Maximum Likelihood Estimation of the Heterogeneity of Substitution Rate among Nucleotide Sites

Xun Gu, Yun-Xin Fu, and Wen-Hsiung Li
Human Genetics Center, University of Texas

This paper presents a maximum likelihood approach to estimating the variation of substitution rate among nucleotide sites. We assume that the rate varies among sites according to an invariant+gamma distribution, which has two parameters: the gamma parameter α and the proportion of invariable sites θ. Theoretical treatments on three, four, and five sequences have been conducted, and computer programs have been developed. It is shown that $\rho = (1+\theta\alpha)/(1+\alpha)$ is a good measure for the rate heterogeneity among sites. Extensive simulations show that (1) if the proportion of invariable sites is negligible, i.e., $\theta = 0$, the gamma parameter α can be satisfactorily estimated, even with three sequences; (2) if the proportion of invariable sites is not negligible, the heterogeneity ρ can still be suitably estimated with four or more sequences; and (3) the distances estimated by the proposed method are almost unbiased and are robust against violation of the assumption of the invariant+gamma distribution.

Introduction

It has been widely observed that the substitution rate varies among nucleotide sites, if the sequence under study is functionally constrained. However, many current methods in molecular evolution do not take rate heterogeneity into consideration. This negligence may have serious consequences for distance estimation, phylogenetic inference, and divergent time dating (Olsen 1987; Palumbi 1989; Jin and Nei 1990; Li et al. 1990; Hasegawa and Fujiwara 1993; Tateno et al. 1994). For these reasons, there is growing interest in this problem (e.g., Sidow et al. 1992; Van de Peer et al. 1993; Wakeley 1993; Yang 1993).

In studying rate heterogeneity, two aspects need be considered. The first one is the model of rate heterogeneity among sites. Several models have been studied. One is the multiclass model, that is, sites are classified into several classes, in each of which the substitution rate is the same (Fitch and Margoliash 1967; Shoemaker and Fitch 1989; Hasegawa et al. 1993). Obviously, the simplest form is the two-class model: one class is invariable, and the other has a constant rate. On the other hand, it has been suggested that the variation of rate among sites can be described by a gamma distribution (Uzzell and Corbin 1971; Holmquist et al. 1983; Kocher and Wilson 1991; Larson 1991; Tamura and Nei 1993; Yang 1993) or by a log-normal distribution (Olsen 1987).

The second aspect is the estimation method. The minimum substitution method uses the parsimony principle to estimate the number of substitutions at each site, assuming that the phylogenetic tree is known. If the substitution at each site is a Poisson process, the minimum number of substitutions, counted by the method of Fitch (1971), follows approximately a multiclass Poisson distribution, if the rate variation over sites follows a multiclass distribution; or it follows approximately a negative binomial distribution, if the rate variation over sites follows a gamma distribution (Johnson and Kotz 1969). After the number of minimum substitutions is inferred for each site, the heterogeneity may be estimated by, e.g., the minimum χ^2 method (Uzzell and Corbin 1971; Holmquist et al. 1983) or the method of moments (Tamura and Nei 1993).

Studies using the minimum substitution method have been conducted on various protein sequences (Uzzell and Corbin 1971; Holmquist et al. 1983), on human mitochondrial DNA sequences (Hasegawa et al. 1993; Tamura and Nei 1993), or on ribosomal DNA sequences (Larson 1991). These empirical studies have shown that the substitution rate is significantly heterogeneous among sites. However, as pointed out by Wakeley (1993), the minimum substitution estimation is always biased. Although it was supported for use in short-term evolution by simulation study (Wakeley 1993), it is unclear for

Key words: nucleotide substitution, rate variation among sites, gamma distribution, invariable sites, distances, maximum likelihood method.

Address for correspondence and reprints: Dr. Wen-Hsiung Li, Human Genetics Center, University of Texas, P.O. Box 20334, Houston, TX 77225. E-mail: gsbs005@utsph.sph.uth.tmc.edu.

Mol. Biol. Evol. 12(4):546–557. 1995.

0737-4038/95/1204-0003$02.00

long-term evolution because the bias is likely to increase with the divergent time.

In principle, the maximum likelihood (ML) method (Felsenstein 1981, 1988) can be suitably extended to the case where the substitution rate (λ) varies among sites. For example, an algorithm was proposed by Yang (1993) under Felsenstein's (1981) model of nucleotide substitution. An important property of the ML criterion is that, for a given data set, different model can be compared on a statistical basis. However, the estimation of rate heterogeneity has not been systematically investigated. Further, it is still unclear how robust the distance estimate is when the assumed distribution for the rate variation over sites is violated.

In this paper, the ML estimation of the rate heterogeneity among sites will be explored under the assumption that the rate variation follows an invariant+gamma distribution. A detailed treatment on the case of three sequences will be used to show the formulations of the estimation problem. Then we will formulate a general approach to estimating the rate heterogeneity and the distances. A computer program is developed under Hasegawa et al.'s (1985) model and Tamura and Nei's (1993) model of nucleotide substitution. Computer simulations are conducted to test the accuracy and robustness of estimates of rate heterogeneity and distance with the method developed in this study.

Model of Rate Variation among Sites

The model assumes that the substitution rate λ is a random variable with the distribution $\Phi(\lambda)$. For a given nucleotide 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 $1 - \theta$. Further, among the sites that are variable, the substitution rate λ varies according to the following gamma distribution

$$\phi(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}, \qquad (1)$$

where $\alpha = \bar{\lambda}^2/V(\lambda)$ and $\beta = \alpha/\bar{\lambda}$; $\bar{\lambda}$ and $V(\lambda)$ are the mean and the variance of the distribution $\phi(\lambda)$ (i.e., of variable sites), respectively. Therefore, $\Phi(\lambda)$, which will be called the invariant+gamma distribution, is a mix of a continuous and a discrete distribution. It can be written as

$$\Phi(\lambda) = \theta\delta_0(\lambda) + (1-\theta)\phi(\lambda), \qquad (2)$$

where $\delta_0(\lambda)$ is the delta function at 0, i.e., $\int_c \delta_0(\lambda)d\lambda = 1$ when $\lambda = 0$ is in c, otherwise $\int_c \delta_0(\lambda)d\lambda = 0$. The rationale of this model is (1) that it may be a reasonable combination of various models used by many authors; and (2) that empirical studies by the minimum substitution method have indicated that it can fit sequence data well (Uzzell and Corbin 1971; Holmquist et al. 1983; Larson 1991; Tamura and Nei 1993).

It can be shown that the mean and variance of λ over all sites under the invariant+gamma distribution are given by

$$\lambda_{\Phi} = (1-\theta)\frac{\alpha}{\beta},$$

$$V_{\Phi} = (1-\theta)\frac{\alpha}{\beta^2} + \theta(1-\theta)\left(\frac{\alpha}{\beta}\right)^2. \qquad (3)$$

The relative strength of the rate variation among sites can be generally evaluated by the coefficient of variation of $\Phi(\lambda)$, that is, $CV_{\Phi} = \sqrt{V_{\Phi}/\lambda_{\Phi}^2}$. Therefore, the heterogeneity of rate among sites can be characterized by

$$\rho = \frac{CV_{\Phi}^2}{1 + CV_{\Phi}^2} = \frac{1 + \theta\alpha}{1 + \alpha}. \qquad (4)$$

Obviously, as ρ varies from 0 to 1, the heterogeneity increases from the uniform rate over sites (ρ=0, or CV_{Φ}=0) to the maximum heterogeneity (ρ=1 or $CV_{\Phi}=\infty$).

It is noteworthy that, for the gamma distribution $\phi(\lambda)$, α is the square of the inverse of the coefficient of variation so that the larger the α, the less heterogeneous the substitution rate among sites. Generally speaking, α = 0.5, 1.0, and 2.0 may represent strong, intermediate, and weak heterogeneity among sites, respectively, and, if $\alpha = \infty$, the substitution rate is uniform among sites. In addition, if $\alpha < 1$, $\phi(\lambda)$ is unbounded near $\lambda = 0$ so that many nucleotide sites are nearly invariable. Therefore, the proportion of invariable sites, θ, the mean rate, $\bar{\lambda}$, and the gamma parameter, α, are important parameters for describing the heterogeneity of substitution rate among sites, while the parameter β is determined by $\bar{\lambda}$ and α.

The Jukes-Cantor Model with Three Sequences

Because one cannot estimate the rate heterogeneity when the number of sequences is <3, the simplest case is the Jukes-Cantor (one parameter) model with 3 sequences. This case is used to illustrate the basic approach to estimating the rate heterogeneity among sites.

The Likelihood Function

Consider three homologous DNA sequences with N nucleotides; it is assumed that the sequences have under gone no deletion or insertion or that gaps can be easily identified and excluded from analysis. Let $h(i, j, k)$ be the probability that the nucleotides at a given site in the three sequences are i, j, and k, respectively (i, j, k=A, G, T, or C). If the substitution process is time reversible, it can be shown that

$$h(i, j, k) = \sum_x \pi_x P(x, i; d_1)P(x, j; d_2)P(x, k; d_3)\,, \tag{5}$$

where π_x is the frequency of nucleotide x at the node X shared by the three sequences in an unrooted tree (see figure 1a), and d_i is the number of substitutions from node X to tip i (Felsenstein 1981). For the Jukes-Cantor model, the transition probability from nucleotide x to y with branch length d is given by

$$\begin{aligned} P(x, y; d) &= {}^1\!/_4(1-e^{-4/3d}) \quad \text{if } x \neq y\,, \\ \mathrm{P}(x, y; d) &= {}^1\!/_4 + {}^3\!/_4e^{-4/3d} \quad \text{if } x = y\,. \end{aligned} \tag{6}$$

A nucleotide configuration at a site is defined as the pattern of nucleotides at the site among the sequences under consideration. For three sequences, there are five configurations based on the nucleotide differences (Saitou 1988): (i, i, i), (i, i, j), (i, j, i), (j, i, i) and (i, j, k). The configuration (i, i, i) means the same nucleotide in the three sequences; (i, i, j) means the same nucleotide in sequences 1 and 2, but a different nucleotide in sequence 3; and so on. The probabilities of the five configurations, denoted by U_i, $i = 1, \ldots, 5$, are given by

$$\begin{aligned} U_1 &= 4h(i, i, i)\,, \\ U_2 &= 12h(i, i, j) \qquad (i{\neq}j)\,, \\ U_3 &= 12h(i, j, i) \qquad (i{\neq}j)\,, \\ U_4 &= 12h(j, i, i) \qquad (i{\neq}j)\,, \\ U_5 &= 24h(i, j, k) \qquad (i{\neq}j{\neq}k)\,, \end{aligned} \tag{7}$$

where the factor 4 in U_1 arises because i can be any of the four nucleotides; the factor 12 in U_2, U3, and U_4 arises because there are four nucleotides for i and three for j; and the factor 24 in U_5 arises because there are four nucleotides for i, three for j, and two for k.

Under the assumption that all sites are independent, Saitou (1988) showed that the likelihood can be derived in terms of the nucleotide configurations, that is,

$$f = C \prod_{i=1}^{5} U_i^{N_i}\,, \tag{8}$$

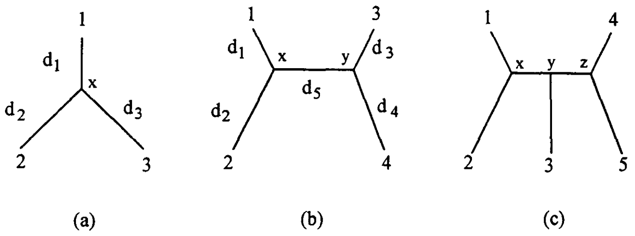

FIG. 1.—Unrooted phylogenies for three, four, and five sequences, respectively.

where $C = N!/(N_1! \ldots N_5!)$; N_i is the observed number of sites belong to the i-th configuration, and $\sum_{i=1}^{5} N_i = N$. When the heterogeneity of substitution rate among sites is taken into consideration, the joint probability density for all sites is given by

$$f = \prod_{k=1}^{N} \prod_{i=1}^{5} U_i^{\varphi_{ik}} \Phi(\lambda_k)\,, \tag{9}$$

where φ_{ik} is an indicator function whose value is 1 if site k has the i-th configuration, otherwise $\varphi_{ik} = 0$. Since the λ_k's are not observable, we need to consider the marginal likelihood of the N_i's, which is given by

$$\begin{aligned} L &= \prod_{k=1}^{N} \prod_{i=1}^{5} \int_0^\infty U_i^{\varphi_{ik}} \Phi(\lambda_k) d\lambda_k\,, \\ &= \prod_{i=1}^{5} \left(\int_0^\infty U_i \Phi(\lambda) d\lambda \right)^{\sum_{k=1}^{N} \varphi_{ik}}\,, \\ &= \prod_{i=1}^{5} \bar{U}_i^{N_i}\,, \end{aligned} \tag{10}$$

where the expected frequency of the i-th configuration, $\bar{U}_i$, is given by

$$\bar{U}_i = \int_0^\infty U_i \Phi(\lambda) d\lambda\,. \tag{11}$$

Under the Jukes-Cantor model, the equilibrium frequency of each nucleotide in node X, π_x, is assumed to be 0.25. We first put Eqs. (5) and (6) into Eq. (7) so that the explicit expressions of U_i's (i=1, . . . , 5) are obtained. For example, the probability of the first configuration is given by

$$\begin{aligned} U_1 &= 4h(i, i, i) \\ &= \sum_{x=A,T,C,G} P(x, i, d_1)P(x, i, d_2)P(x, i, d_3) \\ &= \prod_{k=1}^{3} ({}^1\!/_4 + {}^3\!/_4e^{-4/3d_k}) + 3 \prod_{k=1}^{3} ({}^1\!/_4 - {}^1\!/_4e^{-4/3d_k}) \\ &= {}^1\!/_{16} + {}^3\!/_{16}e^{-4/3(d_1+d_2)} + {}^3\!/_{16}e^{-4/3(d_2+d_3)} \\ &\quad + {}^3\!/_{16}e^{-4/3(d_1+d_3)} + {}^3\!/_8e^{-4/3(d_1+d_2+d_3)} \end{aligned} \tag{12}$$

Then the expected frequency $\bar{U}_i$ is obtained by integrating U_1 according to Eq. (11), i.e., $\bar{U}_i = \int_0^\infty U_i \Phi(\lambda) d\lambda$, which are given by

$$\begin{aligned}
\bar{U}_1 &= {}^1/_{16} + {}^3/_{16}E_{12} + {}^3/_{16}E_{23} + {}^3/_{16}E_{13} + {}^3/_8 E_{123}\,,\\
\bar{U}_2 &= {}^3/_{16} + {}^9/_{16}E_{12} - {}^3/_{16}E_{23} - {}^3/_{16}E_{13} - {}^3/_8 E_{123}\,,\\
\bar{U}_3 &= {}^3/_{16} - {}^3/_{16}E_{12} - {}^3/_{16}E_{23} + {}^9/_{16}E_{13} - {}^3/_8 E_{123}\,,\\
\bar{U}_4 &= {}^3/_{16} - {}^3/_{16}E_{12} + {}^9/_{16}E_{23} - {}^3/_{16}E_{13} - {}^3/_8 E_{123}\,,\\
\bar{U}_5 &= {}^6/_{16} - {}^6/_{16}E_{12} - {}^6/_{16}E_{23} - {}^6/_{16}E_{13} + {}^6/_8 E_{123}\,,
\end{aligned} \tag{13}$$

where

$$\begin{aligned}
E_{ij} &= \int_0^\infty e^{-4/3(d_i+d_j)} \Phi(\lambda) d\lambda\\
&= \theta + (1-\theta)\left\{1 + \frac{4}{3}\frac{d_{ij}}{\alpha}\right\}^{-\alpha},
\end{aligned} \tag{14}$$

$$\begin{aligned}
E_{123} &= \int_0^\infty e^{-4/3(d_1+d_2+d_3)} \Phi(\lambda) d\lambda\\
&= \theta + (1-\theta)\left\{1 + \frac{4}{3}\frac{d_{123}}{\alpha}\right\}^{-\alpha},
\end{aligned}$$

in which $i, j = 1, 2$, or 3 $(i \neq j)$ refer to sequences 1, 2, and 3, respectively, $d_{ij} = d_i + d_j$ and $d_{123} = d_1 + d_2 + d_3$; the branch length d_i is given by the expected number of substitutions over variable sites, $d_i = t_i \alpha / \beta$.

The *ML* Estimates

The maximization of the likelihood function L is achieved if

$$\bar{U}_i = \frac{N_i}{N}, \qquad i = 1, \ldots, 5\,. \tag{15}$$

Therefore, the number of degrees of freedom is four, because $\sum_{i=1}^5 \bar{U}_i = 1$. Because three branch lengths need be estimated, there is only one degree of freedom left for the estimation of the rate heterogeneity. That is, in the Jukes-Cantor model with three sequences, the parameter α or θ in the invariant+gamma distribution of substitution rate can be estimated only if the other is known or assumed.

Substituting Eq. (15) into Eq. (13), we obtain the *ML* estimates of branch lengths d_i:

$$\begin{aligned}
\hat{d}_1 &= {}^1\!/_2(\hat{d}_{12}+\hat{d}_{13}-\hat{d}_{23})\,,\\
\hat{d}_2 &= {}^1\!/_2(\hat{d}_{12}+\hat{d}_{23}-\hat{d}_{13})\,,\\
\hat{d}_3 &= {}^1\!/_2(\hat{d}_{13}+\hat{d}_{23}-\hat{d}_{12})\,,
\end{aligned} \tag{16}$$

where

$$\hat{d}_{ij} = \frac{3}{4}\hat{\alpha}\left\{\left(1 - \frac{4}{3}\frac{p_{ij}}{1-\hat{\theta}}\right)^{-1/\hat{\alpha}} - 1\right\}, \tag{17}$$

in which p_{ij} is the proportion of differences between sequences i and j. The *ML* estimates of α and θ should satisfy the following equation

$$\begin{aligned}
2\left\{1 - \frac{4}{3(1-\hat{\theta})}\left(\frac{p}{2} - \frac{N_5}{N}\right)\right\}^{-1/\hat{\alpha}} &+ 1\\
= \sum_{i<j}^{3}\left(1 - \frac{4}{3}\frac{p_{ij}}{1-\hat{\theta}}\right)^{-1/\hat{\alpha}}&.
\end{aligned} \tag{18}$$

where $p = p_{12} + p_{13} + p_{23}$.

In particular, if λ follows a gamma distribution, i.e., $\theta = 0$, the *ML* estimate of α is the nonnegative solution of the following equation

$$\begin{aligned}
2\left\{1 - \frac{4}{3}\left(\frac{p}{2} - \frac{N_5}{N}\right)\right\}^{-1/\hat{\alpha}} &+ 1\\
= \sum_{i<j}^{3}\left(1 - \frac{4}{3}p_{ij}\right)^{-1/\hat{\alpha}}&.
\end{aligned} \tag{19}$$

On the other hand, if λ follows the two-class model, in which one class is invariable and the other has a constant rate, i.e., $\alpha = \infty$ in the invariant+gamma distribution, the *ML* estimate of θ is given by

$$\hat{\theta} = 1 - \frac{q}{3f_5}\left(1 - \sqrt{1 - \frac{p_{12}p_{13}p_{23}f_5}{q^2}}\right), \tag{20}$$

where q and f_5 are given by

$$\begin{aligned}
f_5 &= \frac{N_5}{N}\\
q &= \left(f_5 - \frac{p}{2}\right)^2 - (p_{12}p_{13} + p_{12}p_{23} + p_{13}p_{23})\,.
\end{aligned} \tag{21}$$

The Likelihood Function under the Hasegawa-Kishino-Yano Model

The Matrix of Transition Probability

Hasegawa et al.'s (1985) model (table 1) assumes that the process of nucleotide substitution is a stationary Markov chain, in which nucleotide i is replaced by nu-

Table 1
The Substitution Pattern of Hasegawa et al.'s (1985) Model

	To			
From	A	T	G	C
A	...	$\pi_T v$	$\pi_G s$	$\pi_C v$
T	$\pi_A v$	...	$\pi_G v$	$\pi_C s$
G	$\pi_A s$	$\pi_T v$	...	$\pi_C v$
C	$\pi_A v$	$\pi_T s$	$\pi_G v$	...

cleotide j in an infinitesmally short time interval, dt, with a probability of

$$\begin{aligned} P(i, j; dt) &= s\pi_j dt\,, \quad \text{for transition,} \\ P(i, j; dt) &= v\pi_j dt\,, \quad \text{for transversion,} \end{aligned} \tag{22}$$

where π_A, π_G, π_T, and π_C are the equilibrium frequencies of nucleotides A, G, T, and C; and s and v are the transitional and transversional rates, respectively.

Hasegawa et al.'s model takes into consideration the transition-transversion bias and the nucleotide composition of the sequence. In the case of $s = v$, it is equivalent to the model of Felsenstein (1981). If the frequencies at equilibrium are assumed to be equal for all nucleotides, i.e., $\pi_A = \pi_G = \pi_T = \pi_C = 0.25$, the model reduces to Kimura's (1980) two-parameter model, which reduces further to the one-parameter model (Jukes and Cantor 1969) if $s = v$. The three-parameter model proposed by Tamura (1992), which extends Kimura's two-parameter model to the case where the GC content is not 50%, is also a special case of Hasegawa et al.'s model with $\pi_A = \pi_T$ and $\pi_G = \pi_C$. Moreover, Hasegawa et al.'s model is very similar to that used in Felsenstein's *DNAML* program (Felsenstein 1991).

Let $P(i, j; t)$ be the probability of changing from nucleotide i at time 0 to nucleotide j at time t, and let $\mathbf{P}(t)$ be the matrix whose ij-th element is $P(i, j; t)$. Let r_{ij} $(i \neq j)$ be the rate of change from nucleotide i to nucleotide j as defined in table 1, and $r_{ii} = -\sum_{j \neq i} r_{ij}$ (i.e., $\sum_j r_{ij} = 0$). Let $\mathbf{R}$ be the rate matrix consisting of r_{ij}'s. In general, we can write

$$\mathbf{P}(t) = \mathbf{e}^{t\mathbf{R}}\,. \tag{23}$$

The transition probability $P(i, j; t)$ has been explicitly derived by Hasegawa et al. (1985). For convenience, let $-Z_k$ $(k=1, \ldots, 4)$ be the eigenvalues of $\mathbf{R}$; they are given by $Z_1 = 0$, $Z_2 = v$, $Z_3 = v\pi_Y + s\pi_R$ and $Z_4 = v\pi_R + s\pi_Y$, respectively, where $\pi_R = \pi_A + \pi_G$ is the frequency of purines, and $\pi_Y = \pi_T + \pi_C$ is the frequency of pyrimidines. Then,

$$P(i, j; t) = \sum_{k=1}^{4} G(i, j, k)e^{-Z_k t}\,, \tag{24}$$

where the coefficients $G(i, j, k)$ are given in table 2.

For n aligned homologous *DNA* sequences without gaps, there are 4^n possible configurations at a site. Number these configurations from 1 to 4^n, and let U_i and N_i be the probability and the observed number of the i-th configuration, respectively. Under the assumption that the heterogeneity of substitution rate among sites is independent of the pattern of nucleotide substitution, the likelihood function, which is the marginal density of the joint distribution over all sites, is given by

$$L = \prod_{i=1}^{4^n} \bar{U}_i^{N_i}\,, \tag{25}$$

where $\bar{U}_i = \int_0^\infty U_i \Phi(\lambda) d\lambda$ is the expected frequency of the i-th configuration. It is noteworthy that the assumption of a molecular clock is not necessary in our model.

The Likelihood Function for Three Sequences

To facilitate computation, the expected frequencies of configurations should be given explicitly. First, let us consider the case of three sequences, as shown in figure 1*a*. To account for unequal rates among branchs, let $t_k^* = c_k t_k$ $(k=1, 2, 3)$ be the effective divergent time of lineage k, where t_k is the absolute divergent time, and c_k is the constant specific to lineage k. Then, the probability $U_m(\mathbf{b})$ of having the m-th nucleotide configuration $\mathbf{b} = (b_1, b_2, b_3)$ is given by

$$U_m(\mathbf{b}) = \sum_x \pi_x P(x, b_1; t_1^*) P(x, b_2; t_2^*) P(x, b_3; t_3^*)\,,$$
$$m = 1, \ldots, 64\,. \tag{26}$$

Table 2
The Coefficients *G*(*i*, *j*, *k*) of Eq. (23)

	$G(i, j, 1)$	$G(i, j, 2)$	$G(i, j, 3)$	$G(i, j, 4)$
$i = j = A, G$	π_j	$\pi_j\left(1 - \frac{1}{\pi_R}\right)$	$1 - \frac{\pi_j}{\pi_R}$	0
$i = j = T, C$	π_j	$\pi_j\left(1 - \frac{1}{\pi_Y}\right)$	0	$1 - \frac{\pi_j}{\pi_Y}$
$i \neq j$ $(A \leftrightarrow G)$	π_j	$\pi_j\left(1 - \frac{1}{\pi_R}\right)$	$-\frac{\pi_j}{\pi_R}$	0
$i \neq j$ $(T \leftrightarrow C)$	π_j	$\pi_j\left(1 - \frac{1}{\pi_Y}\right)$	0	$-\frac{\pi_j}{\pi_Y}$
$i \neq j$ (tranv)[a]	π_j	$-\pi_j$	0	0

[a] tranv = transversion.

Let D_{ijk} and E_{ijk} be

$$\begin{aligned} D_{ijk} &= Z_i t_1^* + Z_j t_2^* + Z_k t_3^* , \\ \mathrm{E}_{ijk} &= \int_0^\infty e^{-D_{ijk}} \Phi(\lambda) d\lambda \\ &= (1-\theta)\left(1+\frac{D_{ijk}}{\alpha}\right)^{-\alpha} + \theta , \end{aligned} \tag{27}$$

where $i, j, k = 1, 2, 3$, or 4 refer to the four eigenvalues of the rate matrix **R**. Substituting Eq. (23) into Eq. (25), one can show that

$$\bar{U}_m(\mathbf{b}) = \sum_{i=1}^4 \sum_{j=1}^4 \sum_{k=1}^4 A_{ijk}(\mathbf{b}) E_{ijk} , \tag{28}$$

where the coefficient $A_{ijk}(\mathbf{b})$ is determined by the coefficients $G(x, y, z)$ in table 2, i.e.,

$$A_{ijk}(\mathbf{b}) = \sum_{x=1}^4 \pi_x G(x, b_1, i) G(x, b_2, j) G(x, b_3, k) . \tag{29}$$

The Likelihood Function for n Sequences

There are $l = 2n - 3$ branches for an unrooted bifurcating tree with n nucleotide sequences. Similar to the case of three sequences, let $t_k^* = c_k t_k$ (k=1, . . . , l) be the effective divergent time of lineage k. Let $\mathbf{b} = (b_1, \ldots, b_n)$ be a nucleotide configuration, numbered by m. To facilitate the derivation of the expected frequency of the m-th configuration $\bar{U}_m(\mathbf{b})$, let $D(i_1, \ldots, i_l)$ be

$$D(i_1, \ldots, i_l) = \sum_{k=1}^{l} Z_{i_k} t_k^* , \tag{30}$$

where for lineage k (k=1, . . . , l), $-Z_{i_k}$ (i_k=1, 2, 3, or 4) is the i_k-th eigenvalue of the rate matrix **R**. Thus in total, we have 4^l different combinations of $D(i_1, \ldots, i_l)$. For example, if $n = 4$ with $l = 2n - 3 = 5$, there are 4^5 = 1,024 combinations: $D(1, 1, 1, 1, 1) = Z_1 \times \Sigma_{k=1}^5 t_k^* = 0$; $D(1, 1, 1, 1, 2) = Z_1 \times \Sigma_{k=1}^4 t_k^* + Z_2 t_5^* = Z_2 t_5^*$; and so on. Further, let $E(i_1, \ldots, i_l)$ be

$$\begin{aligned} E(i_1, \ldots, i_l) &= \int_0^\infty e^{-D(i_1, \ldots, i_l)} \Phi(\lambda) d\lambda \\ &= (1-\theta)\left\{1+\frac{D(i_1, \ldots, i_l)}{\alpha}\right\}^{-\alpha} + \theta . \end{aligned} \tag{31}$$

Then it can be shown that $\bar{U}_m(\mathbf{b})$ is given by

$$\bar{U}_m(\mathbf{b}) = \sum_{i_1=1}^4 \ldots \sum_{i_l=1}^4 A(i_1, \ldots, i_l; \mathbf{b}) E(i_1, \ldots, i_l) , \tag{32}$$

where the coefficient $A(i_1, \ldots, i_l; \mathbf{b})$ depends on the given phylogenetic tree. For the four sequences in figure 1b, we have $l = 5$ and

$$\begin{aligned} &A(i_1, \ldots, i_l; \mathbf{b}) \\ &= \sum_{x=1}^4 \sum_{y=1}^4 \pi_x G(x, b_1, i_1) G(x, b_2, i_2) G(x, y, i_5) \\ &\qquad \times G(y, b_3, i_3) G(y, b_4, i_4) , \end{aligned} \tag{33}$$

and for the five sequences in figure 1c, we have $l = 7$ and

$$\begin{aligned} &A(i_1, \ldots, i_l; \mathbf{b}) \\ &= \sum_{x=1}^4 \sum_{y=1}^4 \sum_{z=1}^4 \pi_x G(x, b_1, i_1) G(x, b_2, i_2) \\ &\quad \times G(x, y, i_6) G(y, b_3, i_3) G(y, z, i_7) \\ &\qquad \times G(z, b_4, i_4) G(z, b_5, i_5) . \end{aligned} \tag{34}$$

The Newton-Raphson Algorithm

The maximum likelihood estimates of α, θ, and branch lengths can be obtained by the Newton-Raphson iteration method. In this algorithm, we set the equilibrium frequency of each nucleotide, π_k ($k = A, G, T$, or C), estimated from the sequences. Thus, under Hasegawa et al.'s (1985) model of substitution, there are two unknown parameters needed to be estimated: the branch length and the rate ratio of transition to transversion. Also one can assume the ratio s/v to be the same in all branches so that the number of unknown parameters can be reduced greatly.

Let **x** be the parameter vector with m elements. The iteration equation is then given by

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \mathbf{H}^{-1}\mathbf{g} , \tag{35}$$

where **H** is the $m \times m$ Hessian matrix and **g** is the gradient vector, which are defined by

$$\begin{aligned} H_{ij} &= \frac{\partial^2 \ln L}{\partial x_i \partial x_j} , \qquad i, j = 1, \ldots, m , \\ g_i &= \frac{\partial \ln L}{\partial x_i} , \qquad i = 1, \ldots, m . \end{aligned} \tag{36}$$

Table 3
Simulation Results for Three Sequences[a]

N	$\hat{\alpha}$	$\hat{\rho}$	$\hat{d}_1$[b]	$\hat{d}_{jc}$[c]
α = 0.5 (ρ=0.667):				
200	1.00 ± 5.10	0.64 ± 0.18	0.25 ± 0.35	0.14 ± 0.03
500	0.59 ± 0.50	0.66 ± 0.11	0.21 ± 0.07	0.14 ± 0.02
1,000	0.53 ± 0.19	0.66 ± 0.07	0.21 ± 0.04	0.14 ± 0.01
2,000	0.51 ± 0.12	0.67 ± 0.05	0.20 ± 0.03	0.14 ± 0.01
5,000	0.50 ± 0.07	0.67 ± 0.03	0.20 ± 0.02	0.14 ± 0.01
α = 1.0 (ρ=0.500):				
200	2.75 ± 21.50	0.46 ± 0.22	0.22 ± 0.11	0.16 ± 0.04
500	1.43 ± 2.98	0.48 ± 0.14	0.21 ± 0.04	0.16 ± 0.02
1,000	1.17 ± 0.76	0.49 ± 0.10	0.20 ± 0.03	0.16 ± 0.02
2,000	1.06 ± 0.32	0.50 ± 0.07	0.20 ± 0.02	0.16 ± 0.01
5,000	1.02 ± 0.17	0.50 ± 0.07	0.20 ± 0.01	0.16 ± 0.01
α = 2.0 (ρ=0.333):				
200	4.64 ± 29.80	0.32 ± 0.22	0.21 ± 0.08	0.18 ± 0.04
500	4.38 ± 22.30	0.31 ± 0.16	0.20 ± 0.04	0.18 ± 0.02
1,000	3.00 ± 7.00	0.33 ± 0.12	0.20 ± 0.03	0.18 ± 0.02
2,000	2.43 ± 3.64	0.33 ± 0.09	0.20 ± 0.02	0.18 ± 0.01
5,000	2.11 ± 0.57	0.33 ± 0.05	0.20 ± 0.01	0.18 ± 0.01

[a] The substitution model used was the Jukes-Cantor model, and the rate variation among sites was generated by a gamma distribution. N is the sequence length, and the branch lengths are $d_1 = d_2 = d_3 = 0.2$.

[b] The length of branch 1 estimated by the proposed method.

[c] The length of branch 1 estimated by the Jukes-Cantor method under the assumption of a uniform rate among sites.

The analytical expressions for **H** and **g** are obtained so that their calculations are not very difficult, though time consuming. Given the appropriate initial values for parameters, $\mathbf{x}_n$ would converge to $\hat{\mathbf{x}}$, the maximum likelihood estimates of **x**, as $n \rightarrow \infty$. Finally, their large sample variance-covariance matrix, **V**, can be approximately estimated by the inverse of Fisher's information matrix, **I**, that is,

$$\mathbf{V}(\mathbf{x}) = \mathbf{I}^{-1}(\mathbf{x}) . \qquad (37)$$

Computer Simulation

To test the performance of the proposed estimation method, a series of simulation studies have been conducted. In particular, we are interested in the estimation of distances between sequences when the substitution rate varies among sites, because, for example, accurate estimation of distances is critical to the neighbor-joining method and other distance matrix methods for tree inference (Tateno et al. 1994). We have developed a computer program for estimating the rate heterogeneity over sites, based on the Newton-Raphson algorithm described above. Large α and low θ (i.e., low heterogeneity) are suggested for their initial values, e.g., α_0 = 3.0 and θ_0 = 0.05 were used in the following simulations. The convergence is generally fast, except when the heterogeneity ρ is high and sequence length N is short. Fortunately, only few replications failed to converge and so they were treated as inapplicable cases. For simplicity, in our simulation model, we assume $\pi_A = \pi_T$ and $\pi_G = \pi_C$ so that Hasegawa et al.'s (1985) model can be characterized by the following two parameters: the $GC\%$ and the ratio of transitional to transversional rate s/v. We have used only three or four sequences, because the simulation time increases enormously with the number of sequences.

Three Sequences

In the case of three sequences, we assumed $\theta = 0$ and used the gamma distribution to simulate the rate variation among sites. We set the gamma distribution parameter α to be 0.5, 1.0, or 2.0, which represent, respectively, strong, intermediate, and weak variation of rate over sites. We set the sequence length N to be 200, 500, 1,000, 2,000, and 5,000 nucleotides. The simulations were conducted for 1,000 replicates for each case. For each replicate, the estimation was based on Hasegawa et al.'s (1985) model of nucleotide substitution and the gamma distribution for the rate heterogeneity.

In the case where the substitution scheme follows the Jukes-Cantor model, table 3 shows that the estimates of the gamma distribution parameter α and the heterogeneity ρ are asymptotically unbiased. When the sequence length was short (say, $N \leqslant 500$), or the heterogeneity was weak (say, α=2), the estimate of α was subject to a large bias and a large sample variance be-

Table 4
Estimate of ρ with Three Sequences under Various Substitution Models[a]

s/v	GC (%)	ρ = 0.333 (α=2.0)	ρ = 0.500 (α=1.0)	ρ = 0.667 (α=0.5)
1 . . .	50	0.355 ± 0.144 (0.139)[b]	0.511 ± 0.127 (0.127)	0.671 ± 0.091 (0.092)
3 . . .	60	0.320 ± 0.130 (0.137)	0.489 ± 0.126 (0.132)	0.666 ± 0.094 (0.089)
8 . . .	40	0.344 ± 0.212 (0.235)	0.484 ± 0.158 (0.160)	0.649 ± 0.133 (0.115)

[a] The rate variation among sites is generated by a gamma distribution and the sequence length N = 1,000. The branch lengths are $d_1 = d_2 = 0.2$, and $d_3 = 0.4$.

[b] The values in parentheses are the asymptotic standard errors.

cause in a few replicates the estimate of α was very large. In contrast, the estimation of ρ was little affected, because ρ becomes close to 0 if α is very large. Therefore, ρ is a better measure than α for the rate heterogeneity among sites. Further, as shown in table 4, estimation of ρ seems robust against various s/v biases and GC content variation.

For distances, the simulation shows that the estimation of distances is almost unbiased, except when the sequence length is very short (N=200) and the heterogeneity is strong (α=0.5); in this case the distance may be overestimated (table 3). For comparison, $\hat{d}_{jc}$, the distance estimated by the Jukes-Cantor method under the assumption of a uniform rate over sites, is also presented in table 3. As expected, in all cases the distance is consistently underestimated by $\hat{d}_{jc}$.

Four Sequences

For four sequences, we first consider the case where the heterogeneity is simulated by a gamma distribution, i.e., θ = 0. We set the gamma distribution parameter α to be 0.5, 1.0, or 2.0; the substitution model was s/v = 1, and GC = 50% or s/v = 3, and GC = 60%; and the sequence length N was 2,000. The number of replications was 300. Similar to the case of three sequences, the estimation was based on Hasegawa et al.'s (1985) model and the gamma distribution for the rate heterogeneity. As summarized in table 5, all estimates, including the gamma distribution parameter α, the heterogeneity ρ, and the distances d_i, are almost unbiased for all cases. Combining these and the simulation results for three sequences, we may conclude that, if the variation of rate over sites does follow a gamma distribution and

Table 5
Simulation Results for Four Sequences, I[a]

	α = 2.0 (ρ=0.333)	α = 1.0 (ρ=0.500)	α = 0.5 (ρ=0.667)
s/v = 1, GC = 50%:			
$\hat{\alpha}$	1.985 ± 0.410 (0.410)[b]	1.012 ± 0.150 (0.146)	0.497 ± 0.058 (0.058)
$\hat{\rho}$	0.335 ± 0.046 (0.046)	0.497 ± 0.037 (0.036)	0.668 ± 0.026 (0.026)
$\hat{d}_1$ (0.30)[c]	0.300 ± 0.027 (0.026)	0.298 ± 0.029 (0.030)	0.304 ± 0.036 (0.037)
$\hat{d}_2$ (0.15)	0.152 ± 0.018 (0.018)	0.153 ± 0.020 (0.020)	0.150 ± 0.025 (0.025)
$\hat{d}_3$ (0.30)	0.301 ± 0.025 (0.026)	0.299 ± 0.030 (0.030)	0.303 ± 0.037 (0.037)
$\hat{d}_4$ (0.15)	0.148 ± 0.017 (0.018)	0.151 ± 0.021 (0.020)	0.147 ± 0.025 (0.024)
$\hat{d}_5$ (0.10)	0.102 ± 0.018 (0.018)	0.101 ± 0.023 (0.020)	0.103 ± 0.025 (0.025)
s/v = 3, GC = 60%:			
$\hat{\alpha}$	2.040 ± 0.453 (0.453)	1.004 ± 0.149 (0.153)	0.502 ± 0.059 (0.063)
$\hat{\rho}$	0.329 ± 0.049 (0.049)	0.499 ± 0.037 (0.038)	0.666 ± 0.026 (0.028)
$\hat{d}_1$ (0.30)	0.301 ± 0.022 (0.023)	0.303 ± 0.037 (0.038)	0.303 ± 0.031 (0.033)
$\hat{d}_2$ (0.15)	0.152 ± 0.016 (0.015)	0.150 ± 0.017 (0.017)	0.150 ± 0.021 (0.021)
$\hat{d}_3$ (0.30)	0.302 ± 0.023 (0.023)	0.303 ± 0.027 (0.027)	0.303 ± 0.033 (0.033)
$\hat{d}_4$ (0.15)	0.152 ± 0.015 (0.015)	0.150 ± 0.019 (0.017)	0.146 ± 0.020 (0.020)
$\hat{d}_5$ (0.10)	0.098 ± 0.016 (0.015)	0.100 ± 0.016 (0.017)	0.103 ± 0.022 (0.021)

[a] The rate variation among sites is simulated by a gamma distribution, and the sequence length N = 2,000.

[b] The asymptotic standard errors are in parentheses.

[c] The branch lengths in the simulation model are given in the parentheses in the first column.

Table 6
Simulation Results for Four Sequences, II[a]

	α = 1.0, θ = 0.1 (ρ=0.550)	α = 1.0, θ = 0.2 (ρ=0.600)
$\hat{\alpha}$	1.164 ± 0.745 (0.867)[b]	1.137 ± 0.881 (0.991)
$\hat{\theta}$	0.112 ± 0.122 (0.170)	0.181 ± 0.159 (0.194)
$\hat{\rho}$	0.541 ± 0.066 (0.070)	0.594 ± 0.068 (0.072)
$\hat{d}_1$ (0.30)[c]	0.307 ± 0.047 (0.056)	0.305 ± 0.059 (0.065)
$\hat{d}_2$ (0.15)	0.155 ± 0.029 (0.032)	0.151 ± 0.030 (0.036)
$\hat{d}_3$ (0.30)	0.311 ± 0.047 (0.057)	0.303 ± 0.053 (0.065)
$\hat{d}_4$ (0.15)	0.152 ± 0.029 (0.031)	0.153 ± 0.033 (0.037)
$\hat{d}_5$ (0.10)	0.103 ± 0.022 (0.025)	0.104 ± 0.029 (0.028)

[a] The rate variation among sites is simulated by an invariant + gamma distribution and the substitution model is GC = 60%, and s/v = 3. The sequence length N = 2,000.

[b] The asymptotic standard errors are in parentheses.

[c] The branch lengths in the simulation model are given in the parentheses in the first column.

if the sequence length is fairly long, then even three or four sequences can give satisfactory estimates.

Second, we consider the case where invariable sites are included, i.e., $\theta \neq 0$. The variation of rate among sites was generated by an invariant+gamma distribution. We set the gamma distribution parameter α to be 1.0; the substitution model was s/v = 3 and GC = 60%; and the proportion of invariable sites was 0.1 or 0.2. The number of nucleotides N was 2,000, and the number of replicates was 300. The model we used for estimation was Hasegawa et al.'s (1985) model and the invariant+gamma distribution for the rate heterogeneity. As summarized in table 6, (1) the estimates of distances are almost unbiased; (2) the heterogeneity ρ, the gamma parameter α, and the proportion of invariable sites θ are slightly biased; and (3) the estimates of α and θ are subject to large sample variances, but the sample variance of ρ is small.

The Robustness of Distance Estimation

For practical reasons, it is important to see how robust our estimation of distance is, if the assumption of an invariant+gamma distribution is violated. We have therefore conducted a simulation in which the rate variation over sites was simulated by an invariant+lognormal distribution, while the model for estimation was assumed to be an invariant+gamma distribution. The lognormal distribution used for simulation was the same as that used by Olsen (1987) and Jin and Nei (1990), which is quite skewed. We set the proportion of the invariable sites θ to be 0 or 0.1; the substitution pattern was s/v = 3, and GC = 60%; and the sequence length was N = 2,000. Three hundred replicates were conducted, and the results are presented in table 7. In both cases, the distances, which are measured by the number of substitutions per variable sites, are only slightly underestimated. Therefore, it seems that the distance estimated by the proposed method is robust against violation of the invariant+gamma distribution.

The Sample Variances of Estimates

The sample variances of estimates can be calculated approximately by the inverse of the information matrix. To test the accuracy of the asymptotic sample variances, they are compared to the observed sample variances. As presented in the parentheses of tables 3–6, the asymptotic standard errors (i.e., the square root of sample variances computed by using the information matrix) are quite close to the observed values in most cases.

An Example: The Rate Variation among Sites in SSU rRNAs

To provide an example of the estimation methods discussed above, we used four aligned sequences of small-subunit rRNAs (data from M. Gouy). These four sequences are *Dictyostelium* (D), a slime mold; *Crithidia* (C), a flagellate; yeast (Y); and human (H). There are a total of 1,734 aligned sites without gaps. Table 8 summarizes the ML estimates of the gamma distribution parameter α, the proportion of invariant sites θ, and the branch lengths when the rate variation among sites follows an invariant+gamma distribution. For comparisons, they are also estimated when the rate variation follows a gamma distribution (i.e., θ=0), and when the uniform rate among sites is assumed (i.e., θ=0 and $\alpha=\infty$). These estimates are based on the Hasegawa et al.'s (1985) model of substitution and the true phylogeny ((D, C), (Y, H)).

The statistical significance of the rate variation among sites can be tested by the likelihood ratio test.

The null hypothesis H_0 is the uniformity of substitution rate among sites, i.e., $\alpha = \infty$ and $\theta = 0$ so that $\rho = 0$; the alternative hypothesis H_a is $\rho > 0$. Let $\hat{L}_0$ and $\hat{L}$ be the maximum likelihood values under H_0 and H_a, respectively. Then it is known that the likelihood ratio statistic δ_h

$$\delta_h = 2(\ln \hat{L} - \ln \hat{L}_0) , \qquad (38)$$

follows asymptotically the χ^2 distribution; the degree of freedom (df) depends on the parameter number of rate heterogeneity among sites. Under the null hypothesis, we have $\ln \hat{L}_0 = -6{,}696.2$. In the case of the invariant+gamma distribution for the rate variation among sites, we have $\ln \hat{L} = -6{,}609.0$ so that $\delta_h = 2(-6{,}609.0+6{,}696.2) = 174.4$. The probability that a χ^2 random variable with df = 2 will exceed this value is $P < 10^{-4}$. This is strong evidence for rate variation among the sites of SSU rRNA. The same test can be applied for the case when a gamma distribution is assumed for the rate variation among sites except that df = 1, which gives $\delta_h = 173.2$ ($P<10^{-4}$).

Further, we can construct a likelihood ratio statistic to test whether θ is significantly larger than zero. Let δ_i be

$$\delta_i = 2(\ln \hat{L} - \ln \hat{L}_{\theta=0}) , \qquad (39)$$

where $\hat{L}_{\theta=0}$ is the maximum likelihood value when a gamma distribution is assumed for rate variation (i.e., $\theta=0$). It is known that the statistic δ_i follows asymptotically the χ^2 distribution with df = 1. Applying this test for the SSU rRNA data, we have $\delta_i = 2(-6{,}609.0+6609.6) = 1.2$ ($0.2<P<0.3$). Therefore, the invariant+gamma distribution fits the SSU rRNA data slightly better than does the gamma distribution.

Consistent with our simulation results, the branch lengths are seriously underestimated under the assumption of uniform rate among sites. The branch lengths under the gamma distribution (Γ-length) are slightly larger (~10%) than that under the invariant+gamma distribution ($I + \Gamma$-length).

Discussion

In this paper, we have studied (1) how to estimate the variation in substitution rate among sites and (2) how to obtain unbiased estimates of the distances between sequences. To describe the rate variation among sites, an invariant+gamma distribution is assumed, which has two parameters, the gamma parameter α and the proportion of invariable sites θ. We have conducted an extensive simulation study under Hasegawa et al.'s (1985) model. For the estimation of rate heterogeneity, the results can be summarized as follows: (1) to describe the rate variation among sites, $\rho = (1+\theta\alpha)/(1+\alpha)$ is a better measure of heterogeneity than α, particularly when the sequence length is short and the variation among sites is weak; (2) if the proportion of invariable sites is negligible, the gamma parameter α and the heterogeneity ρ can be satisfactorily estimated, even with three or four sequences; (3) in the presence of invariable sites, the heterogeneity ρ can still be suitably estimated with four sequences, though the estimates of α and θ are subject to large sample variances. Moreover, our simulation results showed that when the rate varies among sites, the proposed method gives an unbiased estimate of the distance. Indeed, the estimate seems robust against violation of the assumption of the invariant+gamma distribution. In our simulation we have not considered the case of five sequences, because of time limitation, but the

Table 7
Simulation Results with Four Sequences to Show the Robustness of the Distance Estimates when the Rate at Variable Sites Does Not Follow a Gamma Distribution[a]

	$\theta = 0.0$	$\theta = 0.1$
$\hat{\alpha}$	1.137 ± 0.429 (0.575)[b]	1.375 ± 0.710 (0.854)
$\hat{\theta}$	0.027 ± 0.060 (0.131)	0.062 ± 0.085 (0.133)
$\hat{\rho}$	0.487 ± 0.055 (0.057)	0.467 ± 0.060 (0.066)
$\hat{d}_1$ (0.30)[c]	0.281 ± 0.030 (0.040)	0.293 ± 0.033 (0.043)
$\hat{d}_2$ (0.15)	0.141 ± 0.017 (0.024)	0.150 ± 0.022 (0.026)
$\hat{d}_3$ (0.30)	0.278 ± 0.030 (0.040)	0.294 ± 0.037 (0.043)
$\hat{d}_4$ (0.15)	0.145 ± 0.017 (0.024)	0.147 ± 0.021 (0.025)
$\hat{d}_5$ (0.10)	0.091 ± 0.016 (0.019)	0.096 ± 0.017 (0.020)

[a] The rate variation among sites is simulated by an invariant + lognormal distribution, and the substitution model is GC = 60%, and s/v = 3. The sequence length N = 2,000.

[b] The asymptotic standard errors are in parentheses.

[c] The branch lengths in the simulation model are given in the parentheses in the first column.

Table 8
The ML Estimates for SSU rRNA Sequences[a]

Model[b]	$I + \Gamma$	Γ	U
$\hat{\alpha}$	1.381 ± 1.270	0.508 ± 0.067	∞
$\hat{\theta}$	0.263 ± 0.133	0	0
$\hat{\rho}$	0.572 ± 0.092	0.663 ± 0.029	0
$\hat{d}_D$	0.700 ± 0.111	0.794 ± 0.118	0.351 ± 0.020
$\hat{d}_C$	0.263 ± 0.045	0.297 ± 0.047	0.176 ± 0.015
$\hat{d}_Y$	0.104 ± 0.022	0.112 ± 0.021	0.090 ± 0.010
$\hat{d}_H$	0.216 ± 0.034	0.238 ± 0.029	0.158 ± 0.013
$\hat{d}_I$	0.130 ± 0.035	0.148 ± 0.037	0.086 ± 0.013
ln $\hat{L}$ ...	−6609.0	−6609.6	−6696.2

[a] The estimation is based on the tree ((D, C), (Y, H)). The branch lengths are measured by the average number of substitutions per site (over all sites). We denote $\hat{d}_D$, $\hat{d}_C$, $\hat{d}_Y$, and $\hat{d}_H$ by the external branch lengths of *Dictyostelium*, *Crithidia*, yeast, and human, respectively, and $\hat{d}_I$ by the internal branch length.

[b] $I + \Gamma$ is the model of invariant + gamma distribution for rate variation among sites, whereas Γ is the model of gamma distribution, and U is the model of uniform rate among sites.

method is expected to perform better for the case of five sequences than for four or fewer sequences.

In principle, the algorithm developed here can be applied to any number of sequences. However, the maximization of the likelihood function needs much more computational time than in the case of a uniform rate over sites. Therefore, the maximum likelihood method may not be very useful when the number of sequences is six or larger.

How to obtain an accurate estimate for the proportion of invariable sites θ is still an unsolved problem; the reasons are twofold. First, it needs many sequences to reduce the sampling variance, but requires too much computational time. Second, because many sites are nearly invariable when the gamma distribution parameter $\alpha < 1$, the estimation for the true proportion of invariable sites is sensitive to these nearly invariable sites. The problem will be studied in the future.

Because of its simplicity, the minimum substitution method can be applied when the number of sequences is large. Therefore, in practice, we have two choices: (1) the minimum substitution method, which uses all sequences but gives a biased estimate (Wakeley 1993); and (2) the maximum likelihood method, which gives an unbiased estimate but may be able to use only some of the sequences. To give clear-cut practical guidance, an extensive simulation study is necessary. However, the maximum-likelihood method is preferred if the sequence length is long and the number of sequences is five or smaller.

During the preparation of the manuscript, we learned that a new substitution model has been developed by Tamura and Nei (1993), which is an extension of Hasegawa et al.'s model (1985) by allowing the transition rate in purines to be different from that in pyrimidines. Since our simulation study has shown that the properties of all estimates, including α, θ, and distances, are similar under various substitution model, it is reasonable to expect that our conclusions apply to Tamura and Nei's (1993) model. Because this model may fit sequence data better than Hasagawa et al.'s (1985) model (for example, the human mitochondrial DNA D-loop sequences, Tamura 1994), we have modified our computer program to include Tamura and Nei's model. Therefore, in our program, there are two options: Hasegawa et al.'s model (1985) and Tamura and Nei's (1993) model. This program, written in FORTRAN for three, four, and five sequences, is available upon request.

Acknowledgments

This study was supported by NIH grants. We thank Dr. M Gouy for sending us well-aligned SSU sequence data.

LITERATURE CITED

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1988. Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22**:521–565.

———. 1991. PHYLIP (Phylogenetic Inference Package) version 3.4, documentation. University of Washington, Seattle.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

FITCH, W. M., and E. MARGOLIASH. 1967. A method for estimating the number of invariant amino acid codon positions in a gene using cytochrome c as a model case. Biochem. Genet. **4**:579–593.

HASEGAWA, M., A. DI RIENZO, T. D. KOCHER, and A. C. WILSON. 1993. Toward a more accurate time scale for the human mitochondrial DNA tree. J. Mol. Evol. **37**:347–354.

HASEGAWA, M., and M. FUJIWARA. 1993. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-jointing methods for estimating protein phylogeny. Mol. Phylogen. Evol. **2**:1–5.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HOLMQUIST, R., M. GOODMAN, T. CONROY, and J. CZELUSNIAK. 1983. The spatial distribution of fixed mutations within genes codings for proteins. J. Mol. Evol. **19**:437–448.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

JOHNSON, N. L., and S. KOTZ. 1969. Discret distributions. Houghton Mifflin, Boston.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–32 *in* H. R. MUNRO, ed. Mammalian protein metabalism, Vol. III. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:**111–120.

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolution of mitochondrial DNA in human and chimpanzees: control region and a protein-coding region. Pp 41–64 *in* S. OSAWA and T. HONJO, eds. Evolution of life: fossils, molecules and culture. Springer, Tokyo.

LARSON, A. 1991. Evolutionary analysis of length-variable sequences: divergent domains of ribosomal RNA. Pp 221–248 *in* M. M. MIYAMOTO and J. CRACRAFT, eds. Phylogenetic analysis of DNA sequences. Oxford University Press.

LI, W. H., M. GOUY, P. M. SHARP, C. O'HUIGIN, and Y. W. YANG. 1990. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla and Carnivora and molecular clocks. Proc. Natl. Acad. Sci. USA. **87:**6703–6707.

OLSEN, G. J. 1987. Earlist phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Pp. 825–837 *in* Cold Spring Harbor Symposia on Quantitative Biology **52.**

PALUMBI, S. R. 1989. Rates of molecular evolution and the fraction of nucleotide positions free to vary. J. Mol. Evol. **29:**180–187.

SAITOU, N. 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. **27:**261–273.

SHOEMAKER, J. S., and FITCH, W. M. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. **6:** 270–289.

SIDOW, A., T. NGUYEN, and T. SPEED. 1992. Estimating the fraction of invariable codons with a capture-recapture method. J. Mol. Evol. **35:**253–260.

TAMURA, K. 1992. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **9:**678–687.

TAMURA, K., and M. NEI. 1993. Estimating of the number of nucleotide cubstitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10:**512–526.

TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-jointing, and maximum-parsimory methods when substitution rate varies with sites. Mol. Biol. Evol. **11:**261–277.

UZZEL, T., and K. W. CORBIN. 1971. Fitting discrete probability distributions to evolutionary events. Science **172:** 1089–1096.

VAN DE PEER, Y., I. M. NEEFS, P. DE RIJK, and R. D. WACHTER. 1993. Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: calibration of the molecular clock. J. Mol. Evol. **37:**221–232.

WAKELEY, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J. Mol. Evol. **37:**613–623.

YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10:**1396–1401.

MANOLO GOUY, reviewing editor

Received May 19, 1994

Accepted January 20, 1995