Using linear invariants for various models of nucleotide substitution, we developed test statistics for examining the applicability of a specific model to a given dataset in phylogenetic inference. The models examined are those developed by Jukes and Cantor (1969), Kimura (1980), Tajima and Nei (1984), Hasegawa et al. (1985), Tamura (1992), Tamura and Nei (1993), and a new model called the eight-parameter model. The first six models are special cases of the last model. The test statistics developed are independent of evolutionary time and phylogeny, although the variances of the statistics contain phylogenetic information. Therefore, these statistics can be used before a phylogenetic tree is estimated. Our objective is to find the simplest model that is applicable to a given dataset, keeping in mind that a simple model usually gives an estimate of evolutionary distance (number of nucleotide substitutions per site) with a smaller variance than a complicated model when the simple model is correct. We have also developed a statistical test of the homogeneity of nucleotide frequencies of a sample of several sequences that takes into account possible phylogenetic correlations. This test is used to examine the stationarity in time of the base frequencies in the sample. For Hasegawa et al.'s and the eight-parameter models, analytical formulas for estimating evolutionary distances are presented. Application of the above tests to several sets of real data has shown that the assumption of stationarity of base composition is usually acceptable when the sequences studied are closely related but otherwise it is rejected. Similarly, the simple models of nucleotide substitution are almost always rejected when actual genes are distantly related and/or the total number of nucleotides examined is large.

Introduction

In the reconstruction of phylogenetic trees by parametric methods such as distance and maximum likelihood methods it is important to find a mathematical model of nucleotide substitution that fits the dataset used. There have been several attempts to solve this problem using the generalized least squares (Bulmer 1991) and maximum likelihood methods (see, e.g., Felsenstein 1983; Ritland and Clegg 1987; Goldman 1993). These methods are known to have high statistical power, but their application to real datasets usually requires a prohibitive amount of computer time. In addition, there are several restrictive conditions that must be met for successful application of these methods.

First, these methods require the phylogenetic tree to be known for the purpose of model selection except for the case of two sequences (Tamura 1994). Second, in the case of the likelihood ratio test, the conventional asymptotic approximation to the distribution of the $\chi^2$ test statistic usually does not apply for actual sequence data, as noted by Goldman (1993). The reason for this is that the number of possible nucleotide combinations (configurations) in each site of sequence alignment grows exponentially as the number of sequences increases. To ensure that the expected number of observations of each nucleotide configuration is at least five (so-called rule of thumb for $\chi^2$ tests), one needs to study very long sequences. For example, there are only 16 dinucleotide configurations for two sequences compared. If the length of these sequences is greater than 100 (usually the case with real genes), the expected number of each configuration is large enough to ensure good approximation to the asymptotic distribution of the test statistic. However, when three sequences are examined (there are $4^3 = 256$ different dinucleotide configurations), the nucleotide sequence should be at least $256 \times 5 = 1,280$ nucleotides long if all configurations are equiprobable, or longer if the expected frequencies of some configurations are smaller than $\frac{1}{256}$. Clearly, in the case of the large number of sequences the requirement for the minimum length is unrealistically high. To overcome this
Mathematical Models of Nucleotide Substitution

Table 1 includes six models of nucleotide substitution that have often been used for phylogenetic inference. In addition, a new model with eight parameters is included. Note that the new eight-parameter model is the most complicated one and all other models are regarded as special cases of this model. The elements of the transition matrices for these models represent the transition frequencies of the four nucleotides A, T, C, and G remain the same throughout the evolutionary process. We call the former two models the nonstationarity models and the rest of the models the stationarity models. For the nonstationarity models we need test statistics that do not depend on the assumption of \( g_A = g_T = g_C = g_G = \frac{1}{4} \), whereas for the stationarity models we can use statistics that explicitly depend on estimated stationary nucleotide frequencies.

In this article we introduce three different types of tests. The first type is for the nonstationarity models and is based on single invariant properties of the expected nucleotide-pair frequencies that are independent of evolutionary time. We call this type of test a single-invariant test statistic. This approach has clear-cut advantages but requires a complete specification of the null-hypothesis model to be tested. That is, the topology and instantaneous substitution rates must be explicitly defined. The computational time required for this test is also usually enormous.

In this article we present a method for testing the applicability of a given substitution model to a set of DNA sequences without constructing phylogenetic trees. This can be done if we consider model-specific statistics that are independent of evolutionary time and phylogenetic relationships among the sequences. Since the datasets used in phylogenetic analysis typically include many sequences that are not very long (less than 2,000 nucleotides), we consider only statistics that are based on pairwise comparisons of sequences and are applicable to the same dataset, and this causes errors in estimating evolutionary distances. Therefore, an obvious strategy to choose appropriate distance measures is to find the simplest model that fits the data satisfactorily. For this reason, we first consider a simple model and then gradually examine more complicated ones.

### Table 1

<table>
<thead>
<tr>
<th>ORIGINAL</th>
<th>A</th>
<th>T</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
</tr>
<tr>
<td>I</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
</tr>
<tr>
<td>C</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
</tr>
<tr>
<td>G</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
<td>( \lambda )</td>
</tr>
</tbody>
</table>

**NOTE.**—In all matrices the ellipses on the diagonal replace the entry required to ensure that row sums are zero.

#### 1. Jukes-Cantor model:

\[
\begin{array}{cccc}
A & T & C & G \\
A & \lambda & \lambda & \lambda \\
T & \lambda & \lambda & \lambda \\
C & \lambda & \lambda & \lambda \\
G & \lambda & \lambda & \lambda \\
\end{array}
\]

#### 2. Kimura model:

\[
\begin{array}{cccc}
A & T & C & G \\
A & \beta & \alpha & \beta \\
T & \beta & \alpha & \beta \\
C & \beta & \alpha & \beta \\
G & \alpha & \beta & \beta \\
\end{array}
\]

#### 3. Tamura model:

\[
\begin{array}{cccc}
A & T & C & G \\
A & (1-\theta)\beta & \theta \alpha & \theta \beta \\
T & \theta \alpha & (1-\theta)\alpha & \theta \beta \\
C & \theta \beta & \theta \beta & (1-\theta)\alpha \\
G & \theta \beta & \theta \beta & \theta \beta \\
\end{array}
\]

#### 4. Tajima-Nei model:

\[
\begin{array}{cccc}
A & T & C & G \\
A & g_A & g_C & g_C \\
T & g_A & g_T & g_T \\
C & g_C & g_T & g_T \\
G & g_C & g_T & g_T \\
\end{array}
\]

#### 5. Hasegawa et al. model:

\[
\begin{array}{cccc}
A & T & C & G \\
A & g_A & g_C & g_C \\
T & g_T & g_A & g_A \\
C & g_C & g_A & g_A \\
G & g_C & g_A & g_A \\
\end{array}
\]

#### 6. Tamura-Nei model:

\[
\begin{array}{cccc}
A & T & C & G \\
A & g_A & g_C & g_C \\
T & g_T & g_A & g_A \\
C & g_C & g_A & g_A \\
G & g_C & g_A & g_A \\
\end{array}
\]

#### 7. Eight-parameter model:

\[
\begin{array}{cccc}
A & T & C & G \\
A & \beta_2 & \beta_3 & \alpha_4 \\
T & \beta_4 & \beta_4 & \beta_4 \\
C & \alpha_4 & \alpha_4 & \alpha_4 \\
G & \alpha_4 & \alpha_4 & \alpha_4 \\
\end{array}
\]
test. The second type of test is to examine the stationarity of nucleotide frequencies. The third type is based on multiple invariant properties of the expected nucleotide pair frequencies and requires the stationarity of base composition in the sequences. This test is called a multiple-invariant test.

**Single-Invariant Tests**

**Jukes and Cantor’s (1969) Model**

In this model any of the nucleotides A, T, C, and G are replaced by any other nucleotide with the same substitution rate ($\lambda$). Therefore, the expected proportion ($P$) of transitional differences (AG and TC) between a pair of sequences compared is equal to half the expected proportion ($Q$) of the transversional differences (all other differences). In other words, we have $2P - Q = 0$. This property holds true irrespective of the extent of sequence divergence and thus applies to any pair of sequences under investigation (see App. A). Therefore, the following quantity can be used for testing the Jukes-Cantor model.

$$JC = \sum_{i<j}^m (2\hat{P}_{ij} - \hat{Q}_{ij}),$$  \hspace{1cm} (1)

where $\hat{P}_{ij}$ and $\hat{Q}_{ij}$ are the observed values of $P_{ij}$ and $Q_{ij}$ between the $i$th and $j$th sequences, respectively, and $m$ is the number of sequences studied. Under the Jukes-Cantor model the expectation of $JC$ is 0. Following Bulmer (1991), we have

$$V(JC) = \sum_{i<j} \left[ (4P_{ij} + Q_{ij}) - (2P_{ij} - Q_{ij})^2 \right] + 2 \sum_{ij<k<l} \left[ 4(P_{ijkl} - P_{ij}P_{kl}) \right. + (Q_{ijkl} - Q_{ij}Q_{kl}) \\
- \left. (R_{ijkl} - P_{ij}Q_{kl} + R_{ij,kl} - Q_{ij}P_{kl}) \right] / n,$$  \hspace{1cm} (2)

where $P_{ijkl}$ is the expected proportion of sites where sequences $i$ and $j$ and sequences $k$ and $l$ simultaneously show a transitional difference, $Q_{ijkl}$ is the expected proportion of sites where sequences $i$ and $j$ and sequences $k$ and $l$ simultaneously show a transversional difference, $R_{ijkl}$ is the expected proportion of sites where sequences $i$ and $j$ show a transitional difference and sequences $k$ and $l$ show a transversional difference, and $n$ is the number of sites examined. We estimate $V(JC)$ by $\hat{V}(JC)$, obtained by replacing all parameters in the right-hand side of equation (2) by their estimates.

Since the distribution of $JC / \sqrt{V(JC)}$ is close to the standard normal distribution under Jukes and Cantor’s model, we can test the null hypothesis of $2P = Q$ by using

$$|JC| / \sqrt{V(JC)} > z_{a/2},$$  \hspace{1cm} (3)

where $\alpha$ is the significance level and $z_{a/2}$ is the value beyond which the standard normal density curve has area $\alpha/2$. This test is equivalent to testing the null hypothesis $\alpha = \beta$ in Kimura’s model.

**Kimura’s (1980) Model**

This model requires two parameters and assumes that the rate of transitional nucleotide substitution (\(\alpha\)) is different from that of transversional substitution (\(\beta\)). In this model the expected proportion ($T$) of nucleotide pairs AT and GC is equal to the expected proportion ($U$) of nucleotide pairs AC and GT between two sequences. Indeed, this is true even if the initial nucleotide frequencies are not $1/4$ (see App. A). Therefore, we can use the following statistic for testing the applicability of the Kimura model.

$$K = \sum_{i<j} (\hat{T}_{ij} - \hat{U}_{ij}),$$  \hspace{1cm} (4)

where $i$ and $j$ again refer to the $i$th and $j$th sequences. Under Kimura’s model the expectation of $K$ is 0, and the variance is given by

$$V(K) = \sum_{i<j} [Q_{ij} - (T_{ij} - U_{ij})^2] + 2 \sum_{ij<k<l} \left[ T_{ijkl} + U_{ijkl} - Z_{ijkl} \right] / n,$$  \hspace{1cm} (5)

where $T_{ijkl}$ is the proportion of sites where two pairs of sequences $i$ and $j$ and $k$ and $l$ have nucleotide pairs AC or GT, $U_{ijkl}$ is the proportion of sites where the two pairs have AG or CT, and $Z_{ijkl}$ is the proportion of sites where sequences $i$ and $j$ have AC or GT but sequences $k$ and $l$ have AG or CT. As in the case of the JC statistic, we can test the null hypothesis for $K$ using the two-tailed normal deviate test. The above $K$ test is equivalent to testing the equality of two transversional substitution rates (one is $\alpha \leftrightarrow C$ and $G \leftrightarrow T$, and another is $A \leftrightarrow T$ and $C \leftrightarrow G$).

**Test of the Stationarity of Nucleotide Frequencies**

As mentioned earlier, all currently used methods for estimating evolutionary distance (except those for
the Jukes-Cantor and Kimura models) require the assumption of the stationarity of nucleotide frequencies. Therefore, it is important to test the applicability of these methods by examining the stationarity of nucleotide frequencies. Since we cannot test the stationarity in time with present-day data, we use a phylogenetically aware test of the homogeneity of base frequencies across m sequences in the sample. If these base frequencies are not significantly heterogeneous, we assume that the base frequencies are at equilibrium in time.

Consider nucleotide sequences of length n, and let \( \hat{g}_1, \hat{g}_2, \ldots, \hat{g}_m \) be the expected frequencies of nucleotide x in each of the m sequences, where x is A, T, C, or G. If all the sequences evolve following the same model of nucleotide substitution and if the nucleotide frequencies are at equilibrium for the entire evolutionary process, we have

\[
\hat{g}_1 = \hat{g}_2 = \ldots = \hat{g}_m.
\]

This property will be used as the null hypothesis for our test of stationarity.

Under our null hypothesis this vector follows a distribution that approximates the multivariate normal distribution with mean \( \hat{g} = (\hat{g}_1, \ldots, \hat{g}_m) \) and the following test statistic,

\[
I = (\hat{g} - \hat{g}_e)'B'B(\hat{g} - \hat{g}_e).
\]

Here \( I \) follows a \( \chi^2 \) distribution with \( 3m - 3 \) degrees of freedom if the null hypothesis is valid (data not shown). In this case three degrees of freedom have been lost because we have estimated \( \hat{g}_e \) as \( \hat{g} \).

Multiple-Invariant Test

If the stationarity of base composition of the dataset under investigation cannot be rejected by the above test,
the next question is: "What is the simplest model that fits the data set?" We now consider this problem by developing another type of statistical test.

In this test we consider all restrictions imposed by a particular model of nucleotide substitution with respect to the expected frequencies of dinucleotide pairs between two homologous sequences. For two sequences the test is equal to the Wald test, which is asymptotically equivalent to the likelihood ratio test (see Rao 1973, pp. 417-419). For brevity, we describe only a simpler version of the test, where the goodness of fit of each of the six time-reversible models ([I] - [6] in table 1) is compared with that of a general time-reversible model. Since the expected frequency of nucleotide pair xy is equal to that of nucleotide pair yx for any time-reversible model even if the rates of nucleotide substitution vary among evolutionary lineages, we can consider pairs xy and yx as a single class. Therefore, we consider only 10 nucleotide pairs, that is, AA, AT, AC, AG, TT, TC, TG, CC, CG, and GG, rather than 16 pairs.

We denote the expected frequencies of these pairs for a particular pair of sequences by $X_1, X_2, \ldots, X_{10}$ in this order. It is then possible to find linear equations of $X_i$'s of the following form.

$$\sum_s a_s X_i + b = 0,$$

where $s$ refers to the $s$th nucleotide pair and $a$ and $b$ are parameters that are specific to each model. Note that the $K$ statistic for the Kimura model is based on a special case of this equation, where $a_2 = a_7 = 1, a_3 = a_9 = -1$, and all other $a_s$'s and $b$ are equal to 0. That is, for each pair of sequences

$$X_{AT} - X_{AC} - X_{GT} + X_{GC} = 0.$$

Similarly, the JC statistic is based on a special case of equation (12).

Once the relation satisfying (12) is found for a given model, it is possible to compute the following quantity.

$$S = \left( \sum_s a_s X_i \right) + b(m - 1)/2,$$

where $X_i = \sum_{s<i,j} X_{s,j}$, in which $X_{s,j}$ is the observed frequency of the $s$th nucleotide pair for the sequences $i$ and $j$, and $a_s$ and $b$ are unbiased estimates of $a_s$ and $b$, respectively, obtained from the data or specified by the model. Strictly speaking, the expectation of $S$ is not zero, but the deviation of $E(S)$ from zero is so small compared with the standard error that we can assume $E(S) = 0$ for practical purposes. (The deviation of $E(S)$ from zero is $\sum_i \text{Cov}(\hat{a}_i, \hat{X}_i)$.

For large $n$, $\text{Cov}(\hat{a}_i, \hat{X}_i)$ is of the order $1/n$, while $\hat{a}_i - a_i$ and $\hat{X}_i - \sum_{s<i,j} \hat{X}_{s,j}$ are of the order $1/\sqrt{n}$.)

In practice, in every model considered below $X_i$'s are subject to multiple restrictions satisfying equation (12), and it is desirable to find all independent linear equations of this kind. Since $X_{AA} + X_{AT} + \ldots + X_{GG} - 1$, we have only nine independent observations for each pair of sequences. For a complicated model of nucleotide substitution, the number of restrictions ($p$) can be smaller than nine, as will be shown later. At any rate, we can consider the following vector $S$ consisting of $p$ linearly independent relations.

$$S = (S_1, S_2, \ldots, S_p).$$

Let $W$ be the $p \times p$ asymptotic covariance matrix for $S_i$'s. Assuming that vector $S$ is asymptotically normally distributed with mean $(0, 0, \ldots, 0)$ and covariance matrix $W$, we can show that the statistic

$$T = S'W^{-1}S$$

follows a $\chi^2$ distribution. (The number of degrees of freedom, $p$, should be specified separately for each model.) Statistic (15) has a nice feature that if one decides to use a different set of $p$ independent linear restrictions, the value of $T$ still remains the same. This property allows us to arbitrarily choose $p$ independent restrictions for each model from all restrictions available.

Therefore, to test the applicability of a particular model of nucleotide substitution to a given set of data, we can compute $T$ and examine whether $T > \chi^2(p)$, where $\chi^2(p)$ is the value beyond which the density curve of the $\chi^2$ distribution with $p$ degrees of freedom has area $\alpha$. If $T$ is greater than $\chi^2(p)$, we reject the model at the significance level of $\alpha$. For this purpose, we must first compute $S$ from data and $W$ theoretically. The computation of these quantities can be done in the following way.

(1) Compute the following column vectors defined by

$$\hat{X}' = (\hat{X}_{AA}, \hat{X}_{AT}, \hat{X}_{AC}, \hat{X}_{AG}, \hat{X}_{TT}, \hat{X}_{TC}, \hat{X}_{TG}, \hat{X}_{CC}, \hat{X}_{CG}, \hat{X}_{GG})$$

where $\hat{X}_i = \sum_{s<i,j} \hat{X}_{s,j}$ as mentioned above. Here $\hat{g}_k (k = A, T, C, \text{and} G)$ is an average frequency of the $k$th nucleotide for all sequences, that is, $\hat{g}_k = (\hat{g}_{k1} + \hat{g}_{k2} + \ldots + \hat{g}_{km})/m$, where $\hat{g}_k$ is the observed frequency of the
where

\[ V(\hat{X}_{s,ij}) = X_{s,ij}(1 - X_{s,ij})/n. \]

\[ \text{Cov}(\hat{X}_{r,ij}, \hat{X}_{s,kl}) = (X_{r,ijkl} - X_{s,ijkl})/n, \]

and \( X_{r,ijkl} \) is the expected proportion of sites where sequences \( i \) and \( j \) have nucleotide pair \( r \) and sequences \( k \) and \( l \) have nucleotide pair \( s \).

We can compute matrix \( M \) by noting that, for all the models considered below, parameters \( a_i \)'s and \( b_i \)'s depend only on the stationary nucleotide frequencies, \( g_k \)'s. We first compute the matrix \( Y \) with elements \( y_{ik} = \sum_s (\partial a_{ik}/\partial g_k) \hat{X}_s + (\partial b_{ik}/\partial g_k) \), and then compute the matrix \( C \) with elements

\[ c_{ij} = a_{ij} + \sum_k y_{ik} m_{kj}, \]

where \( m_{kj} \) is the element of matrix \( M \) in equation (17). In other words, \( C = A + YM \).

(4) Compute the statistic \( T = S'W^{-1}S \), where \( W^{-1} \) is the inverse of \( W \).

Let us now consider the test statistic for each model.

**Tajima and Nei's (1984) and Jukes and Cantor's (1969) models**

Tajima and Nei's model is a special case of the eight-parameter model (see [41] and [71] in table 1). Using equation (B10) in Appendix B, we obtain the following formulas for the expected frequencies of nucleotides pairs.

\[
\begin{align*}
X_{AA} &= \frac{g_A^2}{g_A(1-g_A)} - \frac{g_A}{1-g_A} + \frac{1}{2g_A g_T} \\
X_{AT} &= \frac{g_A g_T}{g_A(1-g_A)} - \frac{g_T}{1-g_T} \\
X_{AC} &= \frac{2g_A g_C}{g_A(1-g_A)} - \frac{g_C}{1-g_C} \\
X_{AG} &= \frac{g_A g_G}{g_A(1-g_A)} - \frac{g_G}{1-g_G} \\
X_{TT} &= \frac{g_T^2}{g_T(1-g_T)} - \frac{g_T}{1-g_T} \\
X_{TC} &= \frac{2g_T g_C}{g_T(1-g_T)} - \frac{g_C}{1-g_C} \\
X_{TG} &= \frac{2g_T g_G}{g_T(1-g_T)} - \frac{g_G}{1-g_G} \\
X_{CG} &= \frac{2g_C g_G}{g_C(1-g_C)} - \frac{g_G}{1-g_G} \\
X_{GG} &= \frac{g_G^2}{g_G(1-g_G)} - \frac{g_G}{1-g_G}
\end{align*}
\]

where \( t \) is the time since divergence of the two sequences compared. This indicates that \( e^{-2t} \) can be expressed in terms of \( X_i \)'s in 10 different ways.

\[
e^{-2t} = \frac{X_{AA}}{g_A(1-g_A)} - \frac{g_A}{1-g_A} = \frac{X_{AT}}{2g_A g_T} + 1
\]

\[ = \frac{X_{AC}}{2g_A g_C} + 1 = \frac{X_{AG}}{2g_A g_G} + 1
\]

\[ = \frac{X_{TT}}{g_T(1-g_T)} - \frac{g_T}{1-g_T} = \frac{X_{TC}}{2g_T g_C} + 1
\]
Using pairs of expressions for $e^{-2k}$, we can obtain the following $(9 \times 10)$ matrix $A$ and $(9 \times 1)$ vector $B$ that satisfy equation (12). (We have nine independent linear equations. As we mentioned, this is not the only set of equations that satisfy eq. [12] but is one of the simplest ones.)

\[
A = \begin{bmatrix}
\frac{1}{1 - \hat{g}_\alpha} & \frac{1}{2\hat{g}_T} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{\hat{g}_A} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & \frac{1}{\hat{g}_C} & \frac{1}{\hat{g}_G} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & \frac{1}{2\hat{g}_A} & 0 & 0 & \frac{1}{1 - \hat{g}_T} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{\hat{g}_C} & \frac{1}{\hat{g}_G} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{1 - \hat{g}_C} & \frac{1}{2\hat{g}_G} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\hat{g}_T} & -\frac{1}{\hat{g}_G} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{\hat{g}_G} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{1 - \hat{g}_G} \\
\end{bmatrix}
\]

\[
B = \begin{bmatrix}
\hat{g}_A(m - 1)/2 \\
1 - \hat{g}_\alpha \\
0 \\
0 \\
\hat{g}_T(m - 1)/2 \\
1 - \hat{g}_T \\
0 \\
0 \\
\hat{g}_C(m - 1)/2 \\
1 - \hat{g}_C \\
\hat{g}_G(m - 1)/2 \\
1 - \hat{g}_G \\
\end{bmatrix}
\]

The matrix $Y$ is then given by

\[
Y = \begin{bmatrix}
\hat{X}_1 - m(m - 1)/2 & -\hat{X}_2(2\hat{g}_T) & 0 & 0 \\
0 & \hat{X}_2/\hat{g}_T^2 & 0 & -\hat{X}_3/\hat{g}_C^2 \\
0 & 0 & -\hat{X}_3/\hat{g}_C^2 & \hat{X}_4/\hat{g}_C^2 \\
-\hat{X}_2/(2\hat{g}_A^2) & \frac{\hat{X}_5 - m(m - 1)/2}{(1 - \hat{g}_T)^2} & 0 & 0 \\
-\hat{X}_2/\hat{g}_A^2 & 0 & 0 & -\hat{X}_6/\hat{g}_C^2 \\
0 & 0 & -\hat{X}_6/\hat{g}_C^2 & \hat{X}_7/\hat{g}_C^2 \\
0 & 0 & \frac{\hat{X}_8 - m(m - 1)/2}{(1 - \hat{g}_C)^2} & -\hat{X}_9/(2\hat{g}_C^2) \\
0 & 0 & -\hat{X}_7/\hat{g}_T^2 & \hat{X}_9/\hat{g}_C^2 \\
0 & 0 & -\hat{X}_7/(2\hat{g}_T^2) & 0 \frac{\hat{X}_{10} - m(m - 1)/2}{(1 - \hat{g}_C)^2} \\
\end{bmatrix}
\]
Computer simulation showed that under the Tajima-Nei model \( T \) follows a \( \chi^2 \) distribution with nine degrees of freedom.

Substituting \( \frac{1}{4} \) for \( g_A, g_T, g_C, \) and \( g_G \) in equation (22), we obtain the following stationarity-dependent relations for \( X_i \)'s for the Jukes-Cantor model: \( X_{AA} = X_{TT} = X_{CC} = X_{GG}, \) \( X_{AT} = X_{AC} = X_{AG} = X_{TC} = X_{TG} = X_{CG}, \) \( 3X_{AT}/2 + X_{AA} - \frac{1}{4} = 0. \) Using these relations, we can find the following expressions for \( A \) and \( B. \)

\[
A = \begin{bmatrix}
1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \\
1 & \frac{3}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}, \quad B = \begin{bmatrix}
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
-m(m-1)/8 \\
\end{bmatrix}
\]

Since the stationary nucleotide frequencies need not be estimated in this case, matrix \( Y \) contains only zero entries and \( T \) follows a \( \chi^2 \) distribution with nine degrees of freedom.


Using equations (B10) in Appendix B, we obtain the following expected frequencies of nucleotide pairs for Tamura's model:

\[
\begin{bmatrix}
X_{AA} \\
X_{AT} \\
X_{AC} \\
X_{AG} \\
X_{TT} \\
X_{TC} \\
X_{TG} \\
X_{CC} \\
X_{CG} \\
X_{GG} \\
\end{bmatrix} = -\frac{1}{4} \begin{bmatrix}
(1 - \theta)^2 & 2\theta(1 - \theta) & (1 - \theta)^2 \\
2(1 - \theta)^2 & 2\theta(1 - \theta) & 0 \\
2\theta(1 - \theta) & (1 - \theta)^2 & 2\theta(1 - \theta) \\
2\theta(1 - \theta) & 0 & 2\theta(1 - \theta) \\
\theta^2 & 2\theta(1 - \theta) & 0 \\
\theta^2 & 0 & \theta^2 \\
\end{bmatrix} \begin{bmatrix}
1 \\
\text{e}^{-2(\alpha + \beta)t} \\
\text{e}^{-4\beta t} \\
\end{bmatrix}.
\]

Equation (26) provides the following set of restrictions that are specific to Tamura's model.

\[
X_{AA} = X_{TT}, \quad X_{CC} = X_{GG},
\]

\[
X_{AG} = X_{TC}, \quad X_{AC} = X_{TG},
\]

\[
e^{-4\beta t} = -\frac{2X_{AT}}{(1 - \theta)^2} + 1 = -\frac{2X_{AC}}{\theta(1 - \theta)} + 1 = -\frac{2X_{GC}}{\theta^2} + 1 = \frac{4X_{AA}}{1 - \theta} + \frac{2X_{AG}}{1 - \theta} - 1
\]

\[
= \frac{4X_{GG}}{0} + \frac{2X_{AG}}{0} - 1 = \frac{4X_{AA}}{1 - 2\theta} - \frac{4X_{GG}}{1 - 2\theta} - 1,
\]

where \( \theta \neq \frac{1}{2}. \) Substitution of equations (27) into (28) gives other expressions for \( e^{-4\beta t} \) in terms of \( X_i \)'s. In a similar fashion one can find expressions for \( e^{-2(\alpha + \beta)t} \) in terms of \( X_i \)'s. However, we do not consider these equations, because they do not give additional independent statistics. At any rate, using pairs of expressions in (27) and (28), we obtain the following matrix \( A \) and vector \( B \) that satisfy (12).
The corresponding matrix $Y$ is given by

$$
Y = \begin{bmatrix}
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    -\frac{X_2}{2\theta^2} + \frac{X_3(1 - \theta)^2}{[\theta(1 - \theta)]^2} & -\frac{X_3}{2\theta^2} + \frac{X_4(1 - \theta)^2}{[\theta(1 - \theta)]^2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    -\frac{X_2}{2\theta^2} - \frac{m(m - 1)/2}{2\theta^2} + \frac{m(m - 1)/2}{2\theta^2} & -\frac{X_3}{2\theta^2} - \frac{m(m - 1)/2}{2\theta^2} + \frac{m(m - 1)/2}{2\theta^2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}
$$

In this case the test statistic $T$ follows a $\chi^2$ distribution with eight degrees of freedom.

Substituting $\frac{1}{2}$ for $\theta$ in equations (26), we obtain the stationarity-dependent restrictions for the Kimura model: $X_{AA} = X_{TT} = X_{CC} = X_{GG}, X_{AT} = X_{AC} = X_{CG}, X_{AG} = X_{TC}, 4X_{AA} + 4X_{AT} + 2X_{AG} - 1 = 0$. These relations lead to the following expressions for $A$ and $B$.

$$
A = \begin{bmatrix}
    1 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 \\
    0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\
    0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\
    0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    1 & 1 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\
\end{bmatrix}, \quad B = \begin{bmatrix}
    0 \\
    0 \\
    0 \\
    0 \\
    0 \\
    0 \\
    0 \\
    0 \\
    -(1 - \theta)m(m - 1)/2 \\
    -\theta m(m - 1)/2 \\
\end{bmatrix}
$$

In this case matrix $Y$ consists of zeros and $T$ follows the $\chi^2$ distribution with eight degrees of freedom.
Rzhetsky and Nei

Hasegawa, Kishino, and Yano's (1985) and Tamura and Nei's (1993) Models

Invariant relationships of $X_i$'s for these two models appear to be identical. Indeed, we have

$$
\begin{bmatrix}
X_{AA} \\
X_{AT} \\
X_{AC} \\
X_{AG} \\
X_{TT} \\
X_{TC} \\
X_{TG} \\
X_{CC} \\
X_{CG} \\
X_{GG}
\end{bmatrix} =
\begin{bmatrix}
g_A^2 & g_Ag_G/g_R & 0 & g_A^2 g_Y/g_R \\
2g_Ag_T & 0 & 0 & -2g_Ag_T \\
2g_Ag_C & 0 & 0 & -2g_Ag_C \\
2g_Ag_G & -2g_Ag_G/g_R & 0 & 2g_Ag_Gg_Y/g_R \\
g_T^2 & g_Tg_C/g_Y & g_T^2 g_R/g_Y & E
\end{bmatrix}
$$

(31)

where $E$ is a column vector given by

$$
E' = (1, e^{-2(a_{GR}+b_{GY})t}, e^{-2(a_{GR}+b_{GR})t}, e^{-2b_t}),
$$

and

$$
E' = (1, e^{-2(a_{GR}+b_{GY})t}, e^{-2(a_{GR}+b_{GR})t}, e^{-2b_t})
$$

(32)

for Hasegawa et al.'s and Tamura and Nei's models, respectively. Equation (31) gives the following relationships of $X_i$'s:

$$
e^{-2b_t} = \frac{-X_{AT}/(2g_Ag_T) + 1 = -X_{AC}/(2g_Ag_C) + 1 = -X_{AG}/(2g_Ag_G) + 1 = -X_{TT}/(g_Tg_C) + X_{TT}/(g_Tg_R) - g_Y/g_R}{X_{CC}/(g_Cg_R) + X_{TC}/(2g_Cg_R) - g_Y/g_R - X_{TT}/(g_Tg_C) + X_{TT}/(g_Tg_R) - g_Y/g_R}
$$

(33)

(All the identities of $X_i$'s obtainable by using the second and third elements of vector $E$ are linear combinations of eq. [33].) Using the same strategy as before, we obtain

$$
A =
\begin{bmatrix}
0 & -\frac{1}{\hat{g}_T} & \frac{1}{\hat{g}_G} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -\frac{1}{\hat{g}_A} & 0 & 0 & 0 & 0 & \frac{1}{\hat{g}_G} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -\frac{1}{\hat{g}_T} & 0 & \frac{1}{\hat{g}_C} & 0 \\
\frac{1}{\hat{g}_A} & 0 & 0 & \hat{g}_G - \hat{g}_A & 0 & 0 & 0 & 0 & -\frac{1}{\hat{g}_G} & 0 \\
0 & 0 & 0 & 0 & 1 & \hat{g}_C - \hat{g}_T & 0 & 1 & \hat{g}_C & 0 \\
0 & 0 & 0 & 0 & 0 & \frac{1}{\hat{g}_T} & 2\hat{g}_G\hat{g}_T & 0 & 1 & \hat{g}_C & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\hat{g}_R} & 0 & \frac{1}{\hat{g}_R} & 0 \\
\frac{1}{\hat{g}_Y} & \frac{1}{\hat{g}_T} & 0 & 0 & \frac{1}{\hat{g}_Y} & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}
$$

$$
B =
\begin{bmatrix}
0 \\
0 \\
0 \\
-\hat{g}_C(m-1)/(2\hat{g}_R) \\
-\hat{g}_A(m-1)/(2\hat{g}_Y)
\end{bmatrix}
$$

(34)
For the two models considered here, $T$ follows a $\chi^2$ distribution with five degrees of freedom.

Since the pattern of nucleotide pair frequencies is identical for the Hasegawa et al.'s and Tamura and Nei's models, we have to use an additional method to distinguish between these two models. Note that the only difference between these two models (see [5] and [6] in table 1) is that parameters $\alpha_1$ and $\alpha_2$ in Tamura and Nei's model are equal to each other in Hasegawa et al.'s model. (One consequence of this difference is that Hasegawa et al.'s model provides multiple estimators of evolutionary distance [number of nucleotide substitutions] rather than only one; see App. C.) Therefore, we can use the following statistic:

$$HKY = \sum_{i<j} H_{ij}, \quad (36)$$

where $H_{ij}$ is the difference between the estimates of $2\alpha_1t$ and $2\alpha_2t$ for sequences $i$ and $j$ and is expected to be zero under Hasegawa et al.'s model but is different from zero under Tamura and Nei's model. $H_{ij}$ can be estimated (see App. C) by

$$H = \frac{\hat{g}_Y^2 - \hat{g}_R^2}{\hat{g}_R \hat{g}_Y} \log_e \left( 1 - \frac{\hat{Q}}{2\hat{g}_R \hat{g}_Y} \right)$$

$$- \frac{1}{\hat{g}_R} \log_e \left( 1 - \frac{\hat{Q}}{2\hat{g}_R} - \frac{\hat{P} \hat{g}_R}{2\hat{g}_A \hat{g}_G} \right) \quad (37)$$

$$+ \frac{1}{\hat{g}_Y} \log_e \left( 1 - \frac{\hat{Q}}{2\hat{g}_Y} - \frac{\hat{P} \hat{g}_Y}{2\hat{g}_C \hat{G}_T} \right)$$

and

$$\begin{bmatrix}
0 & X_1/\hat{g}^2 & -X_1/\hat{g} \hat{c} & 0 & 0 & 0 \\
X_1/\hat{g}^2 & 0 & -X_2/\hat{g} \hat{c} & 0 & 0 & 0 \\
0 & X_1/\hat{g} \hat{c} & 0 & 0 & -X_3/\hat{g} \hat{o} & 0 \\
-(2\hat{X}_1 + \hat{X}_3)/(2\hat{g} \hat{c}) & 0 & (2\hat{X}_1 + \hat{X}_3)/(2\hat{g} \hat{c}) & 0 & 0 & -(2\hat{X}_10 + \hat{X}_3)/(2\hat{g} \hat{o}) \\
0 & -(2\hat{X}_5 + \hat{X}_8)/(2\hat{g} \hat{o}) & 0 & -m(m-1) & 0 & 0 \\
-\frac{m(m-1)}{2\hat{g}^2} & -\frac{[2\hat{X}_1 + \hat{X}_4 - \hat{g}_o m(m-1)]}{2\hat{g} \hat{c}} & \frac{\hat{X}_4}{2\hat{g} \hat{o}} & -\frac{[2\hat{X}_1 + \hat{X}_4 - \hat{g}_o m(m-1)]}{2\hat{g} \hat{c}} & 0 & 0 \\
\end{bmatrix} \cdot$$

where $\hat{Q} = \hat{X}_{AT} + \hat{X}_{AC} + \hat{X}_{GC} + \hat{X}_{GT}$, $\hat{R}_1 = \hat{X}_{AG}$, and $\hat{R}_2 = \hat{X}_{CT}$. The variance of HKY in (36) is computed in the following way.

$$V(HKY) = \sum_{i<j} V(H_{ij}) + 2 \sum_{i<j<k} Cov(H_{ij}, H_{jk}). \quad (38)$$

Here, the variances and covariances of $H_{ij}$'s can be computed by

$$Cov(H_{ij}, H_{kl}) = a_{ij} U a_{kl}^T, \quad (39)$$

where $U$ is the covariance matrix for vectors ($\hat{Q}_{ij}$, $\hat{P}_{1,ij}$, $\hat{P}_{2,ij}$) and ($\hat{Q}_{kl}$, $\hat{P}_{1,kl}$, $\hat{P}_{2,kl}$), $a_{ij} = (a_{ij}, b_{ij}, c_{ij})$, $a_{kl} = (a_{kl}, b_{kl}, c_{kl})$, and the elements $a_{ij}$, $b_{ij}$, and $c_{ij}$ are given by

$$a = \left[ 2g^\frac{1}{R} \left( 1 - \frac{Q}{2g_R} - \frac{P_{1,GR}}{2g_A \hat{g}_G} \right) \right]^{-1}$$

$$- \left[ 2g^\frac{1}{Y} \left( 1 - \frac{Q}{2g_Y} - \frac{P_{2,GR}}{2g_C \hat{g}_T} \right) \right]^{-1}$$

$$+ \frac{g^\frac{1}{Y} - g^\frac{1}{R}}{2(g_R g_Y)} \left( 1 - \frac{Q}{2g_R g_Y} \right), \quad (40)$$

$$b = \left[ 2g_A g_G \left( 1 - \frac{Q}{2g_R} - \frac{P_{1,GR}}{2g_A \hat{g}_G} \right) \right]^{-1}.$$
and
\[ c = -\left[ 2g_C g_T \left( 1 - \frac{Q}{2g_Y} - \frac{P_2 g_Y}{2g_C g_T} \right) \right]^{-1}, \]

for every pair of sequences \( i \) and \( j \). The variances and covariances of \( \hat{Q}, \hat{P}_1, \) and \( \hat{P}_2 \) are computed by using formulas that are analogous to those in equation (20). Here we treat the estimates of equilibrium nucleotide frequencies, \( \hat{g}'s \), as constants, because they are estimated from all \( m \) sequences. By contrast, \( \hat{Q}'s, \hat{P}_1's, \) and \( \hat{P}_2's \) are estimated each time from a pair of sequences. Therefore, the variances of \( \hat{g}'s \) are much smaller than those of \( \hat{Q}'s, \hat{P}_1's, \) and \( \hat{P}_2's \).

Once the statistic HKY and its variance are computed, we can test the significance of the deviation of HKY from zero using the normal deviate test. Note that the HKY statistic may not be applicable if the sequences under investigation are too divergent, because the argument of logarithm in (37) may become negative.

**Eight-Parameter Model**

This model does not give simple statistics of the form of equation (13), where \( a_i's \) are functions of stationary nucleotide frequencies. However, it is possible to find invariant relations of the form (12), where \( a_i's \) are more sophisticated functions of \( \hat{g}'s \) and \( \hat{X}'s \). For example, the following identity, which is due to Nguyen and Speed (1992), holds for the eight-parameter model

\[ s_1 s_2 X_{AC} - s_1 X_{AT} - s_2 X_{GC} + X_{GT} = 0, \]  

(41)

where \( s_1 = \beta_3/\beta_2, \) \( s_2 = \beta_4/\beta_1, \) and \( \beta_i's \) are the substitution rates for eight-parameter model (see [7] in table 1). It is possible to estimate \( 2\beta_i't's, s_1, \) and \( s_2 \) from a set of sequences (see formulas [B16] in App. B) and then use these estimates to construct a statistic for testing the eight-parameter model. However, since resulting formulas are quite cumbersome, we will not elaborate them in this article.

**Numerical Example**

Using the above approach we analyzed several datasets. We found that the simpler models of nucleotide substitution are applicable only when real datasets (the number of sequences and/or sequence length) are small and sequences are closely related. As the number of sequences and/or their length increases, simple models of nucleotide substitution almost always fail to fit the data. Here we present analysis of three datasets: \( \beta_2 \)-microglobulin genes from mouse, rat, human, chimpanze gorilla, and orangutan; cyclophilin genes from rat, mouse, human, and clawed frog; and mitochondrial encoded NADH dehydrogenase subunit 2 genes from 32 species of cichlid fish. The first two datasets were obtained from GenBank release 72, and the third one was kindly provided by Kocher et al. (1994). After excluding gaps and ambiguous nucleotides, 330, 492, and 1,048 nucleotides were used for the \( \beta_2 \)-microglobulin cyclophilin, and NADH dehydrogenase subunit 2 genes, respectively (\( \beta_2 \)-microglobulin and cyclophilin are th light chain of major histocompatibility class complex class protein and the cellular receptor for the immunsup pressive drug cyclosporin A, respectively). The result obtained are as follows.

The first two datasets provide examples of a relatively rare case when actual sequences are compatibl with one of the simplest model of nucleotide substitutio (Kimura's model). The nucleotide frequencies in bot datasets were close to \( \frac{1}{4} \). That is, \( \hat{g}_A = 0.28, \hat{g}_T = 0.22, \hat{g}_C = 0.24, \) and \( \hat{g}_G = 0.22 \) for \( \beta_2 \)-microglobulin and cyclophilin, respectively. The K statistic was not significant in both data sets. The \( T \) statistic als rejected the Jukes-Cantor model but not the Kimui model. However, it rejected the Tajima-Nei model for cyclophilin dataset (\( T = 40.60 \) with 9 df). All oth models were not rejected by the \( T \) test. Therefore, fit these datasets Kimura's model was the simplest mod applicable.

In the case of the NADH dehydrogenase subunit gene the average frequencies of nucleotides C and T were clearly different from \( \frac{1}{4} \) (\( \hat{g}_A = 0.258, \hat{g}_T = 0.270, \hat{g}_C = 0.353, \) and \( \hat{g}_G = 0.119 \)). The nucleotide frequencies also varied from species to species: \( \hat{g}_{Ai} \) ranged from 0.24 (Perissodus microlepis) to 0.268 (Tanganicodus irsaca), \( \hat{g}_{Ti} \) from 0.241 (Cichlasoma citrinellum) to 0.25 (T. irsaca), \( \hat{g}_{Ci} \) from 0.331 (T. irsaca) to 0.379 (C. citrinellum), and \( \hat{g}_{Gi} \) from 0.109 (Lepidolamprologus elongatus) to 0.129 (P. microlepis). The assumption (stationary nucleotide frequencies was rejected for the data when the entire sequences or third-codon position only were used (\( I = 152.2 \) with 93 df for the entire se quences and \( I = 210.0 \) with 93 df for third-codon p ositions). Since the cichlid fish species considered he: are evolutionarily closely related (Kocher et al. 1994 the base composition in third-codon positions of the NADH dehydrogenase subunit 2 gene seems to hav
changed surprisingly rapidly. However, when only first- and second-codon positions were analyzed, the assumption of the stationarity of nucleotide frequencies was not rejected ($I = 91.0$ with $93 \text{ df}$). For all the three NADH dehydrogenase datasets of entire sequences, third-codon positions, and first- and second-codon positions, the JC test rejected Jukes and Cantor's model ($JC/\sqrt{V(JC)} = 16.3, 17.9, \text{ and } 7.0$, respectively), whereas the K test did not reject Kimura's model ($K/\sqrt{V(K)} = 1.25, 1.54, \text{ and } 0.13$, respectively). Application of the $T$ test showed that the Hasegawa-Kishino-Yano and the Tamura-Nei models are the simplest models that can not be rejected for the dataset of first- and second-codon positions. However, the HKY statistic rejected Hasegawa et al.'s model. Therefore, our tests suggested that Tamura and Nei's model be used for phylogenetic inference for the cichlid fish sequences considered here.

We have conducted the stationarity test for a number of various datasets. Our results indicated that the stationarity assumption is usually acceptable for closely related sequence data. However, for distantly related genes such as those from different orders of mammals and amphibians, this assumption was almost always rejected.

**Discussion**

The $T$ test that we suggest in this article can be regarded as a computationally efficient approximation to the likelihood ratio test comparing each model with a general time-reversible Markov model (A. Rzhetsky, unpublished manuscript). For two sequences the $T$ statistic is equivalent to the Wald statistic, and for more than two sequences it is an approximation to the latter. The Wald test is in turn asymptotically equivalent to the likelihood ratio test. For most datasets, all the tests presented in this article can be executed by a personal computer. By contrast, a rigorous application of the likelihood ratio test may require a supercomputer to analyze even a sizable number of sequences.

Some readers may wonder how the linear invariants we used are related to those of Lake (1987), Cavender (1989), Nguyen and Speed (1992), Evans and Speed (1993), and others. Actually, these authors have identified various linear invariants for a given substitution model and used them for phylogenetic inference rather than for statistical tests of applicability of various models. Therefore, their purpose as well as the linear invariants they used are different from ours.

The objective of our approach is to identify the simplest substitution model that fits the data because a simpler model generally provides estimator of evolutionary distance with a smaller variance than those for a more complicated model, given both models are compatible with the data. The problem is that all simple models tested may not fit a given set of data according to our tests (or the likelihood ratio test; see Yang 1994). What should we do then? One solution to this question would be to use a more complicated model such as those proposed by Lanave et al. (1984) or by Barry and Hartigan (1987) (see also Lake 1994; Stccl 1994; Zharkikh 1994). However, the actual data would almost never fit any mathematical model perfectly. We are then forced to find a model that fit the data sufficiently well for a given practical purpose (such as phylogenetic inference with a particular algorithm), even though the deviation from the model is statistically significant. For this purpose we need a specialized statistical test. Such a test without constructing phylogenetic trees is now under investigation.

**Computer Program**

A computer program for IBM PC-compatible computers for computing all the statistics developed in the article is available on request.

**Acknowledgments**

We are grateful to Koichiro Tamura, Terence P. Speed, and Simon Tavarè for helpful discussions. We thank two anonymous reviewers, and especially Andrey Zharkikh and Stanley Sawyer, for their helpful comments on the manuscript. Special thanks to Margarita Rzhetskaya for assistance with typing mathematical equations. This work was supported by grants from the National Institutes of Health and the National Science Foundation to M.N.

**APPENDIX A**


Consider two nucleotide sequences that diverged $t$ time units ago from an ancestral sequence that had the frequencies of nucleotides $A$, $T$, $C$, and $G$ equal to $g_A$, $g_T$, $g_C$, and $g_G$, respectively. Since the Jukes-Cantor model is a special case of the Kimura model, we consider the latter model and assume that the two sequences evolve following the Kimura model of nucleotide substitution with parameters $\alpha_1$ and $\beta_1$ and $\alpha_2$ and $\beta_2$, respectively. That is, we assume that the substitution rates in two evolutionary lineages are not necessarily equal to each other. If we denote by $f_{xy}$ the probability of observing nucleotide pair $x$ and $y$ at a homologous site of these two sequences, we obtain
\[
\begin{bmatrix}
    f_{AA} \\
    f_{AT} \\
    f_{TA} \\
    f_{AC} \\
    f_{CA} \\
    f_{AG} \\
    f_{GA} \\
    f_{TT} \\
    f_{TC} \\
    f_{CT} \\
    f_{TG} \\
    f_{GT} \\
    f_{CC} \\
    f_{CG} \\
    f_{GC} \\
    f_{GG}
\end{bmatrix} = \frac{1}{16}
\begin{bmatrix}
    1 & 2a & 2a & c & c & 2a & 2a & 4g_R & 1 \\
    1 & 2a & 2b & c & -c & -2b & -2a & 0 & -1 \\
    1 & 2b & 2a & -c & c & -2b & -2a & 0 & -1 \\
    1 & 2a & -2b & 2a & c & c & 2b & 0 & -1 \\
    1 & -2b & 2a & c & c & -2b & -2a & 0 & -1 \\
    1 & 2a & -2a & c & 2a & -2a & 4g_R & 1 \\
    1 & -2a & 2a & c & 2a & -2a & -4g_R & 1 \\
    1 & 2b & 2b & -c & c & -2b & -2a & 0 & -1 \\
    1 & 2b & -2b & 2b & -c & c & -2b & -2a & 0 & -1 \\
    1 & 2b & -2a & c & c & 2b & 0 & -1 \\
    1 & -2b & 2b & -c & c & -2b & -2a & 0 & -1 \\
    1 & 2b & 2b & -c & c & 2b & 0 & -1 \\
    1 & 2b & -2a & c & c & 2b & 0 & -1 \\
\end{bmatrix},
\]

where \( a = g_A - g_G, b = g_T - g_C, g_R = g_A + g_G, g_Y = g_C + g_T, \) and \( c = g_R - g_Y. \) Note that in general, \( f_{xy} \neq f_{yx} \) for \( x \neq y \) under the present model.

Because of the symmetry of the substitution rate matrix (see [1] and [2] in table 1), the expected number of nucleotide substitutions per site \( (d) \) is independent of the base composition in the ancestral sequence. In other words, for estimating \( d \) by the Jukes-Cantor or Kimura method, there is no need of the assumption of the equilibrium nucleotide frequencies, which are all 0.25. Denoting \( X_{xy} = f_{xy} + f_{yx} (x \neq y) \), we find that

\[
P = X_{AG} + X_{TC} = [1 - 2e^{-2(\alpha+\beta)(r_1+r_2)t} + e^{-4\beta(r_1+r_2)t}]/4,
\]

\[
Q = X_{AT} + X_{AC} + X_{TG} + X_{GC} = [1 - e^{-4\beta(r_1+r_2)t}]/2.
\]

The above equations are of the same form as the equivalent formulas derived by Kimura and become identical with the latter when \( r_1 = r_2 = 1. \) Therefore, Kimura's formulas for estimating the number of nucleotide substitutions can be used even when \( r_1 \neq r_2. \) In the Jukes-Cantor model \( \alpha = \beta. \) Therefore, we have

\[
2P - Q = 0,
\]

This justifies the JC statistic in text (eq. [1]). Note that this is also independent of the initial base composition.

From equation (A1), we can derive

\[
X_{AT} + X_{GC} = X_{AC} + X_{TG} = (1 - e^{-4\beta(r_1+r_2)t}))/4.
\]

Therefore, the expected value of the statistic \( K \) in text (expression [4]) is zero.

**APPENDIX B**

**Derivation of Basic Formulas for the Eight-Parameter Model**

Consider a continuous-time Markov model of nucleotide substitution with instantaneous transition matrix, \( Q, \) given in table 1, (7). The eigenvalues, \( \lambda_1, \lambda_2, \lambda_3, \) and \( \lambda_4, \) of \( Q \) are given by

\[
\begin{align*}
\lambda_1 &= 0, \\
\lambda_2 &= -(\alpha_1 + \alpha_4 + \beta_2 + \beta_3), \\
\lambda_3 &= -(\alpha_2 + \alpha_3 + \beta_1 + \beta_4), \\
\lambda_4 &= -(\beta_1 + \beta_2 + \beta_3 + \beta_4).
\end{align*}
\]

The equilibrium nucleotide frequencies corresponding to matrix \( Q \) can be found by solving the following system of equations.

\[
\begin{align*}
\mathbf{g}'Q &= \mathbf{0}, \\
\mathbf{u}g &= 1,
\end{align*}
\]

where \( \mathbf{g}' = (g_A, g_T, g_C, g_G) \) is a column vector of equilibrium frequencies of nucleotides A, T, C, and G, respectively, \( \mathbf{u} = (1, 1, 1, 1), \) and \( \mathbf{0}' = (0, 0, 0, 0). \) In the present case we have

\[
\begin{align*}
g_A &= (\beta_2 + \beta_3)\beta_1 + (\beta_1 + \beta_4)\alpha_1)/\lambda_2\lambda_4, \\
g_T &= (\beta_2 + \beta_3)\alpha_2 + (\beta_1 + \beta_4)\beta_2)/\lambda_3\lambda_4, \\
g_C &= (\beta_2 + \beta_3)\alpha_3 + (\beta_1 + \beta_4)\beta_3)/\lambda_3\lambda_4, \\
g_G &= (\beta_2 + \beta_3)\beta_4 + (\beta_1 + \beta_4)\alpha_4)/\lambda_2\lambda_4, \\
g_R &= g_A + g_G = -(\beta_1 + \beta_4)/\lambda_4, \\
g_Y &= g_C + g_T = -(\beta_2 + \beta_3)/\lambda_4.
\end{align*}
\]

\( g_R \) is independent of the initial base composition.
where $\lambda_1, \lambda_2, \lambda_3,$ and $\lambda_4$ are as given in equation (B1). We can rewrite matrix $Q$ in terms of equilibrium nucleotide frequencies as follows.

$$Q = \begin{pmatrix}
\ldots & \beta_2 g_T & \beta_2 + g_Y \Delta_2 / g_R & \beta_2 + g_Y \Delta_2 / g_R & \alpha_4 g_G \\
[\beta_2 + g_G \Delta_1 / g_R + \alpha_2 g_T] a_2 g_T & \ldots & \alpha_3 g_C & [\beta_2 - g_A \Delta_1 / g_Y + g_C \Delta_2 / g_R] g_A & \ldots \\
[\beta_2 + g_G \Delta_1 / g_R + g_C \Delta_2 / g_R] a_2 g_T & \ldots & [\beta_2 - g_A \Delta_1 / g_Y + g_C \Delta_2 / g_R] g_A & \beta_2 g_T & [\beta_2 + g_Y \Delta_2 / g_R] g_C & \ldots
\end{pmatrix}, \quad (B4)$$

where $\Delta_1 = \alpha_4 - \alpha_1,$ and $\Delta_2 = \alpha_2 - \alpha_3.$ (When $Q$ is presented in this fashion, it is clear that the eight-parameter model does not belong to the class of time-reversible models, i.e., $g_i Q_{ij} \neq g_j Q_{ji}$.)

Let us introduce the following additional notations.

$$v_1 = [g_A, g_C, g_G, g_Y], \quad v_2 = [-1, 0, 1, 0],$$

and

$$v_3 = [0, -1, 1, 0],$$

$$v_4 = [-(g_A + y_1) / g_R, (g_T + y_2) / g_Y, (g_C - y_2) / g_Y, (g_G - y_1) / g_R],$$

where $u_i$ and $v_i$ stand for the right and left eigenvectors corresponding to eigenvalue $\lambda_i,$ respectively. Note that the eigenvectors in equation (B6) are determined to satisfy the equation $u_i v_j = \delta_{ij}$ for any $i$ and $j,$ where $\delta_{ij}$ is Kronecker's delta. Now we compute a matrix $P(t),$ where the element $P(i \rightarrow j; t)$ is the conditional probability for nucleotide $i$ to be replaced by nucleotide $j$ at time $t.$ $P(t)$ is computed according to

$$P(t) = \sum_{i=1}^{4} v_i u_i \exp \{ \lambda_i t \}, \quad (B7)$$

It is convenient to rewrite matrix $P(t)$ in a vector form.
Equations (B8) can be directly used for inferring phylogenetic trees by the maximum likelihood method with the eight-parameter model.

The next step is to compute the expected frequencies, \( X_{xy} \)'s, of nucleotide pair \( xy \) at a homologous site of two present-day sequences that diverged \( t \) time units ago. Assuming the stationarity of base composition in the nearest common ancestor of these two sequences, one can compute \( X_{xy} \)'s by

\[
\begin{bmatrix}
X_{AA} \\
X_{AT} \\
X_{AC} \\
X_{AG} \\
X_{TT} \\
X_{TC} \\
X_{TG} \\
X_{CC} \\
X_{CG} \\
X_{GG}
\end{bmatrix} = \begin{bmatrix}
g_A^2 \\
2g_Ag_T \\
2g_Ag_C \\
2g_Ag_G - 2g_Ag_G + g_Yy_1^2 \\
g_T^2 \\
2g_Tg_C \\
2g_Tg_G \\
g_C^2 \\
2g_Gg_C \\
g_G^2 \\

g_Ag_G + g_Yy_1^2 \\
\frac{g_Ag_G + g_Yy_1^2}{g_R} \\
\frac{g_Ag_G + g_Yy_1^2}{g_R} \\
\frac{g_Ag_G + g_Yy_1^2}{g_R} \\
\frac{g_Ag_G + g_Yy_1^2}{g_R} \\
\frac{g_Ag_G + g_Yy_1^2}{g_R}
\end{bmatrix}
\begin{bmatrix}
z_1^2k \\
0 \\
0 \\
2z_2z_4k \\
z_2^2/k \\
0 \\
0 \\
z_3^2k \\
0 \\
p
\end{bmatrix}
\]

where \( A' = (1, e^{2\lambda t}, e^{2\lambda t}, e^{2\lambda t}, e^{(\lambda_2+\lambda_3)t}, e^{(\lambda_2+\lambda_4)t}, z_1 = g_A + y_1, z_2 = g_T + y_2, z_3 = g_G - y_1, z_4 = g_C - y_2, \) and \( k = g_Y/g_R. \) Let us introduce the following notations.

\( Q = X_{AT} + X_{AC} + X_{GT} + X_{GC}, \quad P_1 = X_{AG}, \quad P_2 = X_{CT}, \quad R_1 = g_A(X_{GC} + X_{GT}) - g_G(X_{AC} + X_{AT}), \quad \) and \( R_2 = g_T(X_{AC} + X_{GC}) - g_C(X_{AT} + X_{GT}). \)

Equation (B10) then gives

\[
X_{xy} = (2 - \delta_{xy}) \sum_k g_k P(k \rightarrow x; t) P(k \rightarrow y; t), \quad (B9)
\]

where \( P(i \rightarrow j; t)'s \) are as given in equations (B8). Note that, because of the assumption of equality of transition matrices for two evolutionary lineages, the expected frequencies of nucleotide pairs \( xy \) and \( yx \) are equal to each other and are treated here as a single class. Substituting equations (B8) into (B9), we obtain

\[
\begin{align*}
\Lambda' &= \begin{bmatrix}
R_1 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0 \\
0
\end{bmatrix}
\begin{bmatrix}
y_1 \\
y_2 \\
y_3 \\
y_4 \\
y_5 \\
y_6 \\
y_7 \\
y_8 \\
y_9 \\
y_{10}
\end{bmatrix} = \begin{bmatrix}
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\\
2g_Rg_Y[1 - e^{2\lambda t}]
\end{bmatrix}
\end{align*}
\]

Combining equations (B10), (B12), and (B13), we obtain

\[
e^{2\lambda t} = 1 - \frac{Q}{2g_Ag_G} = A,
\]

\[
e^{2\lambda t} = 1 - \frac{g_R}{2g_Ag_G} P_1 - \frac{1}{2g_R} Q - \frac{(g_A - g_G)}{2g_Ag_Gg_R} R_1
\]
\[
e^{2\beta_1j} = 1 - \frac{g_Y}{2g_Cg_T} P_2 - \frac{1}{2g_Y} Q - \frac{(g_T - g_C)}{2g_Cg_Tg_Y} R_2
\]
- \frac{1}{4g_Cg_Tg_Rg_Y^2} R_2^2 = B_2.

(B14)

Using equations (B5) and (B14), we also find

\[
2(\alpha_1 g_A - \alpha_4 g_A)t = C_1(\log e A - \log e B_1),
\]
\[
2(\alpha_2 g_C - \alpha_3 g_T)t = C_2(\log e A - \log e B_2),
\]

where

\[
C_1 = R_1/[2g_R(A - \sqrt{AB_1})],
\]
and

\[
C_2 = R_2/[2g_Y(A - \sqrt{AB_2})].
\]

We can now obtain formulas for the rates of nucleotide substitution.

\[
2\alpha_1t = \frac{g_A g_Y + C_1}{g_R} \log e A - \frac{C_1 + g_A}{g_R} \log e B_1,
\]
\[
2\alpha_2t = \frac{g_T g_R + C_2}{g_Y} \log e A - \frac{C_2 + g_T}{g_Y} \log e B_2,
\]
\[
2\alpha_3t = \frac{g_C g_R - C_2}{g_Y} \log e A + \frac{C_2 - g_C}{g_Y} \log e B_2,
\]
\[
2\alpha_4t = \frac{g_G g_Y - C_1}{g_R} \log e A + \frac{C_1 - g_G}{g_R} \log e B_1,
\]
\[
2\beta_1t = (C_1/g_Y - g_A) \log e A - (C_1/g_Y) \log e B_1,
\]
\[
2\beta_2t = (C_2/g_R - g_T) \log e A - (C_2/g_R) \log e B_2,
\]
\[
2\beta_3t = -(C_2/g_R + g_C) \log e A + (C_2/g_R) \log e B_2,
\]
\[
2\beta_4t = -(C_1/g_Y + g_G) \log e A + (C_1/g_Y) \log e B_1.
\]

(B16)

The expected number of nucleotide substitutions per site is then given by

\[
d = -2t \sum_{i=1}^{4} g_i q_{ii},
\]

(B17)

where \(q_{ii}\)'s are the diagonal elements of rate matrix \(Q\), and \(g_i\)'s are the equilibrium nucleotide frequencies. Combining equations (B16) and (B17), we express \(d\) in the following way.

\[
d = \left[2g_A g_G g_Y + g_C g_T g_R g_Y - \frac{g_A g_Y}{g_R} - \frac{g_C g_T g_Y}{g_R} \cdot \frac{g_T - g_C}{g_C g_T g_Y} \frac{C_1 g_Y}{g_R} + \frac{C_2 g_T g_Y}{g_R} \cdot \frac{C_1 g_Y}{g_R} + \frac{C_2 g_T g_Y}{g_R} \cdot \frac{C_1 g_Y}{g_R} \right] \log e A
\]
- \(\left[2g_A g_Y + g_C - g_A \right] \frac{C_1 g_Y}{g_R} \log e B_1
\]
- \(\left[2g_C g_T g_Y + g_C - g_T \right] \frac{C_2 g_T g_Y}{g_R} \log e B_2\),

where

\[
A = [1 - Q/(2g_R g_Y)],
\]
\[
B_1 = [1 - g_R P_1/(2g_A g_G) - Q/(2g_R)
- (g_A - g_G) R_1/(2g_A g_G g_R)
- R_1^2/(4g_A g_G g_R g_Y A)],
\]
\[
B_2 = [1 - g_Y P_2/(2g_C g_T) - Q/(2g_Y)
- (g_C - g_T) R_2/(2g_C g_T g_Y)
- R_2^2/(4g_C g_T g_Y g_T g_Y A)],
\]
\[
C_1 = R_1/[2g_R(A - \sqrt{AB_1})],
\]
\[
C_2 = R_2/[2g_Y(A - \sqrt{AB_2})],
\]
\[
Q = X_{AT} + X_{AC} + X_{GT} + X_{GC},
\]
\[
P_1 = X_{AG},
\]
\[
P_2 = X_{CT},
\]
\[
R_1 = g_A(X_{GC} + X_{GT}) - g_G(X_{AC} + X_{AT}),
\]
and
\[
R_2 = g_T(X_{AC} + X_{GC})
- g_C(X_{AT} + X_{GT}).
\]

The variance of estimate \(\hat{d}\) in equation (B18) can be obtained in the following way. Assume that we have a set of \(m\) sequences with the equilibrium nucleotide frequencies. Then \(g_i\)'s can be obtained from all the sequences in the dataset. Since the variances of \(\hat{g}_i\)'s are much smaller than those of \(\hat{Q}, \hat{P}_i\)'s and \(\hat{R}_i\)'s, we can compute the variance of \(\hat{d}\) by

\[
V(\hat{d}) = a'Z a,
\]

(B20)

where \(a' = (a_1, a_2, a_3, a_4, a_5)\) is a column vector to be defined below, and \(Z\) is the covariance matrix for the vector \([\hat{Q}, \hat{P}_1, \hat{P}_2, \hat{R}_1, \hat{R}_2]\):

\[
V(\hat{Q}) = Q(1 - Q)/n,
\]
\[
V(\hat{P}_1) = P_1(1 - P_1)/n,
\]
\[
V(\hat{P}_2) = P_2(1 - P_2)/n,
\]
\[
V(\hat{R}_1) = [(g_A^2 Q_{GY} + g_G^2 Q_{AY})
- (g_A Q_{GY} + g_G Q_{AY})^2]/n,
\]
\[ V(\hat{R}_2) = \left( g_{QY}^2 Q_{CR} + g_{QY}^2 Q_{TR} \right) \]

\[ = \left( g_{QY}^2 Q_{CR} + g_{QY}^2 Q_{TR} \right)^2 / n, \]

\[ \text{Cov}(\hat{Q}, \hat{P}_1) = -P_1 Q / n, \quad \text{Cov}(\hat{Q}, \hat{P}_2) = -P_2 Q / n, \]

\[ \text{Cov}(P_1, P_2) = -P_1 P_2 / n, \]

\[ \text{Cov}(\hat{Q}, \hat{R}_1) = [R_1 - (g_{QY}^2 Q_{CR} - g_{QY}^2 Q_{TR})] / n, \]

\[ \text{Cov}(\hat{R}_1, \hat{R}_2) = [R_2 - (g_{QY}^2 Q_{CR} - g_{QY}^2 Q_{TR})] / n, \]

\[ \text{Cov}(\hat{R}_1, \hat{P}_1) = -R_1 P_1 / n, \quad \text{Cov}(\hat{R}_1, \hat{P}_2) = -R_1 P_2 / n, \]

\[ \text{Cov}(\hat{R}_2, \hat{P}_1) = -R_2 P_1 / n, \quad \text{Cov}(\hat{R}_2, \hat{P}_2) = -R_2 P_2 / n, \]

\[ \text{Cov}(\hat{R}_1, \hat{R}_2) = \left[ g_{QY}^2 (X_{GC} - Q_{CR} Q_{QY}) \right. \]

\[ - g_{QY}^2 (X_{GC} - Q_{TR} Q_{QY}) \]

\[ - g_{QY}^2 (X_{AC} - Q_{CR} Q_{QY}) \]

\[ + g_{QY}^2 (X_{AT} - Q_{TR} Q_{QY}) \right] / n, \]

where \( Q_{QY} = X_{QY} + X_{GC}, Q_{AV} = X_{AT} + X_{AC}, Q_{CR} = X_{AC} \]

\[ X_{QY} = X_{QY} + X_{GC}, Q_{TR} = X_{AT} + X_{TC}. \]

The elements of vector \( a \) in equation (B20) are as follows.

\[ a_1 = \frac{\partial d}{\partial Q} \]

\[ = \left[ 2 g_{AG}^2 \frac{g_Y}{g_r} + \left( g_C - g_A \right) \frac{C_1}{g_r} + 2 g_{CT} \frac{g_T}{g_Y} \right] D_0 \]

\[ + \left( g_C - g_T \right) \frac{C_2}{g_Y} - 2 g_Y g_T \]

\[ \left[ 2 g_{AG}^2 \frac{g_Y}{g_r} + \left( g_C - g_A \right) \frac{C_1}{g_r} \right] D_1 \]

\[ - \left[ 2 g_{CT} \frac{g_T}{g_Y} + \left( g_C - g_T \right) \frac{C_2}{g_Y} \right] D_2 \]

\[ + \frac{g_C^2 - g_A}{g_Y} D_3 \log_e (A / B_1) \]

\[ + \frac{g_C - g_T}{g_Y} D_4 \log_e (A / B_2), \]

where \( D_0 = \frac{\partial [\log_e A]}{\partial Q} = -(2 g_Y g_A) \),

\[ D_1 = \frac{\partial [\log_e B_1]}{\partial Q} = -\left( \frac{R_1^2}{8 g_{AG} g_{CT} g_Y g_A^2} + \frac{1}{2 g_T} \right) / B_1, \]

\[ D_2 = \frac{\partial [\log_e B_2]}{\partial Q} = -\left( \frac{R_2^2}{8 g_{CT} g_{AG} g_Y g_A^2} + \frac{1}{2 g_Y} \right) / B_2, \]

\[ D_3 = \frac{\partial C_1}{\partial Q} \]

\[ = \frac{R_1}{4 g_Y^2} \left[ 1 - \left( \frac{g_C^2 - g_A}{g_Y} \frac{g_T}{g_Y} + \frac{1}{2 g_T} \right) / \left( \frac{2 g_Y}{2 g_Y} \right) \right] / B_1 \]

\[ - \left( \frac{g_C - g_A}{g_Y} g_T + g_Y \right) / \left( 2 \sqrt{A B_1} \right) \]

\[ (A - \sqrt{A B_1})^2, \]

and

\[ D_4 = \frac{\partial [C_2]}{\partial Q} \]

\[ = \frac{R_2}{4 g_Y^2} \left[ 1 - \left( \frac{g_C^2 - g_A}{g_Y} g_T + g_Y \right) / \left( 2 \sqrt{A B_2} \right) \right] / B_2 \]

\[ (A - \sqrt{A B_2})^2. \]

\[ a_2 = \frac{\partial d}{\partial P_1} = \frac{g_C - g_A}{g_Y} E \frac{log_e (A / B_1)}{E_1} \]

\[ - \left[ \frac{2 g_{AG} g_T}{g_Y} + (g_C^2 - g_A) \frac{C_1}{g_Y} \right] E_2, \]

where

\[ E_1 = \frac{\partial [\log_e B_1]}{\partial P_1} = -\frac{g_Y}{2 g_{AG} B_1}, \]

and

\[ E_2 = \frac{\partial C_1}{\partial P_1} \]

\[ = \frac{g_C^2 - g_A}{g_Y} \left( 1 - \left( \frac{g_C^2 - g_A}{g_Y} \frac{g_T}{g_Y} + \frac{1}{2 g_T} \right) / \left( \frac{2 g_Y}{2 g_Y} \right) \right) / \left( 2 \sqrt{A B_1} \right) \]

\[ - \left( \frac{R_1^2}{4 g_{AG} g_{CT} g_Y g_A^2} \right) / \left( \frac{2 \sqrt{A B_1} (A - \sqrt{A B_1})}{} \right) \]

\[ a_3 = \frac{\partial d}{\partial P_2} = \frac{g_C - g_T}{g_Y} F \frac{log_e (A / B_2)}{F_1} \]

\[ - \left[ \frac{2 g_{CT} g_T}{g_Y} + (g_C^2 - g_T) \frac{C_2}{g_Y} \right] F_2, \]

where

\[ F_1 = \frac{\partial [\log_e B_2]}{\partial P_2} = \frac{g_Y}{2 g_C B_2}, \]
and

\[ F_2 = \frac{\partial C_2}{\partial P_2} \]

\[ = -\frac{g_Y A}{2g_C g_T} \left( 1 - \frac{Q}{2g_Y} - \frac{(g_T - g_C)R_2}{2g_C g_T g_Y} \right) \left[ 2\sqrt{A B_2(A - \sqrt{A B_2})} \right] \]

\[ - \frac{R_2^2}{4g_C g_T g_Y^2} \right) \left[ 2\sqrt{A B_2(A - \sqrt{A B_2})^2} \right]. \]

\[ a_4 = \frac{\partial d}{\partial R_1} = \frac{g_R - g_A}{g_Y} G_1 \log_e(A/B_1) \]

\[ - \left[ \frac{2g_R g_G}{g_Y} + \frac{g_R - g_A}{g_Y} \right] G_2, \]

where

\[ G_1 = \frac{\partial \log_e B_1}{\partial R_1} \]

\[ = -\frac{1}{2g_A g_B g_R B_1} \left[ (g_A - g_G) + \frac{R_1}{g_R g_Y A} \right], \]

and

\[ G_2 = \frac{\partial C_1}{\partial R_1} = \left[ 1 - \frac{\sqrt{A B_1} G_1 R_1}{2(A - \sqrt{A B_1})} \right] \left[ 2g_R(A - \sqrt{A B_1}) \right]^{-1}. \]

Finally,

\[ a_5 = \frac{\partial d}{\partial R_2} = \frac{(g_C - g_1)}{g_Y} H_1 \log_e(A/B_2) \]

\[ - \left[ \frac{2g_C g_T}{g_Y} + \frac{g_C - g_T}{g_Y} \right] H_2, \]

where

\[ H_1 = \frac{\partial \log_e B_2}{\partial R_2} = -\frac{1}{2g_C g_T g_Y B_2} \]

\[ \times \left[ (g_T - g_C) + \frac{R_2}{g_R g_Y A} \right], \]

and

\[ H_2 = \frac{\partial C_2}{\partial R_2} = \left[ 1 - \frac{\sqrt{A B_2} G_2 R_2}{2(A - \sqrt{A B_2})} \right] \left[ 2g_Y(A - \sqrt{A B_2}) \right]^{-1}. \]

APPENDIX C

Multiple estimators for Hasegawa et al.'s (1985) Model

Hasegawa et al.'s (1985) model has five parameters, that is, \( \alpha \) and \( \beta \) and three independent parameters for the expected equilibrium nucleotide frequencies (see [5] in table 1). Therefore, to estimate the evolutionary distance, \( d \), we need at least five independent observations.

Hasegawa et al. (1985) attempted to express the expected number of nucleotide substitutions per site, \( d \), in terms of \( g_i \)'s, \( Q \), and \( P \). In this way they obtained a set of equations that had no explicit solution. However, one can derive multiple explicit expressions for \( d \) by considering the expected frequencies of transitional differences \( AG \) and \( CT \) separately. One of them is Tamura and Nei's (1993) estimator, and it is possible to find many other estimators. Indeed, the expected value of the number of nucleotide substitutions per site can be expressed in terms of \( \alpha \), \( \beta \), and the equilibrium nucleotide frequencies as follows.

\[ d = 4[(g_A g_B + g_C g_T) \alpha + g_R g_Y \beta] t. \]  

where \( g_R = g_A + g_G, g_Y = g_C + g_T \), and \( t \) is the evolutionary time since divergence of two sequences. Therefore, to estimate \( d \) we need to express \( g_1 \) and \( g_2 \) in terms of the expected frequencies of nucleotide pairs, \( X_i \)'s.

From equation (B10) in Appendix B, we can express \( 2g_1 t \) in two different ways:

\[ A_1 = 2g_1 t = \frac{g_Y}{g_R} \log_e \left( 1 - \frac{Q}{2g_R g_Y} \right) \]

\[ - \frac{1}{g_R} \log_e \left( 1 - \frac{Q}{2g_R} - \frac{g_R P_1}{2g_A g_B} \right), \]

and

\[ A_2 = 2g_1 t = \frac{g_R}{g_Y} \log_e \left( 1 - \frac{Q}{2g_R g_Y} \right) \]

\[ - \frac{1}{g_Y} \log_e \left( 1 - \frac{Q}{2g_Y} - \frac{g_Y P_2}{2g_C g_T} \right), \]

where all notations are the same as those in Appendix C. (Formulas in eq. [C2] have been used to distinguish between Hasegawa et al.'s and Tamura and Nei's models in eq. [37] in text.) Computing \( 2g_2 t \) by

\[ A_3 = 2g_2 t = -\log_e \left( 1 - \frac{Q}{2g_R g_Y} \right), \]

we can write a general expression for \( d \):

\[ d = 2(g_A g_B + g_C g_T)[\gamma A_1 + (1 - \gamma) A_2] + 2g_R g_Y A_3, \]
where $\gamma$ is an arbitrary coefficient between 0 and 1. The simplest estimator of $d$ is obtained when $\gamma$ is equal to 1 or 0. For instance, for $\gamma = 1$ we obtain

$$d = 2 \left( \frac{g_R}{g_Y} g_{AGG} + \frac{g_Y}{g_R} g_{CTG} - g_{CTG} - g_{AGG} \right) \log_e \left( 1 - \frac{Q}{2g_R g_Y} \right)$$

$$- 2 \left( \frac{g_{ACG}}{g_R} + \frac{g_{GTC}}{g_Y} \right) \log_e \left( 1 - \frac{Q}{2g_R g_Y} \right).$$

(C5)

The variance of estimate $\hat{d}$ can be computed as

$$V(\hat{d}) = \frac{\left( (a^2Q + b^2P) - (aQ + bP)^2 \right)}{n},$$

where

$$a = \frac{g_{AGG} + g_{CTG}}{2g_R} \left( 1 + \frac{g_{CTG}}{g_{AGG}} \right)$$

$$b = \left( 1 + \frac{g_{CTG}}{g_{AGG}} \right),$$

$$c = \left( 1 - \frac{Q}{2g_R g_Y} \right),$$

and

$$e = \left( 1 - \frac{Q}{2g_R g_Y} \right).$$

(C6)

Clearly, we can find the estimator of form (C4) that provides the minimum variance of the estimate. It is not difficult to find $\gamma$ that minimizes the variance of estimator (C4):

$$\gamma = \frac{V(\hat{A}_2) - \text{Cov}(A_1, A_2)}{V(\hat{A}_1) + V(\hat{A}_2) - 2 \text{Cov}(A_1, A_2)}$$

$$+ \frac{g_R g_Y}{g_{AGG} + g_{CTG}} \frac{\text{Cov}(A_1, A_2) - \text{Cov}(A_1, A_1) + \text{Cov}(A_2, A_2) - 2 \text{Cov}(A_1, A_2)}{V(\hat{A}_1) + V(\hat{A}_2) - 2 \text{Cov}(A_1, A_2)}.$$

(C8)

The variances and covariances required in the latter equation can be computed by

$$V(\hat{A}_1) = \left( \delta^2 Q + \epsilon^2 P_1 \right) - (\delta Q + \epsilon P_1)^2 \right)/n,$$

$$V(\hat{A}_2) = \left( \xi^2 Q + \eta^2 P_2 \right) - (\xi Q + \eta P_2)^2 \right)/n,$$

$$\text{Cov}(\hat{A}_1, \hat{A}_2) = \left( \delta^2 Q (1 - Q) - \delta \eta Q P_2 - \epsilon P_1 P_2 \right)/n,$$

$$\text{Cov}(A_1, A_1) = \nu Q [\delta (1 - Q) - \epsilon P_1]/n,$$

$$\text{Cov}(A_2, A_2) = \nu Q (\xi (1 - Q) - \eta P_2)/n,$$

(C9)

where

$$\delta = (2g_R g_Y e)^{-1} - (2g_R^2 c)^{-1},$$

$$\xi = (2g_Y^2 f)^{-1} - (2g_Y^2 c)^{-1},$$

$$\eta = (2g_Y g_C f)^{-1},$$

$$\nu = (2g_R g_Y c)^{-1},$$

$$f = \left( 1 - \frac{Q}{2g_R g_Y} \right).$$

(C10)

and values of $c$ and $e$ are as given in (C7).

LITERATURE CITED


STANLEY SAWYER, reviewing editor
Received February 5, 1994
Accepted August 25, 1994