program of Exercise 5.4 to estimate the sequence distance under the K80 model. Use the exponential prior $f(\theta) = (1/m)e^{-\theta/m}$ with mean $m = 0.2$ for distance θ and exponential prior with mean 5 for the transition/transversion rate ratio κ. Implement two proposal steps, one for updating θ and another for updating κ. Compare the posterior estimates with the MLEs of Subsection 1.4.2.

9.6 Exercises

9.1 Write a small simulation program to study the *birthday problem*. Suppose that there are 365 days in a year and that one's birthday falls on any day at random. Calculate the probability that at least two people out of a group of $k = 30$ people have the same birthday (that is, they were born on the same day and month but not necessarily in the same year). Use the following algorithm. (The answer is 0.706.)

1. Generate $k = 30$ birthdays, by taking 30 random draws from $1, 2, \ldots, 365$.
2. Check whether any two birthdays are the same.
3. Repeat the process 10^6 times and calculate the proportion of times in which two out of 30 people have the same birthday.

9.2 Monte Carlo integration (Subsection 5.3.1). Write a small program to calculate the integral $f(x)$ in the Bayesian estimation of sequence distance under the JC69 model, discussed in Subsection 5.1.2. The data are $x = 90$ differences out of $n = 948$ sites. Use the exponential prior with mean 0.2 for the sequence distance θ. Generate $N = 10^6$ or 10^8 random variables from the exponential prior: $\theta_1, \theta_2, \ldots, \theta_N$, and calculate

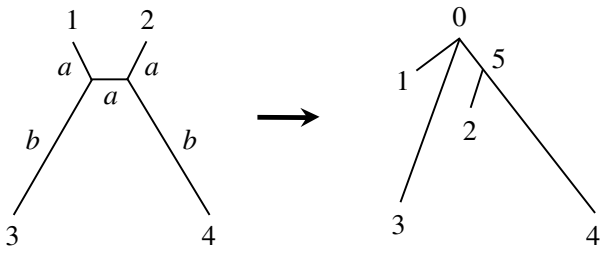
$$f(x) = \int_0^\infty f(\theta) f(x|\theta) \mathrm{d}\theta \simeq \frac{1}{N} \sum_{i=1}^{N} f(x|\theta_i). \tag{9.8}$$

Note that the likelihood $f(x|\theta_i)$ may be too small to represent in the computer, so scaling may be needed. One way is as follows. Compute the maximum log likelihood $\ell_m = \log\{f(x|\hat{\theta})\}$, where $\hat{\theta} = 0.1015$ is the MLE. Then multiply $f(x|\theta_i)$ in equation (9.8) by a big number $e^{-\ell_m}$ so that they are not all vanishingly small before summing them up; that is,

$$\sum_{i=1}^{N} f(x|\theta_i) = e^{\ell_m} \cdot \sum_{i=1}^{N} \exp\left(\log\{f(x|\theta_i)\} - \ell_m\right). \tag{9.9}$$

9.3 Write a small simulation program to study the optimal sequence divergence when two sequences are compared to estimate the transition/transversion rate ratio κ under the K80 model. Assume $\kappa = 2$ and use a sequence length of 500 sites. Consider several sequence distances, say, $d = 0.01, 0.02, \ldots, 2$. For each d, simulate 1000 replicate data sets under the K80 model and analyse it under the same model to estimate d and κ using equation (1.11). Calculate the mean and variance of the estimate $\hat{\kappa}$ across replicate data sets. Each data set consists of a pair of sequences, which can be generated using any of the three approaches discussed in Subsection 9.5.1.

9.4 Long-branch attraction by parsimony. Use the JC69 model to simulate data sets on a tree of four species (Fig. 9.3a), with two different branch lengths $a = 0.1$ and $b = 0.5$. Simulate 1000 replicate data sets. For each data set, count the sites with the three site patterns *xxyy*, *xyxy*, *xyyx*, and determine the most parsimonious tree. To simulate a data set, reroot the tree at an interior node, as in, say, Fig. 9.3(b). Generate



(a) Unrooted tree (b) Rooted tree

Fig. 9.3 (a) A tree of four species, with three short branches (of length a) and two long branches (with length b) for simulating data to demonstrate long-branch attraction. (b) The same tree rerooted at an ancestral node for simulation.

a sequence for the root (node 0) by random sampling of the four nucleotides, and then evolve the sequence along the five branches of the tree. You may also use the approach of multinomial sampling. Consider a few sequence lengths, such as 100, 1000, and 10 000 sites.

9.5 A useful test of a new and complex likelihood program is to generate a few data sets of very long sequences under the model and then analyse them under the same model, to check whether the MLEs are close to the true values used in the simulation. As MLEs are consistent, they should approach the true values when the sample size (sequence length) becomes larger and larger. Use the program written for Exercise 9.4 to generate a few data sets of 10^6, 10^7, or 10^8 sites and analyse them using a likelihood program (such as PHYLIP, PAUP, or PAML) under the same JC69 model, to see whether the MLEs of branch lengths are close to the true values. Beware that some programs may demand a lot of resources to process large data sets; save your important work before this exercise in case of a computer crash.