A New Method of Inference of Ancestral Nucleotide and Amino Acid Sequences

Ziheng Yang, Sudhir Kumar and Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received May 8, 1995
Accepted for publication September 14, 1995

ABSTRACT

A statistical method was developed for reconstructing the nucleotide or amino acid sequences of extinct ancestors, given the phylogeny and sequences of the extant species. A model of nucleotide or amino acid substitution was employed to analyze data of the present-day sequences, and maximum likelihood estimates of parameters such as branch lengths were used to compare the posterior probabilities of assignments of character states (nucleotides or amino acids) to interior nodes of the tree; the assignment having the highest probability was the best reconstruction at the site. The lysozyme *c* sequences of six mammals were analyzed by using the likelihood and parsimony methods. The new likelihood-based method was found to be superior to the parsimony method. The probability that the amino acids for all interior nodes at a site reconstructed by the new method are correct was calculated to be 0.91, 0.86, and 0.73 for all, variable, and parsimony-informative sites, respectively, whereas the corresponding probabilities for the parsimony method were 0.84, 0.76, and 0.51, respectively. The probability that an amino acid in an ancestral sequence is correctly reconstructed by the likelihood analysis ranged from 91.3 to 98.7% for the four ancestral sequences.

IT was suggested many years ago that amino acid sequences of present-day species may be used to reconstruct sequences of their extinct ancestors (*e.g.*, PAULING and ZUCKERKANDL 1963; ECK and DAYHOFF 1966: 161–202). The usefulness of reconstruction of ancestral sequences has been well recognized by evolutionary biologists (PAULING and ZUCKERKANDL 1963; MADDISON and MADDISON 1992; SWOFFORD 1993; LIBERTINI and DI DONATO 1994; STEWART 1995). For example, MALCOLM *et al.* (1990), ADEY *et al.* (1994), STACKHOUSE *et al.* (1994), and JERMANN *et al.* (1995) used the parsimony approach to infer amino acid sequences of extinct ancestral species and synthesized the genes in the laboratory by site-directed mutagenesis, and produced the gene products (proteins) in bacterial or cultured cells. The physico-chemical properties and physiological functions of these molecules were then studied, with a number of interesting findings (see STEWART 1995 for a review). Reconstruction of ancestral sequences also makes it possible to infer the evolutionary pathway of nucleotide or amino acid substitution at each site of the sequence, and this is useful for identifying specific nucleotide or amino acid changes that caused a functional change of the gene and for detecting convergent evolution or positive Darwinian selection at the nucleotide or amino acid level (*e.g.*, STEWART *et al.* 1987; SWANSON *et al.* 1991).

In the inference of ancestral sequences, the method of maximum parsimony has been used almost exclusively (see the references mentioned above). The method assigns character states (nucleotides or amino acids) to the interior nodes of the tree such that the number of character-state changes along the tree at each site is minimized. Algorithms for reconstructing ancestral character states under this criterion have been developed by FITCH (1971) for rooted bifurcating trees and by HARTIGAN (1973) for general tree topologies (see also ECK and DAYHOFF 1966; MADDISON and MADDISON 1992; SWOFFORD 1993). However, the accuracy of the reconstruction is usually unknown, except for the fact that the reconstruction will be reliable if the sequences are closely related. As parsimony generally fails to take into account biased substitution rates between nucleotides or amino acids and different branch lengths in the tree, there is concern about the reliability of the parsimony reconstruction (*e.g.*, COLLINS *et al.* 1994). Furthermore, parsimony often suggests many equally-best reconstructions at a site, and there is no natural way of choosing one of them.

In stochastic models used in the maximum-likelihood method of phylogenetic analysis, character states in ancestral sequences are regarded as random variables (*e.g.*, FELSENSTEIN 1981; GOLDMAN 1990). They do not appear in the likelihood function and are normally not estimated. However, the major reason for the lack of a probabilistic approach to character reconstruction seems to be the perception that there are a great many possible reconstructions at a site so that choosing one of them would be unlikely to be correct. For example, an (unrooted) tree of 10 amino acid sequences has eight interior nodes, so that at each site there are 20^8

Corresponding author: Ziheng Yang, Department of Integrative Biology, University of California, Berkeley, CA 94720-3140.
E-mail: ziheng@mws4.biol.berkeley.edu

$= 2.56 \times 10^{10}$ possible assignments of amino acids to the interior nodes. A method that assigns amino acids to the interior nodes at random would have a probability 0.39×10^{-10} of being correct.

At any rate, knowledge of the ancestral sequences is of great biological importance, and it is worth knowing how accurate the reconstruction can be and what factors are important in influencing the accuracy of reconstruction with real data sets. In this paper we propose a model-based likelihood approach to reconstructing ancestral sequences. The method follows standard statistical theory: given the data at the site, the conditional probabilities of different reconstructions can be compared and the reconstruction having the highest conditional probability is the best estimate at the site. The method allows calculation of the probability that the reconstruction at a site is correct, which provides a natural measure of the accuracy of the reconstruction. Real data will be analyzed to evaluate the accuracy of both the new method of this paper and the parsimony method and to identify factors accounting for the differences between the two methods. The robustness of character reconstruction to the assumed substitution model will also be examined through analysis of the example data set.

THEORY

The problem: The data consist of aligned nucleotide or amino acid sequences of extant species, all gaps being excluded. The phylogenetic tree linking the species will be assumed to be known, and the tree of Figure 1 will be used as an example to develop the theory, where external nodes 1–6 represent extant species whereas interior nodes 7–10 are extinct ancestors. Amino acid sequences will be considered, but the method can be applied to nucleotide sequences as well. The constancy of substitution rates among lineages (*i.e.*, the assumption of a molecular clock) is not assumed and thus unrooted trees are used (*e.g.*, FELSENSTEIN 1981). The branch length in the tree is measured by the average number of amino acid substitutions per site and represents the amount of evolution that has occurred along the branch. The substitution rate is assumed to be the same for all sites.

Let the data at a site be represented by $\mathbf{x} = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, where $x_i (i = 1, 2, \ldots, 6)$ stands for the amino acid in the *i*th extant sequence at the site. Let $\mathbf{y} = \{y_7, y_8, y_9, y_{10}\}$ be a set of amino acids assigned to the interior nodes of the tree, where $y_i (i = 7, 8, \ldots, 10)$ is the amino acid for the *i*th node of the tree (Figure 1). We will refer to $\mathbf{y}$ as an amino acid assignment or a reconstruction at the site. Together with data $\mathbf{x}$ at the site, $\mathbf{y}$ will specify the evolutionary history at the site. Our purpose is to estimate (reconstruct) $\mathbf{y}$ when $\mathbf{x}$ is given.

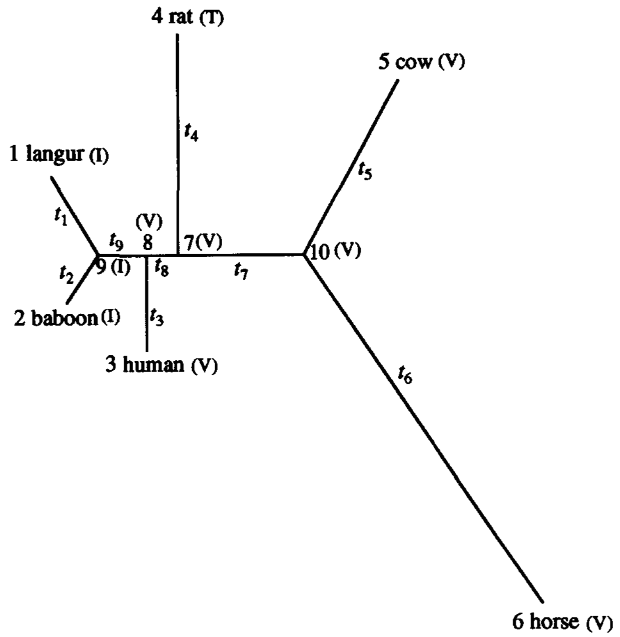

FIGURE 1.—Biological tree (unrooted) of langur (*Presbytis entellus*), baboon (*Papio cynocephalus*), human (*Homo sapiens*), rat (*Rattus norvegicus*), cow (*Bos taurus*), and horse (*Equus caballus*) for the lysozyme *c* sequences of STEWART *et al.* (1987). Branch lengths, measured as the average number of amino acid substitutions per site, were estimated by the maximum-likelihood approach under the empirical model of amino acid substitution of JONES *et al.* (1992); the estimates are $\hat{t}_1 = 0.081$, $\hat{t}_2 = 0.033$, $\hat{t}_3 = 0.064$, $\hat{t}_4 = 0.288$, $\hat{t}_5 = 0.240$, $\hat{t}_6 = 0.630$, $\hat{t}_7 = 0.106$, $\hat{t}_8 = 0.010$, and $\hat{t}_9 = 0.021$. The log-likelihood of this tree is $\ell = -1043.99$. The maximum-likelihood topology, which is incorrect and which is also the maximum-parsimony tree, is ((rat, horse), human, (baboon, (langur, cow))), with a much higher log-likelihood value ($\ell = -1031.60$). The amino acid compositions at site 2 of the six sequences are shown in parentheses, together with the single parsimony assignment of amino acids to the interior nodes (7, 8, 9, and 10). This is also the best reconstruction according to the maximum-likelihood analysis, although it is not very reliable as its posterior probability is only 0.563.

Parsimony reconstruction: The parsimony method assigns amino acids to the interior nodes to minimize the number of amino acid changes at the site along all branches of the tree. For example, the data at site 2 of the lysozyme *c* sequence of STEWART *et al.* (1987) can be represented by $x_1x_2x_3x_4x_5x_6$ = IIVTVV (Figure 1), where the one-letter code of amino acids is used. The reconstruction $y_7y_8y_9y_{10}$ = VVIV involves two changes along branches 8-9 and 7-4 and is the single most parsimonious reconstruction; any other assignment of amino acids to the interior nodes requires more than two changes. At other variable sites, there is often more than one reconstruction at a site that requires the same minimum number of changes. Although it is quite easy to count the number of changes required for a particular reconstruction at each site, enumerating all and only reconstructions that require the minimum number of changes is much more complicated; algorithms for doing this have been developed by FITCH (1971) and HARTIGAN (1973).

Likelihood reconstruction: A likelihood analysis requires formulation of a probabilistic model for amino acid substitution. Estimates of parameters in the model such as branch lengths of the tree will be used to evaluate the possible reconstructions. We use the empirical model of amino acid substitution derived by JONES *et al.* (1992) from analyzing the SwissProt Release 22 protein-sequence data. This is an update to the empirical matrix of amino acid substitution probabilities of DAYHOFF *et al.* (1978) (see KISHINO *et al.* 1990 for implementation of the model in the maximum-likelihood framework). Let π_i be the equilibrium frequency of amino acid i, and $P_{ij}(t)$ be the transition probability from amino acid i to j during time t according to the model of JONES *et al.* (1992). As the empirical model does not involve any free parameters, the parameters in the model are branch lengths in the tree and will be denoted $\theta = \{t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\}$ (see Figure 1).

With the root of the tree fixed (arbitrarily) at the interior node 7 (Figure 1), we can calculate the probability of observing data **x** as a sum over all possibilities of **y** (see, *e.g.*, FELSENSTEIN 1981):

$$f(\mathbf{x};\,\theta) = \sum_{\mathbf{y}} f(\mathbf{y})f(\mathbf{x}\,|\,\mathbf{y};\,\theta) = \sum_{y_7}\sum_{y_8}\sum_{y_9}\sum_{y_{10}} [\pi_{y_7}P_{y_7y_8}(t_8)P_{y_7y_{10}}(t_7)P_{y_8y_9}(t_9) \times P_{y_7x_4}(t_4)P_{y_8x_3}(t_3) \times P_{y_9x_1}(t_1)P_{y_9x_2}(t_2)P_{y_{10}x_5}(t_5)P_{y_{10}x_6}(t_6)], \quad (1)$$

where $f(\mathbf{y})$ represents the prior probability of **y** and $f(\mathbf{x}|\mathbf{y};\theta)$ is the conditional probability of observing **x** given **y**. The notation $f(\mathbf{x};\theta)$ implies that f is a function of **x** with θ to be the parameters. The quantity in the square bracket is the contribution by the assignment $y_7y_8y_9y_{10}$ to the probability, $f(\mathbf{x};\theta)$, of observing the data (**x**) and is equal to the probability of observing amino acid y_7 at node 7, which is the equilibrium frequency (π_{y7}) of amino acid y_7, times the transition probabilities along the nine branches in the tree (Figure 1). Under the assumption of independent substitution among sites, the probability of observing the whole sequence data, *i.e.*, the likelihood function for estimating θ, is the product of $f(\mathbf{x};\theta)$ over all sites.

When we are interested in estimating **y** (the ancestral character states at the site), we study the conditional probability of **y** given data **x**:

$$f(\mathbf{y}\,|\,\mathbf{x};\,\theta) = \frac{f(\mathbf{y})f(\mathbf{x}\,|\,\mathbf{y};\,\theta)}{f(\mathbf{x};\,\theta)}. \quad (2)$$

Since **y** is discrete, we estimate **y** by maximizing this conditional probability. We designate this estimate as $\hat{\mathbf{y}}$. Intuitively, the amino acid assignment that makes the greatest contribution to the probability of observing the data at a site will be the best reconstruction at the site. Parameters θ in (2) have to be replaced by their estimates, and the maximum-likelihood estimates are used in this paper; the method then has an empirical Bayesian interpretation (MARITZ and LWIN 1989).

Reconstruction of ancestral sequences by (2) is closely related to the method of YANG and WANG (1995; see also YANG 1995a) for estimating (predicting) substitution rates at nucleotide sites when these rates are assumed to be gamma distributed (YANG 1993, 1994b, 1995a). In both cases, we are interested in knowing the realized values of certain random variables in the model, and the conditional distribution of the random variables given the data is used to provide the best estimation. The rates for sites are continuous random variables and are thus estimated by the conditional mean of rates given the data or the regression of the rates on the data. This estimator has the highest correlation with the true rate (YANG and WANG 1995). In the current context, the ancestral amino acids are discrete random variables in the model and are estimated by maximizing the posterior probabilities.

Accuracy of a reconstruction at a site: The posterior probability $f(\hat{\mathbf{y}}|\mathbf{x};\theta)$ measures the accuracy of the reconstruction ($\hat{\mathbf{y}}$) at sites with data **x**. The overall accuracy can be calculated as an average of this over **x**'s at different sites. That is,

$$\text{Prob}(\hat{\mathbf{y}} \text{ is correct}) = \sum_{\mathbf{x}} f(\mathbf{x})f(\hat{\mathbf{y}}\,|\,\mathbf{x};\,\theta). \quad (3)$$

To understand the meaning of this measure of accuracy, consider a computer simulation of sequence evolution. Suppose that the sequence at the root of the tree is generated by using the equilibrium amino acid frequencies and is then allowed to "evolve" along the tree under a given substitution model with specified values of parameters (*e.g.*, GOLDMAN 1993; YANG 1995b). The ancestral sequences are recorded in the simulation, and the sequences of extant species are used to reconstruct the ancestral sequences by (2). Equation 3 is then equal to the probability that the reconstructed ancestral amino acids at a site are identical to the true amino acids at the site recorded in the simulation. Note that (2) generates the same ancestral amino acids if the data at different sites are the same, but the same data can be generated by different (true) ancestral amino acids at different sites during the simulation.

Strictly speaking, the measure of accuracy given by (3) should be calculated by enumerating all possibilities for **x**, but this would require extensive computation. A simpler approach, adopted in this paper, is to use the observed frequencies of different site patterns (**x**). This approach is expected to be reliable, especially if a realistic substitution model is used in the analysis. Other interesting measures are the accuracy of reconstruction at the variable (polymorphic) sites only or at the parsimony-informative sites, *i.e.*, sites at which at least two different amino acids are observed across the sequences, each present at least twice (NEI 1987: 315–316). The accuracy of a reconstruction by the parsimony method ($\tilde{\mathbf{y}}$) can be similarly defined:

$\Sigma_{\mathbf{x}} f(\mathbf{x}) f(\hat{\mathbf{y}}|\mathbf{x};\theta)$. When two or more equally-best reconstructions are obtained by parsimony at a site, the accuracy at the site is calculated as the average of these equally-best reconstructions. This is equivalent to the practice of regarding the method as being 10% correct in the computer simulation if one of the 10 equally-best reconstructions is correct.

Computational methods: In calculating $f(\mathbf{x};\theta)$, the summation signs in (1) can be moved rightward, so that an iterative algorithm can be devised; this technique was termed the pruning algorithm by FELSENSTEIN (1981). After maximum-likelihood estimates of θ are obtained, (2) can be used to evaluate the possible reconstructions at each site. Unlike the parsimony reconstruction, for which efficient algorithms are available, the likelihood reconstruction has to rely on comparison of $f(\mathbf{y}|\mathbf{x};\theta)$ for different values of $\mathbf{y}$. There are $20^4 = 160{,}000$ possible reconstructions at each site for the tree of Figure 1, and this number increases explosively with the number of interior nodes of the tree. However, it is unnecessary to enumerate all possibilities for $\mathbf{y}$, since most of the reconstructions have vanishingly small probabilities. For instance, there is little point in evaluating reconstructions involving amino acids that are not observed at the site in any species. Furthermore, the parsimony reconstruction produces a set of equivocal character states for each interior node (*e.g.*, MADDISON and MADDISON 1992). Use of these equivocal states considerably reduces the number of possible reconstructions to be evaluated. The calculation will become even faster if the most-parsimonious reconstructions only are evaluated.

Assignment of a character state at a given node: The posterior probability for an amino acid assignment to a particular interior node can be obtained by summing the contribution to the probability of observing data at the site $f(\mathbf{x};\theta)$ over all reconstructions at the site that assign the same amino acid to the node. For instance, the posterior probability that node 10 in the tree of Figure 1 has valine (Val) at a site with data $\mathbf{x}$ is

$$f(y_{10} = \text{Val}|\mathbf{x};\theta) = \frac{\Sigma_{\mathbf{y}:y_{10}=\text{Val}} f(\mathbf{y}) f(\mathbf{x}|\mathbf{y};\theta)}{f(\mathbf{x};\theta)}. \quad (4)$$

The best assignment at a node will be the amino acid that has the highest posterior probability. It is possible for the best amino acid assignment at a node obtained by (4) to be different from the amino acid assigned for the node in the best reconstruction obtained by (2). However, such cases are rare, and in this paper (2) is used to reconstruct the ancestral sequences, while (4) is used to provide another useful measure of the accuracy of the reconstruction. It is apparent that sequences at different ancestral nodes are reconstructed with different accuracy, and that the reconstruction of all ancestral amino acids ($y_7 y_8 y_9 y_{10}$) at a site [given by (3)] cannot be more reliable than the assignment at any of the interior nodes (say, y_{10}) [given by (4)].

In theory, the method can also be used to reconstruct sequences in extinct species that have not left extant descendants (*i.e.*, dinosaurs). Suppose that the human sequence at node 3 of Figure 1 is not available. Analysis of the remaining five sequences will lead to estimates of $t_1, t_2, \ldots, t_7$ and $t_8 + t_9$. Branching dates estimated from other sources such as fossil records can be used to provide estimates of t_8 (and thus t_9) and t_3. The conditional probability of the character state at node 3 at a site with data $\mathbf{x}$ is

$$f(y_3|\mathbf{x};\theta) = \sum_{y_8} f(y_8|\mathbf{x};\theta) P_{y_8 y_3}(t_3). \quad (5)$$

This probability can be evaluated at different values of y_3 to obtain the best assignment at node 3. Most often, the most likely amino acid at node 3 will be identical to that at node 8, but the accuracy of the reconstruction at node 3 will be lower than that at node 8.

ANALYSIS OF AN EXAMPLE DATA SET

Data: The amino acid sequences of lysozyme *c* of langur, baboon, human, rat, cow, and horse (STEWART *et al.* 1987) were analyzed. The horse sequence involves a deletion at site 70, and the cow sequence lacks an amino acid at site 103. These two sites were removed, with 128 amino acids left in each sequence. The biological tree relating these species is shown in Figure 1. Maximum-likelihood estimates of branch lengths suggest a large amount of evolution along the horse lineage and a greater rate of substitution along the langur lineage than in other primates. The langur and cow sequences are similar; both likelihood and parsimony methods of phylogenetic tree reconstruction produce an incorrect tree with the langur and the cow forming a sister clade. In many mammals such as humans and rats, lysozyme *c* exists mainly in secretions like tears and saliva, as well as in white blood cells and tissue macrophages, where its function is to fight invading bacteria. Colobine monkeys (such as the langur) and ruminants (such as the cow) have fermentative foreguts, where high levels of lysozyme *c* are present and where its function is to digest bacteria that pass from the foreguts into the true stomach. STEWART *et al.* (1987; see also SWANSON *et al.* 1991) suggested that the similar distribution and function of lysozyme *c* in colobine monkeys and ruminants may have led to similar selective pressure on the enzyme, and that positive Darwinian selection may account for the observed similarity in the lysozyme *c* sequences of the cow and the langur.

Reconstruction of ancestral amino acids at individual sites: The method developed in this paper and the parsimony method were applied to the lysozyme *c* data to reconstruct pathways of amino acid substitution for each site in the sequence. In the parsimony analysis, the algorithm of HARTIGAN (1973) was used to determine the minimum number of changes and all most-parsimonious reconstructions at each site. The results are given in Table 1, where the most-parsimonious reconstructions are shown in boldface. Reconstructions shown in column 4 (additional parsimony reconstructions) have posterior probabilities lower than 5%. The number of equally-best parsimony reconstructions at a site shown here is generally smaller than that obtained by using the MacClade program (MADDISON and MADDISON 1992), as MacClade seems to work with rooted trees only.

In the likelihood analysis, the branch lengths of the biological tree were estimated by using the empirical model of amino acid substitution of JONES *et al.* (1992) and are shown in Figure 1. Equation 2 was then used to compare different reconstructions. Because of the small size of the data, we evaluated all reconstructions that can be generated by assigning all observed amino acids at a given site to each interior node. Table 1 shows only sites at which the reconstructions have posterior probabilities in the range of 0.05–0.95. Sites not shown in the table include 46 invariant sites at which all species have identical amino acids and some sites at which only one species has a different amino acid. For these sites, there was a single parsimony reconstruction, which was also the best reconstruction by the likelihood method and had a very high probability (>0.95) of being correct.

Although the two methods agree with each other at the invariant sites, substantial differences exist at the more-variable sites between the parsimony and likelihood reconstructions (Table 1). First, when parsimony suggests one single reconstruction, it may not be reliable. For example, site 2 has data IIVTVV (Figure 1), and the single most-parsimonious reconstruction VVIV requires a minimum of two changes: one change of V ↔ I along branch 8-9 (with length 0.021) and another of V ↔ T along branch 7-4 (with length 0.288). This reconstruction has posterior probability 0.563 and is not very reliable. The second reconstruction (IIIV) requires three changes and has posterior probability 0.203. The third reconstruction (IIII) requires four changes (*i.e.*, I ↔ V, I ↔ T, I ↔ V, and I ↔ V along branches 8-3, 7-4, 10-5, and 10-6, respectively) and has a posterior probability of 0.179. This probability is quite high and is clearly due to the higher substitution rate between Ile (I) and Val (V) and between Ile (I) and Thr (T) than that between Val (V) and Thr (T) according to the substitution model of JONES *et al.* (1992), and due to branches 10-5 and 10-6 being very long (Figure 1). More extreme cases include site 23, at which the two most-parsimonious reconstructions have a total posterior probability of 0.267, and site 117, at which the single most-parsimonious reconstruction has only 18.5% chance of being correct.

Furthermore, parsimony often suggests many equally-best reconstructions, while some of them are much more likely than others according to their posterior probabilities. For example, at site 14, parsimony suggests three most-parsimonious reconstructions, each requiring three changes. However, the posterior probabilities suggest that one of them, RRRR, is actually much more reliable than the other two and is more reliable than the single most-parsimonious reconstruction at site 2. As explained in Table 2, the three reconstructions all involve one R ↔ K change along the branch 9-1 of Figure 1, but the first reconstruction (RRRR) requires two additional frequent changes (R ↔ K and R ↔ A) along two long branches, while the third reconstruction (RRRA) requires two additional rare changes (R ↔ A and A ↔ K) along two short branches, so that their posterior probabilities are 150 times different.

The accuracy of the reconstruction by the likelihood analysis is calculated to be 0.908 when all sites in the sequence are considered, is 0.856 for the variable (polymorphic) sites only, and is 0.733 for the parsimony-informative sites only. The probabilities for the parsimony reconstruction are 0.843, 0.755, and 0.512 for all, variable, and parsimony-informative sites, respectively. At the invariant or less variable sites, the two methods most often produce identical results and the accuracy of the reconstruction is also very high. At the more variable sites, the reconstruction is not very reliable for any method, but the likelihood reconstruction is nevertheless much more reliable than the parsimony reconstruction.

Assignment of amino acids to interior nodes of the tree: Calculation of the posterior probabilities by (4) for an amino acid assignment to a given interior node would require enumeration of all evolutionary pathways at the site, which means excessive computation. Instead, we evaluate only the likely reconstructions, *i.e.*, those that can be generated by assigning all observed amino acids at the site to each interior node. The posterior probabilities for all amino acid assignments to the node are then slightly underestimated, and the comparison among them will not be seriously affected. The results are shown in Table 3. Without exception, the best amino acid assignments to the interior nodes (Table 3) are compatible with the likelihood reconstructions at the site shown in Table 1. However, Table 3 provides information as to which nodes are responsible when the reconstruction is unreliable. For example, the best reconstruction at site 2 (with data IIVTVV) is VVIV, with probability 0.563 [by (2)]. The posterior probabilities for the amino acid assignment calculated by (4) are 0.612 for V (Val) at node 7, 0.602 for V (Val) at node 8, 0.958 for I (Ile) at node 9, and 0.817 for V (Val) at node 10. So the amino acid at node 9 is quite likely to be Ile (I) although the reconstruction for all interior nodes at the site is not so reliable. The overall accuracy for the entire sequence inferred by the likelihood analysis is estimated to be 0.974, 0.982, 0.987, and 0.913, for nodes 7, 8, 9, and 10 in the tree of Figure 1, respectively. In other words, the error rate for the four reconstructed

TABLE 1

Reconstructions of ancestral amino acids by the likelihood and parsimony methods

Site	Data $x_1x_2x_3x_4x_5x_6$	Reconstructions $y_7y_8y_9y_{10}$ and their posterior probabilities (in parentheses)	Additional parsimony reconstructions	# paths (# changes)
2	IIVTVV	**VVIV** (0.563), IIIV (0.203), IIII (0.179)		1 (2)
14	KRRRKA	**RRRR** (0.721), **RRRK** (0.249)	**RRRA**	3 (3)
17	LLMMLM	**MMLM** (0.847), LLLL (0.077)		1 (2)
21	KRRYKG	**RRRR** (0.819), **RRRK** (0.141)	**KKKK, GRRG, KRRK, RRRG, YRRG, YRRK, YRRY**	9 (4)
23	VIIVVY	IIII (0.542), **VVVV** (0.184), IIIV (0.166), **VIIV** (0.083)		2 (3)
29	VVMVLV	**VVVV** (0.911)		1 (2)
32	AAAATA	**AAAA** (0.898), AAAT (0.094)		1 (1)
37	GDGNSN	**NGGN** (0.245), GGGS (0.244), GGGG (0.191), SGGS (0.096)		1 (3)
41	EQRQKR	**QQQQ** (0.404), **QQQK** (0.331), **QQQR** (0.125), **RRQR** (0.090)	**RRER, RRRR**	6 (4)
48	GGGGSA	**GGGG** (0.533), **GGGS** (0.338), **GGGA** (0.114)		3 (2)
49	DDDDSN	**DDDD** (0.445), **DDDN** (0.401), **DDDS** (0.134)		3 (2)
50	EQRQEG	**QQQE** (0.502), **QQQQ** (0.406)	**EEEE, QQQG**	4 (4)
62	RHRRKK	**RRRK** (0.648), RRRR (0.335)		1 (2)
72	GGGRNS	**GGGG** (0.416), **GGGN** (0.286), **GGGS** (0.228)	**NGGN, RGGN, RGGR, RGGS, SGGS**	8 (3)
75	DNNNDN	**NNNN** (0.927), NNND (0.068)		1 (2)
76	AAAAGA	**AAAA** (0.937), AAAG (0.053)		1 (1)
79	IILIVI	**IIII** (0.867), IIIV (0.125)		1 (2)
83	AAAAEK	**AAAE** (0.507), **AAAA** (0.365), **AAAK** (0.066)		3 (2)
85	LLLLML	**LLLL** (0.920), LLLM (0.076)		1 (1)
86	QQQQED	**QQQE** (0.706), **QQQQ** (0.238)	**QQQD**	3 (2)
87	NDDDNE	**DDDD** (0.898), **DDDN** (0.069)	**DDDE**	3 (3)
88	NNNDDN	**NNNN** (0.857), **DNND** (0.075), NNND (0.0627)		2 (2)
90	ATATAD	**AAAA** (0.906)		1 (3)
91	DDDQKD	**DDDD** (0.888)		1 (2)
98	RRRRKR	**RRRR** (0.908), RRRK (0.091)		1 (1)
99	VVVVIV	**VVVV** (0.850), VVVI (0.146)		1 (1)
101	SSRRSR	**RRSR** (0.874), SSSS (0.091)		1 (1)
102	DDDDED	**DDDD** (0.896), DDDE (0.102)		1 (1)
107	RRRRTS	**RRRT** (0.405), **RRRR** (0.318), **RRRS** (0.206)		3 (2)
113	RRRQKV	**RRRR** (0.678), **RRRK** (0.264)	**KRRK, QRRK, QRRQ, QRRV, RRRV, VRRV**	8 (3)
114	NNNRSK	**NNNN** (0.810), **NNNS** (0.113)	**KNNK, NNNK, RNNK, RNNR, RNNS, SNNS**	8 (3)
117	QQQKRK	QQQQ (0.344), QQQR (0.283), **KQQK** (0.185), QQQK (0.107), RQQR (0.052)		1 (2)
118	NNNNDD	**NNND** (0.729), NNNN (0.263)		1 (1)
119	KRRRHK	**RRRR** (0.881), **RRRH** (0.068)	**RRRK**	3 (3)
123	QQQGSE	**QQQQ** (0.724), **QQQE** (0.078), **QQQS** (0.054)	**EQQE, GQQE, GQQG, GQQS, SQQS**	8 (3)
126	KQQREA	**QQQQ** (0.566), **QQQE** (0.376)	**AQQA, EQQA, QQQA, RQQA, RQQE, RQQR**	8 (4)
129	GGGGTN	**GGGG** (0.374), **GGGT** (0.233), **GGGN** (0.162)		3 (2)
130	VVVVLL	**VVVL** (0.778), VVVV (0.193)		1 (1)

The lysozyme *c* sequences of langur, baboon, human, rat, cow, and horse were analyzed, and site refers to the numbering of STEWART *et al.* (1987). Data at a site ($x_1x_2x_3x_4x_5x_6$) are the amino acid compositions in the six extant species (*i.e.*, at nodes 1–6 in the tree of Figure 1), while the reconstruction ($y_7y_8y_9y_{10}$) is specified by the amino acids at interior nodes 7–10 in the tree of Figure 1. Only reconstructions that have posterior probabilities in the range 0.05–0.95 for the likelihood method are shown, and sites at which no reconstruction has probability in this range are not shown. The best reconstruction by the likelihood method is the first reconstruction in column 3. Equally-parsimonious reconstructions are shown in boldface. No. of paths and No. of changes refer to the number of most-parsimonious reconstructions at each site and the minimum number of substitutions required for each of these reconstructions at the site.

TABLE 2

Three most-parsimonious reconstructions and their posterior probabilities (in parentheses) for site 14 of the lysozyme sequence data

Number	Reconstruction	Inferred changes and concerned branches		
1	RRRR (0.721)	R–K, 9–1 (6.46) (0.081)	R–K, 10–5 (6.46) (0.240)	R–A, 10–6 (0.58) (0.630)
2	RRRK (0.249)	R–K, 9–1 (6.46) (0.081)	R–K, 7–10 (6.46) (0.106)	K–A, 10–6 (0.35) (0.630)
3	RRRA (0.005)	R–K, 9–1 (6.46) (0.081)	R–A, 7–10 (0.58) (0.106)	A–K, 10–5 (0.35) (0.240)

The amino acid compositions at the site are KRRRKA (Table 1). The tree topology and estimates of branch lengths are shown in Figure 1. The number in the parentheses (A_{ij}) below the amino acid difference measures the "exchangeability" between the two amino acids i and j, with the mean to be one, calculated from the empirical model of amino acid substitution of JONES *et al.* (1992); $Q_{ij} = A_{ij}\pi_j$ (with $A_{ij} = A_{ji}$) is the rate of substitution from amino acid i to j, where π_j is the equilibrium frequency of amino acid j. The length of the branch is shown below the branch, along which the change is supposed to occur.

ancestral sequences ranges from 1.3 to 8.7%. Node 10 is related with the extant species by long branches, and the reconstructed sequence for this node is much less reliable than those for other nodes.

DISCUSSION

Positive selection and convergent evolution: In the case of lysozyme *c*, STEWART *et al.* (1987: Figure 1) identified seven sites at which the same amino acid arose independently along the lineages leading to the cow and the langur. However, the existence of many equally-best reconstructions inferred by parsimony made it difficult to determine the most likely amino acids at the interior nodes from which the cow and langur amino acids have changed. Using the likelihood approach, we can see that these changes not only led to the same amino acids in cow and langur but also probably came from the same amino acids. These identical changes include an Arg (R) → Lys (K) change at site 14, an Arg

TABLE 3

Maximum-likelihood reconstructions of amino acids at interior nodes of the tree

						1							2					3						4							5				6						7
Site	2	3	4	5	8	0	1	4	5	6	7	8	0	1	3	7	9	1	2	3	4	7	8	1	3	5	6	7	8	9	0	2	6	9	1	2	3	6	7	8	1
Langur	I	F	E	R	L	R	T	K	L	G	L	D	Y	K	V	N	V	L	A	K	W	G	Y	E	T	Y	N	P	G	D	E	T	I	I	S	R	Y	N	N	G	P
Baboon	I	F	E	R	L	R	T	R	L	G	L	D	Y	R	I	N	V	L	A	K	W	D	Y	Q	T	Y	N	P	G	D	Q	T	I	I	S	H	Y	N	D	G	P
Human	V	F	E	R	L	R	T	R	L	G	M	D	Y	R	I	N	M	L	A	K	W	G	Y	R	T	Y	N	A	G	D	R	T	I	I	S	R	Y	N	D	G	P
Rat	T	Y	E	R	F	R	T	R	N	G	M	S	Y	Y	V	D	V	L	A	Q	H	N	Y	Q	R	Y	D	P	G	D	Q	T	I	I	S	R	Y	N	D	G	P
Cow	V	F	E	R	L	R	T	K	L	G	L	D	Y	K	V	N	L	L	T	K	W	S	Y	K	T	Y	N	P	S	S	E	T	I	I	S	K	W	N	D	G	P
Horse	V	F	S	K	L	H	K	A	Q	E	M	D	F	G	Y	N	V	M	A	E	Y	N	F	R	F	G	K	N	A	N	G	S	L	L	N	K	W	K	D	N	R
Node 7	v	**F**	**E**	**R**	**L**	**R**	**T**	**R**	**L**	**G**	**M**	**D**	**Y**	**R**	**i**	**N**	**V**	**L**	**A**	**K**	**W**	?	**Y**	Q	**T**	**Y**	**N**	**P**	**G**	**D**	**Q**	**T**	**I**	**I**	**S**	**R**	**Y**	**N**	**D**	**G**	**P**
Node 8	v	**F**	**E**	**R**	**L**	**R**	**T**	**R**	**L**	**G**	**M**	**D**	**Y**	**R**	**i**	**N**	**V**	**L**	**A**	**K**	**W**	G	**Y**	Q	**T**	**Y**	**N**	**P**	**G**	**D**	**Q**	**T**	**I**	**I**	**S**	**R**	**Y**	**N**	**D**	**G**	**P**
Node 9	**I**	**F**	**E**	**R**	**L**	**R**	**T**	**R**	**L**	**G**	**L**	**D**	**Y**	**R**	**i**	**N**	**V**	**L**	**A**	**K**	**W**	G	**Y**	**Q**	**T**	**Y**	**N**	**P**	**G**	**D**	**Q**	**T**	**I**	**I**	**S**	**R**	**Y**	**N**	**D**	**G**	**P**
Node 10	V	**F**	**E**	**R**	**L**	**R**	**T**	**r**	**L**	**G**	M	**D**	**Y**	R	i	**N**	**V**	**L**	A	**K**	**W**	?	**Y**	?	**T**	**Y**	**N**	**P**	g	?	e	**T**	**I**	**I**	**S**	k	**W**	**N**	**D**	**G**	**P**

																					1					1							1							1
	7							8							9							0					1							2							3
Site	2	3	4	5	6	8	9	0	2	3	5	6	7	8	0	1	2	3	4	8	9	1	2	4	6	7	0	3	4	5	7	8	9	1	2	3	5	6	7	9	0
Langur	G	A	V	D	A	H	I	S	S	A	L	Q	N	N	A	D	A	V	A	R	V	S	D	Q	I	R	V	R	N	H	Q	N	K	V	S	Q	V	K	G	G	V
Baboon	G	A	V	N	A	H	I	S	N	A	L	Q	D	N	T	D	A	V	A	R	V	S	D	Q	I	R	V	R	N	H	Q	N	R	V	S	Q	V	Q	G	G	V
Human	G	A	V	N	A	H	L	S	S	A	L	Q	D	N	A	D	A	V	A	R	V	R	D	Q	I	R	V	R	N	R	Q	N	R	V	R	Q	V	Q	G	G	V
Rat	R	A	K	N	A	G	I	P	S	A	L	Q	D	D	T	Q	A	I	Q	R	V	R	D	Q	I	R	V	Q	R	H	K	N	R	L	S	G	I	R	N	G	V
Cow	N	A	V	D	G	H	V	S	S	E	M	E	N	D	A	K	A	V	A	K	I	S	E	Q	I	T	V	K	S	H	R	D	H	V	S	S	V	E	G	T	L
Horse	S	S	S	N	A	N	I	M	S	K	L	D	E	N	D	D	D	I	S	R	V	R	D	K	M	S	K	V	K	H	K	D	K	L	S	E	L	A	S	N	L
Node 7	**G**	**A**	**V**	**N**	**A**	**H**	**I**	**S**	**S**	**A**	**L**	**Q**	**D**	**N**	**A**	D	**A**	**V**	**A**	**R**	**V**	R	**D**	**Q**	**I**	**R**	**V**	**R**	**N**	**H**	**q**	**N**	**R**	**V**	**S**	Q	**V**	**Q**	**G**	**G**	**V**
Node 8	**G**	**A**	**V**	**N**	**A**	**H**	**I**	**S**	**S**	**A**	**L**	**Q**	**D**	**N**	**A**	**D**	**A**	**V**	**A**	**R**	**V**	R	**D**	**Q**	**I**	**R**	**V**	**R**	**N**	**H**	**Q**	**N**	**R**	**V**	**S**	**Q**	**V**	**Q**	**G**	**G**	**V**
Node 9	**G**	**A**	**V**	**N**	**A**	**H**	**I**	**S**	**S**	**A**	**L**	**Q**	**D**	**N**	**A**	**D**	**A**	**V**	**A**	**R**	**V**	**S**	**D**	**Q**	**I**	**R**	**V**	**R**	**N**	**H**	**Q**	**N**	**R**	**V**	**S**	**Q**	**V**	**Q**	**G**	**G**	**V**
Node 10	?	**A**	**V**	**N**	**A**	**H**	I	**S**	**S**	e	**L**	**e**	D	N	**A**	D	**A**	**V**	**A**	**R**	V	R	D	**Q**	**I**	?	**V**	r	N	**H**	?	**d**	R	**V**	**S**	**q**	**V**	q	**G**	?	**l**

The posterior probability for the reconstructed amino acid at an interior node is calculated according to (4) and is indicated using different typefaces: upper case bold for 0.9–1.0, upper case for 0.8–0.9, lower case bold 0.7–0.8, lower case 0.5–0.7, and question mark (?) for <0.5. Only variable sites in the sequence data are shown.

(R) → Lys (K) change at site 21, an Asn (N) → Asp (D) change at site 75, and an Asp (D) → Asn (N) change at site 87 along both the cow and langur lineages (branches 9-1 and 10-5 of Figure 1); and a Met (M) → Leu (L) change at site 17, an Arg (R) → Ser (S) change at site 101 along both the cow lineage and the lineage leading to langur and baboon (branches 8-9 and 10-5 of Figure 1). At site 50, two reconstructions have very similar posterior probabilities (Table 1). The first (with probability 0.502) suggests a shared Gln (Q) → Glu (E) change along the langur lineage and the lineage leading to cow and horse (branches 9-1 and 7-10 of Figure 1), while the second (with probability 0.406) suggests the same Gln (Q) → Glu (E) change along the langur and cow lineages. The above analysis is tentative, and it is worthwhile to develop statistical methods for testing whether such convergent substitutions occur significantly more often than expected by the model that assumes the same pattern of substitution throughout the tree. GOLDMAN (1993) has made a good start in this direction.

Accuracy of reconstructed ancestral sequences: Branch lengths in the tree and the pattern of amino acid (or nucleotide) substitution appear to be the major factors accounting for the differences between the likelihood and parsimony reconstructions. In a likelihood analysis, information concerning these factors, obtained either from the data being analyzed (such as the branch lengths) or from other sources [such as the empirical matrix of JONES *et al.* (1992)], is used to evaluate possible reconstructions at each site. In contrast, the parsimony analysis fails to use such information. Therefore, the former can be expected to be generally more reliable than the latter.

Nevertheless, it should be noted that there exist a great number of possible reconstructions at each site. By the model assumed in the likelihood analysis, one site in the sequence amounts to one data point, and, in a sense, more parameters than the number of data points are being estimated, so that it is almost certain that some of the reconstructions are wrong, although they are the best we can obtain from the data. The accuracy of reconstruction will not improve very much by increasing the sequence length, although long sequences will lead to more reliable estimates of branch lengths (and of the substitution pattern if this is estimated from the data).

Because the accuracy of a reconstruction depends on the closeness of the sequences and possibly the number of species, we have analyzed another data set with a larger number of sequences and examine here the accuracy of the reconstructions. The data used were the mitochondrial cytochrome *b* sequences of 16 animals (two species of whales, cow, rat, mouse, opossum, chicken, toad, carp, loach, trout, smalltail shark, horn shark, ray, lamprey, and sea urchin). The phylogenetic relationship among these species is well established from fossil and morphological evidence, and the biological tree, which had 14 interior nodes, was used in the analysis. After exclusion of gaps from the alignment, the sequences had 375 amino acid sites, of which 169 were invariant. The total branch length along the tree, estimated by maximum likelihood using the empirical model of JONES *et al.* (1992), was 2.726 amino acid substitutions per site, whereas the parsimony analysis gave a minimum of 1.891 changes per site (709 changes for all sites). Therefore, the sequences involved a considerable amount of evolution. At 20 sites, there were >10 equally-best reconstructions by the parsimony analysis, and at three sites the number of most-parsimonious reconstructions was over 90. The accuracy of the reconstruction at a site by the parsimony approach was 0.795, 0.627, and 0.495 for all, variable (polymorphic), and parsimony-informative sites, respectively, whereas the corresponding probabilities for the likelihood analysis were calculated to be 0.851, 0.728, and 0.633, respectively. The probability that an amino acid in the ancestral sequence reconstructed by the likelihood analysis is correct ranged from 0.870 to 0.928 for the 14 interior nodes in the tree. These posterior probabilities are lower than those for the lysozyme *c* data, but since the sequences were quite different and there were so many possible reconstructions at each site ($20^{14} = 1.64 \times 10^{18}$), this level of accuracy seems to be acceptable.

For both the likelihood and parsimony methods, reconstruction of ancestral character states is more reliable if the extant sequences in the data are similar, if the site is less variable, and if the interior nodes are close to extant species whose sequences are part of the data. When the sequences are very similar, parsimony reconstructions can be expected to be generally reliable and to be identical to the likelihood reconstructions. While caution should be exercised concerning the reliability of the reconstructed sequences, reconstruction of ancestral sequences appears to be generally accurate enough to be useful in evolutionary studies.

Robustness of the reconstruction by the likelihood method to inaccuracies in the assumed model: The calculated posterior probabilities may not be reliable if estimates of parameters (θ in Equation 2) are unreliable or if the assumed substitution model is incorrect. While many complex and realistic models of nucleotide substitution have been developed (see, *e.g.*, YANG 1994), only a few models of amino acid substitution are available. The empirical frequency matrices of amino acid substitution derived by DAYHOFF *et al.* (1978) and JONES *et al.* (1992) represent average substitution patterns over many different proteins and different branch lengths and thus may not reflect the characteristics of the protein being analyzed. To examine the sensitivity of character reconstruction to the assumed model, two additional substitution models were used to analyze the lysozyme *c* data: the empirical model of DAYHOFF *et al.* (1978) and the Poisson-process model that assumes equal substitution rates between any two amino acids. Use of the empirical model of DAYHOFF *et al.* (1978) produced results very similar to those obtained under the model of JONES *et al.* (1992). For example, the likelihood reconstructions were identical to those shown in Table 1 at all but three sites. The exceptions were sites 50, 83, and 107, at which the model of DAYHOFF *et al.* suggested QQQQ, AAAA, and RRRR, respectively, as the best reconstructions. Under the model of JONES *et al.* (Table 1), these reconstructions have slightly lower posterior probabilities than the best reconstructions. The best reconstructions obtained under the Poisson-process model were different from those under the model of JONES *et al.* (1992) at six sites only, although the posterior probabilities under the two models were often quite different. These six sites were sites 23, 50, 83, 86, 107, and 117, where the best reconstructions under the Poisson model were VVVV, QQQQ, AAAA, QQQQ, RRRR, and KQQK, respectively. The two most-parsimonious reconstructions at site 23 and the single most-parsimonious reconstruction at site 117 (Table 1) also had the highest posterior probabilities. As one might expect, the parsimony reconstructions were more similar to the likelihood reconstructions under the Poisson model than to those under the model of JONES *et al.* As the likelihood method takes into account the differences of branch lengths of the tree, use of even the simple Poisson-process model produces more reliable results than parsimony. Reconstruction of ancestral sequences appears to be quite robust to changes to the substitution model assumed, although there is certainly need for developing more realistic models of amino acid substitution.

The model used in this paper for obtaining maximum likelihood parameter estimates assumes that the same model of amino acid substitution applies to the whole phylogenetic tree. This assumption is questionable in the case of lysozyme *c* sequences because special adaptive evolution seems to have occurred in the langur and cow lineages. Models allowing for such differences in the pattern of substitution will need to use different rate matrices or transition-probability matrices for different branches in the tree and may involve too many parameters (BARRY and HARTIGAN 1987; YANG and ROBERTS 1995). Practically useful models are yet to be developed. Nevertheless, the similarity between the results obtained under the model of JONES *et al.* (1992) and those obtained under the Poisson model (in which case the substitution matrices for all branches in the tree are unreliable) suggests that the reconstruction of ancestral sequences is more-or-less robust to violation of this assumption.

Another problem with the model used in this paper is the assumption of constancy of substitution rates across sites. Maximum likelihood models that allow for variable substitution rates among sites have been developed by YANG (1993, 1994b, 1995), and it is straightforward to implement the method of this paper in the discrete-gamma model of YANG (1994b). This is not pursued in this paper, mainly because the reconstruction of ancestral sequences is expected to be insensitive to whether or not the same rate is assumed for all sites. According to the formulation of YANG (1993, 1994b), amino acids at a fast-changing site in effect evolve along proportionally elongated branches in the tree. As can be seen from (1), the relative contributions to the likelihood by different reconstructions at a site is unlikely to change significantly when the branch lengths are multiplied by a constant; they mainly depend on the number of changes required for the reconstruction, the lengths of the branches along which the changes are supposed to occur, and the likelihood of the changes according to the assumed substitution model (see Table 3).

Program availability: The method developed in this paper was implemented in the baseml and aaml programs of the PAML package, which are for analyzing nucleotide and amino acid sequences, respectively. The package is distributed by ZIHENG YANG and can be obtained by anonymous ftp at ftp.bio.indiana.edu under the directory molbio/evolve.

This study was supported by National Institutes of Health grant GM-20293 and National Science Foundation grant DEB9119802 to M.N.

LITERATURE CITED

ADEY, N. B., T. O. TOLLEFSBOL, A. B. SPARKS, M. H. EDGELL and C. A. HUTCHISON, 1994 Molecular resurrection of an extinct ancestral promoter for mouse L1. Proc. Natl. Acad. Sci. USA **91:** 1569–1573.

BARRY, D., and J. A. HARTIGAN, 1987 Statistical analysis of hominoid molecular evolution. Stat. Sci. **2:** 191–210.

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

DAYHOFF, M. O., R. M. SCHWARTZ and B. C. ORCUTT, 1978 A model of evolutionary change in proteins, pp. 345–352 in *Atlas of Protein Sequence and Structure,* Vol. 5, Suppl. 3, edited by M. O. DAYHOFF. National Biomedical Research Foundation, Washington DC.

ECK, R. V., and M. O. DAYHOFF, 1966 *Atlas of Protein Sequence and Structure.* National Biomedical Research Foundation, Silver Spring, MD.

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

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

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

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

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

JERMANN, T. M., J. G. OPITZ, J. STACKHOUSE and S. A. BENNER, 1995 Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature **374:** 57–59.

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

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.

LIBERTINI, G., and A. DI DONATO, 1994 Reconstruction of ancestral sequences by the inferential method, a tool for protein engineering studies. J. Mol. Evol. **39:** 219–229.

MADDISON, W. P., and D. R. MADDISON, 1992 *MacClade: Analysis of Phylogeny and Character Evolution, Version 3.* Sinauer Associates, Sunderland, MA.

MALCOLM, B. A., K. P. WILSON, B. W. MATTHEWS, J. F. KIRSCH and A. C. WISLON, 1990 Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. Nature **345:** 86–89.

MARITZ, J. S., and T. LWIN, 1989 *Empirical Bayes Methods,* Ed. 2. Chapman and Hall, London.

NEI, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

PAULING, L., and E. ZUCKERKANDL, 1963 Chemical paleogenetics: molecular "restoration studies" of extinct forms of life. Acta Chem. Scand. **17:** S9-S16.

STACKHOUSE, J., S. R. PRESNELL, G. M. MCGEEHAN, K. P. NAMBIAR and S. A. BENNER, 1990 The ribonuclease from an ancient bovid ruminant. FEBS Lett. **262:** 104–106.

STEWART, C.-B., 1995 Active ancestral molecules. Nature **374:** 12–13.

STEWART, C.-B., J. W. SCHILLING and A. C. WILSON, 1987 Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature **330:** 401–404.

SWANSON, K. W., D. M. IRWIN and A. C. WILSON, 1991 Stomach lysozyme gene of the langur monkey: tests for convergence and positive selection. J. Mol. Evol. **33:** 418–425.

SWOFFORD, D. L., 1993 *Phylogenetic Analysis Using Parsimony (PAUP), Version 3.1.* University of Illinois, Champaign, IL.

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

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

YANG, Z., 1995b Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. J. Mol. Evol. **40:** 689–697.

YANG, Z., and D. ROBERTS, 1995 On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. **12:** 451–458.

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

Communicating editor: B. GOLDING