CHAPTER 1

Models of nucleotide substitution

1.1 Introduction

Calculation of the distance between two sequences is perhaps the simplest phylogenetic analysis, yet it is important for two reasons. First, calculation of pairwise distances is the first step in distance matrix methods of phylogeny reconstruction, which use cluster algorithms to convert a distance matrix into a phylogenetic tree. Second, Markov process models of nucleotide substitution used in distance calculation form the basis of likelihood and Bayesian methods of phylogeny reconstruction. Indeed, joint analysis of multiple sequences can be viewed as a natural extension of pairwise distance calculation. Thus, besides discussing distance estimation, this chapter introduces the theory of Markov chains used in modelling nucleotide substitutions in a DNA sequence. It also introduces the method of maximum likelihood (ML). Bayesian estimation of pairwise distances and Bayesian phylogenetics are introduced in Chapters 6–8.

The distance between two sequences is defined as the expected number of nucleotide substitutions per site. If the evolutionary rate is constant over time, the distance will increase linearly with the time of divergence. A simplistic distance measure is the proportion of different sites, sometimes called the p distance. If 10 sites are different between two sequences, each 100 nucleotides long, then $p = 10\% = 0.1$. This raw proportion works fine for very closely related sequences but is otherwise a clear underestimate of the number of substitutions that have occurred. A variable site may result from more than one substitution, and even a constant site, with the same nucleotide observed in the two sequences, may harbour back or parallel substitutions (Figure 1.1). Multiple substitutions at the same site or *multiple hits* cause some changes to be hidden. As a result, p is not a linear function of evolutionary time. Thus the raw proportion p is usable only for highly similar sequences, with $p < 5\%$, say.

To estimate the number of substitutions, we need a probabilistic model to describe changes between nucleotides over evolutionary time. Continuous-time Markov chains are commonly used for this purpose. The nucleotide sites in the sequence are assumed to be evolving independently of each other. Substitutions at any particular site are described by a Markov chain, with the four nucleotides to be the *states* of the chain. The main feature of a Markov chain is that it has no memory: 'given the present, the future does not depend on the past'. In other words, the probability with which the chain jumps into other nucleotide states depends on the current state, but not on how the current state is reached. This is known as the *Markovian property*. Besides this basic assumption, we often place further constraints on substitution rates between nucleotides, leading to

Molecular Evolution: A Statistical Approach. Ziheng Yang. © Ziheng Yang 2014.
Published 2014 by Oxford University Press.

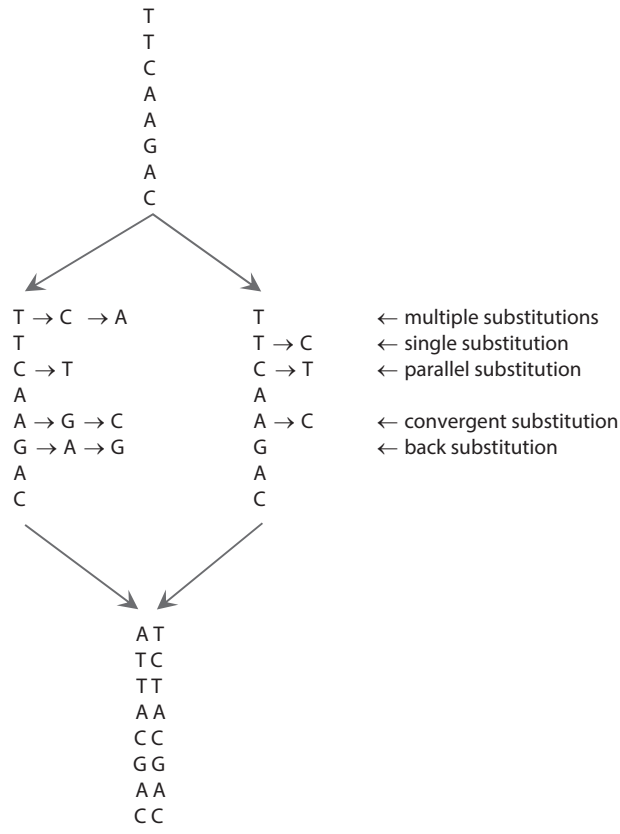

Fig. 1.1 Illustration of multiple substitutions at the same site or multiple hits. An ancestral sequence has diverged into two sequences and has since accumulated nucleotide substitutions independently along the two lineages. Only two *differences* are observed between the two present-day sequences, so that the proportion of different sites is $\hat{p} = 2/8 = 0.25$, while in fact as many as 10 *substitutions* (seven on the left lineage and three on the right lineage) occurred so that the true distance is $10/8 = 1.25$ substitutions per site. Constructed following Graur and Li (2000).

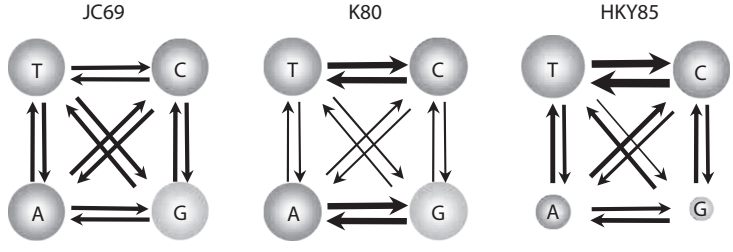

Fig. 1.2 Relative substitution rates between nucleotides under three Markov chain models of nucleotide substitution: JC69, K80, and HKY85. The thickness of the lines represents the substitution rates, while the sizes of the circles represent the steady-state distribution.

different models of nucleotide substitution. A few commonly used models are summarized in Table 1.1 and illustrated in Figure 1.2. These are discussed below.

Table 1.1 Substitution rate matrices for commonly used Markov models of nucleotide substitution

			To			
	p	From	T	C	A	G
JC69 (Jukes and Cantor 1969)	1	T	·	λ	λ	λ
		C	λ	·	λ	λ
		A	λ	λ	·	λ
		G	λ	λ	λ	·
K80 (Kimura 1980)	2	T	·	α	β	β
		C	α	·	β	β
		A	β	β	·	α
		G	β	β	α	·
F81 (Felsenstein 1981)	4	T	·	π_C	π_A	π_G
		C	π_T	·	π_A	π_G
		A	π_T	π_C	·	π_G
		G	π_T	π_C	π_A	·
HKY85 (Hasegawa et al. 1984, 1985)	5	T	·	$\alpha\pi_C$	$\beta\pi_A$	$\beta\pi_G$
		C	$\alpha\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
		A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha\pi_G$
		G	$\beta\pi_T$	$\beta\pi_C$	$\alpha\pi_A$	·
F84 (Felsenstein, DNAML program since 1984)	5	T	·	$(1 + \kappa/\pi_Y)\beta\pi_C$	$\beta\pi_A$	$\beta\pi_G$
		C	$(1 + \kappa/\pi_Y)\beta\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
		A	$\beta\pi_T$	$\beta\pi_T$	·	$(1 + \kappa/\pi_R)\beta\pi_G$
		G	$\beta\pi_T$	$\beta\pi_C$	$(1 + \kappa/\pi_R)\beta\pi_A$	·
TN93 (Tamura and Nei 1993)	6	T	·	$\alpha_1\pi_C$	$\beta\pi_A$	$\beta\pi_G$
		C	$\alpha_1\pi_T$	·	$\beta\pi_A$	$\beta\pi_G$
		A	$\beta\pi_T$	$\beta\pi_C$	·	$\alpha_2\pi_G$
		G	$\beta\pi_T$	$\beta\pi_C$	$\alpha_2\pi_A$	·
GTR (REV) (Tavaré 1986; Yang 1994b; Zharkikh 1994)	9	T	·	$a\pi_C$	$b\pi_A$	$c\pi_G$
		C	$a\pi_T$	·	$d\pi_A$	$e\pi_G$
		A	$b\pi_T$	$d\pi_C$	·	$f\pi_G$
		G	$c\pi_T$	$e\pi_C$	$f\pi_A$	·
UNREST (Yang 1994b)	12	T	·	a	b	c
		C	d	·	e	f
		A	g	h	·	i
		G	j	k	l	·

Note: The diagonals of the matrix are determined by the requirement that each row sums to 0. p is the number of free parameters in the model. If only relative rates are considered (as in a typical likelihood analysis), the number should be reduced by 1. In F84, $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$. The equilibrium distribution is $\pi = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ under JC69 and K80, and $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$ under F81, F84, HKY85, TN93, and GTR. Under the general unrestricted (UNREST) model, it is given by equation (1.61).

1.2 Markov models of nucleotide substitution and distance estimation

1.2.1 *The JC69 model*

The JC69 model (Jukes and Cantor 1969) assumes that every nucleotide has the same instantaneous rate λ of changing into every other nucleotide. We use q_{ij} to denote the substitution rate from nucleotides i to j, with i, j = T, C, A, or G. Thus the *substitution rate matrix* is

$$Q = \{q_{ij}\} = \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix}, \tag{1.1}$$

where the nucleotides are ordered T, C, A, and G. The diagonals are determined by the mathematical requirement that each row of the matrix sums to 0. The total rate of substitution of any nucleotide i is 3λ, which is $-q_{ii}$.

To relate the Markov chain model to sequence data, we need calculate the probability that given the nucleotide i at a site now, it will become nucleotide j time t later. This is known as the *transition probability*, denoted $p_{ij}(t)$. If time t is very small, we have $p_{ij}(t) \approx q_{ij}t$ for $i \neq j$, and $p_{ii}(t) \approx 1 - t\sum_{j\neq i} q_{ij}$. In other words, the *matrix of transition probabilities* is

$$P(t) = \{p_{ij}(t)\} \approx I + Qt = \begin{bmatrix} 1-3\lambda t & \lambda t & \lambda t & \lambda t \\ \lambda t & 1-3\lambda t & \lambda t & \lambda t \\ \lambda t & \lambda t & 1-3\lambda t & \lambda t \\ \lambda t & \lambda t & \lambda t & 1-3\lambda t \end{bmatrix}, \text{ for small } t. \tag{1.2}$$

Suppose a random region of the human genome evolves according to the JC69 model, at the rate of $3\lambda = 2.2 \times 10^{-9}$ substitutions/site/year (Kumar and Subramanian 2002) (Table 1.2). Consider a site occupied by a T right now. The probability that $t = 10^6$ years later this site will have a C will be $\lambda t = 0.00073$, and the probability that it remains to be T will be $1 - 3\lambda t = 0.9978$.

Equation (1.2) does not work well if t is not small. In general,

$$P(t) = e^{Qt} = I + Qt + \frac{1}{2!}(Qt)^2 + \frac{1}{3!}(Qt)^3 + \cdots. \tag{1.3}$$

We will discuss the calculation of this matrix exponential later. For the moment, we simply give the solution for the JC69 model as

$$P(t) = e^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix}, \text{ with } \begin{cases} p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}, \\ p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}. \end{cases} \tag{1.4}$$

Imagine a long sequence with nucleotide i at every site, and let every site evolve for a time period t. Then the proportion of nucleotide j in the sequence will be $p_{ij}(t)$, for j = T, C, A, G. The two different elements of the transition probability matrix, $p_0(t)$ and $p_1(t)$, are plotted in Figure 1.3. A few features of the matrix $P(t)$ are worth noting. First, every row of $P(t)$ sums to 1, because at any time t the chain has to be in one of the four nucleotide states. Second, $P(0) = I$, the identity matrix, reflecting the case of no evolution ($t = 0$). Third, rate λ and time t occur in the transition probabilities only in the form of the product λt. Thus if we are given a source sequence and a target sequence,

Table 1.2 A sample of estimated mutation/substitution rates

Taxa	Genes/genomes	Mutation/substitution rate	Source
Placental mammals	Genomic mutation rate at four-fold degenerate sites	2.2×10^{-9} per site per year	Kumar & Subramanian (2002)
Primates	12 protein-coding genes in the mitochondrial genome	7.9×10^{-9} per site per year for all codon positions, or $2.2, 0.1, 4.2 \times 10^{-9}$ per site per year for positions 1, 2, and 3, respectively.	Yang & Yoder (2003)
Human	Family-based genome sequencing	$1.1–1.2 \times 10^{-8}$ per site per generation	Roach et al. (2010), Kong et al. (2012)
Plants (rice and maize)	Nuclear genome	6×10^{-9}/site/year for synonymous 9×10^{-11}/site/year for nonsynonymous	Gaut (1998)
Plants (rice and maize)	Mitochondrial genome	0.3×10^{-9}/site/year for synonymous 1.3×10^{-11}/site/year for nonsynonymous	Gaut (1998)
Plants (rice and maize)	Chloraplast genome	1.1×10^{-9}/site/year for synonymous 1.8×10^{-11}/site/year for nonsynonymous	Gaut (1998)
HIV virus	HIV-1 *env* V3 region	$2–17 \times 10^{-3}$/site/year	Berry et al. (2007)

it will be impossible to tell whether the source has evolved into the target at rate λ over time t or at rate 2λ over time $t/2$. In fact, the sequences will look the same for any combination of λ and t as long as λt is fixed. With no external information about either the time or the rate, we can estimate only the distance, but not time and rate individually.

Lastly, when $t \to \infty$, $p_{ij}(t) = 1/4$, for all i and j. This represents the case where so many substitutions have occurred at every site that the target nucleotide is random, with probability

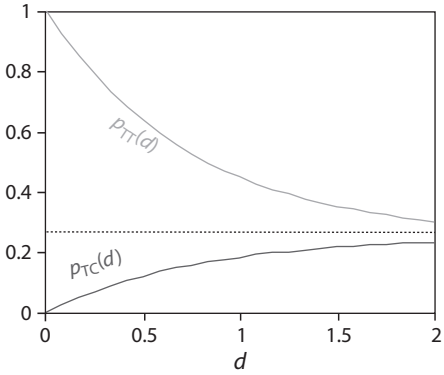

Fig. 1.3 Transition probabilities under the JC69 model (equation (1.4)) plotted against distance $d = 3\lambda t$, measured in the expected number of substitutions per site.

1/4 for every nucleotide, irrespective of the starting nucleotide. The probability that the chain is in state j when $t \to \infty$ is represented by π_j and the distribution $(\pi_T, \pi_C, \pi_A, \pi_G)$ is known as the *limiting distribution* of the chain. For the JC69 model, $\pi_j = 1/4$ for every nucleotide j. If the states of the chain are already in the limiting distribution, the chain will stay in that distribution, so the limiting distribution is also the *steady-state distribution* or *stationary distribution*. In other words, if a long sequence starts with T at every site, the proportions of the four nucleotides T, C, A, and G will drift away from (1, 0, 0, 0) and approach (1/4, 1/4, 1/4, 1/4), as the sequence evolves. If the sequence starts with equal proportions of the four nucleotides, it will continue to have equal proportions of the four nucleotides as the sequence evolves. The Markov chain is said to be stationary, or nucleotide substitutions are said to be in equilibrium. This is an assumption made in almost all models used in phylogenetic analysis, and is violated if the sequences in the data have different base compositions.

How does the Markov chain model correct for multiple hits and recover the hidden changes illustrated in Figure 1.1? This is achieved through the calculation of the transition probabilities using equation (1.3), which accommodates all possible paths the evolutionary process might have taken. In particular, the transition probabilities for a Markov chain satisfy the following equation, known as the Chapman–Kolmogorov equation (e.g. Grimmett and Stirzaker 1992, p. 239):

$$p_{ij}(t_1 + t_2) = \sum_k p_{ik}(t_1)\, p_{kj}(t_2). \tag{1.5}$$

This is a direct application of the *law of total probability*: the probability that nucleotide i will become nucleotide j time $t_1 + t_2$ later is a sum over all possible states k at any intermediate time point t_1 (Figure 1.4).

We now consider estimation of the distance between two sequences. From equation (1.1), the total substitution rate for any nucleotide is 3λ. If the two sequences are separated by time t (for example, if they diverged from a common ancestor time $t/2$ ago), the distance between the two sequences will be $d = 3\lambda t$. Suppose x out of n sites are different between the two sequences, so that the proportion of different sites is $\hat{p} = x/n$. (The hat or caret is used to indicate that the proportion is an estimate from the data.) To derive the expected probability p of different sites, consider one sequence as the ancestor of the other. By the symmetry of the model (equation (1.4)), this is equivalent to considering the two sequences as descendants of an extinct common ancestor. From equation (1.4), the probability that the nucleotide in the descendant sequence is different from the nucleotide in the ancestral sequence is

$$p(d) = 3p_1(t) = \frac{3}{4} - \frac{3}{4}e^{-4\lambda t} = \frac{3}{4} - \frac{3}{4}e^{-4d/3}. \tag{1.6}$$

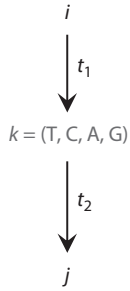

Fig. 1.4 Illustration of the Chapman–Kolmogorov theorem. The transition probability from any nucleotide i to any nucleotide j over time $t_1 + t_2$ is a sum over all possible states k at any intermediate time point t_1.

By equating this to the observed proportion $\hat{p}$, we obtain an estimate of distance as

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\hat{p}\right), \tag{1.7}$$

where the logarithm has base e (sometimes written as ln instead of log). If $\hat{p} \geq \frac{3}{4}$, the distance formula will be inapplicable; two random sequences should have about 75% different sites, and if $\hat{p} \geq \frac{3}{4}$, the distance estimate is infinite. To derive the variance of $\hat{d}$, note that $\hat{p}$ is a binomial proportion with variance $\hat{p}(1-\hat{p})/n$. Considering $\hat{d}$ as a function of $\hat{p}$ and using the so-called delta technique (see Appendix B), we obtain

$$\mathrm{var}(\hat{d}) = \mathrm{var}(\hat{p}) \times \left|\frac{\mathrm{d}\hat{d}}{\mathrm{d}\hat{p}}\right|^2 = \hat{p}(1-\hat{p})/n \times \frac{1}{\left(1 - 4\hat{p}/3\right)^2} \tag{1.8}$$

(Kimura and Ohta 1972).

Example 1.1. The observed sequences of human and orangutan 12s rRNA genes from the mitochondrial genome are summarized in Table 1.3. From the table, $x = 90$ out of the $n = 948$ sites are different, so that $\hat{p} = x/n = 0.09494$. By equation (1.7), $\hat{d} = 0.1015$. Equation (1.8) gives the variance of $\hat{d}$ as 0.0001188 and standard error 0.0109. The approximate 95% confidence interval is thus $\hat{d} \pm 1.96 \times \mathrm{SE} = 0.1015 \pm 1.96 \times 0.0109$ or $(0.0801, 0.1229)$. □

1.2.2 *The K80 model*

Substitutions between the two pyrimidines (T ↔ C) or between the two purines (A ↔ G) are called *transitions*, while those between a pyrimidine and a purine (T, C ↔ A, G) are called *transversions*. In real data, transitions often occur at higher rates than transversions. Thus Kimura (1980) proposed a model that accounts for different transition and transversion rates. Note that the biologist's use of the term transition (as opposed to transversion) has nothing to do with the probabilist's use of the same term (as in transition probability). Typically the usage is clear from the context and there is little risk of confusion.

Let the substitution rates be α for transitions and β for transversions. The model is referred to as K80, also known as Kimura's two-parameter model. The rate matrix is as follows (see also Figure 1.2):

Table 1.3 Numbers and frequencies (in parentheses) of sites for the 16 site configurations (patterns) in human and orangutan mitochondrial 12s rRNA genes

	Human				
Orang	T	C	A	G	Sum (π_i)
T	179 (0.188819)	23 (0.024262)	1 (0.001055)	0 (0)	0.2141
C	30 (0.031646)	219 (0.231013)	2 (0.002110)	0 (0)	0.2648
A	2 (0.002110)	1 (0.001055)	291 (0.306962)	10 (0.010549)	0.3207
G	0 (0)	0 (0)	21 (0.022152)	169 (0.178270)	0.2004
Sum (π_j)	0.2226	0.2563	0.3323	0.1888	1

Note: Genbank accession numbers for the human and orangutan sequences are D38112 and NC_001646, respectively (Horai et al. 1995). There are 954 sites in the alignment, but six sites involve alignment gaps and are removed, leaving 948 sites in each sequence. The average base frequencies in the two sequences are 0.2184 (T), 0.2605 (C), 0.3265 (A), and 0.1946 (G).

$$Q = \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix}. \tag{1.9}$$

The total substitution rate for any nucleotide is $\alpha + 2\beta$, and the distance between two sequences separated by time t is $d = (\alpha + 2\beta)t$. Note that αt is the expected number of transitions per site and $2\beta t$ is the expected number of transversions per site. One can use αt and βt as the two parameters in the model, but it is often more convenient to use the distance d and the transition/transversion rate ratio $\kappa = \alpha/\beta$. The matrix of transition probabilities is given as

$$P(t) = e^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_2(t) & p_2(t) \\ p_1(t) & p_0(t) & p_2(t) & p_2(t) \\ p_2(t) & p_2(t) & p_0(t) & p_1(t) \\ p_2(t) & p_2(t) & p_1(t) & p_0(t) \end{bmatrix}, \tag{1.10}$$

where the three distinct elements of the matrix are

$$\begin{aligned} p_0(t) &= \tfrac{1}{4} + \tfrac{1}{4}e^{-4\beta t} + \tfrac{1}{2}e^{-2(\alpha+\beta)t} = \tfrac{1}{4} + \tfrac{1}{4}e^{-4d/(\kappa+2)} + \tfrac{1}{2}e^{-2d(\kappa+1)/(\kappa+2)}, \\ p_1(t) &= \tfrac{1}{4} + \tfrac{1}{4}e^{-4\beta t} - \tfrac{1}{2}e^{-2(\alpha+\beta)t} = \tfrac{1}{4} + \tfrac{1}{4}e^{-4d/(\kappa+2)} - \tfrac{1}{2}e^{-2d(\kappa+1)/(\kappa+2)}, \\ p_2(t) &= \tfrac{1}{4} - \tfrac{1}{4}e^{-4\beta t} = \tfrac{1}{4} - \tfrac{1}{4}e^{-4d/(\kappa+2)} \end{aligned} \tag{1.11}$$

(Kimura 1980; Li 1986). Note that $p_0(t) + p_1(t) + 2p_2(t) = 1$.

The sequence data can be summarized as the proportions of sites with transitional and transversional differences. Let these be S and V, respectively. Again, by the symmetry of the model (equation (1.10)), the probability that a site is occupied by nucleotides with a transitional difference is $E(S) = p_1(t)$. Similarly $E(V) = 2p_2(t)$. Equating these to the observed proportions S and V leads to two simultaneous equations in two unknowns, which are easily solved to give

$$\begin{aligned} \hat{d} &= -\tfrac{1}{2}\log(1-2S-V) - \tfrac{1}{4}\log(1-2V), \\ \hat{\kappa} &= \frac{2\times\log(1-2S-V)}{\log(1-2V)} - 1 \end{aligned} \tag{1.12}$$

(Kimura 1980; Jukes 1987). Equivalently the transition distance αt and the transversion distance $2\beta t$ are estimated as

$$\begin{aligned} \widehat{\alpha t} &= -\tfrac{1}{2}\log(1-2S-V) + \tfrac{1}{4}\log(1-2V), \\ \widehat{2\beta t} &= -\tfrac{1}{2}\log(1-2V), \end{aligned} \tag{1.13}$$

The distance formula is applicable only if $1-2S-V > 0$ and $1-2V > 0$. As S and V are multinomial proportions with $\mathrm{var}(S) = S(1-S)/n$, $\mathrm{var}(V) = V(1-V)/n$, and $\mathrm{cov}(S, V) = -SV/n$, we can use the delta technique to derive the variance–covariance matrix of $\hat{d}$ and $\hat{\kappa}$ (see Appendix B). In particular, the variance of $\hat{d}$ is

$$\mathrm{var}(\hat{d}) = [a^2S + b^2V - (aS + bV)^2]/n, \tag{1.14}$$

where

$$\begin{aligned} a &= (1-2S-V)^{-1}, \\ b &= \tfrac{1}{2}[(1-2S-V)^{-1} + (1-2V)^{-1}]. \end{aligned} \tag{1.15}$$

Example 1.2. For the 12s rRNA data of Table 1.3, the proportions of transitional and transversional differences are $S = (23 + 30 + 10 + 21)/948 = 0.08861$ and $V = (1 + 0 + 2 + 0 + 2 + 1 + 0 + 0)/948 = 0.00633$. Thus equations (1.12) and (1.14) give the distance and standard error as 0.1046 ± 0.0116 (Table 1.4). The estimate $\hat{\kappa} = 30.836$ indicates that the transition rate is ~30 times higher than the transversion rate. □

1.2.3 *HKY85, F84, TN93, etc.*

1.2.3.1 *TN93*

The JC69 and K80 models have symmetrical substitution rates, with $q_{ij} = q_{ji}$ for all $i \neq j$. Such a Markov chain has $\pi_i = \frac{1}{4}$ for all i as the stationary distribution; that is, when the substitution process reaches equilibrium, the sequence will have equal proportions of the four nucleotides. This assumption is unrealistic for most datasets. Here we consider a few models that accommodate unequal base compositions. The model of Tamura and Nei (1993), referred to as TN93, has most of the commonly used models as special cases. Thus we present detailed results for this model, which also apply to its special cases. The substitution rate matrix under the TN93 model is

$$Q = \begin{bmatrix} -(\alpha_1\pi_C + \beta\pi_R) & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & -(\alpha_1\pi_T + \beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha_2\pi_G + \beta\pi_Y) & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & -(\alpha_2\pi_A + \beta\pi_Y) \end{bmatrix}. \tag{1.16}$$

Table 1.4 Estimates of distance between the human and orangutan 12s rRNA genes

Model and method	$\hat{d}$	Estimates of other parameters	ℓ
Distance formulae			
JC69	0.1015 ± 0.0109		
K80	0.1046 ± 0.0116	$\hat{\kappa} = 30.83 \pm 13.12$	
F81	0.1016		
F84	0.1050	$\hat{\kappa} = 15.548$	
TN93	0.1078	$\hat{\kappa}_1 = 44.228$, $\hat{\kappa}_2 = 21.789$	
Maximum likelihood			
JC69	0.1015 ± 0.0109		−1710.58
K80	0.1046 ± 0.0116	$\hat{\kappa} = 30.83 \pm 13.12$	−1637.90
F81	0.1017 ± 0.0109	$\hat{\pi} = (0.2251, 0.2648, 0.3188, 0.1913)$	−1691.97
F84	0.1048 ± 0.0117	$\hat{\kappa} = 15.640$, $\hat{\pi} = (0.2191, 0.2602, 0.3286, 0.1921)$	−1616.60
HKY85	0.1048 ± 0.0117	$\hat{\kappa} = 32.137$, $\hat{\pi} = (0.2248, 0.2668, 0.3209, 0.1875)$	−1617.27
TN93	0.1048 ± 0.0117	$\hat{\kappa}_1 = 44.229$, $\hat{\kappa}_2 = 21.781$ $\hat{\pi} = (0.2185, 0.2604, 0.3275, 0.1936)$	−1613.03
GTR (REV)	0.1057 ± 0.0119	$\hat{a} = 2.0431$, $\hat{b} = 0.0821$, $\hat{c} = 0.0000$, $\hat{d} = 0.0670$, $\hat{e} = 0.0000$, $\hat{\pi} = (0.2184, 0.2606, 0.3265, 0.1946)$	−1610.36
UNREST	0.1057 ± 0.0120	See equation (1.66) for the estimated Q; $\hat{\pi} = (0.2184, 0.2606, 0.3265, 0.1946)$	−1610.36

Note: ℓ is the log likelihood under the model.

While parameters $\pi_T, \pi_C, \pi_A, \pi_G$ are used to specify the substitution rates, they also give the stationary (equilibrium) distribution, with $\pi_Y = \pi_T + \pi_C$ and $\pi_R = \pi_A + \pi_G$ to be the frequencies of pyrimidines and purines, respectively.

The matrix of transition probabilities over time t is $P(t) = \{p_{ij}(t)\} = e^{Qt}$. A standard approach to calculating an algebraic function, such as the exponential, of a matrix Q, is to *diagonalize* Q (e.g. Schott 1997, Chapter 3). Suppose Q can be written in the form

$$Q = U\Lambda U^{-1}, \tag{1.17}$$

where U is a nonsingular matrix and U^{-1} is its inverse, and Λ is a diagonal matrix $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \lambda_3, \lambda_4\}$. The λs are the eigenvalues (or latent roots) of Q, and columns of U and rows of U^{-1} are the corresponding right and left eigenvectors of Q, respectively. Equation (1.17) is also known as the *spectral decomposition* of Q. The reader should consult a textbook on linear algebra for calculation of eigenvalues and eigenvectors of a matrix (e.g. Schott 1997, Chapter 3).

From equation (1.17), we have $Q^2 = (U\Lambda U^{-1})(U\Lambda U^{-1}) = U\Lambda^2 U^{-1} = U\,\text{diag}\{\lambda_1^2, \lambda_2^2, \lambda_3^2, \lambda_4^2\}\,U^{-1}$. Similarly $Q^m = U\,\text{diag}\{\lambda_1^m, \lambda_2^m, \lambda_3^m, \lambda_4^m\}\,U^{-1}$ for any integer m. In general, any algebraic function h of matrix Q can be calculated as $h(Q) = U\,\text{diag}\{h(\lambda_1), h(\lambda_2), h(\lambda_3), h(\lambda_4)\}\,U^{-1}$ as long as $h(Q)$ exists. Thus

$$P(t) = e^{Qt} = U\,\text{diag}\left\{e^{\lambda_1 t}, e^{\lambda_2 t}, e^{\lambda_3 t}, e^{\lambda_4 t}\right\}U^{-1}. \tag{1.18}$$

For the TN93 model, the spectral decomposition of Q is analytical. We have $\lambda_1 = 0$, $\lambda_2 = -\beta, \lambda_3 = -(\pi_R\alpha_2 + \pi_Y\beta)$, and $\lambda_4 = -(\pi_Y\alpha_1 + \pi_R\beta)$, and

$$U = \begin{bmatrix} 1 & 1/\pi_Y & 0 & \pi_C/\pi_Y \\ 1 & 1/\pi_Y & 0 & -\pi_T/\pi_Y \\ 1 & -1/\pi_R & \pi_G/\pi_R & 0 \\ 1 & -1/\pi_R & -\pi_A/\pi_R & 0 \end{bmatrix}, \tag{1.19}$$

$$U^{-1} = \begin{bmatrix} \pi_T & \pi_C & \pi_A & \pi_G \\ \pi_T\pi_R & \pi_C\pi_R & -\pi_A\pi_Y & -\pi_G\pi_Y \\ 0 & 0 & 1 & -1 \\ 1 & -1 & 0 & 0 \end{bmatrix}. \tag{1.20}$$

Substituting Λ, U, and U^{-1} into equation (1.18) gives

$$P(t) = \begin{bmatrix} \pi_T + \frac{\pi_T\pi_R}{\pi_Y}e_2 + \frac{\pi_C}{\pi_Y}e_4 & \pi_C + \frac{\pi_C\pi_R}{\pi_Y}e_2 - \frac{\pi_C}{\pi_Y}e_4 & \pi_A(1-e_2) & \pi_G(1-e_2) \\ \pi_T + \frac{\pi_T\pi_R}{\pi_Y}e_2 - \frac{\pi_T}{\pi_Y}e_4 & \pi_C + \frac{\pi_C\pi_R}{\pi_Y}e_2 + \frac{\pi_T}{\pi_Y}e_4 & \pi_A(1-e_2) & \pi_G(1-e_2) \\ \pi_T(1-e_2) & \pi_C(1-e_2) & \pi_A + \frac{\pi_A\pi_Y}{\pi_R}e_2 + \frac{\pi_G}{\pi_R}e_3 & \pi_G + \frac{\pi_G\pi_Y}{\pi_R}e_2 - \frac{\pi_G}{\pi_R}e_3 \\ \pi_T(1-e_2) & \pi_C(1-e_2) & \pi_A + \frac{\pi_A\pi_Y}{\pi_R}e_2 - \frac{\pi_A}{\pi_R}e_3 & \pi_G + \frac{\pi_G\pi_Y}{\pi_R}e_2 + \frac{\pi_A}{\pi_R}e_3 \end{bmatrix}, \tag{1.21}$$

where $e_2 = \exp(\lambda_2 t) = \exp(-\beta t), e_3 = \exp(\lambda_3 t) = \exp\{-(\pi_R\alpha_2 + \pi_Y\beta)t\}, e_4 = \exp(\lambda_4 t) = \exp\{-(\pi_Y\alpha_1 + \pi_R\beta)t\}$.

When t increases from 0 to ∞, the diagonal element $p_{jj}(t)$ decreases from 1 to π_j, while the off-diagonal element $p_{ij}(t)$ increases from 0 to π_j, with $p_{ij}(\infty) = \pi_j$, irrespective of the starting nucleotide i. The limiting distribution $(\pi_T, \pi_C, \pi_A, \pi_G)$ is also the stationary distribution. Also the rate of convergence to the stationary distribution, that is, the rate at which $p_{ij}(t) - \pi_j$ approaches zero, is determined by the largest nonzero eigenvalue.

We now consider estimation of the sequence distance under the model. First, we will look at the definition of distance. The substitution rate of nucleotide i is $-q_{ii} = \sum_{j\neq i} q_{ij}$, and

differs among the four nucleotides. When the substitution process is in equilibrium, the amount of time the Markov chain spends in the four states T, C, A, and G is proportional to the equilibrium frequencies π_T, π_C, π_A and π_G, respectively. Similarly, if we consider a long DNA sequence in substitution equilibrium, the proportions of sites occupied by nucleotides T, C, A, and G are π_T, π_C, π_A and π_G, respectively. Thus the average substitution rate, either defined as an average over a long time for one site or as an average over many sites in a long sequence at one time point, is

$$\lambda = -\sum_i \pi_i q_{ii} = 2\pi_T\pi_C\alpha_1 + 2\pi_A\pi_G\alpha_2 + 2\pi_Y\pi_R\beta. \tag{1.22}$$

The distance between two sequences separated by time t is $d = \lambda t$.

To derive a distance estimate, we use the same strategy as for the K80 model discussed above. We call the nucleotides across sequences at a site as a *site configuration* or *site pattern*. Our strategy is to equate the observed proportions of sites with certain site patterns to their expected probabilities. Let S_1 be the proportion of sites occupied by two different pyrimidines (i.e. sites with patterns TC or CT), S_2 the proportion of sites with two different purines (i.e. sites with patterns AG or GA), and V the proportion of sites with a transversional difference.

Next, we need to derive the expected probabilities for those sites: $E(S_1), E(S_2)$, and $E(V)$. We cannot use the symmetry argument as for JC69 and K80 since Q is not symmetrical. However, Q satisfies the following condition:

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ for all } i \neq j. \tag{1.23}$$

Equivalently, $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$, for all t and for all $i \neq j$. Markov chains satisfying such conditions are said to be *time-reversible*. Reversibility means that the process will look the same whether time runs forward or backward; that is, whether we view the substitution process from the present into the future or from the present back into the past. As a result, given two sequences, the probability of data at a site is the same whether one sequence is ancestral to the other or both are descendants of an ancestral sequence. Equivalently, equation (1.23) means that the expected amount of change from i to j is equal to the expected amount of change in the opposite direction. Note that the *rates* of change may be different in the two directions: $q_{ij} \neq q_{ji}$. Now consider sequence 1 to be the ancestor of sequence 2, separated by time t. Then

$$E(S_1) = \pi_T p_{TC}(t) + \pi_C p_{CT}(t) = 2\pi_T p_{TC}(t). \tag{1.24}$$

The first term in the sum, $\pi_T p_{TC}(t)$, is the probability that a site has nucleotide T in sequence 1 and C in sequence 2. This equals the probability of having T in sequence 1, given by π_T, times the transition probability $p_{\mathrm{TC}}(t)$ that T will become C in sequence 2 time t later. Thus $\pi_T p_{TC}(t)$, is the probability of observing site pattern TC. The second term in the sum, $\pi_C p_{CT}(t)$, is the probability for site pattern CT. Similarly $E(S_2) = 2\pi_A p_{AG}(t)$ and $E(V) = 2\pi_T p_{TA}(t) + 2\pi_T p_{TG}(t) + 2\pi_C p_{CA}(t) + 2\pi_C p_{CG}(t)$. Equating the observed proportions S_1, S_2, and V to their expected probabilities leads to three simultaneous equations in three unknowns: e_2, e_3, and e_4 in the transition probability matrix (1.21) or equivalently, $d, \kappa_1 = \alpha_1/\beta$, and $\kappa_2 = \alpha_2/\beta$. Note that the nucleotide frequency parameters π_T, π_C, π_A, and π_G can be estimated using the average observed frequencies. Solving the system of equations gives the following estimates:

$$\begin{aligned}
\hat{d} &= \tfrac{2\pi_T\pi_C}{\pi_Y}(a_1 - \pi_R b) + \tfrac{2\pi_A\pi_G}{\pi_R}(a_2 - \pi_Y b) + 2\pi_Y\pi_R b, \\
\hat{\kappa}_1 &= \tfrac{a_1 - \pi_R b}{\pi_Y b}, \\
\hat{\kappa}_2 &= \tfrac{a_2 - \pi_Y b}{\pi_R b},
\end{aligned} \tag{1.25}$$

where

$$a_1 = -\log\left(1 - \frac{\pi_Y S_1}{2\pi_T\pi_C} - \frac{V}{2\pi_Y}\right),$$
$$a_2 = -\log\left(1 - \frac{\pi_R S_2}{2\pi_A\pi_G} - \frac{V}{2\pi_R}\right), \tag{1.26}$$
$$b = -\log\left(1 - \frac{V}{2\pi_Y\pi_R}\right)$$

(Tamura and Nei 1993).

The formulae are inapplicable whenever π_Y or π_R is 0 or any of the arguments to the logarithm functions are ≤ 0, as may happen when the sequences are divergent. The variance of the estimated distance $\hat{d}$ can be obtained by using the delta technique, ignoring errors in the estimates of nucleotide frequencies and noting that S_1, S_2, and V are multinomial proportions. This is similar to the calculation under the K80 model (equation (1.14)); see Tamura and Nei (1993).

Example 1.3. For the 12s rRNA data of Table 1.3, we have the observed proportions $S_1 = (23 + 30)/948 = 0.05591$, $S_2 = (10 + 21)/948 = 0.03270$, and $V = 6/948 = 0.00633$. Equation (1.25) gives the estimates as $\hat{d} = 0.1078$, $\hat{\kappa}_1 = 44.228$, and $\hat{\kappa}_2 = 21.789$. □

1.2.3.2 *HKY85 and F84 models*

Two models that are commonly used in likelihood and Bayesian phylogenetics are special cases of TN93. The first is due to Hasegawa and colleagues (Hasegawa et al. 1984, 1985). This is now commonly known as HKY85, instead of HYK84, apparently due to my misnaming (Yang 1994b). The model is obtained by setting $\alpha_1 = \alpha_2 = \alpha$ or $\kappa_1 = \kappa_2 = \kappa$ in the TN93 model (Table 1.1). The transition probability matrix is given by equation (1.21), with α_1 and α_2 replaced by α. It is not straightforward to derive a distance formula under this model (Yang 1994b), although Rzhetsky and Nei (1994) suggested a few possibilities.

The second special case of the TN93 model was implemented by Joseph Felsenstein in his DNAML program since Version 2.6 (1984) of the PHYLIP package. This is now known as the F84 model. The rate matrix was first published by Hasegawa and Kishino (1989) and Kishino and Hasegawa (1989). It is obtained by setting $\alpha_1 = (1 + \kappa/\pi_Y)\beta$ and $\alpha_2 = (1 + \kappa/\pi_R)\beta$ in the TN93 model, requiring one fewer parameter (Table 1.1). Under this model, the eigenvalues of the Q matrix become $\lambda_1 = 0$, $\lambda_2 = -\beta$, $\lambda_3 = \lambda_4 = -(1 + \kappa)\beta$. There are only three distinct eigenvalues, as for the K80 model, and thus it is possible to derive a distance formula.

From equation (1.22), the sequence distance is $d = \lambda t = 2(\pi_T\pi_C + \pi_A\pi_G + \pi_Y\pi_R)\beta t + 2(\pi_T\pi_C/\pi_Y + \pi_A\pi_G/\pi_R)\kappa\beta t$. The expected probabilities of sites with transitional and transversional differences are

$$E(S) = 2(\pi_T\pi_C + \pi_A\pi_G) + 2\left(\frac{\pi_T\pi_C\pi_R}{\pi_Y} + \frac{\pi_A\pi_G\pi_Y}{\pi_R}\right) e^{-\beta t} - 2\left(\frac{\pi_T\pi_C}{\pi_Y} + \frac{\pi_A\pi_G}{\pi_R}\right) e^{-(\kappa+1)\beta t},$$
$$E(V) = 2\pi_Y\pi_R(1 - e^{-\beta t}). \tag{1.27}$$

By equating the observed proportions S and V to their expectations, one can obtain a system of two equations in two unknowns, which can be solved to give

$$\hat{d} = 2\left(\frac{\pi_T\pi_C}{\pi_Y} + \frac{\pi_A\pi_G}{\pi_R}\right) a - 2\left(\frac{\pi_T\pi_C\pi_R}{\pi_Y} + \frac{\pi_A\pi_G\pi_Y}{\pi_R} - \pi_Y\pi_R\right) b,$$
$$\hat{\kappa} = a/b - 1, \tag{1.28}$$

where

$$a = \overline{(\kappa+1)\beta t} = -\log\left\{1 - \frac{S}{2\left(\frac{\pi_T\pi_C}{\pi_Y} + \frac{\pi_A\pi_G}{\pi_R}\right)} - \frac{\left(\frac{\pi_T\pi_C\pi_R}{\pi_Y} + \frac{\pi_A\pi_G\pi_Y}{\pi_R}\right)V}{2\left(\pi_T\pi_C\pi_R + \pi_A\pi_G\pi_Y\right)}\right\},$$
$$b = \overline{\beta t} = -\log\left\{1 - \frac{V}{2\pi_Y\pi_R}\right\} \tag{1.29}$$

(Tateno et al. 1994; Yang 1994a). The approximate variance of $\hat{d}$ can be obtained similarly to that under K80 (Tateno et al. 1994). The estimated distance under F84 for the 12s rRNA genes is shown in Table 1.4.

If we assume $\alpha_1 = \alpha_2 = \beta$ in the TN93 model, we obtain the F81 model (Felsenstein 1981) (Table 1.1). A distance formula was derived by Tajima and Nei (1982). Estimates under this and some other models for the 12s rRNA dataset of Table 1.3 are listed in Table 1.4. It may be mentioned that the matrices Λ, U, U^{-1} and $P(t)$ derived for the TN93 model hold for its special cases, such as JC69 (Jukes and Cantor 1969), K80 (Kimura 1980), F81 (Felsenstein 1981), HKY85 (Hasegawa et al. 1984, 1985), and F84. Under some of those simpler models, simplifications are possible (see Problem 1.2).

1.2.4 *The transition/transversion rate ratio*

Unfortunately at least three definitions of the 'transition/transversion rate ratio' are in use in the literature. The first is the ratio of the numbers (or proportions) of transitional and transversional differences between the two sequences, without correcting for multiple hits (e.g. Wakeley 1994). This is $E(S)/E(V) = p_1(t)/(2p_2(t))$ under the K80 model (see equation (1.10)). For highly similar sequences, this is close to $\alpha/(2\beta)$ under K80. At intermediate levels of sequence divergence, $E(S)/E(V)$ increases with $\alpha/(2\beta)$, but the pattern is complex. When the sequences are very different, $E(S)/E(V)$ approaches $1/2$ irrespective of $\alpha/(2\beta)$. Figure 1.5 plots the ratio $E(S)/E(V)$ against the sequence divergence. Thus the ratio is meaningful only for closely related sequences. In real datasets, however, highly similar sequences may not contain much information and the estimate may involve large sampling errors. In general, the $E(S)/E(V)$ ratio is a poor measure of the transition–transversion rate difference and should be avoided.

The second measure is $\kappa = \alpha/\beta$ in the models of Kimura (1980) and Hasegawa et al. (1985), with $\kappa = 1$ meaning no rate difference between transitions and transversions. A third measure may be called the average transition/transversion ratio, and is the ratio

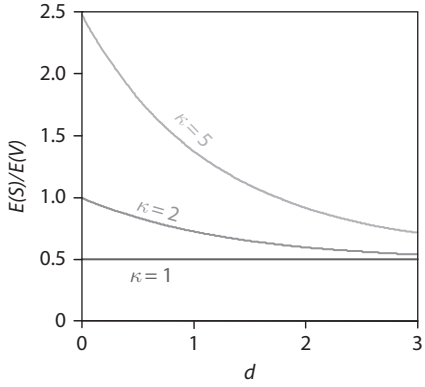

Fig. 1.5 The transition/transversion ratio $E(S)/E(V)$ under the K80 model plotted against sequence divergence d. This is $p_1/(2p_2)$ in equation (1.11) and corresponds to infinitely long sequences.

Table 1.5 Average transition/transversion ratio R

Model	Average transition/transversion rate ratio (R)
JC69	$\frac{1}{2}$
K80	$\frac{\kappa}{2}$
F81	$\frac{\pi_T\pi_C + \pi_A\pi_G}{\pi_Y\pi_R}$
F84	$\frac{\pi_T\pi_C(1+\kappa/\pi_Y) + \pi_A\pi_G(1+\kappa/\pi_R)}{\pi_Y\pi_R}$
HKY85	$\frac{(\pi_T\pi_C + \pi_A\pi_G)\kappa}{\pi_Y\pi_R}$
TN93	$\frac{\pi_T\pi_C\kappa_1 + \pi_A\pi_G\kappa_2}{\pi_Y\pi_R}$
REV (GTR)	$\frac{\pi_T\pi_C a + \pi_A\pi_G f}{\pi_T\pi_A b + \pi_T\pi_G c + \pi_C\pi_A d + \pi_C\pi_G e}$
UNREST	See equation (1.30) in text

of the expected numbers of transitional and transversional substitutions between the two sequences. This is the same measure as the first one, except that it corrects for multiple hits. For a general substitution rate matrix (the UNREST model in Table 1.1 but note that $q_{TC} = a$, $q_{TA} = b$, etc.), this is

$$R = \frac{\pi_T q_{TC} + \pi_C q_{CT} + \pi_A q_{AG} + \pi_G q_{GA}}{\pi_T q_{TA} + \pi_T q_{TG} + \pi_C q_{CA} + \pi_C q_{CG} + \pi_A q_{AT} + \pi_A q_{AC} + \pi_G q_{GT} + \pi_G q_{GC}}. \tag{1.30}$$

Note that the Markov chain spends a proportion π_T of time in state T, while q_{TC} is the rate that T changes to C. Thus $\pi_T q_{TC}$ is the amount of 'flow' from T to C. The numerator in equation (1.30) is the average amount of transitional change, while the denominator gives the amount of transversional change. Table 1.5 gives R for commonly used simple models. Under the model of Kimura (1980), $R = \alpha/(2\beta)$ and equals $1/2$ when there is no transition–transversion rate difference. As from each nucleotide, one change is a transition and two changes are transversions, we expect to see twice as many transversions as transitions, hence the ratio $1/2$.

Note that parameter κ has different definitions under the F84 and HKY85 models (Table 1.1). Without the transition–transversion rate difference, $\kappa_{\text{F84}} = 0$ and $\kappa_{\text{HKY85}} = 1$. Roughly, $\kappa_{\text{HKY85}} \simeq 1 + 2\kappa_{\text{F84}}$. By forcing the average ratio R to be identical under the two models (Table 1.5), one can derive a more accurate approximation (Goldman 1993):

$$\kappa_{\text{HKY85}} \simeq 1 + \frac{\pi_T\pi_C/\pi_Y + \pi_A\pi_G/\pi_R}{\pi_T\pi_C + \pi_A\pi_G}\kappa_{\text{F84}}. \tag{1.31}$$

Overall, R is more convenient to use for comparing estimates under different models while κ is more suitable for formulating the null hypothesis of no transition–transversion rate difference.

Some authors (e.g. Singh et al. 2009) used the measure

$$R^* = \frac{q_{TC} + q_{CT} + q_{AG} + q_{GA}}{q_{TA} + q_{TG} + q_{CA} + q_{CG} + q_{AT} + q_{AC} + q_{GT} + q_{GC}}. \tag{1.32}$$

Under the K80 model this has the ratio $1/2$ if there is no transition–transversion rate difference. For other models, this is hard to interpret and should be avoided.

1.3 Variable substitution rates across sites

All models discussed in §1.2 assume that different sites in the sequence evolve at the same rate. This assumption may be unrealistic in real data. First, the mutation rate may vary among sites (Hodgkinson and Eyre-Walker 2011). Second, different sites may play different roles in the structure and function of the gene and are thus under different selective pressures. Mutations at different sites may thus be fixed in the population at different rates. When the substitution rates vary, the hotspots may accumulate many changes, while the conserved sites remain unchanged. Thus, for the same sequence distance or the same amount of evolutionary change, we will observe fewer differences than if the rate is constant. In other words, ignoring variable rates among sites leads to underestimation of the sequence distance.

One can accommodate the rate variation by assuming that the rate r for any site is a random variable drawn from a statistical distribution. The most commonly used distribution is the gamma distribution. The resulting models are represented by a suffix '+Γ', such as JC69 + Γ, K80 + Γ, etc., and the distances are sometimes called *gamma distances*. The density function of the gamma distribution is

$$g(r;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} e^{-\beta r} r^{\alpha-1}, r > 0, \tag{1.33}$$

where $\alpha > 0$ and $\beta > 0$ are the shape and rate parameters. Here

$$\Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1} dt \tag{1.34}$$

is the gamma function. For integer n, $\Gamma(n) = (n-1)!$. The mean and variance of the gamma distribution are $E(r) = \alpha/\beta$ and var $(r) = \alpha/\beta^2$. To avoid using too many parameters, we set $\beta = \alpha$ so that the mean of the distribution is 1, with variance $1/\alpha$. The shape parameter α is then inversely related to the extent of rate variation at sites (Figure. 1.6). If $\alpha > 1$, the distribution is bell-shaped, meaning that most sites have intermediate rates around 1, while few sites have either very low or very high rates. In particular, when $\alpha \to \infty$, the distribution degenerates into the model of a single rate for all sites. If $\alpha \leq 1$, the distribution has a highly skewed L-shape, meaning that most sites have very low rates or are nearly 'invariable', but there are some substitution hot spots with high rates. Estimation of α from real data requires joint comparison of multiple sequences as it is virtually impossible to do so using only two sequences. We will discuss estimation of α later, in §4.3.1. Here we assume that α is given.

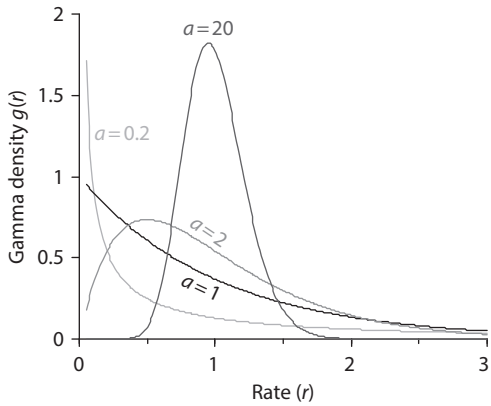

Fig. 1.6 Probability density function of the gamma distribution for variable rates among sites. The rate parameter of the distribution is fixed so that the mean is one; as a result, the density involves only the shape parameter α. The x-axis is the substitution rate, while the y-axis is proportional to the number of sites with that rate.

With variable rates among sites, the sequence distance is defined as the expected number of substitutions per site, averaged over all sites. Here we will derive gamma distances under the JC69 and K80 models, and comment on similar derivations under other models.

First we consider the JC69 + Γ model. Since the mean rate in the gamma distribution is one, the distance averaged across all sites is still d. As in the case of the JC69 model with one rate for all sites, we equate the observed proportion of different sites to the expected probability under JC69 + Γ to derive a distance formula. If a site has rate r, the distance between the sequences at that site is $d{\cdot}r$, and the probability of observing a difference at the site is $p(d{\cdot}r)$, with p given in equation (1.6). However, r is an unknown random variable, so we have to consider contributions from sites with different rates. In other words, we average over the distribution of r to calculate the unconditional probability of different sites:

$$p = \int_0^\infty \left(\frac{3}{4} - \frac{3}{4}e^{-4d\cdot r/3}\right) g(r)\,dr = \frac{3}{4} - \frac{3}{4}\left(1 + \frac{4d}{3\beta}\right)^{-\alpha}. \tag{1.35}$$

Here we made use of the following result for the gamma distribution:

$$\int_0^\infty e^{-cr} g(r)\,dr = \int_0^\infty e^{-cr}\frac{\beta^\alpha}{\Gamma(\alpha)}e^{-\beta r}r^{\alpha-1}\,dr = \left(\frac{\beta+c}{\beta}\right)^{-\alpha}\int_0^\infty \left[\frac{(\beta+c)^\alpha}{\Gamma(\alpha)}e^{-(\beta+c)r}r^{\alpha-1}\right]dr = \left(1+\frac{c}{\beta}\right)^{-\alpha}, \tag{1.36}$$

for any constant $c > 0$. Note that the quantity in the square brackets is the probability density function (PDF) for the gamma distribution with parameters α and $\beta + c$ and thus the integral is 1.

By equating p of equation (1.35) to the observed $\hat{p}$, we obtain the JC69 + Γ distance as

$$\hat{d} = \frac{3}{4}\alpha\left[\left(1 - \frac{4}{3}\hat{p}\right)^{-1/\alpha} - 1\right] \tag{1.37}$$

(Golding 1983), with variance

$$\mathrm{var}(\hat{d}) = \mathrm{var}(\hat{p}) \times \left|\frac{dd}{dp}\right|^2 = \frac{\hat{p}(1-\hat{p})}{n} \times \left(1 - \frac{4}{3}\hat{p}\right)^{-2/\alpha-2}. \tag{1.38}$$

Now we consider the K80 + Γ model. To avoid confusion about the notation, we use d and κ as parameters under the K80 model and α and β as parameters of the gamma distribution. The distance is defined as an average across all sites. As in the case of K80 with one rate for all sites, we equate the observed proportions of transitional and transversional differences to their expected probabilities to derive a distance formula. If a site has rate r, both transition and transversion rates at the site are multiplied by r, with the same transition/transversion rate ratio κ assumed across sites. As in the case of JC69 + Γ, we derive the probability for a transitional difference by averaging $p_1(d \cdot r)$, with p_1 given in equation (1.11), over the gamma distribution:

$$\begin{aligned} E(S) &= \int_0^\infty p_1(d\cdot r)\,g(r)\,dr \\ &= \int_0^\infty \left[\frac{1}{4} + \frac{1}{4}\exp\left(\frac{-4d\cdot r}{\kappa+2}\right) - \frac{1}{2}\exp\left(\frac{-2(\kappa+1)d\cdot r}{\kappa+2}\right)\right] g(r)\,dr \\ &= \frac{1}{4} + \frac{1}{4}\left(1 + \frac{4d}{(\kappa+2)\alpha}\right)^{-\alpha} - \frac{1}{2}\left(1 + \frac{2(\kappa+1)d}{(\kappa+2)\alpha}\right)^{-\alpha}, \end{aligned} \tag{1.39}$$

making use of equation (1.36). Similarly the probability that we observe a transversional difference is

$$E(V) = \int_0^\infty 2p_2(d\cdot r)\,g(r)\,dr = \frac{1}{2} - \frac{1}{2}\left(1 + \frac{4d}{(\kappa+2)\alpha}\right)^{-\alpha}, \tag{1.40}$$

where p_2 is given in equation (1.11). Equating the above to the observed proportions S and V leads to

$$\begin{aligned}\hat{d} &= \frac{\alpha}{2}[(1-2S-V)^{-1/\alpha}-1] + \frac{\alpha}{4}[(1-2V)^{-1/\alpha}-1],\\ \hat{\kappa} &= \frac{2[(1-2S-V)^{-1/\alpha}-1]}{[(1-2V)^{-1/\alpha}-1]} - 1\end{aligned} \tag{1.41}$$

(Jin and Nei 1990). Compared with equation (1.12) for the one-rate model, the only change is that the logarithm function $\log(y)$ becomes $-\alpha(y^{-1/\alpha}-1)$. This is a general feature of gamma distances. The large-sample variance of $\hat{d}$ is given by equation (1.14) except that now

$$\begin{aligned}a &= (1-2S-V)^{-1/\alpha-1},\\ b &= \frac{1}{2}[(1-2S-V)^{-1/\alpha-1} + (1-2V)^{-1/\alpha-1}].\end{aligned} \tag{1.42}$$

In general, note that equation (1.18) can be written equivalently as

$$p_{ij}(t) = \sum_{k=1}^{4} c_{ijk} e^{\lambda_k t} = \sum_{k=1}^{4} u_{ik} u_{kj}^{-1} e^{\lambda_k t}, \tag{1.43}$$

where λ_k is the kth eigenvalue of the rate matrix Q, u_{ik} is the ikth element of U, and u_{kj}^{-1} is the kjth element of U^{-1} in equation (1.18). Thus the probability of observing nucleotides i and j in the two sequences at a site is

$$f_{ij}(t) = \int_0^\infty \pi_i p_{ij}(t \cdot r) g(r)\, dr = \pi_i \sum_{k=1}^{4} c_{ijk}\left(1 - \lambda_k t/\alpha\right)^{-\alpha}, \tag{1.44}$$

by equation (1.36). The exponential functions under the one-rate model are replaced by the power functions under the gamma model. Under the one-rate model, we can view the exponential functions as unknowns to solve the equations, and now we can view those power functions as unknowns. Thus, one can derive a gamma distance under virtually every model for which a one-rate distance formula is available. Those include the F84 model (Yang 1994a) and the TN93 model (Tamura and Nei 1993), among others.

Example 1.4. We calculate the sequence distance between the two mitochondrial 12s rRNA genes under the K80 + Γ model, with $\alpha = 0.5$ fixed. The estimates of the distance and the transition/transversion rate ratio κ are $\hat{d} \pm \text{SE} = 0.1283 \pm 0.01726$ and $\hat{\kappa} \pm \text{SE} = 37.76 \pm 16.34$. Both estimates are much larger than under the one-rate model (Table 1.4). It is well known that ignoring rate variation among sites leads to underestimation of both the sequence distance and the transition/transversion rate ratio (Wakeley 1994; Yang 1996a). The underestimation is more serious at larger distances and with more variable rates (that is, smaller α). □

1.4 Maximum likelihood estimation of distance

In this section, we discuss the ML method for estimating sequence distances. ML is a general methodology for estimating parameters in a model and for testing hypotheses concerning the parameters. It plays a central role in statistics and is widely used in molecular phylogenetics. It forms the basis of much material covered later in this book. We will focus mainly on the JC69 and K80 models, re-deriving the distance formulae discussed

earlier. While discovering what we already know may not be very exciting, it may be effective in helping us understand the workings of the likelihood method. Note that ML is an 'automatic' method, as it tells us how to proceed even when the estimation problem is difficult and our intuition fails. Interested readers should consult a statistics textbook, for example, DeGroot and Schervish (2002), Kalbfleisch (1985), and Edwards (1992) at elementary levels, or Cox and Hinkley (1974) and Stuart et al. (1999) at more advanced levels.

1.4.1 *The JC69 model*

Let X be the observed data and θ the parameter we hope to estimate. The probability of observing data X, when viewed as a function of the unknown parameter θ with the data given, is called the *likelihood function*: $L(\theta; X) = f(X|\theta)$. Probability and likelihood are fundamentally different concepts; see Box 1.1 for a summary. According to the *likelihood principle*, the likelihood function contains all information in the data about θ. The value of θ that maximizes the likelihood, say $\hat{\theta}$, is our best point estimate, called the *maximum likelihood estimate* (MLE). Furthermore, the likelihood curve around $\hat{\theta}$ provides information about the uncertainty in the point estimate. The theory applies to problems with either a single parameter or with multiple parameters. In the later case θ is a vector.

Box 1.1 PROBABILITY VERSUS LIKELIHOOD

Probability and likelihood are fundamentally different concepts.

- Likelihood is defined up to a proportionality constant. Likelihood does not integrate (over the parameter space) to 1. Probability sums (over the sample space) to 1.
- The probability curve should be interpreted by area under the curve, while its height is not meaningful. The likelihood should be compared using the height and point by point, e.g. $L(\theta_1) > L(\theta_2)$ for two points θ_1 and θ_2. The area under the likelihood curve is not meaningful.
- Likelihood is invariant to reparametrization. Suppose we are interested in the size of a crater lake, which is a circle. We can use either α (the area of the circle) or β (the radius) as the parameter in the model, with $\alpha = \pi\beta^2$. Then if $L(\alpha_1) > L(\alpha_2)$, we have $L(\beta_1) > L(\beta_2)$. In general, if α and β are alternative parametrizations, with $\beta = h(\alpha)$ where h is a one-to-one monotonic function, then the MLEs are invariant to reparametrization: $\hat{\beta} = \hat{h}(\alpha) = h(\hat{\alpha})$. In contrast, variable transformation (if it is nonlinear) in general changes the shape of the probability density. For example, Appendix A includes an example in which x has a two-moded distribution while $y = h(x)$ has only one mode.

Here we apply the theory to estimation of the distance between two sequences under the JC69 model. The single parameter is the distance d. The data are two aligned sequences, each n sites long, with x differences. From equation (1.6), the probability of observing different nucleotides at a site between two sequences separated by distance d is

$$p = 3p_1 = \frac{3}{4} - \frac{3}{4}e^{-4d/3}. \tag{1.45}$$

Thus, the probability of observing the data, that is, x differences out of n sites, is given by the binomial probability

$$L(d; x) = f(x|d) = Cp^x(1-p)^{n-x} = C\left(\frac{3}{4} - \frac{3}{4}e^{-4d/3}\right)^x\left(\frac{1}{4} + \frac{3}{4}e^{-4d/3}\right)^{n-x}. \tag{1.46}$$

As the data x are observed, this probability is now considered a function of the parameter d. Values of d with higher L are better supported by the data than values of d with lower L. As multiplying the likelihood by any function of the data that is independent of the parameter θ will not change our inference about θ, the likelihood is defined up to a proportionality constant. We will use this property to introduce two changes to the likelihood of equation (1.46). First, the binomial coefficient $C = n!/[x!(n-x)!]$ is a constant and will be dropped. Second, to use the same definition of likelihood for all substitution models, we will distinguish 16 possible data outcomes at a site (the 16 possible site patterns) instead of just two outcomes (that is, difference with probability p and identity with probability $1-p$) as in equation (1.46). Under JC69, the four constant patterns (TT, CC, AA, GG) have the same probability of occurrence, as do the 12 variable site patterns (TC, TA, TG etc.). This will not be the case for other models. Thus the redefined likelihood is given by the multinomial probability with 16 cells:

$$L(d;x) = \left(\frac{1}{4}p_1\right)^x \left(\frac{1}{4}p_0\right)^{n-x} = \left(\frac{1}{16} - \frac{1}{16}e^{-4d/3}\right)^x \left(\frac{1}{16} + \frac{3}{16}e^{-4d/3}\right)^{n-x}, \tag{1.47}$$

where p_0 and p_1 are from equation (1.4). Each of the 12 variable site patterns has probability $\frac{1}{4}\ p_1$ or $\frac{1}{12}\ p$. For example, the probability for site pattern TC is equal to $\frac{1}{4}$, the probability that the starting nucleotide is T, times the transition probability $p_{TC}(t) = p_1$ from equation (1.4). Similarly, each of the four constant site patterns (TT, CC, AA, GG) has probability $\frac{1}{4}\ p_0$ or $(1-p)/4$. The reader can verify that equations (1.46) and (1.47) differ only by a proportionality constant (Problem 1.4).

Furthermore, the likelihood L is typically extremely small and awkward to work with. Thus its logarithm $\ell(d) = \log\{L(d)\}$ is commonly used instead. As the logarithm function is monotonic, we achieve the same result; that is, $L(d_1) > L(d_2)$ if and only if $\ell(d_1) > \ell(d_2)$. The *log likelihood function* is thus

$$\ell(d;x) = \log\{L(d;x)\} = x\log\left(\frac{1}{16} - \frac{1}{16}e^{-4d/3}\right) + (n-x)\log\left(\frac{1}{16} + \frac{3}{16}e^{-4d/3}\right). \tag{1.48}$$

To estimate d, we maximize L or equivalently its logarithm ℓ. By setting $d\ell/dd = 0$, we can determine that ℓ is maximized at

$$\hat{d} = -\frac{3}{4}\log\left(1 - \frac{4}{3}\times\frac{x}{n}\right). \tag{1.49}$$

This is the MLE of d. It is the distance formula, equation (1.7), which we derived earlier.

We now discuss some statistical properties of MLEs. Under quite mild regularity conditions we will not go into, the MLEs have nice asymptotic (large-sample) properties (see, e.g. Stuart et al. 1999, pp. 46–116). For example, they are consistent, asymptotically unbiased and efficient. Consistency means that the estimate $\hat{\theta}$ converges to the true value θ when the sample size $n \to \infty$. Unbiasedness means that the expectation of the estimate equals the true parameter value: $E(\hat{\theta}) = \theta$. Efficiency means that no other unbiased estimate can have a smaller variance than the MLE. Furthermore, the MLEs are asymptotically normally distributed. These properties are known to hold in large samples. How large the sample size has to be for the approximation to be reliable depends on the particular problem.

Another important property of MLEs is that they are invariant to transformations of parameters or reparametrizations. The MLE of a function of parameters is the same function of the MLEs of the parameters: $\hat{h}(\theta) = h(\hat{\theta})$. Thus if the same model can be formulated using either parameters θ_1 or θ_2, with θ_1 and θ_2 constituting a one-to-one mapping, use of either parameter leads to the same inference. For example, we can use the probability of

a difference between the two sequences p as the parameter in the JC69 model instead of the distance d. The two form a one-to-one mapping through equation (1.45). The log likelihood function for p corresponding to equation (1.47) is $L(p; x) = \left(\frac{p}{12}\right)^x \left(\frac{1-p}{4}\right)^{n-x}$, from which we get the MLE of p: $\hat{p} = x/n$. We can then view d as a function of p and obtain its MLE $\hat{d}$, as given by equation (1.49). Whether we use p or d as the parameter, the same inference is made, and the same log likelihood is achieved: $\ell(\hat{d}) = \ell(\hat{p}) = x \log \frac{x}{12n} + (n-x) \log \frac{n-x}{4n}$.

Two approaches can be used to calculate a confidence interval for the MLE. The first relies on the theory that the MLE $\hat{\theta}$ is asymptotically normally distributed around the true value θ when the sample size $n \to \infty$. The asymptotic variance can be calculated using either the observed information $-\frac{d^2\ell}{d\theta^2}$ or the expected (Fisher) information $-E\left(\frac{d^2\ell}{d\theta^2}\right)$. While both are reliable in large samples, the observed information is preferred in real data analysis (e.g. Efron and Hinkley 1978). This is equivalent to using a quadratic polynomial to approximate the log likelihood around the MLE. Here we state the result for the multivariate case, with k parameters in the model:

$$\hat{\theta} \sim N_k(\theta, -H^{-1}), \text{ with } H = \left\{ \frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j} \right\}. \tag{1.50}$$

In other words, the MLEs $\hat{\theta}$ have an asymptotic k-variate normal distribution, with the mean to be the true values θ, and the variance–covariance matrix to be $-H^{-1}$, where H is the matrix of second derivatives, also known as the Hessian matrix (Stuart et al. 1999, pp. 73–74).

In our example, the asymptotic variance for $\hat{d}$ is

$$\text{var}(\hat{d}) = -\left(\frac{d^2\ell}{dd^2}\right)^{-1} = \frac{\hat{p}(1-\hat{p})}{(1-4\hat{p}/3)^2 n}. \tag{1.51}$$

This is equation (1.8). An approximate 95% confidence interval for d can be constructed as $\hat{d} \pm 1.96\sqrt{\text{var}(\hat{d})}$.

The normal approximation has a few drawbacks. First, if the log likelihood curve is not symmetrical around the MLE, the normal approximation will be unreliable. For example, if the parameter is a probability, which ranges from 0 to 1, and the MLE is close to 0 or 1, the normal approximation may be very poor. Second, the confidence interval constructed this way includes parameter values that have lower likelihood than values outside the interval. Third, even though the MLEs are invariant to reparametrizations, the confidence intervals constructed using the normal approximation are not.

The second approach is called the *likelihood interval*, which avoids all three problems associated with the normal approximation. It is based on the likelihood ratio test (LRT). In large samples, the LRT statistic, $2[\ell(\hat{\theta}) - \ell(\theta)]$, where θ is the true parameter value and $\hat{\theta}$ is the MLE, has a χ^2_k distribution with the degree of freedom k equal to the number of parameters. Thus one can lower the log likelihood from the peak $\ell(\hat{\theta})$ by $\frac{1}{2}\chi^2_{k,\,5\%}$ to construct a 95% confidence (likelihood) region. Here $\chi^2_{k,\,5\%}$ is the 5% critical value of the χ^2 distribution with k degrees of freedom. The likelihood region contains parameter values with the highest likelihood, values that cannot be rejected by an LRT at the 5% level when compared against $\hat{\theta}$. This likelihood ratio approach is known to give more reliable intervals than the normal approximation. The normal approximation works well for some parametrizations but not for others; the likelihood interval automatically uses the best parametrization.

Example 1.5. For the 12s rRNA data of Table 1.3, we have $\hat{p} = x/n = 90/948 = 0.09494$, and $\hat{d} = 0.1015$. The variance of $\hat{d}$ is 0.0001188, so that the 95% confidence interval based on the normal approximation is $(0.0801,\ 0.1229)$. If we use p as the parameter instead, we have $\text{var}(\hat{p}) = \hat{p}(1-\hat{p})/n = 0.00009064$, so that the 95% confidence interval for p is $(0.0763,\ 0.1136)$. These two intervals do not match; for example, if we use the lower bound for p to calculate the lower bound for d, the result will be different. The log likelihood curves are shown in Figure 1.7, with the peak at $\ell(\hat{d}) = \ell(\hat{p}) = -1710.577$. By lowering the log likelihood ℓ by $\frac{1}{2}\chi^2_{1,\,5\%} = 3.841/2 = 1.921$ from its peak, we obtain the 95% likelihood intervals $(0.0817,\ 0.1245)$ for d and $(0.0774,\ 0.1147)$ for p. Compared with the intervals based on the normal approximation, the likelihood intervals are asymmetrical and are shifted to the right, reflecting the steeper drop of log likelihood and thus

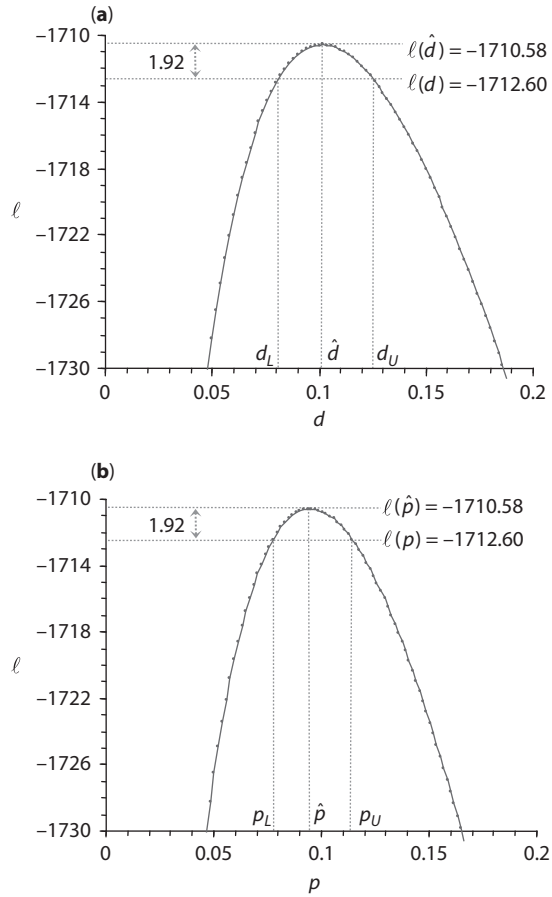

Fig. 1.7 Log likelihood curves and construction of confidence (likelihood) intervals under the JC69 model. The parameter in the model is the sequence distance d in (**a**) and the probability of different sites p in (**b**). The mitochondrial 12s rRNA genes of Table 1.3 are analysed.

more information on the left side of the MLE. Also the likelihood intervals for p and d match each other. The likelihood interval is invariant to reparametrization. □

1.4.2 *The K80 model*

The likelihood theory applies to models with multiple parameters. We apply the method to estimation of the sequence distance d and the transition/transversion rate ratio κ under the K80 model (Kimura 1980). The data are the numbers of sites with transitional (n_S) and transversional (n_V) differences, with the number of constant sites to be $n - n_S - n_V$. In deriving the probabilities of observing such sites, we again consider all 16 site patterns, as for the JC69 model. Thus the probability is $\frac{1}{4}\, p_0$ for any constant site (e.g. TT), $\frac{1}{4}\, p_1$ for any site with a transitional difference (e.g. TC), and $\frac{1}{4}\, p_2$ for any site with a transversional difference (e.g. TA), with p_0, p_1, p_2 given in equation (1.11). The log likelihood is

$$\begin{aligned}\ell(d, \kappa; n_S, n_V) &= \log\{f(n_S, n_V|d, \kappa)\} \\ &= (n - n_S - n_V)\log(p_0/4) + n_S \log(p_1/4) + n_V \log(p_2/4).\end{aligned} \tag{1.52}$$

MLEs of d and κ can be derived from the likelihood equation $\partial\ell/\partial d = 0, \partial\ell/\partial\kappa = 0$. The solution can be shown to be equation (1.12), with $S = n_S/n$ and $V = n_V/n$. A simpler argument relies on the invariance property of the MLEs. Suppose we consider the probabilities of transitional and transversional differences $E(S) = p_1$ and $E(V) = 2p_2$ as parameters in the model instead of d and κ. From the log likelihood (equation (1.52)), the MLEs of $E(S)$ and $E(V)$ are simply S and V. The MLEs of d and κ can be obtained through the one-to-one mapping between the two sets of parameters, which involves the same step taken when we derived equation (1.12) in §1.2.2 by equating the observed proportions S and V to their expected probabilities.

Example 1.6. For the 12s rRNA data of Table 1.3, we have $S = 0.08861$ and $V = 0.00633$. The MLEs are thus $\hat{d} = 0.1046$ for the sequence distance and $\hat{\kappa} = 30.83$ for the transition/transversion rate ratio. These are the same as calculated in Example 1.2. The maximized log likelihood is $\ell(\hat{d}, \hat{\kappa}) = -1637.905$. Application of equation (1.50) leads to the variance–covariance matrix (see Appendix B):

$$\text{var}\begin{pmatrix}\hat{d} \\ \hat{\kappa}\end{pmatrix} = \begin{pmatrix}0.0001345 & 0.007253 \\ 0.007253 & 172.096\end{pmatrix}. \tag{1.53}$$

From this, one can get the approximate SEs to be 0.0116 for $\hat{d}$ and 13.12 for $\hat{\kappa}$. The log likelihood surface contour is shown in Figure 1.8, which indicates that the data are much more informative about d than about κ. One can lower the log likelihood from its peak by $\frac{1}{2}\chi^2_{2,\,5\%} = 5.991/2 = 2.996$, to construct a 95% confidence (likelihood) region for the two parameters (Figure 1.8). □

1.4.3 *Likelihood ratio test of substitution models*

We may ask whether the transition and transversion rates are indeed different or whether K80 fits the data much better than JC69. The LRT provides a general framework for testing hypotheses concerning model parameters in the likelihood framework. Most of the tests we have learned in a biostatistics course are its special cases or approximations, such as the t test comparing two population means, the t test of the linear regression coefficient, and χ^2 test of association in a 2×2 contingency table. The two hypotheses under comparison are nested, with one being a special case of the other. Suppose the simpler hypothesis (the

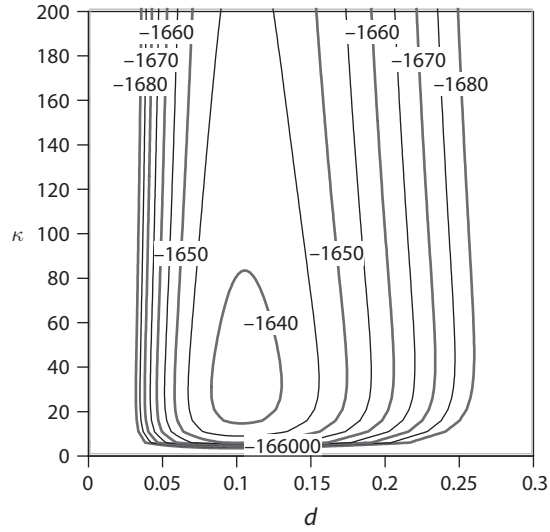

Fig. 1.8 Log likelihood contours for the sequence distance d and the transition/transversion rate ratio κ under the K80 model. The mitochondrial 12s rRNA genes of Table 1.3 are analysed. The peak of the surface is at the MLEs $\hat{d} = 0.1046$, $\hat{\kappa} = 30.83$, with $\ell = -1637.905$. The 95% likelihood region is surrounded by the contour line at $\ell = -1637.905 - 2.996 = -1640.901$ (not shown).

null hypothesis H_0) involves q parameters and the more general hypothesis (the *alternative hypothesis* H_1) has p parameters. Let the maximum log likelihood values under the two models be $\ell_0 = \log\{L(\hat{\theta}_0)\} = \ell(\hat{\theta}_0)$ and $\ell_1 = \log\{L(\hat{\theta}_1)\} = \ell(\hat{\theta}_1)$, where $\hat{\theta}_0$ and $\hat{\theta}_1$ are the MLEs under the two models, respectively. Then under certain regularity conditions, the LRT statistic

$$2\Delta\ell = 2\log\left(\frac{L_1}{L_0}\right) = 2(\ell_1 - \ell_0) \tag{1.54}$$

is asymptotically distributed as χ^2_{p-q} if H_0 is true. In other words, if the null model is true, twice the log likelihood difference between the null and alternative models is approximately χ^2 distributed with the degree of freedom equal to the difference in the number of parameters between the two models. The approximation applies to large samples.

Example 1.7. We use the LRT to compare JC69 and K80 using the 12s rRNA data of Table 1.3. The (maximized) log likelihood under H_0 (JC69) is $\ell_0 = -1710.58$ and that under H_1 (K80) is $\ell_1 = -1637.90$ (Table 1.4). The LRT statistic is $2\Delta\ell = 2(\ell_1 - \ell_0) = 2[-1637.90 - (-1710.58)] = 145.36$. Note that JC69 is equivalent to K80 with parameter $\kappa = 1$ fixed, so that JC69 is nested within K80. Thus $2\Delta\ell$ should be compared with the χ^2 distribution with one degree of freedom (df = 1), with the significance values to be 3.84 at 5% and 6.63 at 1%. K80 fits the dataset much better. It has been observed that we can often easily reject simpler models using molecular sequence data, possibly because the datasets are typically large. □

*1.4.4 *Profile and integrated likelihood methods*

Suppose we are interested in the sequence distance d under the K80 model (Kimura 1980) but not in the transition/transversion rate ratio κ. However we want to consider κ in the model as transition and transversion rates are known to differ and the rate difference may affect our estimation of d. Parameter κ is thus appropriately called a *nuisance parameter*, while d is our parameter of interest. Dealing with nuisance parameters is commonly considered a weakness of the likelihood method. The approach we described above, estimating both d and κ with ML and using $\hat{d}$ while ignoring $\hat{\kappa}$, is known variously as the *relative likelihood*, *pseudo likelihood* or *estimated likelihood*, since the nuisance parameters are replaced by their estimates.

A more respected approach is the *profile likelihood*. This defines a log likelihood for the parameters of interest only, which is calculated by optimizing the nuisance parameters at fixed values of the parameters of interest. In other words, the profile log likelihood for d is $\ell(d) = \ell(d, \hat{\kappa}_d)$, where $\hat{\kappa}_d$ is the MLE of κ for the given d. This is a pragmatic approach that most often leads to reasonable answers. The likelihood interval for $\hat{d}$ is constructed from the profile likelihood in the usual way.

Example 1.8. For the 12s rRNA genes, the highest likelihood $\ell(\hat{d}) = -1637.905$ is achieved at $\hat{d} = 0.1046$ and $\hat{\kappa} = 30.83$. Thus the point estimate of d is the same as before. We fix d at different values. For each fixed d, the log likelihood (1.52) is a function of the nuisance parameter κ, and is maximized to estimate κ. Let the estimate be $\hat{\kappa}_d$, with the subscript indicating it is a function of d. It does not seem possible to derive $\hat{\kappa}_d$ analytically, so we use a numerical optimization algorithm instead (as discussed later in §4.5). The optimized likelihood is the profile likelihood for d: $\ell(d) = \ell(d, \hat{\kappa}_d)$. This is plotted against d in Figure 1.9a, together with the estimate $\hat{\kappa}_d$. We lower the log likelihood by $\frac{1}{2}\chi^2_{1,\,5\%} = 1.921$ to construct the profile likelihood interval for d: $(0.0836,\ 0.1293)$. □

If the model involves many parameters, and in particular, if the number of parameters increases without bound with the increase of the size of the data, the likelihood method may run into deep trouble, so deep that the MLEs may not even be consistent (e.g. Kalbfleisch and Sprott 1970; Kalbfleisch 1985, pp. 92–96). A useful strategy in this case is to assign a statistical distribution to describe the variation or uncertainties in the *nuisance parameters*, and integrate them out in the likelihood. This is known as *integrated likelihood* or *marginal likelihood* and has the flavour of a Bayesian approach.

Here we apply the idea to deal with the nuisance parameter κ. Let $f(\kappa)$ be the distribution assigned to κ, also known as a prior. Then the integrated likelihood is

$$\begin{aligned} L(d) &= \int_0^\infty f(\kappa)\, f(n_S,\, n_V \,|\, d, \kappa)\, \mathrm{d}\kappa \\ &= \int_0^\infty f(\kappa) \times \left(\frac{p_0}{4}\right)^{n-n_S-n_V} \left(\frac{p_1}{4}\right)^{n_S} \left(\frac{p_2}{4}\right)^{n_V} \mathrm{d}\kappa, \end{aligned} \tag{1.55}$$

where $p_0, p_1,$ and p_2 are from equation (1.11). For the present problem, it is possible to use an improper prior: $f(k) = 1,\ 0 < \kappa < \infty$. The prior is *improper* as it does not integrate to 1 and is not a proper probability density. The integrated likelihood is then

$$L(d) = \int_0^\infty f(n_S, n_V|d, \kappa)\, \mathrm{d}\kappa = \int_0^\infty \left(\frac{p_0}{4}\right)^{n-n_S-n_V} \left(\frac{p_1}{4}\right)^{n_S} \left(\frac{p_2}{4}\right)^{n_V} \mathrm{d}\kappa. \tag{1.56}$$

* indicates a more difficult or technical section.

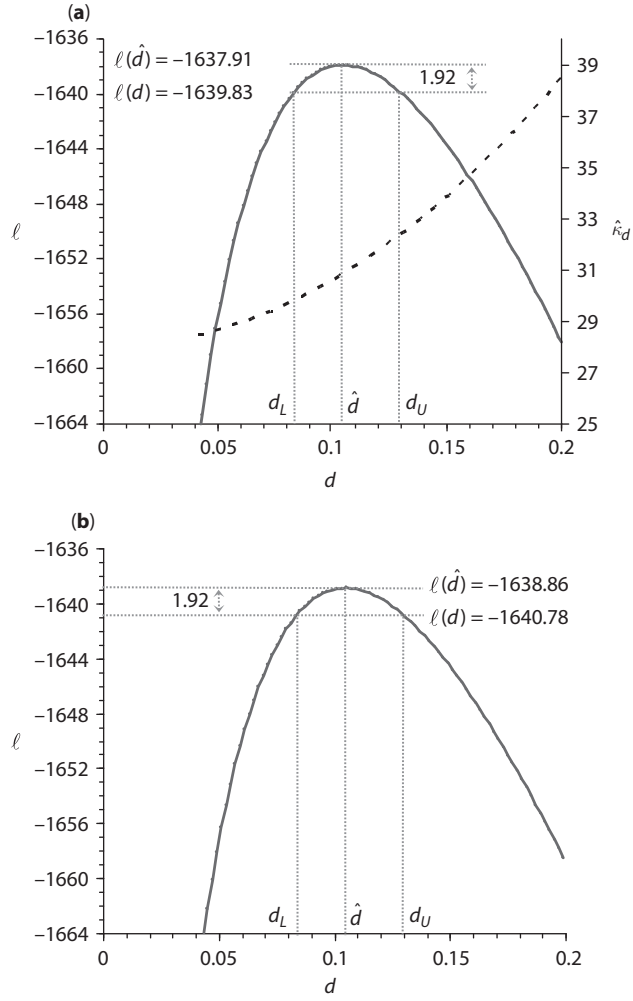

Fig. 1.9 Profile (**a**) and integrated (**b**) log likelihood for distance d under the K80 model. The mitochondrial 12s rRNA genes of Table 1.3 are analysed. (**a**) The profile likelihood $\ell(d) = \ell(d, \hat{\kappa}_d)$ is plotted against d. The estimated nuisance parameter $\hat{\kappa}_d$ at fixed d is also shown. The profile log likelihood is lowered from the peak by 1.921 to construct a likelihood interval for parameter d. (**b**) The integrated likelihood for d is calculated by integrating over the nuisance parameter κ using equation (1.55), with a uniform prior $\kappa \sim U(0, 99)$.

Example 1.9. We apply equation (1.55) to the 12s rRNA data of Table 1.3. We use a uniform prior $\kappa \sim U(0, c)$, with $c = 99$ so that $f(\kappa) = 1/c$ for $0 < \kappa < c$. Analytical calculation of the integral appears awkward, so a numerical method is used instead. The log integrated likelihood $\ell(d) = \log\{L(d)\}$, with $L(d)$ given by equation (1.55), is plotted in Figure 1.9b. This is always lower than the profile log likelihood (Figure 1.9a). The MLE of d is obtained

numerically as $\hat{d} = 0.1048$, with the maximum log likelihood $\ell(\hat{d}) = -1638.86$. By lowering ℓ by 1.921, we construct the likelihood interval for d to be (0.0837, 0.1295). For this example, the profile and integrated likelihood methods produced very similar MLEs and likelihood intervals. □

1.5 Markov chains and distance estimation under general models

We have discussed most of the important properties of continuous-time Markov chains useful in this book. In this section we provide a more systematic overview, and also discuss two general Markov chain models: the general time-reversible (GTR) model and the general unconstrained model. The theory will be applied in a straightforward manner to model substitutions between amino acids and between codons in Chapter 2. Note that Markov chains (processes) are classified according to whether time and state are discrete or continuous. In the Markov chains we consider in this chapter, the states (the four nucleotides) are discrete while time is continuous. In Chapter 7, we will encounter Markov chains with discrete time and either discrete or continuous states. Interested readers should consult a textbook on Markov chains and stochastic processes (e.g. Karlin and Taylor 1975; Grimmett and Stirzaker 1992; Ross 1996; Norris 1997). Note that some authors use the term Markov chain if time is discrete, and Markov process if time is continuous.

1.5.1 *Markov chains*

Let the state of the chain at time t be $X(t)$. This is one of the four nucleotides T, C, A, or G. We assume that different sites in a DNA sequence evolve independently, and the Markov chain model is used to describe nucleotide substitutions at any site. The Markov chain is characterized by its *generator* or the substitution rate matrix $Q = \{q_{ij}\}$, where $q_{ij}, i \neq j$, is the instantaneous rate of change from i to j; that is, $\Pr\{X(t + \Delta t) = j | X(t) = i\} = q_{ij}\Delta t$, for any $j \neq i$. If q_{ij} does not depend on time, as we assume here, the process is said to be *time-homogeneous*. The diagonals q_{ii} are specified by the requirement that each row of Q sums to zero, that is, $q_{ii} = -\sum_{j \neq i} q_{ij}$. Thus $-q_{ii}$ is the substitution rate of nucleotide i, the rate at which the Markov chain leaves state i. The general model without any constraint on the structure of Q will have 12 free parameters.

The dynamics of a Markov chain with only a finite number of states is fully determined by its Q matrix. For example, the Q matrix specifies the *transition probability matrix* over any time $t > 0$: $P(t) = \{p_{ij}(t)\}$, where $p_{ij}(t) = \Pr\{X(t) = j | X(0) = i\}$. Indeed $P(t)$ is the solution to the following differential equation:

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = P(t)Q, \tag{1.57}$$

with the boundary condition $P(0) = I$, the identity matrix (e.g. Grimmett and Stirzaker 1992, p. 242). This has the solution

$$P(t) = \mathrm{e}^{Qt}. \tag{1.58}$$

(e.g. Lang 1987, Chapter 8).

As Q and t occur only in the form of a product, it is conventional to multiply Q by a scale factor so that the average rate is 1. Time t will then be measured by distance, that

is, the expected number of substitutions per site. Thus we use Q to define the relative substitution rates only.

If the Markov chain $X(t)$ has the initial distribution $\pi^{(0)} = (\pi_T^{(0)}, \pi_C^{(0)}, \pi_A^{(0)}, \pi_G^{(0)})$, then time t later the distribution $\pi^{(t)} = (\pi_T^{(t)}, \pi_C^{(t)}, \pi_A^{(t)}, \pi_G^{(t)})$ will be given by

$$\pi^{(t)} = \pi^{(0)} P(t). \tag{1.59}$$

If a long sequence initially has the four nucleotides in proportions $\pi_T^{(0)}, \pi_C^{(0)}, \pi_A^{(0)}, \pi_G^{(0)}$, then time t later the proportions will become $\pi^{(t)}$. For example, consider the frequency of nucleotide T in the target sequence: $\pi_T^{(t)}$. Such a T can result from any nucleotide in the source sequence at time 0. Thus $\pi_T^{(t)} = \pi_T^{(0)} p_{TT}(t) + \pi_C^{(0)} p_{CT}(t) + \pi_A^{(0)} p_{AT}(t) + \pi_G^{(0)} p_{GT}(t)$. Written in matrix notation, this is equation (1.59).

If the initial and target distributions are the same, $\pi^{(0)} = \pi^{(t)}$, the chain will stay in that distribution forever. The chain is then said to be stationary or at equilibrium, and the distribution (let it be π) is called the *stationary* or *steady-state distribution*. Our Markov chain can move from any state to any other state in finite time with positive probability. Such a chain is called *irreducible* and has a unique stationary distribution, which is also the *limiting distribution* when time $t \to \infty$. As indicated above, the stationary distribution is given by

$$\pi P(t) = \pi. \tag{1.60}$$

This is equivalent to

$$\pi Q = 0 \tag{1.61}$$

(e.g. Grimmett and Stirzaker 1992, p. 244). This can also be written as $\sum_i \pi_i q_{ij} = 0$ or $\sum_{i \neq j} \pi_i q_{ij} = -\pi_j q_{jj}$ for any j. The total amount of flow into any state j is $\sum_{i \neq j} \pi_i q_{ij}$, while the total amount of flow out of state j is $-\pi_j q_{jj}$. Equation (1.61) states that the two are equal when π is the stationary distribution. Equation (1.61), together with the obvious constraints $\pi_j \geq 0$ and $\sum_j \pi_j = 1$, allows us to determine the stationary distribution from Q for any Markov chain.

*1.5.2 *Distance under the unrestricted (UNREST) model*

In the most general model of nucleotide substitution, all the non-diagonal elements of the rate matrix Q are free parameters. This model, referred to as UNREST, was implemented by Yang (1994b) for estimating the pattern of nucleotide substitution, in comparison with the GTR (REV) model. The rate matrix Q is shown in Table 1.1. Note that the equilibrium nucleotide frequencies $\{\pi_T, \pi_C, \pi_A, \pi_G\}$ are given by equation (1.61), as functions of the rate parameters in Q; they should not be counted as additional parameters in the model. In this regard, note that the rate matrix for this model is often given incorrectly in the phylogenetics literature (e.g. Swofford et al. 1996, eq. 3). In general Q may have complex eigenvalues and eigenvectors, so its implementation requires care.

We mention here an interesting special case of the UNREST model, the strand-symmetry model, proposed by Sueka (1995) and implemented by Bielawski and Gold (2002; see also Singh et al. 2009). This assumes that the mutation rates are the same on the two strands of the DNA, so that in the comparison of homologous sequences on the same strand, we have, say, $q_{TC} = q_{AG}$. The rate matrix involves six parameters:

* indicates a more difficult or technical section.

$$Q = \begin{bmatrix} -(b+c+e) & b & c & e \\ a & -(a+d+f) & d & f \\ c & e & -(b+c+e) & b \\ d & f & a & -(a+d+f) \end{bmatrix}. \quad (1.62)$$

Perhaps by coincidence, all eigenvalues of this matrix are real: $\lambda_0 = 0, \lambda_1 = -(a+b+d+e)$, and $\lambda_{2,3} = -\frac{1}{2}[a+b+d+e+2c+2f \pm [(a+b+d+e+2c+2f)^2 - 8(ac+ae+cd+bd+bf+2cf)]^{1/2}]$. The equilibrium distribution is given analytically by Singh et al. (2009) as

$$(\pi_T, \pi_C, \pi_A, \pi_G) = \left(\frac{a+d}{2(a+b+d+e)}, \frac{b+c}{2(a+b+d+e)}, \frac{a+d}{2(a+b+d+e)}, \frac{b+c}{2(a+b+d+e)}\right). \quad (1.63)$$

To estimate the sequence distance under the UNREST model, note that the model can in theory identify the root of the two-sequence tree (Figure. 1.10a), so that two branch lengths (t_1 and t_2) are involved. In addition there are 11 relative-rate parameters in the Q matrix (suppose we fix $q_{GA} = l = 1$ in Q, Table 1.1). The likelihood is given by the multinomial probability with 16 cells, corresponding to the 16 possible site patterns. Let $f_{ij}(t_1, t_2, Q)$ be the probability for the ijth cell, that is, the probability that any site has nucleotide i in sequence 1 and j in sequence 2. Since such a site can result from all four possible nucleotides in the ancestor, we have to average over them:

$$f_{ij}(t_1, t_2, Q) = \sum_k \pi_k p_{ki}(t_1) p_{kj}(t_2). \quad (1.64)$$

Here π_k is the equilibrium frequency of nucleotide k, given by equation (1.61), together with the constraint $\sum_j \pi_j = 1$, as a function of Q. Let n_{ij} be the number of sites in the ijth cell. The log likelihood is then

$$\ell(t_1, t_2, Q) = \sum_{i,j} n_{ij} \log\{f_{ij}(t_1, t_2, Q)\}. \quad (1.65)$$

The model involves 13 parameters: 11 relative rates in Q plus two branch lengths t_1 and t_2. There are two problems with this unconstrained model. First, numerical methods are necessary to find the MLEs of parameters as no analytical solution seems possible. The eigenvalues of Q may be complex numbers. Second, and more importantly, typical datasets may not have enough information to estimate so many parameters. In particular, even though t_1 and t_2 are identifiable, their estimates are highly correlated. For this reason the model is not advisable for use in distance calculations.

Example 1.10. For the 12s rRNA data of Table 1.3, the log likelihood appears flat when t_1 and t_2 are estimated as separate parameters. We thus force $t_1 = t_2$ during the numerical

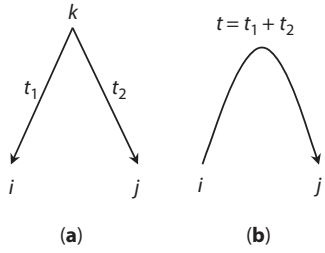

Fig. 1.10 A tree for two sequences, showing the observed nucleotides i and j at one site and the direction of evolution. (**a**) Two sequences diverged from a common ancestor (root of the tree) t_1 and t_2 time units ago; time is measured by the distance or the amount of sequence change. (**b**) Sequence 1 is ancestral to sequence 2. Under time-reversible models, we cannot identify the root of the tree, as the data will look the same whether both sequences were descendants of a common ancestor (as in **a**), or one sequence is ancestral to the other (as in **b**), or wherever the root of the tree is along the single branch connecting the two sequences.

maximization of the log likelihood. The estimate of the sequence distance $t = (t_1 + t_2)$ is 0.1057, very close to estimates under other models (Table 1.4). The MLE of rate matrix Q is

$$Q = \begin{pmatrix} -1.46 & 1.40 & 0.06 & 0.00 \\ 1.16 & -1.22 & 0.06 & 0.00 \\ 0.05 & 0.04 & -0.60 & 0.51 \\ 0.00 & 0.00 & 0.85 & -0.85 \end{pmatrix}, \tag{1.66}$$

scaled so that the average rate is $-\sum_i \pi_i q_{ii} = 1$. The steady-state distribution is calculated from equation (1.61) to be $\hat{\pi} = (0.2184, 0.2606, 0.3265, 0.1946)$, virtually identical to the observed frequencies (Table 1.3). The log likelihood is –1610.36. □

*1.5.3 *Distance under the general time-reversible model*

A Markov chain is said to be time-reversible if and only if

$$\pi_i q_{ij} = \pi_j q_{ji}, \text{ for all } i \neq j. \tag{1.67}$$

Note that π_i is the proportion of time the Markov chain spends in state i, and $\pi_i q_{ij}$ is the amount of 'flow' from states i to j, while $\pi_i q_{ij}$ is the flow in the opposite direction. Equation (1.67) states that the flow between any two states is the same in the opposite directions and is known as the *detailed-balance* condition. There does not appear to be any biological reason to expect the substitution process to be reversible, so reversibility is a mathematical convenience. Models discussed in this chapter, including JC69 (Jukes and Cantor 1969), K80 (Kimura 1980), F84, HKY85 (Hasegawa et al. 1985), and TN93 (Tamura and Nei 1993), are all time-reversible. Equation (1.67) is equivalent to

$$\pi_i p_{ij}(t) = \pi_j p_{ji}(t), \text{ for all } i \neq j \text{ and for any t.} \tag{1.68}$$

Another equivalent condition for reversibility is that the rate matrix can be written as a product of a symmetrical matrix multiplied by a diagonal matrix; the diagonal elements in the diagonal matrix will then specify the equilibrium frequencies. Thus the rate matrix for the GTR model of nucleotide substitution is

$$Q = \{q_{ij}\} = \begin{bmatrix} . & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & . & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & . & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & . \end{bmatrix} = \begin{bmatrix} . & a & b & c \\ a & . & d & e \\ b & d & . & f \\ c & e & f & . \end{bmatrix} \begin{bmatrix} \pi_T & 0 & 0 & 0 \\ 0 & \pi_C & 0 & 0 \\ 0 & 0 & \pi_A & 0 \\ 0 & 0 & 0 & \pi_G \end{bmatrix}, \tag{1.69}$$

with the diagonals of Q given by the requirement that each row of Q sums to 0. This matrix involves nine free parameters: the rates a, b, c, d, e, and f and three frequency parameters. The model was first applied by Tavaré (1986) to sequence distance calculation and by Yang (1994b) to estimation of relative substitution rates (substitution pattern) between nucleotides using ML. It is commonly known as GTR or REV.

Keilson (1979) discussed a number of nice mathematical properties of reversible Markov chains. One of them is that all eigenvalues of the rate matrix Q are real (see §2.6). Thus efficient and stable numerical algorithms can be used to calculate the eigenvalues of Q. Alternatively, it appears possible to diagonalize Q of equation (1.69) analytically: one eigenvalue is 0, so that the characteristic equation that the eigenvalues should satisfy is a cubic equation (e.g. Lang 1987, Chapter 8), which is solvable. Even so, analytical calculation appears too tedious to be practical.

* indicates a more difficult or technical section.

In the phylogenetic analysis of sequence data, reversibility leads to an important simplification to the likelihood function. The probability of observing site pattern *ij* in equation (1.64) becomes

$$\begin{aligned} f_{ij}(t_1, t_2) &= \sum_k \pi_k p_{ki}(t_1) p_{kj}(t_2) \\ &= \sum_k \pi_i p_{ik}(t_1) p_{kj}(t_2) \\ &= \pi_i p_{ij}(t_1 + t_2). \end{aligned} \tag{1.70}$$

The second equality is because of the reversibility condition $\pi_k p_{ki}(t_1) = \pi_i p_{ik}(t_1)$, while the third equality is due to the Chapman–Kolmogorov theorem (equation (1.5)).

Two remarks are in order. First, f_{ij} depends on $t_1 + t_2$ but not on t_1 and t_2 individually; thus we can estimate $t = t_1 + t_2$ but not t_1 and t_2 separately. Equation (1.70) thus becomes

$$f_{ij}(t) = \pi_i p_{ij}(t). \tag{1.71}$$

Second, while we defined f_{ij} as the probability of a site when both sequences are descendants of a common ancestor (Figure 1.10a), $\pi_i p_{ij}(t)$ is the probability of the site if sequence 1 is ancestral to sequence 2 (Figure 1.10b). The probability is the same if we consider sequence 2 as the ancestor of sequence 1, or wherever we place the root along the single branch linking the two sequences. Thus under the model, the log likelihood (1.65) becomes

$$\ell(t, a, b, c, d, e, \pi_T, \pi_C, \pi_A) = \sum_i \sum_j n_{ij} \log\{f_{ij}(t)\} = \sum_i \sum_j n_{ij} \log\{\pi_i p_{ij}(t)\}. \tag{1.72}$$

We use Q to represent the relative rates, with $f = 1$ fixed in Q, and multiply the whole matrix by a scale factor so that the average rate is $-\sum_i \pi_i q_{ii} = 1$. Time t is then the distance: $d = -t \sum_i \pi_i q_{ii} = t$. The model thus involves nine parameters, which can be estimated numerically by solving a nine-dimensional optimization problem. Sometimes the base frequency parameters are estimated using the average observed frequencies, in which case the dimension is reduced to six.

Note that the log likelihood functions under the JC69 and K80 models, that is, equations (1.48) and (1.52), are special cases of equation (1.72). Under these two models, the likelihood equation is analytically tractable, so that numerical optimization is not needed. Equation (1.72) also gives the log likelihood for other reversible models such as F81, HKY85, F84, and TN93. MLEs under those models were obtained through numerical optimization for the 12s rRNA genes of Table 1.3 and listed in Table 1.4. Note that the distance formulae under F81, F84, TN93 etc., discussed in §1.2, are not MLEs, despite claims to the contrary. First, the observed base frequencies are in general not MLEs of the base frequency parameters. Second, all 16 site patterns have distinct probabilities under those models and are not collapsed in the likelihood function (1.72), but collapsed site patterns are used in the distance formulae (such as the constant patterns TT, CC, AA, GG). Nevertheless, it is expected that the distance formulae will give estimates very close to the MLEs (see, e.g. Table 1.4).

Under GTR + Γ, with gamma-distributed rates among sites, the log likelihood is still given by equation (1.72) but with $f_{ij}(t)$ given by equation (1.44). Thus the distance can be estimated by maximizing the likelihood. This is the ML method for distance estimation under the GTR + Γ model described by Gu and Li (1996) and Yang and Kumar (1996).

Besides the ML estimation, a few distance formulae have been suggested in the literature for the GTR and even the UNREST models. We consider the GTR model first. Note that in matrix notation, equation (1.71) becomes

$$F(t) = \{f_{ij}(t)\} = \Pi P(t), \tag{1.73}$$

where $\Pi = \text{diag}\{\pi_T, \pi_C, \pi_A, \pi_G\}$. As $P(t) = e^{Qt}$, we can estimate Qt by

$$\overline{Qt} = \log\{\hat{P}\} = \log\{\hat{\Pi}^{-1}\hat{F}\}. \tag{1.74}$$

where we use the average observed frequencies to estimate Π and use $\hat{f}_{ij} = \hat{f}_{ji} = (n_{ij} + n_{ji})/n$ to estimate the F matrix. The logarithm of $\hat{P}$ is computed by diagonalizing $\hat{P}$. When Q is defined as the relative substitution rates with the average rate to be 1, both t and Q can be recovered from the estimate of Qt. Note that the sequence distance can be defined as $d = -\sum_i \pi_i q_{ii} t = -\text{trace}\{\Pi Q t\}$, where trace $\{A\}$ is the sum of the diagonal elements of matrix A. An estimate is thus:

$$\hat{d} = -\text{trace}\{\hat{\Pi}\ \log(\hat{\Pi}^{-1}\hat{F})\}. \tag{1.75}$$

This approach was first suggested by Tavaré (1986, Equation 3.12), although Rodriguez et al. (1990) were the first to publish equation (1.75). A number of authors (e.g. Gu and Li 1996; Yang and Kumar 1996; Waddell and Steel 1997) apparently rediscovered the distance formula, and also extended the distance to the case of gamma-distributed rates among sites, using the same idea for deriving gamma distances under JC69 and K80 (see §1.3).

The distance (1.75) is inapplicable when any of the eigenvalues of $\hat{P}$ is ≤ 0, which can occur often at high sequence divergences. This is similar to the inapplicability of the JC69 distance when more than 75% of sites are different. As there are nine free parameters in the model and also nine free observables in the symmetrical matrix $\hat{F}$, the invariance property of MLEs suggests that equation (1.75), if applicable, should give the MLEs.

Next we describe a distance suggested by Barry and Hartigan (1987a), which works without the reversibility assumption and even without assuming a stationary model:

$$\hat{d} = -\frac{1}{4}\log\{\text{Det}(\hat{\Pi}^{-1}\hat{F})\}, \tag{1.76}$$

where Det(A) is the determinant of matrix A, which is equal to the product of the eigenvalues of A. The distance is inapplicable when the determinant is ≤ 0 or when any of the eigenvalues of $\hat{\Pi}^{-1}\hat{F}$ is ≤ 0. Barry and Hartigan (1987a) referred to equation (1.76) as the *asynchronous distance*. It is now commonly known as the *Log-Det distance*.

Let us consider the behaviour of the distance under simpler stationary models in very long sequences. In such a case, $\hat{\Pi}^{-1}\hat{F}$ will approach the transition probability matrix $P(t)$, and its determinant will approach $\exp\left(\sum_k \lambda_k t\right)$, where the λ_ks are the eigenvalue of the rate matrix Q (see equation (1.18)). Thus $\hat{d}$ in equation (1.76) will approach $-\frac{1}{4}\sum_k \lambda_k t$. For the K80 model, the eigenvalues of the rate matrix (1.9) are $\lambda_1 = 0, \lambda_2 = -4\beta, \lambda_3 = \lambda_4 = -2(\alpha + \beta)$, so that $\hat{d}$ approaches $(\alpha + 2\beta)t$, which is the correct sequence distance. Obviously this will hold true for the simpler JC69 model as well. However, for more complex models with unequal base frequencies, $\hat{d}$ of equation (1.76) does not estimate the correct distance, even though it grows linearly with time. For example, under the TN93 model, $\hat{d}$ approaches $\frac{1}{4}(\pi_Y\alpha_1 + \pi_R\alpha_2 + 2\beta)t$.

Barry and Hartigan (1987a) defined $\hat{f}_{ij} = n_{ij}/n$, so that $\hat{F}$ is not symmetrical, and interpreted $\hat{\Pi}^{-1}\hat{F}$ as an estimate of $P_{12}(t)$, the matrix of transition probabilities from sequences 1 to 2. The authors argued that the distance should work even if the substitution process is not homogeneous or stationary, that is, if there is systematic drift in base compositions during the evolutionary process. Evidence for the performance of the distance when different sequences have different base compositions is mixed. The distance appears to have acquired a paranormal status when it was rediscovered or modified by Lake (1994), Steel

(1994b), Zharkikh (1994), among others. For a more recent discussion of the distance, see Massingham and Goldman (2007).

1.6 Discussions

1.6.1 *Distance estimation under different substitution models*

One might expect more complex models to be more realistic and to produce more reliable distance estimates. However, the situation is more complex. At small distances, the different assumptions about the structure of the Q matrix do not make much difference, and simple models such as JC69 and K80 produce very similar estimates to those under more complex models. The two 12s rRNA genes analysed in this chapter are different at about 10% of the sites. The different distance formulae produced virtually identical estimates, all between 0.10 and 0.11 (Table 1.4). This is the case despite the fact that simple models like JC69 are grossly wrong, judged by the log likelihood values achieved by the models (see §1.4.3 and Problem 1.7). The rate variation among sites has much more impact, as seen in §1.3.

At intermediate distances, for example, when the sequences are about 20% or 30% different, model assumptions become more important. It may be favourable to use realistic models for distance estimation, especially if the sequences are not short. At large distances, for example, when the sequences are >40% different, the different methods often produce very different estimates, and the estimates, especially those under more complex models, involve large sampling errors. Sometimes the distance estimates become infinite or the distance formulae become inapplicable. This happens far more often under more complex models than under simpler models. In such cases, a useful approach is to add more sequences to break down the long distances and to use a likelihood-based approach to compare all sequences jointly on a phylogeny.

1.6.2 *Limitations of pairwise comparison*

If there are only two sequences in the whole dataset, pairwise comparison is all we can do. If we have multiple sequences, however, pairwise comparison may be hampered as it ignores the other sequences, which should also provide information about the relatedness of the two sequences being compared. Here, brief comment will be made on two obvious limitations of the pairwise approach. The first is the lack of internal consistency. Suppose we use the K80 model for pairwise comparison of three sequences: a, b, and c. 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 multiple sequences are compared 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 Problems

1.1 Use the transition probabilities under the JC69 model (equation (1.4)) to confirm the Chapman–Kolmogorov theorem (equation (1.5)). 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 $p_{TT}(t_1 + t_2) = p_{TT}(t_1)p_{TT}(t_2) + p_{TC}(t_1)p_{CT}(t_2) + p_{TA}(t_1)p_{AT}(t_2) + p_{TG}(t_1)p_{GT}(t_2)$.

1.2 Derive the transition probability matrix $P(t) = \mathrm{e}^{Qt}$ for the JC69 model. Set $\pi_T = \pi_C = \pi_A = \pi_G = 1/4$ and $\alpha_1 = \alpha_2 = \beta$ in the rate matrix (1.16) for the TN93 model to obtain the eigenvalues and eigenvectors of Q under JC69, using results of §1.2.3. Alternatively you can derive the eigenvalues and eigenvectors from equation (1.1) directly. Then apply equation (1.18).

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

$$Q = U\Lambda U^{-1} = \begin{bmatrix} 1 & -u \\ 1 & v \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & -u-v \end{bmatrix} \begin{bmatrix} \frac{v}{u+v} & \frac{u}{u+v} \\ -\frac{1}{u+v} & \frac{1}{u+v} \end{bmatrix}, \tag{1.77}$$

so that

$$P(t) = \mathrm{e}^{Qt} = U\mathrm{e}^{\Lambda t}U^{-1} = \frac{1}{u+v} \begin{bmatrix} v + u\mathrm{e}^{-(u+v)t} & u - u\mathrm{e}^{-(u+v)t} \\ v - v\mathrm{e}^{-(u+v)t} & u + v\mathrm{e}^{-(u+v)t} \end{bmatrix}. \tag{1.78}$$

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

$$P(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2t} & \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-2t} \\ \frac{1}{2} - \frac{1}{2}\mathrm{e}^{-2t} & \frac{1}{2} + \frac{1}{2}\mathrm{e}^{-2t} \end{bmatrix}. \tag{1.79}$$

This is the binary equivalent of the JC69 model.

1.4 Confirm that the two likelihood functions for the JC69 model, equations (1.46) and (1.47), are proportional and the proportionality factor is a function of n and x but not of d. Confirm that the likelihood equation, $\frac{\mathrm{d}\ell}{\mathrm{d}d} = \frac{\mathrm{d}\log\{L(d)\}}{\mathrm{d}d} = 0$, is the same whichever of the two likelihood functions is used.

1.5 Derive the equilibrium nucleotide frequencies for the K80 model. Solve the system of linear equations generated by equation (1.61) and the constraint $\sum_j \pi_j = 1$.

1.6 A large genomic region evolves neutrally according to the JC69 model, at the rate 2×10^{-8} substitutions/site/year (this is roughly the rate in the mitochondria in mammals). (**a**) Suppose initially the sequence consists of Ts only. What will be the proportions of T, C, A, and G in the sequence 10^6 and 10^8 years later? (**b**) Do the same calculation assuming that the sequence initially had Cs only. (**c**) Do the same calculation if the initial proportions of T, C, A, and G are $\pi_0 = (0.4,\ 0.3,\ 0.2,\ 0.1)$.

1.7 Use the 12s rRNA data of Table 1.3 to conduct the LRT to compare K80 against HKY85, and HKY85 against GTR. The numbers of parameters under the models are listed in Table 1.1, and the log likelihood values are listed in Table 1.4, but you may prefer running a program (such as BASEML in the PAML package, Yang 2007b) to do the calculation yourself.

1.8 Use the protein-coding DNA sequences from the human and gibbon mitochondrial genomes to calculate the sequence distance at the three codon positions. Download the sequences from GenBank (accession numbers X93334 for *Homo sapiens* and X99256 for *Hylobates lar*) and concatenate the 12 protein-coding genes encoded on the same H strand of the genome (that is, excluding NADH6, which is coded on the other strand with very different base compositions). Align the sequences using an alignment program such as CLUSTAL or PRANK and make manual adjustments if necessary. Separate the three codon positions into three independent datasets, and calculate the distance between the two species under various substitution models: JC69, K80, GTR, and JC69 + Γ_5, K80 + Γ_5, and GTR + Γ_5 (with $\alpha = 0.5$ fixed for the gamma models). Discuss the impact of model assumptions on distance estimation.

1.9* 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

$$f(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}. \tag{1.80}$$

(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

$$f(x|\theta) = \binom{r+x-1}{x} \theta^x (1-\theta)^{n-x}. \tag{1.81}$$

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

* indicates a more difficult or technical problem.