# Approximate Methods for Estimating the Pattern of Nucleotide Substitution and the Variation of Substitution Rates Among Sites

*Ziheng Yang and Sudhir Kumar*

Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University

We propose two approximate methods (one based on parsimony and one on pairwise sequence comparison) for estimating the pattern of nucleotide substitution and a parsimony-based method for estimating the gamma parameter for variable substitution rates among sites. The matrix of substitution rates that represents the substitution pattern can be recovered through its relationship with the observable matrix of site pattern frequences in pairwise sequence comparisons. In the parsimony approach, the ancestral sequences reconstructed by the parsimony algorithm were used, and the two sequences compared are those at the ends of a branch in the phylogenetic tree. The method for estimating the gamma parameter was based on a reinterpretation of the numbers of changes at sites inferred by parsimony. Three data sets were analyzed to examine the utility of the approximate methods compared with the more reliable likelihood methods. The new methods for estimating the substitution pattern were found to produce estimates quite similar to those obtained from the likelihood analyses. The new method for estimating the gamma parameter was effective in reducing the bias in conventional parsimony estimates, although it also overestimated the parameter. The approximate methods are computationally very fast and appear useful for analyzing large data sets, for which use of the likelihood method requires excessive computation.

## Introduction

Estimation of the pattern of nucleotide substitution, i.e., the relative probabilities of substitution between different nucleotides, is useful for revealing the evolutionary dynamics of different genes or genomes. For example, transitions (T $\leftrightarrow$ C, A $\leftrightarrow$ G) are known to occur more often than transversions (T, C $\leftrightarrow$ A, G) in animal mitochondrial DNAs (e.g., Brown et al. 1982), and human immunodeficiency viral genes have an unusual pattern of nucleotide substitution in that A $\leftrightarrow$ G changes occur much more frequently than any other changes (e.g., Moriyama et al. 1991). Information concerning the pattern of nucleotide substitution in real sequences can also be incorporated into methods for phylogenetic tree reconstruction (e.g., Yang 1994*a*).

The variation of substitution rates across nucleotide or amino acid sites appears to be a common characteristic of sequence evolution. The major reason seems to be the different selective constraints exerted on different sites, which lead to rate variation among sites. It is well known that ignoring such rate variation causes underestimation of sequence divergences (e.g., Gillespie 1986; Takahata 1991), and that the underestimation is more serious for large distances than for small ones (e.g., Yang et al. 1994). This unproportional underestimation of distances leads to biased estimations of branching dates (e.g., Adachi and Hasegawa 1995) and appears to account for the observation that reconstruction of the tree topology can also be misleading when the rate variation is ignored (Kuhner and Felsenstein 1994; Tateno et al. 1994).

Variable substitution rates at sites have most often been described by a gamma distribution. The distribution involves a shape parameter, which is inversely related to the extent of rate variation. Given the knowledge of this parameter, a number of formulas have been suggested for estimating sequence divergence under the gamma model of rates at sites (e.g., Golding 1983; Jin and Nei 1990; Tamura and Nei 1993; Rzhetsky and Nei 1994; Yang 1994*b*). Reliable estimation of the shape parameter can not only reveal characteristics of the gene but also improves the accuracy of phylogenetic tree reconstruction (e.g., Kuhner and Felsenstein 1994; Tateno et al. 1994).

Both parsimony and likelihood methods have been employed to estimate the pattern of nucleotide substitution and the shape parameter of the gamma distribution for rates among sites. In a likelihood analysis, a substitution model is constructed and relevant parameters are estimated by maximizing the likelihood function. For instance, a gamma distribution can be assumed for rates among sites to estimate the shape parameter of the distribution (Yang 1993, 1994*b*) and a general Markov-process model of nucleotide substitution can be assumed for estimating the substitution pattern (Yang 1994*a*). In a parsimony analysis, the nucleotides in ancestral sequences are inferred by minimizing the number of changes (differences) for each site along the tree (Fitch 1971; Hartigan 1973). Then the counted (minimum) numbers of changes at sites are used to estimate the shape parameter of the gamma distribution for rates among sites (e.g., Uzzell and Corbin 1971; Holmquist et al. 1983; Larson and Wilson 1989; Kocher and Wilson 1991; Tamura and Nei 1993; Wakeley 1993). The inferred ancestral sequences are also used to calculate a matrix of frequencies of changes between nucleotides; this matrix is often interpreted as an estimation of the substitution pattern. This approach was previously used to construct the empirical matrix of amino acid substitution probabilities by Dayhoff et al. (1978; see also Jones et al. 1992) and was later used to compare nucleotide sequences for estimating the pattern of nucleotide substitution (Gojobori et al. 1982; Li et al. 1984; Go-

Key words: likelihood, parsimony, pairwise comparison, substitution pattern, rate variation among sites, Markov models, reversibility, generalized sequence distance, molecular evolution.

Address for correspondence and reprints: Ziheng Yang, Department of Integrative Biology, University of California, Berkeley, California 94720-3140. E-mail: ziheng@mws4.biol.berkeley.edu.

*Mol. Biol. Evol.* 13(5):650–659. 1996
© 1996 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

jobori and Yokoyama 1987; Moriyama et al. 1991; Imanishi and Gojobori 1992; Tamura and Nei 1993).

However, the parsimony analysis generally fails to account for nucleotide-frequency bias, transition/transversion rate bias, rate variation across sites, and unequal branch lengths in the tree, so that the reliability of estimates obtained from such analyses may be questionable. For instance, Wakeley (1994) pointed out that if variation of substitution rates among sites is ignored, the transition/transversion rate bias will be underestimated (see also Yang et al. 1994). Nucleotide-frequency biases were also found to mislead parsimony reconstructions of ancestral sequences and lead to biased estimation of the substitution pattern by parsimony (Collins et al. 1994; Perna and Kocher 1995). However, not much is known about the reliability of the parsimony analysis or about the differences between the likelihood and parsimony analyses of the two problems.

The original purpose of this paper was to analyze several real data sets by the likelihood and parsimony methods in order to characterize their differences and to examine the direction and magnitude of systematic errors involved in the parsimony analysis. In our attempts to obtain comparable results, we derived approximate methods for estimating the substitution pattern and the gamma parameter for variable rates across sites. The new methods are described below and are compared with the likelihood methods developed previously.

## Data and Methods

### Data

We analyzed two data sets of nucleotide sequences and one data set of protein sequences. All three data sets were used for estimating the gamma parameter for rates at sites, while the two data sets of nucleotide sequences were used for estimating the pattern of nucleotide substitution. Whenever the well-established phylogeny was available, it was assumed in the analysis; the effects of the tree topology on the estimations were examined.

#### *Human DNA Sequences of the Control Region of the Mitochondrial Genome*

Twenty-five sequences were extracted from the data of 136 different sequences published by Vigilant et al. (1991) for the control region of the human mitochondrial genome. Only sites 1–358 and 604–937 were used (see Hedges et al. 1992), and sites involving gaps were removed. There are 601 sites in each sequence, out of which 510 are constant sites, occupied by identical nucleotides across species. The maximum-likelihood tree for the 25 sequences, obtained by using a divisive algorithm to perform a heuristic tree-search, has many multifurcations, and the data appear to contain little information about the phylogenetic relationship among the sequences (e.g., Hedges et al. 1992). This maximum-likelihood tree (not shown) will be assumed in later analyses.

#### *Small-subunit rRNAs*

The 16S-like rRNAs of 10 animals (*Placopecten magellanicus, Herdmania momus, Xenopus laevis, Artemia salina, Tripedalia cystophora, Anemonia sulcata, Trichoplax adhaerens, Mnemiopsis leidyi, Microciona prolifera, Scypha lingua*), four fungi (*Aureobasidium pullulans, Saccharomyces cerevisiae, Athelia bombacina, Blastocladiella emersonii*), and three plants (*Zea mays, Zamia pumila, Chlamydomonas reinhardtii*) in the data of Wainright et al. (1993) were analyzed. After exclusion of gaps from the alignment of those authors, 1,477 nucleotides are left in each sequence. The relevant part of the phylogenetic tree of the authors was used.

#### *Mitochondrial Cytochrome* b *Sequences*

A set of mitochondrial cytochrome *b* sequences, extracted from GenBank, were used to estimate the rate variation among amino acid sites. The species include two whales (*Balaenoptera physalus* and *B. musculus*), cow (*Bos taurus*), rat (*Rattus norvegicus*), mouse (*Mus musculus*), opossum (*Didelphis virginiana*), chicken (*Gallus gallus*), African clawed frog (*Xenopus laevis*), carp (*Cyprinus carpio*), loach (*Crossostoma lacustre*), trout (*Oncorhynchus mykiss*), smalltail shark (*Carcharhinus porosus*), horn shark (*Heterodontus fransci*), ray (*Urolopus concentricus*), sea lamprey (*Petromyzon marinus*), and sea urchin (*Paracentrotus lividus*). Each sequence has 375 amino acids after exclusion of sites involving gaps. The phylogenetic relationship of the species is well established from morphological and fossil evidences, and the "correct" tree topology (not shown) was assumed.

### Methods

#### *Estimation of the Pattern of Nucleotide Substitution*

Nucleotide substitution is described by a continuous-time stationary Markov process. The pattern of nucleotide substitution is represented by the "rate matrix" of the process, $Q = \{Q_{ij}\}$, where $Q_{ij}$ ($j \neq i$) is the rate of substitution from nucleotide $i$ to $j$. The row sums of this matrix is zero, and $-Q_{ii} = \Sigma_{j \neq i} Q_{ij}$ is the substitution rate of nucleotide $i$. The matrix of transition probabilities over time $t$, which is needed in the likelihood calculation (e.g., Felsenstein 1981), can be generated for any $t$ by

$$P(t) = \{P_{ij}(t)\} = e^{tQ}, \quad (1)$$

where $P_{ij}(t)$ is the probability that a given nucleotide $i$ will become $j$ after time $t$. Because $t$ and $Q$ appear in the form $tQ$ only, $Q$ is multiplied by a constant so that the average rate of substitution is 1:

$$-\sum_i \pi_i Q_{ii} = 1, \quad (2)$$

while time $t$ is measured by the average number of substitutions per site. Depending on the context, we use $t$ to mean time, the distance between two extant sequences, or the branch length in a phylogenetic tree.

Maximum-likelihood estimation of the substitution pattern was performed using the method of Yang

(1994*a*), with the general reversible Markov-process model of nucleotide substitution assumed. This model makes a mild restriction about the structure of $Q$, that is,

$$\pi_i Q_{ij} = \pi_j Q_{ji}, \text{ for any } i, j. \quad (3)$$

This reversibility restriction, together with equation (2), reduces the number of free parameters in the rate matrix from 11 to 8. The rate matrix $Q$, as well as branch lengths in the tree, was estimated by maximum likelihood. Estimation of the substitution pattern under the gamma model of variable rates among sites was performed using the discrete-gamma model of Yang (1994*b*), with eight rate categories used.

Consider two sequences separated by time (distance) $t$. Let $F_{ij}(t)$ be the probability of observing a site with nucleotides $i$ and $j$ in the two sequences. $F_{ij}(t)$ is equal to the probability of observing nucleotide $i$ in sequence 1, which is the equilibrium frequency ($\pi_i$) of nucleotide $i$, times the transition probability from nucleotide $i$ to $j$ in the time interval $t$. In other words,

$$F_{ij}(t) = \pi_i P_{ij}(t) \text{ or } F(t) = \Pi P(t), \quad (4)$$

where $F(t) = \{F_{ij}(t)\}$ and $\Pi = \text{diag}\{\pi_T, \pi_C, \pi_A, \pi_G\}$. Equations (1) and (4) can then be used to obtain estimates of $t$ and $Q$:

$$tQ = \log\{P(t)\} = \log\{\Pi^{-1}F(t)\}. \quad (5)$$

Because of the requirement of equation (2), both $t$ and $Q$ can be uniquely estimated. The reversibility of the process, which we assume here, implies that $F_{ij}(t) = F_{ji}(t)$, and so the observed $F(t)$ matrix is made symmetrical before equation (5) is applied. The calculation can be carried out by the spectral decomposition (diagonalization) of $P(t)$. Let

$$P(t) = \Pi^{-1}F(t) = U\text{diag}\{\delta_1, \delta_2, \delta_3, \delta_4\}U^{-1}, \quad (6)$$

where the $\delta$s are the eigenvalues of $P(t)$ and the columns of $U$ are the corresponding (right) eigenvectors. Then, substituting (6) in (5), we get

$$tQ = U\text{diag}\{\log(\delta_1),\log(\delta_2),\log(\delta_3),\log(\delta_4)\}U^{-1}. \quad (7)$$

When substitution rates among sites are assumed to follow a gamma distribution with shape parameter $\alpha$, equations 5 and 7 do not hold anymore. The eigenvalues ($\delta$s) of $\Pi^{-1}F(t)$ (or $P(t)$) and those ($\lambda$s) of $Q$ are then related by $\delta_k = (\alpha/(\alpha-\lambda_k))^\alpha$. Thus the transformation to be used will be

$$tQ = U\text{diag}\{\alpha(1 - \delta_1^{-1/\alpha}), \alpha(1 - \delta_2^{-1/\alpha}), \alpha(1 - \delta_3^{-1/\alpha}), \alpha(1 - \delta_4^{-1/\alpha})\}U^{-1}, \quad (8)$$

with columns of $U$ the (right) eigenvectors of $\Pi^{-1}F(t)$ (e.g., Yang 1994*b*). Note that $\alpha(1 - \delta_k^{-1/\alpha}) \to \log(\delta_k)$ when $\alpha \to \infty$.

We estimate $F_{ij}(t)$ by $(N_{ij} + N_{ji})/N$, where $N_{ij}$ is the number of sites occupied by nucleotides $i$ and $j$ in the two sequences respectively and $N = \Sigma_{i,j}N_{ij}$ is the total number of sites in the sequence. $F(t)$ constructed this way is symmetrical with $\Sigma_{i,j}F_{i,j}(t) = 1$. Two possibilities were considered: one is to take the average of the $F(t)$s over all the pairwise comparisons, and the other is to take the average of the $Q$s calculated from each pairwise comparison.

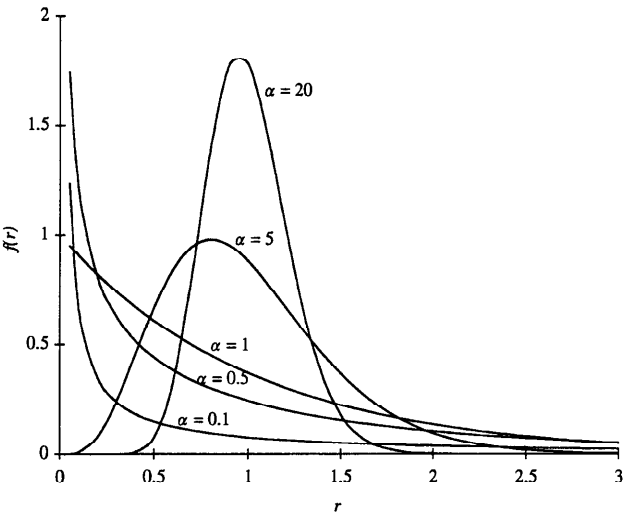


FIG. 1.—The gamma distribution with different shape parameter $\alpha$. The mean and variance of the distribution are 1 and $1/\alpha$, respectively.

In a parsimony analysis, the ancestral sequences inferred by parsimony (Hartigan 1973) were used, so one of the two sequences in comparison is ancestral and the other is descendent. All sites including the invariant sites were used in the analysis. When many equally best pathways were possible at a site, an equal weight was assigned to each of them; if there existed 100 equally best pathways, each was given a weight of 1/100. The average of $F(t)$ taken over all branches of the tree was used to estimate $t$ and $Q$ by equation (5). The estimate of $t$ may be considered an average of branch lengths in the phylogenetic tree.

### *Estimation of the α Parameter of the Gamma Distribution*

The gamma distribution has been used to describe variable substitution rates across sites. The distribution with its mean fixed to be one (Yang 1993) has density

$$f(r) = \frac{\alpha^\alpha}{\Gamma(\alpha)} r^{\alpha-1} e^{-\alpha r}, \qquad \alpha > 0, \qquad r > 0. \quad (9)$$

The shape parameter $\alpha$ is inversely related to the extent of rate variation at sites (fig. 1). The distribution with $\alpha > 1$ is bell-shaped ($\cap$), which means that most sites have intermediate rates, while few sites have very low or very high rates. An infinitely large $\alpha$ means a constant rate for all sites (Yang 1993). The distribution with $\alpha \leq 1$ is highly skewed, has an L shape, and means that most sites have very low rates or are almost "invariable," while mutational "hot spots" exist in the sequence that change at very high rates. So the $\alpha$ parameter, when estimated from the sequence data, reflects characteristics of the gene.

Estimation of the $\alpha$ parameter by the likelihood method was described by Yang (1993, 1994*b*). The discrete-gamma model of Yang (1994*b*) was used in this

paper to achieve computational efficiency, with eight categories of rates used to approximate the continuous gamma. Different substitution models were assumed in combination with the gamma model of rates among sites.

In a parsimony analysis, the (minimum) numbers of changes at sites were inferred by the algorithm of Hartigan (1973). Such numbers (if we ignore possible errors in the inference) will follow a Poisson distribution if the rate is constant at sites, or a negative-binomial distribution if the rate is gamma-distributed across sites. This observation has been employed to estimate $\alpha$ and to test for rate constancy across sites. The method is described below. By the negative-binomial distribution, the probability that $k$ changes occur at a site is

$$\text{prob}(k) = \binom{k + \alpha - 1}{\alpha - 1}\left(\frac{1}{1 + \mu/\alpha}\right)^{\alpha}\left(\frac{\mu/\alpha}{1 + \mu/\alpha}\right)^{k}, \quad (10)$$

where $\alpha$ is the shape parameter of the gamma distribution, and $\mu$ is the average number of substitutions per site. Let $N_k$ be the number of sites at which $k$ changes are inferred to have occurred along the tree. The sample mean and variance of the number of changes at sites can be equated to the mean and variance of the negative-binomial distribution to give

$$\mu = \bar{x} = \frac{\Sigma_k k N_k}{\Sigma_k N_k}$$

$$\alpha = \bar{x}^2/(s^2 - \bar{x}), \quad (11)$$

where $s^2$ is the sample variance. Parameter $\alpha$ cannot be estimated if $s^2 < \bar{x}$. This method, commonly used, is known as "the method of moments," and tends to overestimate $\alpha$ when $\alpha$ is small even if the numbers of changes at sites inferred by parsimony were accurate (Sullivan et al. 1995; see also Johnson et al. 1992:214–220). A maximum-likelihood method for fitting the negative-binomial distribution was suggested by Sullivan et al. (1995), in which $\alpha$ and $\mu$ were estimated by numerical maximization of the log-likelihood function:

$$\log(L) = \sum_k N_k \log\{\text{prob}(k)\}$$

$$= \sum_k N_k \left\{\log\binom{k + \alpha - 1}{\alpha - 1} - \alpha \log(1 + \mu/\alpha) + k \log\left(\frac{\mu/\alpha}{1 + \mu/\alpha}\right)\right\}. \quad (12)$$

The estimate of $\mu$ from equation (12) is identical to $\bar{x}$ (equation 11).

Nevertheless, we suggest that, given the ancestral sequences inferred by parsimony, the number of "changes" at a site should better be interpreted as the number of "differences" between the compared sequences at the ends of branches along the tree. Unlike the number of substitutions, the number of differences does not follow a negative-binomial distribution when rates at sites are gamma-distributed. In the following, we derive its correct distribution, and thus a new parsimony-based method for estimating $\alpha$. Assuming a Poisson-process model of substitution, we obtain the probability that a nucleotide (or amino acid) $i$ will become $j$ after time $t$ as

$$P_{ij}(t) = \begin{cases} \frac{1}{c} + \frac{c-1}{c}\exp\left(-\frac{c}{c-1}t\right), & \text{if } i = j, \\ \frac{1}{c} - \frac{1}{c}\exp\left(-\frac{c}{c-1}t\right), & \text{if } i \neq j, \end{cases} \quad (13)$$

where the number of character states $c$ is 4 for nucleotide sequences or 20 for amino acid sequences. The conditional probability of observing a site with a specific character pattern with $k$ differences along all the $b$ branches of the tree, given that the substitution rate for the site is $r$ (which is a gamma variable), is

$$\text{prob}(k|r) = [P_{ii}(rt)]^{b-k}[P_{ij}(rt)]^{k}. \quad (14)$$

The unconditional probability is then

$$\text{prob}(k) = \int_0^{\infty} \text{prob}(k \mid r) f(r)\, dr, \quad (15)$$

where $f(r)$ is given in equation (9). When we use the discrete-gamma model (Yang 1994*b*), with several categories used to approximate the continuous gamma, we have

$$\text{prob}(k) = \sum_i f(r_i) \times \text{prob}(k \mid r_i), \quad (16)$$

where $f(r_i) = 1/8$ is the frequency of the category with rate $r_i$. The rate $r_i$ for category $i$ is a function of $\alpha$ and can be calculated by the method of Yang (1994*b*). The log-likelihood function is then

$$\log(L) = \sum_k N_k \log\{\text{prob}(k)\}$$

$$= \sum_k N_k \log\left\{\sum_i \left[f(r_i)\left(\frac{1}{c} + \frac{c-1}{c}\exp\left(-\frac{c}{c-1}r_i t\right)\right)^{b-k} \cdot \left(\frac{1}{c} - \frac{1}{c}\exp\left(-\frac{c}{c-1}r_i t\right)\right)^{k}\right]\right\}. \quad (17)$$

Note that the summation over $k$ is equivalent to summing over sites in the sequence or over all the possible site patterns. Parameter $\alpha$ can be estimated by numerical maximization of (17). Time $t$ needed in equation (17) can be calculated as $\hat{\mu}/b$ (equation 11), where $b$ is the number of branches in the tree. Although this is clearly an underestimate of the average branch length in the tree, using a larger value does not seem to improve the estimation of $\alpha$. In this paper, this estimate of $t$ was used.

## Results

### Estimation of the Pattern of Nucleotide Substitution

#### *Human Mitochondrial DNAs*

The estimated rate matrices by the likelihood methods are given in table 1. The sequences are very similar,

**Table 1**
**Estimates of the Pattern of Nucleotide Substitution for the Control Region of the Human Mitochondrial Genome (601 bp) by the Likelihood Method**

| | To | | | |
|---|---|---|---|---|
| From | T | C | A | G |
| (*a*) Assuming the maximum-likelihood (ML) tree: | | | | |
| T ... | −1.3346 | 1.3056 | 0.0144 | 0.0146 |
| C ... | 1.0409 | −1.1226 | 0.0817 | 0.0000 |
| A ... | 0.0103 | 0.0733 | −0.5692 | 0.4856 |
| G ... | 0.0253 | 0.0000 | 1.1707 | −1.1960 |
| (*b*) Assuming the star tree: | | | | |
| T ... | −1.4677 | 1.4290 | 0.0093 | 0.0293 |
| C ... | 1.1393 | −1.2093 | 0.0700 | 0.0000 |
| A ... | 0.0066 | 0.0629 | −0.4703 | 0.4008 |
| G ... | 0.0506 | 0.0000 | 0.9664 | −1.0171 |
| (*c*) Assuming the ML tree and gamma rates at sites: | | | | |
| T ... | −1.3677 | 1.3405 | 0.0123 | 0.0149 |
| C ... | 1.0687 | −1.1413 | 0.0726 | 0.0000 |
| A ... | 0.0088 | 0.0651 | −0.5440 | 0.4700 |
| G ... | 0.0258 | 0.0001 | 1.1333 | −1.1591 |

NOTE.—The nucleotide frequencies are 0.2361 (T), 0.2961 (C), 0.3313 (A), and 0.1363 (G). The general reversible Markov-process model of nucleotide substitution (Yang 1994*a*) is assumed. In (*c*), eight categories of rates were assumed to approximate the continuous gamma distribution of rates for sites (Yang 1994*b*). The transition/transversion rate ratio averaged over nucleotide frequencies is 15.08, 15.69, and 16.86 for (*a*), (*b*), and (*c*), respectively.

**Table 2**
**Estimation of the Substitution Pattern for the Control Region of the Human Mitochondrial Genome by the Parsimony Method**

| | To | | | |
|---|---|---|---|---|
| From | T | C | A | G |
| (*a*) The matrix of counts [$=601 \times 39 \times F(t)$]: | | | | |
| T ... | 5,490.5 | 44 | 0.5 | 0.5 |
| C ... | 44 | 6,902.5 | 3.5 | 0 |
| A ... | 0.5 | 3.5 | 7,723.5 | 23 |
| G ... | 0.5 | 0 | 23 | 3,179.5 |
| (*b*) The transition-probability matrix $P(t)$: | | | | |
| T ... | 0.99187 | 0.00795 | 0.00009 | 0.00009 |
| C ... | 0.00633 | 0.99317 | 0.00050 | 0.00000 |
| A ... | 0.00007 | 0.00045 | 0.99652 | 0.00297 |
| G ... | 0.00016 | 0.00000 | 0.00718 | 0.99266 |
| (*c*) The rate matrix $Q$: | | | | |
| T ... | −1.33354 | 1.30432 | 0.01441 | 0.01480 |
| C ... | 1.03886 | −1.12108 | 0.08240 | 0.00000 |
| A ... | 0.01029 | 0.07389 | −0.57012 | 0.48593 |
| G ... | 0.02559 | 0.00000 | 1.17585 | −1.20106 |

NOTE.—Sequences of 25 humans were analyzed, with the maximum-likelihood tree assumed. The transition/transversion rate ratio averaged over nucleotide frequencies is 14.98. See Note to table 1.

and most changes are transitions; the transition/transversion rate ratio, averaged over nucleotide frequencies, is 15.08. Without any substitution-rate bias, the diagonal elements of $Q$ would be close to −1, the off-diagonal elements close to 1/3, and the transition/transversion rate ratio would be close to 0.5. Substitution rates in the mitochondrial control region are thus highly biased. As noted by Yang (1994*a*), estimation of the substitution pattern does not seem to be sensitive to the assumed tree topology; use of the star phylogeny (table 1*b*) is seen to produce very similar results. Substitution rates are highly variable among sites in these sequences ($\hat{\alpha} = 0.172 \pm 0.037$). Compared with the substitution pattern estimated under the gamma model of rates among sites (table 1*c*), ignoring the rate variation among sites (table 1*a*) leads to more similar substitution rates and to underestimated transition/transversion rate ratio. Nevertheless, the estimated rate matrices are quite similar to each other whether or not the rate variation among sites is taken into account in the model. This similarity appears to be partly due to the similarity of the sequences.

Estimation of the substitution pattern by the parsimony method (equation 5 or 7) is explained in table 2. Counting the site patterns at each site in all pairs of sequences along branches in the tree produces the matrix of counts shown in table 2*a*. The sum of elements of this matrix is the number of sites (601) times the number of branches in the (multifurcating) tree (39). Dividing this matrix by 23,439 ($= 601 \times 39$) produces the matrix $F(t)$. The row sums of $F(t)$ give estimates of the nucleotide frequencies, that is, 0.23615 (T), 0.29651 (C), 0.33067 (A), and 0.13665 (G). Equation (5) was then used to calculate $P(t)$ and $Q$. The estimated rate matrix $Q$ (table 2*c*) is similar to that obtained by the likelihood analysis (table 1*a*). The average transition/transversion rate ratio estimated by parsimony (14.98) is also close to that obtained from the likelihood analysis (15.08) (table 1*a*).

Results obtained from pairwise comparisons of sequences are given in table 3. The matrix in table 3*a* was obtained by averaging $F(t)$ over all pairwise comparisons before equation (5) was applied, while that in table 3*b* is an average of the estimated $Q$s for each pairwise comparison. The latter approach was not used in the parsimony analysis because the estimated $Q$s for short

**Table 3**
**Estimation of the Substitution Pattern for the Control Region of the Human Mitochondrial Genome by the Pairwise Comparison**

| | To | | | |
|---|---|---|---|---|
| From | T | C | A | G |
| (*a*) $Q$ from average $F(t)$: | | | | |
| T ... | −1.4028 | 1.3639 | 0.00084 | 0.0306 |
| C ... | 1.0874 | −1.1693 | 0.0812 | 0.0007 |
| A ... | 0.0060 | 0.0729 | −0.5154 | 0.4366 |
| G ... | 0.0527 | 0.0014 | 1.0527 | −1.1069 |
| (*b*) Average of $Q$s from pairwise comparison: | | | | |
| T ... | −1.4264 | 1.3865 | 0.0126 | 0.0288 |
| C ... | 1.1049 | −1.1699 | 0.0649 | 0.0011 |
| A ... | 0.0090 | 0.0581 | −0.5013 | 0.4354 |
| G ... | 0.0498 | 0.0023 | 1.0500 | −1.1002 |
| (*c*) $Q$ from average $F(t)$ with gamma rates at sites ($\alpha = 0.17$): | | | | |
| T ... | −1.4374 | 1.4083 | 0.0000 | 0.0313 |
| C ... | 1.1228 | −1.1966 | 0.0781 | 0.0000 |
| A ... | 0.0000 | 0.0701 | −0.4877 | 0.4191 |
| G ... | 0.0540 | 0.0000 | 1.0104 | −1.0551 |

NOTE.—The transition/transversion rate ratios averaged over nucleotide frequencies are 13.94, 16.49, and 16.38 for (*a*), (*b*), and (*c*), respectively. See Note to table 1.

**Table 4**
**Estimates of the Pattern of Nucleotide Substitution for the 16S-like rRNAs**

| | To | | | |
|---|---|---|---|---|
| From | T | C | A | G |
| (*a*) Likelihood, constant rate at sites: | | | | |
| T ... | −1.0952 | 0.6778 | 0.2143 | 0.2031 |
| C ... | 0.8655 | −1.2876 | 0.2235 | 0.1985 |
| A ... | 0.2020 | 0.1650 | −0.8460 | 0.4790 |
| G ... | 0.1984 | 0.1519 | 0.4964 | −0.8466 |
| (*b*) Likelihood, gamma rates at sites: | | | | |
| T ... | −1.0997 | 0.7364 | 0.7160 | 0.1873 |
| C ... | 0.9404 | −1.3435 | 0.2141 | 0.1809 |
| A ... | 0.1659 | 0.1580 | −0.8127 | 0.4887 |
| G ... | 0.1830 | 0.1446 | 0.5064 | −0.8340 |
| (*c*) Parsimony: | | | | |
| T ... | −1.0757 | 0.6453 | 0.2185 | 0.2119 |
| C ... | 0.8468 | −1.2916 | 0.2345 | 0.2104 |
| A ... | 0.2166 | 0.1689 | −0.8452 | 0.4697 |
| G ... | 0.2112 | 0.1598 | 0.4952 | −0.8662 |
| (*d*) Pairwise, constant rate for sites: | | | | |
| T ... | −1.0979 | 0.6693 | 0.2067 | 0.2219 |
| C ... | 0.8547 | −1.2937 | 0.2308 | 0.2082 |
| A ... | 0.1948 | 0.1704 | −0.8291 | 0.4639 |
| G ... | 0.2168 | 0.1593 | 0.4807 | −0.8568 |
| (*e*) Pairwise, gamma rates for sites with $\alpha = 0.3$: | | | | |
| T ... | −1.1590 | 0.8060 | 0.1639 | 0.1892 |
| C ... | 1.0292 | −1.3961 | 0.1988 | 0.1681 |
| A ... | 0.1545 | 0.1468 | −0.7607 | 0.4595 |
| G ... | 0.1848 | 0.1286 | 0.4761 | −0.7896 |

NOTE.—The average transition/transversion rate ratios are 1.59, 1.85, 1.48, 1.50, and 2.02 for (*a*), (*b*), (*c*), (*d*), and (*e*), respectively. See Note to table 1.

branches were unreliable. Table 3*c* was obtained from the average $F(t)$ over branches by equation (8) with $\alpha$ fixed at 0.17. The pairwise comparison produced results similar to the likelihood results, although for this data set, the parsimony results appear to be better.

*Small-subunit rRNAs*

Results obtained from the analysis of the 16S-like rRNAs are presented in table 4. Substitution rates between different pairs of nucleotides are much more similar to one another than those in human mtDNAs (see tables 1, 2, and 3), and the average transition/transversion rate ratio is estimated to be 1.59. Allowing for rate variation at sites by assuming the gamma model of rates at sites leads to more extreme substitution rates, with the average transition/transversion rate ratio estimated at 1.85 (table 4*b*). This parallels the previous finding that ignoring the rate variation among sites leads to underestimation of the transition/transversion rate bias (Wakeley 1994; Yang et al. 1994).

The rate matrices estimated by the parsimony and pairwise analyses are given in tables 4*c*, *d*, and *e*. As one expects, the parsimony results are closer to those obtained from the likelihood analysis assuming a single rate for all sites (table 4*a*) than to those obtained from likelihood assuming gamma rates for sites (table 4*b*). The transition/transversion rate ratio is underestimated by parsimony. The pairwise estimates of $Q$ shown in tables 4*d* and *e* were obtained from the average $F(t)$ over all pairwise sequence comparisons. The practice of taking the average of the $Q$s from each pairwise comparison gave very similar results (not shown). Unlike the case of the mtDNAs, results of the pairwise comparison are closer to the likelihood results than are those of the parsimony analysis. Overall, the approximate methods gave quite reliable estimates of the substitution pattern for this data set.

## Estimation of the Rate Variation Across Sites

*Human Mitochondrial DNAs*

The human control-region DNA sequences were analyzed using the likelihood and parsimony methods to estimate the gamma parameter for variable substitution rates among sites. The results are shown in table 5*a*. The parsimony algorithm (Hartigan 1973) suggests no change at 510 invariant sites in the data, and one, two, three, and four changes (differences) at 62, 13, 9, and 7 sites, respectively. The mean and variance of the number of changes at a site are $\hat{\mu} = 0.238$ and $s^2 = 0.455$, from which the $\alpha$ parameter can be estimated by the method of moments (equation 11) as $\hat{\alpha} = 0.261$ (table 5a). The maximum-likelihood method of Sullivan et al. (1995) (equation 12) applied to these numbers of changes inferred by parsimony gave an estimate $\hat{\alpha} = 0.234$, and the new method of this paper (equation 17) produced the estimate $\hat{\alpha} = 0.179$.

The maximum-likelihood method applied to the original sequence data (Yang 1994*b*) produced estimates 0.172 and 0.269 for $\alpha$ and $\mu$, respectively, when the general reversible-process model of nucleotide substitution (Yang 1994*a*) was assumed, $\mu$ being calculated as the sum of branch lengths along the tree. For these data, parsimony (the method of moments) overestimates $\alpha$ by 52% and underestimates the amount of evolution ($\mu$) by 12%, while the new method proposed in this paper is very effective in reducing the bias in the parsimony estimate of $\alpha$. The similarity of the sequences appears to be the major reason for the similarity of the $\alpha$ estimates obtained by the likelihood method when different models were assumed. In these data, the nucleotide frequencies are biased, and the transition/transversion rate ratio is very high, so that the substitution model of Jukes and Cantor (1969) is quite unrealistic (see the log-likelihood values of different models in table 5*a*). Yet, this simple model produced an estimate of $\alpha$ that is quite close to that obtained under more realistic models.

*Small-subunit rRNAs*

Much greater differences were found between likelihood and parsimony methods when the sequences are more different. Table 5*b* lists estimates of $\mu$ and $\alpha$ by the two methods for the rRNA sequences. Reconstruction of ancestral sequences by parsimony suggests 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 changes at 888, 256, 105, 84, 59, 43, 26, 9, 4, 2, and 1 site(s), respectively. By a similar analysis, we obtained the following estimates

**Table 5**
**Estimates of the α Parameter of the Gamma Distribution and the Average Number of Substitutions per Site Along the Tree (μ) by Different Methods**

| Method | $\hat{\alpha}$ | $\hat{\mu}$ | $\ell$ |
|---|---|---|---|
| (*a*) For the control region of mtDNA of 25 humans: | | | |
| MP (method of moments) | 0.261 | 0.238 | |
| MP (Sullivan et al. 1995) | 0.234 | 0.238 | |
| MP (equation 17) | 0.179 | 0.238 | |
| ML (Jukes and Cantor 1969) | 0.175 ± 0.038 | 0.254 | −1,777.71 |
| ML (Kimura 1980) | 0.168 ± 0.036 | 0.262 | −1,658.76 |
| ML (Yang 1994*a*) | 0.172 ± 0.037 | 0.269 | −1,620.40 |
| (*b*) For the 16S-like rRNAs of 17 species: | | | |
| MP (method of moments) | 0.572 | 0.980 | |
| MP (Sullivan et al. 1995) | 0.446 | 0.980 | |
| MP (equation 17) | 0.373 | 0.980 | |
| ML (Jukes and Cantor 1969) | 0.326 ± 0.021 | 1.199 | −9,049.49 |
| ML (Kimura 1980) | 0.311 ± 0.020 | 1.270 | −8,813.01 |
| ML (Yang 1994*a*) | 0.312 ± 0.020 | 1.295 | −8,794.05 |
| (*c*) For the cytochrome *b* of 16 species: | | | |
| MP (method of moments) | 0.928 | 1.891 | |
| MP (Sullivan et al. 1995) | 0.606 | 1.891 | |
| MP (equation 17) | 0.553 | 1.891 | |
| ML (Poisson) | 0.455 ± 0.048 | 2.256 | −5,259.81 |
| ML (equal-input) | 0.401 ± 0.042 | 2.612 | −4,932.48 |
| ML (Jones et al. 1992) | 0.435 ± 0.045 | 2.726 | −4,645.96 |

NOTE.—The parsimony analysis (MP) infers the numbers of changes (differences) at sites and uses the method of moments (equation 11), the likelihood method of Sullivan et al. (1995) (equation 12), or the new method of this paper (equation 17) to estimate α and μ. In the likelihood analysis (ML), a (discrete-) gamma model of rates across sites was assumed (Yang 1994*b*), in combination with a Markov-process model of nucleotide or amino acid substitution. Substitution models assumed for the two data sets of nucleotide sequences are those of Jukes and Cantor (1969), Kimura (1980), and the general reversible-process model (Yang 1994*a*), while those for the mitochondrial cytochrome *b* (375 amino acids) are the Poisson-process model, the equal-input (or proportional) model, and the empirical model of Jones et al. (1992). $\ell$ is the log-likelihood value under the model, calculated from the original sequence data.

based on these numbers of changes inferred by parsimony: $\hat{\mu} = 0.980$, and $\hat{\alpha} = 0.572$ by the method of moments (equation 11), $\hat{\alpha} = 0.446$ by the method of Sullivan et al. (1995) (equation 12), and $\hat{\alpha} = 0.373$ by the new method of this paper (equation 17).

The maximum-likelihood method applied to the original sequence data produces estimates $\hat{\mu} = 1.295$ and $\hat{\alpha} = 0.312$ under the general reversible-process model (Yang 1994*a*) (table 5*b*). Parsimony (method of moments) is seen to underestimate μ by 24% and to overestimate α by 83%. In these data, the transition/transversion rate ratio is not very high, and the nucleotide-frequency bias does not seem to have a large effect on the estimation of α, as estimates obtained from the simple models of Jukes and Cantor (1969) and Kimura (1980) are quite similar to that obtained under the general reversible-process model.

### *Mitochondrial Cytochrome* b *Sequences*

Results obtained from analyzing the mitochondrial cytochrome *b* sequences are presented in table 5*c*. The parsimony algorithm suggests 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 amino acid changes at 169, 52, 39, 30, 20, 23, 19, 12, 5, 4, and 2 sites, respectively. These numbers of changes at sites give $\hat{\mu} = 1.891$, and $\hat{\alpha} = 0.928$ by the method of moments (equation 11), $\hat{\alpha} = 0.606$ by the method of Sullivan et al. (1995) (equation 12), and $\hat{\alpha} = 0.553$ by the new method of this paper (equation 17).

In the likelihood analysis of the original sequence data (Yang 1994*b*), three models of amino acid substitution were assumed: the Poisson-process model, which assumes equal substitution rate between any amino acids; the "equal-input" model, by which the rate of substitution is proportional to the frequency of the target amino acid (i.e., $Q_{ij} = \pi_j$); and the empirical model of amino acid substitution derived by Jones et al. (1992) from the SwissProt data bank Release 22. The likelihood values under the three substitution models are drastically different, and the empirical model of Jones et al. (1992) fits the data much better than the other two models, while the fit of the Poisson-process model is the poorest (table 5*c*). As the estimates obtained under the empirical model of Jones et al. (1992) appear to be most reliable, the results suggest that for these data, parsimony (method of moments) overestimates α by 113% and underestimates the amount of evolution (μ) by 31%.

Substitution rates at sites "estimated" by the approach of Yang (1995*a*) and Yang and Wang (1995) under the model of Jones et al. (1992) were plotted along the sequence in fig. 2. Substitution rates at sites appear to be related to the functional domains of the protein; sites in the transmembrane regions tend to have higher rates than other sites in the protein.

In comparison with estimation of the substitution pattern, estimation of the α parameter appears to be more sensitive to the assumed tree topology; the use of

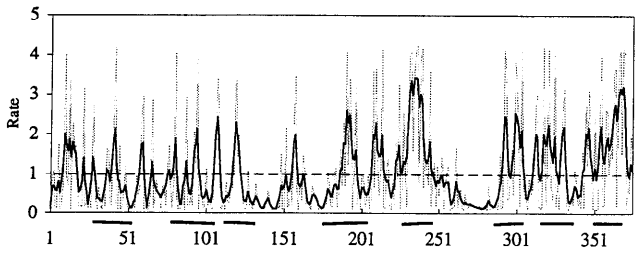


FIG. 2.—Amino acid substitution rates at sites plotted along the mitochondrial cytochrome *b* sequence. The cytochrome *b* sequences of 16 species were analyzed assuming the empirical model of Jones et al. (1992) and a discrete-gamma model of rates across sites. Branch lengths in the known tree topology and the α parameter of the gamma distribution were estimated by the maximum-likelihood approach, and substitution rates at sites were then "estimated" by the method of Yang (1995) and Yang and Wang (1995). The mean of rates across sites is 1. The correlation between the true underlying rate for a site and its estimated rate is calculated to be 0.78 (Yang and Wang 1995). Rates for sites were then smoothed using a moving average: $\hat{r}_h = (r_{h-2} + 2r_{h-1} + 4r_h + 2r_{h+1} + r_{h+2})/10$, where $r_h$ (the dotted line) is the estimated rate for site $h$, and $\hat{r}_h$ (the real line) is the rate for site $h$ after the smoothing. The eight thick bars mark the transmembrane helics of cytochrome *b* (Esposti et al. 1993).

totally wrong trees such as the star phylogeny tends to underestimate α (Yang 1994*b*). The reason seems to be that a small α means more mutational "hot spots" in the sequence, the existence of which would explain some of the variable site patterns by the wrong tree, which would otherwise be incompatible with the topology. Nevertheless, it appears possible to obtain reliable estimates of α even when the true phylogeny is unknown, as estimates obtained from using more-or-less reasonable trees have been found to be very similar (Yang 1994*b*, 1995*a*; Yang et al. 1994). For example, application of various tree-reconstruction methods implemented in the MEGA program package (Kumar et al. 1993) to the cytochrome *b* data produced four estimated (incorrect) tree topologies; estimates of α obtained from those trees under the empirical model of Jones et al. (1992) are in the range 0.436–0.446, and these estimates are very similar to that (0.435) obtained by using the correct topology (table 5*b*).

## Discussions

### Generalized Sequence Distances

As indicated before, estimate of $t$ from equation (5) can be considered a measure of pairwise distance under the general reversible-process model (Yang 1994*a*), and equation (8) is an extension of the distance formula to the case of gamma rates at sites. For the small-subunit rRNA data analyzed in this paper, estimates of pairwise distances by equation (5) are identical to those obtained under the maximum-likelihood criterion (Yang 1994*a*) at the seventh or eighth decimal points (results not shown), although in general the method is not a maximum-likelihood estimator. The distance measure will reduce to the appropriate distances when the true model is simpler than the general reversible model, such as those of Jukes and Cantor (1969) and Kimura (1980).

Estimation of sequence distance ($t$) by equation (5) appears to be first suggested by Rodriguez et al. (1990) (see also Tavaré 1986), although the method of Lanave et al. (1984) can be expected to produce similar results (Zharkikh 1994; Yang 1995*b*). Rodriguez and colleagues' version of the formula is

$$\hat{t} = \text{trace}(\Pi \log\{\Pi^{-1}F(t)\}), \tag{18}$$

where the notation trace($A$) means the sum of the diagonal elements of the matrix $A$. Those authors, however, did not make $F(t)$ symmetrical, so their method is numerically less stable; neither does it lead to a legitimate estimate of the substitution pattern ($Q$).

This distance measure is also very similar to that suggested by Barry and Hartigan (1987), known more recently as the LogDet or "paralinear" distance (Lake 1994; Steel 1994; Zharkikh 1994):

$$\hat{d} = -\frac{1}{4}\log\{\text{Det}(F(t))\}, \tag{19}$$

where the notation Det($A$) means the determinant of $A$, which is equal to the product of the eigenvalues of $A$. Consider the limiting case of infinitely long sequences, where the data matrix $F(t)$ will be the expected frequencies under the model. Since Det($AB$) = Det($A$)Det($B$), we have, in expectation,

$$\begin{aligned}\hat{d} &= -\frac{1}{4}\log\{\text{Det}(\Pi P(t))\}\\ &= -\frac{1}{4}\log\{\text{Det}(\Pi)\text{Det}(P(t))\}\\ &= -\frac{1}{4}\log\left\{\prod_k \pi_k \prod_k e^{\lambda_k t}\right\}\\ &= -\frac{1}{4}\sum_k \log\{\pi_k\} - t\left(\frac{1}{4}\sum_k \lambda_k\right),\end{aligned} \tag{20}$$

where the $\lambda_k$s are the eigenvalues of $Q$. Equation (20) suggests that $\Sigma_k \log\{\pi_k\}/4$ should be subtracted from $\hat{d}$. The factor $\Sigma_k \lambda_k/4 = \text{trace}(Q)/4 = \Sigma_k Q_{kk}/4$ is equal to 1 if the nucleotides have equal frequencies 1/4 (for example, under the models of Jukes and Cantor [1969] or Kimura [1980]) but is not in presence of nucleotide frequency bias (cf: equation 2). So although $\hat{d}$ is a linear function of $t$, it generally does not converge to the expected number of nucleotide substitutions per site.

Both distances (equations 18 and 19) are applicable to amino acid sequences. Both distances are, however, inapplicable when any of the eigenvalues of $\Pi^{-1}F(t)$ (for equation 5 or 8) or $F(t)$ (for equation 19) are negative. Approaches like that of Tajima (1993) may then be helpful. With protein sequences, the absence of some amino acids in the data (i.e., $\pi_i = 0$ for some $i$) also cause problems, and ad hoc treatment is necessary. Barry and Hartigan (1987; see also Steel 1994) argue that the LogDet formula estimates an average distance when the substitution process is nonstationary. However, such arguments can be applied to any distance measure currently used, in the sense that the estimated distance (or pattern) can be interpreted as an average of a variable substitution rate, and they do not provide evidence that the LogDet

formula is superior to other distances when nucleotide frequencies are unequal in different sequences.

### Estimation of the Substitution Pattern

The parsimony method lacks a clearly specified model with well-defined parameters. This makes it difficult to interpret results and to assess biases in estimates obtained from a parsimony analysis. For example, estimates of $F(t)$ obtained from the parsimony analysis have often been inaccurately interpreted as estimates of substitution rates ($Q$). The failure to explicitly consider branch lengths in the parsimony analysis also poses difficulties in the interpretation of the parsimony estimates of the substitution pattern. Clearly, $P(t)$ or $F(t)$ is dependent on the amount of evolution as reflected in $t$. If $t$ is close to zero, $P(t)$ will be close to the unit matrix, while if $t$ is very large, $P(t)$ will merely reflect the equilibrium nucleotide frequencies as $P_{ij}(\infty) = \pi_j$. The biases in the estimated rate matrix $Q$ caused by taking the average of many $P(t)$ or $F(t)$ that correspond to different branch lengths are unclear. The same problem exists with the pairwise comparison approach, although the behaviors of the two methods are different. For example, similar branch lengths in the tree are favorable to parsimony, but star-like phylogenies are favorable to the pairwise approach. It may be worthwhile to explore methods for weighting $F(t)$s according to the branch length or sequence distance ($t$).

Because the parsimony inference of ancestral sequences ignores reconstructions that require more changes than the most parsimonious reconstructions, the method underestimates the off-diagonal elements of $F(t)$ or $P(t)$ and thus the amount of evolution ($t$). However, biases in estimates of $Q$ by equation (5) are not very clear. The parsimony reconstruction ignores the biases in substitution rates between nucleotides when the ancestral sequences are inferred, and it may be expected to give more similar estimates of elements of $Q$, or underestimated transition/transversion rate ratio. This seems to be the case for the two data sets analyzed in this paper, although the biases do not seem to be very large. Another difference between the two methods is that the pairwise method does not depend on the phylogeny and involves much less computation.

### Estimation of the Gamma Parameter for Variable Rates Among Sites

The (minimum) numbers of changes at sites inferred by parsimony are underestimates, as the method ignores, at least, multiple substitutions within one lineage. The underestimation is obviously more serious at fast-changing sites than at slow-changing sites. Thus, the method of moments and the method of maximum likelihood of Sullivan et al. (1995) (equations 11 and 12) are expected to underestimate the extent of rate variation and to overestimate $\alpha$. Wakeley (1993) pointed out that parsimony underestimates both the mean and the variance of the number of changes at sites, but the underestimation of the variance is more serious than that of the mean, so $\alpha$ will be overestimated by equation (11). By reinterpreting the number of "changes" inferred by parsimony as the number of "differences," the new method of this paper (equation 17) considerably reduced the bias in parsimony estimates of $\alpha$. However, the new method also overestimates $\alpha$, as the number of differences is more severely underestimated in fast-changing sites than in slow-changing sites. Unequal branch lengths in the tree also causes biases in the estimate of $\alpha$ as an average branch length is used in equation (17). It can be expected that adding sequences to the data to break long branches in the tree and to increase occurrences of different changes will reduce the bias of parsimony estimates of the $\alpha$ parameter.

### Program Availability and Performance

Maximum-likelihood analyses of this paper were performed using the PAML program package, which is distributed by Z. Yang and can be obtained by anonymous file transfer protocol at ftp.bio.indiana.edu under the directory molbio/evolve. Two programs, baseml and aaml, are for analyzing nucleotide and amino acid sequences, respectively. Approximate methods described in this paper were implemented in a separate program (pamp) in the same package. The likelihood calculations for each of the three data sets analyzed in this paper took a few hours on a SUN Sparc Station, while the parsimony calculations took a few seconds on the same machine.

## Acknowledgments

We thank Masatoshi Nei, Jack Sullivan, and the anonymous reviewers for comments on the manuscript. This study was supported by National Institutes of Health and National Science Foundation grants to M. Nei.

### LITERATURE CITED

ADACHI, J., and M. HASEGAWA. 1995. Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J. Mol. Evol. **40**:622–628.

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

BROWN, W. M., E. M. PRAGER, A. WANG, and A. C. WILSON. 1982. Mitochondrial DNA sequences of primates, tempo and mode of evolution. J. Mol. Evol. **18**:225–239.

COLLINS, T. M., P. H. WIMBERGER, and G. J. P. NAYLOR. 1994. Compositional bias, character-state bias, and character-state reconstruction using parsimony. Syst. Biol. **43**:482–496.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1978. A model of evolutionary change in proteins. Pp. 345–352 *in* Atlas of protein sequence and structure, Vol 5, Suppl. 3. National Biomedical Research Foundation, Washington, D. C.

ESPOSTI, M. D., S. DE VRIES, M. CRIMI, A. GHELLI, T. PATARNELLO, and A. MEYER. 1993. Mitochondrial cytochrome *b*: evolution and structure of the protein. Biochimica et Biophysica Acta **1143**:243–271.

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

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

GILLESPIE, J. H. 1986. Rates of molecular evolution. Ann. Rev. Ecol. Syst. **17**:637–665.

GOJOBORI, T., and S. YOKOYAMA. 1987. Molecular evolutionary rates of oncogenes. J. Mol. Evol. **26**:148–156.

GOJOBORI, T., W.-H. LI, and D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:360–369.

GOLDING, G. B. 1983. Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol. Biol. Evol. **1**:125–142.

HARTIGAN, J. A. 1973. Minimum evolution fits to a given tree. Biometrics **29**:53–65.

HEDGES, A. B., S. KUMAR, K. TAMURA, and M. STONEKING. 1992. Human origins and analysis of mitochondrial DNA sequences. Science **255**:737–739.

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

IMANISHI, T., and T. GOJOBORI. 1992. Patterns of nucleotide substitutions inferred from the phylogenies of the class I major histocompatibility complex genes. J. Mol. Evol. **35**: 196–204.

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

JOHNSON, N. J., S. KOTZ, and A. W. KEMP. 1992. Univariate discrete distributions. 2nd ed. Wiley, New York.

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. The rapid generation of mutation data matrices from protein sequences. Comp. Appl. Biosci. **8**:275–282.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–123 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates 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 humans and chimpanzees: control region and protein-coding region. Pp. 391–413 *in* S. OSAWA and T. HONJO, eds. Evolution of life: fossils, molecules and culture. Springer-Verlag, Tokyo.

KUHNER, M. K., and J. FELSENSTEIN. 1994. A simulation of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis, version 1.0. The Pennsylvania State University, University Park, Pa. 16802.

LAKE, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc. Natl. Acad. Sci. USA **91**:1455–1459.

LARSON, A., and A. C. WILSON. 1989. Patterns of ribosomal RNA evolution in salamanders. Mol. Biol. Evol. **6**:131–154.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J. Mol. Evol. **21**:58–71.

MORIYAMA, E. N., Y. INA, K. IHEO, N. SHIMIZU, and T. GOJOBORI. 1991. Mutation pattern of human immunodeficiency virus genes. J. Mol. Evol. **32**:360–363.

PERNA, N. T., and T. D. KOCHER. 1995. Unequal base frequencies and the estimation of substitution rates. Mol. Biol. Evol. **12**:359–361.

RODRIGUEZ, F., J. F. OLIVER, A. MARIN, and J. R. MEDINA. 1990. The general stochastic model of nucleotide substitutions. J. Theor. Biol. **142**:485–501.

RZHETSKY, A., and M. NEI. 1994. Unbiased estimates of the number of nucleotide substitutions when substitution rates varies among different sites. J. Mol. Evol. **38**:295–299.

STEEL, M. 1994. Recovering a tree from the leaf colourations it generates under a Markov model. Appl. Math. Lett. **7**: 19–23.

SULLIVAN, J., K. E. HOLSINGER, and C. SIMON. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmontine rodents. Mol. Biol. Evol. **12**:988–1001.

TAKAHATA, N. 1991. Overdispersed molecular clock at the major histocompatibility complex loci. Proc. R. Soc. Lond. B **243**:13–18.

TAJIMA, T. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **10**:677–688.

TAMURA, K., and M. NEI. 1993. Estimation of the number of nucleotide substitutions 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-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **11**:261–277.

TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Pp. 57–86 *in* Lectures in mathematics in the life sciences. Vol. 17. American Mathematical Society, Providence, R.I.

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

VIGILANT, L., M. STONEKING, H. HARPENDING, K. HAWKES, and A. C. WILSON. 1991. African populations and the evolution of human mitochondrial DNA. Science **253**:1503–1507.

WAINRIGHT, P. O., G. HINKLE, M. L. SOGIN, and S. K. STICKEL. 1993. Monophyletic origins of the metazoa: an evolutionary link with fungi. Science **260**:340–342.

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

———. 1994. Substitution rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. **11**:436–442.

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

———. 1994*a*. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

———. 1994*b*. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1995*a*. A space-time process model for the evolution of DNA sequences. Genetics **139**:993–1005.

———. 1995*b*. On the general reversible Markov-process model of nucleotide substitution: a reply to Saccone et al. J. Mol. Evol. **41**:254–255.

YANG, Z., and T. WANG. 1995. Mixed model analysis of DNA sequence evolution. Biometrics **51**:552–561.

YANG, Z., N. GOLDMAN, and A. E. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol. Biol. Evol. **11**: 316–324.

ZHARKIKH, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. **39**:315–329.

PAUL SHARP and DAN GRAUR, reviewing editors

Accepted January 15, 1996