Comparison of Models for Nucleotide Substitution Used in Maximum-Likelihood Phylogenetic Estimation

Ziheng Yang, Nick Goldman,† and Adrian Friday**

*Department of Zoology, University of Cambridge, and †Laboratory of Mathematical Biology, National Institute for Medical Research

Using real sequence data, we evaluate the adequacy of assumptions made in evolutionary models of nucleotide substitution and the effects that these assumptions have on estimation of evolutionary trees. Two aspects of the assumptions are evaluated. The first concerns the pattern of nucleotide substitution, including equilibrium base frequencies and the transition/transversion-rate ratio. The second concerns the variation of substitution rates over sites. The maximum-likelihood estimate of tree topology appears quite robust to both these aspects of the assumptions of the models, but evaluation of the reliability of the estimated tree by using simpler, less realistic models can be misleading. Branch lengths are underestimated when simpler models of substitution are used, but the underestimation caused by ignoring rate variation over nucleotide sites is much more serious. The goodness of fit of a model is reduced by ignoring spatial rate variation, but unrealistic assumptions about the pattern of nucleotide substitution can lead to an extraordinary reduction in the likelihood. It seems that evolutionary biologists can obtain accurate estimates of certain evolutionary parameters even with an incorrect phylogeny, while systematists cannot get the right tree with confidence even when a realistic, and more complex, model of evolution is assumed.

Introduction

Models of nucleotide substitution are important for estimation of evolutionary trees and for understanding of the evolutionary process of DNA sequences. As more and more sequences are determined, attempts to refine models seem ever more worthwhile. Better models will lead to more accurate estimates of the evolutionary history of the species concerned and to a better understanding of the forces and mechanisms that affected the evolution of the sequences.

The method of maximum likelihood, proposed by Felsenstein (1981) for tree estimation from DNA sequences, is undoubtedly the method with the best-understood statistical basis. The assumptions made in this approach are all explicit and so can be checked against real data. Goldman (1991, 1993) devised a method for testing the general adequacy of models used in conjunction with maximum-likelihood estimation. Application of this test to real data reveals that, for most of the data analyzed, widely used models should be rejected. It has been speculated that the most worrying unrealistic assumption made in approaches to phylogenetic estimation based on Felsenstein's (1981) formulation is that of constancy of substitution rates over all nucleotide sites. This assumption must be unrealistic for gene sequences coding for products with biological functions. There have been many attempts to model rate variation over sites by using the gamma distribution (e.g., see Uzzell and Corbin 1971; Holmquist et al. 1983; Nei and Gojobori 1986; Jin and Nei 1990; Kocher and Wilson 1991). Recently Yang (1993) suggested an extension to the method of Felsenstein (1981), in which a gamma distribution is used to model rate variation over sites. The rate at a specific site is assumed to be a random variable drawn from a gamma distribution. We examine here whether this speculation is supported and whether the revised model, which takes into account such spatial rate variation, is adequate for the data.

We analyzed some real data in order to address the following questions: (1) How much can the model be improved, as judged by the increase in likelihood, by assuming a gamma distribution of rates over sites instead of a single rate? (2) Is the revised model then adequate for describing the data? This question is especially interesting for those data sets for which Goldman (1991, 1993) found previous models to be inadequate. (3) How different are the reconstructed trees obtained using different models, or, put another way, how robust is the maximum-likelihood approach to violations of its assumptions?

We make a distinction between two types of statistical test of models of nucleotide substitution. The first

Key words: maximum likelihood, models, nucleotide substitution, spatial rate variation, the gamma distribution.

Address for correspondence and reprints: Dr. Ziheng Yang, Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, United Kingdom.

Mol. Biol. Evol. 11(2):316–324. 1994.

0737-4038/94/1102-0014$02.00

is a comparison of two different models, to judge whether one is significantly better than another. For example, tests have been proposed to compare a model including a "molecular clock" hypothesis to one without (Felsenstein 1981; Goldman 1991, 1993) and to compare the suitability of two different patterns of nucleotide substitution rates (Goldman 1991, 1993). Generally, the models compared will be closely related, and the tests are of value in determining which of the assumptions of the models are most important in improving their description of the data at hand. Second, it is possible to test the overall adequacy of a model—that is, does the model fit the data? Such tests have been described by Navidi et al. (1991), Goldman (1991, 1993), and Reeves (1992). Clearly, the results of such tests are of great importance for assessing the quality of inferences based on the models. Examples of both types of test appear below.

Data and Methods

Sequences

Insertions and deletions are not accommodated by the models examined here: only aligned gapless sequences were analyzed. The following data sets were used:

Data Set 1

This data set contains 895-bp mitochondrial DNA (mtDNA) sequences of human, chimpanzee, gorilla, orangutan, and gibbon (Brown et al. 1982; position 560 removed). These sequences have previously been analyzed by a number of authors using various methods. There is now little doubt as to the phylogeny of these species (e.g., see Hasegawa 1991). For computational reasons, some results in this paper are presented using only four species represented in the data set, those of human, chimpanzee, gorilla, and orangutan. We designate this smaller dataset as 1′.

Data Set 2

This data set contains α- and β-globin gene sequences of a primate (human), an artiodactyl (goat for the α-globin and cow for the β-globin sequences), a lagomorph (rabbit), and a rodent (rat), as aligned by Yang (1992). Only the first- and second-codon positions in the coding regions are used, with nucleotides both at the third position and within introns excluded. There are 570 nucleotides in all, $2 \times 141 = 282$ for the α-globin gene and $2 \times 144 = 288$ for the β-globin gene.

Data Set 3

This data set contains small-subunit RNA (ssRNA) sequences of *Sulfolobus solfatarius, Halobacterium salinarium, Escherichia coli,* and *Homo sapiens* analyzed by Navidi et al. (1991). There are 1,352 nucleotides in the sequence.

Data Set 4

This data set contains glutamine synthetase genes of *E. coli, Salmonella typhimurium, Thiobacillus ferrooxidans,* and an unrecorded species of *Anabaena* (Pesole et al. 1991). The sequence length is 928 nucleotides.

Data Set 5

This data set contains ψη-globin pseudogenes of human, chimpanzee, gorilla, and orangutan (Miyamoto et al. 1987). The sequence length is 6,166 nucleotides.

Methods

With s sequences there are 4^s possible site patterns. When independent evolution at different sites is assumed, the likelihood function is proportional to the multinomial probability of observing the data given the model and tree:

$$P = \frac{N!}{\prod_i n_i!} \prod_{i=1}^{4^s} p_i^{n_i}, \qquad (1)$$

where $N = \sum_i n_i$ is the sequence length, p_i is the probability of observing the ith site pattern, and n_i is the observed number of occurrences of the ith site pattern. In real analysis the logarithm is used and the constant term in equation (1) is ignored, to give the log-likelihood

$$\ell = \log(L) = \sum_{i=1}^{4^s} n_i \cdot \log(p_i). \qquad (2)$$

We consider two sorts of assumptions made in calculating p_i in this paper. The first concerns rate variation over nucleotide sites. We consider two models for this; one assumes a single rate, and the other assumes rates drawn from a gamma distribution. The gamma distribution with parameters α and β has mean α/β and variance α/β^2. As described by Yang (1993), in the current context β is a trivial scale factor and can be fixed equal to α to give mean 1 (and variance $1/\alpha$). Values of α less than approximately 0.5 mean the gamma distribution has a reverse-J shape and imply strong rate variation, while values of α larger than 1 or 2 imply a more-or-less-constant rate over sites. By choosing different values of α, rate variation can be accommodated in a variety of real examples, as described below.

The second assumption concerns the pattern of nucleotide substitution. The Markov process model of Hasegawa et al. (1985) is adopted. We designate this as "HKY85." In this model, the probability of nucleotide

i changing into nucleotide *j* ($j \neq i$) in a very small time interval Δt is given by

$$Q_{ij}\Delta t = \begin{cases} \kappa\mu\pi_j\Delta t & \text{for transitions: T}\leftrightarrow\text{C, A}\leftrightarrow\text{G} \\ \mu\pi_j\Delta t & \text{for transversions: T, C}\leftrightarrow\text{A, G} \end{cases} \quad (3)$$

where π_j is the frequency of nucleotide *j* when the process is in equilibrium and μ is a scale factor, chosen so that the average rate of substitution is $-\sum_j \pi_j Q_{jj} = 1$. When $\kappa = 1$, the model reduces to that of Felsenstein (1981), designated "F81." When equal frequencies are assumed for the four nucleotides, that is, $\pi_T = \pi_C = \pi_A = \pi_G = 1/4$, then the model reduces to Kimura's (1980) two-parameter model ("K80"). When $\kappa = 1$ and $\pi_T = \pi_C = \pi_A = \pi_G = 1/4$, the model is equivalent to that of Jukes and Cantor (1969) ("JC69"). Thus JC69, F81, K80, and HKY85 contain 0, 3, 1, and 4 parameters, respectively, and JC69, F81, and K80 are all special cases of HKY85.

The four nucleotide-substitution models (JC69, F81, K80, and HKY85) may each be combined with the gamma distribution for rate variation across sites (to give models denoted "JC69+Γ," "F81+Γ," "K80+Γ," and "HKY85+Γ"). We have compared these models to determine the suitability and effects of the different assumptions that they make.

All the above models are reversible, and as we do not assume the existence of a molecular clock, only unrooted trees can be estimated (Felsenstein 1981). For the F81 and HKY85 models, we estimate equilibrium frequencies of nucleotides by averaging over the sequences. These estimates should be very similar to the maximum-likelihood estimates (Goldman 1993). The ratio of transition to transversion rate, κ, in the HKY85 and K80 models is determined by maximum-likelihood estimation when a single rate over sites is assumed, and values from such calculations are used in models assuming the gamma distribution, in order to reduce computation. The estimates of κ specified in this way are often found to be very near to those obtained by iteration, if κ is not very large, say, $\kappa < 8$ (but see fig. 1).

Results

The Increase in Likelihood Obtained by Assuming Variable Rates of Substitution over Nucleotide Sites

Likelihood values and maximum-likelihood estimates of parameters under two models are listed in table 1. Model 0 assumes a single rate of substitution over sites, while Model 1 assumes the gamma distribution. In both cases the HKY85 model of nucleotide substitution is assumed. The gamma distribution reduces to the single-rate model when its parameter α approaches infinity. In practice, the likelihood values and paramete estimates (such as branch lengths) are almost indistir guishable when α is as large as 20. Model 0 is a speci case of Model 1, and a standard likelihood-ratio test ca be used to test whether Model 1 is significantly bette than Model 0. This provides a test for rate constanc over nucleotide sites. Goldman (1993) has noted tha while the distributional approximations of this test ai not known to be certainly valid, they seem to be rea sonable.

When we compare $2(\ell_1 - \ell_0)$ with a critical χ^2 valu with df = 1, the difference is extremely significant fc the mtDNA sequences, the α- and β-globin genes, th ssRNAs, and the glutamine synthetase genes (data se 1, 1′, 2, 3, and 4, all $P < 0.01$). On the other hand fc the ψη-globin pseudogenes, the difference is barely si nificant ($0.01 < P < 0.05$). The mtDNA dataset for fiv species was analyzed by Goldman (1991), who foun that Model 0 is inadequate. The same conclusion drawn when the test is applied to the smaller dataset (results not shown). The ssRNA sequences were an lyzed by Navidi et al. (1991) and Goldman (199 1993), both of whom suggested that Model 0 be rejecte for these data. We may assume that this is also the ca for the α- and β-globin genes and the glutamine synth tase gene. With another ψη-globin pseudogene data se which contains more, but shorter, sequences, Goldma (1993) found that Model 0 could not be rejected. Indee the preference for Model 1 over Model 0 for this gei is marginally significant ($0.01 < P < 0.05$) (table 1).

In summary, for data sets for which the HKY8 model is found to be inadequate, addition of the gamn distribution (HKY85+Γ) results in an impressive in provement. It seems likely that the component lackii in the original model is that of rate variation over n cleotide sites.

The Adequacy of Models

Navidi et al. (1991) suggested a method for testi the adequacy of models used in the maximum-likeliho approach. This test uses an unconstrained model which $4^s - 1$ parameters are used to describe the dat that is, all the p_i in equation (1) are taken as independe parameters, with only the restriction that $\sum_i p_i = 1$. A mittedly even this unconstrained model involves the a sumption that data at different sites are independent ai identically distributed. The log-likelihood under tl model is given by

$$\ell_{max} = \sum_i n_i \log(n_i) - N \log(N) . \quad ($$

The adequacy of a model can be evaluated by e amining $\ell_{max} - \ell$. Navidi et al. (1991) suggested the u of a χ^2 test to evaluate the statistical significance. Wh

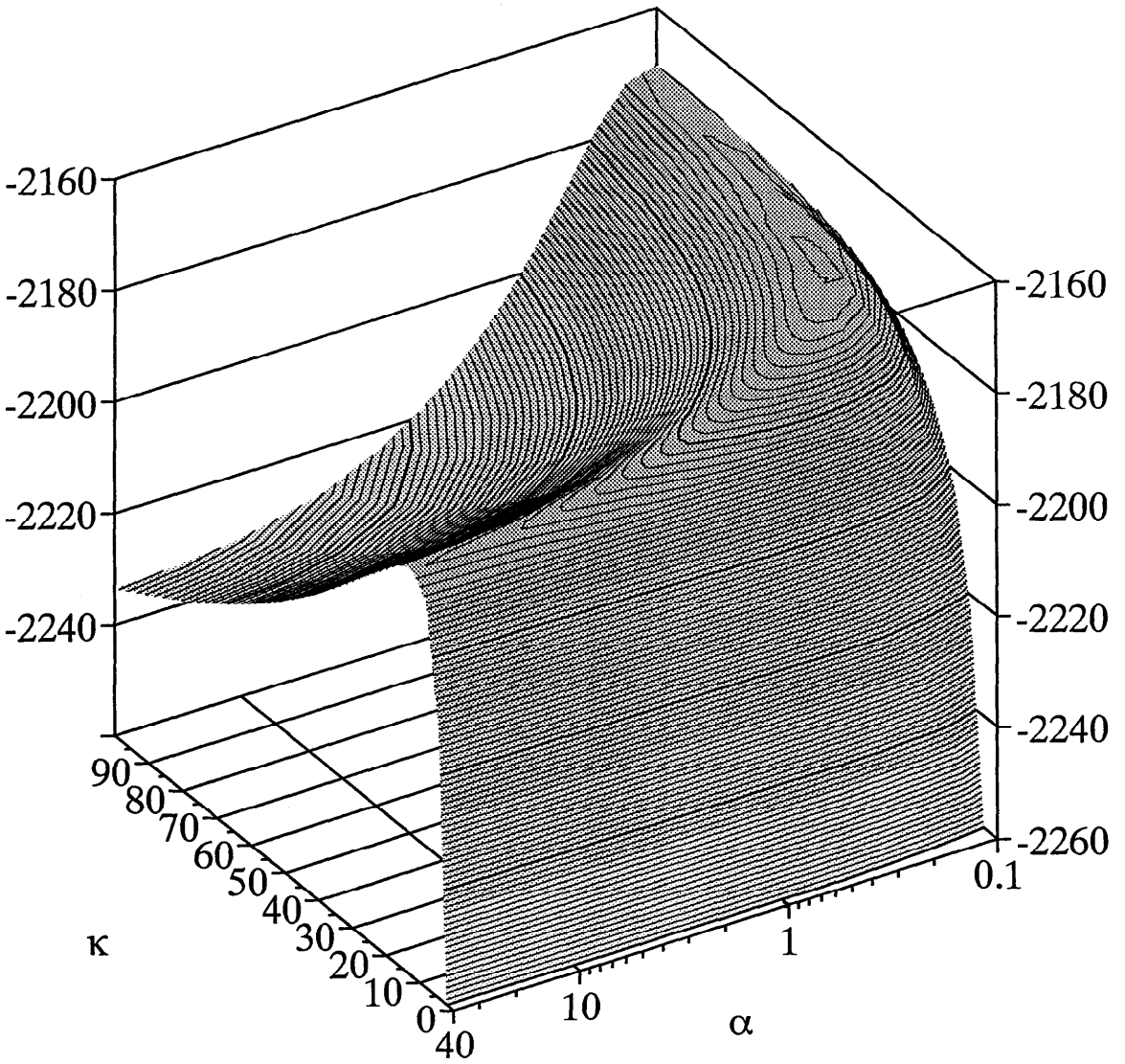

FIG. 1.—Log-likelihood as a function of the transition/transversion ratio, κ, and the α parameter of the gamma distribution. The 895-bp mtDNA sequences of human, chimpanzee, gorilla, and orangutan are used. The maximum-likelihood estimates of the two parameters are $\hat{\kappa} = 33.05 \pm 14.67$ and $\hat{\alpha} = 0.20 \pm 0.07$, with $\ell = -2{,}168.47$. The graph shows that the estimate of κ involves large sampling error, while that of α can be much more accurate. Also, the estimates of the two parameters are negatively correlated.

calculating the number of df, however, Navidi et al. (1991) ignored the fact that the maximum-likelihood tree is selected from among all the possible tree topologies. This problem is serious, particularly when more sequences are used (Goldman 1993). Goldman (1991, 1993) suggested a Monte Carlo method to derive the distribution of the test statistic under the model to be tested and to compare the observed value with this distribution. Goldman's results show that Navidi et al.'s (1991) approximation can be misleading and tends to favor the model being tested. As the method of Goldman (1991, 1993) is computationally very intensive, we have only used this test with the mtDNA sequences for four species (data set 1′).

Using the maximum-likelihood tree and branch lengths for data set 1′ under the HKY85+Γ model (table 3), we simulated data sets conforming to this model. Each simulated data set was analyzed in the same manner as the original sequences, giving simulated values of $\ell_{max} - \ell$ whose distribution is shown in figure 2. For the true data, the attained value of $\ell_{max} - \ell$ is $(-2104.19) - (-2173.69) = 69.5$ (table 1). This value falls in the middle of the distribution obtained by simulation, indicating that the HKY85+Γ model gives a good description of the evolution of these sequences.

Likelihoods of Different Trees under Different Models

Table 2 lists the likelihoods for different tree topologies under different models for the mtDNA for four

Table 1
Maximum Likelihoods With and Without the Assumption of a Gamma Distribution of Rates over Sites

Data Set and Description	Model 0: HKY85		Model 1: HKY85+Γ			Unconstrained Model	
	ℓ_0	$\hat{\kappa} \pm$ SE	ℓ_1	$\ell_1 - \ell_0$	$\hat{\alpha} \pm$ SE	ℓ_{max}	$\ell_{max} -$
(1): Five-species mtDNA	−2,665.42	9.39 ± 1.26	−2,632.11	33.32**	0.47 ± 0.09	−2,476.97	155.14
(1′): Four-species mtDNA	−2,187.60	12.23 ± 2.13	−2,173.69	13.90**	0.46 ± 0.12	−2,104.19	69.50
(2): 1st and 2d positions of the α- and β-globin gene	−1,451.01	1.48 ± 0.27	−1,434.58	16.43**	0.29 ± 0.09	−1,338.14	96.44
(3): ssRNA	−5,837.58	1.80 ± 0.13	−5,796.18	41.40**	0.94 ± 0.14	−5,591.06	205.12
(4): Glutamine synthetase gene	−2,958.05	0.96 ± 0.12	−2,948.70	9.35**	0.89 ± 0.30	−2,862.62	86.08
(5): ψη-globin gene	−10,130.14	5.35 ± 0.69	−10,127.36	2.64*	0.66 ± 0.39	−10,060.49	67.01

NOTE.—κ is the transition/transversion ratio, and α is the parameter of the gamma distribution. The HKY85 scheme of nucleotide substitution is assumed both models. Rate constancy over sites is assumed in Model 0, while a gamma distribution of rates is assumed in Model 1. Standard errors (SE) are estimated the curvature method.

* $P < 0.05$; $\chi^2_{0.05}$ (1 df) = 3.84.

** $P < 0.01$; $\chi^2_{0.01}$ (1 df) = 6.63.

species (dataset 1′). All the models produce the same maximum-likelihood tree, i.e., the tree ((human, chimpanzee), gorilla) with orangutan as the outgroup. Such agreement is most often the case for other data sets, although there are exceptions. One such case is that of the ψη-globin pseudogenes, where HKY85 and F81+Γ favor different trees. Another case is that of the ssRNAs, where HKY85 and HKY85+Γ lead to different tree topologies. In both cases the likelihoods for the different trees are almost the same, implying that not enough information is contained in the data to estimate reliably the branching order of the species.

The likelihood of the best tree under a given model is an indication of that model's goodness of fit. As

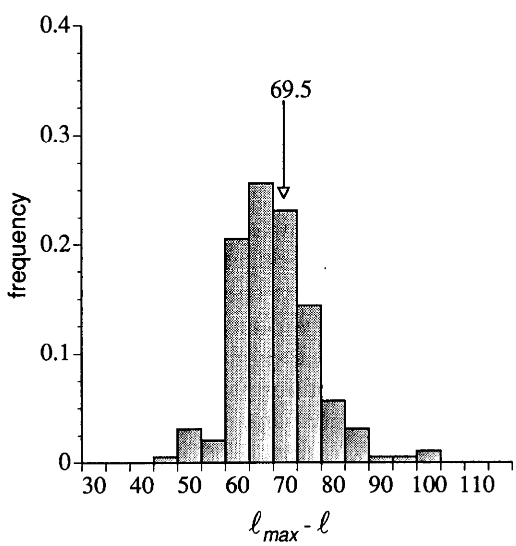

FIG. 2.—Monte Carlo distribution of $\ell_{max} - \ell$ for the test of the HKY85+Γ model applied to the mtDNA for four sequences (data set 1′). The attained value (69.5) falls well within this distribution; the HKY85+Γ model is accepted.

JC69+Γ, F81+Γ, and K80+Γ are all special cases o HKY85+Γ, the likelihood-ratio test can be used to te whether these simpler models are acceptable compare with HKY85+Γ. For the mtDNA for four species, th differences in likelihoods are 232.65 (df = 4), 156.5 (df = 1), and 88.9 (df = 3) for the three models, re spectively (table 2). All three simpler models appea totally unacceptable. JC69+Γ and K80+Γ are unrealisti because nucleotide frequencies are not equal in this dat set (π_T = 0.254, π_C = 0.331, π_A = 0.311, π_G = 0.104 while JC69+Γ and F81+Γ are unrealistic because th transition/transversion ratio is not 1 (κ = 12.23). How ever it is striking that unrealistic assumptions about th pattern of nucleotide substitution can lead to such con siderable reductions in likelihood. Even for ψηn-globi genes (data set 5) the difference between HKY85+Γ (= 5) and F81+Γ (κ = 1) is $\Delta\ell$ = 90.45. Judging by th likelihood, the assumptions about relative rates of di ferent nucleotide substitutions seem even more impo tant than do the assumptions about rate variation ove sites.

The comparison of likelihoods between tree (2) c table 2, the maximum-likelihood tree, and tree (1), th "star" tree, is a test of positivity of the middle branc of the maximum-likelihood tree. The differences in like lihoods between these two trees, $\ell_{(2)} - \ell_{(1)}$, are 9.3 8.90, 6.12, 4.11, and 2.63 for the HKY85, JC69+I F81+Γ, K80+Γ, and HKY85+Γ models, respectivel When $2(\ell_{(2)} - \ell_{(1)})$ is compared with a χ^2 distributio with 1 df, the difference is highly significant ($P < 0.01$ under all the models, except for HKY85+Γ, for whic the test is only marginally significant ($0.01 < P < 0.05$ As tree (2) is quite likely to be the true tree (Hasegaw 1991), we might tentatively take this test as a measur of the discriminating power of the model (Bishop an Friday 1985), although we note that this test is not a

Table 2
Likelihoods and Estimates of Parameters for Different Trees under Different Models

Model and Parameter	Tree (1): (H, C, G)	(2): ((H, C), G)	(3): ((H, G), C)	(4): (H, (C, G))	$\ell_{(2)} - \ell_{(1)}$
HKY85	−2,196.96	−2,187.60	−2,196.96	−2,194.74	9.37**
$\hat{\kappa}$	11.89 ± 2.00	12.23 ± 2.13	11.89 ± 2.00	11.55 ± 1.95	
JC69+Γ	−2,415.24	−2,406.34	−2,414.06	−2,409.81	8.90**
$\hat{\alpha}$	0.75 ± 0.23	1.47 ± 0.74	0.88 ± 0.31	1.11 ± 0.45	
F81+Γ	−2,336.34	−2,330.22	−2,335.77	−2,332.29	6.12**
$\hat{\alpha}$	0.64 ± 0.19	1.10 ± 0.48	0.72 ± 0.24	0.88 ± 0.33	
K80+Γ (κ = 12)	−2,266.70	−2,262.59	−2,266.70	−2,266.59	4.11**
$\hat{\alpha}$	0.49 ± 0.14	0.67 ± 0.24	0.49 ± 0.14	0.51 ± 0.15	
HKY85+Γ (κ = 12)	−2,176.32	−2,173.69	−2,176.32	−2,176.29	2.63*
$\hat{\alpha}$	0.39 ± 0.09	0.46 ± 0.12	0.39 ± 0.09	0.39 ± 0.09	

NOTE.—Data are the 895-bp mtDNA sequences of human (H), chimpanzee (C), gorilla (G), and orangutan (O). ((H, C), G, O) (tree 2) is the maximum-likelihood tree under all the models.

* $P < 0.05$; $\chi^2_{0.05}$ (1 df) = 3.84.

** $P < 0.01$; $\chi^2_{0.01}$ (1 df) = 6.63.

evaluation of the reliability of tree (2). Thus simpler models, such as JC69+Γ, F81+Γ, K80+Γ, and HKY85, all seem to have stronger discriminating power than does HKY85+Γ. This is generally true for other data sets. For example, without exception, adding the gamma distribution to HKY85 leads to reduction of differences in likelihood among the possible tree topologies. We believe that the apparent discriminating power of simple but unrealistic models is an artefact and thus is unreliable (see below).

We note that estimates of κ for the HKY85 model remain almost constant for different tree topologies. This is also the case for other data sets analyzed; the range of κ estimates over the 15 bifurcating tree topologies and the star tree is 8.66–9.39 for the mtDNA sequences for five species and is 11.55–12.23 in the case of four species, 1.47–1.52 for the α- and β-globin genes, 1.78–1.80 for the ssRNAs, 0.96–0.97 for the bacterial glutamine synthetase genes, and 5.14–5.26 for the ψη-globin pseudogenes. The difference in estimates of κ between data sets 1 and 1′ is because the ratio is higher in branches leading to human, chimpanzee, and gorilla but is lower in those leading to gibbon and orangutan (results not shown).

The estimates of the α parameter of the gamma distribution, though more variable over trees under the three unrealistic substitution models, are also very stable under the HKY85+Γ model. For example, the range is 0.39–0.46 for the mtDNA sequences for four species, 0.22–0.29 for the α- and β-globin genes, 0.91–0.95 for the ssRNAs, and 0.51–0.68 for the ψη-globin genes. It is noteworthy that estimates of α are usually different for the F81+Γ model (κ = 1) and for the HKY85+Γ model (κ ≠ 1), which implies that estimates of α and κ are correlated. Figure 1 shows the likelihood surface for these two parameters for the mtDNA sequences for four species.

Estimates of Branch Lengths under Different Models

As different models most often give the same best tree for given data, it is interesting to see whether the estimates of branch lengths are also stable under different models. Table 3 lists the estimates of branch lengths for the maximum-likelihood tree for the mtDNA sequence data for four species, ((human, chimpanzee), gorilla, orangutan). If we take estimates from the HKY85+Γ model as the correct values, we see that all simpler models of nucleotide substitution—i.e., JC69+Γ, F81+Γ, and K80+Γ—will lead to underestimated branch lengths. The underestimation caused by ignoring the variation in substitution rates over sites (HKY85) is even more serious. In both cases the bias is more serious for longer branches. For data sets with more distantly related species, such errors are much more pronounced. The results here are consistent with findings from computer simulations on the estimation of sequence divergence. When a more complex model of nucleotide substitution is used to generate the two sequences and a simpler model is used in the estimation of the distance, then the estimate is always biased downward (e.g., as found by Tamura 1992). It has also been found that ignoring spatial rate variation leads to even more serious underestimation of sequence divergence (e.g., see Gillespie 1986).

Frequencies of Different Site Patterns

In the following section we examine the frequencies of different site patterns in order to understand how the

Table 3
Maximum-Likelihood Estimates of Branch Lengths under Different Models

	Branch Length (Ratio of Given Model to HKY85+Γ)[a]				
Models	→H	→C	HC↔GO	→G	→O
HKY85	0.0436 (0.83)	0.0522 (0.81)	0.0191 (0.84)	0.0529 (0.78)	0.1535 (0.56)
JC69+Γ	0.0438 (0.83)	0.0520 (0.81)	0.0195 (0.86)	0.0526 (0.78)	0.1585 (0.61)
F81+Γ	0.0446 (0.85)	0.0534 (0.83)	0.0192 (0.85)	0.0543 (0.81)	0.1579 (0.65)
K80+Γ	0.0482 (0.92)	0.0572 (0.88)	0.0193 (0.85)	0.0602 (0.89)	0.1915 (0.79)
HKY85+Γ	0.0525	0.0644	0.0227	0.0675	0.2416

NOTE.—Data are the 895-bp mtDNA sequences from human (H), chimpanzee (C), gorilla (G), and orangutan (O). The tree topology is ((H, C), G, O), which is supported by all the models.

[a] Branch lengths represent the expected number of nucleotide substitutions per site. Numbers in brackets are the ratio of branch length under the present model to that under the best model, HKY85+Γ.

HKY85 model is improved by adding the gamma distribution and how the discriminating power is reduced. The mtDNA data set for four species exhibits only 46 of $4^4 = 256$ possible site patterns; many site patterns simply do not appear. In table 4 we list these 46 site patterns, their observed frequencies, and their predicted frequencies under the HKY85 and HKY85+Γ models. Results from two tree topologies are listed for each of the two models. We can see that for frequent sites, such as the patterns of identity (TTTT, CCCC, etc.), $\log(p_i)$ is almost the same under the two models. The difference between the two models lies in rare and highly variable site patterns such as CCAA, ACAT, and GAAC, for which the likelihood values under HKY85+Γ are higher than those under HKY85. With the assumption of rate variation over sites, the expected frequencies of observation of such mutational hot spots are increased to match more closely their occurrence in the data. Comparison of the results for the two trees reveals that the decrease in discriminating power of HKY85+Γ is similarly caused by these highly variable sites. By assuming high rates of substitution at the sites, the rare site patterns can be easily explained even with a wrong tree topology.

Since it has been shown above that the HKY85+Γ model is significantly better than the HKY85 model for the mtDNA sequences for four species and that the HKY85+Γ model is an adequate description of the sequences' evolution, the values of $\log(p_i)$ for the HKY85+Γ model in table 4 can be expected to be more reliable than those for the HKY85 model. Consequently, the apparent decrease in discriminating power of the HKY85+Γ model is not actually due to weaker statistical properties: it is a more accurate reflection of an illusory discriminating power seen when an inadequate model (e.g., HKY85) is used. This finding also applies to other sequences and models, in which case the apparent discriminating power of simple but unrealistic and inadequate models is an artefact.

It is noteworthy that the contribution to the total likelihood by a site, $\log(p_i)$, depends greatly on how variable the site is. More-variable sites have low probabilities of occurrence and thus contribute more to the total likelihood (eq. [2]). Recently Williams and Fitch (1990) proposed a scheme for weighting sites in parsimony analyses. They judge how informative a site is by counting how frequent the site pattern is: the more frequent, the less informative. Schöniger and von Haeseler (1993) have found that such "combinational weighting" can also improve the performance of the neighbor-joining method with molecular sequence data. Goldman (1990) pointed out that parsimony analyses can be framed in terms of likelihood estimation under a simple stochastic model of change, but with some unreasonable assumptions. For example, there is no assumption of time structure in this formulation: the probability of a substitution occurring in a long time interval may be the same as that in a short time interval. The weighting of sites proposed by Williams and Fitch (1990) has a similar effect to the use of time structure, as in the Markov models currently used in maximum-likelihood analyses, which, as seen above, give greater weights to highly variable, less common, site patterns. Further study of the relationship between the contribution to the likelihood by each site and the variability of the site could lead to an even better weighting scheme. This would bring the treatment of data in the parsimony approach closer to that of the maximum-likelihood approach, although the justification for using the former is that it represents a computationally feasible approximation to the latter (Cavalli-Sforza and Edwards 1967).

Discussion

The results in this paper show that different models frequently produce the same best-supported tree for the same data: the maximum-likelihood approach seems robust to violation of some assumptions, at least as far

Table 4
Observed Frequencies of Different Site Patterns and Their Expected Values under Different Models and Tree Topologies

Site Pattern (Observed Frequency)	Expected Frequency (log p_i) for			
	HKY85		HKY85+Γ	
	((H, C), G)	((H, G), C)	((H, C), G)	((H, G), C)
AAAA (222)	234.10 (−1.34)	232.74 (−1.35)	232.29 (−1.35)	231.09 (−1.35)
CCCC (217)	215.00 (−1.43)	213.13 (−1.43)	221.73 (−1.40)	221.38 (−1.40)
TTTT (167)	150.69 (−1.78)	149.00 (−1.79)	159.57 (−1.72)	159.62 (−1.72)
GGGG (71)	61.40 (−2.68)	60.62 (−2.69)	64.40 (−2.63)	64.31 (−2.63)
TTTC (23)	28.68 (−3.44)	29.86 (−3.40)	22.54 (−3.68)	22.16 (−3.70)
CCCT (17)	30.13 (−3.39)	31.34 (−3.35)	24.84 (−3.58)	24.53 (−3.60)
AAAG (16)	13.00 (−4.23)	13.55 (−4.19)	12.14 (−4.30)	12.17 (−4.30)
CTCC (16)	10.67 (−4.43)	10.12 (−4.48)	8.17 (−4.70)	8.17 (−4.70)
CCTC (13)	11.28 (−4.37)	13.63 (−4.18)	9.21 (−4.58)	9.93 (−4.50)
TCCC (12)	9.04 (−4.59)	9.04 (−4.60)	7.02 (−4.85)	6.78 (−4.88)
AAGA (11)	4.92 (−5.20)	5.99 (−5.01)	4.63 (−5.26)	5.07 (−5.17)
GAAA (9)	3.97 (−5.42)	3.98 (−5.42)	3.55 (−5.53)	3.48 (−5.55)
CTTC (8)	2.07 (−6.07)	2.26 (−5.98)	3.30 (−5.60)	3.55 (−5.53)
AAAC (7)	3.78 (−5.47)	4.09 (−5.39)	4.64 (−5.26)	4.95 (−5.20)
GGGA (7)	11.39 (−4.36)	11.90 (−4.32)	9.13 (−4.58)	9.02 (−4.60)
CCCA (7)	3.49 (−5.55)	3.77 (−5.47)	3.84 (−5.45)	4.04 (−5.40)
TTCC (6)	5.97 (−5.01)	2.92 (−5.73)	6.01 (−5.00)	4.63 (−5.27)
TTCT (6)	10.31 (−4.46)	12.62 (−4.26)	7.56 (−4.77)	8.20 (−4.69)
AGAA (6)	4.68 (−5.25)	4.45 (−5.30)	4.13 (−5.38)	4.19 (−5.36)
CCTT (5)	5.57 (−5.08)	2.48 (−5.89)	5.44 (−5.10)	4.05 (−5.40)
TCTT (4)	9.82 (−4.51)	9.27 (−4.57)	6.72 (−4.89)	6.62 (−4.91)
AAAT (4)	2.90 (−5.73)	3.14 (−5.65)	3.56 (−5.53)	3.80 (−5.46)
CTCT (4)	2.00 (−6.10)	3.15 (−5.65)	3.27 (−5.61)	3.59 (−5.52)
TCTC (3)	2.28 (−5.97)	3.43 (−5.56)	3.64 (−5.50)	3.99 (−5.41)
CCAA (3)	0.45 (−7.59)	0.03 (−10.24)	0.45 (−7.59)	0.14 (−8.79)
GGAA (3)	2.34 (−5.95)	1.03 (−6.77)	2.49 (−5.89)	1.82 (−6.20)
TCCT (3)	1.87 (−6.17)	2.06 (−6.07)	3.03 (−5.69)	3.30 (−5.60)
TTTA (3)	2.55 (−5.86)	2.76 (−5.78)	2.61 (−5.84)	2.71 (−5.80)
GAGA (2)	0.77 (−7.06)	1.23 (−6.59)	1.34 (−6.50)	1.49 (−6.40)
ATAA (2)	0.95 (−6.85)	0.92 (−6.88)	0.88 (−6.93)	0.91 (−6.89)
AACA (2)	1.26 (−6.56)	1.67 (−6.28)	1.23 (−6.59)	1.53 (−6.37)
CCTG (2)	0.08 (−9.31)	0.08 (−9.28)	0.25 (−8.19)	0.28 (−8.06)
CCAC (1)	1.12 (−6.68)	1.48 (−6.40)	0.97 (−6.83)	1.19 (−6.62)
ACAT (1)	0.02 (−10.91)	0.04 (−10.01)	0.07 (−9.39)	0.10 (−9.11)
TCTA (1)	0.17 (−8.55)	0.20 (−8.42)	0.52 (−7.44)	0.63 (−7.25)
AGGA (1)	0.67 (−7.20)	0.73 (−7.12)	1.15 (−6.65)	1.24 (−6.58)
TTCA (1)	0.23 (−8.25)	0.24 (−8.22)	0.66 (−7.21)	0.74 (−7.10)
GAAG (1)	0.45 (−7.60)	0.51 (−7.47)	0.77 (−7.06)	0.88 (−6.93)
GAGG (1)	3.75 (−5.47)	3.52 (−5.54)	2.32 (−5.96)	2.25 (−5.99)
TAAG (1)	0.04 (−9.90)	0.05 (−9.81)	0.11 (−9.00)	0.12 (−8.90)
GGGC (1)	1.11 (−6.69)	1.20 (−6.61)	1.10 (−6.70)	1.13 (−6.67)
AGAC (1)	0.08 (−9.34)	0.09 (−9.21)	0.31 (−7.97)	0.38 (−7.76)
CTTT (1)	8.28 (−4.68)	8.27 (−4.68)	5.68 (−5.06)	5.37 (−5.12)
GAAC (1)	0.07 (−9.49)	0.07 (−9.39)	0.28 (−8.07)	0.34 (−7.87)
CAAA (1)	1.04 (−6.76)	1.08 (−6.72)	0.93 (−6.87)	0.92 (−6.88)
CCTA (1)	0.24 (−8.22)	0.25 (−8.19)	0.74 (−7.09)	0.84 (−6.97)

as reconstruction of tree topology is concerned. As long as the true tree is simple, in that neither great rate variation over lineages nor very long branches exist, and the data contain plenty of information, any reasonable approach might choose the right tree. However, caution is needed because false or too-simple models can be misleading about the reliability of the estimated tree, tending to suggest that the tree is significantly supported when, in fact, it cannot be.

Realistic formulation of the models is much more important when branch lengths need to be estimated and when evolutionary events are to be dated. For most

of the data sets analyzed in this paper, the HKY85+Γ model is significantly better than the others tested and gives substantially greater branch length estimates. It is interesting that in estimating some important parameters of molecular evolution, such as the transition/transversion ratio, κ, and the α parameter of the gamma distribution, knowledge of the true phylogeny is not very important, as long as a sufficiently realistic model of evolution is adopted.

Acknowledgments

We thank Drs. Masami Hasegawa and William Navidi for sending some of the sequence data analyzed in this paper.

LITERATURE CITED

BISHOP, M. J., and A. E. FRIDAY. 1985. Evolutionary trees from nucleic acid and protein sequences. Proc. R. Soc. Lond. [Biol.] **226**:271–302.

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.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedures. Evolution **32**:550–570.

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

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

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.

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

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

HASEGAWA, M. 1991. Molecular phylogeny and man's place in Hominoidea. J. Anthropol. Soc. Nippon **99**:49–61.

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

HOLMQUIST, R., M. GOODMAN, T. CONRY, and J. CZELUSNIAK. 1983. The spatial distribution of fixed mutations within genes coding for proteins. J. Mol. Evol. **19**:137–448.

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

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

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

KOCHER, T. D., and A. C. WILSON. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding region. Pp. 391–413 *in* S. OSAWA and T. HONJO, eds. Evolution of life: fossils, molecules and culture. Springer, Tokyo.

MIYAMOTO, M. M., J. L. SLIGHTON, and M. GOODMAN. 1987. Phylogenetic relations of humans and African apes from DNA sequences in the ψη-globin region. Science **238**:369–373.

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.

PESOLE, G., M. P. BOZZETTI, C. LANAVE, G. PREPARATA, and C. SACCONE. 1991. Glutamine synthetase gene evolution: a good molecular clock. Proc. Natl. Acad. Sci. USA **88**:522–526.

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.

SCHÖNIGER, M., and A. VON HAESELER. 1993. A simple method to improve the reliability of tree reconstructions. Mol. Biol. Evol. **10**:471–483.

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.

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

WILLIAMS, P. L., and W. M. FITCH. 1990. Phylogeny determination using dynamically weighted parsimony method. Methods Enzymol. **183**:615–626.

YANG, Z. 1992. Variations of substitution rates and estimation of evolutionary distances of DNA Sequences. Ph.D. thesis, Beijing Agricultural University, Beijing.

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

MASATOSHI NEI, reviewing editor

Received May 23, 1993

Accepted November 15, 1993