Syst. Biol. 44(3):384–399, 1995

MAXIMUM LIKELIHOOD TREES FROM DNA SEQUENCES: A PECULIAR STATISTICAL ESTIMATION PROBLEM

ZIHENG YANG,[1,4] NICK GOLDMAN,[2,5] AND ADRIAN FRIDAY[3]

[1]*Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, England*
[2]*Laboratory of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, England*
[3]*Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, England*

Abstract.—The parameter space of the phylogenetic tree estimation problem consists of three components, T, $\mathbf{t}$, and θ. The tree topology T is a discrete entity that is not a proper statistical parameter but that can nevertheless be estimated using the maximum likelihood criterion. Its role is to specify the branch length parameters and the form of the likelihood function(s). Branch lengths $\mathbf{t}$ are conditional on T and are meaningful only for specific values of T. Parameters θ in the model of nucleotide substitution are common to all the tree topologies and represent such values as the transition/transversion rate ratio. T and $\mathbf{t}$ thus represent the tree, and θ represents the model. With typical DNA sequence data, differences in T have only a small effect on the likelihood, but changing θ will influence the likelihood greatly. Estimates of θ are also found to be insensitive to T, making it possible to obtain reliable estimates of θ and to perform tests concerning the model (θ) even if knowledge of the evolutionary relationship (T) is not available. In contrast, tests concerning $\mathbf{t}$, such as testing the existence of a molecular clock, appear to be more difficult to perform when the true topology is unknown. In this paper, we explore the peculiarity of the parameter space of the tree estimation problem and suggest methods for overcoming some difficulties involved with tests concerning the model. We also address difficulties concerning hypothesis testing on T, i.e., evaluation of the reliability of the estimated tree topology. We note that estimation of and particularly tests concerning T depend critically on the assumed model. [Maximum likelihood; models; parameter space; consistency; sampling errors; hypothesis testing; nucleotide substitution; phylogeny estimation; molecular systematics; molecular clock.]

> Acceptance of the theory of evolution as the means of explaining observed similarities and differences among organisms invites the construction of trees of descent purporting to show evolutionary relationships.
>
> —Cavalli-Sforza and Edwards, 1967:550

Phylogenetic tree estimation has been considered a statistical estimation problem since the pioneering work of Edwards and Cavalli-Sforza (Edwards and Cavalli-Sforza, 1963, 1964; Cavalli-Sforza and Edwards, 1964, 1966, 1967; Edwards, 1970) and has been identified as producing novel statistical problems almost from the start (Neyman, 1971). Analyzing gene frequency data from different human populations, Edwards and Cavalli-Sforza used a Yule process to model the branching pattern and Brownian motion to model the drift of (transformed) gene frequencies within populations. In this formulation, the tree topology and its branch lengths were considered random variables, the distributions of which were specified by the parameters in the model. The parameters of the Yule process and the Brownian motion process could in principle be estimated from the likelihood function, and the tree topology and branch lengths could be estimated using their conditional distributions given the data (Edwards, 1970; Thompson, 1975).

Although the Yule process is a very simple description of the branching process, it caused insurmountable computational problems. Perceiving that the likelihood function was a sum over all tree topologies, which were discrete, Thompson (1975) concluded that the model was also theoretically infeasible—certainly it remained unanalyzed. The Yule process was dropped by Felsenstein (1973, 1981) when

[4] Present address: Institute of Molecular Evolutionary Genetics, 328 Mueller Laboratory, University Park, Pennsylvania 16802, USA.

[5] To whom correspondence should be addressed. E-mail: n_goldma@nimr.mrc.ac.uk.

he addressed the problem of evolutionary tree estimation from discrete morphological characters and, later, from nucleic acid sequence data and by Thompson (1975) working with gene frequency data. For discrete characters, a Markov process was used (Felsenstein, 1973, 1981) to model the changes among the possible character states in place of the Brownian motion model used for gene frequency data. In this formulation, the trees and branch lengths are not random variables; they are parameters and are estimated from the likelihood function(s) (Felsenstein, 1973, 1981; Thompson, 1975; Goldman, 1990). The result is a maximum likelihood (ML) method of tree estimation.

Nei (1987) appears to have been the first to point out the difference between this ML tree estimation method and more traditional ML methodology. Seeing that different tree topologies have different likelihood functions, Nei implicitly questioned the statistical consistency of the method in this context. Comparative studies using both computer simulations (Hasegawa and Yano, 1984; Fukami-Kobayashi and Tateno, 1991; Hasegawa et al., 1991; Kuhner and Felsenstein, 1994; Huelsenbeck, 1995; Gaut and Lewis, 1995) and numerical analysis (Yang, 1994b) suggest that the ML method of tree estimation is preferable to other tree reconstruction methods that are currently available. Yang (1994b) indicated the nature of a formal proof of its statistical consistency.

Although ML point estimation of the tree topology and other parameters is consistent and efficient, hypothesis testing concerning both the adequacy of the model and the reliability of the ML tree seems to be full of difficulties, both theoretical and practical. In view of the complexities caused by the tree topology parameter and the peculiarity of the data, which put traditional distributional approximations into question, Goldman (1993a) used computer simulations to derive the distributions of the statistics used to test models. Such an approach by computer simulation involves very heavy computation.

In this paper, we examine the irregularity of the parameter space of the tree estimation problem, study the effects of this irregularity on tests of models, and suggest approximate methods for testing a model's general adequacy and for comparing two parametric models. Difficulties in the evaluation of the reliability of the ML tree are also examined.

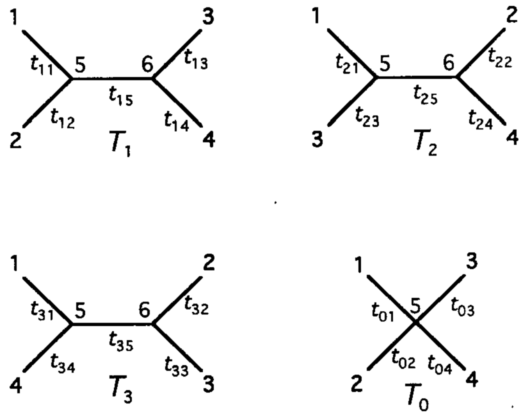

FIGURE 1. The possible unrooted tree topologies and their branch lengths for four species, used for describing the parameter space of the phylogenetic tree estimation problem. When the trees are used in Figure 3a, species 1, 2, 3, and 4 correspond to human, chimpanzee, gorilla, and orangutan, respectively. The branch length is measured by the expected number of nucleotide substitutions per site accumulated along the branch.

PARAMETER SPACE OF THE TREE ESTIMATION PROBLEM

Figure 1 shows all possible unrooted tree topologies for four species, where T_0 is the star tree or "big bang" tree. Different symbols are used for branch lengths in different trees. Branch lengths in T_j will be denoted as $\mathbf{t}_j$; for example, $\mathbf{t}_1 = \{t_{11}, t_{12}, t_{13}, t_{14}, t_{15}\}$. The true tree for any four given species is assumed to be one of T_1, T_2, or T_3 and will be denoted T^*.

We base our discussions on the Markov model of nucleotide substitution proposed by Hasegawa et al. (1985), referred to as HKY85. According to this model, the probability of nucleotide i changing into nucleotide j ($j \neq i$) in a small time interval Δt is given by

$$Q_{ij}\Delta t = \begin{cases} \kappa\pi_j\mu\Delta t \text{ (for transitions: T} \leftrightarrow \text{C, A} \leftrightarrow \text{G)} \\ \pi_j\mu\Delta t \text{ (for transversions: T, C} \leftrightarrow \text{A, G),} \end{cases} \quad (1)$$

where π_j is the frequency of nucleotide j when the process is in equilibrium, with $\Sigma_j\pi_j = 1$, and the scale factor μ is chosen such that the average rate of substitution is 1. Parameter κ (α/β in the notation of Hasegawa et al., 1985) represents the instantaneous transition/transversion rate ratio; $\kappa > 1$ means that transitions occur at higher rates than transversions. Other models used for comparison are those of Jukes and Cantor (1969), Kimura (1980), and Felsenstein (1981), referred to as JC69, K80, and F81, respectively. They are all special cases of HKY85, which is itself a special case of the most general reversible process model (REV), which involves eight independent parameters (see Yang, 1994a). These models assume a single rate of substitution over nucleotide sites.

It is also possible to make the assumption that rates of substitution are drawn from a gamma distribution. This assumption permits modeling of substitution rate heterogeneity. The gamma distribution with parameters α and β has a mean of α/β and a variance of α/β^2. In the current context, β is a trivial scale factor and can be fixed equal to α to give a mean of 1 and a variance of $1/\alpha$. Values of α less than approximately 0.5 mean the gamma distribution has a reverse-J shape and imply strong rate variation, whereas values greater than 1 or 2 imply a mostly constant rate over sites (Fig. 2). Yang (1993) described the computational implementation of this model, which in conjunction with the above Markov models is represented, for example, as HKY85+Γ.

Parameters in the substitution model are common to all the tree topologies and will be collectively denoted θ. For example, with HKY85+Γ, $\theta = \{\pi_T, \pi_C, \pi_A, \kappa, \alpha\}$, where α is the shape parameter of the gamma distribution (π_G need not be included because it is defined by the requirement $\Sigma_j\pi_j = 1$).

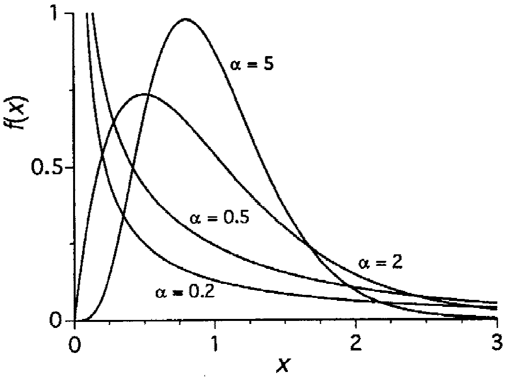

FIGURE 2. Examples of the gamma distribution, used to model heterogeneity of substitution rates across sites. Four values of the shape parameter α are illustrated; all distributions have a mean of 1 and a variance of 1/α.

Irregularity of the Parameter Space

Let s be the number of species and n the length of sequence. Then there are 4^s possible site patterns (Cavender, 1989; Goldman, 1993a). Let the observed number of occurrences of the ith site pattern be n_i, with $\Sigma n_i = n$. Under the assumption that sites evolve independently, data for different sites follow a multinomial distribution. Given the data, the likelihood function(s) is proportional to the probability of observing the data given the model and tree and can be written as

$$L(T_j, \mathbf{t}_j, \theta) = C \cdot \prod_{i=1}^{4^s} [p_i^{(T_j)}(\mathbf{t}_j, \theta)]^{n_i}, \quad (2)$$

where p_i, the probability of observing the ith site pattern, is a function of $\mathbf{t}_j$ and θ. The functional form of p_i depends on the tree topology T_j. T_j is discrete, and its role is to specify the set of branch lengths $\mathbf{t}_j$ and to change p_i and thus the likelihood function. $C = n!/\Pi(n_i!)$ is a proportionality constant; in practical implementation (e.g., Felsenstein, 1981), C is ignored and $\ell_j = \log[L(T_j, \mathbf{t}_j, \theta)]$ is maximized with respect to $\mathbf{t}_j$ and θ for each of the tree topologies T_j, leading to as many (maximum) likelihood values as the number of trees. The tree with the highest (maximum) likelihood, say $\hat{T}$, is taken as the estimate of T^*, and estimates $\hat{t}$ and $\hat{\theta}$ are obtained from the ML tree $\hat{T}$.

The irregularity of this parameter space may be better understood by considering an even more irregular case. Suppose that a random sample of size n is drawn from one of the following distributions with unknown parameters: the normal $N(\mu, \sigma^2)$ with a mean of μ and a variance of σ^2, the log-normal $LN(\mu, \sigma)$ with scale parameter μ and shape parameter σ, and the gamma $G(\alpha, \beta)$ with shape parameter α and scale parameter β. The problem is to estimate which distribution the sample is drawn from and to estimate its parameters. An extra problem is to provide a probabilistic statement concerning the reliability of the estimated distribution. The distributions in this example correspond to tree topologies, and the parameters of each distribution correspond to the branch lengths in each tree. The distribution does not correspond to a parameter; in the same way, the tree topology is not a proper statistical parameter of the phylogenetic tree estimation problem. The ML method does not appear directly usable to estimate the true distribution because the likelihood function is not defined.

The tree estimation problem is more conventional, however, because the probability functions of observing the data for different topologies share a general form (Equation 2) and therefore share the same proportionality constant C. Thus, the likelihood function(s) is well defined, the likelihood values for different trees are comparable, and the ML estimator of T^* is consistent (i.e., as $n \rightarrow \infty$, Prob$(\hat{T} = T^*) \rightarrow 1$ [Yang, 1994b]).

Likelihood Surface as a Function of θ

Another way to look at the irregularity of the parameter space is through the likelihood surface as a function of the common parameters in the model, θ. An example is shown in Figure 3 using the $\psi\eta$-globin pseudogenes of human (H), chimpanzee (C), gorilla (G), and orangutan (O) (Miyamoto et al., 1987). The sequences have 6,166 sites. The HKY85+Γ model is used, with the frequency parameters estimated directly using the averages of observed frequencies: $\hat{\pi}_T = 0.3074$, $\hat{\pi}_C = 0.1852$, $\hat{\pi}_A = 0.3073$, $\hat{\pi}_G = 0.2001$. The likelihood is calculated as a function of parameters κ and α, maximized over tree topologies and their branch lengths.

The likelihood surface (Fig. 3a) may be divided into three regions in the plane specified by κ and α, within each of which one tree topology is optimal with different estimates of branch lengths. We denote the possible (unrooted) tree topologies as T_0 = (H, C, G, O), T_1 = ((H, C)G, O), T_2 = ((H, G)C, O), and T_3 = (H, O(C, G)) (see Fig. 1). In the region where T_0 is best, to the left of the red curve, none of the three bifurcating trees is better than the star tree, whereas traversing the white curve, a delicate balance is maintained such that tree topologies T_1 (optimal above the white curve) and T_3 (optimal between the red and white curves) have identical likelihood values and both are ML trees. The superimposed color contours represent the length of the interior branch in the best tree topology: t_{15} in the region for T_1 and t_{35} in the region for T_3, but in the region for T_0 no interior branch exists (see Fig. 1). Intuitively, the interior branch length in the ML tree appears to measure the confidence we can put in the estimated topology, but this has been shown not to be the case (Yang, 1994b). When α gets smaller, the interior branch lengths in T_1 and T_3 get smaller, indicating that use of a small α leads to reduction in the difference of likelihood values among the trees.

Figure 3b shows a schematic cross section through the likelihood surface of Figure 3a. The likelihood curves for all the trees are shown, indicating that their maximum is a continuous function of α but is not smooth (differentiable) at the red and white boundary curves of Figure 3a because likelihood values in different regions are calculated from different likelihood functions.

Figure 3 illustrates that the likelihood surface, here shown as a function of κ and α, is not smooth everywhere. Also, it is possible for the star tree to be the best tree and for two bifurcating trees to have iden-

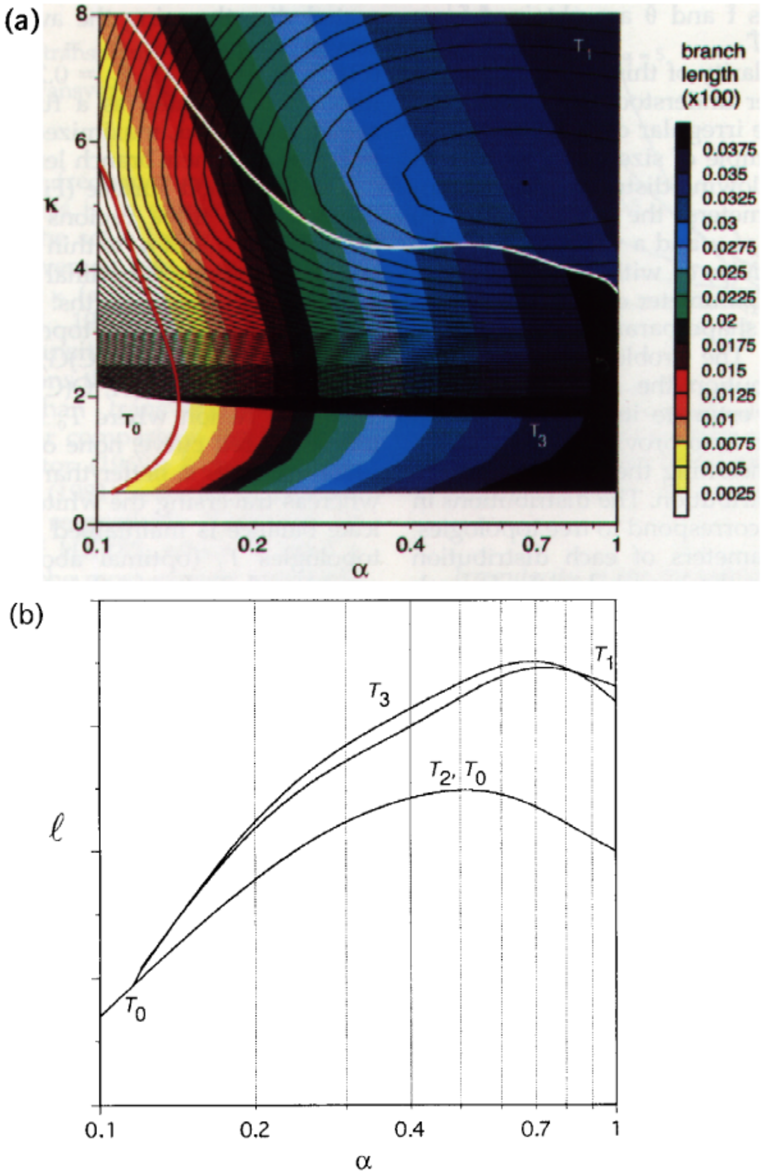

FIGURE 3. The irregular likelihood surface of the phylogenetic tree estimation problem. (a) A "five-dimensional" representation of the likelihood surface and interior branch length estimate. The plane is determined by parameters κ and α. Black contours show likelihoods, maximized over all the tree topologies and branch lengths; those for very low likelihood values (roughly for the region $\kappa < 2$) are omitted to improve clarity. The overall highest likelihood value occurs at $\hat{\kappa} = 5.35$, $\hat{\alpha} = 0.66$ ($\ell = -10127.36$). The three regions delineated by the red and white curves indicate that different tree topologies are supported by the data for different given values of κ and α. (b) A schematic cross section through the likelihood surface of Figure 3a. Parameter κ is fixed at 4, and the ℓ_i, as functions of α, are maximized over the branch lengths in each tree.

tical highest likelihood values (at the white boundary). Although in this example there is a unique local maximum to the likelihood surface as a function of κ and α, it is not certain that this will always be the case. This question is analogous to the question of whether there is a unique maximum to the likelihood surface as a function of trees and branch lengths. In this latter problem, there may be multiple maxima spread among different trees (Bishop and Friday, 1988; Holmes, 1991) or even for a single tree (Steel, 1994; contra Fukami and Tateno, 1989). The problem of multiple maxima rarely presents itself in the analysis of real data, particularly if maximization of the likelihood is performed independently for different trees. Furthermore, different model assumptions can lead to support of different tree topologies: $\alpha = \infty$ corresponds to the model of a single rate for all sites, whereas the HKY85 model reduces to the F81 model if $\kappa = 1$.

The estimated interior branch length remains quite similar as the ML topology changes (crossing the white boundary). However, this is only a subjective judgment, with no theoretical results to support it.

Criteria for Evaluating the Performance of a Tree Estimation Method

Although the tree topology T is not a proper statistical parameter, it can be thought of as one, and appropriate measures for evaluating the properties of its estimators, i.e., tree estimation methods, can be defined. Of course, its difference from a conventional statistical parameter should be kept in mind.

Consistency and Robustness

The consistency of $\hat{T}$ can be defined in the usual statistical sense, i.e., $\hat{T}$ is a consistent estimator of T^* if Prob($\hat{T} = T^*$) $\rightarrow$ 1 as $n \rightarrow \infty$ (Felsenstein, 1978). As far as estimation of the tree topology alone is concerned, it is possible for an estimator (a tree estimation method) to be consistent even if the model used for data analysis is wrong (DeBry, 1992). The variety of conditions for which this is the case gives a measure of the robustness of the method when its assumptions are violated. Yang (1994b) presented examples in which ML (joint) analysis was more robust than a least squares (pairwise) comparison.

Sampling Error or Efficiency

Some properties for an estimator of a regular statistical parameter, such as unbiasedness, do not seem to make sense for an estimator of the tree topology. However, when an estimator of the tree topology is consistent, its efficiency or sampling error can be measured by the probability that the estimated tree is the true tree, Prob($\hat{T} = T^*$); this measurement is analogous to the variance of an estimator of a regular parameter. Nevertheless, a method assuming a wrong model may still be consistent and may have smaller sampling errors than one using the right model. When the model is misspecified, however, parameters **t** and θ will be estimated with systematic errors, and more importantly, hypothesis testing concerning both trees and models can be misleading. The robustness of a tree estimation method is not a justification for the use of wrong models.

In computer simulation studies, inconsistency of the estimation method, sampling errors due to finite sequence length, and sampling errors due to limited number of repetitions of the simulation are confounded. Few attempts have been made to discriminate among these different sources of error.

Estimating the Sampling Error of the ML Tree

The sampling error of an estimated tree topology, as defined above, appears very difficult to estimate in practice. The only approach currently available for this purpose is nonparametric bootstrapping (Felsenstein, 1985). A natural measure of the sampling error of $\hat{T}$ is the probability, P, that $\hat{T}$ is the true tree. Then $1 - P$ will be the probability of committing a Type I error if we draw the conclusion that the estimated tree is the true tree. In our discussion, one probability is assigned to one whole tree topology; the sum of these probabilities over all trees is 1. Many authors assign such probabilities to every interior node in an estimated tree. Although such probabilities do seem to provide some measure of the confidence in the monophyletic relationship at the nodes, their exact meaning is not clear. Because many probabilities are calculated from one analysis, it seems too demanding to expect all of them simultaneously to mean what each one independently is supposed to mean.

Other than in simulations, the true tree is not known. Instead, we can examine the probability that the estimated tree will be chosen as the best tree in a finite sample. If we could draw samples from the real process (the original data constitute one such sample), we would be able to analyze each of the samples and calculate this probability. The bootstrap draws samples from the observed data, and each sample is analyzed in exactly the same way as the original data. The bootstrap probability, P_B, is then determined as the proportion of bootstrap samples for which the tree concerned is chosen as the best tree. Recent theoretical study (Zharkikh and Li, 1992) and computer simulations (Hillis and Bull, 1993) suggest that P_B is a poor, sometimes very poor, estimator of P; it tends to overestimate P when P is small and to underestimate P when P is large. Felsenstein and Kishino (1993) argued that, although P_B is a poor estimator of P, $1 - P_B$ can be a good estimator of $1 - P$, which seems to need more justification.

The difficulty appears to be due to the irregularity of the tree estimation problem. For other problems, more powerful parametric methods may be constructed, but this seems very difficult for the tree estimation problem. In fact, even simulations (parametric bootstrapping) cannot easily be performed if there are more than two trees that could be the true one because of the complexity of the null hypothesis; at any rate, we cannot simply undertake simulation with the null assumption that the ML tree is the true tree.

For the ML method, Kishino et al. (1990) suggested two approximate methods for calculating P_B. The first, resampling estimated log likelihoods (RELL), uses the parameter estimates from the original data to calculate the likelihood for each bootstrap sample instead of estimating parameters from every bootstrap sample separately, a significant saving in computation. This practice, however, can have a profound effect (Hall and Wilson, 1991), considerably reducing the power of the test. The second method makes use of a multivariate normal distribution (MND) to approximate the distribution of the calculated likelihoods of the trees. Although we have found the asymptotics of estimates of parameters **t** and θ to be quite reliable, an acceptable normal approximation to the calculated likelihood values seems to need much more data. The accuracy of approximation seems to be influenced by the skewness of the distribution of the data, which is largely determined by the overall amount of evolution (Yang, 1994b). This case is similar to using a normal distribution to approximate a binomial distribution.

The relationships among the many probabilities mentioned above are not clear, nor is it clear what factors affect the approximations. Overall, the RELL and MND methods seem to do a better job in approximating P_B than P_B does in approximating P (Hasegawa and Kishino, 1994).

Tests of Models

Test of the General Adequacy of a Model

Traditionally, the fit of a model to data in the case of the multinomial distribution can be tested by using the likelihood ratio statistic,

$$D = \sum_{i=1}^{k} 2n_i \cdot \log \frac{n_i}{np_i}, \qquad (3)$$

or the Pearson X^2 statistic,

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad (4)$$

where $k = 4^s$ is the number of categories. Both statistics are asymptotically χ^2_{k-1} distributed when the model is correct.

However, with phylogenetic analysis of DNA sequence data, there are two difficulties with the χ^2 approximation (Goldman, 1991, 1993a). The first problem is caused by the tree topology. If the true tree were known, its use (whatever the ML tree derived from the data) would avoid uncertainty over the distribution of the statistic concerned. Most often, we do not have such knowledge and the available alternative is the use of the ML tree, but in this case the choice among tree topologies contributes to the likelihood. Goldman (1993a) pointed out the difficulty of counting the number of degrees of freedom involved with the tree topology parameter. If the likelihood is maximized over the trees, neither D nor X^2 can be expected to be χ^2 distributed because the regularity conditions required for this asymptotic approximation (e.g., Kendall and Stuart, 1979:445–446) are not met; in cases when the ML tree is not the true tree, the expected values (np_i) in Equations 3 and 4 will be calculated using the wrong formulae. When the amount of data is so large that we have virtually no doubt about the tree topology (i.e., when the ML tree is the true tree with $P \cong 1$), the difficulty effectively disappears. However, we will almost always have enough data to reject wrong models long before we have enough data to estimate the true topology with confidence.

Despite this theoretical difficulty, the problem can almost be ignored in practice. When the tree topology is in doubt, the likelihoods of several reasonable trees, including the ML tree and (presumably) the true tree, are very similar (Yang et al., 1994; Yang, Goldman, and Friday, unpubl.), which is why it is so difficult to estimate the true tree with confidence. The approximation of the likelihood of the ML tree to that of the true tree is acceptable; the effect of this close approximation is that the test will be conservative, slightly favoring the model under test.

The second difficulty with using a χ^2 distribution concerns a peculiarity of the data. These are assumed to be a sample from a multinomial distribution, yet the number of categories of the distribution, 4^s, is often larger than the number of data points, n. Furthermore, for typical DNA sequence data most of the data points are clustered in the four "constant" categories characterized by the occurrence of identical nucleotides for all the species. For closely related sequences, those four categories can account for >90% of the data points. A consequence is that we have very many categories with very few or no data points assigned to them, which seems to have a drastic effect on the χ^2 approximation, tending to result in rejection of the model much too readily (Reeves, 1992; Goldman, 1993a). An appropriate approach to this problem is to combine some of the categories so that the expected number in each category is, say, larger than 5.

One strategy is to combine into larger categories those data points (site patterns) that have similar probabilities, which will maintain as much power as possible in the resulting test. Our strategy refers to the JC69 model, under which many different site patterns have equal probabilities (Saitou, 1988; corrected by Yang, 1994b). The procedure is illustrated in Table 1, using as an example tests of the HKY85 and REV models for the evolution of the $\psi\eta$-globin genes of human, chimpanzee, gorilla, orangutan, rhesus monkey, and spider monkey (Miyamoto et al., 1987, 1988; Fitch et al., 1988). The first four sequences were used in Figure 3.

The first category includes all the site patterns where there are more than two different nucleotides in different species and therefore includes the most variable sites. This large category includes $4^s - [4 + (2^{s-1} - 1) \times 12] = 4^s - 3 \times 2^{s+1} + 8$ different site patterns. Categories 2–5 consist of the four "constant" patterns. All the remaining site patterns have exactly two different nucleotides and may be arranged into categories according to which sequences have one nucleotide and which the other, irrespective of what those nucleotides are. For the six-species example, these categories may be represented 000001, 000010, 000011, . . . , 011111, with 0 and 1 standing for any pair of different nucleotides. Each of these categories includes 12

TABLE 1. Goodness-of-fit tests of the HKY85 model and the general reversible process model (REV), using the six species ψη-globin gene sequences (6,166 nucleotides each). The categories are labeled using T, C, A, and G when they comprise one site pattern or using 0 and 1 standing for any pair of different nucleotides when they comprise a combination of 12 site patterns. The observed numbers of occurrences (n_i) are shown with the expected numbers (np_i) under the two models. The residuals are calculated as $r_i = (np_i - n_i)/(np_i)^{1/2}$.

Category no. and pattern	n_i	HKY85 np_i	HKY85 r_i	REV np_i	REV r_i
Most variable patterns					
1 ------[a]	45	30.48	−2.630	32.82	−2.126
Constant patterns					
2 TTTTTT	1,670	1635.77	−0.846	1661.08	−0.219
3 CCCCCC	903	919.31	0.538	897.66	−0.178
4 AAAAAA	1,670	1628.37	−1.032	1647.68	−0.550
5 GGGGGG	966	991.51	0.810	971.40	0.173
Two different nucleotides, one site pattern per category					
6 TTTTTC	46	67.09	2.575	67.14	2.580
7 TTTTTA	17	24.53	1.520	13.77	−0.870
8 TTTTTG	16	15.81	−0.048	16.74	0.181
9 CCCCCT	67	64.50	−0.311	63.13	−0.487
10 CCCCCA	16	14.16	−0.489	16.23	0.057
11 CCCCCG	16	9.13	−2.274	16.52	0.128
12 AAAAAT	12	24.45	2.518	13.69	0.457
13 AAAAAC	15	14.66	−0.089	17.14	0.517
14 AAAAAG	56	72.02	1.888	72.87	1.976
15 GGGGGT	18	15.25	−0.704	15.81	−0.551
16 GGGGGC	19	9.14	−3.261	16.57	−0.597
17 GGGGGA	76	69.55	−0.773	69.09	−0.831
Two different nucleotides, many site patterns per category					
18 000010	213	220.70	−0.035	219.77	−0.031
19 000011	74	78.53	0.511	77.58	0.406
20 000100	78	83.40	0.591	82.78	0.525
21 ------[b]	16	7.71	−2.986	7.69	−2.997
22 000111	27	33.74	1.160	33.57	1.134
23 001000	36	45.30	1.382	44.97	1.338
24 ------[c]	13	6.91	−2.317	6.88	−2.333
25 010000	45	47.31	0.336	46.98	0.289
26 ------[d]	11	7.25	−1.393	7.20	−1.416
27 011111	25	29.44	0.818	29.25	0.786
Totals	6,166		X^2 = 66.87 D = 62.04		X^2 = 39.22 D − 36.51

[a] More than two different nucleotides at a site.

[b] Patterns 000101 and 000110 combined.

[c] Patterns 001001, 001010, 001011, 001100, 001101, 001110, and 001111 combined.

[d] Patterns 010001–011110 combined.

site patterns. In the example of Table 1, some of these categories are further combined so that the expected number in each category exceeds 5 (Table 1, categories 21, 24, 26). When the sequences are arranged in such an order that the outgroups appear last, as in Table 1, categories such as 000001, 000010, and 000100 will typically have higher frequencies than the others and may be separated to achieve more categories. In Table 1, 000001 is separated into 12 categories representing its 12 component site patterns.

The problem emerges of what estimates of **t** and θ should be used to calculate the expected numbers (np_i) in Equations 3 or 4. One method is to reestimate the parameters **t** and θ using the combined data, minimizing either the statistic D or X^2 (the maximum likelihood criterion or the minimum chi-square criterion). Estimates of parameters obtained in this way will be different from those obtained from the original data. Then D or X^2 can be compared with a χ^2 critical value with df = $k - p - 1$, where k is the number of categories of the combined data and p is the number of parameters in the model being tested.

An alternative approach, adopted here because it involves less computation, is to use estimates of **t** and θ obtained from the original data. In this case, X^2 or D are no longer χ^2 distributed but rather are bounded between χ^2_{k-p-1} and χ^2_{k-1} (Kendall and Stuart, 1979:450–455). The test is then considered significant only if the observed value is larger than χ^2_{k-1}. The test will be conservative if p is much larger than 1, when χ^2_{k-p-1} and χ^2_{k-1} are quite different.

Because there are $k = 27$ categories in Table 1, we have $\chi^2_{0.01} = 45.64$ for df = $k - 1 = 26$. The HKY85 model is rejected when either of the two statistics is used ($X^2 = 66.87$, $D = 62.04$), whereas REV is not rejected ($X^2 = 39.22$, $D = 36.51$). The same conclusion holds when one (000010) or two (000010, 000100) further categories are separated into their component site patterns.

Comparison of Two Parametric Models

Tests in this category include, for example, comparison of F81 and HKY85 or of HKY85 and HKY85+Γ, etc. Traditionally, such a test is performed using the χ^2 approximation of the likelihood ratio statistic. Taking the comparison of F81 with HKY85 as an example, the null and alternative hypotheses are H_0: $\kappa = 1$, H_1: $\kappa \geq 0$. Suppose the likelihood under F81 is ℓ_0 and that under HKY85 is ℓ_1. Then $D = 2\Delta\ell = 2(\ell_1 - \ell_0)$ will be asymptotically χ^2 distributed with df = 1 if H_0 is true.

With DNA sequence data, different models may lead to support for different tree topologies, and even if the different models favor the same tree it may still be a wrong one. When ℓ_0 and ℓ_1 are not both calculated using the true tree, D can no longer be expected to be χ^2 distributed. Strictly speaking, the asymptotic properties of ML estimates of θ, obtained always from the ML tree, are now questionable because the regularity conditions leading to such properties (e.g., Kendall and Stuart, 1979:38–81) are not met. Using the analogy of distributions once again, suppose that all the samples are drawn from a normal distribution but that some samples are analyzed under the assumption that the sample is gamma distributed, depending on our best estimate of the distribution. In this case, the model is misspecified, and estimates of parameters involve systematic errors. Nevertheless, the likelihood values of a reasonable tree, say the ML tree under H_1, are in practice acceptable approximations of the likelihood values of the true tree, and therefore the difficulty caused by the lack of confidence in the tree can be ignored.

Goldman (1991, 1993a) performed such tests using a parametric bootstrapping method, generating samples of data by Monte Carlo simulation under H_0 and for each sample maximizing the likelihoods over trees for both models. In all cases, 100 simulations were done to derive the theoretical distributions of the statistics. We reexamined Goldman's (1991, 1993a) results and found that the χ^2 distribution does appear to give good approximations. Figure 4 shows an example in which F81 is compared with HKY85 using the ψη-globin pseudogenes of human, chimpanzee, gorilla, and orangutan, the same data analyzed in Figure 3. The calculated likelihood values under the two models are $\ell_0 = -10130.14$ under F81 and $\ell_1 = -10221.81$ under HKY85, with $D = 2\Delta\ell = 2 \times 91.67 = 183.34$. Monte Carlo samples were generated by "evolving" the sequences along the tree ((H, C)G, O), the ML tree under HKY85, using estimates from the data of **t** and θ for this tree under F81. Each of the 500 samples was analyzed in the same way as the original set of data. Results in Figure 4 were obtained using a single tree for both models. The results (not shown) obtained using the likelihood of the ML tree under each model for each sample are virtually the same, as expected.

The sparseness of the data does not seem to influence tests concerning two

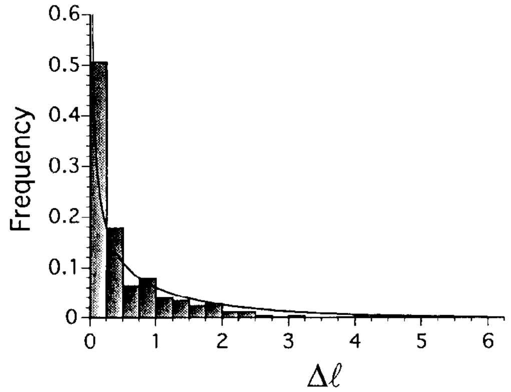

FIGURE 4. The distribution of the likelihood ratio statistic, $\Delta\ell = (\ell_1 - \ell_0)$, for comparison of the F81 and HKY85 models, obtained by 500 Monte Carlo simulations. The 6,166-bp ψη-globin genes of human (H), chimpanzee (C), gorilla (G), and orangutan (O) are used, with the likelihood calculated using only one tree under both models, i.e., ((H, C)G, O), the ML tree under HKY85. The continuous curve is the χ^2 distribution with df = 1, scaled appropriately, which seems to be an acceptable approximation of the simulated distribution. The observed value of the test statistic is $\Delta\ell = -10221.81 \quad (\ 10130.14) = 91.67$; the F81 model is rejected.

parametric models, possibly because the likelihoods under both models are affected in roughly the same way.

When H_0 and H_1 differ in only one parameter, as is the case for comparison of F81 with HKY85, the two models can be compared by examining the variance of that parameter, as estimated by the curvature method (Kendall and Stuart, 1979:45–49), which relies on the asymptotic normality of ML estimates. The difficulty caused by the tree topology can once again be ignored, the justification being the observation that estimates of θ are very similar for all trees (Yang et al., 1994). For the four-species ψη-globin genes, κ is estimated to be 5.26 ± 0.69 under the HKY85 model. A z-statistic can be compared to a standard normal distribution to see if the estimated κ differs from 1: $z = (\hat{\kappa} - 1)/\mathrm{SE}(\hat{\kappa}) = 6.164$, $P < 0.01$, and we conclude that it does.

Comparison between HKY85 and HKY85+Γ is a test of rate constancy over sites, which can be formulated as H_0: $\alpha^{-1} = 0$, H_1: $\alpha^{-1} \geq 0$, because the single rate is the limiting case of the gamma distribution with $\alpha \rightarrow \infty$. For the ψη-globin genes of four species, the α parameter is estimated to be 0.66 ± 0.38 under HKY85+Γ. By the δ technique, $\mathrm{SE}(\alpha^{-1}) = \mathrm{SE}(\alpha)/\alpha^2$, and the normal approximation suggests the comparison of $z = \hat{\alpha}^{-1}/\mathrm{SE}(\hat{\alpha}^{-1}) \approx \hat{\alpha}/\mathrm{SE}(\hat{\alpha}) = 1.75$ with a standard normal distribution. H_0 is just rejected by this test: $P = 0.041$ (one-tail normal probability). In comparison, the likelihood ratio test involves comparison of $2\Delta\ell = 2 \times [-10127.36 - (-10130.14)] = 5.55$ with a χ^2 distribution with df = 1, and again $0.01 < P < 0.05$.

The normal approximation has also been found to agree well with the χ^2 approximation for other data sets, implying that apart from the tree topology problem, the amount of data is sufficient for the asymptotics of ML estimates of parameters to be reliable. This conclusion is consistent with Tajima's (1993) demonstration that even with only two sequences, the biases of the ML estimators of sequence divergence under JC69 and K80 are negligible when the sequences are as short as only 100–500 bases. Nevertheless, we do not expect the normal approximation to be as reliable as the χ^2 approximation because the relatedness among estimates of parameters is ignored.

The Test of a Molecular Clock

Although tests concerning θ can be carried out without knowledge of the true tree, tests concerning branch lengths t appear more problematic. Unless we are certain of the true tree topology, we do not even know which branches exist and tests regarding them will include unknown uncertainties. (An analogy would be an attempt to make inferences regarding the shape and scale parameters, α and β, of a gamma distribution from data drawn from a normal distribution.) One important test of this sort concerns the existence of a molecular clock, which represents a set of restrictions on the branch lengths of the true tree, and the likelihood ratio test would be expected to apply if the likelihood values

TABLE 2. Test of the existence of a molecular clock for the four ψη-globin genes of human (H), chimpanzee (C), gorilla (G), and orangutan (O). Likelihood values (ℓ) and estimates ($\hat{\kappa}$) of the transition/transversion rate ratio are shown for all tree topologies. The HKY85 model is assumed either with or without a molecular clock.

Without clock			With clock		
Unrooted tree	ℓ	$\hat{\kappa}$	Rooted tree	ℓ	$\hat{\kappa}$
(H, C(G, O))	−10130.14	5.26	(O(G(H, C)))	−10132.47	5.26
			(G(O(H, C)))	−10167.21	5.25
			((H, C)(G, O))	−10167.21	5.25
			(C(H(G, O)))	−10171.47	5.24
			(H(C(G, O)))	−10171.47	5.24
(H, G(C, O))	−10133.20	5.14	(O(C(H, G)))	−10135.71	5.14
			(C(O(H, G)))	−10171.70	5.11
			((H, G)(C, O))	−10171.70	5.11
			(G(H(C, O)))	−10173.68	5.15
			(H(G(C, O)))	−10173.68	5.15
(H, O(C, G))	−10130.32	5.21	(O(H(C, G)))	−10134.05	5.21
			((H, O)(C, G))	−10169.56	5.20
			(H(O(C, G)))	−10169.56	5.20
			(G(C(H, O)))	−10172.07	5.19
			(C(G(H, O)))	−10172.07	5.19
(H, C, G, O)	−10133.48	5.16	(H, C, G, O)	−10173.67	5.15

could be calculated using the true topology under both models.

As in the case of tests concerning models, we examined whether the likelihood values for the several best trees are similar as compared with the likelihood difference resulting from the clock assumption. The ψη-globin pseudogenes of human, chimpanzee, gorilla, and orangutan were analyzed (the same data as analyzed in Fig. 3). The HKY85 model was assumed. With the assumption of a molecular clock, the position of the root of the tree can be identified (Felsenstein, 1981). The likelihood values of all bifurcating trees are listed in Table 2 under both models. Because there are $2s - 3$ branch lengths in a unrooted bifurcating tree and $s - 1$ branching times in a rooted bifurcating tree for s species, the likelihood ratio statistic should be compared to a χ^2 distribution with df $= s - 2$ (Felsenstein, 1981; see also Goldman, 1993a). For our example in Table 2, the best unrooted tree without the clock assumption is ((H, C)G, O), with $\ell_1 = -10130.14$, and the best (rooted) tree with the clock assumption is (((H, C)G)O), with $\ell_0 = -10132.47$. This comparison gives $2\Delta\ell = 2 \times 2.33 = 4.66$, which is not significant ($\chi^2_{0.001} = 9.21$, df $= 2$). Other plausible tree topologies, such as ((C, G)H, O), give very similar results for this data set. Removal of the molecular clock assumption therefore does not seem to improve the fit of model to data, i.e., substitution rates are more or less constant along different lineages. Similar results (not shown) are obtained for the mitochondrial (mtDNA) sequences of the same species (Brown et al., 1982). We therefore conclude that the test of a molecular clock can still be performed even if the true topology is unknown. However, because the likelihood values with or without the clock assumption are not very different compared with the likelihood differences caused by the tree topology, we suggest that likelihood values of the several best trees under both models be examined, regardless of whether the ML trees under the two models are compatible with each other.

Estimates of κ are very stable across tree topologies under both models, and estimates from both models are very similar (Table 2). Also, the likelihood values of different rooted trees that correspond to the same unrooted topology are quite different. The three topologies in which orang-

TABLE 3. The dependence of evaluation of reliability of the ML tree upon the assumed model. The 895-bp mtDNA sequences of human, chimpanzee, gorilla, orangutan, and gibbon are used. $\ell_{(1)}$ is the likelihood of the best tree under the model, ℓ_0 is the likelihood of the star tree, and $\ell_{\max} = \Sigma_i n_i \log(n_i) - n \log(n) = -2476.97$ is the maximum possible likelihood. Four (unrooted) tree topologies are examined: T_0 = (H, C, G), T_1 = ((H, C)G), T_2 = ((H, G)C), and T_3 = (H(C, G)); in all cases, orangutan and gibbon are the outgroups. Estimates of κ and α in the models, where present, are obtained from T_1.

			Order of	Bootstrap probabilities[b]				
Model[a]	$\ell_{(1)} - \ell_{\max}$	$\ell_{(1)} - \ell_0$	likelihoods	P_1	P_2	P_3	$\hat{\kappa}$	$\hat{\alpha}$
JC69 (0)	−436.77	13.68	312	46.05	6.00	47.95		
F81 (3)	−365.71	13.64	312	42.90	5.10	52.00		
K80 (1)	−271.45	11.43	132	71.10	7.00	21.90	8.651	
HKY85 (4)	−188.46	10.52	132	71.60	5.85	22.55	9.390	
REV (8)	−181.26	9.09	132	67.90	6.85	25.25		
JC69+Γ (1)	−425.68	8.95	132	49.80	3.65	46.55		1.062
F81+Γ (4)	−350.08	7.44	312	46.20	2.80	51.00		0.802
K80+Γ (2)	−249.47	6.37	132	83.10	3.80	13.10	11.165	0.553
HKY85+Γ (5)	−146.11	3.64	123	85.45	8.45	6.10	22.270	0.252

[a] Number of common parameters θ are given in parentheses.
[b] Estimated by the RELL method (Kishino et al., 1990).

utan diverged first have the highest likelihood values although they correspond to different unrooted tree topologies. The results suggest that the data contain much information concerning the root (the earliest divergence), although the topology is uncertain, and that by removing the clock assumption the model may lose power for discriminating among trees.

DISCUSSION

Test of Trees and Use of Models

In this section, we will use the bootstrap P_B, approximated by the RELL method, as a measure of the sampling error of the ML tree to study the dependence upon the assumed model of the evaluation of the reliability of the estimated tree. The 895-bp mtDNA sequences of human, chimpanzee, gorilla, orangutan, and gibbon (Brown et al., 1982) were analyzed using different models. Only the four (unrooted) trees involving the human–chimpanzee–gorilla separation were evaluated, with orangutan and gibbon as outgroups: T_0 = (H, C, G); T_1 = ((H, C)G); T_2 = ((H, G)C); T_3 = (H(C, G)).

Likelihood values, estimates of parameters, and the estimated P_B values are listed in Table 3. P_B depends critically on the model assumed in the analysis. Even the best (ML) tree depends on the model. In examining the power of a model in discriminating among tree topologies, Yang et al. (1994) used the difference in likelihood between the ML tree and the star tree, $\ell_{(1)} - \ell_0$, and found that more complex and reasonable models invariably gave lower values than did simpler and inappropriate models. (Note the use of parentheses here to indicate the greatest likelihood value, i.e., that ranked 1, as opposed to a particular hypotheses or tree numbered 1.) This result is confirmed in Table 3; in particular, adding the gamma distribution causes substantial reduction in the likelihood difference $\ell_{(1)} - \ell_0$. Nevertheless, $\ell_{(1)} - \ell_0$ is not a measure of the ML tree's reliability at all; two or more bifurcating trees can often be significantly better supported than the star tree using this measure. The bootstrap P_B, approximated by the RELL method, may be a better measure for this purpose. If P_B is used, use of more complex and realistic models does not necessarily mean loss of discriminating power. For the mtDNA sequence data analyzed in Table 3, better models appear to have higher power in discriminating among tree topologies. Similar patterns have been found in other data sets we have examined. Nevertheless, the likelihood method discriminates among the tree topologies by comparing their likelihood values, and because the likelihood values get more and more similar as the model is improved ($\ell_{(1)} - \ell_0$ is the difference of the highest and lowest likelihood values), we feel unable to draw a decisive conclusion concerning the discriminating powers of different models. Furthermore, our application of the P_B(RELL) measure to analyze many real data sets has rarely led to significant support for the estimated topology, and it is worth asking whether it is the data or the measure that lacks discriminating power. We suggest it is both.

Whereas an adequate model is important in estimating tree topology and especially in evaluating the reliability of the estimated tree, correct formulation of the model is of vital importance for understanding the evolutionary process. When a relatively simple model of nucleotide substitution is assumed, the α parameter of the gamma distribution is overestimated and thus the variation of rates over sites is underestimated (Table 3). For example, α is estimated as 1.06, 0.80, 0.55, and 0.25 when the assumed model is JC69+Γ, F81+Γ, K80+Γ, and HKY85+Γ, respectively. The same also holds for the κ parameter; when either the differences in nucleotide frequencies (JC69 and K80 compared with HKY85) or the rate variation over sites (K80 or HKY85 compared with K80+Γ or HKY85+Γ) is ignored, κ is underestimated. Generally, when one aspect of the complexity of the process is ignored, we get the (wrong) impression that other aspects are not important either.

Another way to look at the results is through the improvement in likelihood obtained by adding parameters into the model one after another. For example, the improvement on adding α into the model is greater if κ is already in the model (comparison between K80+Γ and K80, $\Delta\ell$ = 21.98) than if κ is not (comparison between JC69+Γ and JC69, $\Delta\ell$ = 11.09). The improvement is even greater if the frequency parameters have also been included in the model (comparison between HKY85+Γ and HKY85, $\Delta\ell$ = 42.35). The same is true concerning the effects of adding the κ parameter or the frequency parameters. We examined three other data sets and found this pattern to hold for all cases. These results may appear peculiar, but they obviously suggest that all the aspects represented by these parameters are characteristic of the evolutionary process and should not be ignored.

What Models Are Generally Acceptable in Phylogenetic Analysis?

In this paper, results obtained from analyses of two data sets, i.e., the ψη-globin genes (either with four or six species) and the 895-bp mtDNA sequences, are presented. These two data sets are well known and have been intensively analyzed, although few studies have paid attention to the adequacy of the models used.

For the ψη-globin pseudogenes, HKY85 has to be rejected when compared with REV even with only four sequences (human, chimpanzee, gorilla, orangutan): $2\Delta\ell = 2 \times 9.31 = 18.62$ compared with $\chi^2_{0.01} = 13.28$ (df = 4). With all six sequences in the data set, the difference is even greater ($2\Delta\ell = 60.96$). The HKY85+Γ model is slightly better than HKY85 with four sequences ($2\Delta\ell = 5.55$; $0.01 < P < 0.05$), whereas with all six sequences the difference is more significant ($2\Delta\ell = 18.04$; $P < 0.01$). For computational reasons, the REV+Γ model was not implemented. However, use of a discrete distribution of rates over sites to approximate the gamma distribution, in combination with the REV model of nucleotide substitution, leads to significant improvement in likelihood over REV (results not shown), indicating that substitution rates are indeed variable among sites. In Table 1, the observed frequencies in the identical and variable categories are higher than those expected by REV, which is a classical symptom of rate variation over sites (e.g., see Reeves, 1992). The reasons for such rate variation over sites in this pseudogene are unknown.

For the five-species mtDNA sequences, both HKY85+Γ and REV are better than HKY85, with likelihood ratio statistics of $2\Delta\ell = 84.69$ ($P < 0.01$) for the comparison between HKY85+Γ and HKY85 and $2\Delta\ell$

= 14.40 ($P < 0.01$) for the comparison between REV and HKY85 (Table 3). Rate variation over sites seems to be the most important factor accounting for the lack of fit of HKY85. Judging from the likelihood, REV+Γ would be an acceptable model and HKY85+Γ may be expected to be a good approximation.

Similar analyses have been done with many other data sets to find out which models are generally usable. The HKY85 model appears to be the simplest that we can hope to be generally acceptable; simpler models such as JC69, F81, or K80 are not acceptable for any of the data sets examined (Yang et al., 1994; Yang, Goldman, and Friday, unpubl.). These models are typically so poor that we seldom need to look at statistical tables to perform the tests. Two models that have been used in comparison with HKY85 are REV and HKY85+Γ. REV has most often been significantly better than HKY85 except for a few data sets where the number of sequences is small and the sequences are short. HKY85+Γ is not significantly better than HKY85 for the spacer between ψη- and δ-globin genes (Maeda et al., 1988), the ribosomal internal transcribed spacer (Gonzalez et al., 1990), and 28S ribosomal RNA (rRNA) sequences (Gonzalez et al., 1990) but is significantly better for other genes, including a small rRNA (Hixson and Brown, 1986).

These results suggest that for most genes REV should be used, but HKY85 does give very similar results in the estimation of T and $\mathbf{t}$. Supplemented with biological knowledge, possible rate variation over sites can easily be revealed by a residual analysis as performed by Goldman (1991, 1993b) (Table 1). When such variation does exist, a model like HKY85+Γ should be used because ignoring rate variation over sites has drastic effects on the estimation of $\mathbf{t}$ and θ (Yang et al., 1994).

Acknowledgments

We thank Dr. M. Hasegawa for sending some of the sequence data analyzed in this paper and Dr. M. Nei for assistance in meeting publication costs for this paper. Z.Y. was supported by a grant from Department of Zoology, The Natural History Museum (London), while this study was performed.

References

Bishop, M. J., and A. E. Friday. 1988. Molecular sequences and hominoid phylogeny. Pages 150–156 *in* Major topics in primate and human evolution (B. Wood, L. Martin, and P. Andrews, eds.). Cambridge Univ. Press, Cambridge, England.

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. 1964. Analysis of human evolution. Proc. XI Int. Congr. Genet. 3:923–933.

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1966. Estimation procedures for evolutionary branching processes. Bull. Int. Stat. Inst. 41:803–808.

Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution 21:550–570.

Cavender, J. A. 1989. Mechanized derivation of linear invariants. Mol. Biol. Evol. 6:301–316.

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

Edwards, A. W. F. 1970. Estimation of the branch points of a branching diffusion process. J. R. Stat. Soc. B 32:155–174.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1963. The reconstruction of evolution. Heredity 18:553.

Edwards, A. W. F., and L. L. Cavalli-Sforza. 1964. Reconstruction of evolutionary trees. Phenetic and phylogenetic classification. Syst. Assoc. Publ. 6:67–76.

Felsenstein, J. 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22:240–249.

Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

Felsenstein, J., and H. Kishino. 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. 42:193–200.

Fitch, D. H. A., C. Mainone, J. L. Slightom, and M. Goodman. 1988. The spider monkey ψη-globin gene and surrounding sequences: Recent or ancient insertions of LINEs or SINEs. Genomics 3:237–255.

Fukami, K., and Y. Tateno. 1989. On the maximum likelihood method for estimating molecular trees: Uniqueness of the likelihood point. J. Mol. Evol. 28: 460–464.

Fukami-Kobayashi, K., and Y. Tateno. 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitution. J. Mol. Evol. 32:79–91.

Gaut, B. S., and P. O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

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

Goldman, N. 1991. Statistical estimation of evolutionary trees. Ph.D. Thesis, Univ. Cambridge, Cambridge, England.

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

Goldman, N. 1993b. Simple diagnostic statistical tests of models for DNA substitution. J. Mol. Evol. 37:650–661.

Gonzalez, I. L., J. E. Sylvester, T. F. Smith, D. Stambolian, and R. D. Schmickel. 1990. Ribosomal RNA gene sequences and hominoid phylogeny. Mol. Biol. Evol. 7:203–219.

Hall, P., and S. R. Wilson. 1991. Two guidelines for bootstrap hypothesis testing. Biometrics 47:757–762.

Hasegawa, M., and H. Kishino. 1994. Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree. Mol. Biol. Evol. 11:142–145.

Hasegawa, M., H. Kishino, and N. Saitou. 1991. On the maximum likelihood method in molecular phylogenetics. J. Mol. Evol. 32:443–445.

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.

Hasegawa, M., and T. Yano. 1984. Maximum likelihood method of phylogenetic inference from DNA sequence data. Bull. Biometr. Soc. Jpn. 5:1–7.

Hillis, D. M., and J. J. Bull. 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42:182–192.

Hixon, J. E., and W. M. Brown. 1986. A comparison of the small ribosomal RNA genes from the mitochondrial DNA of the great apes and humans: Sequence, structure, evolution, and phylogenetic implications. Mol. Biol. Evol. 3:1–18.

Holmes, E. C. 1991. Different rates of substitution may produce different phylogenies of the eutherian mammals. J. Mol. Evol. 33:209–215.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. Syst. Biol. 44:17–48.

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

Kendall, M., and A. Stuart. 1979. The advanced theory of statistics, 4th edition, Volume 2. Charles Griffin, London.

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.

Kuhner, M., and J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11: 459–468. Erratum, 1995, Mol. Biol. Evol. 12:525.

Maeda, N., C.-I. Wu, J. Bliska, and J. Reneke. 1988. Molecular evolution of intergenic DNA in higher primates: Pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol. Biol. Evol. 5:1–20.

Miyamoto, M. M., B. F. Koop, J. L. Slightom, M. Goodman, and M. R. Tennant. 1988. Molecular systematics of higher primates—Genealogical relations and classification. Proc. Natl. Acad. Sci. USA 85:7627–7631.

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

Nei, M. 1987. Molecular evolutionary genetics. Columbia Univ. Press, New York.

Neyman, J. 1971. Molecular studies of evolution: A source of novel statistical problems. Pages 1–27 *in* Statistical decision theory and related topics (S. S. Gupta and J. Yackel, eds.). Academic Press, New York.

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.

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

Steel, M. 1994. The maximum likelihood point for a phylogenetic tree is not unique. Syst. Biol. 43:560–564.

Tajima, F. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. 10:677–688.

Thompson, E. A. 1975. Human evolutionary trees. Cambridge Univ. Press, Cambridge, England.

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

Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. J. Mol. Evol. 39:105–111.

Yang, Z. 1994b. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Syst. Biol. 43:329–342.

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.

Zharkikh, A., and W.-H. Li. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. 9:1119–1147.

Received 15 February 1994; accepted 22 November 1994