JOURNAL OF MOLECULAR EVOLUTION

Accuracies of Ancestral Amino Acid Sequences Inferred by the Parsimony, Likelihood, and Distance Methods

Jianzhi Zhang, Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802, USA

Received: 11 July 1996 / Accepted: 29 August 1996

Abstract. Information about protein sequences of ancestral organisms is important for identifying critical amino acid substitutions that have caused the functional change of proteins in evolution. Using computer simulation, we studied the accuracy of ancestral amino acids inferred by two currently available methods (maximum-parsimony [MP] and maximum-likelihood [ML] methods) in addition to a distance method, which was newly developed in this paper. All three methods give reliable inference when the divergence of amino acid sequences is low. When the extent of sequence divergence is high, however, the ML and distance methods give more accurate results than the MP method, particularly when the phylogenetic tree includes long branches. The accuracy of inferred ancestral amino acids does not change very much when a few present-day sequences are added or eliminated. When an incorrect model of amino acid substitution is used for the ML and distance methods, the accuracy decreases, but it is still higher than that for the MP method. When the tree topology used is partially incorrect, the accuracy in the correct part of the tree is virtually unaffected. The posterior probability of inferred ancestral amino acids computed by the ML and distance methods is an unbiased estimate of the true probability when a correct substitution model is used but may become an overestimate when a simpler model is used.

Key words: Ancestral protein sequences — Likelihood method — Parsimony method — Distance method

Introduction

To understand critical amino acid substitutions in the evolution of protein function, it is important to know the amino acid sequences of the proteins of ancestral organisms (e.g., Jermann et al. 1995; Chandrasekharan et al. 1996). Several statistical methods have been developed to infer the ancestral amino acid sequences from the sequences of present-day species when the phylogenetic relationship is known (Eck and Dayhoff 1966; Libertini and Di Donato 1994; Schluter 1995; Yang et al. 1995; Koshi and Goldstein 1996). Among these methods, the maximum-parsimony (MP) method has been used most frequently. In the MP method, each amino acid site is considered separately, and the amino acid at each interior node of the tree is determined so as to make the total number of amino acid changes at the site minimal (Eck and Dayhoff 1966; Fitch 1971; Hartigan 1973; Maddison and Maddison 1992). Because no information about the branch lengths and the pattern of amino acid substitution is used in the MP method, the inferred ancestral sequences are expected to be somewhat unreliable (e.g., Collins et al. 1994). To infer the ancestral sequences more accurately, Yang et al. (1995) introduced a maximum-likelihood (ML) method. In this method, the branch lengths of the phylogenetic tree are estimated by the maximum-likelihood method, and the posterior probability of each assignment of amino acids at ancestral nodes is computed at each site by using the Bayesian approach. The amino acid assignment that has the highest posterior probability is chosen as the best set of ancestral amino acids.

Analyzing six mammalian lysozyme *c* sequences,

Correspondence to: M. Nei; e-mail: nxm2@psu.edu

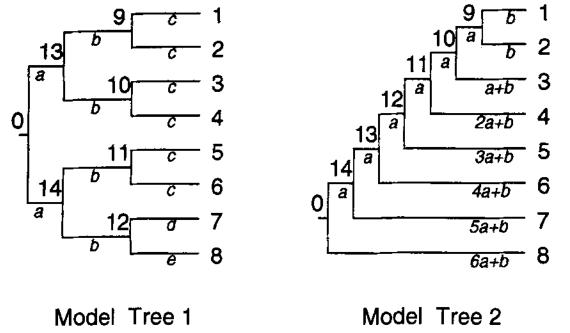

Fig. 1. Model trees used in computer simulation. The branch lengths are measured in terms of the expected number of amino acid substitutions per site. The actual values of *a, b, c, d,* and *e* used are given in Table 1.

Yang et al. (1995) suggested that the ML method gives more accurate results than the MP method, but their computation of the accuracy depended on a number of assumptions about the amino acid frequencies and substitution pattern. In practice, their assumptions may not hold, so the accuracy of the ancestral amino acids inferred by the ML and MP methods remains unclear. We have therefore studied the accuracy of inferred ancestral amino acids by using computer simulation. In the process of this study, however, we came to realize that the branch length estimation in the ML method is time-consuming and that this estimation can be done much more efficiently by using a least squares method. We therefore developed a distance method of inferring ancestral sequences, in which the branch lengths are estimated by the least squares method and the ancestral amino acids are inferred by the Bayesian approach. In this paper we first present our results of computer simulation concerning the ML and MP methods and then discuss the new distance method of inference of ancestral amino acids and its efficiency.

Computer Simulation

In this computer simulation, we considered two different model trees, each with eight protein sequences (Fig. 1). Jones et al.'s (1992) empirical model (JTT model) of amino acid substitution was used to simulate the evolutionary change of amino acid sequences. The method of our computer simulation was as follows. First, a random sequence of 100 amino acids at node 0 in Fig. 1 was generated with the expected frequency of each amino acid equal to its equilibrium frequency specified by the JTT model. This sequence then evolved according to the predetermined branching pattern of the model tree. Random amino acid substitutions (mutations) were introduced following the JTT model, with the expected number of substitutions per amino acid site for a branch being equal to the branch length assigned. Thus, the ancestral protein sequences at all interior nodes and the "present-day" sequences at all exterior nodes were generated and recorded.

Once the eight present-day sequences were generated, we used the MP and ML methods to infer the ancestral sequences at all interior nodes, assuming that the unrooted topology was known. We used our own computer program for the MP method but Yang (1995)'s program CODEML of the PAML package for the ML method. At each amino acid site, the ancestral amino acids were inferred for all interior nodes simultaneously, and they were compared with the ancestral amino acids recorded in the simulation. In the MP method, several evolutionary pathways (sets of inferred amino acids for all interior nodes) that were equally parsimonious were often obtained. When there were *n* equally parsimonious pathways, the accuracy of pathway reconstruction (inference) was defined as 1/*n* if the correct pathway was included. The accuracy of the inferred amino acid at each interior node was also computed. This accuracy was defined as *m*/*n*, where *m* is the number of parsimonious pathways in which the correct amino acid was obtained at the particular interior node examined. In the ML method, only the pathway that had the highest posterior probability (called the best pathway) was considered. The accuracy of the ML pathway reconstruction was defined to be 1 if the best pathway was correct; otherwise it was 0. The accuracy of inferred amino acid for a given interior node was defined as 1 if the amino acid in the best pathway was correct for the node; otherwise it was 0. For both the MP and ML methods, 200 replicate simulations were conducted, unless otherwise mentioned.

The average accuracy for the entire sequence can be computed by considering all sites, all variable sites, or all parsimony informative sites (Kumar et al. 1993). However, since the ancestral amino acids at nonparsimony informative sites were almost always correctly inferred by both methods, we present only the average accuracies for parsimony informative sites in this paper. These accuracies are obviously lower than those for all variable sites or all sites.

Tree Topologies and Levels of Sequence Divergence. Table 1 shows the average accuracies of inferred amino acids by the MP and ML methods for the two different model trees (Fig. 1). In both model trees, we considered three levels of sequence divergence, and the expected number of substitutions per site between two most distantly related sequences was 0.2, 0.6, and 1.2 for the low, intermediate, and high levels of sequence divergence, respectively. It is clear that for both model trees the average accuracy of amino acid reconstruction is lower for an interior node located close to the tree root than for an interior node close to the exterior nodes, as expected. When the level of sequence divergence is low, however, even the nodes closest to the root have an accuracy of 90% or higher. Therefore, the amino acid reconstruction for a node is quite accurate. However, when the level of sequence divergence is high, the accuracy is rather low, particularly when the MP method is used.

The accuracy of pathway reconstruction is obviously lower than that of amino acid reconstruction for individual nodes because in this case the amino acids for all interior nodes must be correct. This is particularly so when the level of sequence divergence is high. When this level is low, however, the accuracy is only slightly lower than that for individual nodes.

The accuracy of inferred ancestral amino acids is always higher in the ML method than in the MP method. However, the difference in accuracy between the two methods is smaller in model tree 1 than in tree 2. In the former tree, the difference is nearly the same for different levels of sequence divergence, whereas in the latter the difference is substantial for the high level of sequence divergence. This difference is probably caused by the fact that in the MP method no consideration is given to branch lengths, and thus the accuracy of inferring amino acid change from node to node is low. Note that some branches in tree 2 are much longer than the branches in tree 1.

Effects of Branch Lengths. We have seen that the accuracy of inferred ancestral amino acids is affected by the level of sequence divergence. However, the pattern and extent of the effect of branch lengths are unclear. We therefore studied the accuracy at interior node 12 of model tree 1 by changing branch length *e* (from interior node 12 to exterior node 8; see Fig. 1). In this study, all other branch lengths were the same as those in the case of high sequence divergence. The results obtained are shown in Fig. 2. When branch length *e* is 0, exterior node 8 converges to interior node 12, and thus the accuracy at node 12

Table 1. Average accuracies of inferred ancestral amino acids by the maximum-parsimony, maximum-likelihood, and distance methods for different model trees and different levels of sequence divergence[a]

	Tree 1									Tree 2								
Divergence	Low			Intermediate			High			Low			Intermediate			High		
	MP	ML	D	MP	ML	D	MP	ML	D	MP	ML	D	MP	ML	D	MP	ML	D
Node 9	99	99	99	96	97	97	84	87	87	98	99	99	95	97	97	87	91	91
10	99	99	99	96	97	97	84	87	87	97	99	99	92	96	96	85	90	90
11	99	99	99	96	97	97	84	87	87	97	99	99	90	95	95	81	88	88
12	99	99	99	96	97	97	84	87	86	96	98	98	88	94	93	78	86	85
13	97	98	98	90	92	92	77	81	81	94	98	96	86	92	92	73	83	83
14	97	98	98	90	92	92	77	82	81	93	97	95	84	89	89	68	79	78
Pathway	93	94	94	77	82	81	47	55	55	83	92	90	65	79	77	43	64	63

[a] The percent accuracies for parsimony informative sites are given. MP: maximum-parsimony method. ML: maximum-likelihood method. D: distance method. In model tree 1, we used branch lengths $a = 0.02$ (amino acid substitutions per site), $b = 0.03$, $c = d = e = 0.05$; $a = 0.1$, $b = 0.1$, $c = d = e = 0.1$; and $a = 0.1$, $b = 0.2$, $c = d = e = 0.3$ for the low, intermediate, and high levels of sequence divergence, respectively. In model tree 2, $b = 4a$, and $a = 0.01$, 0.03, and 0.06 were used for the low, intermediate, and high levels of sequence divergence, respectively

for the ML method becomes 1.00. In the case of the MP method, however, the accuracy is 0.94. The reason for this is that in the MP method amino acid substitutions are assumed to occur for each branch even if the branch length is 0. This is clearly incorrect and therefore lowers the accuracy. In both methods, the accuracy decreases as the branch length increases, and when e is 0.9, it becomes 0.78 and 0.73 for the ML and MP methods, respectively.

Effects of the Number of Sequences Used. When we reconstruct a protein sequence of the common ancestor for a group of organisms, will the accuracy depend on the number of sequences used? One would expect that the accuracy will increase with the number of sequences used, because more information is available. We studied this problem by eliminating or adding one or two present-day sequences for model tree 1.

Let us first consider the case of elimination of one sequence from model tree 1. All branch lengths except e were assumed to be the same as those in the case of high sequence divergence. For various values of branch length e, we compared the accuracies of inferred amino acids for the case where sequence 8 was included or excluded in the reconstruction (see Fig. 3A). Figure 4A shows that the accuracy for interior node 14 decreases when sequence 8 is excluded but the decrease is small when e is large. This is reasonable because as e increases sequence 8 becomes less informative for the inference of ancestral amino acids at interior node 14. However, the effect of exclusion of sequence 8 is generally small. Even when e is as small as 0.1, which is one-third of branch length d (from interior node 12 to exterior node 7), the decrease of the accuracy for interior node 14 is only 0.07 for both the MP (from 0.78 to 0.71) and ML (from 0.84 to 0.77) methods. The effect of elimination of sequence 8 on the accuracies for other interior nodes (e.g., see the result for interior node 11 in Fig. 4A) is even smaller, since sequence 8 is distantly related to them.

We now consider the case of addition of two sequences to model tree 1 (Fig. 3B). The two sequences added are denoted by 7′ and 8′. The branch lengths from interior node 12 to exterior nodes 7, 7′, 8, and 8′ were all assumed to be h, whereas all other branch lengths were the same as those in the case of high divergence. We compared the accuracies of ancestral sequences for the case where sequences 7′ and 8′ are included or excluded in the reconstruction. Figure 4B shows that the accuracy for interior node 12 increases when sequences 7′ and 8′ are included, and the amount of increase is greater when h is large than when h is small. The reason for this difference is that when h is small, sequences 7, 7′, 8, and 8′ are the same for most amino acid sites, and the two additional sequences do not provide extra information. When h is large, however, the two sequences give useful information for the inference of ancestral amino acids. Nevertheless, the effect of addition of two sequences is generally small. Even when h is as large as 1.0, the accuracy of inferred amino acids for node 12 increases only by 0.09 (from 0.55 to 0.64 for the MP method; Fig. 4B). The effect on the accuracies for other interior nodes is even smaller (e.g., see the result for node 11 in Fig. 4B). Therefore, at least in this situation, use of two additional sequences does not increase the accuracy very much. We did not study this effect for the ML method, because it required an enormous amount of computer time. However, we believe that the effect of additional two sequences will be similar for both the ML and MP methods, as in the case of Fig. 4A.

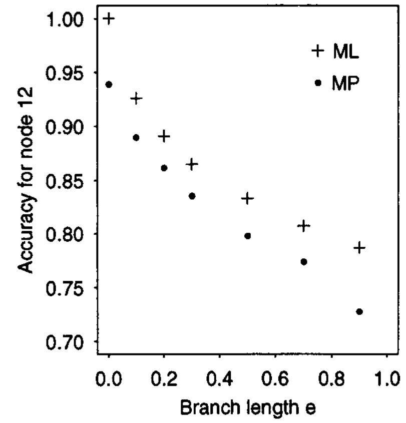

Fig. 2. Relationships of the accuracy of inferred ancestral amino acids and branch length e. Model tree 1 is used with branch lengths $a = 0.1$, $b = 0.2$, and $c = d = 0.3$.

These results suggest that addition or elimination of one or a few present-day sequences does not change the accuracy of inferred ancestral sequences very much.

Effects of Substitution Models. To reconstruct ancestral sequences by the ML method, a model of amino acid substitution is required, whereas no such model is needed for the MP method. Since we usually do not know the real pattern of amino acid substitution for a given protein, we will have to use an available model even if it may not be an appropriate one. Therefore, it is important to know the accuracy of the ML method when an incorrect model is used. To study this problem, we used the JTT model to generate the present-day sequences and

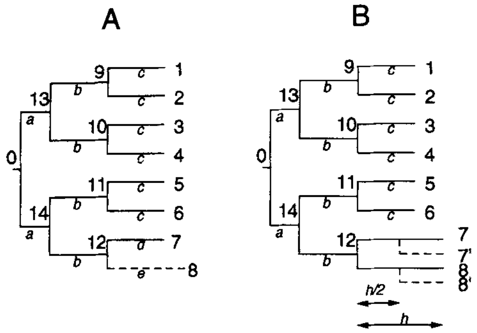

Fig. 3. Model trees used for studying the effects of the number of present-day sequences on the accuracies of inferred ancestral amino acids. **A** A model tree to show the elimination of sequence 8. The branch lengths used are $a = 0.1$, $b = 0.2$, and $c = d = 0.3$. **B** A model tree to show the addition of two sequences 7′ and 8′. The branch lengths used are $a = 0.1$, $b = 0.2$, and $c = 0.3$. The lengths from interior node 12 to exterior nodes 7, 7′, 8, and 8′ are all h.

applied the JTT, Dayhoff (Dayhoff et al. 1978), or Poisson (equal probability of any amino acid change) model to infer the ancestral sequences. The average accuracies of inferred evolutionary pathways are given in Table 2. The ML reconstruction with the JTT model is only slightly more accurate than that with the Dayhoff model. The substitution (transition) matrix for the JTT model is based on a large data set of amino acid substitutions and is somewhat different from that for the Dayhoff model. Yet, the accuracy of inferred amino acids is nearly the same for the two models. This suggests that minor differences in substitution model do not affect the results seriously. The very simple Poisson model, which is clearly unrealistic, gives accuracies which are somewhat lower than those for the JTT and Dayhoff models. However, even this model gives better results than the MP method particularly in model tree 2. From these results, we may conclude that the effect of differences in substitution model in the ML method is relatively small but that the correct substitution model gives the best results as expected.

Effects of Wrong Topologies. To reconstruct the ancestral sequences from present-day sequences, we have to know the phylogenetic relationship (tree topology) of the sequences. However, the tree topology is not always well established. It is therefore interesting to know whether we can still obtain reliable ancestral sequences when a wrong topology is used. Of course, it is meaningless to reconstruct the ancestral sequence of an interior node that does not exist in the real phylogeny. We therefore examined the accuracy of inferred amino acids only for the correct part of the tree when the branching pattern of other parts of the tree is incorrect. For this purpose, we generated the present-day sequences according to model tree 2 (Fig. 1) with the high level of sequence divergence, interchanged sequences 2 and 3, and then reconstructed ancestral amino acids by using this partially incorrect topology. We assessed the accuracies for all the nodes other than node 9. For both the ML and MP methods, the accuracy for interior node 10 decreased a little (by less than 0.05), but the accuracies for other interior nodes were hardly affected. We obtained similar results when we interchanged sequences 5 and 6 in model tree 2 or sequences 2 and 3 in model tree 1 (data not shown). These results show that even when the topology is not entirely correct, the inferred amino acids are still reliable for the correct part of the tree.

Distribution of the Number of Equally Parsimonious Pathways. In the MP method, it is difficult to identify the correct pathway when there are many equally parsimonious pathways. To have some idea of the number of equally parsimonious pathways the MP method generates, we studied the distribution of this number. For a set of amino acid sequences, let $N(i)$ be the number of sites that have i equally parsimonious pathways. The proportion of sites that have i pathways is then given by

$$f(i) = N(i) / \sum_{i=1}^{M} N(i) \tag{1}$$

where M is the largest value of i observed. Chandrasekharan et al. (1996) used the proportion of sites having a single MP pathway [$f(1)$] as a measure of the reliability of MP reconstruction. However, the MP reconstruction with a single pathway is not necessarily correct. Let $n(i)$ be the number of sites that have i equally parsimonious pathways when the correct pathway is included. The probability of including the correct pathway at a site that has i pathways is then given by

$$g(i) = n(i)/N(i) \tag{2}$$

Note that $N(i)$, $n(i)$, $f(i)$, and $g(i)$ can be defined for all sites, variable sites, or parsimony informative sites.

Figure 5A–C shows the distributions of $f(i)$ and $g(i)$ for parsimony informative sites when model tree 1 is used with the low, intermediate, and high levels of sequence divergence. When the level of sequence divergence is low, $f(1)$ is over 90% and $g(i)$ is nearly 1 irrespective of i. When the divergence level is high, $f(1)$ becomes about 50%, and in this case $g(i)$ is close to 75% irrespective of i. These results indicate that when sequence divergence is low the correct pathway is almost always included in the parsimonious pathways identified, but for a high divergence level the correct pathway may not be included whether the number of pathways is small or large. The same results were obtained for model tree 2 as well.

Distribution of the Posterior Probability of the Best ML Pathway. In the ML method, the best pathway is determined by computing the posterior probability (p). As the extent of sequence divergence increases, the probability is expected to decrease, and the inferred ancestral amino acids become unreliable. This can be seen by examining the distribution of p values. For a set of amino acid sequences, let $K(x)$ be the number of sites whose p is within the probability interval $(x, x + 0.05)$, where $x = 0, 0.05, 0.10, \ldots, 0.95$. The proportion of sites whose p is within the interval $(x, x + 0.05)$ is therefore given by

$$u(x) = K(x) / \sum_{x=0}^{0.95} K(x) \tag{3}$$

It is also important to know whether the p value obtained by the ML method is equal to the true probability of the best pathway. Let $k(x)$ be the number of sites whose p is within the interval $(x, x + 0.05]$ when the best pathway is correct. Then, the probability that the best pathway is correct when the p value is within the interval $(x, x + 0.05)$ is given by

$$v(x) = k(x)/K(x) \tag{4}$$

If the posterior probability p computed in the ML method is an unbiased estimator of the true probability of the best pathway, $v(x)$ should be approximately equal to $x + 0.025$. Note that $K(x)$, $k(x)$, $u(x)$, and $v(x)$ can again be defined for all sites, variable sites, or parsimony informative sites.

Figure 6A–C shows the distributions of $u(x)$ and $v(x)$ for parsimony informative sites when model tree 1 is used with the low, intermediate, and high levels of sequence divergence. For the case of low sequence divergence, $u(0.95)$ is 81%, but it decreases as the divergence level increases and becomes 3% for the case of high divergence. This indicates that when the divergence level is high most of the best pathways chosen are not very reliable. However, $v(x)$ is still very close to $x + 0.025$. Therefore, the posterior probability of the best pathway is a good estimator of the true accuracy. Note that $v(x)$ is subject to large sampling errors when $u(x)$ is very small, as is clear from Fig. 6A.

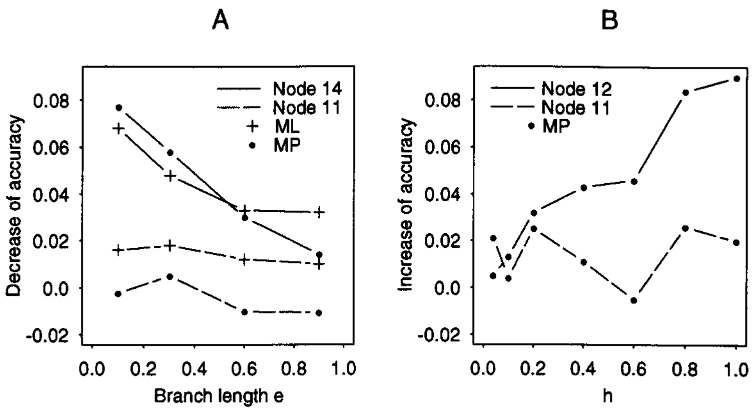

Fig. 4. Effects of elimination or addition of present-day sequences on the accuracy of inferred amino acids. **A** Decrease of the accuracy due to elimination of sequence 8. The model tree used is shown in Fig. 3A. **B** Increase of the accuracy due to the addition of sequences 7′ and 8′. The model tree used is shown in Fig. 3B.

Table 2. Average accuracies of pathway (set of ancestral amino acids) reconstructions when various substitution models are used[a]

	Tree 1			Tree 2		
Sequence divergence	Low	Intermediate	High	Low	Intermediate	High
MP	93	77	47	83	65	43
ML-P	93	77	48	90	75	60
ML-D	94	82	54	91	77	64
ML-JTT	94	82	55	92	79	64

[a] The percentage accuracies for parsimony informative sites are given. Present-day sequences were generated by the JTT model, and the ancestral amino acids were inferred by the MP method or the ML method with the Poisson (ML-P), Dayhoff (ML-D), or JTT (ML-JTT) model

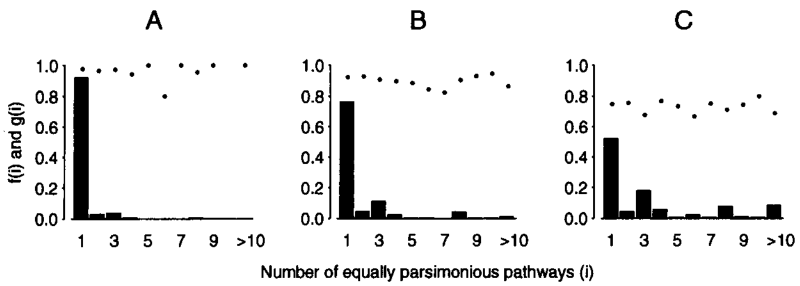

Fig. 5. Distribution of the proportion of sites that have i equally parsimonious pathways ($f[i]$; shown by *bars*) and the distribution of the probability of including the correct pathway when there are i equally parsimonious pathways ($g[i]$; shown by *dots*). The distributions are for the parsimony informative sites when model tree 1 is used with the (**A**) low, (**B**) intermediate, and (**C**) high levels of sequence divergence. In each case, 1,000 replicate simulations were conducted.

However, when a wrong model of amino acid substitution is used in the ML method, the posterior probability p may be a biased estimator. We investigated this problem by using a model that is simpler or more complex than the real substitution model in ML reconstruction. We first used the Poisson model in ML reconstruction when the present-day sequences were generated according to the JTT model. The distributions of $u(x)$ and $v(x)$ for the case of high divergence level are presented in Fig. 6D. We can see that $v(x)$ is smaller than $x + 0.025$, which means p gives an overestimate under a simpler model. And the extent of overestimation is high when p is around 0.6–0.8. We then simulated sequence evolution according to the Poisson model but inferred ancestral amino acids under the JTT model. The distributions of $u(x)$ and $v(x)$ for this case indicate that p gives an underestimate when it is around 0.7 (Fig. 6E). Because it is usually difficult to know the real pattern of amino acid substitution for a given protein, the posterior probability computed may give a biased estimate. Therefore, we must be careful in the interpretation of p values in ML reconstruction.

Distance Method

We have seen that the ML method generally gives more reliable reconstructions of ancestral amino acids than the MP method and that this is primarily due to the fact that in the ML method the branch lengths are taken into account. Unfortunately, estimation of branch lengths by the ML method is time-consuming, and in our experience it often gives zero branch lengths. However, branch lengths need not be estimated by the ML method; they can be estimated by various statistical methods including the least squares (LS) method. We therefore propose that branch lengths be estimated by the LS method but that the ancestral amino acids be inferred by the same posterior probability method as used by Yang et al. (1995).

To explain this method (distance method), let us consider an unrooted tree of five sequences (Fig. 7). In the inference of ancestral sequences we must assume that the topology of the tree is known, but the branch lengths are to be estimated. One problem with the standard LS method of branch length estimation (e.g., Rzhetsky and Nei 1993) is that some branch length estimates may become negative. We therefore suggest that the ordinary LS method with the constraint of nonnegative branches be used (e.g., Lawson and Hanson 1974; Felsenstein 1995).

Once all the branch lengths are estimated, we can compute the probability of observing the data for each amino acid site in the following way. We denote the amino acids of the five exterior nodes of the tree (Fig. 7) by the vector $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5)$ and the amino acids at the three interior nodes by the vector $\mathbf{y} = (y_1, y_2, y_3)$. If we use a

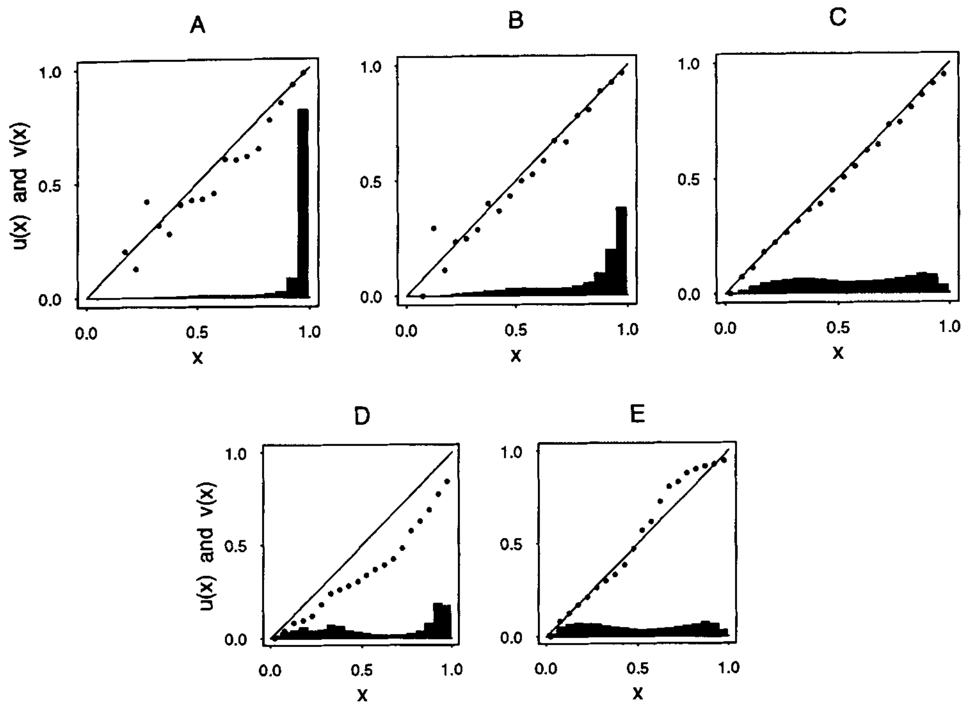

Fig. 6. Distribution of the proportion of sites whose p values are within the probability interval $(x, x + 0.05]$ ($u[x]$; shown by *bars*) and the distribution of the probability that the best pathway at a site is correct when the p value is within $(x, x + 0.05)$ ($v[x]$; shown by *dots*). The *diagonal line* shows the prediction of $v(x)$ ($= x + 0.025$) when p is an unbiased estimator. **A–C** Same distributions when the JTT model is used for generating the present-day sequences and inferring ancestral amino acids. Model tree 1 was used with the (**A**) low, (**B**) intermediate, and (**C**) high levels of sequence divergence. **D** Same distributions when the JTT model is used to generate the present-day sequences but the Poisson model is used to infer ancestral amino acids and the level of sequence divergence is high. **E** Same distributions when the Poisson model is used to generate the present-day sequences, whereas the JTT model is used to infer ancestral amino acids and the level of sequence divergence is high. All the plots are for parsimony informative sites. In each case, 500 replicate simulations were conducted.

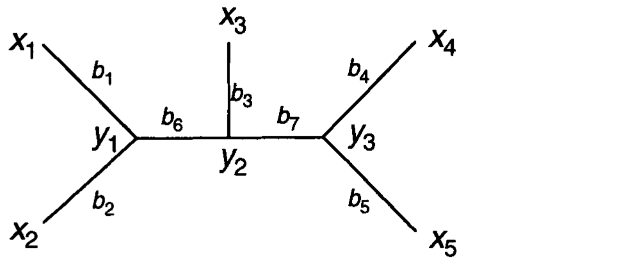

Fig. 7. A model tree for explaining the distance method. x_1–x_5: present-day amino acids. y_1–y_3: ancestral amino acids. b_1–b_7: branch lengths.

time-reversible model of amino acid substitution, we can start the evolutionary change of amino acids at any node, but here we use y_1 as the starting amino acid. The probability of observing data **x** is given by

$$\begin{aligned} f(\mathbf{x};\mathbf{b}) &= \sum_{\mathbf{y}} f(\mathbf{y})f(\mathbf{x}|\mathbf{y};\mathbf{b}) \\ &= \sum_{y_1}\sum_{y_2}\sum_{y_3}[\pi_{y_1}P_{y_1x_1}(b_1)P_{y_1x_2}(b_2)P_{y_1y_2}(b_6)P_{y_2x_3}(b_3) \\ &\quad P_{y_2y_3}(b_7)P_{y_3x_4}(b_4)P_{y_3x_5}(b_5)] \end{aligned} \tag{5}$$

where $\mathbf{b} = (b_1, b_2, b_3, b_4, b_5, b_6, b_7)$ is the vector of estimated branch lengths of the tree and $P_{ij}(b_k)$ is the probability of change from amino acid i to amino acid j when the branch length b_k is given. Note that $P_{ij}(b_k)$ can be computed by the method described by Dayhoff et al. (1978). In equation (5), $f(\mathbf{y})$ is the prior probability of **y**, which is given by

$$f(\mathbf{y}) = \pi_{y_1}P_{y_1y_2}(b_6)P_{y_2y_3}(b_7) \tag{6}$$

where π_{y_1} is assumed to be equal to the relative frequency of amino acid y_1 as is usually done. The other element of equation (5) is $f(\mathbf{x}|\mathbf{y};\mathbf{b})$, which is the conditional probability of observing data **x**, given the ancestral amino acids **y**, and is given by

$$f(\mathbf{x}|\mathbf{y};\mathbf{b}) = P_{y_1x_1}(b_1)P_{y_1x_2}(b_2)P_{y_2x_3}(b_3)P_{y_3x_4}(b_4)P_{y_3x_5}(b_5) \tag{7}$$

Although equation (5) has a form similar to the likelihood function of Felsenstein (1981) or Yang et al. (1995), it is not used to estimate any parameter. It is just probability of observing data **x**, when the substitution model and branch lengths are known. Here, we are particularly interested in estimating **y** and use the Bayesian approach to compute the following posterior probability for each set of ancestral amino acids y_1, y_2, and y_3 (evolutionary pathway).

$$f(\mathbf{y}|\mathbf{x};\mathbf{b}) = \frac{f(\mathbf{y})f(\mathbf{x}|\mathbf{y};\mathbf{b})}{f(\mathbf{x};\mathbf{b})} \tag{8}$$

There are 20 different kinds of amino acids so that theoretically y_1, y_2, and y_3 can each take 20 different character states. In practice, the amino acids that are not observed at the amino acid site under consideration are unlikely to have ever appeared at the site. In fact, if we consider such amino acids, $f(\mathbf{y}|\mathbf{x};\mathbf{b})$ is usually vanishingly small. Therefore, we exclude them from consideration to speed up the computation. This procedure is adopted in Yang et al.'s method (1995) as well.

At any rate, if we compute $f(\mathbf{y}|\mathbf{x};\mathbf{b})$ for all combinations of amino acids y_1, y_2, and y_3, we will know the set of amino acids that have the highest posterior probability. We can then infer that these amino acids are the ancestral amino acids. If this is done for all sites of the amino acid sequences under consideration, we can determine the ancestral sequence at each interior node.

In the above formulation we considered a tree for five sequences, but the same method can be used for any number of sequences.

We conducted a computer simulation to examine the accuracy of ancestral amino acids inferred by this distance method. The simulation of sequence evolution was the same as that described before. After the present-day sequences were generated, we computed pairwise distances between the sequences by using the amino acid gamma distance with the shape parameter of $\alpha = 2.4$ (Ota and Nei 1994) and estimated branch lengths by using the LS method with the constraint of nonnegative branches. In practice, we used Felsenstein's (1995) FITCH algorithm to estimate the branch lengths. Once the branch length estimates were obtained, we inferred the ancestral amino acids and assessed the accuracies of the amino acids obtained. Here we used the gamma distance with $\alpha = 2.4$, because it is very close to the true distance (equivalent to PAMs of Dayhoff et al. 1978) for sequence divergence under the JTT model (J. Zhang, unpublished). Our simulation results showed that the accuracy of inferred ancestral sequences by this new distance method is virtually the same as that obtained by the ML method for every case examined (Table 1). We also compared the computational time required for the inference of ancestral amino acid sequences by the ML and distance methods. This computational time depends on the number, length, and divergence level of the sequences used. For example, in the case of the computer simulation for model tree 2 (eight sequences) with the intermediate level of sequence divergence, the distance method was about 10 times faster than the ML method. In an analysis of 20 sequences of abalone sperm lysin (Lee et al. 1995), we found that the distance method was about 200 times faster than the ML method.

It is obvious that the distance method can also be used for inferring ancestral nucleotides when DNA sequence data are given. In this case we can use any substitution model that is time-reversible, as long as evolutionary distances are estimable (e.g., Yang 1994; Rzhetsky and Nei 1995).

Discussion

Substitution Model

Although information on branch lengths is important for improving the accuracy of inferred amino acids, use of an appropriate substitution model also contributes to the improvement. Cao et al. (1994) and Adachi and Hasegawa (1995) have shown that the JTT model better fits actual data when it is modified so as to make the equilibrium amino acid frequencies equal to the observed frequencies. This modified model is called the JTT-f model. Collins et al. (1994) also suggested that when the observed amino acid frequencies of the present-day sequences are different from the equilibrium values specified by the model used, the amino acid frequencies of inferred ancestral sequences may be different from those of the present-day sequences. We therefore recommend that in general the JTT-f model be used in actual data analysis. This model is incorporated into Yang's (1995) PAML program package and our program of the distance method.

Other Methods of Ancestral Sequence Reconstruction

Besides the MP, ML, and distance methods discussed in this paper, there are a few other methods available for reconstructing ancestral amino acid or nucleotide sequences, though they are not used very often. Libertini and Di Donato (1994) introduced the inferential method, which is essentially the same as the MP method. In this method, however, amino acid sequences are reverse-translated into nucleotide sequences to carry out the reconstruction. This is expected to decrease the reliability of inferred amino acids, because the accuracy of inferred ancestral nucleotides is expected to be relatively low and the reverse-translation introduces ambiguity due to redundancy of the genetic code. Libertini and Di Donato (1994)'s computer simulation also showed that the inferred ancestral amino acid sequences by their method are not as reliable as the inferred nucleotide sequences.

Assuming a probabilistic model of character state change, Maddison (1995) also presented a method of computing the accuracy of inferred ancestral states by the MP method. However, his method does not take into account branch lengths, so the applicability of his method may be limited.

Schluter (1995) developed a likelihood method to reconstruct ancestral sequences. This method is different from Yang et al.'s (1995). In Schluter's method, the substitution model is assumed to be unknown, and the parameters in the model are estimated by maximizing the likelihood function of observing the present-day amino acids at a given site. The ancestral states that give the highest likelihood, conditional on the estimated substitution parameters, are regarded to be the best reconstructions. Since in this method the substitution parameters are estimated by using information at a single site, stochastic errors are expected to be very large, and this would make the inference of ancestral states unreliable. However, if a site is subject to positive selection and the substitution pattern is very different from a general model such as the JTT model, Schluter's method may give better results since it is free from the assumption about the substitution model. At present, the reliability of this method remains unclear.

After completion of our simulation work, Koshi and Goldstein (1996) published a Bayesian method for ancestral sequence reconstruction, which is similar to Yang et al.'s (1995) method. Generally speaking, Koshi and Goldstein's method is expected to perform well, since it also takes into account both the branch lengths and the substitution pattern. However, there are some differences between this method and Yang et al.'s method. The major one is that in Koshi and Goldstein's method, the tree topology and branch lengths are estimated by the neighbor-joining method (Saitou and Nei 1987), whereas in Yang et al.'s method, the topology is predetermined and branch lengths are estimated by the ML method. Another difference is that in Koshi and Goldstein's method, reconstruction of amino acids is done node by node rather than for all interior nodes simultaneously as in the case of the ML, MP, and distance methods we considered. Although the simultaneous reconstruction gives the same results as those obtained node by node in most cases, they are not always the same (Yang et al. 1995). Koshi and Goldstein (1996) also conducted some computer simulations to investigate the reliability of their method. The results obtained are generally consistent with our results. However, they did not compare the accuracy of their method with the MP method. The model tree and the substitution model used in their simulation were different from ours. In most of their simulations, they assumed that both the tree topology and branch lengths are known. In our simulation for the ML and distance methods, the topology was assumed to be known, but the branch lengths were estimated.

Reconstructions at Sites Under Positive Selection

We are usually interested in knowing ancestral amino acids at sites where the amino acid changes have affected the protein function. It is possible that at these sites the pattern of amino acid substitution is different from the JTT or Dayhoff model, and this difference may introduce some bias in our estimates of the posterior probabilities (p values) of the inferred amino acids. However, our computer simulation suggests that the accuracy of inferred amino acids would not be affected very much, and even when an incorrect substitution model is used the ML and distance methods give more reliable results than the MP method. Nevertheless, it is advisable to use the MP method in addition to the ML or distance method. If the MP method, which does not assume any substitution model, gives the same inferred amino acids as those obtained by the ML or distance method, the results would be more reassuring.

Program availability: A computer program for inferring ancestral amino acid sequences by using the distance method is available upon request.

Acknowledgments. We are grateful to Ziheng Yang for modification of his computer program CODEML to fit our simulations and helpful discussion, to Sudhir Kumar for discussions, and to Naoko Takezaki and Tatyana Sitnikova for their comments on earlier versions of this manuscript. This work was supported by research grants from the NIH and the NSF to M.N.

References

Adachi J, Hasegawa M (1995) Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. J Mol Evol 40:622–628

Cao Y, Aadachi J, Janke A, Pääbo S, Hasegawa M (1994) Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J Mol Evol 39:519–527

Chandrasekharan UM, Sanker S, Glynias MJ, Karnik SS, Husain A (1996) Angiotensin II-forming activity in a reconstructed ancestral chymase. Science 271:502–505

Collins TM, Wimberger PH, Naylor GJP (1994) Compositional bias, character-state bias, and character-state reconstruction using parsimony. Syst Biol 43:482–496

Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO (ed) Atlas of protein sequence and structure, vol 5, suppl 3. National Biomedical Research Foundation, Washington, DC, pp 345–352

Eck RV, Dayhoff MO (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

Felsenstein J (1995) PHYLIP: phylogeny inference package. Version 3.57c. University of Washington, Seattle

Fitch WM (1971) Toward defining the course of evolution: minimum change for a specific tree topology. Syst Zool 20:406–416

Hartigan JA (1973) Minimum evolution fits to a given tree. Biometrics 29:53–65

Jermann T, Opitz JG, Stackhouse J, Benner SA (1995) Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. Nature 374:57–59

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Koshi JM, Goldstein RA (1996) Probabilistic reconstruction of ancestral protein sequences. J Mol Evol 42:313–320

Kumar S, Tamura K, Nei M (1993) MEGA: molecular evolutionary genetics analysis, version 1.01. Pennsylvania State University, University Park

Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs, NJ, pp 158–165

Lee Y-H, Ota T, Vacquier VD (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol Biol Evol 12:231–238

Libertini G, Di Donato A (1994) Reconstruction of ancestral sequences by the inferential method, a tool of protein engineering studies. J Mol Evol 39:219–229

Maddison WP (1995) Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. Syst Biol 44:474–481

Maddison WP, Maddison DR (1992) MacClade: analysis of phylogeny and character evolution. Version 3. Sinauer, Sunderland, MA

Ota T, Nei M (1994) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J Mol Evol 38:642–643

Rzhetsky A, Nei M (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol 10: 1073–1095

Rzhetsky A, Nei M (1995) Tests of applicability of several substitution models for DNA sequence data. Mol Biol Evol 12:131–151

Saitou N, Nei M (1987) The neighbor joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Schluter D (1995) Uncertainty in ancient phylogenies. Nature 377:108–109

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

Yang Z (1995) PAML: phylogenetic analysis by maximum likelihood. Version 1.1. Pennsylvania State University, University Park

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650