J Mol Evol (1994) 39:105–111

Journal of Molecular Evolution

Estimating the Pattern of Nucleotide Substitution

Ziheng Yang*

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, United Kingdom

Received: 14 June 1993 / Accepted: 11 November 1993

Abstract. Knowledge of the pattern of nucleotide substitution is important both to our understanding of molecular sequence evolution and to reliable estimation of phylogenetic relationships. The method of parsimony analysis, which has been used to estimate substitution patterns in real sequences, has serious drawbacks and leads to results difficult to interpret. In this paper a model-based maximum likelihood approach is proposed for estimating substitution patterns in real sequences. Nucleotide substitution is assumed to follow a homogeneous Markov process, and the general reversible process model (REV) and the unrestricted model without the reversibility assumption are used. These models are also applied to examine the adequacy of the model of Hasegawa et al. (*J. Mol. Evol.* 1985;22:160–174) (HKY85). Two data sets are analyzed. For the ψη-globin pseudogenes of six primate species, the REV model fits the data much better than HKY85, while, for a segment of mtDNA sequences from nine primates, REV cannot provide a significantly better fit than HKY85 when rate variation over sites is taken into account in the models. It is concluded that the use of the REV model in phylogenetic analysis can be recommended, especially for large data sets or for sequences with extreme substitution patterns, while HKY85 may be expected to provide a good approximation. The use of the unrestricted model does not appear to be worthwhile.

Key words: Substitution patterns — Models — Markov process — Reversible process — Sequence divergence — Maximum likelihood — DNA sequences

**Present address*: Biometrics Section, Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom

Introduction

In comparative analysis of homologous DNA sequences, nucleotide substitution is commonly assumed to follow a (stationary) homogeneous Markov process. The Markov process is specified by a rate matrix, **Q**, whose elements represent instantaneous substitution rates among the four nucleotides. For mathematical simplicity and ease of computation extra restrictions have normally been placed on the structure of **Q**, leading to various parametric models. For example, the model of Jukes and Cantor (1969), designated "JC69," is the simplest in that all the changes among the four nucleotides are assumed to occur with equal probability. Kimura's (1980) model (K80) allows transitions and transversions to occur with different rates, while Felsenstein's (1981) model (F81) allows the four nucleotides to have unequal frequencies at equilibrium. The model of Hasegawa et al. (1985) (HKY85) merits special attention as it allows both different rates for transitions and transversions and different nucleotide frequencies, which is a natural extension to all the above three. Other models include, for example, those of Tamura (1992) and Tamura and Nei (1993); both are very similar to HKY85. Li et al. (1985), Tavare (1986), and Rodriguez et al. (1990) provided reviews on the substitution models available at the time. Such models are used to construct estimators of evolutionary distances in pairwise sequence comparisons and are used in the maximum-likelihood joint comparison of all the sequences. The evolutionary distance between two sequences, either extant or extinct, is then proportional to the time that separates them.

While many models of nucleotide substitution have been proposed in the literature, relatively few attempts

have been made to refine methods for estimating substitution patterns in real sequences. The most commonly used method for this purpose is parsimony analysis, which was originally used to compare protein sequences, for example, to construct the PAM matrix by Dayhoff et al. (1978), and was later used in comparing nucleotide sequences (Gojobori et al. 1982b; Li et al. 1984; Gojobori and Yokoyama 1987; Moriyama et al. 1991; Imanishi and Gojobori 1992). In this approach, the phylogeny connecting the sequences is estimated and the nucleotide compositions at the interior nodes of the estimated tree—that is, sequences of extinct ancestors—are inferred. Nucleotide changes along the tree are then counted to produce a matrix of relative frequencies of changes among the nucleotides.

The above approach, however, has several problems. First, the parsimony method is known to have no time structure; the probability of a substitution occurring in a short time interval is assumed to be the same as that in a long time interval (Thompson 1975; Goldman 1990). Although a tree topology is used in the analysis, branch lengths, which are obviously very important in determining the number of changes along the branches, are not taken into account. Second, the estimated matrix of frequencies of changes has no clear meaning. It is at best a reflection (distorted average) of many different matrices of transition probabilities for different branches in the tree, and therefore depends on the overall amount of evolution. For instance, Tamura and Nei (1993) used this approach to estimate the substitution pattern of the control region of human mitochondrial DNA as a justification of their new model. However, the relationship between the estimated matrix of frequencies of changes on one hand, and the proposed probabilistic model on the other, is not clear. This problem was also ignored by Kishino et al. (1990) when they used Dayhoff et al.'s (1978) PAM matrix as a direct estimate of the instantaneous replacement rates among the amino acids. (See also Wilbur 1985 for criticisms of the PAM matrix.)

Other problems include, third, that parallel or backward substitutions are ignored in the parsimony analysis, and fourth, that the inferred tree topology may not be the correct one. Nevertheless, the third problem has virtually been avoided by choosing only closely related sequences (Dayhoff et al. 1978; Gojobori et al. 1982b). We also note that using a wrong tree will not cause serious errors in estimates of rate parameters (Yang et al. submitted; see also below), and therefore the fourth does not appear to be a big problem either.

In this paper we propose an approach for estimating the rate matrix, **Q**, using the maximum likelihood method, which avoids the first three problems mentioned above. The model to be used is the general reversible Markov process model. An "unrestricted model," which makes no restriction about the structure of **Q**, is also used for comparison. Our second purpose is to examine the adequacy of simpler models, and, if they are not acceptable, to provide a new substitution model for use in the maximum likelihood phylogenetic estimation (Felsenstein 1981). In a recent study we compared the JC69, K80, and F81 models against HKY85 and found that the three simpler models were totally unacceptable for all the datasets analyzed (Yang et al. in press). It would be interesting to examine whether the HKY85 model is acceptable when compared to more general ones.

Markov Process Models of Nucleotide Substitution

The General Reversible Process Model (REV). The rate matrix for a reversible homogeneous Markov process has the following general form

$$Q = \begin{bmatrix} \cdot & a\pi_C & b\pi_A & c\pi_G \\ a\pi_T & \cdot & d\pi_A & e\pi_G \\ b\pi_T & d\pi_C & \cdot & f\pi_G \\ c\pi_T & e\pi_C & f\pi_A & \cdot \end{bmatrix} \quad (1)$$

where the diagonals are given as $Q_{ii} = -\sum_{j \neq i} Q_{ij}$ and the nucleotides are ordered T, C, A, G. $Q_{ij}\Delta t$ $(i \neq j)$ is the probability that nucleotide i will change into nucleotide j in an infinitesimal time interval Δt. It is easy to confirm that π_T, π_C, π_A, π_G $(\pi_T + \pi_C + \pi_A + \pi_G = 1)$ are the equilibrium distribution and that the reversibility condition holds, i.e., $\pi_i Q_{ij} = \pi_j Q_{ji}$. We will call *a, b, c, d, e, f* "rate parameters," and the πs "frequency parameters." It is unclear whether it is biologically reasonable to consider these two sets of parameters as representing different forces that affect nucleotide substitution, but this distinction is mathematically convenient. One of the rate parameters is redundant. Thus f is set to 1 and **Q** is multiplied by a constant so that the average rate of substitution at equilibrium is 1, i.e., $2a\pi_T\pi_C + 2b\pi_T\pi_A + 2c\pi_T\pi_G + 2d\pi_C\pi_A + 2e\pi_C\pi_G + 2f\pi_A\pi_G = 1$. Parameters *a, b, c, d, e* are then "rate ratios," and branch lengths in the tree, or times t, are defined as the expected number of nucleotide substitutions per site accumulated during that time period.

The reversible Markov process model (REV) was introduced into nucleotide sequence analysis by Tavare (1986). Almost all the models proposed in the literature are special forms of (1). Especially, if $a = f = \kappa\mu$ and $b = c = d = e = \mu$, the REV model reduces to HKY85. The substitution model that has been implemented in the DNAML program of Joe Felsenstein's PHYLIP package since 1984 ("F84," Felsenstein pers. comm.) is obtained by setting $a = [1 + \kappa/(\pi_T + \pi_C)]\mu$, $f = [1 + \kappa/(\pi_A + \pi_G)]\mu$, and $b = c = d = e = \mu$ (Goldman pers. comm.). The model of Tamura and Nei (1993) makes only the restriction that $b = c = d = e$; both HKY85 and F84 are special cases of this. An exception is the six-parameter model suggested by Kimura (1981), the correct solution of which was given in Gojobori et al. (1982a). This model, however, is very sensitive to small perturbations in the data and is often inapplicable if the true distance is not small (Yang 1992). The model of Lanave et al. (1984), which was alleged by those authors to be general, also involves the assumption of reversibility (Yang and Goldman in press). This fact was ignored by those authors, which led to inconsistent results; for example, the estimated **Q** matrix in Table 6 of Lanave et al. (1984: 92) did not satisfy the mathematical requirement that the row sums of **Q** should be zero.

The matrix of transition probabilities is then $\mathbf{P}(t) = \exp(-\mathbf{Q}t)$, from which the likelihood for a given tree topology can be calculated following Felsenstein (1981). In this study a general-purpose program is used to calculate the eigenroots and eigenvectors of **Q**, although it is possible to use the algorithm for solving a cubic equation

as one of the four eigenroots is zero. My numerical experiments suggest that the rate matrix for a reversible process has only real eigenroots, although no proof of this assertion is known.

In order to construct a simple formula for estimating the distance between two sequences, the substitution model needs to satisfy two mathematical requirements: (1) the eigenvectors of **Q** are functions of only the frequency parameters and are free from the rate parameters; (2) the number of unknowns, not including the frequency parameters, is the same as the number of nonzero distinctive eigenroots of **Q.** As the frequency parameters can be estimated using the averages of the observed frequencies in the two sequences, these two conditions will ensure as many simple equations as the number of unknowns and hence a simple solution. This conclusion also holds when a known distribution, such as a gamma distribution with known shape parameter, is used to approximate rate variation over sites. The two requirements are met by the JC69, K80, F81 models, the models of Tamura (1992) and Tamura and Nei (1993), and the model underlying Felsenstein's DNAML program (F84). The HKY85 model, however, does not meet the second requirement, as its **Q** matrix has three nonzero eigenroots while there are only two unknowns (the distance and the transition/transversion rate ratio κ). The REV model does not satisfy either of the two requirements, and therefore a simple formula for estimating sequence divergence is not available. In fact, the model of Tamura and Nei (1993) can be expected to be the last which can lead to such a simple solution.

The Unrestricted Model. In principle, a **Q** matrix without the reversibility restriction, called the "unrestricted" model, can be used in a similar way as REV. The model will then be a special case of the parameter-rich model of Barry and Hartigan (1987), who used one general **Q** matrix for each branch in the tree. With the unrestricted model, **Q** may have complex eigenroots, and calculations involving complex numbers have to be carried out to calculate $\mathbf{P}(t)$, though $\mathbf{P}(t)$ itself is real. Furthermore, rooted trees with the same unrooted topology will have different likelihood values even though no assumption of the existence of a molecular clock is made (Felsenstein 1981). However, this may not be expected to be a useful way to root the tree, because, for real data, the likelihood values for those trees will be very similar, and the method will have little discriminating power. Inaccuracies in other aspects of the model may well cause larger differences in likelihood than the tree topology differences. The REV model already contains eight free parameters while an unrestricted **Q** will have 11. The improvement over the REV model by adding three extra parameters is expected to be marginal. I believe that reversibility is a restriction that leads to nice mathematical properties without sacrificing much of the biological reality. However, the unrestricted model will be applied to the same data sets, mainly to show that it cannot be a worthwhile attempt.

Data Analysis

We concentrate on the REV model for estimating the pattern of nucleotide substitution, but use the HKY85 model to perform similar analysis to examine the adequacy of the latter. The unrestricted model is also used to see whether it can provide a better fit than the REV model. Two data sets, one of $\psi\eta$-globin pseudogenes from six primate species and the other of a segment of mtDNA genomes from nine primate species, will be analyzed.

$\psi\eta$-Globin Pseudogenes

The $\psi\eta$-globin pseudogenes of human, chimpanzee, gorilla, orangutan, rhesus monkey, and spider monkey (Miyamoto et al. 1987) are analyzed. There are 6,166 nucleotides in each sequence. The observed nucleotide frequencies are quite homogeneous across species, with averages $\hat{\pi}_T = 0.308$, $\hat{\pi}_C = 0.185$, $\hat{\pi}_A = 0.308$, $\hat{\pi}_G = 0.199$. These values will be used directly as estimates in the HKY85 and REV models. The upper limit of the likelihood (Navidi et al. 1991; Reeves 1992; Goldman 1993) is $\ell_{max} = -13{,}597.68$. Using a smaller dataset which contains only the first four sequences, we previously examined the possible rate variation across nucleotide sites (Yang et al. in press). Assuming gamma-distributed rates over sites did not lead to extremely significant improvement over a model assuming a single rate ($P > 0.01$), suggesting that substitution rates at different sites are more or less equal. The HKY85 model was used in the comparison. In an even earlier study using six sequences of the same gene, each 2,040 bases long, the fit to the data by the HKY85 model is found to be acceptable (Goldman 1993). In sum, both HKY85 and REV can be expected to be good candidates for describing the evolution of these sequences.

Tree estimation is not the purpose of this study. However, the three trees concerning the human-chimpanzee-gorilla separation, with all the others as outgroups in the order of orangutan, rhesus monkey, spider monkey, are evaluated to examine the effect of errors in tree estimation on the estimation of rate parameters in the REV model. The best branching order among these three is ((human, chimpanzee), gorilla) (Fig. 1). The estimated rate matrix obtained from this tree is shown in Table 1. The likelihood is $\ell = \ln(L) = -13{,}803.63$, with parameter estimates $\hat{a} = 0.987 \pm 0.075$, $\hat{b} = 0.110 \pm 0.015$, $\hat{c} = 0.218 \pm 0.027$, $\hat{d} = 0.243 \pm 0.030$, and $\hat{e} = 0.395 \pm 0.048$. (Standard errors are estimated by the curvature method.) Results obtained using the other tree topologies confirm our previous observation that a wrong tree can be used to get reliable estimates of rate parameters (Yang et al. in press). For example, the estimated **Q** matrix from the six-species star tree (Table 1) is very similar to that obtained from the best tree.

The HKY85 model also chooses the same best tree from the three, with likelihood $\ell = -13{,}833.92$. The transition/transversion rate ratio is estimated as $\hat{\kappa} = 4.79 \pm 0.32$—this is an estimate of $1/b$ under the restrictions $a = f = 1$ and $b = c = d = e$ in the REV model. (See above.) Estimates of branch lengths are roughly the same under the two models. The likelihood ratio test means comparison of $2\Delta\ell = 2 \times 30.29 = 60.58$ with $\chi^2_{.001} = 13.28$ with df = 4, and the difference is significant. Estimates of rate parameters from REV also suggest that HKY85 is not describing the data very well. In particular, the assumption $b = c = d = e$ is unrealistic.

We use the unrestricted model to fit the data, assuming the same best (unrooted) tree. The "root"—that is, the starting point for the calculation—is (arbitrarily)

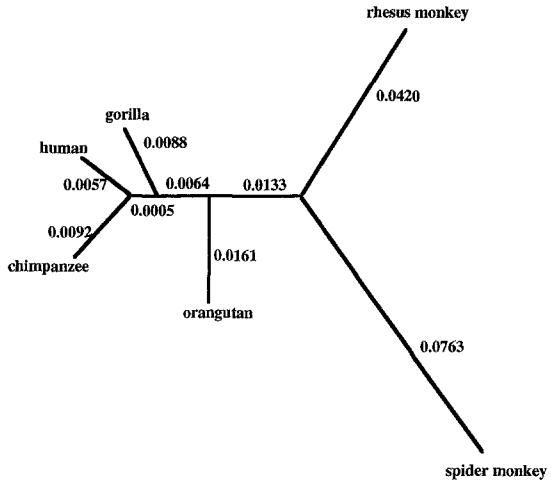

Fig. 1. Estimates of branch lengths for the ψη-globin pseudogenes. The REV model of nucleotide substitution is assumed. The averages of nucleotide frequencies among the species are $\hat{\pi}_T = 0.308$, $\hat{\pi}_C = 0.185$, $\hat{\pi}_A = 0.308$, $\hat{\pi}_G = 0.199$ and they are used as estimates for the model. The rate parameters are estimated by iteration, giving $\hat{a} = 0.987$, $\hat{b} = 0.110$, $\hat{c} = 0.218$, $\hat{d} = 0.243$, $\hat{e} = 0.395$. [See equation (1).] The estimated rate matrix is shown in Table 1. The likelihood of this tree is $\ell = -13{,}803.63$.

Table 1. Estimates of the rate matrix **Q** for the primate ψη-globin pseudogenes (6,166 nucleotides)[a]

(1) Using the best tree and the REV model			
−0.765	0.537	0.100	0.128
0.897	−1.349	0.221	0.231
0.100	0.132	−0.818	0.586
0.198	0.215	0.909	−1.322
(2) Using the six-species star tree and the REV model			
−0.771	0.534	0.103	0.135
0.890	−1.337	0.220	0.227
0.102	0.131	−0.817	0.583
0.209	0.211	0.905	−1.325
(3) Using the best tree and the unrestricted model			
−0.765	0.548	0.089	0.128
0.881	−1.351	0.220	0.250
0.111	0.133	−0.814	0.570
0.197	0.199	0.935	−1.332

[a] The element of the matrix, Q_{ij} $(i \neq j)$ is the rate of substitution from nucleotide i to j. The matrix is scaled so that the average rate in equilibrium is 1

set at the node connecting rhesus monkey, spider monkey, and the group of all the other species (Fig. 1). Branch lengths and all the 11 parameters in the **Q** matrix are estimated by iteration. The estimated rate matrix is shown in Table 1, which is very similar to that obtained by using the REV model. The likelihood is now −13,802.80, with only slight improvement over REV ($2\Delta\ell = 2 \times 0.83 = 1.66$, $\chi^2_{001} = 11.35$, df = 3). The equilibrium distribution estimated from the unrestricted model is $\hat{\pi}_T = 0.308$, $\hat{\pi}_C = 0.185$, $\hat{\pi}_A = 0.310$, $\hat{\pi}_G = 0.197$, virtually the same as the observed frequencies.

mtDNA Sequences

The second dataset consists of a segment of mitochondrial genomes of human, chimpanzee, gorilla, orangutan, gibbon, crab-eating macaque, squirrel monkey, tarsier and lemur, which is an expanded dataset of that of Brown et al. (1982). The sequences were aligned by eye by Adrian Friday, and the length of sequence is 888 after sites involving insertions or deletions are excluded. The most "reasonable" (unrooted) tree separates the species in the order human, chimpanzee, gorilla, orangutan, gibbon, crab-eating macaque, squirrel monkey, tarsier, and lemur (Fig. 2). Nucleotide frequencies in different species are similar but some systematic differences are apparent (Table 2). The frequencies of G are quite homogeneous, while, compared to the average values, those of T and A are higher and those of C are lower in the squirrel monkey, tarsier, and lemur sequences. The orangutan sequence has lower T and higher C frequencies. This suggests that the stationarity and homogeneity assumptions may not be satisfied in this dataset. We therefore estimate the frequency parameters for the HKY85 and REV models by iteration, so that a fair comparison with the unrestricted model can be performed. As the frequencies are not very different among species, we expect that the comparison of models will not be biased too much.

Another problem with this dataset is that substitution rates are highly variable across nucleotide sites, as the beginning and ending parts of this segment code for parts of two proteins respectively and the middle part codes for three tRNAs (Brown et al. 1982). The method of Yang (1993), which uses a gamma distribution to describe the rate variation over sites, is computationally unfeasible for datasets with more than a few species. To account for such rate variation, we will instead make use of a "discrete gamma" model. The continuous gamma distribution is "discretized" into k categories each of equal probability, and for each category the mean of that portion of the distribution is used to represent the rates in that category. For several datasets, $k = 4$ is found to give both an optimum or near-optimum fit to data and an acceptable approximation to the continuous gamma distribution. This value will be used in this study. Full details of the method will be described elsewhere; here we use it to compare the three models of nucleotide substitution.

We first fit the models under the assumption that the rate is constant over sites. The most reasonable tree (Fig. 2) is assumed. The HKY85 model produces the following estimates of parameters: $\hat{\pi}_T = 0.286$, $\hat{\pi}_C = 0.297$, $\hat{\pi}_A = 0.313$, $\hat{\pi}_G = 0.103$, $\hat{\kappa} = 4.17$, with $\ell = -5{,}232.71$. Using the REV model leads to the following results: $\hat{\pi}_T = 0.291$, $\hat{\pi}_C = 0.275$, $\hat{\pi}_A = 0.305$, $\hat{\pi}_G = 0.130$, and $\hat{a} = 1.368$, $\hat{b} = 0.249$, $\hat{c} = 0.029$, $\hat{d} = 0.506$,

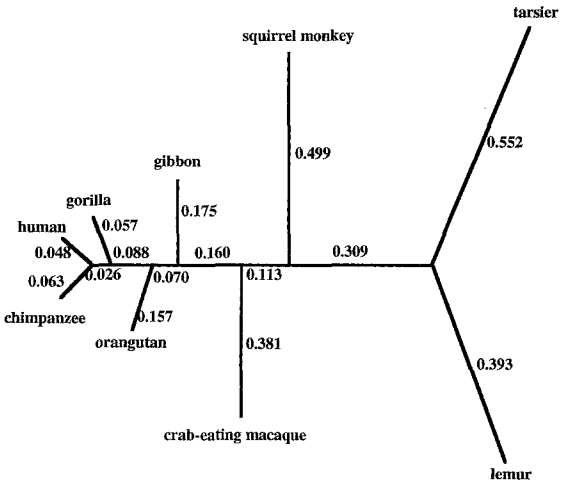

Fig. 2. Estimates of branch lengths for the mtDNA sequences. A discrete gamma model is used to account for rate variation over sites while the substitution pattern is assumed to follow the HKY85 model. All the parameters in the model are estimated by iteration, giving $\hat{\pi}_T = 0.241$, $\hat{\pi}_C = 0.319$, $\hat{\pi}_A = 0.356$, $\hat{\pi}_G = 0.084$, $\hat{\kappa} = 10.60$, $\hat{\alpha} = 0.36$, with $\ell = -5{,}030.80$.

$\hat{e} = 0.149$, with $\ell = -5{,}189.32$. The estimated frequency parameters are quite different from the averages of the observed ones, and this is found to have a large effect on the likelihood values; for example, using the observed averages (Table 2) would give $\ell = -5{,}197.73$ under REV. The estimated **Q** matrix from the REV model is given in Table 3, showing the effect of ignoring rate variation over sites on the estimation of rate parameters. The estimate of the rate matrix from the unrestricted model, which is not shown, is similar to that obtained from the REV model, with the equilibrium distribution estimated as $\hat{\pi}_T = 0.291$, $\hat{\pi}_C = 0.275$, $\hat{\pi}_A = 0.305$, $\hat{\pi}_T = 0.130$; these estimates are identical to those obtained from REV at the third decimal points. The likelihood is $\ell = -5{,}187.86$. The "root" of the (unrooted) tree has been placed to the node that connects tarsier, lemur, and the group of the remaining species (Fig. 2). Using the likelihood ratio test, HKY85 would be rejected when compared to REV ($2\Delta\ell = 2 \times 43.39 = 86.78$, $\chi^2_{001} = 13.28$ with df = 4), while the unrestricted model does not give a better fit than REV ($2\Delta\ell = 2 \times 1.46 = 2.93$, $\chi^2_{001} = 11.35$ with df = 3). However, as rates are known to be variable across sites, these comparisons cannot be expected to be reliable.

We now use the discrete gamma model to account for variable rates over sites and estimate the shape parameter of the gamma distribution, α, from the data. Estimates of parameters from the HKY85 + Gamma model are $\hat{\pi}_T = 0.241$, $\hat{\pi}_C = 0.319$, $\hat{\pi}_A = 0.356$, $\hat{\pi}_G = 0.084$, $\hat{\kappa} = 10.60 \pm 1.43$, $\hat{\alpha} = 0.36 \pm 0.04$, with $\ell = -5{,}030.80$. When rate variation over sites was ignored in the model, κ was seriously underestimated ($\hat{\kappa} = 4.17$ from HKY85), as noted by Yang et al. (in press). Estimates of the frequency parameters are also very different from HKY85 without using the gamma distribution. Results obtained from the REV + Gamma model are $\hat{\pi}_T = 0.250$, $\hat{\pi}_C = 0.315$, $\hat{\pi}_A = 0.348$, $\hat{\pi}_G = 0.087$, and $\hat{a} = 0.961 \pm 0.215$, $\hat{b} = 0.091 \pm 0.028$, $\hat{c} = 0.012 \pm 0.019$, $\hat{d} = 0.133 \pm 0.037$, $\hat{e} = 0.087 \pm 0.038$, and $\hat{\alpha} = 0.385 \pm 0.042$, with $\ell = -5{,}026.86$. The estimated **Q** matrix is shown in Table 3. Estimates of both the rate parameters and the frequency parameters are different from those obtained when a single rate over sites was assumed, although the whole rate matrix looks more similar. The estimated **Q** matrix obtained by using the nine-species star tree is very similar to that obtained by using the best tree (Table 3). It is apparent that estimation of the substitution pattern is affected much more by ignoring rate variation over sites than by assuming a wrong tree. The unrestricted model, when combined with the gamma distribution of rates over sites, produced very similar results to those from REV + Gamma (Table 3). The equilibrium distribution is estimated as $\hat{\pi}_T = 0.250$, $\hat{\pi}_C = 0.315$, $\hat{\pi}_A = 0.348$, $\hat{\pi}_G = 0.087$, and $\hat{\alpha} = 0.385$, being identical to those obtained from the REV + Gamma model at this level of accuracy. The likelihood is $\ell = -5{,}026.52$. REV + Gamma is not significantly better than HKY85 + Gamma ($2\Delta\ell = 2 \times 3.95 = 7.89$, $\chi^2_{001} = 13.28$ with df = 4). The unrestricted model also gives trivial improvement over REV ($2\Delta\ell = 2 \times 0.34 = 0.68$) or HKY85 ($2\Delta\ell = 2 \times 4.29 = 8.58$, $\chi^2_{001} = 18.48$ with df = 7). Estimates of branch lengths under HKY85 + Gamma are shown in Fig. 2.

Table 2. The observed nucleotide frequencies in the mtDNA sequences (888 nucleotides)

	T	C	A	G
Human	0.2579	0.3300	0.3041	0.1081
Chimpanzee	0.2658	0.3232	0.3086	0.1025
Gorilla	0.2579	0.3255	0.3097	0.1070
Orangutan	0.2376	0.3446	0.3142	0.1036
Gibbon	0.2511	0.3187	0.3153	0.1149
Crab-eating macaque	0.2680	0.3063	0.3187	0.1070
Squirrel monkey	0.2804	0.2646	0.3378	0.1171
Tarsier	0.2905	0.2545	0.3469	0.1081
Lemur	0.2849	0.2725	0.3423	0.1002
Average	0.2660	0.3044	0.3220	0.1076

Discussion

In a traditional statistical setting, the overall adequacy of a model, or its goodness of fit, can be examined by comparing its likelihood with the upper limit, ℓ_{max}. However, naive use of the χ^2 approximation in this context can be quite misleading due to the many possible site patterns that simply do not appear in the data or appear with very low frequencies (Reeves 1992; Goldman 1993). Such a test is therefore not performed here. Based on previous analyses using more rigorous tests

Table 3. Estimates of the rate matrix **Q** for the primate mtDNA sequences (888 nucleotides)[a]

(1) Using the best tree and the REV model			
−1.037	0.856	0.173	0.009
0.905	−1.300	0.351	0.044
0.165	0.317	−0.776	0.295
0.019	0.093	0.694	−0.806
(2) Using the best tree and the REV + Gamma model[b]			
−1.278	1.154	0.121	0.004
0.916	−1.122	0.176	0.029
0.086	0.159	−0.579	0.333
0.011	0.105	1.329	−1.445
(3) Using the nine-species star tree and the REV + Gamma model[b]			
−1.324	1.248	0.073	0.002
0.943	−1.040	0.082	0.016
0.047	0.069	−0.542	0.425
0.007	0.061	1.903	−1.970
(4) Using the best tree, with the unrestricted + Gamma model[b]			
−1.279	1.147	0.124	0.009
0.919	−1.118	0.173	0.026
0.087	0.160	−0.580	0.333
0.000	0.112	1.339	−1.451

[a] See the note to Table 1

[b] Four equal-probable categories as an approximation to the continuous gamma distribution are used to account for variable rates over sites

(Yang et al. in press), it can be expected that the fit of the REV model to the pseudogene data is statistically acceptable.

Application of the χ^2 approximation to comparison of two parametric models, such as the comparison between HKY85 and REV, appears to be quite reliable and the results are found to be consistent with parameter estimates. In fact, comparison of two parametric models appears to be more powerful than the overall goodness-of-fit test. This may be the main reason why the HKY85 model was not rejected for the ψη-globin pseudogenes by the overall test (Goldman 1993), while it is when compared against the REV model.

For the mtDNA sequences, HKY85 + Gamma appears to be acceptable, as concluded by Yang et al. (in press) when an overall goodness-of-fit test was applied to a subset of the present data containing the first four sequences. The extreme rate variation over sites is manifest from the tremendous improvement in likelihood by adding the gamma distribution (only one parameter) to either HKY85 or REV or the unrestricted model. Indeed, when rate variation over sites is properly accounted for in the model, neither REV nor the unrestricted model can give much improvement over HKY85. While HKY85 + Gamma can be expected to have described many of the characteristics of the evolution of these sequences, we do notice some peculiarities: for example, some of the estimated frequency parameters lie outside the range of the observed ones (Table 2). This seems to suggest that the patterns of substitution are not quite the same along different lineages, while all the models considered here assume the same substitution pattern for all the lineages.

The approach taken in this study is statistical model fitting, pioneered by Ritland and Clegg (1987) in the context of phylogenetic analysis. It is unfortunate that many phylogenetic analyses appear to have paid little attention to what assumptions concerning the evolutionary process are being made. For some methods based principally on intuitive arguments, such as parsimony analyses, it is not even clear what assumptions are made. In a few cases where some aspects of the assumptions were examined, commonly used models were found to be totally unacceptable (Reeves 1992; Goldman 1993). Recent studies suggest that phylogenetic estimation can be substantially affected by the model assumed in the analysis: different models may support different tree topologies; estimates of branch lengths are particularly sensitive; evaluation of the reliability of the estimated tree, by whatever methods, can be quite misleading if the model is wrong (Yang et al. in press and unpublished results).

I believe that the models examined in this paper and those commonly used in phylogenetic analysis are "descriptive" rather than "interpretative." For example, on their own they do not tell whether the process of nucleotide substitution is mainly driven by mutation or selection. The estimate of a branch length in the tree may be better interpreted as an average, over time, of a variable rate than a reflection of a constant rate. Models used in phylogenetic analysis have commonly been formulated at the level of nucleotide substitution, which is the observed product of a complicated process driven by many factors, notably mutation and selection, the effects of which are still not well understood or cannot be accurately measured in practice. However, this limitation does not mean that there is no pattern of nucleotide substitution at all. Neither does it justify ambiguous, incomplete formulations or inadequate analyses, for they may lead to uninterpretable or misleading results. It is apparent that an adequate description of the substitution process is essential to an understanding of its underlying mechanisms. It is my belief that we can gain insights into molecular sequence evolution by rejecting wrong models and constructing more realistic ones, using knowledge of the biology of the sequences.

Acknowledgments. I wish to thank Tianlin Wang for sending me his C programs for calculating the eigenroots and eigenvectors of a general real matrix. I am grateful to Adrian Friday for preparing the mtDNA sequence data, and for informing me of the biologically reasonable phylogenies linking those species. I thank Nick Goldman for discussions and comments on an earlier version of the manuscript. I was supported by a grant from the Department of Zoology, The Natural History Museum (London), during the revision of the manuscript, when substantial modifications were made.

References

Barry D, Hartigan JA (1987) Statistical analysis of hominoid molecular evolution. Stat Sci 2:191–210

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates, tempo and mode of evolution. J Mol Evol 18:225–239

Dayhoff MO (1978) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington, DC, pp 347

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Gojobori T, Yokoyama S (1987) Molecular evolutionary rates of oncogenes. J Mol Evol 26:148–156

Gojobori T, Ishii K, Nei M (1982a) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotides. J Mol Evol 18:414–423

Gojobori T, Li WH, Graur D (1982b) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18: 360–369

Goldman N (1990) Maximum likelihood inference of phylogenetic trees, with special reference to Poisson process models of DNA substitution and to parsimony analysis. Syst Zool 39:345–361

Goldman N (1993) Statistical tests of models of DNA substitution. J Mol Evol 36:182–198

Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Imanishi T, Gojobori T (1992) Patterns of nucleotide substitutions inferred from the phylogenies of the class I major histocompatibility complex genes. J Mol Evol 35:196–204

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

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

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78: 454–458

Kishino H, Miyata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 31:151–160

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86-93

Li W-H, Wu C-I, Luo C-C (1984) Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. J Mol Evol 21:58–71

Li W-H, Wu C-I, Luo C-C (1985) Evolution of DNA sequences. In: MacIntyre J (ed) Molecular evolutionary genetics. Plenum Press, New York, pp 1–94

Miyamoto MM, Slighton JL, Goodman M (1987) Phylogenetic relations of humans and African apes from DNA sequences in the ψη-globin region. Science 238:369–373

Moriyama EN, Ina Y, Iheo K, Shimizu N, Gojobori T (1991) Mutation pattern of human immunodeficiency virus genes. J Mol Evol 32:360–363

Navidi WC, Churchill GA, von Haeseler A (1991) Methods for inferring phylogenies from nucleotide acid sequence data by using maximum likelihood and linear invariants. Mol Biol Evol 8: 128–143

Reeves JH (1992) Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. J Mol Evol 35:17–31

Ritland K, Clegg MT (1987) Evolutionary analysis of plant DNA sequences. Am Nat 130:S74–S100

Rodriguez F, Oliver JF, Marin A, Medina JR (1990) The general stochastic model of nucleotide substitutions. J Theor Biol 142: 485–501

Tamura K (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G + C-content biases. Mol Biol Evol 9:678–687

Tamura K, Nei M (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

Tavare S (1986) Some probabilistic and statistical problems on the analysis of DNA sequences. In: Lectures in mathematics in the life sciences, vol 17. pp 57–86

Thompson E (1975) Human evolutionary trees. Cambridge University Press, Cambridge

Wilbur WJ (1985) On the PAM matrix model of protein evolution. Mol Biol Evol 2:434–447

Yang Z (1992) Variations of substitution rates and estimation of evolutionary distances of DNA sequences. PhD Thesis, Beijing Agricultural University, Beijing

Yang Z (1993) Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10:1396–1401

Yang Z, Goldman N (in press) Evaluation and extension of Markov process models of nucleotide substitution. Acta Genetica Sinica

Yang Z, Goldman N, Friday AE (in press) Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation. Mol Biol Evol