from: Computational Molecular Evolution
Ziheng Yang (2006) Oxford.
ISBN: 0198567022

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 analysis of multiple sequences on a phylogeny. Indeed, joint multiple sequence analysis under the same model 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 substitution in a DNA sequence. It also introduces the method of maximum likelihood (ML). Bayesian estimation of pairwise distances and Bayesian phylogenetics are introduced in Chapter 5.

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 in the two sequences, may harbour back or parallel substitutions (Fig. 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. Continuous-time Markov chains are commonly used for this purpose. The nucleotide sites in the sequence are normally 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

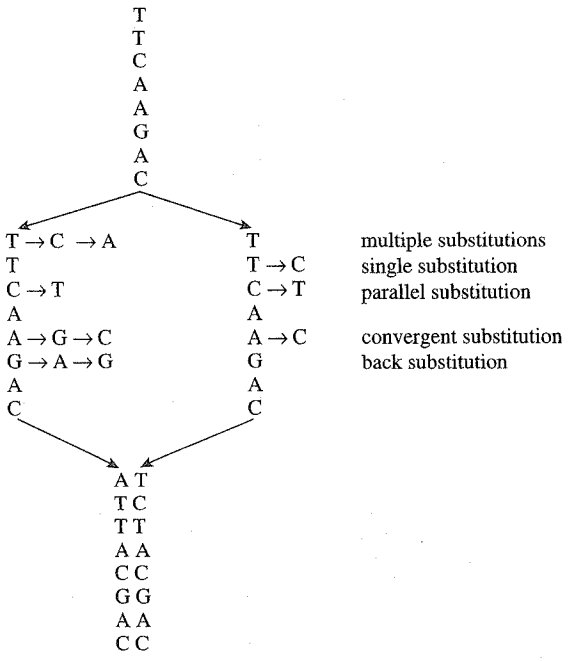


Fig. 1.1 Illustration of multiple substitutions at the same site or multiple hits. An ancestral sequence diverged into two sequences and has since accumulated nucleotide substitutions independently along the two lineages. Only two *differences* are observed between the two present-day sequences, so that 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).

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

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 rate λ of changing into any other nucleotide. We use q_{ij} to denote the instantaneous

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

	From	To			
		T	C	A	G
JC69 (Jukes and Cantor 1969)	T	·	λ	λ	λ
	C	λ	·	λ	λ
	A	λ	λ	·	λ
	G	λ	λ	λ	·
K80 (Kimura 1980)	T	·	α	β	β
	C	α	·	β	β
	A	β	β	·	α
	G	β	β	α	·
F81 (Felsenstein 1981)	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)	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)	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	$\beta\pi_T$	$\beta\pi_C$	·	$(1+\kappa/\pi_R)\beta\pi$
	G	$\beta\pi_T$	$\beta\pi_C$	$(1+\kappa/\pi_R)\beta\pi_A$	·
TN93 (Tamura and Nei 1993)	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 1994*b*; Zharkikh 1994)	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 1994*b*)	T	·	q_{TC}	q_{TA}	q_{TG}
	C	q_{CT}	·	q_{CA}	q_{CG}
	A	q_{AT}	q_{AC}	·	q_{AG}
	G	q_{GT}	q_{GC}	q_{GA}	·

The diagonals of the matrix are determined by the requirement that each row sums to 0. The equilibriu distribution is $\pi = (1/4, 1/4, 1/4, 1/4)$ under JC69 and K80, and $\pi = (\pi_T, \pi_C, \pi_A, \pi_G)$ under F81, F8 HKY85, TN93, and GTR. Under the general unrestricted (UNREST) model, it is given by the equatio $\pi Q = 0$ under the constraint $\sum_i \pi_i = 1$.

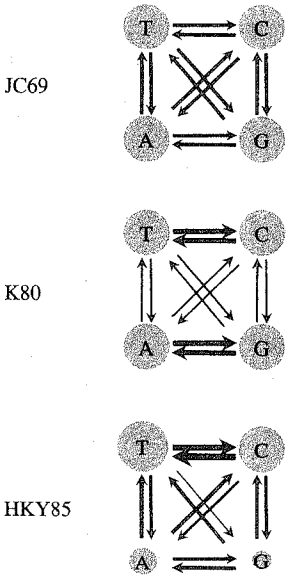


Fig. 1.2 Relative substitution rates between nucleotides under three Markov-chain models of nucleotide substitution: JC69 (Jukes and Cantor 1969), K80 (Kimura 1980), and HKY85 (Hasegawa *et al.* 1985). The thickness of the lines represents the substitution rates while the sizes of the circles represent the steady-state distribution.

rate of substitution from nucleotide i to nucleotide 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. Each row of the matrix sums to 0. The total rate of substitution of any nucleotide i is 3λ, which is $-q_{ii}$.

Note that $q_{ij}\Delta t$ gives the probability that any given nucleotide i will change to a different nucleotide j in an infinitely small time interval Δt. To characterize the Markov chain, we need a similar probability over any time $t > 0$. This is the *transition probability*; $p_{ij}(t)$ is the probability that a given nucleotide i will become j time t later. The matrix $P(t) = \{p_{ij}(t)\}$ is known as the *transition-probability matrix*. As will be discussed later in Section 1.5,

$$P(t) = e^{Qt}. \tag{1.2}$$

Calculation of this matrix exponential is discussed later. For the moment, we simply give the solution

$$P(t) = e^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix}, \quad \text{with} \begin{cases} p_0(t) = \frac{1}{4} + \frac{3}{4}e^{-4\lambda t}, \\ p_1(t) = \frac{1}{4} - \frac{1}{4}e^{-4\lambda t}. \end{cases} \tag{1.3}$$

Imagine a long sequence with nucleotide i at every site; we 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 Fig. 1.3. A few features of the matrix are worth noting. First, every row of $P(t)$ sums to 1, because the chain has to be in one of the four nucleotide states at time t. 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 a product λt. Thus if we are given a source sequence and a target sequence, it will be impossible to tell whether one sequence has evolved into the other 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 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

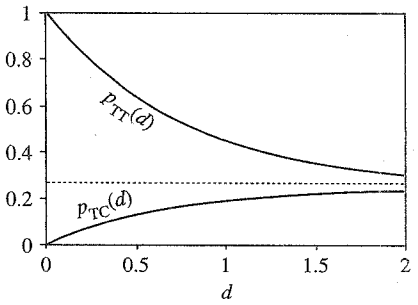


Fig. 1.3 Transition probabilities under the JC69 model (equation 1.3) plotted against distance $d = 3\lambda t$, measured in the expected number of substitutions per site.

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 the limiting distribution (1/4, 1/4, 1/4, 1/4), as the sequence evolves. If the sequence starts with equal proportions of the four nucleotides, the sequence 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 in phylogenetic analysis and is clearly 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 Fig. 1.1? This is achieved through the calculation of the transition probabilities using equation (1.2), which accommodates all the 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 theorem (e.g. Grimmett and Stirzaker 1992, p. 239)

$$p_{ij}(t_1 + t_2) = \sum_k p_{ik}(t_1)p_{kj}(t_2). \tag{1.4}$$

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 (Fig. 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 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 1.3), this is

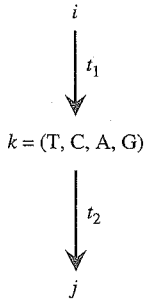


Fig. 1.4 Illustration of the Chapman–Kolmogorov theorem. The transition probability from any nucleotide i to any nucleotide j over time $t_1 + t_2$ is a sum over all possible states k at any intermediate time point t_1.

equivalent to considering the two sequences as descendants of an extinct common ancestor. From equation (1.3), the probability that the nucleotide in the descendant sequence is different from the nucleotide in the ancestral sequence is

$$p = 3p_1(t) = \tfrac{3}{4} - \tfrac{3}{4}e^{-4\lambda t} = \tfrac{3}{4} - \tfrac{3}{4}e^{-4d/3}. \tag{1.5}$$

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

$$\hat{d} = -\tfrac{3}{4}\log\left(1 - \tfrac{4}{3}\hat{p}\right), \tag{1.6}$$

where the base of the logarithm is the constant e. If $\hat{p} > 3/4$, the distance formula will be inapplicable; two random sequences should have about 75% different sites, and if $\hat{p} > 3/4$, the estimated distance 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

$$\mathrm{var}(\hat{d}) = \mathrm{var}(\hat{p}) \times \left|\frac{\mathrm{d}\hat{d}}{\mathrm{d}\hat{p}}\right|^2 = \frac{\hat{p}(1-\hat{p})}{n} \times \frac{1}{(1 - 4\hat{p}/3)^2} \tag{1.7}$$

(Kimura and Ohta 1972).

Example. Consider the sequences of human and orangutan 12s rRNA genes from the mitochondrial genome, summarized in Table 1.2. From the table, $x = 90$ out of the $n = 948$ sites are different, so that $\hat{p} = x/n = 0.09494$. By equation (1.6), $\hat{d} = 0.1015$. Equation (1.7) gives the variance of $\hat{d}$ as 0.0001188 and standard error 0.0109. The approximate 95% confidence interval is thus $0.1015 \pm 1.96 \times 0.0109$ or $(0.0801, 0.1229)$. □

Table 1.2 Numbers and frequencies (in parentheses) of sites for the 16 site configurations (patterns) in human and orangutan mitochondrial 12s rRNA genes

	Human				Sum (π_i)
Orangutan	T	C	A	G	
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

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).