Bias-Corrected Paralinear and LogDet Distances and Tests of Molecular Clocks and Phylogenies Under Nonstationary Nucleotide Frequencies

Xun Gu and Wen-Hsiung Li
Human Genetics Center, University of Texas at Houston

The statistical properties of the paralinear and LogDet distances under nonstationary nucleotide frequencies were studied. First, we developed formulas for correcting the estimation biases of the paralinear and LogDet distances, i.e., the bias-corrected distance is estimated by \( d_e = d - 2\text{var}(d) \), where \( d \) and \( \text{var}(d) \) are the estimated distance and sampling variance, respectively. The performances of these formulas and the formulas for sampling variances were examined by computer simulation. Second, we developed a method for estimating the variance–covariance matrix of paralinear distances, so that statistical tests of DNA phylogenies can be conducted in the nonstationary case. Third, a new LogDet-based method for testing the molecular clock hypothesis was developed under nonstationary nucleotide frequencies.

Introduction

The stationarity of nucleotide frequencies is one of the most common assumptions made in estimating evolutionary distances (see Lanave et al. 1984; Zharkikh 1994; Gu and Li 1996). It assumes that the expectations of nucleotide frequencies in a sequence do not change with time and are equal to those in the ancestral sequence. Therefore, to estimate the distance between two sequences, the nucleotide frequencies in the ancestral sequence are estimated by the averages of the nucleotide frequencies in the two extant sequences. If nucleotide frequencies vary with time so that stationarity does not hold, the estimated distance may not be accurate. As a consequence, a distance-matrix method for phylogeny reconstruction can be misleading, i.e., it tends to group sequences of similar nucleotide frequencies irrespective of the true evolutionary relationships (Hasegawa and Itzhak 1993; Sogin, Hinkle, and Leipe 1993; Steel 1994). In addition, nonstationary nucleotide frequencies can affect the statistical testing of a molecular clock.

The paralinear (Lake 1994) and LogDet (Steel 1994; Lockhart et al. 1994) distances have been proposed to deal with the nonstationarity problem. They are based on the most general model of nucleotide substitutions, and the additivity of the paralinear distance holds under nonstationarity. Historically, these methods can be traced back to Barry and Hartigan (1987) and Cavender and Felsenstein (1987).

The purpose of this paper is to study the statistical properties of the paralinear and LogDet distances, especially with respect to the following problems. First, our preliminary computer simulation indicated that these methods give biased estimates when sequences are not long. We shall develop methods for correcting the bias. Second, we will derive the variance–covariance matrices of paralinear distances, so that some statistical tests of DNA phylogenies can be applied. Third, we will develop a new method for testing molecular clocks under nonstationary nucleotide frequencies.

Methods

The Paralinear and LogDet Distances

Consider two sequences (denoted by 1 and 2, respectively) that evolved from 0, their common ancestor, \( t \) time units ago (see fig. 1). We denote the diagonal matrix of nucleotide frequencies in sequence \( k \) \( (k = 0, 1, 2) \) by \( \mathbf{F}^{(k)} = \text{diag}(f_1^{(k)}, f_2^{(k)}, f_3^{(k)}, f_4^{(k)}) \), where the subscript \( j \) refers to nucleotide \( j \) \( (j = 1, \ldots , 4) \) for \( A, G, T \), and \( C \), respectively). Let \( \mathbf{J} \) be the data matrix whose \( ij \)-th element \( J_{ij} \) is the proportion of sites at which the nucleotide is \( i \) in sequence 1 and \( j \) in sequence 2. Then, the paralinear distance (between sequences 1 and 2) is defined as

\[
d = -\frac{1}{4} \ln \frac{\text{det}[\mathbf{J}]}{\sqrt{\det[\mathbf{F}^{(1)}] \det[\mathbf{F}^{(2)}]}},
\]

where \( \ln \) denotes the natural logarithm, \( \text{det} \) means the determinant of a matrix, and \( \det[\mathbf{F}^{(k)}] = \prod_{j=1}^{4} f_j^{(k)}, \ k = 1, 2 \) (Lake 1994). A related measure is the LogDet distance (Steel 1994; Lockhart et al. 1994), which is defined as

\[
d = -\frac{1}{4} \ln \det[\mathbf{J}] - \ln 4.
\]

In equation (2), we added the constant \(-\ln 4\), which does not change any property of the original LogDet distance but makes the biological interpretation of equation (2) easier (see further discussion below).

By using the delta method (Barry and Hartigan 1987), the sampling variance of the paralinear distance (eq. 1) is found to be

\[
\text{var}(d) \approx \frac{1}{16L} \sum_{i=1}^{4} \left( \frac{4}{\sum_{j=1}^{4} M_{ij}^2 J_{ij}} - 1/\sqrt{f_i f_j} \right),
\]

where \( L \) is the sequence length and \( M_{ij} \) is the \( ij \)-th element of \( \mathbf{M} = \mathbf{J}^{-1} \). Note that, when the nucleotide frequencies are stationary, i.e., \( f_i^{(0)} = f_i^{(1)} = f_i^{(2)} \) for \( i = 1, \ldots , 4 \), equation (3) reduces to the formula given by Barry and Hartigan (1987). The sampling variance of the LogDet distance (eq. 2) is given by
FIG. 1.—Two DNA sequences diverged \( t \) time units ago.

\[
\text{var}(d) = \frac{1}{16L} \sum_{i=1}^{4} \sum_{j=1}^{4} (M_{ij}^2 J_{ij} - 1)
\]

(Lockhart et al. 1994).

Assume that nucleotide substitution follows a Markov process, which can be characterized by a rate matrix \( R \) (Barry and Hartigan 1987; Zharkikh 1994). The \( ij \)-th element of \( R \), \( r_{ij} \), is the substitution rate per unit time from nucleotide \( i \) to \( j \), \( i \neq j \) (\( i, j = 1, 2, 3, 4 \)), and the diagonal elements are given by \( r_{ii} = -\sum_{j \neq i} r_{ij} \). Then, the matrix of transition probabilities for \( t \) time units is given by \( P(t) = e^{Rt} \).

The paralinear and LogDet distances are potentially very useful in the study of DNA evolution because they have the following properties:

1. Both distances are based on the most general model of nucleotide substitution, i.e., the 12-parameter model (Lake 1994; Steel 1994; Lockhart et al. 1994; Zharkikh 1994). Moreover, they are valid even if the rate matrix \( R \) varies along and among lineages.

2. Under the assumption of the same substitution rate among sites, the paralinear and LogDet distances are very useful for phylogenetic reconstruction when nucleotide frequencies are nonstationary. As will be shown in the appendix, for some distance matrix methods of phylogenetic reconstruction, e.g., the neighbor-joining method (Saitou and Nei 1987), the LogDet distance (eq. 2) and the paralinear distance (eq. 1) lead to the same tree topology.

3. The branch lengths can be estimated in terms of paralinear distances, so that many statistical tests of DNA phylogenies (see Li and Zharkikh 1995 for a review) are applicable. On the other hand, the LogDet distance is useful for testing the molecular clock hypothesis under nonstationary frequencies (see further discussion below).

4. The biological interpretation of the two distances can be described as follows (Eqs. 7 and 8).

For two DNA sequences (1 and 2) that diverged from the ancestral sequence \( t \) time units ago (fig. 1), let \( R^{(1)} = \{ r_{ij}^{(1)} \} \) and \( R^{(2)} = \{ r_{ij}^{(2)} \} \) be the rate matrix in lineages 1 and 2, respectively. Note that \( R^{(1)} \) and \( R^{(2)} \) can be different. It can be shown (e.g., Zharkikh 1994) that

\[
J = P^{(1)}J^{(0)}P^{(2)},
\]

where \( P^{(1)} = e^{R^{(1)}t} \) and \( P^{(2)} = e^{R^{(2)}t} \) are the transition probability matrices from node \( O \) to node 1 and node 2, respectively; \( ' \) means transpose; and \( J^{(0)} \) is the diagonal matrix of nucleotide frequencies at the ancestral node \( O \).

Let \( \mu^{(k)} = -\sum_{i=1}^{4} r_{ii}^{(k)}/4 \) be the arithmetic mean rate in lineage \( k \) \( (k = 1, 2) \). Barry and Hartigan (1987) and Zharkikh (1994) showed that

\[
-\frac{1}{4} \ln \det[P^{(k)}] = \mu^{(k)} t, \quad k = 1, 2.
\]

From equation (5), we have \( \ln \det[J] = \ln \det[P^{(1)}] + \ln \det[P^{(2)}] \). Therefore, by equation (6), and noting that \( \det[J^{(0)}] = \prod_{i=1}^{4} f_{i}^{(0)} \), we can show that the paralinear distance (eq. 1) is equal to

\[
d = 2\mu t - \frac{1}{4} \sum_{i=1}^{4} (\ln f_{i}^{(1)} + \ln f_{i}^{(2)} - 2 \ln f_{i}^{(0)})
\]

and the LogDet distance (eq. 2) is equal to

\[
d = 2\mu t - \frac{1}{4} \sum_{i=1}^{4} \ln f_{i}^{(0)} - 4,
\]

where \( \mu = (\mu^{(1)} + \mu^{(2)})/2 \) is the mean rate over the two lineages and \( f_{i}^{(k)} \) is the frequency of nucleotide \( i \) at node \( k \), \( (i = 1, \ldots, 4 \text{ and } k = 0, 1, 2) \).

Since both distance measures depend on nucleotide frequencies, they are not linear in time \( t \) if the nucleotide frequencies change with time. However, if the nucleotide frequencies are stationary and equal to \( 1/4 \), i.e., \( f_{i}^{(0)} = f_{i}^{(1)} = f_{i}^{(2)} = 1/4 \), the paralinear and LogDet distances are equal to \( K = 2\mu t \), the number of substitutions per site (Lake 1994; Lockhart et al. 1994; Zharkikh 1994).

Bias-Corrected Paralinear and LogDet Distances

The estimation procedure for the paralinear and LogDet distances is straightforward: The matrix \( J \) and the nucleotide frequencies can be directly estimated from the sequence data and then the distances can be estimated from equations (1) and (2) (Lake 1994; Lockhart et al. 1994). However, our preliminary simulation showed that both distances are on average overestimated when the sequences are short (see further discussion below). As the sequence length becomes longer than 2,000 bp, the estimation bias becomes trivial. Since for many cases in practice the sequence length is shorter than 2,000 bp, it is useful to develop a method for correcting the estimation bias. Related work, although by a different approach, on the bias of the LogDet distance has been done by Bar-Hen and Penny (1996).

Assume that random variables \( J_{ij} \) follow a multinomial distribution with parameters \( E[J_{ij}] = \bar{J}_{ij} \) (\( i, j = 1, \ldots, 4 \)). Let \( \bar{J}_{ij} \) be the estimate of \( J_{ij} \) from sequence data. The matrix forms of \( J_{ij} 's, \bar{J}_{ij} 's, \) and \( J_{ij} 's \) are \( J, \bar{J}, \) and \( \bar{J} \), respectively. Let \( V_{ij} \) be the sampling variance of \( J_{ij} \) and \( \text{cov}_{i,j,k,l} \) be the covariance between \( J_{ij} \) and \( J_{kl} \). Under the assumption of a multinomial distribution, we have

\[
V_{ij} = \frac{\bar{J}_{ij}(1 - \bar{J}_{ij})}{L},
\]

\[
\text{cov}_{i,j,k,l} = -\frac{\bar{J}_{ij}\bar{J}_{kl}}{L},
\]

where \( L \) is the sequence length.
Consider the LogDet distance first. To simplify our notation, define the function $g$ as $g(J) = -\log \det(J)$. Then, the LogDet distance (eq. 2) can be written as $d = g(J)/4 - \ln 4$. Expanding $d$ around $J = \tilde{J}$ for the first three terms, we have

$$
d = \tilde{d} + \frac{1}{4} \sum_{\ell} \left[ \frac{\partial^2 g}{\partial J_{\ell}^2} \right] (J_{\ell} - \tilde{J}_{\ell}) + \frac{1}{8} \sum_{\ell, kl} \left[ \frac{\partial^2 g}{\partial J_{\ell} \partial J_{kl}} \right] (J_{\ell} - \tilde{J}_{\ell})(J_{kl} - \tilde{J}_{kl}). \tag{10}
$$

where $\tilde{d}$ is given by

$$
\tilde{d} = -\frac{1}{4} \log \det(\tilde{J}) - \ln 4, \tag{11}
$$

which is the true value of the LogDet distance. By taking expectations of both sides of equation (10), one can show that

$$
E[d] = \tilde{d} + \frac{1}{8} \sum_{\ell} \left[ \frac{\partial^2 g}{\partial J_{\ell}^2} \right] V_{\ell} + \frac{1}{8} \sum_{\ell, kl} \left[ \frac{\partial^2 g}{\partial J_{\ell} \partial J_{kl}} \right] \text{cov}_{\ell, kl}, \tag{12}
$$

because $E[J_{\ell}] = J_{\ell}$. Let $a_{ij}$ and $b_{ijkl}$ be

$$
a_{ij} = -\left[ \frac{1}{\det(J)} \frac{\partial \det(J)}{\partial J_{ij}} \right] J_{ij},
$$

$$
b_{ijkl} = -\left[ \frac{1}{\det(J)} \frac{\partial^2 \det(J)}{\partial J_{ij} \partial J_{kl}} \right] J_{ij} J_{kl}. \tag{13}
$$

It is easy to show that

$$
\left[ \frac{\partial^2 g}{\partial J_{ij} \partial J_{kl}} \right] = a_{ij} a_{kl} - b_{ijkl}. \tag{14}
$$

Therefore, noting that $b_{ijkl} = 0$, we have

$$
E[d] = \tilde{d} + \frac{1}{8} \sum_{ij} a_{ij}^2 V_{ij} + \frac{1}{8} \sum_{ijkl} (a_{ij} a_{kl} - b_{ijkl}) \text{cov}_{ij, kl}. \tag{15}
$$

It can be shown that the sampling variance of $d$ derived by the delta method (i.e., eq. 4) is approximately given by

$$
\text{var}(d) = \frac{1}{16} \sum_{ij} a_{ij}^2 V_{ij} + \frac{1}{16} \sum_{ijkl} a_{ij} a_{kl} \text{cov}_{ij, kl}. \tag{16}
$$

Thus, equation (13) can be simplified as

$$
E[d] = \tilde{d} + 2 \text{var}(d) - \frac{1}{8} \sum_{ijkl} b_{ijkl} \text{cov}_{ij, kl}. \tag{17}
$$

It is not easy to compute the constants $b_{ijkl}$. However, since the third term in the right-hand side of equation (17) is generally small, equation (17) can be approximated by

$$
E[d] = \tilde{d} + 2 \text{var}(d). \tag{18}
$$

Let $\hat{d}$ and $\text{var}(\hat{d})$ be the estimates of the LogDet distance (eq. 2) and the sampling variance (eq. 4), respectively. Then, from equation (18), the bias-corrected LogDet distance $\hat{d}_c$ can be estimated by

$$
\hat{d}_c = \hat{d} - 2 \text{var}(\hat{d}). \tag{19}
$$

For the paralinear distance, equation (1) can be written as $d = g(J)/4 - g(F(1)F(2))/8$; the second term is treated as a constant. Thus, similar to the derivation of equations (10) to (18), we can show that the bias-corrected paralinear distance is also estimated by equation (19), where $\hat{d}$ and $\text{var}(\hat{d})$ are obtained from equation (1) and equation (3), respectively.

Note that equation (19) is only approximate for correcting the estimation bias. Therefore, its performance will be examined by computer simulation (see further discussion below).

The Variance–Covariance Matrix of Paralinear Distances

Statistical testing of a phylogenetic hypothesis (Nei, Stephens, and Saitou 1985; Bulmer 1991; Rzhetsky and Nei 1992; Gu and Li 1996) or a molecular clock (Wu and Li 1985) based on distance-matrix methods requires the variance–covariance matrix of distances. When the paralinear distance is used, the variance of a distance can be computed by equation (3) and the covariance between two distances can be computed as follows.

Similar to above, let $g^{(kl)} = -\ln \det(J^{(kl)})$, where $J^{(kl)}$ is the data matrix of distance $kl$ (e.g., $k = 1, l = 2$ means the distance between sequences 1 and 2). Let $J_{ij}^{(kl)}$ be the $ij$-th element of $J^{(kl)}$. According to the matrix theory, we have

$$
\frac{\partial g^{(kl)}}{\partial J_{ij}^{(kl)}} = -M_{ij}^{(kl)}, \tag{20}
$$

where $M_{ij}^{(kl)}$ is the $ij$-th element of matrix $M^{(kl)}$, which is the inverse of matrix $J^{(kl)}$.

To compute the covariance between two distances, we need to distinguish between two situations: (1) three sequences (denoted by 1, 2, and 3) are involved, e.g., the covariance between $d_{13}$ and $d_{23}$; and (2) four sequences (denoted by 1, 2, 3, and 4) are involved, e.g., the covariance between $d_{12}$ and $d_{34}$. By the delta method, the covariance for the first situation is given by

$$
\text{cov}(d_{13}, d_{23}) = \frac{1}{16} \text{cov}(g^{(13)}, g^{(23)}),
$$

$$
= \frac{1}{16} \sum_{ijkl} \text{cov}(J_{ij}^{(13)}, J_{lk}^{(23)}) \frac{\partial g^{(13)}}{\partial J_{ij}^{(13)}} \frac{\partial g^{(23)}}{\partial J_{lk}^{(23)}},
$$

$$
= \frac{1}{16} \sum_{ijkl} \text{cov}(J_{ij}^{(13)}, J_{lk}^{(23)}) M_{ij}^{(13)} M_{lk}^{(23)}.
$$

Let $f_{ijk}$ be the frequency of sites at which the nucleotides in sequences 1, 2, and 3 are $i$, $j$, and $k$, respectively. Under the assumption of a multinomial distribution, it can be shown that
FIG. 2.—The phylogenetic tree used for molecular clock testing.

\[
\text{cov}(J_{ij}^{(13)}, J_{ij}^{(23)}) = \frac{1}{L} (f_{ijk} - J_{ij}^{(13)} J_{ij}^{(23)}).
\]

Therefore, the covariance between \( d_{12} \) and \( d_{23} \) for the first situation can be estimated by

\[
\text{cov}(d_{13}, d_{23}) = \frac{1}{16L} \sum_{i,j,k,l=1}^{4} (f_{ijk} - J_{ij}^{(13)} J_{ij}^{(23)}) M_{ij}^{(12)} M_{ij}^{(24)}.
\]

For the second situation, let \( f_{ijkl} \) be the frequency of sites at which the nucleotides in sequences 1, 2, 3, and 4 are \( i, j, k, \) and \( l, \) respectively. The covariance between \( d_{12} \) and \( d_{34} \) is given by

\[
\text{cov}(d_{12}, d_{34}) = \frac{1}{16L} \sum_{i,j,k,l=1}^{4} \text{cov}(J_{ij}^{(12)}, J_{ij}^{(24)}) M_{ij}^{(12)} M_{ij}^{(24)}
\]

\[
= \frac{1}{16L} \sum_{i,j,k,l=1}^{4} (f_{ijkl} - J_{ij}^{(12)} J_{ij}^{(24)}) M_{ij}^{(12)} M_{ij}^{(24)}.
\]

Statistical Tests of a Molecular Clock Under Nonstationarity

The relative rate test of rate constancy (i.e., a molecular clock) (Wu and Li 1985) can be described as follows. Consider three species as shown in figure 2, where species 3 is an outgroup. To test whether the evolutionary rate in lineage \( O1 \) is the same as that in lineage \( O2 \) (i.e., the molecular clock hypothesis), one needs to test whether the difference

\[
D = d_{13} - d_{23}
\]

is significantly different from zero. When the nucleotide frequencies are nonstationary, \( d_{ij} \) may be measured by the paralinear or LogDet distance. As can be seen from equations (7) and (8), both distances have two components: \( 2k \mu t \) and a term involving nucleotide frequencies. A test of equal rates under the nonstationary situation can be regarded as a test of the difference in \( \mu t \) between the two lineages.

We show that the LogDet distance is suitable for this purpose. Let \( \mu^{(i)} \) (\( i = 1, 2, 3 \)) be the (arithmetic) mean of substitution rates in lineage \( i, \) let \( t \) be the divergent time between species 1 and 2, and let \( T \) be the divergent time between species 1 (or 2) and 3 (see fig. 2). From equation (8), we can write \( d_{13} \) as \( d_{13} = \mu^{(1)} t + \mu^{(3)} (2T - t) - \Sigma_{j=1}^{4} (\ln f_{ij}^{(1)} - \ln (4j/4)) \) and \( d_{23} \) as \( d_{23} = \mu^{(2)} t + \mu^{(3)} (2T - t) - \Sigma_{j=1}^{4} (\ln f_{ij}^{(1)} - \ln (4j/4)), \)

where \( f_{ij}^{(1)} \) is the frequency of nucleotide \( j \) at root \( R. \) Therefore, we have

\[
D = d_{13} - d_{23} = (\mu^{(1)} - \mu^{(2)}) t.
\]

That is, the difference between \( d_{13} \) and \( d_{23} \) is due to the rate difference between lineages \( O1 \) and \( O2. \) Thus, based on equation (25), we can construct a new relative rate test of the molecular clock hypothesis. However, if \( d_{ij} \) is measured by the paralinear distance, then \( D' = d_{13} - d_{23} \) is given by

\[
D' = (\mu^{(1)} - \mu^{(2)}) t + \sum_{i=1}^{4} (\ln f_{ij}^{(1)} - \ln f_{ij}^{(2)}),
\]

which is not suitable because it does not distinguish between the rate difference and the nucleotide frequency difference.

To test whether \( D \) is significantly different from zero, one needs to estimate the sampling variance of \( D, \) which is given by

\[
V(D) = V(d_{13}) + V(d_{23}) - 2 \text{cov}(d_{13}, d_{23}).
\]

where \( V(d_{13}) \) and \( V(d_{23}) \) can be estimated by equation (4), and \( \text{cov}(d_{13}, d_{23}) \) can be estimated by equation (23). By putting them together, we have

\[
V(D) = \frac{1}{16L} \sum_{i,j,k,l=1}^{4} [(M_{ij}^{(13)})^2 J_{ij}^{(13)} - 1]
+ \frac{1}{16L} \sum_{i,j,k,l=1}^{4} [(M_{ij}^{(23)})^2 J_{ij}^{(23)} - 1]
- \frac{1}{8L} \sum_{i,j,k,l=1}^{4} (f_{ijkl} - J_{ij}^{(13)} J_{ij}^{(24)}) M_{ij}^{(12)} M_{ij}^{(24)}.
\]

The statistic \( Z = D/\sqrt{V(D)} \) can be computed. When the sequence is long, \( Z \) follows approximately the standard normal distribution, and the standard normal test can be used to determine the significance level (Wu and Li 1985).

Simulation

We use computer simulation to study the performances of the bias-corrected paralinear and LogDet distances (eq. 19). Denote nucleotides A, G, T, and C by 1, 2, 3, and 4, respectively. In the following, we use \( (f_1, f_2, f_3, f_4) \) to denote the nucleotide frequencies at a particular node. In the case of two-lineage evolution (fig. 1), we set the divergence parameter \( 2\mu t \) (see eqs. 7 and 8 for definition) equal to 0.2, 0.5, or 0.8, and the sequence length equal to \( L = 200, 500, \) or 2,000 nucleotides. Two models of nucleotide substitution are used for simulation; the first model (TR) is time reversible and the second (NR) is time irreversible. The rate matrices \( R \) of TR and NR are as follows (the scale is arbitrary):

(1) The TR model

\[
\begin{pmatrix}
-1.5 & 1.5 & 0.2 & 0.8 \\
0.5 & -3.1 & 0.6 & 2.0 \\
0.1 & 0.9 & -2.6 & 1.6 \\
0.2 & 1.5 & 0.8 & -2.5 \\
\end{pmatrix}
\]

with the equilibrium frequencies \((0.1, 0.4, 0.2, 0.3).\)
Case 1 assumes that the model of nucleotide substitution evolution can be simulated. We studied five cases (table 1).

In case 2, the substitution models in both lineages are fixed. In each lineage with the equilibrium frequencies (0.47, 0.11, 0.36, 0.06), the LogDet distances (d) are computed from equations (7) and (8). The percentage values presented in the parentheses (0.47, 0.11, 0.36, 0.06) are the true values of the paralinear distance shown in table 2. The estimation bias (d - d) increases with increasing divergence 2μ. The bias can be substantially reduced by the bias-corrected paralinear distance (d₆). For example, in the case of 2μ = 0.5, the bias of the paralinear distance is on average reduced from 0.016 to 0.004 when L = 200, and reduced from 0.006 to 0.002 when L = 500. Of course, when L = 2000, the estimation bias becomes trivial so that d₆ = d = d for all cases.

Table 3 shows the performance of the bias-corrected LogDet distance; the notations are the same as in table 2 except that d is estimated by equation (2), var(d) is estimated by equation (4), and d is computed by equation (8). The performance of the bias-corrected LogDet distance is similar to that of the bias-corrected paralinear distance shown in table 2.

The percentage values presented in the parentheses in tables 2 and 3 were the average biases of the estimates. In the case of the paralinear distance with sequence length 200 bp, the average biases of the uncorrected estimates (d) were about 2%, 3.5%, and 5% for 2μ = 0.2, 0.5, and 0.8, respectively, but they were reduced, on average, to 0.4%, 0.6%, and 2%, respectively, by the bias-corrected method (table 2). For the LogDet distance with L = 200 bp, the average biases of the uncorrected estimates were about 5%, 4%, and 5% for 2μ = 0.2, 0.5, and 0.8, respectively, but were reduced to, on average, 1%, 0.5%, and 1%, respectively, by the bias-corrected method. It is clear that, when the sequence length is short, the estimation bias can be effectively corrected by equation (19).

The mean of standard errors computed by equation (3) or equation (4) is presented in parentheses in tables 2 and 3. The standard error estimated by equation (3) or equation (4) is quite close to the observed one in most cases. However, when the sequences are short (e.g., L = 200) and the distance is large (2μ ≥ 0.8), the sampling variance could be overestimated. We have examined many cases with L ≥ 200 and 2μ ≤ 0.8 and found that the estimated sampling variances for the paralinear and LogDet distances are accurate.
Table 2
Statistical Performances of the Bias-Corrected Paralinear Distance

<table>
<thead>
<tr>
<th>Case</th>
<th>L</th>
<th>d</th>
<th>( \mu_d \pm \text{SE} )</th>
<th>( \mu_d \pm \text{SE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>0.200</td>
<td>0.199 (0.5%) \pm 0.036 (0.040)</td>
<td>0.203 (1.5%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.200</td>
<td>0.201 (0.5%) \pm 0.024 (0.025)</td>
<td>0.202 (1.0%)</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>0.179</td>
<td>0.180 (0.6%) \pm 0.036 (0.037)</td>
<td>0.183 (2.2%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.179</td>
<td>0.178 (0.6%) \pm 0.021 (0.023)</td>
<td>0.178 (0.6%)</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>0.157</td>
<td>0.157 (0.0%) \pm 0.034 (0.035)</td>
<td>0.157 (1.5%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.157</td>
<td>0.157 (0.0%) \pm 0.021 (0.022)</td>
<td>0.157 (0.5%)</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>0.235</td>
<td>0.236 (0.4%) \pm 0.043 (0.048)</td>
<td>0.240 (2.1%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.235</td>
<td>0.234 (0.4%) \pm 0.027 (0.029)</td>
<td>0.236 (0.4%)</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>0.263</td>
<td>0.262 (0.4%) \pm 0.058 (0.062)</td>
<td>0.273 (3.8%)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.263</td>
<td>0.262 (0.4%) \pm 0.020 (0.026)</td>
<td>0.263 (0.9%)</td>
</tr>
</tbody>
</table>

2\( \mu_d = 0.5 \)

<table>
<thead>
<tr>
<th>Case</th>
<th>L</th>
<th>d</th>
<th>( \mu_d \pm \text{SE} )</th>
<th>( \mu_d \pm \text{SE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>0.500</td>
<td>0.498 (0.4%) \pm 0.075 (0.078)</td>
<td>0.510 (2.0%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.500</td>
<td>0.500 (0.0%) \pm 0.046 (0.048)</td>
<td>0.505 (0.1%)</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>0.446</td>
<td>0.447 (0.6%) \pm 0.068 (0.075)</td>
<td>0.455 (2.9%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.446</td>
<td>0.446 (0.0%) \pm 0.042 (0.046)</td>
<td>0.451 (1.1%)</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>0.486</td>
<td>0.488 (0.4%) \pm 0.063 (0.068)</td>
<td>0.497 (2.3%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.486</td>
<td>0.489 (0.6%) \pm 0.039 (0.042)</td>
<td>0.492 (1.2%)</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>0.555</td>
<td>0.556 (0.2%) \pm 0.081 (0.087)</td>
<td>0.572 (3.1%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.555</td>
<td>0.557 (0.4%) \pm 0.049 (0.052)</td>
<td>0.563 (4.1%)</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>0.607</td>
<td>0.608 (0.2%) \pm 0.066 (0.072)</td>
<td>0.613 (1.0%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.607</td>
<td>0.609 (0.3%) \pm 0.030 (0.035)</td>
<td>0.611 (0.7%)</td>
</tr>
</tbody>
</table>

2\( \mu_d = 0.8 \)

<table>
<thead>
<tr>
<th>Case</th>
<th>L</th>
<th>d</th>
<th>( \mu_d \pm \text{SE} )</th>
<th>( \mu_d \pm \text{SE} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>200</td>
<td>0.800</td>
<td>0.791 (1.1%) \pm 0.115 (0.144)</td>
<td>0.838 (4.8%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.800</td>
<td>0.797 (0.4%) \pm 0.076 (0.079)</td>
<td>0.810 (1.3%)</td>
</tr>
<tr>
<td>2</td>
<td>200</td>
<td>0.712</td>
<td>0.697 (2.1%) \pm 0.107 (0.144)</td>
<td>0.743 (4.4%)</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>0.712</td>
<td>0.710 (0.2%) \pm 0.072 (0.077)</td>
<td>0.723 (5.9%)</td>
</tr>
<tr>
<td>3</td>
<td>200</td>
<td>0.770</td>
<td>0.766 (0.5%) \pm 0.094 (0.110)</td>
<td>0.791 (2.7%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.770</td>
<td>0.768 (0.3%) \pm 0.059 (0.065)</td>
<td>0.777 (9.9%)</td>
</tr>
<tr>
<td>4</td>
<td>200</td>
<td>0.858</td>
<td>0.842 (1.9%) \pm 0.123 (0.158)</td>
<td>0.890 (3.7%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.858</td>
<td>0.854 (0.5%) \pm 0.075 (0.082)</td>
<td>0.868 (1.2%)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.858</td>
<td>0.859 (0.1%) \pm 0.039 (0.049)</td>
<td>0.862 (0.5%)</td>
</tr>
<tr>
<td>5</td>
<td>200</td>
<td>0.926</td>
<td>0.880 (5.0%) \pm 0.167 (0.250)</td>
<td>0.986 (6.5%)</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>0.926</td>
<td>0.918 (0.9%) \pm 0.094 (0.114)</td>
<td>0.946 (1.2%)</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.926</td>
<td>0.925 (0.1%) \pm 0.046 (0.051)</td>
<td>0.930 (0.5%)</td>
</tr>
</tbody>
</table>

*Note.—L is the sequence length; \( \bar{d} \) is the true value of the paralinear distance; \( \mu_d \pm \text{SE} \) is the mean and standard error of \( d \) estimated by the bias-corrected paralinear method (the mean of estimated standard error is given in parentheses); and \( \bar{d} \) is the mean of the \( d \) estimated by the method without bias-correction. The percentage values presented in parentheses are the biases of corrected (\( \bar{d} \)) and uncorrected (\( \bar{d} \)) paralinear estimates, respectively. For each case, the number of replications is 1,000. The simulation models are given in table 1.*

In conclusion, when nucleotide frequencies vary among sequences, the paralinear and LogDet distances are suitable. When the sequence length is short, the bias-corrected paralinear or LogDet distance is recommended.

**Discussion**

The main results of our paper can be summarized as follows. (1) Formulas for correcting the estimation biases of the paralinear and LogDet distances were developed and shown by simulation to perform well. (2) The sampling variance for the paralinear distance was developed and examined by simulation. (3) A method for estimating the variance-covariance matrix of paralinear distances was developed, so that statistical tests of phylogenies can be conducted under nonstationary nucleotide frequencies. (4) A new method for testing the molecular clock hypothesis was developed under nonstationary nucleotide frequencies.

Although our method for correcting the estimation bias (equation 19) is simple and efficient, there are two factors that can affect accuracy: (1) equation (18) is ap-
proximate so that the bias may not be completely corrected; and (2) when the sequences are short and very divergent, the bias can be overcorrected. Fortunately, the two biases cancel each other to some extent. Indeed, our examination of many cases with \( L \geq 200 \) and \( 2\mu t \lesssim 0.8 \) showed that the correction performs well. However, in extreme cases such as \( 2\mu t = 1.0 \) and \( L = 100 \), we found that the distances were overcorrected, mainly due to overestimation of the sampling variance of \( d \). In general, as the divergence between two sequences \( (2\mu t) \) increases, the sequence length required for accurate correction also increases. For example, if \( 2\mu t \leq 0.5 \), the corrected estimate is significantly better than that of uncorrected if \( L > 100 \), while if \( 2\mu t = 1.0 \), the required length \( L \) needs to be 500 or larger.

The paralinear and LogDet distances were developed for phylogenetic inference when nucleotide frequencies are nonstationary. The performance of phylogenetic reconstruction can be improved by our new method for correcting the estimation bias when the sequences are short.

There are several analytical methods for testing DNA phylogenies (e.g., Nei, Stephens, and Saitou 1985; Bulmer 1991; Rzhetsky and Nei 1992). In the following, we use the minimum-evolution (ME) criterion (Rzhetsky and Nei 1992) as an example to show how to apply our variance–covariance matrix. Under the ME criterion, tree \( A \) is significantly better than tree \( B \) if the (estimated) total branch length \( (S_A) \) for tree \( A \) is significantly smaller than that \( (S_B) \) for tree \( B \). For example, in the case of four sequences (fig. 3), it can be shown that \( D_{BA} = S_B - S_A \) is equal to

\[
D_{BA} = \frac{1}{4}(d_{13} + d_{24} - d_{12} - d_{34}),
\]

where \( d_{ij} \) is the paralinear distance between sequences \( i \) and \( j \). To test whether \( D \) is significantly larger than zero, we need to calculate the Z score, \( Z = D/\sqrt{V(D)} \), which approximately follows a normal distribution. From equation (29), one can show that

\[
V(D) = [V(d_{13}) + V(d_{24}) + V(d_{12}) + V(d_{34})]/16 + [\text{cov}(d_{13}, d_{24}) + \text{cov}(d_{12}, d_{34})]
\]

\[
- \text{cov}(d_{12}, d_{24}) - \text{cov}(d_{24}, d_{34})]/8,
\]

which are easily computed by equations (21) and (22). Therefore, the method developed in this paper for estimating the variance–covariance matrix of paralinear distances makes it possible to apply the ME criterion under nonstationary nucleotide frequencies.

If the nucleotide frequencies vary considerably among sequences, current methods for testing a molecular clock (e.g., Wu and Li 1985; Gu and Li 1992) may not be suitable because the observed differences may be due to changes in nucleotide frequencies. Our test based on the LogDet distance is more suitable in this case because it is not affected by changes in nucleotide composition.

In principle, our new method can be easily extended to more complex models where the dimension of the rate matrix \( R \) is more than 4 (Lake 1994; Lockhart et al. 1994). Two interesting examples are the amino-acid-based model (a general \( 20 \times 20 \) rate matrix) (Dayhoff, Schwartz, and Orcutt 1978) and the codon-based model (a general \( 61 \times 61 \) rate matrix) (Goldman and Yang 1994; Muse and Gaut 1994). However, our preliminary simulation showed that, even for the amino-acid-based model, the paralinear and LogDet distances are subject to large sampling variances unless the sequence is very long, say, longer than 2,000 amino acids; the sampling variance would be much larger for the codon-based model. Indeed, because there are too many unknown parameters, the distance cannot be estimated accurately. Thus, one should be cautious when applying our bias-corrected methods to analyze amino acid sequence data.

Acknowledgments

This research was supported by NIH grants. We thank Dr. Michael A. Steel for comments.

APPENDIX

Some Properties of the Paralinear and LogDet Distances for Phylogenetic Inference

Denote the joint-probability matrix of nodes \( i \) and \( j \) by \( J^{(i)} \), whose \( kl \)-th element is the probability that the nucleotide is \( k \) at node \( i \) and \( l \) at node \( j \). Note that, if both nodes \( i \) and \( j \) are external (i.e., representing sequences \( i \) and \( j \), \( J^{(i)} \) is also called the data matrix (see eqs. 1 and 3). If the evolution is from node \( i \) to node \( j \), we have

\[
J^{(i)} = F^{(i)} P^{(i)},
\]

where \( P^{(i)} \) is the transition probability matrix from node \( i \) to node \( j \) and \( F^{(i)} \) is the diagonal matrix of nucleotide frequencies at node \( i \). How-
ever, if the common ancestor of nodes $i$ and $j$ is node
0, the relationship is different: $J(0) = P(O_i)'F(0)P(O_i)$ (see also eq. 5).

First, we show that the additivity of the LogDet distance does not extend to internal nodes or branches.
Consider the case that node $k$ is between nodes $i$ and $j$ and the evolution is from node $i$ to node $k$ to node j. By equation (2), we have $d_{ik} = -0.25 \ln \det[J_{ik}] - \ln 4$, $d_{kj} = -0.25 \ln \det[J_{kj}] - \ln 4$, and $d_{ij} = -0.25 \ln \det[J_{ij}] - \ln 4$. According to the Markov property, $P(G) = P(ik)P(w)$, we can show that $d_{ik} + d_{kj} = -0.25 \ln \det[J_{ik}] - \ln 4$ = $d_{ij}$.

Therefore, unless $\det[J_{ik}] = (1/4)^4$, $d_{ik} + d_{kj} \neq d_{ij}$ and so the additivity does not hold. For the case where $k$ is the common ancestor of nodes $i$ and $j$, one can also show $d_{ki} + d_{kj} \neq d_{ij}$.

Second, we show that the four-point condition holds under the paralinear and LogDet distances. If the true tree is as shown in figure 4, this condition is expressed by

\[ Q = d_{15} + d_{24} - d_{12} - d_{34} \geq 0 \]

We will show that the four-point condition holds when $d_{ij}$'s are paralinear distances, indicating that the two distance measures can reconstruct the correct tree. In the case of the LogDet distance, Steel (1994) reached this conclusion by another approach.

Here we only show the first equation of equation (A.2); the second can be shown in a similar manner. For both distance measures,

\[ Q = -\frac{1}{4} \ln \frac{\det[J^{(3)}P_{56}]\det[J^{(56)}]}{\det[P_{56}]}. \]  

For the tree in figure 4b, the following relations hold by the Markov property: $J^{(13)} = P^{(13)}P_{56}P^{(56)}P_{61}$, $J^{(24)} = P^{(24)}P_{66}P^{(66)}P_{64}$, $J^{(12)} = P^{(12)}P_{64}P^{(64)}P_{62}$, and $J^{(34)} = P^{(34)}P_{62}P^{(62)}P_{66}$, where $'$ means transpose. Putting these into equation (A.3), we have

\[ Q = -\frac{1}{4} \ln \frac{\det[J^{(13)}P_{56}]\det[J^{(56)}]}{\det[P_{56}]}. \]  

Similarly, for the tree in figure 4a, we have $J^{(13)} = P^{(13)}P_{55}P^{(55)}P_{56}$, $J^{(24)} = P^{(24)}P_{65}P^{(65)}P_{64}$, $J^{(12)} = P^{(12)}P_{54}P^{(54)}P_{52}$, and $J^{(34)} = P^{(34)}P_{64}P^{(64)}P_{66}$, and

\[ Q = -\frac{1}{2} \ln \frac{\det[J^{(05)}]}{\sqrt{\det[J^{(05)}]\det[F^{(05)}]}} - \frac{1}{2} \ln \frac{\det[J^{(06)}]}{\sqrt{\det[J^{(06)}]\det[F^{(06)}]}}. \]

\[ Q = 2b_{56} \geq 0. \]

Therefore, in both cases (fig. 4a and b), $Q \geq 0$.

Finally, we show that the neighbor-joining (NJ) algorithm (Saitou and Nei 1987) gives the same topology for the paralinear and LogDet distances. By definition, the paralinear distance between sequences $i$ and $j$, $d_{ij}$, can be expressed as

\[ d_{ij} = d'_{ij} + A_i + A_j, \]

where $d'_{ij}$ is the LogDet distance between sequences $i$ and $j$, and $A_i = \ln \det[F_i]/8 + \ln 2$, $i = 1, 2$. In terms of the paralinear distance, the criterion of the NJ algorithm for choosing sequences 1 and 2 as neighbors is that

\[ S_{1,2} = (r - 2)d_{12} - \sum_{i=1}^{r} (d_{1i} + d_{2i}) \]

is the smallest among all $S_{ij}$, where $r$ is the number of sequences (Studier and Keppler 1988). Let $S'_{1,2}$ be the NJ criterion in terms of the LogDet distance. Then,

\[ S_{1,2} = S'_{1,2} - 2 \sum_{i=1}^{r} A_i. \]

Thus, $S_{1,2}$ and $S'_{1,2}$ are equivalent because they differ only by the constant $-2 \Sigma_{i=1}^{r} A_i$.

Without loss of generality, we assume that the first iteration of the NJ algorithm chooses sequences 1 and 2 as a pair of neighbors. So in the second iteration, the paralinear distance between the composite (1, 2) and sequence $i$, denoted by $d_{1(2)i}$, is computed by $d_{1(2)i} = (d_{1i} + d_{2i})/2$, and for the LogDet distance, $d'_{1(2)i} = (d'_{1i} + d'_{2i})/2$. The tree is then recomputed.
+ d'_{2k}/2$, so that $d_{1,2k} = d'_{1,2k} + A_{1,2} + A_t$, where
$A_{1,2} = (A_t + A_2)/2$. Therefore, by the same argument
as in the first iteration, one can show that the same new
pair of neighbors will be chosen regardless of whether
the paralinear or the LogDet distance is used. Therefore,
the paralinear and LogDet distances both give the same
tree topology when the NJ method is used.

Note that the gradual neighborliness algorithm
(Fitch 1981) is equivalent to the NJ algorithm (Gascuel
1994), and that the method of Sattath and Tversky
(1977) is an application of the four-point condition.
Therefore, the paralinear and LogDet distances should
give the same tree topology under these two methods.

LITERATURE CITED

LogDeterminant transformation for evolutionary trees. Appl.

BARRY, D., and J. A. HARTIGAN. 1987. Asynchronous distanc-
eses between homologous DNA sequences. Biometrics 43:
261–276.

BISHOP, M. J., and A. E. FRIDAY. 1988. Estimating the inter-
relationship of tetrapod groups on the basis of molecular
sequence data. Pp. 35–38 in M. J. BENTON, ed. The phy-
genetics and classification of tetrapods. Vol. 1. Clarendon,
Oxford.

squares in reconstructing phylogenies from sequence data.

CAYNDER, J. A., and J. FELSENSTEIN. 1987. Invariants of
phylogenies: simple case with discrete states. J. Classif. 4:
57–71.

A model of evolutionary change in proteins. Pp. 345–352
in M. O. DAYHOFF, ed. Atlas of protein sequence structure.
Vol. 5, suppl. 3. National Biomedical Research Foundation,
Washington, D.C.

FITCH, W. M. 1981. A non-sequential method for constructing
trees and hierarchical classifications. J. Mol. Evol. 18:30–
37.

GASCUEL, O. 1994. A note on Sattath and Tversky's, Saitou
and Nei's, and Studier and Keppler's algorithms for infer-
ing phylogenies from evolutionary distances Mol Biol
Evol. 11:961–963.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of
nucleotide substitution for protein-coding DNA sequences.

GU, X., and W. H. LI. 1992. Higher rates of amino acid sub-

1996. A general additive distance with time-revers-
ibility and rate variation among nucleotide sites. Proc. Natl.
Acad. Sci. USA 93:4671–4676.

HASEGAWA, M., and T. HASHIMOTO. 1993. Ribosomal RNA

HEDGES, S. B., K. D. MOBERG, and L. R. MAXSON. 1990. Tet-
rapod phylogeny inferred from 18S and 28S ribosomal se-
quences and a review of the evidence for amniote relation-

molecules. Pp. 21–32 in H. R. MUNRO, ed. Mammalian pro-

KIMURA, M. 1980. A simple method for estimating evolution-
ary rate of base substitutions through comparative studies

LAKE, J. A. 1994. Reconstructing evolutionary trees from DNA
Acad. Sci. USA 91:1455–1459.

A new method for calculating evolutionary substitution

LI, W. H., and A. ZHARKIKH. 1995. Statistical tests of DNA

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, and D. PENNY.
1994. Recovering evolutionary trees under a more realistic

comparing synonymous and nonsynonymous nucleotide
substitution rates, with application to the chloroplast ge-

computing the standard errors of branching points in an
evolutionary tree and their application to molecular data

mating and testing minimum-evolution trees. Mol. Biol.
Evol. 9:945–967.

SAIOUT, N., and M. NEI. 1987. The neighbor-joining method:
a new method for reconstruction of phylogenetic trees. Mol.


STEEL, M. A. 1994. Recovering a tree from the leaf coloura-
7:19–24.

bor-joining method of Saitou and Nei. Mol. Biol. Evol. 5:
729–731.

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of
nucleotide substitution in rodents than in man. Proc. Natl

ZHARKIKH, A. 1994. Estimation of evolutionary distances be-

MANOLO GOY, reviewing editor

Accepted August 20, 1996