4. Distance Estimation

The evolutionary distance between a pair of sequences is usually measured by the number of nucleotide or amino acid substitutions between them. Evolutionary distances are fundamental for the study of molecular evolution and are useful for phylogenetic reconstruction and estimation of divergence times. In MEGA, most of the widely used methods for distance estimation for nucleotide and amino acid sequences are included. In the following, they are presented in three sections: nucleotide substitutions, synonymous-nonsynonymous substitutions, and amino acid substitutions. For advice in the use of these methods, see Guidelines for Choosing Distance Measures in section 4.4. The treatment of alignment gaps and missing-information sites in distance computation is explained in section 4.5.

4.1Nucleotide Substitutions

The evolutionary distances that are computed from DNA sequence data are primarily estimates of the number of nucleotide substitutions per site (d) between two sequences. There are many methods for estimating evolutionary distances, depending on the pattern of nucleotide substitutions (see Nei 1987, Gojobori *et al*. 1990, Saccone *et al*. 1990, and others). Here we have included only methods that are relatively simple and frequently used by molecular evolutionists. Two methods, i.e., the Tamura and Tamura-Nei methods, are new and their utility has not been well tested, but they are included here because they seem to be useful for analyzing mitochondrial DNA data, which are now often used for phylogenetic inference. In the following we first present the simplest method and then discuss gradually more complicated ones.

p-distance

This distance is merely the proportion (p) of nucleotide sites at which the two sequences compared are different. This is obtained by dividing the number of nucleotide differences (n_d) by the total number of nucleotides compared (n). Thus,

$$p = n_d/n. \tag{4.1}$$

The variance of p is given by

$$V(p) =[p(1 -p)]/n. \tag{4.2}$$

The p-distance is approximately equal to the number of nucleotide substitutions per site (d) only when it is small, say $p < 0.1$. However, the computation of this distance is simple, and for constructing phylogenetic trees it gives essentially the same results as the more complicated distance measures mentioned below, as long as all pairwise distances are small. Actually, when the rate of nucleotide substitution is the same for all evolutionary lineages, the p-distance gives the correct topology slightly more often than the Jukes-Cantor and Kimura distances mentioned below, because it has a smaller variance (Saitou and Nei 1987, Saitou and Imanishi 1989, Schöniger and von Haeseler 1993, Tajima and Takezaki, 1994). Of course, for estimating the divergence times of two sequences, this is not a good measure.

Under certain circumstances, one may want to compute the proportion of sites with transitional and transversional nucleotide differences. In MEGA, the proportions of transitional differences (P) and transversional differences (Q) are computed by

$$P = n_s/n \text{ and } Q = n_v/n, \qquad (4.3)$$

respectively, where ns and n_v are the numbers of transitional and transversional differences between the two sequences, with $n_s + n_v = n_d$. The variances of P and Q are computed by equations analogous to (4.2). In addition, the ratio of transitional to transversional differences (R_d) and its variance are given by

$$R_d = P/Q, \qquad (4.4)$$

$$V(R_d) = [c_1^2P + c_2^2Q - (c_1P + c_2Q)^2]/n, \qquad (4.5)$$

where $c_1 = 1/Q$ and $c_2 = - P/Q^2$.

Jukes-Cantor distance

This method (Jukes and Cantor 1969) was developed under the assumption that the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G (see Table 4.1), and it gives a maximum likelihood estimate of the number of nucleotide substitutions (d) between two sequences. It is given by

$$d = -3\log_e(1 - 4p/3)/4 , \qquad (4.6)$$

where p is computed by using equation (4.1). The variance of this estimate is given by

$$V(d) = p(1 -p)/[(1 - 4p/3)^2n] \qquad (4.7)$$

(Kimura and Ohta 1972).

The Jukes-Cantor distance can be computed if $p < 0.75$; otherwise it is not applicable because the argument of the logarithm becomes negative. This distance gives a good estimate of the number of nucleotide substitutions if (1) the frequency of each nucleotide is close to 0.25, (2) there is no transition/transversion bias (*i.e.*, the transition/transversion ratio is nearly equal to 0.5), and (3) d is not very large (say $d < 1.0$). However, when the number of nucleotides examined is small, say $n < 100$, the Jukes-Cantor distance tends to give overestimates of the true number of nucleotide substitutions (Tajima 1993).

Table 4.1 Models of nucleotide substitution

Original	Mutant			
	A	T	C	G
A. Jukes-Cantor model				
A	-	λ	λ	λ
T	λ	-	λ	λ
C	λ	λ	-	λ
G	λ	λ	λ	-

	λ is the rate of substitution.			
B. Tajima-Nei model				
A	-	β	γ	δ
T	α	-	γ	δ
C	α	β	-	δ
G	α	β	γ	-
	α, β, γ, and δ are the rates of substitution.			
C. Kimura 2-parameter model				
A	-	β	β	α
T	β	-	α	β
C	β	α	-	β
G	α	β	β	-
	α and β are the rates of transitional and transversional substitution, respectively.			
D. Tamura model				
A	-	$(1-\theta)\beta$	$\theta\beta$	$\theta\alpha$
T	$(1-\theta)\beta$	-	$\theta\alpha$	$\theta\beta$
C	$(1-\theta)\beta$	$(1-\theta)\alpha$	-	$\theta\beta$
G	$(1-\theta)\alpha$	$(1-\theta)\beta$	$\theta\beta$	-
	α and β are the rates of transitional and transversional substitution, respectively, and θ is the G+C content.			
E. Hasegawa *et al*. model				
A	-	$g_T\beta$	$g_C\beta$	$g_G\alpha$
T	$g_A\beta$	-	$g_C\alpha$	$g_G\beta$
C	$g_A\beta$	$g_T\alpha$	-	$g_G\beta$
G	$g_A\alpha$	$g_T\beta$	$g_C\beta$	-
	α and β are the rates of transitional and transversional substitution, respectively, and g_i denotes nucleotide frequencies (i = A, T, C, G).			
F. Tamura-Nei model				
A	-	$g_T\beta$	$g_C\beta$	$g_G\alpha_1$
T	$g_A\beta$	-	$g_C\alpha_2$	$g_G\beta$
C	$g_A\beta$	$g_T\alpha_2$	-	$g_G\beta$

G	$g_A\alpha_1$	$g_T\beta$	$g_C\beta$	-
	α_1 and α_2 are the rates of transitional substitution between purines and between pyrimidines, respectively, β is the rate of transversional substitution, and g_i denotes nucleotide frequencies (i = A, T, C, G).			

Tajima-Nei distance

In real data, nucleotide frequencies often deviate substantially from 0.25. In this case the Tajima-Nei distance (Tajima and Nei 1984) gives a better estimate of the number of nucleotide substitutions than the Jukes-Cantor distance. This estimator is based on the equal-input model of Nei and Tajima (1981) (see Table 4.1). The estimator (*d*) and its variance [V(*d*)] are given by the following equations.

$$d = -b \log_e(1 - p/b), \tag{4.8}$$

$$V(d) = p(1 - p)/[(1 - p/b)^2 n], \tag{4.9}$$

where

$$b = \frac{1}{2}(1 - \sum_{i=1}^{4} g_i^{\,2} + p^2 / c)$$

$$c = \sum_{i=1}^{3} \sum_{j=i+1}^{4} \frac{x_{ij}}{2 g_i g_j}$$

Here, g_i and g_j are the frequencies of the ith and jth nucleotides, respectively (i, j = A, T, C, G), and x_{ij} is the relative frequency of nucleotide pair *i* and *j*. Computer simulations have shown that this estimate is quite robust and is applicable to a wide variety of cases unless the number of nucleotide substitutions is very large, say more than 1.0 per site.

Kimura 2-parameter distance

In actual sequence data the rate of transitional nucleotide substitution is often higher than that of transversional substitution. This is particularly so for animal mitochondrial DNA (Brown *et al.* 1982). In this case, the Jukes-Cantor distance is expected to give an underestimate of *d* unless *d* is quite small, say $d < 0.1$. A maximum likelihood estimate of d for this case is given by Kimura's (1980) 2-parameter method (Table 4.1). This estimate and its variance are given by

$$d = -\log_e(1 - 2P - Q)/2 - \log_e(1 - 2Q)/4, \tag{4.10}$$

$$V(d) = [c_1^2 P + c_3^2 Q - (c_1 P + c_3 Q)^2]/n, \tag{4.11}$$

where $c_1 = 1/(1 - 2P - Q)$, $c_2 = 1/(1 - 2Q)$, and $c_3 = (c_1 + c_2)/2$.

With Kimura's model, it is possible to compute the numbers of transitional (*s*) and transversional (*v*) nucleotide substitutions per site and their variances.

Transitional substitutions:

$$s = -\log_e(1 - 2P - Q)/2 + \log_e(1 - 2Q)/4, \tag{4.12}$$

$$V(s) = [c_1^2P + c_4^2Q - (c_1P + c_4Q)^2]/n, \tag{4.13}$$

where $c_4 = (c_1 - c_2)/2$.

Transversional substitutions:

$$v = -\log_e(1 - 2Q)/2, \tag{4.14}$$

$$V(v) = c_2^2Q(1 - Q)/n. \tag{4.15}$$

Transition/transversion ratio (R = s/v):

The ratio (R) of the number of transitional substitutions (s) to that of transversional substitutions (v) is called the transition/transversion ratio. Note that R is different from R_d defined earlier. In the present case, R and its variance [V(R)] are given by

$$R = \log_e(1 - 2P - Q)/\log_e(1 - 2Q) -1/2, \tag{4.16}$$

$$V(R) = [c_5^2P + c_6^2Q - (c_5P + c_6Q)^2]/n, \tag{4.17}$$

where $c_5 = -2c_1/\log_e(1 - 2Q)$ and $c_6 = (c_5 + 4c_2\log_e(1 - 2P - Q)/[\log_e(1 - 2Q)]^2)/2$.

Tamura distance

Kimura's 2-parameter distance is based on the assumption that the nucleotide frequencies are all equal to 0.25 throughout the evolutionary process. In practice, however, this assumption rarely holds. In particular, the G+C content of *Drosophila* mitochondrial DNA is much lower than 0.5. Tamura (1992) developed a maximum likelihood estimator of d, which is suitable for this case (Table 4. 1). The estimator and its variance are given by

$$d = -2\theta(1 - \theta)\log_e(1 - P/[2\theta(1 - \theta)] - Q) -[1 -2\theta(1 -\theta)]\log_e(1 - 2Q)/2, \tag{4.18}$$

$$V(d) = [c_12P + c_32Q - (c_1P + c_3Q)2]/n, \tag{4.19}$$

where $c_1 = 1/(1 - P/[2\theta(1 - \theta)] - Q)$, $c_2 = 1/(1 - 2Q)$, $c_3 = 2\theta(1 - \theta)(c_1 - c_2) + c_2$, and θ is the G+C content.

The estimates of the numbers of transitional (s) and transversional (v) substitutions per site are obtained by the following equations.

Transitional substitutions:

$$s = -2\theta(1 - \theta)\log_e(1 - P/[2\theta(1 - \theta)] - Q) + \theta(1 - \theta)\log_e(1 - 2Q), \tag{4.20}$$

$$V(s) = [c_1^2P + c_4^2Q - (c_1P + c_4Q)^2]/n, \tag{4.21}$$

where $c_4 = 2\theta(1 - \theta)(c_1 - c_2)$.

Transversional substitutions:

$$v = -\log_e(1 - 2Q)/2, \tag{4.22}$$

$$V(v) = c_2^2Q(1 - Q)/n. \tag{4.23}$$

Transition/transversion ratio (R = s/v):

$$R = 4\theta(1 - \theta)\log_e(1 - P/[2\theta(1 - \theta)] - Q)/\log_e(1 - 2Q) - 2\theta(1 - \theta), \tag{4.24}$$

$$V(R) = [c_5^2P + c_6^2Q - (c_5P + c_6Q)^2]/n, \tag{4.25}$$

where $c_5 = -2c_1/\log_e(1 - 2Q)$ and $c_6 = 2\theta(1 - \theta)\{c_5 + 4c_2\log_e(1 - P/[2\theta(1 - \theta)] - Q)/[\log_e(1 - 2Q)]^2\}$.

In MEGA, the average G+C content for the pair of sequences compared is used for θ. Therefore, different pairwise comparisons may have different values of θ. There are other ways of computing θ, but the distance estimates obtained are usually very similar.

Tamura-Nei distance

One of the useful mathematical models for analyzing mitochondrial DNA is that of Hasegawa *et al.*'s (1985). This model (Table 4.1) has been used for phylogenetic inference by the maximum likelihood method. However, no analytical formula for estimating *d* has been derived for this model.

Tamura and Nei (1993) noted that model F in Table 4.1 is more realistic than model E. In model E, $\alpha_1 = \alpha_2$ is assumed, but actual data indicates that the rates of transitional substitution between purines (A and G) and between pyrimidines (T and C) are often different. For model F, Tamura and Nei (1993) derived the following formula for estimating *d*.

$$\begin{aligned} d = & -\frac{2g_Ag_G}{g_R}\log_e(1 - \frac{g_R}{2g_Ag_G}P_1 - \frac{1}{2g_R}Q) \\ & -\frac{2g_Tg_C}{g_Y}\log_e(1 - \frac{g_Y}{2g_Tg_C}P_2 - \frac{1}{2g_Y}Q) \\ & -2(g_Rg_Y - \frac{g_Ag_Gg_Y}{g_R} - \frac{g_Tg_Cg_R}{g_Y})\log_e(1 - \frac{1}{2g_Rg_Y}Q) \end{aligned} \tag{4.26}$$

where P_1 and P_2 are the proportions of transitional differences between A and G and between T and C, respectively, and Q is the proportion of transversional differences.

They also derived the variance of *d*, but we are not going to present it here because it is somewhat complicated. The computation of the variance is included in the computer program.

The estimates of the numbers of transitional (*s*) and transversional (*v*) substitutions per site are obtained by the following equations.

Transitional substitutions:

$$\begin{aligned} s = & -\frac{2g_A g_G}{g_R}\log_e\left(1-\frac{g_R}{2g_A g_G}P_1-\frac{1}{2g_R}Q\right) \\ & -\frac{2g_T g_C}{g_Y}\log_e\left(1-\frac{g_Y}{2g_T g_C}P_2-\frac{1}{2g_Y}Q\right) \\ & +2\left(\frac{g_A g_G g_Y}{g_R}-\frac{g_T g_C g_R}{g_Y}\right)\log_e\left(1-\frac{1}{2g_R g_Y}Q\right) \end{aligned} \qquad (4.27)$$

Transversional substitutions:

$$v = -2g_R g_Y \log_e\left(1-\frac{1}{2g_R g_Y}Q\right) \qquad (4.28)$$

The transition/transversion ratio is given by $R = s/v$. The computation of this ratio and its variance as well as the variances of *s* and *v* is included in MEGA.

Gamma distances

In the distance measures discussed so far, the rate of nucleotide substitution is assumed to be the same for all nucleotide sites. In actual data this assumption rarely holds, and the rate varies from site to site. Statistical analyses of the distribution of the number of substitutions at different sites have suggested that the rate varies approximately according to the gamma distribution (Uzzell and Corbin 1971, Kocher and Wilson 1991, Tamura and Nei 1993, Wakeley 1993). The gamma distribution can be specified by the parameter *a*, which is the inverse of the coefficient of variation of the substitution rate (λ). The smaller the parameter *a*, the higher the extent of variation in λ. In one hypervariable segment of the control region of mitochondrial DNA, *a* has been estimated to be 0.47 (Wakeley 1993), whereas Uzzell and Corbin (1971) obtained $a = 2$ for amino acid sequence data for cytochrome c.

In the following gamma distances the rate of nucleotide substitution is assumed to follow the gamma distribution specified by parameter a. They are due to Jin and Nei (1990) and Tamura and Nei (1993). The default option of MEGA assumes $a = 1.0$ for nucleotide substitution, except for the gamma distance for the Tamura-Nei model. When a is small ($a < 1$) and the number of nucleotides examined is small ($n < 100$), the following formula tends to give underestimates of the true number of nucleotide substitutions (Rzhetsky and Nei 1994). It is therefore important to use a large number of nucleotides.

Gamma distance for the Jukes-Cantor model:

When the rate of substitution in the Jukes-Cantor model varies with the gamma distribution, the gamma distance and its variance are given by

$$d = 3a[(1 - 4p/3)^{-1/a} - 1]/4, \tag{4.29}$$

$$\mathrm{V}(d) = p(1 - p)[(1 - 4p/3)^{-2(1/a + 1)}]/n. \tag{4.30}$$

Gamma distance for the Kimura's 2-parameter model:

In this case the gamma distance and its variance are given by

$$d = (a/2)[(1 - 2P - Q)^{-1/a} + (1/2)(1 - 2Q)^{-1/a} - 3/2], \tag{4.31}$$

$$\mathrm{V}(d) = (\mathrm{c}_1^2 P + \mathrm{c}_3^2 Q - (\mathrm{c}_1 P + \mathrm{c}_3 Q)^2)/n, \tag{4.32}$$

where $\mathrm{c}_1 = (1 - 2P - Q)^{-(1/a + 1)}$, $\mathrm{c}_2 = (1 - 2Q)^{-(1/a + 1)}$, $\mathrm{c}_3 = (\mathrm{c}_1 + \mathrm{c}_2)/2$, and P and Q are the same as those of the Kimura 2-parameter model.

Transitional substitutions:

$$s = (a/2)[(1- 2P - Q)^{-1/a} -(1/2)(1 - 2Q)^{-1/a} -1/2], \tag{4.33}$$

$$\mathrm{V}(s) = (\mathrm{c}_1^2 P + \mathrm{c}_4^2 Q - (\mathrm{c}_1 P + \mathrm{c}_4 Q)^2)/n, \tag{4.34}$$

where $\mathrm{c}_4 = (\mathrm{c}_1 - \mathrm{c}_2)/2$.

Transversional substitutions:

$$v = (a/2)[(1 - 2Q)^{-1/a} -1], \tag{4.35}$$

$$\mathrm{V}(v) = \mathrm{c}_2^2 Q(1 - Q)/n. \tag{4.36}$$

Transition/Transversion ratio (R = s/ v):

$$R = [(1 - 2P - Q)^{-1/a} -(1/2)(1 - 2Q)^{-1/a} - 1/2]/[(1 - 2Q)^{-1/\mathrm{a}} -1]. \tag{4.37}$$

The formula for the variance of *R* is rather complicated and is not presented here, but it is computed in MEGA.

Gamma distance for the Tamura-Nei model:

In the control region of mammalian mitochondrial DNA, the rate of nucleotide substitution is known to vary extensively from site to site, and there is a strong transition/transversion bias. The gamma distance for the Tamura-Nei model was developed primarily for the sequence data for this region. There are two hypervariable segments (5' and 3' segments), and the middle section is highly conserved. Using human data, Kocher and Wilson (1990) and Tamura and Nei (1993) estimated that *a* is about 0.11 for the entire control region, whereas Wakeley (1993) obtained *a* = 0.47 for the 5' hypervariable segment. Since most investigators use only the 5' hypervariable segment, we have decided to use *a* = 0.5 for the default option of MEGA. The gamma distance for the Tamura-Nei model is given by

$$
\begin{aligned}
d = 2a[&\frac{g_A g_G}{g_R}(1-\frac{g_R}{2g_A g_G}P_1 - \frac{1}{2g_R}Q)^{-\frac{1}{a}} \\
&+ \frac{g_T g_C}{g_Y}(1-\frac{g_Y}{2g_T g_C}P_2 - \frac{1}{2g_Y}Q)^{-\frac{1}{a}} \\
&+ (g_R g_Y - \frac{g_A g_G g_Y}{g_R} - \frac{g_T g_C g_R}{g_Y})(1-\frac{1}{2g_R g_Y}Q)^{-\frac{1}{a}} \\
&- g_A g_G - g_T g_C - g_R g_Y]
\end{aligned}
\qquad (4.38)
$$

A formula for the variance of this estimate is available (Tamura and Nei 1993) but is not reproduced here. It is incorporated in the computer program. Tamura and Nei also derived formulas for the estimates of the average numbers of transitional (*s*) and transversional (*v*) nucleotide substitutions per site, their variances, and the variance of the *s*/*v* ratio. These formulas are incorporated in MEGA.

4.2 Synonymous and Nonsynonymous Substitutions

Nucleotide substitutions in coding genes can be subdivided into two classes, *i.e.*, synonymous and nonsynonymous substitutions. Synonymous (or silent) substitutions are the nucleotide substitutions that do not result in amino acid changes, whereas nonsynonymous substitutions are those that change amino acids. The former substitutions are likely to be subject to little purifying selection except in lower organisms (however, see Britten 1993), while a majority of nonsynonymous changes are eliminated by purifying selection. Therefore, the rate of synonymous substitution is usually higher than that of nonsynonymous substitution (Miyata *et al*. 1980, Kimura 1983). Under certain conditions, however, nonsynonymous substitution may be accelerated by positive Darwinian selection (Hughes and Nei 1988, Lee and Vacquier 1992, and others). It is therefore interesting to examine the number of synonymous substitutions per synonymous site and the number of nonsynonymous substitutions per nonsynonymous site.

There are several methods for estimating these numbers (Miyata and Yasunaga 1980, Li *et al*. 1985, and others). In MEGA, however, we have included the simple method given by Nei and Gojobori (1986), since all methods give essentially the same results unless there are strong transition/transversion and G+C content biases. In Nei and Gojobori's (1986) method the numbers of synonymous (*S*) and nonsynonymous (*N*) sites are first computed. Here synonymous and nonsynonymous sites are the sites at which synonymous and nonsynonymous substitutions potentially occur, respectively (see Nei 1987, Pp. 73-76 for the method of computation). The sum of *S* and *N* is equal to the total number of nucleotides, *n*, and *N* is usually much larger than *S*. The numbers of synonymous (S_d) and nonsynonymous (N_d) substitutions that have occurred between two sequences are then computed by considering all pathways of nucleotide substitution between each pair of codons compared.

Using these quantities, we can compute the proportion of synonymous (p_S) and nonsynonymous (p_N) nucleotide differences per synonymous and nonsynonymous site, respectively. They are

$$p_S = S_d/S, \qquad (4.39)$$

$$p_N = N_d/N, \qquad (4.40)$$

with variances

$$V(p_S) = p_S(1 - p_S)/S, \qquad (4.41)$$

$$V(p_N) = p_N(1 - p_N)/N. \qquad (4.42)$$

Approximate estimates of the number of synonymous substitutions per synonymous site (d_S) and the number of nonsynonymous substitutions per nonsynonymous site (d_N) can be obtained by applying the Jukes-Cantor formula.

$$d_S = -3\log_e(1 - 4p_S/3)/4, \qquad (4.43)$$

$$d_N = - 3\log_e(1 - 4p_N/3)/4, \qquad (4.44)$$

with variances

$$V(d_S) = p_S(1 - p_S)/[(1 - 4p_S/3)^2S], \qquad (4.45)$$

$$V(d_N) = p_N(1 - p_N)/[(1 - 4p_N/3)^2N]. \qquad (4.46)$$

Computer simulations have shown that the above equations give good estimates if there is no transition/transversion bias (Nei and Gojobori 1986). However, when this bias is large, d_S tends to underestimate the true number of substitutions (Kondo *et al.* 1993). Li (1993) and Pamilo and Bianchi (1993) developed a method to take care of this problem. It is also possible to extend Nei and Gojobori's method to this case. We plan to include these methods in future versions of MEGA.

It should be noted that d_S and d_N are not reliable when p_S and p_N are large, say greater than 0.4, because their variances are large. In this case one may use p_S and p_N directly, particularly for studying positive Darwinian selection (Tanaka and Nei 1989).

In the study of adaptive evolution at the nucleotide level it is often necessary to compare the average values of d_S and d_N or p_S and p_N for a group of related sequences (*e.g.*, Hughes and Nei 1988). In this case we have to know the variances of average d_S and d_N or average ps and pN. These variances can be computed by Nei and Jin's (1989) method, and this computation is implemented in MEGA.

Once these values are computed, the statistical significance of the difference (d) between average d_S and d_N or average p_S and p_N can be tested by the *t*-test with an infinite degrees of freedom. That is, *t* is given by

$t = d/s(d)$, (4.47)

where $s(d)$ is the standard error of d and is given by [$V(d_S) + V(d_N)]^{1/2}$ or by $[V(p_S) + V(p_N)]^{1/2}$. Here, $V(d_S)$, $V(d_N)$, $V(p_S)$, and $V(p_N)$ are the variances of average d_S, d_N, p_S, and pN , respectively.

4.3 Amino Acid Substitutions

The methods for estimating the number of amino acid substitutions are similar to those for estimating the number of nucleotide substitutions except that there are 20 different states for the former rather than four states. The distance measures presented below can be computed either from amino acid sequences or from the coding regions of nucleotide sequences. In MEGA nucleotide sequences are translated into amino acid sequences by using one of the four genetic code tables ("universal" code and mammalian, Drosophila, and yeast mitochondrial genetic codes). Presence of a stop codon aborts the translation process and produces an error message. The treatment of missing nucleotides (or amino acids) and alignment gaps is discussed in the following section.

p-distance

As in the case for nucleotide sequences, the p-distance is merely the proportion of different amino acids between two sequences compared. Therefore, the statistical properties of this distance are the same as those of the p-distance for nucleotide sequence data.

$p = n_d/n$, (4.48)

$V(p) = p(1 - p)/n$ (4.49)

Here nd and n are the number of amino acid differences and the total number of amino acids compared, respectively.

Poisson-correction distance

This distance is for estimating the number of amino acid substitutions per site under the assumption that the number of amino acid substitutions at each site follows the Poisson distribution. This estimator (a) and its variance are given by

$d = -\log_e(1 - p)$ (4.50)

$V(d) = p/[(1 - p)n]$, (4.51)

where p is estimated by equation (4.48).

Gamma distance

This distance is an estimate of the number of amino acid substitutions per site under the assumption that the rate of amino acid substitution varies from site to site and follows the gamma distribution with parameter a. This distance and its variance can easily be computed from Nei *et al.*'s (1976) work.

$$d = a[(1 - p)^{-1/a} - 1], \qquad (4.52)$$

$$V(d) = p[(1 - p)^{-(1 + 2/a)}]/n. \qquad (4.53)$$

In the default option of MEGA, $a = 2$ is used. When $a = 2$ is used, d is close to Dayhoff's (1978) PAM distance per site (0.01 PAM) (Tatsuya Ota, personal communication).

4.4 Guidelines for Choosing Distance Measures

In the above three sections, we have discussed various distance measures considering different situations. In general, a complex mathematical model fits data better than a simple one. However, a complex model requires the estimation of many parameters, and this increases the variance of the estimate of *d*. Theoretically, it is possible to choose a distance measure most appropriate for a given set of data by using certain statistical criteria. Such statistical criteria are now under investigation (Bulmer 1991, Goldman 1993, Tamura 1994), but it seems to be premature to include these model-selection methods in this version of MEGA. Without such model-selection methods, it is possible to write some guidelines for choosing distance measures for the purpose of phylogenetic inference (modified from Nei 1991). They are as follows:

(1)

When the Jukes-Cantor estimate of the number of nucleotide substitutions per site (d) between different sequences is about 0.05 or less ($d < 0.05$), use the Jukes-Cantor distance whether there is a transition/transversion bias or not or whether the substitution rate (λ) varies with nucleotide site or not. In this case, the Kimura distance or the gamma distance gives essentially the same value as the Jukes-Cantor distance. One may also use the p-distance for constructing a topology.

(2)

When $0.05 < d < 0.3$, use the Jukes-Cantor distance unless the transition/transversion ratio (R) is high, say $R < 2$. When this ratio is high and the number of nucleotides examined is large, use the Kimura distance or the gamma distances for Kimura's 2-parameter model.

(3)

When $0.3 < d< 1$ and there is evidence that λ varies extensively with site, use gamma distances. In general, one may choose different gamma distances, estimating a from data.

(4)

When $0.3 < d < 1$ and the frequencies of the four nucleotides (A, T, C, G) deviate substantially from equality but there is no strong transition/transversion bias, use the Tajima-Nei distance. When there are strong transition/transversion and G+C content biases, use the Tamura or Tamura-Nei distance.

(5)

When $d > 1$ for many pairs of sequences, the phylogenetic tree estimated is not reliable for a number of reasons (*e.g.*, large standard errors of *d*'s and sequence alignment errors). We therefore suggest that these sets of data should not be used. In this case one may eliminate the portion(s) of the gene that evolves very fast and use only the remaining region(s) as is often done in studies of the evolution of different kingdoms or phyla using ribosomal RNA genes. If a coding region of DNA is examined, we suggest that amino acid sequences rather than DNA sequences be used. One may also use a different gene that evolves more slowly.

In the study of evolution of multigene families, it is often necessary to examine phylogenetic relationships (topologies) of distantly related sequences with $d < 1$. In this case the nucleotide or amino acid *p*-distance is helpful because this distance has a smaller variance and it generates no inapplicable cases (*e.g.*, Burke *et al.* 1993).

(6)

When a phylogenetic tree is constructed from the coding regions of a gene, the distinction between synonymous (d_S) and nonsynonymous (d_N) substitutions may be helpful because the rate of synonymous substitutions is usually much higher than that of nonsynonymous substitution. When relatively closely related species with *d*s < 1 are studied for a large number of codons, one may use ds for constructing a tree. This procedure is expected to reduce the effect of variation in substitution rate among different sites, because synonymous substitutions are apparently largely neutral in higher organisms. However, when relatively distantly related species are studied, d_N or amino acid distances should be used.

(7)

As a general rule, if two distance measures give similar distance values for a set of data, use the simpler one because it has a smaller variance. When the rate of nucleotide substitution is the same for all evolutionary lineages and the number of nucleotides used is relatively small, the *p* or Jukes-Cantor distance seems to give a correct tree more often than the Kimura distance even if there is some extent of transition/transversion bias (Schöniger and von Haeseler 1993 Tajima and Takezaki, 1994). When the substitution rate varies with evolutionary lineage, however, this is not the case.

Note that the above guidelines are for constructing phylogenetic trees. For estimating evolutionary times or for testing the reliability of branch lengths, unbiased estimators are better than biased estimators, though unbiased estimators may vary with the data set used.

4.5 Alignment Gaps and Sites with Missing Information

Gaps are often inserted during the alignment of homologous regions of sequences and represent deletions or insertions (indels). These gaps introduce some complications in distance estimation. Furthermore, sites with missing information can sometimes occur because of experimental difficulties, and they create the same problems as that for gaps. In the following discussion both of these sites are treated in the same way.

In MEGA, gap sites are ignored in distance estimation, but there are two different ways to treat these sites. One way to deal with this problem is to delete all of these sites from data analysis. This option, which is called the *Complete-Deletion* option in MEGA, is generally desirable because different regions of DNA or amino acid sequences often evolve under different evolutionary forces. However, if the number of nucleotides involved in a gap is small and gaps are distributed more or less at random, one may compute a distance for each pair of sequences ignoring only those gaps that are involved in the comparison. This option is called the Pairwise-Deletion option. Table 4.2 illustrates the effect of these options on distance estimation with the following three sequences:

```
         1       10        20
seq1     A-AC-GGAT-AGGA-ATAAA
seq2     AT-CC?GATAA?GAAAAC-A
seq3     ATTCC-GA?TACGATA-AGA    Total sites = 20.
```

Here, the alignment gaps are indicated with a hyphen (-) and the missing information sites are denoted by a question mark (?).

Table 4 . 2 Complete-Deletion and Pairwise-Deletion options

Option	Sequence Data	Differences/comparisons		
		(1,2)	(1,3)	(2,3)
Complete-Deletion	1. A C GA A GA A A 2. A C GA A GA A C A 3. A C GA A GA A A A	1/10	0/10	1/10
Pairwise-Deletion	1. A-AC-GGAT-AGGA-ATAA 2. AT-CC?GATAA?GAAAAC-A 3. ATTCC-GA?TACGATA-AGA	2/12	3/13	3/14

In Table 4.2, the number of sites compared varies with pairwise comparison in the Pairwise-Deletion option, but it remains the same for all pairwise comparisons in the Complete-Deletion option. In this particular data set, more information can be obtained by using the Pairwise-Deletion option. In practice, however, different regions of nucleotide or amino acid sequences often evolve differently. In this case the Complete-Deletion option is preferable.

[Next] [Table of Contents]