A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences

Nick Goldman and Ziheng Yang†*

*Laboratory of Mathematical Biology, National Institute for Medical Research; and †Biometrics Section, Department of Zoology, The Natural History Museum

A codon-based model for the evolution of protein-coding DNA sequences is presented for use in phylogenetic estimation. A Markov process is used to describe substitutions between codons. Transition/transversion rate bias and codon usage bias are allowed in the model, and selective restraints at the protein level are accommodated using physicochemical distances between the amino acids coded for by the codons. Analyses of two data sets suggest that the new codon-based model can provide a better fit to data than can nucleotide-based models and can produce more reliable estimates of certain biologically important measures such as the transition/transversion rate ratio and the synonymous/nonsynonymous substitution rate ratio.

Introduction

The importance of having accurate models for the evolution of molecular sequences is increasingly clear. First, accurate models can help clarify some of the most important processes of evolution, by the biological interpretation of their parameters. Second, it is becoming apparent that the reliability of phylogenetic analyses is strongly dependent on having accurate models (Thorne et al. 1991, 1992; Yang et al. 1994). Until recently, very few attempts were made to test the adequacy of models. This is no longer the case, as statistical tests have been described and improved on until they are accurate and practical (Ritland and Clegg 1987; Navidi et al. 1991; Reeves 1992; Goldman 1993*a*, 1993*b*; Yang et al. 1994), and consequently it can no longer be acceptable for models to be used without test and improvement when this is needed and is possible. Virtually all studies that have used appropriate tests to check the fit of models for the evolution of coding DNA and protein sequences have found the models to be statistically unacceptable (Reeves 1992; Goldman 1993*b*).

Previous models for the evolution of protein-coding sequences that are used in phylogenetic analysis work either on the mononucleotide level in DNA sequences (e.g., see Jukes and Cantor 1969; Kimura 1980; Felsenstein 1981; Hasegawa et al. 1985) or on the amino acid level in protein sequences (e.g., see Kishino et al. 1990). The unit of data under consideration is either the nucleotide or the amino acid, and these nucleotides or amino acids are assumed to evolve independently. At the DNA level more information is present, and closely related sequences can be more easily distinguished, e.g., by substitutions that are synonymous and thus invisible at the amino acid level (Miyata and Yasunaga 1980; Li et al. 1985; Nei and Gojobori 1986; Li 1993). For more-distantly related species, amino acid sequences might be more useful than DNA: the translation of DNA to amino acids by using knowledge of the genetic code may act as a filter in which some stochastic noise is removed. Information is lost, too (Li et al. 1985), and it is not clear at what levels of divergence the removal of noise might outweigh the loss of information and increase the accuracy of phylogenetic estimation.

Obviously, nucleotide sequence studies should be preferable (Miyata and Yasunaga 1980), and to this end we have devised a model of nucleotide substitution that uses simultaneously the nucleotide-level information in DNA sequences and knowledge of the genetic code and hence the amino acid–level information of synonymous (silent) and nonsynonymous (replacement) nucleotide substitutions. This is achieved by modeling at the codon level, instead of at the nucleotide or amino acid levels. In addition to the advantage that more of the information in the sequences is used, this model enables us to build more-realistic models incorporating previously ignored effects, e.g., the lack of independence at neighboring sites within a codon triplet and differences in evolutionary rate at different codon positions.

Key words: models, nonsynonymous substitutions, nucleotide substitution, phylogenetic estimation, protein-coding sequences, synonymous substitutions.

Address for correspondence and reprints: Dr. Nick Goldman, Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom. E-mail: n_goldma@nimr.mrc.ac.uk

Mol. Biol. Evol. 11(5):725–736. 1994.

0737-4038/94/1105-0002$02.00

In this paper we characterize our codon-based model, which incorporates biologically meaningful factors such as transition/transversion bias, variability of a gene (e.g., as indicated by synonymous and nonsynonymous substitutions rates), and amino acid differences. We describe the model's use in maximum-likelihood (m.l.) estimation of phylogenies and apply the model to two sample data sets. The results are discussed in comparison with previous nucleotide-based models, and we emphasize the importance of m.l. parameter estimation and testing of the adequacy of models as tools for understanding molecular-sequence evolution.

Methods

A Markov-Process Model of Codon Substitution

Our model is suitable for homologous protein-coding DNA sequences with no gaps (insertions/deletions) or with such gaps removed. It does not incorporate any processes for insertion or deletion. While these are part of the biological process of sequence evolution and should ideally be included in models in order to use the information contained in patterns of gaps (Thorne et al. 1991, 1992), this would add greatly to the complexity of analysis, and we have chosen to omit it in this initial study. It is our hope that if the new model proves useful it could be modified to allow for insertions and deletions.

We use a continuous-time Markov process to model substitutions among the codons within a protein-coding sequence. The general genetic code is used in this study, but the principle can be applied to other genetic codes, such as the code for the mammalian mitochondrial genome. The states of the Markov process are the 61 sense codons. The three nonsense (stop) codons are not considered in the model, as mutations to or from stop codons can be assumed to affect drastically the structure and function of the protein and therefore will rarely survive. We use a 61×61 rate matrix $Q = (Q_{ij})$, where $Q_{ij}\Delta t$ ($i \neq j$) represents the probability that codon i will change to codon j in a small time interval Δt. (Throughout this paper, codons are written as i taking values from 1 to 61, corresponding to the triplet $i_1 i_2 i_3$, where each of i_1, i_2, and i_3 is a nucleotide A, C, G, or T). In the usual manner, elements Q_{ii} are fixed so that the row sums of Q equal zero, allowing the solution

$$P(t) = \exp(tQ) \qquad (1)$$

for the matrix $P(t)$ whose elements $P_{ij}(t)$ give the probabilities that codon j replaces codon i after time t (Cox and Miller 1977).

We assume that mutations occur at the three codon positions independently, and therefore only single-nucleotide substitutions are permitted to occur instantaneously, as mutations involving more than one position (as well as those involving more than one change at one position) will have probabilities of occurrence of order $(\Delta t)^2$ and should be ignored. However, from equation (1), substitution between any two codons is possible for any $t > 0$. Each codon has at most nine "neighbors" to which it may change instantaneously (fig. 1). The rate at which each particular allowed substitution occurs is proportional to the (equilibrium) frequency (π_j) of the codon (j) being changed to; this allows either different nucleotide frequencies at the three codon positions or codon usage information to be incorporated in analyses. Rates of substitutions involving a transition (A↔G or C↔T) are multiplied by a factor κ, as with the nucleotide-based models of Kimura (1980) and Hasegawa et al. (1985). The parameter κ directly affects the ratio of transition and transversion substitutions and is incorporated to allow for the empirical finding that transitions often occur more frequently than do transversions (Brown et al. 1982).

To account for selective restraints at the amino acid level, substitution rates are further modified by a multiplicative factor if the two codons code for different

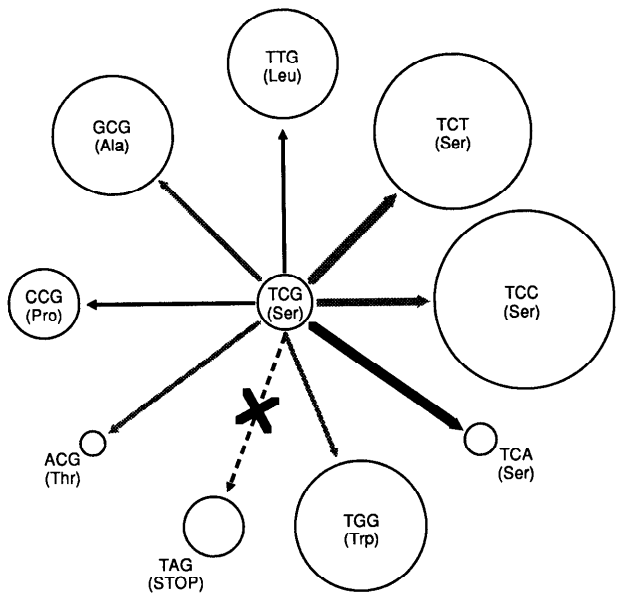

FIG. 1.—Example of the "neighbors" to which a codon (here TCG) may evolve instantaneously through a nucleotide substitution at one position. TCG has eight neighbors, substitution of A for C at the second position being disallowed, as it results in the stop codon TAG. Transitions are marked with black arrows, transversions with gray arrows. Substitutions involving no change in amino acid (generally occurring at a higher rate in this model) are marked with thicker arrows. The size of each circle (except the stop codon TAG) represents the (equilibrium) frequency of that codon, in this case taken from the pooled α- and β-globin gene sequences.

amino acids. In defining these factors, we use the matrix of physicochemical distances between the 20 amino acids given by Grantham (1974). These distances are based on the side-chain composition, polarity, and molecular volume of the amino acids and range from 5 for the pair Ile-Leu to 210 for the pair Cys-Phe, with an average of 100. We multiply the rate of substitution from codon i to codon j by $\exp(-d_{aa_i,aa_j}/V)$, where, following Li et al. (1985), aa_i represents the amino acid coded for by codon i, and d_{aa_i,aa_j} is the distance between amino acids aa_i and aa_j, as given by Grantham (1974). We set $d_{aa_i,aa_i} = 0$. V is a parameter representing the variability of the gene or its tendency to undergo nonsynonymous substitution: V is negatively correlated with the synonymous/nonsynonymous substitution rate ratio. Codons that code for very different amino acids will rarely substitute for one another (large d_{aa_i,aa_j} gives a low exponential term), whereas more-similar amino acids replace each other more readily (small d_{aa_i,aa_j} gives a relatively high exponential term). The term $\exp(-d_{aa_i,aa_j}/V)$ also implicitly enables the model to account for different substitution rates and nonindependence of substitutions at different codon positions; for example, changes at third positions are less likely to lead to amino acid alterations than are changes at first or second positions.

Last, elements Q_{ii} are fixed so that row sums of Q equal zero, and Q is scaled so that the average rate of substitution at equilibrium equals 1:

$$-\sum_{i=1}^{61} \pi_i Q_{ii} = 1 . \qquad (2)$$

This scaling means that time t and branch lengths in a tree are effectively measured as expected numbers of nucleotide substitutions per codon.

Formally, for codons $i = i_1 i_2 i_3$ and $j = j_1 j_2 j_3$ ($i \neq j$), we set

$$Q_{ij} = \begin{cases} 0 & \text{if 2 or 3 of the pair } (i_1,j_1),(i_2,j_2),(i_3,j_3) \text{ are different} \\ \mu\pi_j \cdot \exp(-d_{aa_i,aa_j}/V) & \text{if exactly 1 of the pairs } (i_1,j_1),(i_2,j_2),(i_3,j_3) \text{ is different, and that difference is a transversion} \\ \mu\kappa\pi_j \cdot \exp(-d_{aa_i,aa_j}/V) & \text{if exactly 1 of the pairs } (i_1,j_1),(i_2,j_2),(i_3,j_3) \text{ is different, and that difference is a transition} \end{cases} \qquad (3)$$

and $Q_{ii} = -\sum_{j \neq i} Q_{ij}$. The scaling factor μ is chosen to satisfy equation (2).

The Markov process specified by equation (3) is reversible, as Q is a special form of the rate matrix for a general reversible process (Yang 1994), and the equilibrium distribution is given by the frequencies π_j; that is, $\pi_i Q_{ij} = \pi_j Q_{ji}$ for any i and j. Furthermore, Q appears to have no complex eigenvalues and eigenvectors, as claimed (but not proved) by Yang (1994). We use standard numerical algorithms for the diagonalization (eigensolution) of Q, to calculate $P(t)$ in equation (1).

To demonstrate the effect of the parameter V, we calculate the ratio of the instantaneous rates (per codon) of synonymous and nonsynonymous nucleotide substitutions. The synonymous substitution rate per codon is

$$\rho_s = \sum_{i=1}^{61} \sum_{\substack{j=1 \\ j \neq i \\ aa_i = aa_j}}^{61} \pi_i Q_{ij} . \qquad (4)$$

Similarly, the nonsynonymous rate per codon ρ_a can be calculated by summing $\pi_i Q_{ij}$ over all codons i, j coding for different amino acids. In fact, since the overall rate is normalized by equation (2), we have $\rho_a = 1 - \rho_s$. This allows the calculation of the ratio ρ_s/ρ_a as a function of V and of κ and the codon frequencies π_i. An example is shown in the left-hand panel of figure 2, which uses the observed codon frequencies of the α- and β-globin genes studied below. Values as $V \to \infty$, i.e., $\exp(-d_{aa_i,aa_j}/V) = 1$, are equal to expected values of ρ_s/ρ_a if there are no selective effect caused by amino acid differences and if synonymous and nonsynonymous substitutions occur at equal rates.

The relationship of the parameter κ to the ratio of instantaneous transition rate to transversion rate, T_s/T_v, is shown similarly, by summing $\pi_i Q_{ij}$ over all codons i, j differing by a transition (T_s) and differing by a transversion (T_v). (Again, $T_v = 1 - T_s$.) An example is shown in the right-hand panel of figure 2.

Figure 2 clarifies the relationship between parameters V and κ and the biologically meaningful quantities ρ_s/ρ_a and T_s/T_v. We draw attention to the interpretation of V as a measure of the variability of the gene: high values of V giving low ρ_s/ρ_a (high variability gives relatively many replacement substitutions) and vice versa. We prefer the use of these instantaneous rate ratios to the use of estimates of the *numbers* of silent/replacement substitutions or transitions/transversions derived from "most parsimonious" character-state reconstructions, which will almost certainly underestimate numbers of changes.

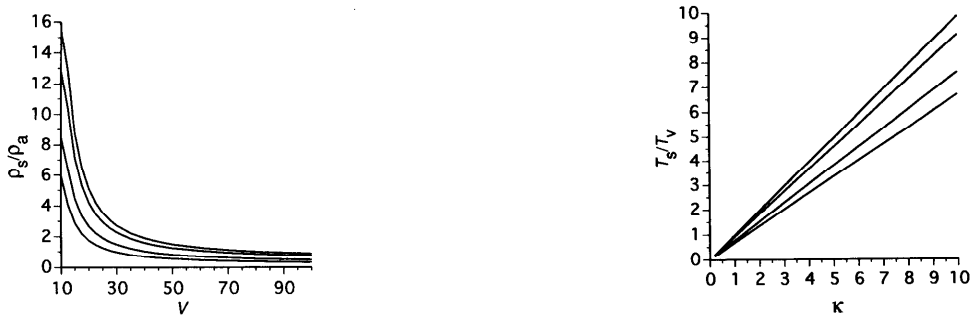

FIG. 2.—*Left,* Ratio of synonymous to nonsynonymous substitution rates per codon, ρ_s/ρ_a, plotted as a function of the parameter V, for various values of the parameter κ. The curves are for (bottom to top) κ = 0.2, 1, 4, and 10. *Right,* Transition/transversion rate ratio, T_s/T_v, plotted as a function of the parameter κ, for various values of the parameter V. The curves are for (top to bottom) V = 10, 30, 50, and 80. Both types of ratios depend also on the frequency parameters π_i, which in these examples were taken from the observed codon frequencies for the α- and β-globin gene data set (see text for details).

From the probabilities $P(t)$ given by equation (1), the log-likelihood for a given tree topology can be calculated following Felsenstein (1981). The differences are that we have now 61 states in the Markov process instead of 4, that one "site" means one codon (triplet), and that much more computation is involved. As usual, we can obtain estimates of $\pi_1, \pi_2, \ldots, \pi_{61}$ from the averages of the observed codon frequencies. The other parameters, κ, V, and branch lengths in the tree, are estimated by maximizing the likelihood, and standard deviations may be estimated by the curvature method (Felsenstein 1981). The same process can be repeated for other tree topologies, and the tree corresponding to the highest likelihood is the m.l. tree. As we assume that the substitution process is homogeneous and stationary, and as we do not assume the existence of a molecular clock (i.e., we do not assume rate constancy over lineages), only unrooted trees can be estimated (Felsenstein 1981; Goldman 1991). As with almost every phylogenetic estimation method, the assumptions of homogeneity and stationarity are made more for expediency than for their accurate reflection of molecular evolution.

Test of Goodness-of-Fit of Models

Statistical tests are available for examining the adequacy of models used for phylogenetic estimation while allowing for nonindependence of sequences due to their common ancestry. The test proposed by Goldman (1993*b*) (see also Navidi et al. 1991; Reeves 1992) compares the log-likelihood for an estimated phylogeny attained under the model in question with that attained under an "unconstrained" model that has no phylogenetic components. The difference in log-likelihoods represents the cost of the phylogenetic inferences drawn. Too high a cost implies that the model is inadequate to explain the data. An accurate statistical test is achieved by using Monte Carlo methods to simulate data sets according to the estimated phylogeny and parameters and analyzing each of these to estimate the distribution of the test statistic under the null hypothesis. Data sets are simulated by random generation of ancestral sequences, according to the equilibrium codon frequencies, and by the simulated evolution of these sequences along descendant lineages according to the model under evaluation (see Goldman 1993*b* for full details).

This test is very time consuming, because of the analysis performed on each simulated data set. A. von Haeseler (personal communication) has suggested to us a time-saving approximation, which is to evaluate the likelihood of simulated data sets by using the original estimated tree, instead of finding the m.l. tree for each. This means that the simulated values of the log-likelihood difference statistic may be too high, affecting the statistical size and power of the test (Hall and Wilson 1991). If these approximate values imply rejection of a model, however, it is certain that the correct values would do so also. Only in the case of acceptance of a model by the approximate test would full reanalysis be necessary. In our experience, the approximation is very good. This approximation is used on one occasion below, as noted.

Diagnostic Tests

Goldman (1993*a*) shows how the probabilities of observing each site pattern, in light of the estimated tree and parameters, may be used to perform diagnostic statistical tests to determine reasons for failure of models to fit data. Two statistics that have been found particularly useful are the number of different site patterns observed in the data and the number of constant sites (Goldman 1993*a*). In this study we have used the observed number of different codon-site patterns as a diagnostic statistic. However, instead of considering the

number of constant codon sites, we have divided the possible patterns into 33 categories, according to the number of different nucleotides observed at each codon position. The categories are shown in table 1. The lower-numbered categories represent patterns with few substitutions (especially at second and first codon positions), and higher-numbered categories represent patterns with many changes. Category 1 contains all the "constant" patterns. We have found it informative to compare the observed numbers of codon sites, in these categories, with the expected numbers under the model's assumptions.

For nucleotide-based models, there are 4^s possible site patterns observable for s sequences, and it is often possible to calculate the frequencies of each site pattern expected under the model, to derive the theoretical distributions of the test statistics. For our codon-based model, this number rises to 61^s—for $s = 5$, this is equal to approximately 8.45×10^8. This is too large a number to analyze in the manner described by Goldman (1993*a*). Instead, Monte Carlo simulations can be used to estimate the distributions of these statistics under the model's assumptions.

Results

Mammalian α- and β-Globin Genes

We have analyzed the α- and β-globin genes of primate (human), artiodactyl (goat for α-globin and cow for β-globin), lagomorph (rabbit), rodent (rat), and marsupial (the eastern quoll, *Dasyurus viverrinus,* for α-globin and opossum for β-globin), previously studied by Yang et al. (1994). Two codons are missing at the beginning of the cow β-globin gene, and these codons in all the other species are excluded. Start and stop codons and introns are also excluded. Previous experience has shown these α- and β-globin genes to have similar evolutionary dynamics. The two genes were combined into one data set, with 285 codons in each combined sequence (855 nucleotides; 3×141 for the α-globin gene and 3×144 for the β-globin gene). The codon frequencies in different species are very similar, implying that the homogeneity and stationarity assumptions are acceptable.

m.l. Estimation

We have taken the codon frequencies π_i (equation [3]) as 60 ($= 61-1$) free parameters, restricted only by $\sum_i \pi_i = 1$. A second model, using the nucleotide frequencies at each of the three codon positions to calculate the expected codon frequencies, was found to produce a much poorer fit to the data. We conclude that nucleotide frequencies at the three codon positions cannot account for codon frequencies, and the following analyses are all based on the first model.

Table 1
The Observed Numbers of Codon-Site Patterns for the α- and β-Globin Genes and Their Expected Numbers under Different Models

Category	b_2	b_1	b_3	Observed	Expected, under Various Models[a] (1)	(2)	(3)
1	1	1	1	72	42.5	93.5	79.7
2	1	1	2	79	55.9	45.5	43.7
3	1	1	3	13	11.4	7.3	7.2
4	1	1	4	1	0.5	0.3	0.4
5	1	2	1	14	33.2	27.5	22.4
6	1	2	2	25	43.9	28.5	29.7
7	1	2	3	10	9.1	6.9	9.8
8	1	2	4	1	0.3	0.4	0.7
9	1	3	1	1	7.4	4.6	4.5
10	1	3	2	1	8.8	7.3	10.2
11	1	3	3	3	1.7	2.2	4.0
12	1	3	4	0	0.1	0.1	0.3
13	1	4	1	0	0.3	0.2	0.3
14	1	4	2	0	0.4	0.7	0.8
15	1	4	3	0	0.1	0.2	0.5
16	1	4	4	0	0.0	0.0	0.0
17	2	1	1	5	13.2	11.1	7.9
18	2	1	2	12	18.3	11.7	12.1
19	2	1	3	5	4.3	3.1	3.8
20	2	1	4	3	0.2	0.2	0.4
21	2	2	1	5	9.0	5.9	5.0
22	2	2	2	12	12.7	11.2	13.9
23	2	2	3	6	2.9	4.1	6.9
24	2	2	4	0	0.1	0.2	0.6
25	2	3	1	1	1.7	1.3	1.1
26	2	3	2	2	2.2	3.3	5.3
27	2	3	3	2	0.5	1.3	3.7
28	2	3	4	0	0.0	0.1	0.2
29	2	4	1	0	0.1	0.1	0.1
30	2	4	2	0	0.1	0.1	0.6
31	2	4	3	0	0.0	0.1	0.4
32	2	4	4	0	0.0	0.0	0.0
33	≥3	any	any	12	4.4	6.3	9.0
Goodness of fit[b]					136.7	118.5	79.6

NOTE.—The possible codon-site patterns are divided into 33 categories, according to the numbers of different nucleotides across species at the three codon positions. For example, category 10 includes all sites at which the second position has the same nucleotide across all species ($b_2 = 1$); the first position has $b_1 = 3$ different nucleotides; and the third position has $b_3 = 2$ different nucleotides.

[a] The three models are as follows: (1) the new codon-based model described in this paper, which assumes constant rate across codon sites; (2) 1 plus a gamma distribution of rates, using shape parameter $\alpha = 1$ and previous estimates of other parameters; and (3) 1 plus a gamma distribution of rates, with all parameters estimated under the discrete gamma model.

[b] Pearson $\chi^2 = \Sigma(\text{obs.} - \text{exp.})^2/\text{exp.}$ For model (1), category 16 is omitted from calculation, as its expectation is indistinguishable from zero, with the accuracy available.

The m.l. tree using our codon-based model is shown in figure 3. Denoting the sequences "P" for primate, "A" for artiodactyl, "L" for lagomorph, and "R" for rodent, and assuming the marsupial (M) to be the out-

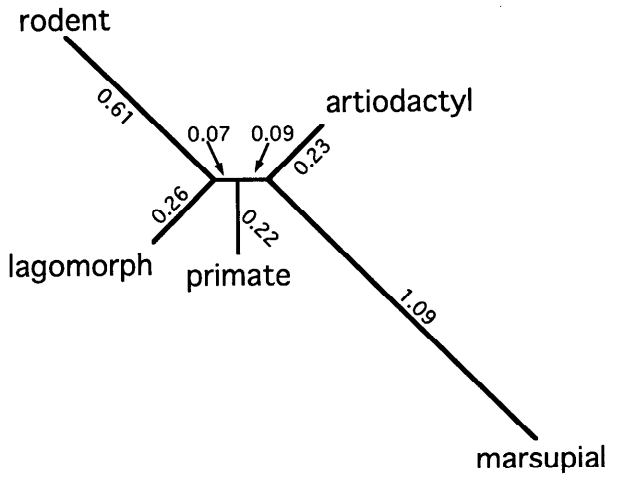

FIG. 3.—Maximum-likelihood tree for the α- and β-globin genes, using the codon-based model of the present paper. Branch lengths are expected numbers of nucleotide substitutions per codon. The parameters in the model are estimated as $\hat{\kappa} = 1.45$; $\hat{V} = 43.99$.

group, this tree may be written as (((R,L),P),A). The tree conforms well with current biological thinking (e.g., Novacek 1993). Parameter estimates are $\hat{\kappa} = 1.45 \pm 0.15$ and $\hat{V} = 43.99 \pm 3.06$, with $\ell = -2915.35$. The estimate of κ agrees closely with the value 1.48 found by Yang et al. (1994) in analysis of the first and second codon positions of the same genes by using the model of Hasegawa et al. (1985). From calculations like those giving the graphs in figure 2, these parameter values give instantaneous rate ratio estimates of $T_s/T_v = 1.15$ and $\rho_s/\rho_a = 1.05$. The ratio ρ_s/ρ_a is 0.28 for $V \to \infty$, and we see that the ratio is raised by a factor or $1.05/0.28 = 3.73$ by selective constraints at the amino acid level.

Because of the lack of reliable methods, we make no attempt here to evaluate the significance of the m.l. tree and suggest that the support from the data for this tree is probably insignificant. We point out that any test of positivity of middle branches in the m.l. tree, effectively a comparison between the m.l. tree and the star tree, is not an evaluation of the reliability of the m.l. tree (Yang et al. 1994); for example, the likelihood-ratio test suggests that at least the five best trees for these data are all "significantly better" than the star tree (star tree $\ell = -2930.29$ with $\hat{\kappa} = 1.44$, $\hat{V} = 45.92$; other results not shown). Similar to the findings of Yang et al. (1994), estimates of parameters κ and V obtained from different tree topologies are very similar. Over all 15 possible bifurcating trees, the range of $\hat{\kappa}$ is 1.42–1.48, and the range of $\hat{V}$ is 43.99–47.17.

The Adequacy of the Model

We assess the goodness of fit of the model to data by using the statistical test of Goldman (1993*b*). When the unconstrained model is applied to the observed codon-site pattern frequencies, the maximum possible log-likelihood for a codon-based model is −1493.92. Compared with −2915.35, the value attained using our new model, we obtain the goodness-of-fit statistic $\delta = -1493.92 - (-2915.35) = 1421.43$. The distribution against which this value is compared was obtained by Monte Carlo means as described above, and the results are shown in the top panel of figure 4. This indicates rejection of the model.

To place this result in context, we performed an analysis of these data, using the nucleotide-based model of Hasegawa et al. (1985). This model supports a different topology, (((P,L),A),R), with marsupial as the outgroup, which conforms less well with current biological thinking (e.g., Novacek 1993). Goldman's (1993*b*) goodness-of-fit test not only indicates rejection of the model, but implies rejection by a much greater margin (fig. 4, *bottom*). The codon-based model fits better to these data than does the nucleotide model of Hasegawa et al. (1985).

Diagnostic Tests

One of the major features of molecular sequence evolution unaccounted for in previous models is the variation of substitution rates over nucleotide sites. This was found to have a profound effect on models' fits to data (Yang et al. 1994). Our codon-based model allows for different rates at the three positions within a codon, by its use of genetic code information, as, for example, changes at third codon positions are least likely to cause amino acid replacements and so are effectively given higher rates than are second- or first-position changes. However, it does not make any allowance for different rates at different codon sites. As selective restraints exerted on amino acids in different domains of the protein must be very different owing to the different roles of these amino acids in the structure and function of the protein (e.g., see Overington et al. 1992), we expect that rate heterogeneity over codon sites is a major reason for the rejection of the model.

To search for the reasons for the discrepancy between the model and the data, we used two diagnostic tests. The first considers the observed number of different (codon) site patterns. For the α- and β-globin genes, 223 patterns are observed at the 285 sites. The theoretical distribution of this statistic, estimated by generating 500 Monte Carlo samples and using parameter estimates obtained from the codon-based model, has mean 254.9 and 95% confidence interval (244, 265). The data exhibit too few different patterns, compared with expectations under the model's assumptions.

Second, the 285 codon sites of the α- and β-globin genes were placed in categories according to the numbers

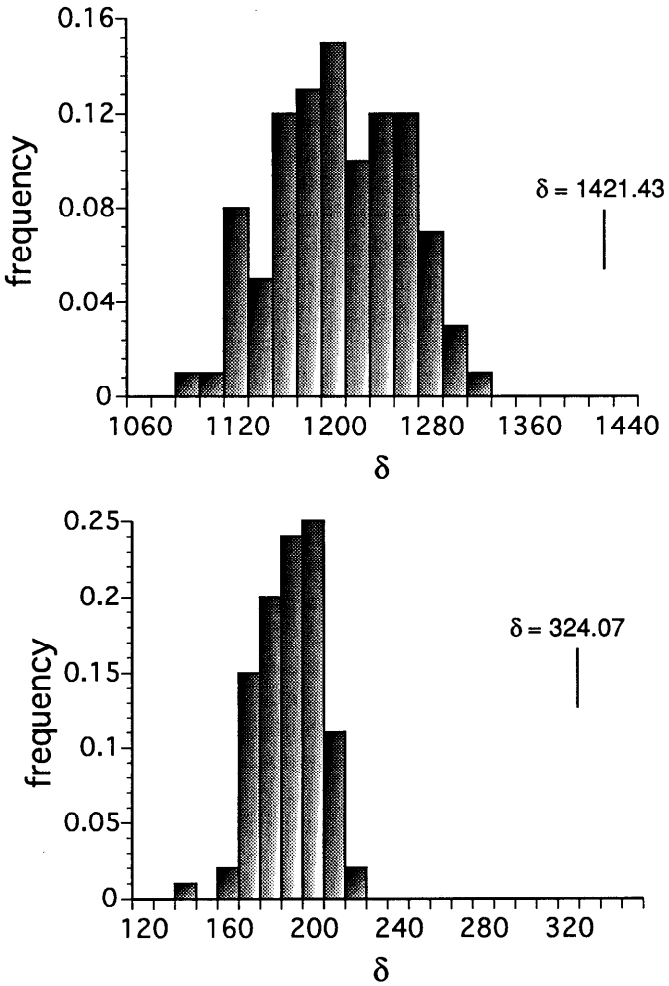

FIG. 4—*Top,* Statistical test of the adequacy of the codon-based model for the α- and β-globin genes. The test statistic (δ = 1421.43) falls above the 95% point of the simulated distribution, and the model is rejected. *Bottom,* Statistical test of the adequacy of the HKY85 model (Hasegawa et al. 1985) for the same data set: the test statistic falls beyond the 95% point of the simulated distribution, and the model is rejected by a greater margin than in the test of the codon-based model.

of bases observed at each position (see Methods) and were compared with expected numbers given the estimated tree and parameter values, again found by Monte Carlo simulations (table 1). The results show that the model predicts too few constant or very slowly evolving codon sites and too few quickly evolving codon sites. This clearly indicates rate variation across codon sites, not accounted for by the model.

Accounting for Rate Heterogeneity across Codon Sites

Following the successful use of a gamma distribution to model rate variation across nucleotide sites (e.g., see Yang 1993; Yang et al. 1994), we studied the effect of allowing rate variation across codon sites according to a gamma distribution. The computational burden of a full m.l. treatment of the gamma distribution in the manner of Yang (1993) is too great for us to attempt. Instead, we took the estimated tree (and associated parameter estimates) from figure 3 and simulated data sets using this tree but adding a gamma distribution with shape parameter $\alpha = 1$ (selected by trial and error). Ideally, parameters κ and V and branch lengths should be estimated from a full m.l. reanalysis of the data, by assuming the gamma distribution. Even without this, these parameter values appear to provide a better fit to the data (table 1). The expected number of different codon-site patterns has mean 219.4 and 95% confidence interval (204.5, 236), obtained from 100 simulations.

A more satisfactory approach has been described by Yang (in press), in which a discrete distribution with k categories, each with probability $1/k$, is used to approximate the gamma distribution. Higher values of k give closer approximations, and, to obtain an accurate estimate of the parameter α, we used $k = 8$. Parameter

estimates obtained from this "discrete gamma" model are $\hat{\kappa} = 1.44 \pm 0.20$, $\hat{V} = 35.58 \pm 2.66$, $\hat{\alpha} = 0.76 \pm 0.12$, with $\ell = -2{,}850.58$. The existence of rate variation over codon sites can be seen from the improvement in likelihood on adding the discrete gamma distribution ($2\Delta\ell = 129.5$; compare $\chi^2_{1,1\%} = 6.63$). The tree and branch lengths are estimated to be (((R: 0.99, L: 0.38): 0.12, P: 0.29): 0.11, A: 0.39, M: 2.11) (numbers give the lengths of the branches that they follow), the same topology as before (fig. 3) but with markedly increased branch lengths. Using these parameter values to generate 100 simulations by assuming a (continuous) gamma distribution of rates over codon sites, we obtain the expected numbers of codon-site patterns falling within the 33 categories shown in table 1. The distribution of the number of different site patterns, estimated by simulation, has mean 232.3 and 95% confidence interval (220, 243.5). Although, owing to computational reasons, we have not performed a test on the general adequacy of the model, these two diagnostic tests suggest that the codon-based model, when combined with a gamma distribution of rates over codon sites, appears to provide a satisfactory description of the evolution of these genes.

ADP-Glucose Pyrophosphorylase Genes

Smith-White and Preiss (1992) aligned DNA sequences coding for ADP-glucose pyrophosphorylase proteins. Their analysis suggested that among plant ADP-glucose pyrophosphorylase sequences the major division is into large and small-subunit sequences (created by a duplication event) and that within the small-subunit sequences the major division is into monocots and dicots (Smith-White and Preiss 1992). We selected sequences from wheat (large subunit), potato (small subunit; dicot), and rice and maize (small subunit; monocot) as having long homologous regions. The sequences are those referred to by Smith-White and Preiss (1992) as "WE7," "potato (T O)," "rice seed," and "*bt*-2," respectively. After removal of codons sites involving insertions or deletions, the aligned sequences are 432 codons (1296 nucleotides) long. We notice that codon frequencies are quite different in different species, a feature not accommodated by our model.

Maximum-likelihood phylogenetic estimation using the codon-based model gives the tree ((rice: 0.23, maize: 0.28): 0.49, potato: 0.32, wheat: 5.05). This is in agreement with the findings of Smith-White and Preiss (1992). The length of the branch leading to the (large subunit) wheat sequence supports the major division between large- and small-subunit sequences, and within the small-subunit sequences the division between monocots and dicots is confirmed. Other parameter estimates from the codon-based model are $\hat{\kappa} = 1.65 \pm 0.14$ and $\hat{V} = 32.26 \pm 1.75$. These values give instantaneous rate ratios $T_s/T_v = 1.35$ and $\rho_s/\rho_a = 1.50$. The ratio ρ_s/ρ_a is 0.32 for $V \to \infty$; selective constraints have raised this ratio by a factor of 4.64.

For comparison, we also analyzed these data by using the nucleotide-substitution model of Hasegawa et al. (1985). The best tree from this model has $\kappa = 2.36$ and has the same topology, but with different branch lengths: ((rice: 0.069, maize: 0.084): 0.12, potato: 0.098, wheat: 0.61). Note that branch lengths in the latter tree, measured by the number of nucleotide substitutions per nucleotide site, are all less than three times the corresponding branch lengths in the tree obtained from the codon-based model. This is presumably because the model of Hasegawa et al. (1985) fails to allow rate variation among the three codon positions; ignoring rate variation over sites has been found to lead to serious underestimation of branch lengths (Yang et al. 1994).

Both models were again assessed using the goodness-of-fit test of Goldman (1993*b*). Results are shown in figure 5. For the analysis of the ADP-glucose pyrophosphorylase genes using our codon-based model, the approximate test proposed by A. von Haeseler (see Methods) was used. Comparison of values of the test statistic obtained with and without this approximation suggested that the effect in this case is to increase the test statistic by about 3.5; in other words, the true distribution for the test of the codon-based model is approximately that shown in the top panel of figure 5 but shifted 3.5 units to the left. This has a negligible effect, and all results remain qualitatively the same.

Both models are rejected, that of Hasegawa et al. (1985) quite dramatically but the new model only just so. Again, the new model appears to be preferable.

Discussion

Estimation of Synonymous and Nonsynonymous Rates of Substitution

Miyata and Yasunaga (1980), Li et al. (1985), Nei and Gojobori (1986), Li (1993), and Pamilo and Bianchi (1993) proposed methods for estimating the numbers of synonymous and nonsynonymous substitutions from comparison of two protein-coding DNA sequences. Here we examine the relationship between parameters in our codon-based model and results from these pairwise methods, which lack an explicit model of substitutions between codons. The method of Nei and Gojobori (1986), an approximation of that of Miyata and Yasunaga (1980), is the simplest but normally produces results very similar to those obtained by the methods of Miyata and Yasunaga (1980) and Li et al. (1985) (Nei and Gojobori 1986). We thus use the method of Nei and Gojobori (1986) as an example. This method performs a codon-by-codon comparison. For each pairwise comparison, the 285 × 3 nucleotide sites are separated

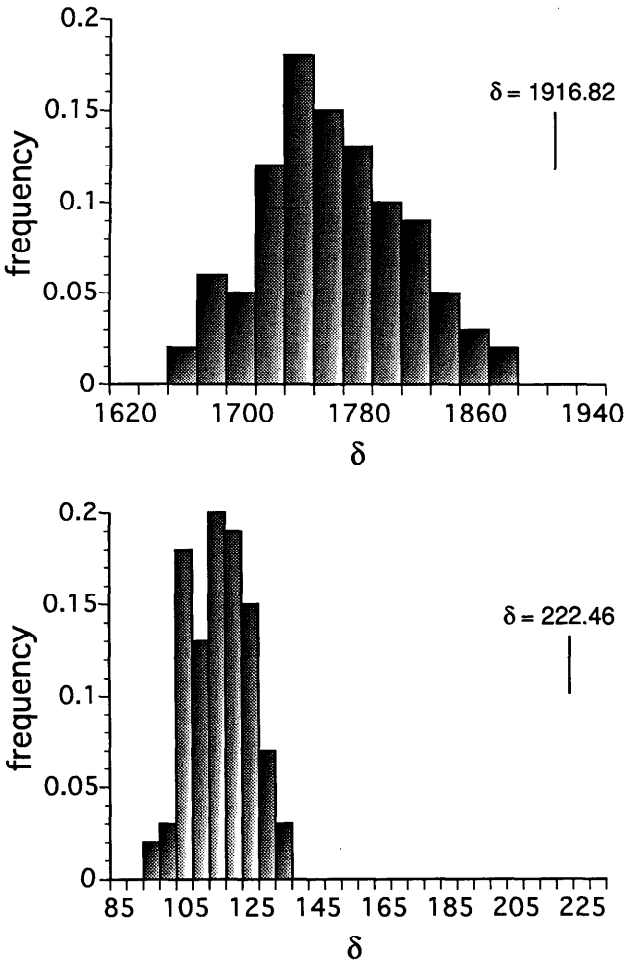

FIG. 5.—*Top,* Statistical test of the adequacy of the codon-based model for the ADP-glucose pyrophosphorylase genes. The test statistic falls above the 95% point of the simulated distribution, and the model is rejected. *Bottom,* Statistical test of the adequacy of the HKY85 model (Hasegawa et al. 1985) for the same data set: the test statistic falls far beyond the 95% point of the simulated distribution, and the model is rejected by a much greater margin than in the test of the codon-based model.

into the numbers of synonymous and nonsynonymous sites, S and N. The observed differences between sequences are similarly separated into the numbers of synonymous and nonsynonymous differences, S_d and N_d. The argument of Jukes and Cantor (1969) is then applied to correct for multiple substitutions that have occurred at one site (Nei and Gojobori 1986), giving

$$K_s = -{}^3/_4 \log(1 - {}^4/_3 \times S_d/S) \qquad (5a)$$

$$K_a = -{}^3/_4 \log(1 - {}^4/_3 \times N_d/N) \qquad (5b)$$

as the number of synonymous substitutions *per synonymous site* and the number of nonsynonymous substitutions *per nonsynonymous site,* respectively. In table 2, we list results for all the pairwise comparisons for the α- and β-globin genes, obtained from the method of Nei and Gojobori (1986) and calculated using the MEGA package (Kumar et al. 1993).

We note some problems with this strategy. First, use of the model of Jukes and Cantor (1969) ignores transition/transversion rate bias and differences in the frequencies of the four nucleotides. The methods of Li et al. (1985) and Li (1993), which are based on the model of Kimura (1980), also assume that all the four nucleotides have equal frequency ¼. Second, the classification of sites and differences into synonymous and nonsynonymous categories does not appear to be natural. This is especially true when the two codons compared are quite different and there is more than one pathway of changing from one codon to the other; there is no guarantee that a site has always been a "synonymous site" or "nonsynonymous site" throughout its evolutionary history. Third, although the correction formula of Jukes and Cantor (1969) is valid when the model is used at the mononucleotide level, equations (5*a*) and (5*b*) appear to be logically flawed when used at the codon level. For example, it is not reasonable to assume that a nonsynonymous site will have equal probability of changing into *three* other nonsynonymous sites (Lewontin 1989).

As discussed above, ρ_s and ρ_a represent rates of synonymous and nonsynonymous substitutions *per codon* in our codon-based model. If ρ_s^∞ and ρ_a^∞ are used for the values of ρ_s and ρ_a corresponding to $V \to \infty$, then ρ_s^∞ and ρ_a^∞ represent "potentials" of synonymous and nonsynonymous substitutions when no selective constraints exist at the amino acid level; $3\rho_s^\infty$ and $3\rho_a^\infty$ therefore represent the numbers of synonymous and nonsynonymous nucleotide sites per codon, respectively. (Under the assumption of equal rates of substitution between nucleotides and of equal codon frequencies, $\rho_s^\infty = 0.21$ and $\rho_a^\infty = 0.79$, as pointed out by Nei and Gojobori [1986]. The values calculated from the α- and β-globin gene sequences are $\rho_s^\infty = 0.22$ and $\rho_a^\infty = 0.78$ for our codon-based model.) If we write ρ_s^* and ρ_a^* for the m.l. values (i.e., those calculated using the m.l. estimates of V and κ), the numbers of synonymous substitutions per synonymous site and nonsynonymous substitutions per nonsynonymous site are $K_s = \rho_s^*/3\rho_s^\infty$ and $K_a = \rho_a^*/3\rho_a^\infty$, and $K_s/K_a = \rho_s^*\rho_a^\infty/\rho_s^\infty\rho_a^*$. K_s and K_a calculated in this way are listed in table 2.

Our estimates of K_s and K_a (or ρ_s and ρ_a) are free from the problems, described above, of the method of Nei and Gojobori (1986). First, the frequency parameters π_i account for different nucleotide frequencies at

Table 2
Number of Synonymous Substitutions per Synonymous Site (K_s) and Number of Nonsynonymous Substitutions per Nonsynonymous Site (K_a) for All the Pairwise Comparisons of the Mammalian α- and β-Globin Genes, Estimated Using the Method of Nei and Gojobori (1986) and Using the Codon-based Model of the Present Paper

	K_s/K_a (K_s, K_a)			
	Primate	Artiodactyl	Lagomorph	Rodent
Nei and Gojobori (1986):				
Artiodactyl	4.017 (0.346, 0.086)	...	...	...
Lagomorph	3.831 (0.332, 0.087)	3.431 (0.361, 0.105)	...	...
Rodent	4.918 (0.619, 0.126)	4.098 (0.611, 0.149)	4.629 (0.623, 0.134)	...
Marsupial	5.319 (1.025, 0.193)	5.376 (1.024, 0.191)	4.651 (0.979, 0.211)	3.725 (0.894, 0.240)
Codon-based Model:				
Artiodactyl	3.237 (0.363, 0.112)	...	...	...
Lagomorph	3.599 (0.385, 0.107)	3.030 (0.396, 0.130)	...	...
Rodent	3.831 (0.664, 0.173)	3.436 (0.669, 0.195)	4.016 (0.732, 0.182)	...
Marsupial	4.280 (1.219, 0.285)	3.746 (1.055, 0.282)	3.930 (1.193, 0.303)	2.971 (0.958, 0.323)

the three codon positions; even codon usage information can be incorporated in the model. The transition/transversion rate bias is accounted for by the parameter κ. Second, as the model is formulated in terms of instantaneous rates of substitutions between codons, there is no confusion as to whether a substitution is synonymous or nonsynonymous. Third, since K_s and K_a are specified as functions of parameters in our model, the invariance property of m.l. estimators ensures that our estimates of K_s and K_a, and their ratio, are also m.l. estimates. We advocate the use of our method to calculate K_s/K_a, in preference to that of Nei and Gojobori (1986).

Estimates of K_s and K_a by the method of Nei and Gojobori (1986) are smaller than estimates from the codon-based model; particularly those of K_a, leading to overestimation of the ratio K_s/K_a. Underestimation of both K_s and K_a can be attributed to the observation that simpler substitution models produce smaller distance estimates (e.g., see Yang et al. 1994). Although the differences between the two estimation methods described above may be responsible for the differences in estimates of K_s and K_a, the major factor may be the way that Nei and Gojobori (1986) count the numbers of synonymous and nonsynonymous sites (S and N). Li (1993) and Pamilo and Bianchi (1993) have pointed out that ignoring transition/transversion rate bias leads to smaller estimates of S and greater estimates of N and thus to overestimates of K_s/K_a.

The Distances between Amino Acids

In this study we modeled selective constraints on protein-coding genes by using knowledge of the genetic code and the amino acid distance matrix of Grantham (1974). Li et al. (1985, p. 165) noted that "Grantham's (1974) indices are quite adequate in predicting amino acid exchangeability." Indeed, we found that for the α- and β-globin genes the parameter V in our model, in combination with Grantham's (1974) distances, can explain the rate differences among the three positions within a codon better than can a nucleotide-based model which allows independent rates at each codon position (results not shown). However, the distances of Grantham (1974) do not incorporate any information concerning the higher-dimensional structures of the amino acids. Distances involving Cys, Trp, etc. may not be very reliable, as these amino acids have special functions in proteins, as discussed by Grantham (1974). It may be worthwhile to improve our model by incorporating better distance measures. Taylor (1989) and Taylor and Jones (1993) discuss measures other than that of Grantham (1974). Looked at another way, investigation of alternative distance measures could indicate which most closely represent amino acids' tendencies for mutual substitution.

We have avoided using amino acid "interchange probability matrices," e.g., the PAM matrices of Dayhoff et al. (1978), used by Kishino et al. (1990), and Jones et al. (1992). We are cautious about the use of matrices that essentially consist of amino acid substitution probabilities, $P(t)$, averaged over large numbers of alignments. These averages may be over large time scales, from very small t, when $P(t)$ will be close to the identity matrix, to very large t, when $P(t)$ simply reflects the amino acid frequencies. In these conditions, an "average" probability has little meaning. Even if these matrices do represent meaningful probabilities, each only applies to a single evolutionary time (distance) t (Schöniger et al. 1990), and it is not clear how they may be converted to time-independent evolutionary rates. We consider it advantageous instead to work in terms of rate parame-

ters, Q_{ij}, which are independent of time but from which probabilities, $P(t)$, may be calculated for any value of t.

The Codon-based Model of Nucleotide Substitution

Our codon-based model of nucleotide substitution provides a better fit to the example data sets we have studied than do the nucleotide-based models. Both intuition and recent research (Yang et al. 1994) show that this improvement is likely to lead to more accurate estimation of phylogenetic relationships, and for this reason we suggest that the new model is of value in the analysis of protein-coding sequences. We believe that this improvement is due to the model's ability to use information from both the nucleotide- and amino acid–level interpretations of protein-coding sequences. Other advantages of this model are its relative biological plausibility and its use of parameters controlling biologically important features of DNA sequence evolution, e.g., transition/transversion rate bias, codon usage, and silent/replacement substitution rate bias, including a measure of the difference between amino acids. These parameters implicitly allow rate variation and lack of independence of substitutions within codons to be incorporated. In addition, the model permits the separation of the effects of rate variation within codons and between codons. Our results indicate that both these effects are important.

Allied with maximum-likelihood estimation, our model is computationally very slow, especially when the (discrete) gamma distribution is used to describe rate variation over codon sites. Even when a full m.l. analysis is not possible, we suggest that the model may still be used to evaluate a few candidate trees obtained by other methods, to perform a finer comparison of the tree topologies and to obtain more reliable estimates of parameters. Noting that improved models sometimes give quite different, presumably also improved, estimates of evolutionary distances, we suggest that our model could also be of practical use in pairwise sequence comparison, e.g., to produce better distance estimates for use in distance-matrix methods. The importance of good distance measures is known to be high (DeBry 1992; Charleston et al. 1993).

Acknowledgments

We thank Adrian Friday for many useful discussions. Comments from Paul Sharp and from two anonymous referees enabled us to make many improvements to this paper, in particular regarding the presentation of our examples and rates of synonymous/nonsynonymous substitutions. Z.Y. was supported by a grant from Department of Zoology, The Natural History Museum (London), during this study.

LITERATURE CITED

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.

CHARLESTON, M. A., M. D. HENDY, and D. PENNY. 1993. Neighbor-joining uses the optimal weight for net divergence. Mol. Phylogenet. Evol. **2**:6–12.

COX, D. R., and H. D. MILLER. 1977. The theory of stochastic processes. Chapman & Hall, London.

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

DEBRY, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. Mol. Biol. Evol. **9**:537–551.

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

GOLDMAN, N. 1991. Statistical estimation of evolutionary trees. Ph.D. thesis, University of Cambridge.

———. 1993*a*. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. **37**:650–661.

———. 1993*b*. Statistical tests of models of DNA substitution. J. Mol. Evol. **36**:182–198.

GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. Science **185**:862–864.

HALL, P., and S. R. WILSON. 1991. Two guidelines for bootstrap hypothesis testing. Biometrics **47**:757–762.

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

JONES, D. T., W. R. TAYLOR, and J. M. THORNTON. 1992. Updating the Dayhoff matrix. Comput. Appl. Biosci. **8**:275–282.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. Munro, ed. Mammalian protein metabolism. Vol. 3. 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.

KISHINO, H., T. MIYATA, and M. HASEGAWA. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J. Mol. Evol. **31**:151–160.

KUMAR, S., K. TAMURA, and M. NEI. 1993. MEGA: molecular evolutionary genetics analysis, version 1.01. The Pennsylvania State University, University Park.

LEWONTIN, R. C. 1989. Inferring the number of evolutionary events from DNA coding sequence differences. Mol. Biol. Evol. **6**:15–32.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MIYATA, T., and T. YASUNAGA. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16**:23–36.

NAVIDI, W. C., G. A. CHURCHILL, and A. VON HAESELER. 1991. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. Mol. Biol. Evol. **8**:128–143.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.

NOVACEK, M. J. 1993. Reflections on higher mammalian phylogenetics. J. Mammal. Evol. **1**:3–30.

OVERINGTON, J., D. DONNELLY, M. S. JOHNSON, A. SALI, and T. L. BLUNDELL. 1992. Environment-specific amino-acid substitution tables—tertiary templates and prediction of protein folds. Protein Sci. **1**:216–226.

PAMILO, P., and N. O. BIANCHI. 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. Mol. Biol. Evol. **10**:271–281.

REEVES, J. H. 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., and M. T. CLEGG. 1987. Evolutionary analysis of plant DNA sequences. Am. Nat. **130**:S74–S100.

SCHÖNIGER, M., G. L. HOFACKER, and B. BORSTNIK. 1990. Stochastic traits of molecular evolution—acceptance of point mutations in native actin genes. J. Theor. Biol. **143**: 287–306.

SMITH-WHITE, B. J., and J. PREISS. 1992. Comparison of proteins of ADP-glucose pyrophosphorylase from diverse sources. J. Mol. Evol. **34**:449–464.

TAYLOR, W. R. 1989. A template based method of pattern matching in protein sequences. Prog. Biophys. Mol. Biol. **54**:159–252.

TAYLOR, W. R., and D. T. JONES. 1993. Deriving an amino acid distance matrix. J. Theor. Biol. **164**:65–83.

THORNE, J. L., H. KISHINO, and J. FELSENSTEIN. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. **33**:114–124. (Erratum, J. Mol. Evol. **34**:91, 1992).

———. 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. **34**:3–16.

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

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

———. Maximum likelihood phylogenetic estimation from DNA sequences when substitution rates differ over sites: approximate methods. J. Mol. Evol. (in press).

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

PAUL SHARP, reviewing editor

Received February 1, 1994

Accepted May 16, 1994