Estimation of Average Number of Nucleotide Substitutions When the Rate of Substitution Varies with Nucleotide

Takashi Gojobori, Kazushige Ishii, and Masatoshi Nei

Center for Demographic and Population Genetics, The University of Texas at Houston, Houston, Texas 77025, USA

**Summary.** A formal mathematical analysis of Kimura's (1981) six-parameter model of nucleotide substitution for the case of unequal substitution rates among different pairs of nucleotides is conducted, and new formulae for estimating the number of nucleotide substitutions and its standard error are obtained. By using computer simulation, the validities and utilities of Jukes and Cantor's (1969) one-parameter formula, Takahata and Kimura's (1981) four-parameter formula, and our six-parameter formula for estimating the number of nucleotide substitutions are examined under three different schemes of nucleotide substitution. It is shown that the one-parameter and four-parameter formulae often give underestimates when the number of nucleotide substitutions is large, whereas the six-parameter formula generally gives a good estimate for all the three substitution schemes examined. However, when the number of nucleotide substitutions is large, the six-parameter and four-parameter formulae are often inapplicable unless the number of nucleotides compared is extremely large. It is also shown that as long as the mean number of nucleotide substitutions is smaller than one per nucleotide site the three formulae give more or less the same estimate regardless of the substitution scheme used.

**Key words:** Molecular evolution — Nucleotide substitution — Evolutionary distance

---

**Introduction**

Studying the nucleotide substitution in preproinsulin and globin genes, Perler et al. (1980) noted that the accumulation of silent changes as measured by “percent corrected divergence” is not linear with evolutionary time. Their method of “percent corrected divergence” is essentially the same as the Jukes-Cantor (1969) method, and assumes that the rates of substitution among the four different nucleotides are the same. Kimura (1980, 1981) and Takahata and Kimura (1981) introduced new methods of estimating the number of nucleotide substitutions, taking into account the unequal rates of substitution between certain pairs of nucleotides. Using these methods, Takahata and Kimura (1981) have shown that the Jukes-Cantor method may give underestimates of nucleotide substitutions when distantly related sequences are compared.

The Jukes-Cantor method is based on a single parameter about the rate of nucleotide substitution. When unequal rates of substitution among different nucleotides are considered, however, more than one parameter is required. Kimura's (1980) first model included two parameters for substitution rate, but later he (Kimura 1981) introduced a three-parameter model (3ST model) and six-parameter model (2PC model). On the other hand, Takahata and Kimura's (1981) model requires 4 parameters (see Tables 1 and 2). Jukes and Cantor's (1969) model is a special case of \( \alpha = \beta = \gamma \) and \( \theta = 1 \) in Table 1. If all these parameters are known for each set of data and remain constant as assumed in the model, a model with a larger number of parameters is expected to be better than the one with a fewer number of parameters. However, if the model setup is not realistic, a model with many parameters may give a poor estimate of nucleotide substitutions. At the present time, we have very little information about the parameters for these models, so that it is difficult to decide which method is better. It should also be noted that the effect of stochastic errors on the estimate of nucleotide substitutions is often greater for a model with many
parameters than for a single-parameter model (Nei and Tateno 1978). It is therefore important to know the relative merits of these methods under various conditions. The purpose of this paper is to study this problem by using computer simulation. We shall also present a formal mathematical analysis of the six-parameter model, since Kimura's (1981) treatment involves some intuitive arguments and approximations.

**Six-Parameter Model**

Let us consider the evolution of a nucleotide sequence consisting of \( L \) codons. We consider nucleotide substitutions at the first, second, and third positions of codons separately, assuming that substitution occurs with the rates given in Table 2. Thus, we are in effect considering a sequence of \( L \) nucleotides, though \( L \) codons contain \( 3L \) nucleotides. We do this because the three different positions often show different rates of substitution. In the following we denote the four nucleotides A, T, C, and G by 1, 2, 3, and 4, respectively. Starting from a common ancestor sequence with an infinitely large number of codons (\( L = \infty \)), the nucleotide frequencies \( \{ q_i(t), i = 1, 2, 3, 4 \} \) at each of its descendant sequences at time \( t \) may be expressed as

\[
q_i(t) = \sum_{k=1}^{4} q_k(0) P_{ki}(t) \quad (i = 1, 2, 3, 4)
\]

in terms of the conditional probability, \( P_{ki}(t) \), that the \( k \)-th nucleotide changes to the \( i \)-th nucleotide during the period of \( t \). Here, we assume that the nucleotide substitutions in different codons occur independently. When \( L \) is finite, a random fluctuation from the deterministic equation (1) occurs, but we shall study this problem later. Now consider two descendant sequences at time \( t \). The probability that at a given nucleotide site the first and second sequences have nucleotides \( i \) and \( j \), respectively, is given by

\[
x_{ij}(t) = \sum_{k=1}^{4} q_k(0) P_{ki}(t) P_{kj}(t) \quad (i, j = 1, 2, 3, 4).
\]

Thus, in order to know \( x_{ij}(t) \) and \( q_i(t) \) we must know \( q_i(0) \) and \( P_{ki}(t) \). Under the assumption of constant rate of nucleotide substitution, \( P_{ki}(t) \)'s satisfy the following set of ordinary differential equations

\[
dP_{ki}(t)/dt = \sum_{j=1}^{4} P_{kj}(t) R_{ij} \quad (k, i = 1, 2, 3, 4).
\]

where \( R_{ij} \) is the rate of substitution for the \( j \)-th nucleotide by the \( i \)-th nucleotide as given in Table 2, while \( R_{ij} \) is given by \( \sum_{i=1}^{4} R_{ij} \) for \( i \neq j \). Solution of (3) with the initial condition

\[
P_{ki}(0) = \begin{cases} 1 & (i=k) \\ 0 & \text{(otherwise)} \end{cases}
\]

gives

\[
P_{ki}(t) = q_i + \sum_{r=1}^{3} u_{rk} v_{ri} \lambda_r^t
\]

where \( \lambda_r \)'s are the eigenvalues of the \( 4 \times 4 \) matrix \( R \) in Table 2. It can be shown that the matrix \( R \) has the eigenvalues of zero and

\[
\lambda_1 = -2(\alpha + \beta), \quad \lambda_2 = -(2\alpha + \alpha_1 + \beta_1), \quad \lambda_3 = -2(\beta + \alpha_2 + \beta_2).
\]

In (5) \( q_i \) denotes the \( i \)-th element of the left eigenvector corresponding to the eigenvalue of zero, and is given by

\[
q_1 = \xi_1, \quad q_2 = \xi(1 - \xi_1), \quad q_3 = (1 - \xi) \xi_2, \quad q_4 = (1 - \xi)(1 - \xi_2)
\]

with

\[
\xi = \beta/(\alpha + \beta), \quad \xi_1 = (\alpha + \beta)/(2\alpha + \alpha_1 + \beta_1), \quad \xi_2 = (\beta + \alpha_2)/(2\beta + \alpha_2 + \beta_2),
\]

whereas its right eigenvector is given by \( (1, 1, 1, 1)^t \) where \( t \) denotes the transpose of the vector. The parameter \( q_i \) in (7) is equal to the equilibrium value of \( q_i(t) \), i.e., \( q_i(\infty) \). In (5) \( u_{rk} \) and \( v_{ri} \) are, respectively, the \( k \)-th and \( i \)-th elements of the right and left eigenvectors corresponding to the eigenvalue \( \lambda_r \) (\( r = 1, 2, 3 \)) in (6). They are given by
\[ u_{11} = u_{12} = 1 + u_{13} = 1 + u_{14} = 1 - \xi \ , \]
\[ v_{11} = 1 - v_{12} = (a_1 - a_2) / (2a + a_2 - a_1) \ , \]
\[ v_{13} = -1 - v_{14} = -(a_1 - a_2) / (2a - a_2 - a_1) \ , \]
\[ u_{21} = 1 + u_{22} = (1 - \xi_1) - (1 - \xi) u_{23} / \xi \ , \]
\[ u_{23} = u_{24} = \beta_1 - a_1 / (a_1 + \beta_1 - 2\beta) (2a + a_1 + \beta_1) \ , \]
\[ v_{21} = -v_{22} = v_{23} + 1 = v_{24} + 1 = 1 \ , \]
\[ u_{31} = u_{32} = \alpha (a_2 - a_1) / (a_1 + \beta_2 - a_2 (2a + a_2 + \beta_2) \ , \]
\[ u_{33} = 1 + u_{34} = 1 - \xi_2 - \xi u_{31} / (1 - \xi) \ , \]
\[ v_{31} = v_{32} = v_{33} - 1 = v_{34} + 1 = 0 \ . \]

In the present paper we assume that the common ancestor sequence at time 0 has the equilibrium nucleotide frequencies, i.e., \( q_i(0) = q_i \). Substituting \( P_{ki} \) in (5) into (1) and (2), we then have

\[ q_i(t) = q_i \]

(12)

and

\[ x_{ij}(t) = q_i q_j + \frac{1}{2} \sum_{r=1}^{3} \sum_{s=1}^{3} C_{rs} v_{ri} v_{sj} e^{(r+s) t} \]

(13)

where \( C_{rs} = \sum_{k=1}^{4} u_{rk} u_{sk} q_k \). It is clear from (12) that the nucleotide frequencies remain at the equilibrium value for the entire evolutionary process.

In the present model the average rate of nucleotide substitution per site is given by

\[ k = \sum_{j=1}^{4} q_i \sum_{i 
eq j} R_{ji} = q_1 (2a + a_1) + q_2 (2a + \beta_1) + q_3 (2\beta + a_2) + q_4 (2\beta + \beta_2) \]

(14)

whereas the average number of nucleotide substitutions between two species for a period of \( t \) is

\[ \delta = 2kt \]

(15)

In practice, we do not know the values of \( \alpha, a_1, a_2, \beta, \beta_1, \) and \( \beta_2 \), but these quantities can be expressed in terms of observable quantities \( q_i \) and \( x_{ij}(t) \). To do this, we first note that \( \delta \) can be written as

\[ \delta = -2\xi(1-\xi)\lambda_1 t - 4\xi_1(1-\xi_1)\lambda_2 t \]
\[ -4(1-\xi)\xi_2(1-\xi_2)\lambda_3 t \ , \]

(16)

if we use the relations \( \alpha + \lambda_1 / 2 = -\xi \lambda_1 / 2, \alpha + a_1 + \lambda_2 = (\alpha + \beta_1) = \xi_1 \lambda_2, \) and \( \beta + a_2 + \lambda_3 = (\beta + \beta_2) = \xi_2 \lambda_3 \), which are obtainable from (6) and (8). We also note that \( \xi, \xi_1, \) and \( \xi_2 \) can be expressed in terms of observable quantities \( q_i \), i.e., \( \xi = q_1 + q_2, \xi_1 = q_1 / (q_1 + q_2) \), and \( \xi_2 = q_3 / (q_3 + q_4) \). Therefore, if we know \( d_1 = -2\lambda_1 t, d_2 = -2\lambda_2 t, \) and \( d_3 = -2\lambda_3 t, \) the value of \( \delta \) in (16) is given by

\[ \delta = h_1 d_1 + 2h_2 d_2 + 2h_3 d_3 \]

(17)

where \( h_1 = \xi(1 - \xi), h_2 = \xi_1(1 - \xi_1), \) and \( h_3 = (1 - \xi) \xi_2(1 - \xi_2) \).

To derive \( d_i \), we note that the quantities \( C_{rs} \) in (13) can be written as \( C_{11} = \xi(1 - \xi), C_{12} = -(1 - \xi)u_{23}, C_{13} = \xi_1(1 - \xi_1), \) \( C_{22} = \xi_1(1 - \xi_1) + (1 - \xi)u_{23}^2 / \xi, C_{23} = -u_{23} u_{31}, \) and \( C_{33} = (1 - \xi) \xi_2(1 - \xi_2) + u_{31}^2 / (1 - \xi) \) from (9) to (11). Substituting these into (13) and making use of the values of \( v_{ri} \) in (10) and (11), we can see that \( x_{ij}(t) \) is given by

\[ x_{ij} = \begin{cases} (-1)^{i+j} c_2 + q_i q_j + a_i a_j / h_1, & (i = 1, 2) \quad (18) \\ (-1)^{i+j} c_3 + q_i q_j + a_i a_j / h_1, & (i = 1, 2; j = 3, 4) \quad (19) \\ q_i q_j + a_i a_j / h_1, & (i = 1, 2; j = 3, 4) \quad (20) \end{cases} \]

where the argument of \( x_{ij}(t) \) is dropped, and

\[ c_2 = h_2 e^{-d_2}, c_3 = h_3 e^{-d_3} \]

(21)

\[ a_i = \begin{cases} h_1 v_{11} e^{-d_1 / 2} + (-1)^{i} (1 - \xi) u_{23} e^{-d_2 / 2}, & (i = 1, 2) \quad (22) \\ h_1 v_{12} e^{-d_1 / 2} - (-1)^{i} u_{31} e^{-d_3 / 2}, & (i = 3, 4) \quad (23) \end{cases} \]

We also have the following relationship from (9).

\[ a_1 + a_2 = - (a_3 + a_4) = h_1 e^{-d_1 / 2} \]

(24)

We are now ready to write down the equations for computing \( d_i \) from observable quantities only. First, from (18) to (20) and (24) we have

\[ a_1 e^{-d_1 / 2} = A_1 = ((1 - 2\xi) q_1 + x_{11} + x_{12} - x_{13} \]
\[ -x_{14}) / 2 \] \quad (25)

Making use of relation (24), we then obtain \( h_1 e^{-d_1} = A_1 \)
\[ + A_2 = A - A_3 - A_4 \]. Therefore, \( d_1 \) is given by

\[ d_1 = -\ln (B_1 / h_1) \]

(26)
\[ B_1 = \left( A_1 + A_2 - A_3 - A_4 \right)/2 \]
\[ = h_1 - \sum_{i=1}^{2} \sum_{j=3}^{4} x_{ij} \]

To get \( d_2 \), we use the following equation
\[ h_2 e^{-d_2} = (-1)^{i+j} (x_{ij} - q_i q_j - A_i A_j / B_1) \quad (i,j=1,2) \]
which is obtainable from (18), (25), and (26). We then have
\[ d_2 = -\ln \left( B_2 / h_2 \right) \quad (27) \]
with
\[ B_2 = \frac{1}{3} \left( x_{11} + x_{22} - x_{12} - q_1^2 - q_2^2 + q_1 q_2 \right) - (A_1^2 + A_2^2 - A_1 A_2) / B_1 \]
Similarly, \( d_3 \) is given by
\[ d_3 = -\ln \left( B_3 / h_3 \right) \quad (28) \]
with
\[ B_3 = \frac{1}{3} \left( x_{33} + x_{44} - x_{34} - q_3^2 - q_4^2 + q_3 q_4 \right) - (A_3^2 + A_4^2 - A_3 A_4) / B_1 \]
Therefore, \( \delta \) in (17) can be expressed in terms of \( q_i \) and \( x_{ij} \).

In the above formulation we assumed \( L = \infty \). In actual practice, \( L \) is always finite, and thus the estimate of \( \delta \) obtained by (17) is subject to a sampling error. The sampling variance \( (\sigma^2) \) of \( \delta \) can be computed in the following way. We first note that \( \delta \) can be expressed in terms of \( x_{ij} \)'s only and the small deviation of \( \delta \) is given by
\[ \Delta \delta = \sum_{i=1}^{4} \sum_{j=1}^{4} D_{ij} \Delta x_{ij} \]
approximately, where \( D_{ij} = \partial \delta / \partial x_{ij} \).

Therefore, we have
\[ \sigma^2 = \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{k=1}^{4} \sum_{e=k}^{4} D_{ij} D_{ke} E(\Delta x_{ij} \Delta x_{ke}) \]
\[ = \frac{1}{L} \left( \sum_{i=1}^{4} D_{ii}^2 x_{ii} + \frac{1}{2} \sum_{i=1}^{4} \sum_{j=i+1}^{4} D_{ij}^2 x_{ij} \right) \]
\[ - \left( \sum_{i=1}^{4} \sum_{j=1}^{4} D_{ij} x_{ij} \right)^2 \]  
(29)

where \( E(\Delta x_{ij}^2) = x_{ij} (1 - x_{ij}) / L, E(\Delta x_{ij}^2) = x_{ij} (1/2 - x_{ij}) / L, \) and \( E(\Delta x_{ij} \Delta x_{ke}) = -x_{ij} x_{ke} / L \).

Although the above method of computing \( \delta \) and \( \sigma^2 \) is theoretically straightforward, it is quite complicated. Therefore, we have developed a computer program, which can be used for any set of data on \( \{ q_i; i = 1, 2, 3, 4 \} \) and \( \{ x_{ij}; i < j \} \). Since \( x_{ij} = x_{ji} \) does not generally hold in actual data, we use the average of \( x_{ij} \) and \( x_{ji} \) in the computation. For estimating \( \delta \) from actual data, however, the following simpler formula may be used.

\[ \delta = -pq \ln \left( \frac{B_1}{pq} \right) \]
\[ - \frac{2q_1 q_2}{p} \ln \left( \frac{-p}{3q_1 q_2} (F_{12} - B_1 + 3E_{12}/B_1) \right) \]
\[ - \frac{2q_3 q_4}{q} \ln \left( \frac{q}{3q_3 q_4} (F_{34} - B_1 + 3E_{34}/B_1) \right) \quad (30) \]

where \( p = q_1 + q_2, q = q_3 + q_4, B_1 = pq - (x_{13} + x_{14} + x_{23} + x_{24}), E_{12} = (q_1 q_4 - x_{13} - x_{14}) (q_2 q_4 - x_{23} - x_{24}), E_{34} = (q_3 p - x_{13} - x_{23}) (q_4 p - x_{14} - x_{24}), F_{12} = x_{11} + x_{22} - x_{12} - p^2 + 3q_1 q_2, \) and \( F_{34} = x_{33} + x_{44} - x_{34} - q_3^2 + 3q_3 q_4 \). The above formula gives exactly the same value as that from (17).

Kimura's Formula

It should be noted that our formula does not give the same estimate of \( \delta \) as that obtained by Kimura's (1981) for this 6-parameter model. This can be seen by considering the simplest case of equal substitution rates for all pairs of nucleotides. In this case \( q_i = 1/4, x_{ii} = x_{11} \) for all \( i \)'s, and \( x_{ij} = x_{12} \) for \( i \neq j \). We also have \( 4x_{11} + 12x_{12} = 1, p = q = 1/2, B_1 = 1/4, 2E_{12} = 12x_{12} = (1/8 \times 2x_{12})^2, \) and \( F_{12} = F_{34} = 2x_{11} - x_{12} - 1/16. \) Therefore, \( \delta \) in (30) becomes
\[ \delta = -\frac{1}{4} \ln (1 - 16x_{12}) - \frac{1}{2} \ln (16x_{11} - 1)/3 \]
\[ = -\frac{3}{4} \ln (1 - \frac{4}{3} \pi) \quad , (31) \]

where \( \pi = 12x_{12} \) is the proportion of different nucleotides between the two homologous sequences compared. The above equation is identical with the Jukes-Cantor formula, as expected.

Kimura's formula does not reduce to (31). His general formula is given by
\[ \delta = -\theta_b \ln (1 - \frac{Z_b}{\theta_b}) - \theta_b \ln (1 - \frac{Z_w}{\theta_w}) \]
\[ - q\theta_w \ln (1 - \frac{Z_w}{\theta_w}) \quad , \]
(32)
where $Z_b = X_{13} + X_{14} + X_{23} + X_{24}, Z_{w1} = X_{12}/(X_{11} + X_{12} + X_{11} + x_{22}), Z_{w2} = X_{34}/(X_{34} + X_{33} + x_{44}), \theta_1 = 2pq, \theta_{w1} = 2q_{14}, \theta_{w2} = 2q_{34}, p = q_1 + q_2, q = q_3 + q_4, q_{14} = q_1/p, q_{34} = q_3/p, q_3 = q_3/q, and q_4 = q_4/q. Here X_{ij} = x_j + x_i (i \neq j). In the present case Z_b = 8x_{32}, Z_{w1} = x_{12}/(x_{11} + x_{12}), \theta_b = \theta_{w1} = \theta_{w2} = 1/2. Therefore, (32) reduces to

$$
\delta = -\frac{1}{2} \ln \left[ \frac{1 - 16x_{12}}{1 - 2x_{12}/(x_{11} + x_{12})} \right]
$$

$$
= -\frac{1}{4} \ln \frac{(1 - 16x_{12})^4}{1 - 16x_{12} + 64x_{12}^2}
$$

(33)

This is different from Jukes and Cantor's formula, though theoretically it should agree with the latter in this case. This difference has occurred because in Kimura's formulation nucleotides are first classified into two groups, i.e., A + T and G + C, and the substitutions within and between groups are separately considered. His approach does not seem to be correct, since in the computation of within-group substitution the contribution from the other group is neglected. Formula (33) gives a slight overestimate. When the number of nucleotide substitutions is small, however, the extent of the overestimate is small. Indeed, when $x_{12}$ is small, $1 - 16x_{12} + 64x_{12}^2$ in (33) is approximately $1 - 16x_{12}$, and thus (33) becomes approximately equal to (31).

Identically, our formulae (17) and (30) reduce to Kimura's (1980) formula for the two-parameter model, where $\alpha = \beta$ and $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = \gamma$ and A, G, C, and T are denoted by 1, 2, 3, and 4, respectively. Note that the six-parameter model in Table 2 is qualitatively different from Takahata and Kimura's (1981) four-parameter model or Kimura's (1981) 3ST model, so that it cannot be compared with these models.

### Numerical Example

To illustrate our method of computation, let us consider the nucleotide sequences of the mouse and rabbit b-globin genes (Konkel et al. 1979, Hardison et al. 1979). There are 146 codons in the b-globin gene, and we consider only the third positions of the codons. Comparison of the 146 third positions between the two sequences gives the following relative nucleotide pair frequencies: $x_{11} = 0.007 (AA), x_{22} = 0.178 (TT), x_{33} = 0.226 (CC), x_{44} = 0.274 (GG), 2x_{12} = 0.028 (AT), 2x_{13} = 0.014 (AC), 2x_{14} = 0.062 (AG), 2x_{23} = 0.136 (TC), 2x_{34} = 0.041, and 2x_{44} = 0.034 (CG). Therefore, the nucleotide frequencies are given by

$q_1 = x_{11} + x_{12} + x_{13} + x_{14} = 0.059,$

$q_2 = x_{12} + x_{22} + x_{23} + x_{24} = 0.281,$

$q_3 = x_{13} + x_{23} + x_{33} + x_{34} = 0.318,$

$q_4 = x_{14} + x_{24} + x_{34} + x_{44} = 0.342.$

Furthermore, we have $p = q_1 + q_2 = 0.34, q = q_3 + q_4 = 0.66, B_1 = 0.0074, E_{12} = 7.8 \times 10^3, E_{34} = 0.0021, F_{12} = 0.1053, and F_{34} = 0.3732. Thus, we have $\delta = 0.56. This is larger than the values obtained by Kimura's formula (0.49) and Jukes and Cantor's formula (0.41).

### Computer Simulation

#### Methods

To see the accuracies of the one-parameter (1-p), four-parameter (4-p), and six-parameter (6-p) methods of estimating nucleotide substitutions, we have conducted a computer simulation under three different substitution schemes, i.e., 1-p scheme, 4-p scheme, and 6-p scheme. In this simulation we considered the evolutionary changes of nucleotide sequences analogous to the actual changes of the sequences of third positions of codons in the globin genes. In practice, we generated a total of 15 sequences, of which seven were the ancestral or intermediate sequences and eight were the contemporary sequences (Fig. 1). In Fig. 1, distances $\delta_1$, $\delta_2$, and $\delta_3$ represent the expected numbers of nucleotide substitutions for the cases corresponding to the divergence of different orders of mammals, the mammal-bird divergence, and the $\alpha$-globin and $\beta$-globin divergence, respectively. For computational convenience we have used discrete time rather than continuous time in this simulation. We denote by $P_{ij}$ the probability of substitution of the i-th nucleotide by the j-th nucleotide during a small but finite time unit. In the 1-p scheme $P_{ij}$ is the same for all $i \neq j$. We have considered the nucleotide substitution as a Markov process with 4 states, whose transition probabilities are $P_{ij} = R_{ij}$ for $i \neq j$ and $P_{ii} = 1 - \sum_{j \neq i} R_{ij}$.

![Fig. 1. The evolutionary tree used for the computer simulation](image-url)
Table 3. Means (δ) and standard deviations (σδ) of δ over all replications and the frequencies of inapplicable cases. One-parameter (1-p), four-parameter (4-p), and six-parameter (6-p) methods are used for three different substitution schemes. L = 144

<table>
<thead>
<tr>
<th>Substitution scheme</th>
<th>True distance (δ)</th>
<th>1-p method</th>
<th>4-p method</th>
<th>6-p method</th>
<th>Inapplicable cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>δ</td>
<td>σδ</td>
<td>δ</td>
<td>σδ</td>
<td>1/p method</td>
</tr>
<tr>
<td>1-p scheme</td>
<td>0.5</td>
<td>0.52</td>
<td>0.07</td>
<td>0.53</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>1.03</td>
<td>0.23</td>
<td>1.07</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>2.08</td>
<td>0.55</td>
<td>1.91</td>
<td>0.33</td>
</tr>
<tr>
<td>4-p scheme</td>
<td>0.5</td>
<td>0.48</td>
<td>0.07</td>
<td>0.51</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.90</td>
<td>0.19</td>
<td>1.00</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>1.84</td>
<td>0.44</td>
<td>1.71</td>
<td>0.36</td>
</tr>
<tr>
<td>6-p scheme</td>
<td>0.5</td>
<td>0.47</td>
<td>0.08</td>
<td>0.56</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.81</td>
<td>0.13</td>
<td>1.00</td>
<td>0.24</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>1.24</td>
<td>0.18</td>
<td>1.36</td>
<td>0.30</td>
</tr>
</tbody>
</table>

*The number of replications was ten, but since there were eight "contemporary sequences", the number of possible comparisons was larger than ten. However, all comparisons are not independent.

We computed the equilibrium nucleotide frequencies (qij) by squaring the matrix [Pij] repeatedly. Squaring was continued until the ratio of the standard deviation of the elements in each column of the matrix to the mean became less than 0.001. We note that all the elements in each column should have the same value at equilibrium. The average rate of nucleotide substitution per site per unit time was then given by

\[ k = \sum_{i=1}^{4} q_{ij} \sum_{j \neq i} P_{ij} \]  \hspace{1cm} (34)

In practice, the values of Pij's were chosen so as to be k = 0.01 for all substitution schemes. This was achieved by using Pij = 0.0033333 for all i ≠ j for the 1-p scheme, α = 0.00125, β = 0.0025, γ = 0.005, θ = 2.0 for the 4-p scheme (see Table 1), and α = 0.00125, α1 = 0.008, α2 = 0.0118, β = 0.005, β1 = 0.004, β2 = 0.0059 for the 6-p scheme (see Table 2). The number of time units (T) corresponding to δi was computed by T = δi/(2k). The ancestral sequence of 144 nucleotides equivalent to the third positions of the globin genes was obtained by generating a random nucleotide sequence for each substitution scheme. The probability, q(T), that a given nucleotide site is occupied by the i-th nucleotide at time t = T was then computed by (1), where one of q1(0), q2(0), q3(0), and q4(0) is 1 and the others are 0. Using the probabilities q(T) thus obtained, we generated the intermediate sequences α, β, α1, β1, ..., β2 as well as the contemporary sequences α1, α2, ..., α6. We then compared every pair of contemporary sequences obtained and computed nucleotide pair frequencies (xij) for each sequence pair. We applied the 1-p, 4-p, and 6-p methods to these nucleotide pair frequency data to estimate evolutionary distances. In addition to the case of L = 144, we also considered the case of L = 500 and L = 3000 to see the effect of the sampling error.

Result

In the case of L = 144, we used δ1 = 0.5, δ2 = 1.0, and δ3 = 2.0, and repeated the simulation ten times. The nucleotide frequencies were computed by

\[ q_{ij} = \sum_{j=1}^{4} x_{ij} \]  \hspace{1cm} (35)

for each pair of sequences. The effect of using the nucleotide frequencies averaged over 8 extant sequences will be discussed later. The results obtained are given in Table 3. When δ is large, particularly when δ = 2.0, there are many cases in which the 4-p and 6-p methods are not applicable because of the negative value of the argument of logarithm (e.g., B1/h1 in (26)). For example, in the case of δ = 2.0 the 4-p and 6-p methods are not applicable in 108 and 129 comparisons out of a total of 160, respectively, when the 6-p substitution scheme is used. On the other hand, the 1-p method is not applicable only in 22 cases of all comparisons examined. In all substitution schemes used, inapplicable cases occur with a slightly higher frequency for the 6-p method than for the 4-p method. There is no indication that the 6-p method gives a smaller number of inapplicable cases when the 6-p substitution scheme is used than when the other substitution schemes are used. The same thing can be said for the 1-p and 4-p methods.

Table 3 includes the means and standard deviations of the estimates (δ) of the number of nucleotide substitutions over all replications. These values were computed by excluding inapplicable cases. When the 1-p substitution scheme is used, all three methods give good estimates of expected values. This is of course expected, since the 4-p and 6-p methods are essentially the same as the 1-p method in this case. When the 4-p substitution scheme is used, the 4-p and 6-p methods give good esti-
Fig. 2. The frequency distributions of the estimates $\hat{\delta}$ of nucleotide substitutions, excluding inapplicable cases. $\delta = 1.0$ and $L = 144$

...imates for $\delta_1 = 0.5$ and $\delta_2 = 1.0$, but the 1-p method tends to give slight underestimates. In the case of the 4-p substitution scheme the 4-p method is expected to give the best result, since it is designed to be used for this case. In practice, this is not necessarily the case. Indeed, when $\delta_3 = 2.0$, the 4-p method is inferior to the 1-p method. This is apparently due to the fact that the 4-p method is not applicable in a high proportion of cases and the estimate obtained is based on a small number of applicable cases. To see this point, we computed the estimate of $\delta_3$ for the 43 cases in which both 1-p and 4-p methods were applicable. It was 1.61 for the 1-p method and 1.72 for the 4-p method. This indicates that exclusion of inapplicable cases gives underestimates as pointed out by Kimura (1981).

When the 6-p substitution scheme is used, the 4-p and 6-p methods give good estimates for $\delta_1 = 0.5$ and $\delta_2 = 1.0$ but underestimates for $\delta_3 = 2.0$. These underestimates are clearly caused by exclusion of inapplicable cases. The difference between the expected and estimated values is smaller in the 6-p method than in the 4-p or 1-p method. This suggests that although the 6-p method is often inapplicable, it gives a better estimate than the other two methods when it is applicable.

In actual data analysis there are no replicate sequences, so that the estimate ($\bar{\delta}$) of $\delta$ has a large variance. This can be seen from Fig. 2, where the distribution of $\hat{\delta}$ is given for the nine cases (three methods for three substitution schemes) for $\delta = 1$. It is clear that in all cases $\hat{\delta}$ is distributed widely and thus the individual value of $\hat{\delta}$ may be considerably different from $\delta = 1$. The variance or standard deviation of $\hat{\delta}$ is nearly the same in all cases except in the 1-p method for the 4-p and 6-p substitution schemes (see Table 3). The small standard deviations for these two cases are caused by the fact that $\delta$ is underestimated in these cases.

Table 3 indicates that the standard deviation (not standard error) increases with increasing $\delta$, as expected. The standard deviation for the 6-p method agrees reasonably well with the theoretical value given by (29) when the 6-p substitution scheme is used and $\delta$ is equal to or smaller than 1.0. Namely, the theoretical standard deviations for $\delta = 0.5$ and $\delta = 1$ are 0.093 and 0.244, respectively. These are close to their corresponding observed values in Table 3. When $\delta$ is as large as 2, however, the observed value (0.522) is considerably smaller than the theoretical value (2.41). This is probably caused by elimination of inapplicable cases.

The high proportion of inapplicable cases and underestimation of $\delta$ in the 4-p and 6-p methods are of course due to the relatively small number of nucleotides used (L). If L is infinitely large, there will be no inapplicable cases and no underestimation. Therefore, to see the effect of L on the estimate of $\delta$, we conducted another simulation, assuming $L = 500$ and $L = 3000$. In the former case simulation was repeated only twice because of the smaller variance of $\delta$ for this case, whereas in the latter there was only one replication. The results obtained are given in Table 4. As expected, the proportion of inapplicable cases declines as L increases, but even...
Table 4. Mean estimates (\( \hat{\delta} \)) of nucleotide substitutions and the frequencies of inapplicable cases for \( L = 500 \) and \( L = 3000 \). One-parameter (1-p), four-parameter (4-p), and six-parameter (6-p) methods are used for three different substitution schemes.

<table>
<thead>
<tr>
<th>Substitution scheme</th>
<th>True distance (( \delta ))</th>
<th>( \delta ) mean</th>
<th>Inapplicable cases</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-p method</td>
<td>4-p method</td>
<td>6-p method</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>3000</td>
<td>500</td>
</tr>
<tr>
<td>1-p scheme</td>
<td>0.5</td>
<td>0.51</td>
<td>0.49</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.99</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>2.17</td>
<td>1.99</td>
</tr>
<tr>
<td>4-p scheme</td>
<td>0.5</td>
<td>0.48</td>
<td>0.48</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.92</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>1.71</td>
<td>1.80</td>
</tr>
<tr>
<td>6-p scheme</td>
<td>0.5</td>
<td>0.46</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>1.0</td>
<td>0.78</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>2.0</td>
<td>1.23</td>
<td>1.20</td>
</tr>
</tbody>
</table>

with \( L = 3000 \) the proportion is substantially high when the 6-p method is used. The mean of \( \hat{\delta} \) for the case of 1-p substitution scheme is again close to the expected value for all the estimation methods. For the 4-p and 6-p substitution schemes, however, the 1-p method gives an underestimate, as expected. The 4-p method gives a good estimate except for the 6-p substitution scheme, whereas the 6-p method gives good estimates for all substitution schemes.

One reviewer of this paper has suggested that in the case of the 4-p substitution scheme we should examine the case of a smaller \( \delta \) value, because Takahata and Kimura's (1981) estimates of \( \delta \) from actual data are often smaller than 1. To see whether the \( \delta \) value affects our conclusion seriously or not, we conducted the same simulation study as the above by using a different set of parameters for the 4-p substitution scheme. The parameters used are \( \alpha = 0.00375 \), \( \beta = 0.015 \), \( \gamma = 0.001 \), and \( \theta = 0.5 \). We also examined another case of the 6-p substitution scheme in which the substitution parameters were considerably different from those used above. Namely, we used \( \alpha = 0.006 \), \( \beta = 0.002 \), \( \alpha_1 = 0.0011 \), \( \alpha_2 = 0.0025 \), \( \beta_1 = 0.0044 \), and \( \beta_2 = 0.01 \). The average substitution rate (k) was again 0.01, and the cases of \( \delta_1 = 0.5 \), 1.0, and 2.0 with \( L = 500 \) and \( L = 3000 \) were examined. The results obtained from these simulations have shown that the absolute values of the estimates of \( \delta \) are affected by these parameter values to some extent but the relative values obtained by the 1-p, 4-p, and 6-p methods remain virtually the same as those in our previous simulations. The proportions of inapplicable cases for the three methods were also nearly the same. Therefore, our conclusions seem to be quite general.

Discussion

From the present study it is clear that the 1-p or Jukes-Cantor method gives underestimates of \( \delta \) when \( \delta \) is large. Thus, it is possible, as was already noted by Kimura (1981), that Perler et al.'s (1980) nonlinear relationship between the estimate of \( \delta \) and evolutionary time is caused by their assumption of equal substitution rates among the four nucleotides (see also Gojobori et al. 1981). Therefore, when \( \delta \) is large, it is preferable to use the 4-p or 6-p method. However, when the number of nucleotides involved is small, the 4-p and 6-p methods are often inapplicable. In this case we will have to use the 1-p method. Fortunately, as long as \( \delta \) is smaller than 1, all the methods give more or less the same estimate.

Kimura (1981) suggests that when a number of homologous nucleotide sequences are to be compared the nucleotide frequencies (\( q_i \)) be estimated by using all the sequences rather than only the pair of sequences to be compared. This suggestion has been made to reduce the sampling error of the estimates of \( q_i \)'s. However, the nucleotide frequencies for the pair of nucleotide sequences to be compared and their common ancestral sequence may be substantially different from those for other sequences because of chance effects or selection. Therefore, it is not clear whether Kimura's suggestion or our suggestion (i.e., estimation of \( q_i \) from the pair of sequences to be compared) is better. In practice, however, our simulation has indicated that either method gives nearly the same estimate of \( \delta \). Therefore, this problem does not seem to be very important.

Earlier we mentioned that Kimura's 6-p method (32) gives an overestimate when the rate of nucleotide substitution is the same for all nucleotide pairs. The magnitude of the overestimate can be examined by considering the cases of \( \delta = 0.5 \), 1, and 2. When \( L = \infty \), his formula gives 0.527 for \( \delta = 0.5 \), 1.104 for \( \delta = 1 \), and 2.35 for \( \delta = 2 \). Therefore, the extent of overestimation seems to be about 15 percent or less when \( \delta < 2 \). When the 6-p substitution scheme is used, his formula gives 0.511, 1.035, and 2.087 for \( \delta = 0.5 \), 1, and 2, respectively. Therefore, the extent of overestima-
tion is smaller in this case. Needless to say, our formula (17) or (30) gives the correct value in these cases.

Acknowledgement. This work was supported by grants from the National Science Foundation and the National Institutes of Health.

References


Received January 5, 1982/Revised May 26, 1982