Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 4671–4676, May 1996
Evolution

A general additive distance with time-reversibility and rate variation among nucleotide sites

(substitution models/linear distance/rate heterogeneity/Markovian)

XUN GU AND WEN-HSIUNG LI*

Human Genetics Center, School of Public Health, University of Texas, P.O. Box 20334, Houston, TX 77225

Communicated by Michael T. Clegg, University of California, Riverside, CA, January 2, 1996 (received for review February 22, 1995)

ABSTRACT **As additivity is a very useful property for a distance measure, a general additive distance is proposed under the stationary time-reversible (SR) model of nucleotide substitution or, more generally, under the stationary, time-reversible, and rate variable (SRV) model, which allows rate variation among nucleotide sites. A method for estimating the mean distance and the sampling variance is developed. In addition, a method is developed for estimating the variance–covariance matrix of distances, which is useful for the statistical test of phylogenies and molecular clocks. Computer simulation shows (*i*) if the sequences are longer than, say, 1000 bp, the SR method is preferable to simpler methods; (*ii*) the SR method is robust against deviations from time-reversibility; (*iii*) when the rate varies among sites, the SRV method is much better than the SR method because the distance is seriously underestimated by the SR method; and (*iv*) our method for estimating the sampling variance is accurate for sequences longer than 500 bp. Finally, a test is constructed for testing whether DNA evolution follows a general Markovian model.**

Additivity is a highly desirable property for evolutionary distances. In fact, if additivity does not hold, all distance matrix methods of tree reconstruction may become statistically inconsistent, leading to an erroneous tree with a probability approaching 1 as the sequence length increases to infinity (1). A distance measure may be nonadditive if an oversimplified model of nucleotide substitution is used (2). Therefore, a general model of nucleotide substitution is desirable. We consider the stationary time reversible (SR) model which assumes only that the substitution process is stationary and reversible in time; it includes many models as special cases—e.g., the models of Jukes and Cantor (3), Hasegawa *et al.* (4), and Tamura and Nei (5).

The SR model has been studied by several authors (6–11). However, much remains to be investigated. First, the SR model assumes that the substitution rate is the same for all sites. This assumption is unrealistic for most genes because the functional constraints are usually different at different sites (12). As a distance measure may become nonadditive if the rate varies among sites, it is important to extend the SR model to the SRV model (stationary, time-reversible, and rate-variable). Second, the sampling variance of distance has not been well developed. Third, it is not clear how to estimate the variance–covariance matrix of a distance matrix under the SR or SRV model, though some simple cases have been studied (13–15). To solve these problems, we first develop a general distance measure under the SR model because the existing framework (6, 9) does not seem suitable for our purpose. Our SR method gives estimates similar to those of Lanave *et al.* (6) and Rodriguez *et al.* (9), but our formulation is more convenient for generalization, particularly for extension to include rate variation among sites. Extensive simulation will be conducted to study the biasness and robustness of our SR and SRV methods and to check the accuracy of our method for estimating the sampling variance.

METHODS

General Additive Distance Under the SR Model. Suppose that nucleotide substitution follows a stationary Markov process. Denote A, G, T, and C as 1, 2, 3, and 4, respectively. Let **R** be the rate matrix whose ijth element r_{ij} is the rate of change from nucleotide i to nucleotide j if $i \neq j$ ($i, j = 1, 2, 3, 4$); the diagonal elements are given by $r_{ii} = -\sum_{j \neq i} r_{ij}$. Then the matrix of transition probabilities **P** for t time units is given by $\mathbf{P}(t) = e^{\mathbf{R}t}$, where the ijth element of $\mathbf{P}(t)$ is $P_{ij}(t)$—i.e., the probability of transition from nucleotide i to nucleotide j after t time units (6).

The substitution process is reversible in time if and only if

$$\pi_i r_{ij} = \pi_j r_{ji}, \qquad i,j = 1, 2, 3, 4, \qquad \textbf{[1]}$$

where π_i is the equilibrium frequency of nucleotide i. Eq. **1** implies that the off-diagonal elements of rate matrix **R** can be expressed as

	A	G	T	C
A		$\pi_2 s_1$	$\pi_3 v_1$	$\pi_4 v_2$
G	$\pi_1 s_1$		$\pi_3 v_3$	$\pi_4 v_4$
T	$\pi_1 v_1$	$\pi_2 v_3$		$\pi_4 s_2$
C	$\pi_1 v_2$	$\pi_2 v_4$	$\pi_3 s_2$	

Thus, the SR model is a nine-parameter model (10, 11).

Now let us consider two sequences X and Y that have evolved from O, a common ancestor, t time units ago under the SR model. Time-reversibility means that the substitution process from the common ancestor O to sequences X and Y is equivalent to the substitution process from X to Y (or from Y to X), whose transition probability matrix (for $2t$ time units) is given by

$$\mathbf{P}(2t) = e^{2t\mathbf{R}}. \qquad \textbf{[2]}$$

Let λ_k ($k = 1, 2, 3, 4$) be the ith eigenvalue of **R**; one of them is zero, say $\lambda_4 = 0$. Let z_k be the kth eigenvalue of $\mathbf{P}(2t)$. Eq. **2** implies that λ_k and z_k satisfy

$$z_i = e^{2t\lambda_i} \text{ or } \ln z_i = 2\lambda_i t. \qquad \textbf{[3]}$$

Since $\lambda_4 = 0$ and $z_4 = 1$, there are only three nontrivial equations in Eq. **3**. So, under the SR model, a general additive distance can be defined as

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: SR, stationary time-reversible; SRV, stationary, time-reversible, rate variable.

*To whom reprint requests should be addressed.

$$d = -\sum_{k=1}^{3} c_k \ln z_k \qquad [4]$$

for some constants c_k ($k = 1, 2, 3$).

Keilson (16) showed that all eigenvalues z_k (or, equivalently, λ_k) are real under the SR model. A simple proof is given below. Introduce matrix $\mathbf{M}$ whose ijth element is defined by $m_{ij} = \sqrt{\pi_i} r_{ij} / \sqrt{\pi_j}$, $i, j = 1, \ldots, 4$. Using Eq. **1**, one can show that

$$m_{ij} - m_{ji} = \frac{\pi_i r_{ij} - \pi_j r_{ji}}{\sqrt{\pi_i \pi_j}} = 0. \qquad [5]$$

Thus, $\mathbf{M}$ is symmetric and all eigenvalues of $\mathbf{M}$ are real. Let $\mathbf{T} = \text{diag}(\sqrt{\pi_1}, \sqrt{\pi_2}, \sqrt{\pi_3}, \sqrt{\pi_4})$ be the diagonal matrix of the square roots of nucleotide frequencies. Then, one can verify that

$$\mathbf{R} = \mathbf{T}\mathbf{M}\mathbf{T}^{-1}. \qquad [6]$$

Therefore, $\mathbf{R}$ and $\mathbf{M}$ have the same eigenvalues because $\det(\lambda\mathbf{I} - \mathbf{R}) = \det(\lambda\mathbf{I} - \mathbf{M})$.

Eq. **4** gives a general form that includes many distance measures so that the sampling variance and the extension to the SRV model can be treated generally. The appropriate constants c_k in Eq. **4** for the number of nucleotide substitutions per site (K), the number of transitional substitutions per site (A), the number of transversional substitutions per site (B), and the number of substitutions from nucleotides i to j (D_{ij}) are presented in Table 1, where u_{ij} and v_{ij} are the ijth elements of eigenmatrix $\mathbf{U}$ and $\mathbf{V} = \mathbf{U}^{-1}$, respectively. These results can be derived from the spectral decomposition of $\mathbf{R}$—i.e., the ijth element of $\mathbf{R}$ can be expressed as

$$r_{ij} = \sum_{k=1}^{3} u_{ik} v_{kj} \lambda_k = \frac{1}{2t} \sum_{k=1}^{3} u_{ik} v_{kj} \ln z_k \qquad [7]$$

for $i \neq j$. For example, the constants c_k for K can be obtained from the definition of $K = 2t \sum_{i=1}^{4} \pi_i \sum_{j \neq i} r_{ij}$.

General Additive Distance Under the SRV Model. We first assume that the rate variation among sites follows a gamma distribution—i.e., $r_{ij} = a_{ij} u$ ($i \neq j$), where a_{ij} is a constant and u varies according to the following gamma distribution

$$\phi(u) = \frac{\beta^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\beta u}. \qquad [8]$$

That is, $\mathbf{R} = u\mathbf{A}$ where matrix $\mathbf{A}$ consists of a_{ij}. Let $\bar{\mathbf{R}} = \bar{u}\mathbf{A}$ be the expected rate matrix where $\bar{u} = \alpha/\beta$ is the mean of u. Then, for any $k = 1, 2, \ldots$, we have

$$\int_0^\infty \mathbf{R}^k \phi(u) du = \mathbf{A}^k \int_0^\infty u^k \phi(u) du = \bar{\mathbf{R}}^k \frac{\Gamma(k+\alpha)}{\alpha^k \Gamma(\alpha)}. \qquad [9]$$

Table 1. Constants c_k in the general additive distance under the SR model (Eq. **4**) or the SRV model (Eq. **12**)

Distance	c_k ($k = 1, 2, 3$)
K	$-\sum_{i=1}^{4} \sum_{j \neq i} \pi_i u_{ik} v_{kj}$
A	$-\sum_{i=1}^{4} \sum_{j \neq i \in \text{Ts}} \pi_i u_{ik} v_{kj}$
B	$-\sum_{i=1}^{4} \sum_{j \neq i \in \text{Tv}} \pi_i u_{ik} v_{kj}$
D_{ij}	$-\pi_i u_{ik} v_{kj}$

K is the number of substitutions per site, A is the number of transitional substitutions per site, B is the number of transversional substitutions per site, and D_{ij} is the number of substitutions from nucleotides i to j per site. The subscripts $j \neq i \in$ Ts and $j \neq i \in$ Tv mean that the differences between nucleotides i and j are transitional (Ts) and transversional (Tv), respectively.

Let $\bar{\mathbf{P}}(2t) = \int_0^\infty e^{2\mathbf{R}t} \phi(u) du$ be the matrix of average transition probabilities, whose ijth element is $\bar{P}_{ij}(2t) = \int_0^\infty P_{ij}(2t) \phi(u) du$. Since $e^{2\mathbf{R}t} = \mathbf{I} + \sum_{k=1}^{\infty} (2t\mathbf{R})^k / k!$,

$$\bar{\mathbf{P}}(2t) = \int_0^\infty e^{2\mathbf{R}t} \phi(u) du = \mathbf{I} + \sum_{k=1}^{\infty} \frac{(2t)^k}{k!} \int_0^\infty \mathbf{R}^k \phi(u) du$$

$$= \mathbf{I} + \sum_{k=1}^{\infty} \frac{(2t)^k}{k!} \bar{\mathbf{R}}^k \frac{\Gamma(k+\alpha)}{\alpha^k \Gamma(\alpha)}. \qquad [10]$$

Therefore, the ith eigenvalue of $\bar{\mathbf{P}}(2t)$, $\bar{z}_i$, is given by

$$\bar{z}_i = 1 + \sum_{k=1}^{\infty} \frac{(2t)^k}{k!} \bar{\lambda}_i^k \frac{\Gamma(k+\alpha)}{\alpha^k \Gamma(\alpha)} = \left(1 - \frac{2\bar{\lambda}_i t}{\alpha}\right)^{-\alpha}, \qquad [11]$$

where $\bar{\lambda}_i$ is the ith eigenvalue of $\bar{\mathbf{R}}$. From Eq. **11**, we have $-2\bar{\lambda}_i t = \alpha(\bar{z}_i^{-1/\alpha} - 1)$, ($i = 1, \ldots, 4$). It follows that the general additive distance under the SRV model is given by

$$\bar{d} = \alpha \sum_{k=1}^{3} c_k (\bar{z}_k^{-1/\alpha} - 1). \qquad [12]$$

Obviously, $\bar{d} \to d$ of Eq. **4** as $\alpha \to \infty$—i.e., if the substitution rate is uniform among sites. The constants c_k are determined in the same manner as above (see Table 1).

Eq. **12** can be generalized to any distribution $f(u)$ of rate variation among sites. Let $G(s) = \int_0^\infty e^{su} f(u) du$ be the moment-generating function of $f(u)$. Then, similar to the derivation of Eq. **11**, one can show that

$$\bar{z}_k = G(2\bar{\lambda}_k t), \qquad k = 1, 2, 3, 4. \qquad [13]$$

Thus, the general additive distance is given by

$$\bar{d} = -\sum_{k=1}^{3} c_k G^{-1}(\bar{z}_k), \qquad [14]$$

where G^{-1} is the inverse function of the moment-generating function G. For example, let us consider the case where the substitution rate varies according to an invariant + gamma distribution (17), which can be described as follows. For a given site, the probability of being invariable (i.e., the substitution rate at this site is 0) is θ, while the probability of being variable is $1 - \theta$. Furthermore, among the sites that are variable, the substitution rate follows a gamma distribution. By applying Eq. **13**, one can show that

$$\bar{z}_k = \theta + (1-\theta)\left\{1 - \frac{2\bar{\lambda}_k t}{(1-\theta)\alpha}\right\}^{-\alpha}, \qquad k = 1, 2, 3, 4. \qquad [15]$$

Then, the general additive distance under the invariant + gamma distribution is given by

$$\bar{d} = (1-\theta)\alpha \sum_{k=1}^{3} c_k \left\{\left(\frac{\bar{z}_k - \theta}{1-\theta}\right)^{-1/\alpha} - 1\right\}. \qquad [16]$$

Estimation of Distances and Sampling Variances. Since the SR model is a special case of the SRV model, the two models can be treated in the same way. At a particular site, let J_{ij} be the probability of having nucleotide i in sequence X and nucleotide j in sequence Y. Then, $J_{ij} = \sum_{k=1}^{4} \pi_k P_{ki}(t) P_{kj}(t)$. Since Eq. **1** (time-reversibility) implies $\pi_i P_{ij}(t) = \pi_j P_{ji}(t)$, J_{ij} can be simplified as $J_{ij} = \pi_i P_{ij}(2t)$. Thus, by taking expectation over all sites, the expected frequency $\bar{J}_{ij}$ is equal to

$$\bar{J}_{ij} = \pi_i \bar{P}_{ij}(2t). \qquad [17]$$

From Eq. **17**, the ijth element of $\bar{\mathbf{P}}(2t)$ can be estimated by

$$\hat{P}_{ij} = \frac{\hat{J}_{ij}}{\hat{\pi}_i}, \qquad i, j = 1, \ldots 4, \qquad [18]$$

where $\hat{\pi}_i$ is the frequency of nucleotide i estimated by taking (simple) average between sequences X and Y, and $\hat{J}_{ij} = (N_{ij} + N_{ji})/(2L)$; N_{ij} is the number of sites at which the nucleotide is i in sequence X and j in sequence Y, and L is the sequence length. Note that, if the substitution process is stationary and reversible, then $J_{ij} = J_{ji}$, so that $E(N_{ij}) = E(N_{ji})$ (E means taking expectation). However, because of sampling effects, this condition holds only approximately. For this reason, we use $(N_{ij} + N_{ji})/(2L)$ to estimate J_{ij}, so that the condition $\hat{J}_{ij} = \hat{J}_{ji}$ always holds. This treatment makes our method approximately applicable even when the substitution process is not time reversible.

Let matrix $\hat{\mathbf{P}}$ consist of $\hat{P}_{ij}$. Then, the eigenvalues $\hat{z}_i$ ($i = 1, \ldots, 4$) can be computed by a standard algorithm; the corresponding eigenmatrix $\mathbf{U}$ and its inverse matrix $\mathbf{V}$ are also obtained simultaneously (18). Thus, the SR distance can be estimated rather quickly. Under the SRV model, however, the parameters for the rate variation among sites (e.g., the gamma distribution parameter α and/or the proportion of invariable sites θ) should be estimated first. One may use either a parsimony method (5, 12) or a maximum likelihood method (17, 19).

Now we consider the sampling variance. Suppose N_{ij} follows a multinomial distribution. By the delta method, the approximate sampling variance of d under the SR model (Eq. **4**) is

$$\mathrm{Var}(d) = \sum_{i=1}^{3} q_i^2 \mathrm{Var}(z_i) + 2 \sum_{i<j} q_i q_j \mathrm{Cov}(z_i, z_j), \qquad [19]$$

where $q_i = c_i/\hat{z}_i$, and the approximate sampling variance of $\bar{d}$ under the SRV model (Eq. **12**) is

$$\mathrm{Var}(\bar{d}) = \sum_{i=1}^{3} \bar{q}_i^2 \mathrm{Var}(z_i) + 2 \sum_{i<j} \bar{q}_i \bar{q}_j \mathrm{Cov}(z_i, z_j) + \bar{q}_\alpha^2 \mathrm{Var}(\alpha), \qquad [20]$$

where $\bar{q}_i = c_i/\hat{z}_i^{1+1/\alpha}$ and $\bar{q}_\alpha = \Sigma_{i=1}^3 c_i[(1 + \ln \hat{z}_i/\alpha)\hat{z}_i^{-1/\alpha} - 1]$. We use the likelihood function to compute $\mathrm{Var}(z_i)$ and $\mathrm{Cov}(z_i, z_j)$, which are approximate because we do not use the maximum likelihood estimates of z_is. Let ℓ be the log-likelihood function of the P_{ij}s between the two sequences. It is easy to show that $\ell = \text{const.} + \Sigma_{i=1}^4 \Sigma_{j=1}^4 N_{ij} \ln P_{ij}$. Since the transition probability P_{ij} can be decomposed as $P_{ij} = \pi_j + \Sigma_{k=1}^3 u_{ik} v_{kj} z_k$, the klth element of the information matrix $\mathbf{I}_f$ can be computed by

$$I_{kl} = -\frac{\partial^2 \ell}{\partial z_k \partial z_l} = \sum_{i,j} \frac{N_{ij}}{\hat{P}_{ij}^2} u_{ik} u_{il} v_{kj} v_{lj}, \qquad k,l = 1, 2, 3. \qquad [21]$$

The variance–covariance matrix of z_is can be obtained by inverting $\mathbf{I}_f$.

Variance–Covariance Matrix of the Distance Matrix. Statistical testing of a phylogenetic hypothesis (13) or the molecular clock (14) based on distance-matrix methods requires the variance–covariance matrix of distances. We propose a simple method to estimate the covariance between two distances (say, d_1 and d_2). Using Eqs. **4** and **12**, one can show that under the SR model

$$\mathrm{Cov}(d_1, d_2) = \sum_{i,j=1}^{3} q_i^{(1)} q_j^{(2)} \mathrm{Cov}(z_i^{(1)}, z_j^{(2)}) \qquad [22]$$

and under the SRV model

$$\mathrm{Cov}(\bar{d}_1, \bar{d}_2) = \sum_{i,j=1}^{3} \bar{q}_i^{(1)} \bar{q}_j^{(2)} \mathrm{Cov}(z_i^{(1)}, z_j^{(2)}) + \bar{q}_\alpha^{(1)} \bar{q}_\alpha^{(2)} \mathrm{Var}(\alpha), \qquad [23]$$

where the superscripts indicate distances 1 and 2, respectively.

To compute $\mathrm{Cov}(z_i^{(1)}, z_j^{(2)})$, we note that the transition probability matrix can be written as $\mathbf{P}(2t) = \mathbf{UZV}$, where $\mathbf{Z} = \mathrm{diag}(z_1, z_2, z_3, z_4)$. Thus, $\mathbf{Z} = \mathbf{VP}(2t)\mathbf{U}$. From the relation $J_{ij} = \pi_i P_{ij}(2t)$, it follows that

$$z_k = \sum_{i,j} \frac{v_{ki} u_{jk}}{\pi_i} J_{ij} = \sum_{i,j} b_{ijk} J_{ij}, \qquad k = 1, 2, 3, \qquad [24]$$

where $b_{ijk} = v_{ki} u_{jk}/\pi_i$.

We distinguish two situations: (*i*) three sequences are involved—e.g., d_1 is for sequences 1 and 2 and d_2 is for sequences 2 and 3; and (*ii*) four sequences are involved—e.g., d_1 is for sequences 1 and 2 and d_2 is for sequences 3 and 4. Suppose N_{ij} follows a multinomial distribution. For the first situation, $\mathrm{Cov}(z_\nu^{(1)}, z_\mu^{(2)})$ can be estimated by

$$\mathrm{Cov}(z_\nu^{(1)}, z_\mu^{(2)}) = \sum_{i,j,k} b_{ij\nu} b_{jk\mu} \mathrm{Cov}(J_{ij}^{(1)}, J_{jk}^{(2)}) = \sum_{i,j,k} b_{ij\nu} b_{jk\mu} \frac{\hat{f}_{ijk} - \hat{J}_{ij}\hat{J}_{jk}}{L}, \qquad [25]$$

where $\hat{f}_{ijk}$ is the frequency that the nucleotides in sequences 1, 2, and 3 are i, j, k, respectively. Similarly, for the second situation, we have

$$\mathrm{Cov}(z_\nu^{(1)}, z_\mu^{(2)}) = \sum_{i,j,k,l} b_{ij\nu} b_{kl\mu} \mathrm{Cov}(J_{ij}^{(1)}, J_{kl}^{(2)}) = \sum_{i,j,k,l} b_{ij\nu} b_{kl\mu} \frac{\hat{f}_{ijkl} - \hat{J}_{ij}\hat{J}_{kl}}{L}, \qquad [26]$$

where $\hat{f}_{ijkl}$ is the frequency that the nucleotides in sequences 1, 2, 3, and 4 are i, j, k, and l, respectively. Note that the variance and covariance of z_is in Eqs. **19** and **20** can also be computed by

$$\mathrm{Cov}(z_\nu, z_\mu) = \sum_{i,j,k,l} b_{ij\nu} b_{kl\mu} \mathrm{Cov}(J_{ij}, J_{kl}) = \sum_{i,j} b_{ij\nu} b_{ij\mu} \frac{\hat{J}_{ij}(1 - \hat{J}_{ij})}{L} - 2 \sum_{ij>kl} b_{ij\nu} b_{kl\mu} \frac{\hat{J}_{ij}\hat{J}_{kl}}{L} \qquad [27]$$

SIMULATION STUDY

By computer simulation, the SR method was compared with Jukes and Cantor's method (3), Kimura's two-parameter method (20), Tajima and Nei's method (21), and Tamura and Nei's method (5). As expected, when the sequences were long, the SR method performed well, while a simpler method gave biased estimates if some of the assumptions of the method were violated (results not shown); these are consistent with the previous results (9, 10). Because the actual substitution pattern of DNA evolution may be complex, the SR method is preferred in data analysis, if the sequences are long, say $L \geq 1000$ bp.

The performance of the SR method under various models is shown in Table 2. We set the expected number of nucleotide substitutions per site as $K = 0.2$, 0.5, or 1.0, and the sequence length as $L = 200$, 500, or 2000. The first two models are time reversible and the last two are not (see footnote of Table 2). For each case, the observed standard error and the mean of standard errors estimated by the new method (in parentheses) are presented. The results can be summarized as follows. First, the SR method may give large biases when $L \leq 200$. However, it is asymptotically unbiased—i.e., the estimation bias decreases to zero as L increases; this property was examined by increasing L to 10,000 bp where no estimation bias was

Table 2. Mean and standard error of the K value estimated by the SR method

Models	Sequence length, L		
	200	500	2000
$K = 0.2$			
TmN	0.206 ± 0.038 (0.036)	0.203 ± 0.025 (0.025)	0.200 ± 0.012 (0.013)
SR	0.204 ± 0.038 (0.035)	0.202 ± 0.024 (0.023)	0.200 ± 0.012 (0.012)
NR1	0.203 ± 0.039 (0.035)	0.202 ± 0.023 (0.023)	0.200 ± 0.011 (0.012)
NR2	0.207 ± 0.039 (0.034)	0.202 ± 0.025 (0.023)	0.200 ± 0.012 (0.012)
$K = 0.5$			
TmN	0.521 ± 0.083 (0.083)	0.510 ± 0.052 (0.051)	0.502 ± 0.025 (0.025)
SR	0.518 ± 0.078 (0.078)	0.503 ± 0.048 (0.047)	0.502 ± 0.024 (0.023)
NR1	0.519 ± 0.082 (0.082)	0.503 ± 0.052 (0.050)	0.502 ± 0.025 (0.025)
NR2	0.521 ± 0.088 (0.083)	0.504 ± 0.054 (0.053)	0.502 ± 0.026 (0.026)
$K = 1.0$			
TmN	1.080 ± 0.262 (0.387)	1.043 ± 0.167 (0.184)	1.011 ± 0.072 (0.067)
SR	1.101 ± 0.253 (0.343)	1.039 ± 0.134 (0.129)	1.008 ± 0.057 (0.056)
NR1	1.093 ± 0.264 (0.366)	1.034 ± 0.139 (0.141)	1.008 ± 0.067 (0.063)
NR2	1.093 ± 0.300 (0.448)	1.042 ± 0.164 (0.165)	1.010 ± 0.074 (0.072)

TmN = Tamura and Nei's model (5); the equilibrium nucleotide frequencies are $\pi_1 = 0.1$, $\pi_2 = 0.3$, $\pi_3 = 0.2$, and $\pi_4 = 0.4$; and the two transition/transversion ratios are $s_1/v_1 = 5$ and $s_2/v_1 = 4$. SR = π_i are the same as those in TmN and $s_1/v_1 = s_2/v_1 = 3$, $v_2/v_1 = 2$, $v_3/v_1 = 3$, and $v_4/v_1 = 5$. NR1 = Let R_{ij}^{nr} and R_{ij} be the substitution rates from nucleotides i to j under NR1 and SR, respectively. Let $R_{ij}^{nr} = R_{ij}$ if $j = A$ or T but $R_{ij}^{nr} = 2R_{ij}$ if j = G or C so that NR1 is not time reversible. NR2 = Similar to NR1 except that $R_{ij}^{nr} = 4R_{ij}$ if j = G or C so that the irreversibility is more serious than NR1. The number of replications is 1000. The percent of inapplicable cases was ≤10% in the case of $K = 1.0$ and $L = 200$, ≤1% in the case of $K = 0.5$ and $L = 200$, but was 0% for all other cases. Inapplicable cases were excluded from computation. Note that the mean of estimated standard errors is given in parentheses.

observed. Second, under the NR1 and NR2 models, which are not time reversible, the SR method gives accurate estimates when $L \geq 500$, suggesting that the method is robust against violations of time-reversibility. Therefore, the assumption of time-reversibility, which simplifies the estimation problem greatly, may not have serious effects on the distance estimation. Third, the standard error estimated by the new method (Eqs. **19** and **20**) is quite close to the observed one in most cases. The only cases with a poor performance are those with a short sequence ($L < 200$) and a large distance ($K \geq 1.0$); the sampling variance may be overestimated in such a case. We have examined many cases with $L \geq 200$ and $K \leq 1.0$ and found that the estimated sampling variances were accurate.

Table 3 compares the performances of the SR method (Eq. **4**) and the SRV method (Eq. **12**) when the substitution rate varies among sites according to a gamma distribution. We set the gamma distribution parameter α equal to 0.5, 1.0, or 2.0, which may represent strong, intermediate, and weak rate variation among sites, respectively. The sequence length L is 200, 500, or 2,000 and the expected number of substitutions per site is $K = 0.3$ or 0.5. The SRV method appears to be asymptotically unbiased, though it tends to overestimate the distance when the sequence length is short. In all cases, the distance is seriously underestimated by the SR method. When $K = 0.3$, the SR method gives about 27%, 17%, and 10% underestimation of the distance for $\alpha = 0.5$, 1.0, and 2.0, respectively. As K increases to 0.5, it gives about 38%, 26%, and 15% underestimation of K for $\alpha = 0.5$, 1.0, and 2.0,

Table 3. Means and standard errors of the K value estimated by the SR and SRV methods

L	K	SR method	SRV method
$\alpha = 0.5$			
200	0.3	0.219 ± 0.039 (0.028)	0.320 ± 0.085 (0.063)
	0.5	0.316 ± 0.050 (0.042)	0.545 ± 0.170 (0.136)
500	0.3	0.216 ± 0.024 (0.023)	0.307 ± 0.049 (0.047)
	0.5	0.310 ± 0.031 (0.030)	0.513 ± 0.082 (0.081)
2000	0.3	0.215 ± 0.012 (0.012)	0.303 ± 0.023 (0.023)
	0.5	0.308 ± 0.015 (0.015)	0.506 ± 0.039 (0.039)
$\alpha = 1.0$			
200	0.3	0.254 ± 0.042 (0.033)	0.312 ± 0.064 (0.051)
	0.5	0.383 ± 0.060 (0.052)	0.535 ± 0.117 (0.103)
500	0.3	0.250 ± 0.027 (0.026)	0.304 ± 0.041 (0.039)
	0.5	0.377 ± 0.036 (0.035)	0.510 ± 0.066 (0.065)
2000	0.3	0.248 ± 0.014 (0.013)	0.301 ± 0.020 (0.019)
	0.5	0.374 ± 0.018 (0.018)	0.503 ± 0.031 (0.032)
$\alpha = 2.0$			
200	0.3	0.277 ± 0.046 (0.036)	0.310 ± 0.058 (0.045)
	0.5	0.434 ± 0.066 (0.061)	0.522 ± 0.096 (0.089)
500	0.3	0.271 ± 0.028 (0.027)	0.302 ± 0.034 (0.034)
	0.5	0.431 ± 0.040 (0.040)	0.503 ± 0.057 (0.056)
2000	0.3	0.271 ± 0.014 (0.014)	0.300 ± 0.017 (0.017)
	0.5	0.425 ± 0.060 (0.021)	0.502 ± 0.029 (0.028)

Note: For all models, the number of replications is 1000. L is the sequence length. The simulation model is the SR model in Table 2. Note that the mean of estimated standard errors is given in parentheses.

respectively. Note that this bias is systematic and can not be eliminated by increasing the sequence length. Interestingly, even in the case of weak rate variation (say, $\alpha = 2.0$), the estimation bias of the SR method is much larger than any other type of bias we have examined. Therefore, the SR method is sensitive to rate variation among sites and the SRV method is preferable.

To compare the relative performances of two methods, one should consider the estimation bias and the sampling variance simultaneously, because for a simpler model, the sampling variance tends to be smaller but the bias tends to be larger. The mean square error (MSE) ratio is suitable for this purpose. For a given method, the MSE is defined by $\text{MSE} = bias^2 + \text{Var}(d)$. To compare the SR method with, say, Jukes and Cantor's (JC) method (3), the SR method is better than JC if the MSE ratio, SR/JC, is less than 1, and vice versa. In Fig. 1, curves 1 and 2 are the MSE ratios SR/JC and SR/TmN, respectively, and the simulation model is SR; curve 3 is the MSE ratio SRV/SR, and the simulation model is SRV with $\alpha = 1$ [TmN = Tamura and Nei's model (5)]. SR is superior to JC when $L > 500$ (curve 1) and superior to TmN when $L > 1200$ (curve 2). The MSE ratio SRV/SR (curve 3) shows that SRV is always superior to SR if the rate varies among sites. Further simulation indicated that the performance of SR relative to JC (or TmN) increases as K increases. For example, if $K = 1.0$, SR is superior to TmN if $L \geq 800$.

DISCUSSION

It is worth mentioning that, if $c_i = 1/4$, Eq. **4** is the LogDet distance (22–24). Let matrix $\mathbf{J}$ consist of J_{ij}, and $\mathbf{F} = \text{diag}(\pi_1, \pi_2, \pi_3, \pi_4)$ be the diagonal matrix of the nucleotide frequencies. Then, we have $\mathbf{J} = \mathbf{FP}$, or $\mathbf{P} = \mathbf{F}^{-1}\mathbf{J}$. Since $z_1z_2z_3z_4 = \det(\mathbf{P}) = \det(\mathbf{F}^{-1}\mathbf{J}) = \det(\mathbf{J})/\det(\mathbf{F})$, we have

$$d = -\frac{1}{4}\ln(z_1z_2z_3z_4) = -\frac{1}{4}\ln\frac{\det(\mathbf{J})}{\det(\mathbf{F})}. \qquad [28]$$

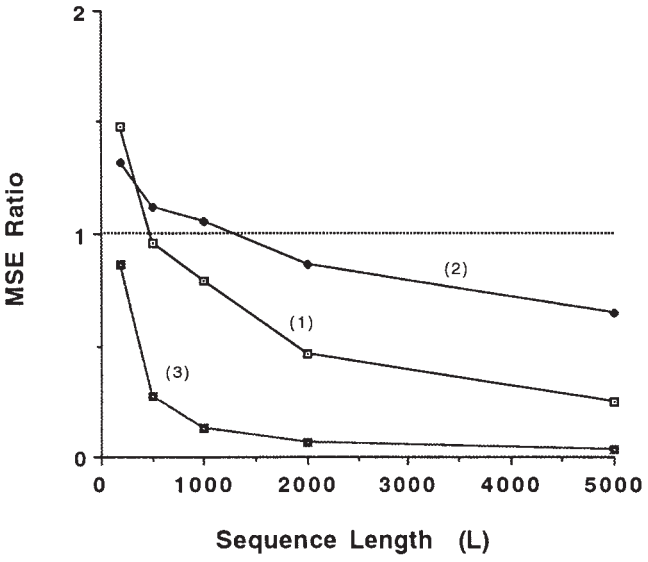

FIG. 1. Mean square error (MSE) ratios for SR/JC (curve 1), SR/TmN (curve 2), and SRV/SR (curve 3), plotted against the sequence length [JC, Jukes and Cantor's method (3); TmN, Tamura and Nei's method (5)]. The simulation model is SR for curves 1 and 2, but SRV for curve 3. The SR model is the same as that in Table 2. The SRV model is the same as the SR model except the rate varies among sites according to a gamma distribution with $\alpha = 1$. The expected number of nucleotide substitutions is $K = 0.5$. The number of replications is 1000.

The Logdet distance was recommended for phylogenetic reconstruction because its additivity holds even when the substitution model is neither stationary nor time-reversible. So, it is based on the general (12-parameter) substitution model. However, the LogDet distance is an unbiased estimate of the number of substitutions per site (K), only if the equilibrium frequencies of the four nucleotides are 1/4 and if the substitution process is stationary (10, 24). More seriously, additivity does not hold if the substitution rate varies among sites. To our knowledge, the SRV method is to date the most general method that includes the effect of rate variation among sites. Since rate variation is a common phenomenon, the SRV method may be preferable over the LogDet distance if the nucleotide frequencies do not change with time very much.

The additivity of a distance can be nullified by a large sampling variance. Therefore, the distance measure that minimizes the sampling effect may be useful in practice. Let $z_M = \text{Max}(z_1, z_2, z_3)$. Then, the minimum distance, d_m, is defined by

$$d_m = -\frac{1}{4}\ln z_M. \qquad [29]$$

The sampling variance of d_m is given by

$$\text{Var}(d_m) = \left(\frac{1}{4z_M}\right)^2 \text{Var}(z_M), \qquad [30]$$

where $\text{Var}(z_M)$ can be computed as above. The minimum distance is additive, less affected by saturation and easily computed. Note that d_m is more general than the transversion distance. Indeed, under Tamura and Nei's model (5), which is a special case of the SR model, the transversional distance is the minimum distance d_m if the rates of transversion are smaller than those of transition.

Following Kelly (25), we can assess whether the substitution process follows a general Markovian model by testing whether the smallest eigenvalue of $\hat{\mathbf{P}}$ is positive. We suggest the Z test defined by

$$Z = \frac{z_m}{\sqrt{\text{Var}(z_m)}} \qquad [31]$$

where $z_m = \text{Min}(\hat{z}_1, \hat{z}_2, \hat{z}_3)$. The procedure is the same as above (see Eqs. **17–21**) except that J_{ij} is now estimated by N_{ij}/L–i.e., the estimated matrix $\hat{\mathbf{J}}$ is not symmetrized.

Note that in the case of $z_m < 0$, the SR and SRV methods are inapplicable. If z_m is not significantly smaller than zero, the inapplicability may be regarded as sampling effects and the minimum distance (Eq. **29**) or a simpler method may be used instead. However, if z_m is significantly smaller than 0, the substitution process may not be Markovian (25).

(A computer program for the SR and SRV methods is available upon request from the authors. The program gives the estimated distance matrix and the standard errors.)

We thank the reviewers for valuable suggestions and Dr. A. Zharkikh for discussion. This study was supported by National Institutes of Health Grant GM30998.

1. Felsenstein, J. (1988) *Annu. Rev. Genet.* **22,** 521–565.
2. DeBry, R. W. (1992) *Mol. Biol. Evol.* **9,** 537–551.
3. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 21–123.
4. Hasegawa, M., Kishino, H. & Yano, T. (1985) *J. Mol. Evol.* **22,** 160–174.
5. Tamura, K. & Nei, M. (1993) *Mol. Biol. Evol.* **10,** 512–526.
6. Lanave, C., Preparata, G., Saccone, C. & Serio, G. (1984) *J. Mol. Evol.* **20,** 86–93.
7. Gillespie, J. H. (1986) *Annu. Rev. Ecol. Evol.* **17,** 637–665.
8. Tavaré, S. (1986) *Lect. Math. Life Sci.* **17,** 57–86.

9. Rodriguez, F., Oliver, J. F., Marin, A. & Medina, J. R. (1990) *J. Theor. Biol.* **142,** 485–501.
10. Zharkikh, A. (1994) *J. Mol. Evol.* **39,** 315–329.
11. Yang, Z. (1994) *J. Mol. Evol.* **39,** 105–111.
12. Uzzel, T. & Corbin, K. W. (1971) *Science* **172,** 1089–1096.
13. Nei, M., Stephens, J. C. & Saitou, N. (1985) *Mol. Biol. Evol.* **2,** 66–85.
14. Wu, C. I. & Li, W. H. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 1741–1745.
15. Bulmer, M. (1989) *Genetics* **123,** 615–619.
16. Keilson, J. (1979) *Markov Chain Models: Rarity and Exponentially* (Springer, New York).
17. Gu, X., Fu, X. Y. & Li, W. H. (1995) *Mol. Biol. Evol.* **12,** 546–557.
18. Searle, S. (1982) *Matrix Algebra Useful for Statistics* (Wiley, New York).
19. Yang, Z. (1993) *Mol. Biol. Evol.* **10,** 1396–1401.
20. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
21. Tajima, F. & Nei, M. (1984) *Mol. Biol. Evol.* **1,** 269–285.
22. Barry, D. & Hartigan, J. A. (1987) *Biometrics* **43,** 261–276.
23. Lake, J. A. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 1455–1459.
24. Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994) *Mol. Biol. Evol.* **11,** 605–612.
25. Kelly, C. (1994) *Biometrics* **50,** 653–664.