J Mol Evol (1982)18:387–404

Journal of Molecular Evolution

Accuracy of Estimated Phylogenetic Trees from Molecular Data

I. Distantly Related Species

Yoshio Tateno[1], Masatoshi Nei, and Fumio Tajima

Center for Demographic and Population Genetics, University of Texas at Houston, Houston, Texas 77025, USA

Summary. The accuracies and efficiencies of four different methods for constructing phylogenetic trees from molecular data were examined by using computer simulation. The methods examined are UPGMA, Fitch and Margoliash's (1967) (F/M) method, Farris' (1972) method, and the modified Farris method (Tateno, Nei, and Tajima, this paper). In the computer simulation, eight OTUs (32 OTUs in one case) were assumed to evolve according to a given model tree, and the evolutionary change of a sequence of 300 nucleotides was followed. The nucleotide substitution in this sequence was assumed to occur following the Poisson distribution, negative binomial distribution or a model of temporally varying rate. Estimates of nucleotide substitutions (genetic distances) were then computed for all pairs of the nucleotide sequences that were generated at the end of the evolution considered, and from these estimates a phylogenetic tree was reconstructed and compared with the true model tree. The results of this comparison indicate that when the coefficient of variation of branch length is large the Farris and modified Farris methods tend to be better than UPGMA and the F/M method for obtaining a good topology. For estimating the number of nucleotide substitutions for each branch of the tree, however, the modified Farris method shows a better performance than the Farris method. When the coefficient of variation of branch length is small, however, UPGMA shows the best performance among the four methods examined. Nevertheless, any tree-making method is likely to make errors in obtaining the correct topology with a high probability, unless all branch lengths of the true tree are sufficiently long. It is also shown that the agreement between patristic and observed genetic distances is not a good indicator of the goodness of the tree obtained.

Key words: Nucleotide substitution – Genetic distance – Species tree – Gene tree – UPGMA – Fitch/Margoliash method – Farris method – Modified Farris method

Introduction

One of the important subjects in the study of molecular evolution is how to construct a phylogenetic tree from molecular data. This subject has been called molecular taxonomy (e.g., Nei 1978). Various methods have been proposed for constructing phylogenetic trees from amino acid sequences, nucleotide sequences, and electrophoretic data. These methods can be classified into two groups. In the first group, genetic distances for all pairs of species are computed, and a tree is constructed from these distance data. In the second the property and relationship of amino acid or nucleotide sequences from different species are used for the construction of a tree. The first group includes the unweighted pair-group method (UPGMA, Sneath and Sokal 1973), Edwards and Cavalli-Sforza's (1965) method, the least squares method (Cavalli-Sforza and Edwards 1967), the additive tree method (Cavalli-Sforza and Edwards 1967), Fitch and Margoliash's (F/M) (1967) method, Farris' (1972) method, Moore et al.'s (1973a) method, and others. UPGMA was originally proposed for phenetic classification by Sokal and Michener (1958), but Nei (1975) applied this method for making a phylogenetic tree under the assumption that the expected number of gene substitutions is proportional to the evolutionary time (constant rate of substitution). Chakraborty (1977)

[1] *Present address:* Institute of Physical and Chemical Research, Rikagaku Kenkyusho, Hirosawa, Wako-shi, Saitama, 351, Japan

Offprint requests to: M. Nei

0022–2844/82/0018/0387/$ 03.60

showed that if the topology is known, this method gives least squares estimates of branch lengths. The second group includes Dayhoff's (1969) ancestral sequence method, Moore et al.'s (1973b) maximum parsimony method, and others.

We know, however, very little about the accuracies of these methods. Actual data are virtually useless for studying the accuracy of tree-making methods, since the true tree is not known. Fossil records are quite uninformative for this type of study simply because there are not enough data in most cases. Thus, the only possible way to examine the accuracy is to conduct computer simulation using certain model trees. Peacock and Boulter (1975) studied this problem by using computer simulation, but they employed an amino acid substitution model rather than a nucleotide substitution model. They examined only two methods, i.e., Dayhoff's ancestral sequence method and Moore et al.'s (1973a) method. On the other hand, Prager and Wilson (1978) studied the discrepancy between the observed and estimated (Farris' (1972) patristic) distances for the phylogenetic trees constructed by UPGMA, Fitch and Margoliash's method, and Farris' method using empirical data. Their criterion of a best tree was that the percent squared deviation of the patristic distances from the observed distances be minimal. However, since the exact phylogenetic tree is unknown, this type of study does not necessarily give correct information.

In this series of papers we shall investigate the accuracies of tree-making methods in the first group. Except in one case, we assume that the expected rate of gene substitution is constant, though the actual number of substitutions for a given period of time may vary because of stochastic errors. We make this assumption, since most molecular data satisfy this assumption approximately (Wilson et al. 1977). In this study we have chosen three methods, i.e., UPGMA, Fitch and Margoliash's method, and Farris' method, since these three methods are more often used than others in molecular taxonomy. In addition to these methods we have modified Farris' method for the case where a distance measure proportional to evolutionary time is used, and included this modified Farris method in the present study. In this paper we shall study the phylogenetic trees constructed from nucleotide sequences, whereas in the second paper the phylogenetic trees obtained from gene frequency data will be examined. Since we are interested in tree-making methods based on genetic distances, we shall not use the second group of tree-making methods, though some of these may produce a better tree than the first group. We note that the first group of methods are much simpler than the second group and in some types of data such as immunological distance or the genetic distance estimated from restriction-site data, the second group of methods are not applicable. We shall first describe our modified Farris method, and then examine the accuracies of the four methods.

Modified Farris Method

Farris' method is intended to construct a parsimonious phylogenetic tree but requires a metric that complies with the triangle inequality. No consideration is made about the effect of stochastic errors in the process of evolution, and in the presence of this effect his method tends to give overestimates of branch lengths, as will be shown later. On the other hand, many measures of gene substitutions such as Jukes and Cantor's (1969) estimate of the number of nucleotide substitutions, Sarich and Wilson's (1966) immunological distance, and Nei's (1972) genetic distance are not metrics and often violate the triangle inequality because of backward and parallel mutations. They are also subjected to a large extent of stochastic errors. Therefore, it is necessary to modify his method to make it applicable to these measures. Just like Farris' method, our modified Farris method is a heuristic one and does not necessarily give the most parsimonious tree. However, it generally gives better estimates of branch lengths compared with Farris' method. Before discussing or method, let us first describe Farris' method briefly.

Consider the distance matrix given in Table 1. Suppose that distance D_{12} is the smallest in the matrix. OTUs 1 and 2 are then combined first. The distance between this combined OTU (1,2) and each of the three remaining OTUs is computed by taking the average of the distance between OTU 1 and a third OTU and that between OTU 2 and the third OTU. Suppose that the distance between OTU (1,2) and OTU 3 is the smallest among the distances thus obtained. Then, OTU 3 is combined with OTU (1,2) as shown in Fig. 1a. This figure represents a network rather than a rooted tree, because this method cannot determine the most ancestral point (the evolutionary origin of all OTUs concerned). In this figure X is a branching point. Each branch length of the network is computed by the following formulas:

$$L(3,X) = (D_{13} + D_{23} - D_{12})/2, \quad (1a)$$

$$L(1,X) = D_{13} - L(3,X), \quad (1b)$$

$$L(2,X) = D_{23} - L(3,X), \quad (1c)$$

where L(a,b) represents the length between points a and b. In practice, these values are computed for every pair

Table 1. Distance matrix for five OTUs

OTU	2	3	4	5
1	D_{12}	D_{13}	D_{14}	D_{15}
2		D_{23}	D_{24}	D_{25}
3			D_{34}	D_{35}
4				D_{45}

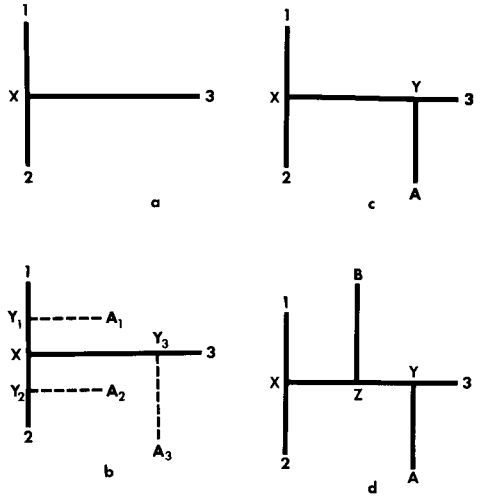

Fig. 1 a-d. Trees represented by networks; **a** Tree for three OTUs. Numbers 1 to 3 refer to three OTUs and X to a branching point; **b** Three possible ways by which OTU A is connected to the tree in (a). A_1, A_2, and A_3 are the three possibilities, and Y_1, Y_2, and Y_3 are the corresponding branching points; **c** Tree for four OTUs. Numbers 1 to 3 and A refer to four OTUs, and X and Y to branching points; **d** One possible way to add OTU B to the tree in (c). Z is a branching point

of OTU (1,2) and the remaining OTUs, and the OTU which shows the smallest length to X is chosen.

We now proceed to the next step where one more OTU is added to the network. There are three possibilities for one (A, say) of the remaining OTUs (4 and 5) to be connected to the network. Namely, OTU A may be connected at point Y_1, Y_2, or Y_3 in Fig. 1b. The subscripts of A in the figure correspond to the three possibilities. The branch lengths $L(A_1,Y_1)$, $L(A_2,Y_2)$, and $L(A_3,Y_3)$ are then computed. This computation is done for all remaining OTUs (4 and 5), and the OTU which gives the smallest branch length is chosen to be connected to the network. In practice, $L(A_i,Y_i)$'s are computed by the following formulas:

$$L(A_1,Y_1) = [L(A_1,1) + L(A_1,X) - L(1,X)]/2 \quad (2a)$$

$$L(A_2,Y_2) = [L(A_2,2) + L(A_2,X) - L(2,X)]/2 \quad (2b)$$

$$L(A_3,Y_3) = [L(A_3,3) + L(A_3,X) - L(3,X)]/2. \quad (2c)$$

In these formulas $L(A_1,1)$, $L(A_2,2)$, and $L(A_3,3)$ are directly obtained from the distance matrix, whereas L(1,X), L(2,X), and L(3,X) have already been computed by using (1). On the other hand, $L(A_1,X)$, $L(A_2,X)$, and $L(A_3,X)$ are computed by the following formulas:

$$L(A_1,X) = L(A_1,2) - L(2,X) = L_1, \text{ or}$$
$$= L(A_1,3) - L(3,X) = L_2, \quad (3a)$$

$$L(A_2,X) = L(A_2,1) - L(1,X) = L_3, \text{ or}$$
$$= L_2, \quad (3b)$$

$$L(A_3,X) = L_3 \text{ or } L_1. \quad (3c)$$

Among L_1, L_2, and L_3 Farris chooses the largest value and uses it for all of $L(A_i,X)$ in (3). Suppose that $L(A_3,Y_3)$ was the smallest. Then OTU A is connected to the branch 3-X, as shown in Fig. 1c.

The last OTU (B which is either OTU 4 or 5) is then added to the network in Fig. 1c. The connecting procedure is the same as the above, except that there are five possible ways of connection in this case. However, there is one problem. To see this, let us consider the case where OTU B is connected to the branch X-Y, as given in Fig. 1d. According to Farris, L(B,Z) is given by

$$L(B,Z) = [L(B,X) + L(B,Y) - L(X,Y)]/2. \quad (4)$$

The problem is that L(B,X) or L(B,Y) cannot be obtained from the distance matrix directly, because X and Y are not OTUs but branching points. Therefore, Farris does not use (4) for obtaining L(B,Z) but uses the result of the previous computation in which OTU A was added to the network. In this computation every member of the remaining OTUs was tested to find the OTU to be added to the network in Fig. 1a. The value of L(B,Z) was computed in this step of testing. At any rate, the final network, which includes all OTUs concerned, is produced in this way.

As mentioned above, we cannot decide the most ancestral point in this method. One of Farris' suggestions is that this point be determined by assuming that the evolutionary rates of the two most divergent branches in the network are equal. In this paper we follow this suggestion.

There seem to be two problems in applying Farris' method to molecular data. The first is the estimation of branch lengths $L(A_i,X)$'s by the largest value of L_1, L_2, and L_3 in (3). Since estimates of genetic distances are generally subject to large sampling errors, this procedure is expected to lead to an overestimate of $L(A_i,X)$, and subsequently an overestimate of $L(A_i,Y_i)$, which may be serious particularly when the number of OTUs is large. In fact, our computer simulation has shown that the distance between two OTUs estimated by this method is often much larger than the actual value (see below). In practice, many measures of molecular changes of genes tend to give an underestimate when the values of the measures are extremely large. Farris' method may correct this underestimation under certain circumstances, but this type of correction often leads to an overcorrection because of stochastic errors even if some caution is exercised (Tateno and Nei 1978; Nei and Tateno 1978). At any rate, in this paper we are primarily interested in the phylogenetic trees constructed by using distance measures of which the expected value is proportional to evolutionary time. For these measures Farris' procedure does not seem to be appropriate. The second problem is the replacement of L(B,Z) in (4) by

the value obtained in the previous computation. Since the network in the previous computation is different from the one in which L(B,Z) is computed, this replacement is not justified.

To make Farris' method appropriate to molecular data, we introduce different ways of computing $L(A_i,X)$ and L(B,Z). We consider a general network given in Fig. 2. We propose that the branch lengths L(F,G) in this network be estimated by the following formula:

$$L(F,G) = [L(F,S) + L(F,T) - L(S,T)]/2. \quad (5)$$

This formula corresponds to (2) or (4), but L(F,S) and L(F,T) are obtained in a different way. In Fig. 2 we note that L(F,S) can be computed by any of the following three formulas:

$$L(F,S) = L(1,F) - L(1,S),$$
$$= L(2,F) - L(2,S),$$
$$= L(4,F) - L(4,S).$$

Note that all six quantities on the right hand side of the equations are already known. We estimate L(F,S) by taking the average of the three values obtained by the above formulas. Similarly, L(F,T) may be estimated either by L(3,F) - L(3,T) or by L(5,F) - L(5,T). We again take the average of these values to estimate L(F,T). L(S,T) is already known from the previous computation.

Since we use an average of two or more estimates to compute a branch length which is not directly obtainable, we can avoid the overestimation that may occur in Farris' method. This procedure also helps to reduce the effect of random errors. Furthermore, our estimation of a branch length is always based on the relevant network, so that the second deficiency in Farris' method is also eliminated. We call this method the modified Farris method, and include it in the following study. (A computer program for this modified Farris method may be obtained by writing to M. Nei.)

Model and Method of Computer Simulation

The computer simulation used is briefly as follows: A hypothetical gene of 300 ancestral nucleotide sequences (100 codons) was duplicated at each branching point of the model tree used and subjected to nucleotide substitution, and all descendant nucleotide sequences were examined at the terminal points of the tree. From these sequences genetic distances (numbers of nucleotide substitutions) for all pairs of the sequences were estimated. These distances were then used as input data for making trees, and the tree reconstructed was compared with the model tree to see how accurately each of the four methods reconstructs the tree.

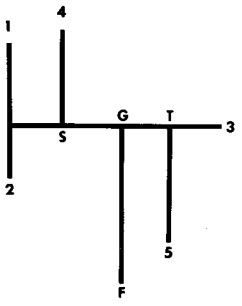

Fig. 2. A network for six OTUs. Numbers 1 to 5 and F refer to six OTUs, and G, S, and T to branching points

Nucleotide substitution was assumed to occur purely at random. Thus, each of the 300 nucleotides was replaced by any of the three other nucleotides with an equal probability (1/3). When a nonsense codon occurred as a new mutation, it was eliminated, and another mutation was generated to substitute the nonsense mutation. We used three different methods to determine the actual number of nucleotide substitutions for a given branch of the model tree. In the first method the number was determined by using the Poisson distribution, whereas in the second a negative binomial distribution was used. In these two cases the expected rate of nucleotide substitution was constant over all branches. The actual number of substitutions for a given branch was generated by using pseudorandom numbers, and thus varied from branch to branch even though the expected number was the same. In the third method the expected rate of substitution varied with unit evolutionary time in each branch, following a gamma distribution. In all cases the actual number of nucleotide substitutions for each branch was recorded.

To set up a model tree, we must determine the number of OTUs and the topology. Since a large-scale computer simulation with many replications was needed, we could not use a large number of OTUs. We decided to use 8 OTUs except in the study of the effect of the number of OTUs. With respect to the topology, there are basically two different types as shown in Fig. 3. In these trees eight OTUs are represented by numbers 1 to 8, and M or L represents the expected number of nucleotide substitutions for the shortest branch. Both of these trees are topologically extreme, and the topologies of actual trees are generally somewhere between them. Preliminary studies have shown that when the entire evolutionary time (T in tree a and T' in tree b) is fixed, tree a is subject to more errors than tree b. This is because tree a includes shorter branches than tree b, and the shorter the branch length the larger the coefficient of variation of the number of substitutions. Since we are interested in the errors that occur in tree-making, we have decided to use tree a in our study.

To obtain a distance matrix to be used as the input data for reconstructing a tree, we computed the number of nucleotide differences between every pair of the eight nucleotide sequences. However, this number does not represent the actual number of nucleotide substitutions,

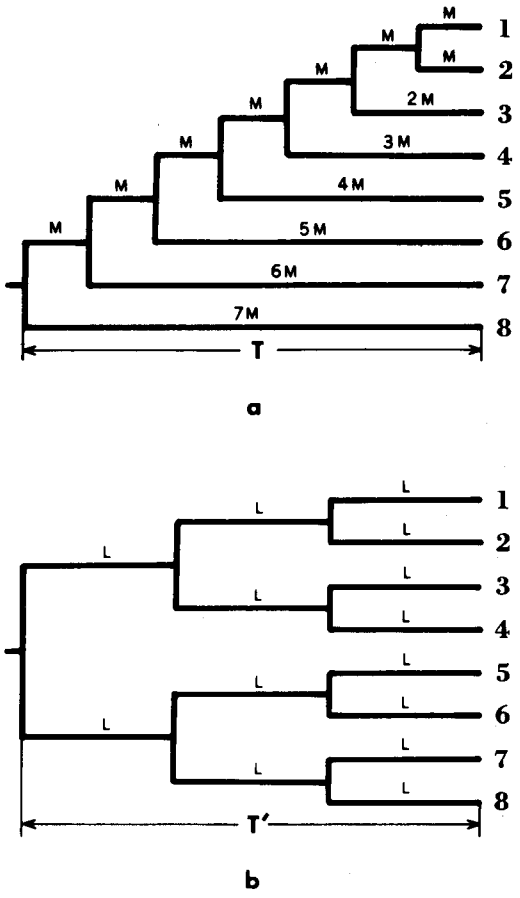

Fig. 3. Two types of model trees considered in computer simulation. Numbers 1 to 8 refer to 8 OTUs. M or L represents the *expected* number of nucleotide substitutions in the shortest branch, whereas T or T' the evolutionary time considered

since multiple substitutions might have occurred at the same nucleotide site. Therefore, we used Jukes and Cantor's (1969) method to estimate the number of nucleotide substitutions from the data on the number of nucleotide differences. (Note that this estimate does not necessarily satisfy triangle inequality). The matrix of the estimated numbers of nucleotide substitutions thus obtained was used for reconstructing a tree. Except in the F/M method, the phylogenetic tree is uniquely determined for a given set of genetic distance data. In the case of the F/M method, several trees are constructed for each data set, and the best tree is chosen. In our study we constructed 15 trees for each replication and chose the best tree according to Fitch and Margoliash's (1967) criterion.

Measure of Deviation of a Reconstructed Tree from the Model Tree

Two criteria are important in measuring the deviation of a reconstructed tree from the model tree. One is the degree of distortion of the topology of the reconstructed tree, and the other is the amount of deviation of the estimated branch lengths from the true lengths.

1. Topological Errors

In this study we used Robinson and Foulds' (1981) method to measure the degree of topological errors. In this method the minimum number of operations (contractions and decontractions of branches) that transform one tree into another is used as a measure of topology distortion, and it can be obtained by counting the number of unmatched branches (edges in graph theory) between the two trees to be compared. In this paper we call this the distortion index and denote it by d_T. This is approximately twice the minimum number of nearest branch interchanges that are required for changing one tree to the other, when d_T is small. Therefore, it has a high correlation with Waterman and Smith's (1978) nearest neighbor interchange metric or Tateno's (1978) DI value, but its computation is straightforward and simple. This method has been developed for unrooted trees, but it can be used for rooted trees as well if we regard the root as one OTU.

In addition to d_T we have also computed the proportion (P) of replications in which the correct topology was obtained. This P can also be used for comparing the accuracies of different tree-making methods.

2. Errors in the Estimates of Branch Lengths

The best way to measure the deviation of estimated branch lengths from true lengths would be to compute the squared deviation of a branch length of the reconstructed tree from that of the model tree, and take the average over all branches. However, this is not practicable, because the topology of a reconstructed tree is often different from that of the model tree. The second choice is to use the matrices of expected and estimated (patristic) distances for pairs of OTUs. In the present case the expected distances can be obtained by adding the expected numbers of nucleotide substitutions for all branches concerned, whereas the patristic distances are obtained from the estimated branch lengths of the reconstructed tree in the same way. Since the agreement between the sum of expected numbers and the patristic distance does not necessarily mean the agreement between the expected and estimated distances for each branch, this comparison is less satisfactory than the first comparison. However, the matrix of patristic distances is obtainable for any type of topology, so that we can compute the average squared deviation of patristic distances from expected distances irrespective of the topology. We would expect that if a reconstructed tree is

close to the model tree, then the average squared deviation is small. In practice, it is more convenient to use the square root of the average squared deviation. Namely,

$$S_E = \sqrt{\underset{i>j}{2\Sigma} (D_{ij} - D_{ij}')^2/(n(n-1))} \quad (6)$$

where D_{ij} and D_{ij}' are the patristic and expected distances between OTUs i and j, respectively, and n is the number of OTUs. We call this the average deviation of patristic distances from expected distances.

In the present computer simulation we know the expected branch lengths, so that we can compute the S_E value. In actual data, the true topology and branch lengths are unknown, and S_E is not computable. In this case, however, the following quantity may be computed.

$$S_0 = \sqrt{\underset{i>j}{2\Sigma} (D_{ij} - D_{ij}'')^2/(n(n-1))}, \quad (7)$$

where D_{ij}'' is the observed distance between OTUs i and j. We call this the average deviation of patristic distances from observed distances. This quantity is similar to Fitch and Margoliash's (1967) percent standard deviation and Ferris' (1972) homoplasy measure. It should be noted that D_{ij}'' can deviate considerably from the true expected distance by chance or by the varying rate of nucleotide substitution and thus a small value of S_0 does not necessarily mean the closeness of the estimated branch lengths to the true branch lengths measured in evolutionary time. However, if many genes are used for estimating D_{ij}'', the error introduced in this way will be averaged out and D_{ij}'' is expected to become close to D_{ij}'. S_0 would then be used as an indicator of the closeness of branch lengths and true branch lengths. At any rate, it is very important to know the correlation between S_E and S_0.

When an evolutionary tree is constructed from amino acid or nucleotide sequence data, the length of a tree branch is sometimes represented by an estimate of the actual number of nucleotide or amino acid substitutions that occurred in that branch. If this estimation of actual numbers is a part of the purpose of tree construction, S_0 is a better measure of the errors in the estimates of branch lengths than S_E. It should be noted, however, that this type of tree (both topology and branch lengths) varies considerably with the gene (or protein) used, as will be seen from the trees of vertebrate species constructed by Goodman et al. (1974) by using hemoglobin α and β chains. This is so despite the fact that the real evolutionary tree must be the same for all genes and there must be only one true tree. Actually, what is important is to estimate this true tree which represents the actual pathways of evolution of a group of species expressed in terms of geological time (Nei 1977). We call this the species-tree in contrast to the protein-tree or gene-tree, in which the branch length is equated to an estimate of the actual number of amino acid or nucleotide substitutions. Among the four methods of tree-making to be studied here, UPGMA is intended to construct a species-tree (or population tree), where the lengths of two descendant branches from an ancestral stock are assumed to be the same, as they should be. All other methods are primarily concerned with a gene-tree. In the following we consider both types of trees. S_E measures the deviation from the true (species) tree, whereas S_0 measures the deviation from a gene-tree, which is specific to each gene (each replicate in our simulation).

Results

1. Topological Errors

Nucleotide Substitutions Following the Poisson Distribution

We shall first present the results of our computer simulation in which the actual number of nucleotide substitutions for each branch of the model tree was assumed to follow the Poisson distribution. The mean of this distribution (expected number of nucleotide substitutions) is given by M or its multiple, as shown in Fig. 3a. We used three different M values, 2, 4, and 8. For each of these M values 20 replicate computations were made and 20 sets of the distance matrices were produced. Each of the four methods mentioned earlier was then used to reconstruct 20 trees using the 20 distance matrices. The reconstructed trees were compared with the model tree and the distortion indices (d_T) were computed. In this paper we are primarily interested in rooted trees, but we have also examined unrooted trees to see the extent of topological errors. The average d_T values ($\bar{d}_T$) over the 20 replications are presented in Table 2. In addition to $\bar{d}_T$, this table includes the proportion (P) of replications in which the tree topology was correctly reconstructed.

It is clear from Table 2 that in the case of M = 2 the $\bar{d}_T$ values for rooted trees are considerably smaller in the Farris and modified Farris methods than in UPGMA and Fitch and Margoliash's (F/M) method. It is also noted that the $\bar{d}_T$ values are more or less the same for the former two methods and also for the latter two methods. P is generally high when $\bar{d}_T$ is low. These results suggest that the Farris and modified Farris methods are somewhat better than the other two methods for the case of M = 2. In the case of M = 4, however, all the four methods give virtually the same values $\bar{d}_T$ and P. In UPGMA and the F/M method the increase in M improves the topology of a reconstructed tree considerably, and in the case of M = 8 the topology is correct

Table 2. Mean distortion indices ($\bar{d}_T$) and proportions of correct trees (P) when the actual number of nucleotide substitutions followed the Poisson distribution. These results are based on 20 replications. The number of OTUs is 8

		UPGMA	F/M method	Farris method	Modified Farris method
			(i) Rooted Tree		
M = 2	$\bar{d}_T$	3.60 ± 0.52	3.65 ± 0.54	2.20 ± 0.54	2.60 ± 0.46
	P	0.15	0.10	0.45	0.20
M = 4	$\bar{d}_T$	1.40 ± 0.36	1.40 ± 0.33	1.40 ± 0.34	1.40 ± 0.33
	P	0.50	0.45	0.45	0.45
M = 8	$\bar{d}_T$	0.40 ± 0.18	0.40 ± 0.18	1.40 ± 0.29	1.20 ± 0.27
	P	0.80	0.80	0.40	0.45
			(ii) Unrooted Tree		
M = 2	$\bar{d}_T$	2.90 ± 0.39	2.95 ± 0.42	1.20 ± 0.39	1.60 ± 0.34
	P	0.15	0.10	0.60	0.40
M = 4	$\bar{d}_T$	1.20 ± 0.34	1.10 ± 0.31	0.70 ± 0.26	0.90 ± 0.27
	P	0.55	0.55	0.70	0.60
M = 8	$\bar{d}_T$	0.30 ± 0.16	0.10 ± 0.10	0.30 ± 0.22	0.50 ± 0.25
	P	0.85	0.95	0.90	0.80

with a probability of 80%. This is, of course, expected, since the coefficient of variation of the number of nucleotide substitutions becomes smaller as M increases under the Poisson law. Unlike our expectation, however, the topologies obtained by the Farris and modified Farris methods do not improve when M increases from 4 to 8, the $\bar{d}_T$ and P values being virtually the same. Consequently, they are less accurate than those obtained by UPGMA and the F/M method in the case of M = 8.

The poor performance of the Farris and modified Farris methods when M is large is caused mainly by the error that occurs at the time of putting the root to the tree. As mentioned earlier, we put the root at the midpoint of the line connecting the two most divergent OTUs in these two methods. Because of the stochastic errors involved, however, this midpoint did not always occur on the longest branch. We have therefore computed $\bar{d}_T$ and P for unrooted trees (Table 2). It is clear that the topology of unrooted trees is much better than that for rooted trees in terms of both $\bar{d}_T$ and P. However, compared with the other two methods, the Farris and modified Farris methods do not necessarily produce a better topology when M = 8. It is noted that in the case of unrooted trees the Farris method tends to give a slightly better topology than the modified Farris method, but the difference is not statistically significant.

In practice, almost every phylogenetic tree obtained from amino acid or nucleotide sequence data is based on a single protein or gene, and includes many branches whose estimated numbers of nucleotide substitutions are small. Our results suggest that a reconstructed tree including branches with a small number of substitutions (say less than 4) is quite erroneous. However, the error can be reduced by increasing the number of genes if the expected rates of nucleotide substitution in evolution are constant. If we note that the sum of two or more Poisson variables is also a Poisson variable, it is clear that the result for M = 8, for instance, is the same as that for the case where two or more genes are involved but the total number of nucleotide substitutions for all genes is equal to 8. Therefore, if a number of different genes are used and the estimated number of nucleotide substitutions for each branch is 8 or more, we can expect that the topology of an unrooted tree is correct with a high probability, P being equal to or higher than 0.8 (see Table 2). When a rooted tree is to be constructed, however, the accuracy of the topology does not necessarily increase with increasing M if the Farris or modified Farris method is used. In this case UPGMA or the F/M method is preferable.

Another important factor that affects the topology of a reconstructed tree is the number of OTUs. To get some idea about the effect of this factor, we conducted a small-scale computer simulation, increasing the number of OTUs from 8 to 32 but keeping the topology of the model tree similar to that shown in Fig. 3a. The M value used was 4. In this study we did not include Fitch and

Table 3. Mean distortion indices ($\bar{d}_T$) and proportions of correct trees (P). The number of nucleotide substitutions followed the Poisson distribution. These results are based on 10 replications. The number of OTUs used is 32 and the unit branch length (M) is 4

	UPGMA	Farris method	Modified Farris method
		(i) Rooted Tree	
$\bar{d}_T$	28.4 ± 1.5	29.0 ± 1.6	29.4 ± 1.3
P	0	0	0
		(ii) Unrooted Tree	
$\bar{d}_T$	27.0 ± 1.4	23.6 ± 1.6	24.8 ± 1.0
P	0	0	0

Table 4. Means of the actual number of nucleotide substitutions (n_a), the estimated number (n_F) by the Farris method, and the estimated number (n_M) by the modified Farris method. n_e is the expected number. Comparisons are made between OTU 1 and eight other OTUs

OTUs compared	n_e	n_a	n_F	n_M
OTUs 1 and 4	24	23.8 ± 1.5	24.3 ± 1.6	24.1 ± 1.5
OTUs 1 and 8	56	55.5 ± 2.0	59.6 ± 2.4	54.4 ± 2.1
OTUs 1 and 12	88	87.1 ± 3.6	99.3 ± 4.6	88.1 ± 4.5
OTUs 1 and 16	120	116.2 ± 3.3	137.7 ± 4.1	114.6 ± 3.7
OTUs 1 and 20	152	151.2 ± 3.3	190.2 ± 6.7	154.6 ± 3.1
OTUs 1 and 24	184	185.1 ± 3.0	238.3 ± 4.2	186.8 ± 4.8
OTUs 1 and 28	216	221.0 ± 3.9	294.9 ± 14.5	220.8 ± 5.3
OTUs 1 and 32	248	252.2 ± 5.2	361.0 ± 10.6	262.4 ± 4.8

Margoliash's method, since this method requires a large amount of computer time when the number of OTUs is large. The number of replications was 10 to save computer time. The results obtained are shown in Table 3. It is seen that $\bar{d}_T$ is large in all tree-making methods, and there are no significant differences among them, though UPGMA tends to show a larger value than the others when unrooted trees are made. The $\bar{d}_T$'s for unrooted trees are slightly smaller than those for rooted trees, but they are still large. In no cases was a correct topolgy obtained. When the number of OTUs is large, the effect of stochastic errors in nucleotide substitution is so large, that all tree-making methods seem to commit errors in topology construction with a high probability.

Earlier we pointed out that when the number of OTUs is large, Farris' method is expected to give an overestimate of the number of nucleotide substitutions. To confirm this, we compared the estimated (patristic) distances obtained by the Farris and modified Farris methods with the actual number of nucleotide substitutions in the case of 32 OTUs. The comparison was made between OTU 1 and each of eight other OTUs. The results obtained are given in Table 4 together with the expected number (n_e) of nucleotide substitutions. The table shows that when the expected number is small, the two estimates are close to the actual number (n_a), but as the expected number increases the estimate (n_F) by Farris' method becomes larger than the actual number, whereas the estimate (n_M) obtained by the modified Farris method is still close to the actual number. The standard deviation of n_F is also much larger than that of n_M when n_e is large. These findings clearly indicate that Farris' method gives gross overestimates when the number of OTUs is large.

One might think that this conclusion is valid only for nonmetric distance measures such as the one we used. However, as will be published elesewhere, our study on tree-making from gene frequency data has shown that overestimation of branch lengths in the Farris method occurs even with metric distances.

Nucleotide Substitutions Following the Negative Binomial Distribution

In the above studies the Poisson process of nucleotide substitution was assumed. In this case the variance of nucleotide substitutions is equal to the mean. There is, however, evidence that the variance of the number of nucleotide substitutions is roughly twice as large as the mean (Ohta and Kimura 1971; Langley and Fitch 1974). We have therefore conducted another simulation in which nucleotide substitution followed a negative binomial distribution. In this distribution we can easily adjust the ratio of the variance to the mean by changing the parameters. We have chosen the parameters so that the variance is twice as large as the mean. Note that in this case a sum of negative binomial variables is also a negative binomial variable. All other aspects of the simulation were the same as those for the previous study, and M = 2, 4, and 8 were used.

The results obtained are given in Table 5. It is clear that the Farris and modified Farris methods produce slightly more accurate trees than Fitch and Margoliash's or UPGMA on the average. However, each method produces less accurate trees than in the case of the Poisson distribution, whether the trees are rooted or unrooted. This occurs apparently because the coefficient of variation of the number of nucleotide substitutions is larger under the negative binomial distribution than under the Poisson distribution. This is confirmed by examining the relationship between the coefficient of variation (c.v.) and the mean distortion index. Note that the c.v. for the Poisson distribution with M = 2 is equal to that of the negative binomial distribution with M = 4, and the c.v. for the Poisson distribution with M = 4 is the same as that for the negative binomial distribution with M = 8. We can therefore compare the $\bar{d}_T$ values for the two types of distributions with the same c.v. values, using the data in Tables 2 and 5. This comparison shows that in both cases of rooted and unrooted trees the $\bar{d}_T$ values for the two distributions

Table 5. Mean distortion indices ($\bar{d}_T$) and proportions of correct trees (P) when the actual number of nucleotide substitutions followed the negative binomial distribution. These results are based on 20 replications. The number of OTUs is 8

		UPGMA	F/M method	Farris method	Modified Farris method
			(i) Rooted Trees		
M = 2	$\bar{d}_T$	4.20 ± 0.50	3.90 ± 0.57	3.90 ± 0.40	3.50 ± 0.38
	P	0.10	0.20	0.0	0.0
M = 4	$\bar{d}_T$	4.20 ± 0.56	3.80 ± 0.58	2.65 ± 0.55	2.35 ± 0.41
	P	0.10	0.10	0.25	0.15
M = 8	$\bar{d}_T$	2.00 ± 0.32	2.10 ± 0.31	1.60 ± 0.31	1.70 ± 0.26
	P	0.25	0.20	0.35	0.25
			(ii) Unrooted Trees		
M = 2	$\bar{d}_T$	3.30 ± 0.51	2.90 ± 0.51	2.30 ± 0.44	2.30 ± 0.42
	P	0.15	0.25	0.25	0.20
M = 4	$\bar{d}_T$	3.10 ± 0.55	3.00 ± 0.53	0.80 ± 0.37	0.90 ± 0.27
	P	0.25	0.25	0.75	0.60
M = 8	$\bar{d}_T$	1.10 ± 0.34	1.20 ± 0.30	0.40 ± 0.18	0.60 ± 0.21
	P	0.60	0.50	0.80	0.70

Table 6. Mean distortion indices ($\bar{d}_T$) and proportion of correct trees (P) when the expected number of nucleotide substitutions for each unit evolutionary time varied following the gamma distribution. These results are based on 20 replications. The number of OTUs is 8

		UPGMA	F/M method	Farris method	Modified Farris method
			(i) Rooted Tree		
M = 2	$\bar{d}_T$	5.65 ± 0.43	5.30 ± 0.51	6.40 ± 0.61	6.50 ± 0.58
	P	0	0	0	0
M = 4	$\bar{d}_T$	4.80 ± 0.57	4.60 ± 0.54	4.20 ± 0.46	4.10 ± 0.49
	P	0.05	0	0.05	0.10
			(ii) Unrooted Tree		
M = 2	$\bar{d}_T$	4.15 ± 0.46	3.90 ± 0.51	3.80 ± 0.54	4.50 ± 0.50
	P	0.05	0.05	0.05	0.05
M = 4	$\bar{d}_T$	3.80 ± 0.48	3.50 ± 0.43	2.10 ± 0.42	2.20 ± 0.41
	P	0.05	0	0.35	0.30

are more or less the same as long as the c.v. remains the same. This indicates that the coefficient of variation rather than the mean of the number of nucleotide substitutions is the important factor determining the accuracy of the topology of a reconstructed tree. It should also be noted that as the number of genes used increases, the accuracy of a reconstructed tree increases as in the case of the Poisson distribution. However, in the present case Farris' method and the modified Farris method show a slightly better performance than the other two methods even when M = 8.

Nucleotide Substitution with Varying Rates

In the two studies mentioned above we assumed that the rate of nucleotide substitution was constant. Although the constant rate is approximately correct (e.g. Zuckerkandl and Pauling 1962, 1965, Doolittle and Blombäck 1964, Margoliash and Smith 1965, Kimura 1969, King and Jukes 1969), the rate is not strictly constant (Ohta and Kimura 1971; Langley and Fitch 1974). We therefore investigated the effect of varying rate of substitution on reconstructed trees.

In this study we used a method similar to Ohta's (1976). We assumed that the expected rate of substitution varies with unit evolutionary time in each branch. The unit evolutionary time is the time corresponding to one M in the model tree. For each unit evolutionary time the number of nucleotide substitutions was determined by a Poisson distribution with parameter λ, and this Poisson parameter varied at random following a gamma distribution. We assumed that the variance of λ is twice the mean. Other aspects of the simulation were the same as those of the previous study. We studied two cases, i.e. $\bar{M} = 2$ and $\bar{M} = 4$, where $\bar{M}$ is the mean number

of nucleotide substitutions over all unit evolutionary times in all branches. In each case 20 replicate computations were made.

The results obtained are given in Table 6. It is seen that in the case of M = 4 the Farris and modified Farris methods are again slightly better than UPGMA and the F/M method. In the case of unrooted trees the former methods produced the correct topology in 30 or 35% of the 20 replications, whereas the latter methods in none or only one of the 20 replications. However, the differences in $\bar{d}_T$ and P among the four tree-making methods for rooted trees are rather small. When M = 2, all methods make errors in the construction of topology with a high probability in both rooted and unrooted trees. If we compare the $\bar{d}_T$ and P values in this table with the corresponding values in Table 2, it is clear that the varying rate of nucleotide substitution reduces the accuracy of the reconstructed tree substantially in all methods used. Comparison of Tables 5 and 6 is interesting, because the negative binomial distribution studied above corresponds to the case where the Poisson parameter varies spacially (among nucleotide sites) rather than temporally following the gamma distribution (e.g., Ohta 1976). This comparison shows that the temporal variation of substitution rate generally disturbs the topology of reconstructed trees more often than the spacial variation.

2. Errors in the Estimates of Branch Lengths

Nucleotide Substitution Following the Poisson Distribution

Figure 4 shows one example (replication) of computer simulation of nucleotide substitution and reconstructed trees for the case of M = 2. It is seen that although the model of constant rate of substitution was used the actual number of nucleotide substitutions for a branch is considerably different from the expected number (Fig. 3a). This is of course due to the stochastic nature of nucleotide substitution, and this clearly shows that the different numbers of nucleotide substitutions for a given evolutionary period do not necessarily mean the non-constant rate of substitution, as is often claimed by some authors. The genetic distances used for reconstructing trees are given in Table 7. These distances were estimated by Jukes and Cantor's (1969) formula, i.e.,

$$D = -300 \times (3/4)\log_e \{1 - (4/3)\pi\}$$

where π is the proportion of different nucleotides per site between the two sequences compared. These estimates are very close to the actual number of substitutions (Table 7).

Figure 4 shows that the topologies of the trees (d and e) reconstructed by the Farris and modified Farris methods are correct, but those (b and c) obtained by UPGMA and the F/M method are incorrect. The estimates of branch lengths obtained by the first two methods are generally close to the actual numbers of nucleotide substitutions, but those from the modified Farris method are slightly better than those from the Farris method, S_0 being 0.87 for the former compared with 1.03 for the latter. The S_E value, however, suggests that the estimated branch lengths from the former are slightly less close to the expected branch lengths than those from the latter. Since UPGMA and the F/M method produced an incorrect topology, it is difficult to compare the estimated branch lengths with the actual or expected lengths. However, if we use S_0 as a criterion, UPGMA is inferior to the other three methods. On the other hand, if we use S_E as a criterion, it is better than the

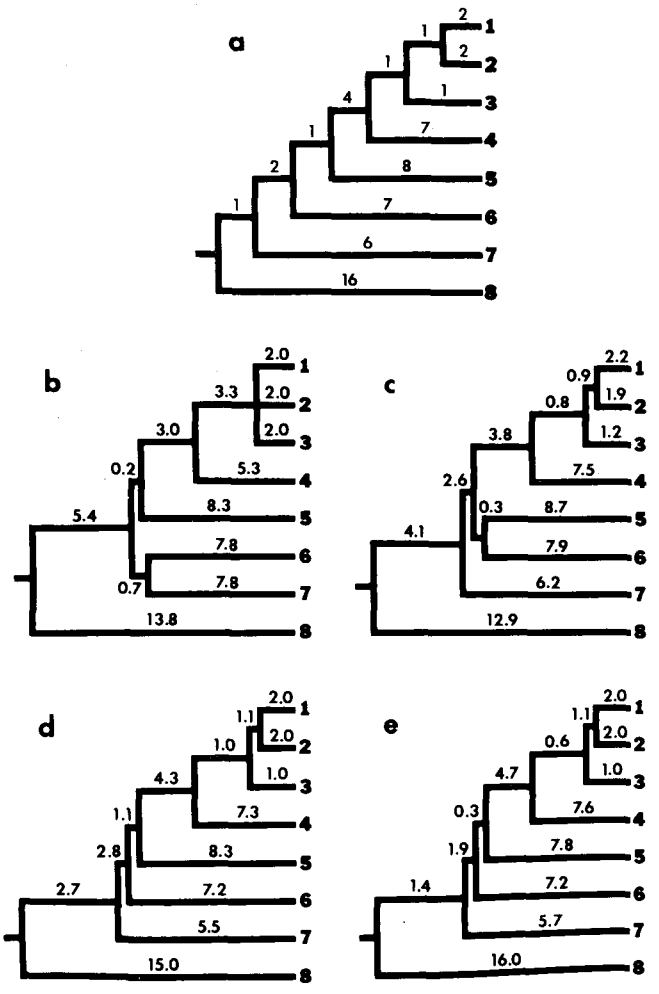

Fig. 4 a-e. One example (replication) of computer simulation. **a** True tree with the actual number of nucleotide substitutions for each branch. The nucleotide substitution followed the Poisson distribution with M = 2. **b** Tree reconstructed by UPGMA. $d_T = 3$; $S_0 = 1.91$; $S_E = 3.82$. **c** Tree reconstructed by the F/M method. $d_T = 2$; $S_0 = 1.09$; $S_E = 4.24$. **d** Tree reconstructed by the Farris method. $d_T = 0$; $S_0 = 1.03$; $S_E = 4.03$. **e** Tree reconstructed by the modified Farris method. $d_T = 0$; $S_0 = 0.87$; $S_E = 4.33$

Table 7. Estimated numbers (below diagonal) and observed numbers (above diagonal) of nucleotide substitutions (genetic distances) for each pair of OTUs. This represents the results of one replication in our computer simulation. The estimated genetic distances were obtained by Jukes and Cantor's formula. The reconstructed trees presented in Fig. 4 were obtained by using the estimated genetic distances in this table

OTU	1	2	3	4	5	6	7	8
1		4.0	4.0	11.0	16.0	16.0	17.0	28.0
2	4.0		4.0	11.0	16.0	16.0	17.0	28.0
3	4.0	4.0		9.0	14.0	14.0	15.0	26.0
4	11.3	11.3	9.2		19.0	19.0	20.0	31.0
5	15.5	16.6	14.5	19.9		16.0	17.0	28.0
6	16.6	14.5	14.5	19.9	16.6		15.0	26.0
7	17.7	15.5	15.5	20.9	17.7	15.5		23.0
8	28.8	26.5	26.5	31.0	29.9	27.6	23.2	

Table 8. Means of the average deviation (S_0) of patristic distances from observed distances and the average deviation (S_E) of patristic distances from expected distances. The actual number of nucleotide substitutions followed the Poisson distribution. The number of replications is 20

	UPGMA	F/M method	Farris method	Modified Farris method
		S_0		
M = 2	2.01 ± 0.13	1.04 ± 0.08	1.12 ± 0.13	0.71 ± 0.06
M = 4	3.54 ± 0.18	2.09 ± 0.16	2.87 ± 0.24	1.67 ± 0.15
M = 8	5.60 ± 0.32	3.70 ± 0.21	6.44 ± 0.55	3.16 ± 0.19
		S_E		
M = 2	4.44 ± 0.32	4.92 ± 0.29	5.07 ± 0.32	4.81 ± 0.30
M = 4	5.76 ± 0.44	6.80 ± 0.38	7.53 ± 0.58	6.61 ± 0.38
M = 8	8.27 ± 0.84	10.42 ± 0.85	12.01 ± 1.02	9.83 ± 0.71

others (see Fig. 4). The small value of S_E for UPGMA is of course due to the fact that this method is designed to give the same evolutionary distance for a pair of OTUs after they diverged. The S_0 and S_E for the F/M method are similar to those for the Farris method.

Figure 5 shows another example of simulated nucleotide substitution and reconstructed trees. In this case UPGMA and the F/M methods produced the correct topology, whereas the Farris and modified Farris methods gave an incorrect one. However, the modified Farris and Farris methods give smaller values of S_0 than those of the other two methods, as in the previous example. The S_E value is also smaller in UPGMA than in the other methods. In other words the relative performance of the four tree-making methods as judged by S_0 and S_E remains nearly the same as those in the previous example, even if the accuracies of the topologies made have changed. This casts some doubt about the utility of S_0 and S_E for judging the accuracy of branch lengths. Indeed, as will be shown later, the small values of these quantities do not necessarily mean that the estimates of branch lengths are close to the true values, though they are still a rough indicator of the accuracy of a tree. Before going into the detail of this problem, however, let us examine the general pattern of S_0 and S_E in our simulation studies.

Since the values of S_0 and S_E varied considerably with replication, we computed the means of these quantities over all replications. The results obtained are given in Table 8. It is clear that the modified Farris method gives the smallest value of S_0 for all three different values of M. Indeed, the S_0 value for the modified Farris method was significantly smaller in all cases except in the comparison of this method and the F/M method for M = 8. (The statistical test was conducted by taking the difference for each replication rather than by using the standard errors given in Table 8.) According to this criterion, the second best method is the F/M method, whereas the third is the Farris method for the cases of M = 2 and M = 4 but UPGMA for the case of M = 8. The poor performance of the Farris method when M = 8 is apparently caused by the overestimation of branch length discussed earlier. If we use S_E, however, UPGMA shows a significantly better performance compared with the other methods. With this criterion the second best result is obtained by the modified Farris method, whereas the Farris method shows the worst performance.

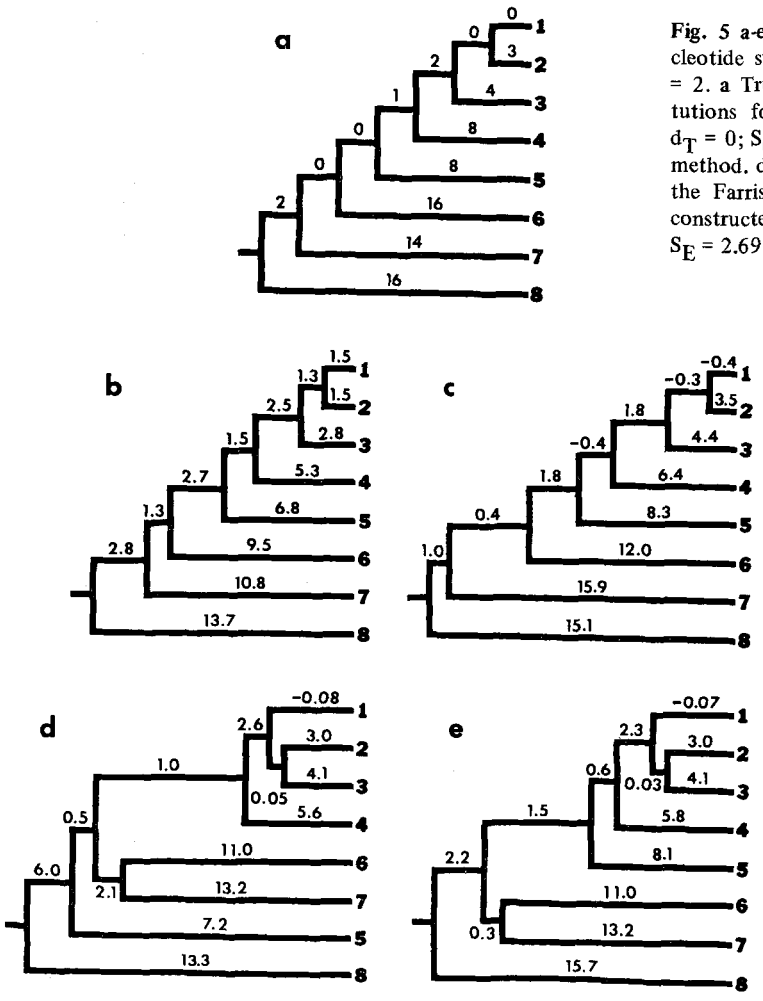

Fig. 5 a-e. Another example of computer simulation. The nucleotide substitution followed the Poisson distribution with M = 2. **a** True tree with the actual number of nucleotide substitutions for each branch. **b** Tree reconstructed by UPGMA. $d_T = 0$; $S_0 = 2.38$; $S_E = 1.70$. **c** Tree reconstructed by the F/M method. $d_T = 0$; $S_0 = 1.51$; $S_E = 4.02$. **d** Tree reconstructed by the Farris method. $d_T = 6$; $S_0 = 0.92$; $S_E = 2.75$. **e** Tree reconstructed by the modified Farris method. $d_T = 4$; $S_0 = 0.69$; $S_E = 2.69$

S_0 and S_E as a Measure of the Accuracy of Branch Lengths

Let us now study the reliability of our quantities S_0 and S_E as a measure of the accuracy of branch lengths estimated. This reliability can be studied by examining the agreement of estimated branch lengths with observed or expected branch lengths for the cases where the correct topology is obtained. In Example 1 the Farris and modified Farris methods give the correct topology, so that we can compute the average deviation (S_{0B}) of estimated branch lengths from observed branch lengths similar to S_0 and the average deviation (S_{EB}) of estimated branch lengths from expected branch lengths similar to S_E. The S_{0B} values obtained for the Farris and modified Farris methods are 1.830 and 1.784, respectively, whereas the S_{EB} values are 2.574 and 2.570. Therefore, with these criteria the former is slightly less accurate than the latter. To compare the accuracies of the four tree-making methods we must use only those replications in which the correct (unrooted) topology was obtained for all the methods. Unfortunately, there are only such replications in the case of M = 2, so that no reasonable comparison can be made (Table 9). In the case of M = 4 there were seven such replications, and the values of S_{0B} and S_{EB} are given in Table 9.

It is clear that UPGMA tends to give somewhat larger values of S_{0B} compared with the other methods, but there are no significant differences in the values of S_{0B} among the four different methods. With respect to S_{EB}, UPGMA gives the smallest value in all the replications examined, whereas the differences among the remaining three methods do not appear to be significant. A similar computation of S_{0B} and S_{EB} was made for the case of M = 8, where there were 13 replications in which the correct (unrooted) topology was obtained by all the four methods. The means of S_{0B} for these 13 replications in the UPGMA, F/M, Farris, and modified Farris methods were 3.48 ± 0.24, 3.25 ± 0.19, 4.19 ± 0.31, and 3.45 ± 0.26, respectively, whereas the means of S_{EB} were 3.30 ± 0.34, 4.52 ± 0.37, 5.11 ± 0.39, and 4.67 ± 0.32. It is clear that the Farris method shows a rather poor performance in both S_{0B} and S_{EB} values, apparently because of the overestimation of branch lengths that occurs when M is large. When S_{0B} is used as a criterion, the

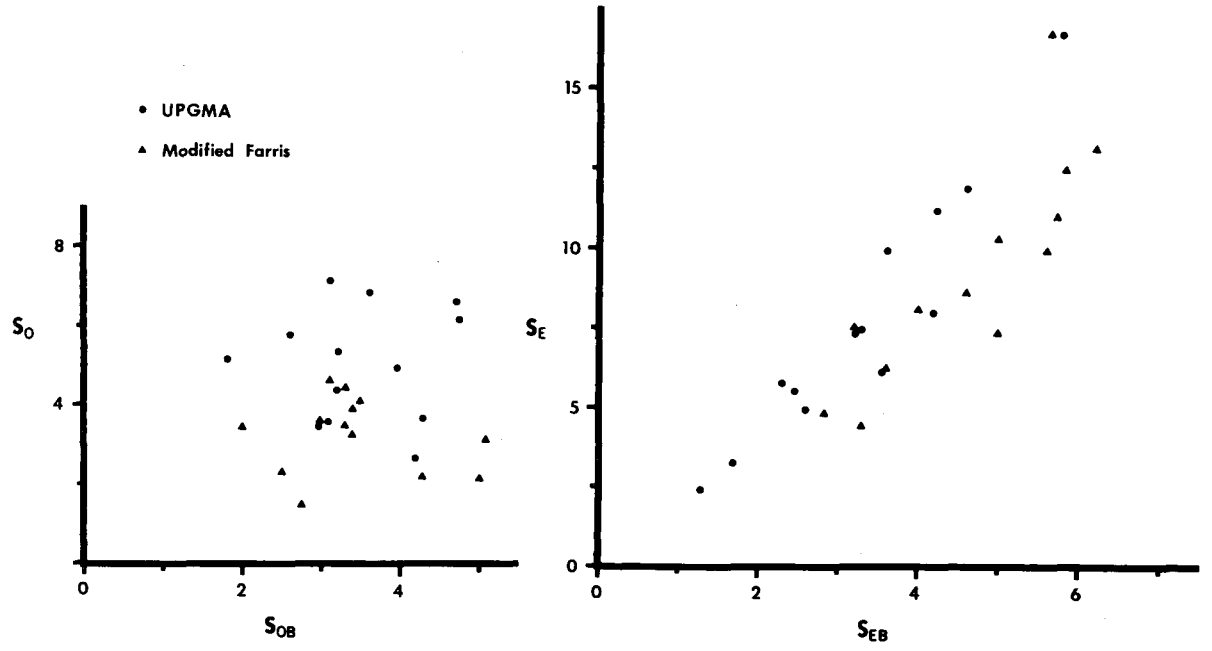

Fig. 6. Correlations between S_0 and S_{0B} and between S_E and S_{EB} in the trees where the topology was correctly reconstructed; M = 8

Table 9. Average deviations of estimated branch lengths from observed distances (S_{0B}) and average deviations of estimated branch lengths from expected branch lengths (S_{EB}) for the reconstructed trees of which the topology is correct

Replication	S_{0B}				S_{EB}			
	UPGMA	F/M	Farris	Modified Farris	UPGMA	F/M	Farris	Modified Farris
				M = 2				
7	0.80	0.78	0.74	0.64	1.76	1.72	1.56	1.56
20	0.93	0.86	1.28	0.85	1.15	1.27	1.80	1.66
Average	0.87	0.82	1.01	0.75	1.46	1.50	1.68	1.61
				M = 4				
1	1.71	1.05	0.95	1.11	1.49	2.42	2.27	2.30
5	1.67	1.80	1.80	1.73	2.48	3.23	3.49	3.23
7	2.58	2.35	2.72	2.32	3.08	3.20	3.80	3.28
12	1.75	1.07	0.79	1.10	1.78	2.78	2.74	2.50
15	2.67	2.99	2.57	2.87	1.63	1.91	2.02	1.86
19	2.24	1.25	1.73	1.32	1.67	2.36	2.25	2.11
20	3.17	2.40	3.40	3.17	2.29	3.30	4.00	3.50
Average	2.25	1.85	1.99	1.94	2.06	2.74	2.94	2.68

remaining three methods show nearly the same performance. However, UPGMA again shows a significantly smaller value of S_{EB} compared with the other methods.

Comparison of the means of S_{0B} and S_{EB} with those of S_0 and S_E indicates that they are roughly correlated, but the correlation does not seem to be high. To see this point in more detail, we examined the correlations of S_0 with S_{0B} and of S_E with S_{EB} for M = 4 and M = 8. In the case of M = 4 there are only seven replications in which S_{0B} and S_{EB} can be compared with S_0 and S_E. The correlation between S_0 and S_{0B} is not significant in any of the four methods, whereas the correlation between S_E and S_{EB} is quite high. For example, the latter correlations for UPGMA and the modified Farris method are 0.90 and 0.74, respectively. A more meaningful study can be done for the case of M = 8, in which 13 replications are available. Figure 6 shows the correlations between S_0 and S_{0B} and between S_E and S_{EB} for UPGMA and the modified Farris method. It is clear that there is virtually no correlation between S_0 and S_{0B}, whereas the correlation between S_E and S_{EB} is significantly high, the correlation coefficients for UPGMA and

the modified Farris method being 0.95 and 0.83, respectively. Essentially the same results were obtained for the other tree-making methods.

The low correlation between S_0 and S_{0B} creates a problem in using S_0 as a criterion of the accuracy of branch lengths. This is particularly unfortunate, because S_0 is the only quantity that can be computed from actual data. Sneath and Sokal (1973) and Farris (1979) suggested that the correlation coefficient between patristic and observed distances be used as a criterion. However, our study has shown that the property of this quantity is virtually the same as that of our S_0. In our simulations the correlation coefficient (r) between the patristic and observed distances (D_{ij} and D_{ij}'' in (7)) was generally very high, the mean for a given M value varying from 0.96 to 0.98 in different cases, but there was little correlation between the values of r and S_{0B}. Therefore, we cannot use the agreement of patristic and observed distances as a reliable measure of the accuracy of a reconstructed tree. Nevertheless, we note that the mean of S_0 for each value of M is correlated with that of S_{0B} (see Tables 8 and 9 and the mean values of S_{0B}). Therefore, S_0 can still be used as a rough measure of the accuracy of branch lengths. In the following we shall use S_0 with this understanding.

The relatively high correlation between S_E and S_{EB} indicates that S_E can be used as a measure of the deviation of estimated branch lengths from expected branch lengths. Although S_E cannot be computed from actual data, it can be used at least for a theoretical study like ours. The computation of S_E is much simpler than that of S_{EB}, so that it facilities theoretical investigation.

In the above computation of S_{0B} and S_{EB} we ignored the trees with erroneous topologies. However, we would expect that even in these trees the same conclusion is obtained if we eliminate all OTUs which are involved in the erroneous part of the topology.

Nucleotide Substitution Following the Negative Binomial Distribution

The values of S_0 and S_E for this case are presented in Table 10. It is noted that these values are somewhat larger than those for the case of Poisson distribution. This is of course expected, since the number of nucleotide substitutions for a given period of evolutionary time has a larger variance in this case than in the case of Poisson distribution. Except this difference, the general

Table 10. Means of the average deviation (S_0) of patristic distances from observed distances and the average deviation (S_E) of patristic distances from expected distances. The actual number of nucleotide substitutions followed the negative binomial distribution. The number of replications is 20

	UPGMA	F/M method	Farris method	Modified Farris method
		S_0		
M = 2	2.91 ± 0.19	1.45 ± 0.19	0.94 ± 0.07	1.01 ± 0.15
M = 4	4.89 ± 0.28	2.91 ± 0.34	2.55 ± 0.26	1.78 ± 0.20
M = 8	6.84 ± 0.43	4.24 ± 0.26	7.00 ± 0.50	3.94 ± 0.38
		S_E		
M = 2	5.25 ± 0.28	5.90 ± 0.29	5.94 ± 0.31	5.57 ± 0.28
M = 4	8.96 ± 0.81	9.99 ± 0.76	10.76 ± 0.90	9.86 ± 0.77
M = 8	11.76 ± 0.73	13.28 ± 0.63	15.60 ± 1.15	13.04 ± 0.67

Table 11. Means of the average deviation (S_0) of patristic distances from observed distances and the average deviation (S_E) of patristic distances from expected distances for the case of varying rates of nucleotide substitution. The number of replications is 20

	UPGMA	F/M method	Farris method	Modified Farris method
		S_0		
M = 2	3.38 ± 0.30	1.94 ± 0.23	1.15 ± 0.13	0.70 ± 0.08
M = 4	6.24 ± 0.45	3.74 ± 0.41	2.42 ± 0.24	1.40 ± 0.10
		S_E		
M = 2	7.11 ± 0.61	7.65 ± 0.61	7.99 ± 0.69	7.67 ± 0.61
M = 4	12.39 ± 1.17	13.38 ± 1.10	14.35 ± 1.17	13.37 ± 1.08

property of S_0 and S_E is the same as that for the case of Poisson distribution.

Nucleotide Substitution with Varying Rates

Table 11 shows the values of S_0 and S_E for the case of varying rates of nucleotide substitution. As expected, both S_0 and S_E are generally greater than those for the cases of Poisson distribution and negative binomial distribution. However, the relative values of these quantities among the four different tree-making methods remain the same as those for the latter cases.

Effect of the Number of OTUs

As mentioned earlier, we conducted a small-scale study of the effect of the number of OTUs on the accuracy of reconstructed trees, using 32 OTUs and M = 4. In this study the F/M method was not included. The mean values of S_0 and S_E obtained are given in Table 12. UPGMA again shows the smallest value of S_E among the three methods examined. Unexpectedly, however, it also shows the smallest value of S_0. Thus, in terms of both S_0 and S_E UPGMA is better than the Farris and modified Farris method. The poorest performance is shown by the Farris method. This is of course expected, because this method gives overestimates of branch lengths in the present case (Table 4). The reason for the good performance of UPGMA seems to be that the procedure of distance-averaging used in this method reduces the effect of stochastic errors substantially when the number of OTUs is large (Nei 1975).

Table 12. Means of the average deviation (S_0) of patristic distances from observed distances and the average deviation (S_E) of patristic distances from expected distances for the case of 32 OTUs and M = 4. The actual number of nucleotide substitutions followed the Poisson distribution. The number of replications is 10

	UPGMA	Farris method	Modified Farris method
S_0	14.0 ± 0.6	52.7 ± 3.0	16.3 ± 1.1
S_E	13.7 ± 1.0	57.3 ± 4.0	20.1 ± 0.7

Correlation of d_T with S_0 and S_E

As noted earlier, S_0 is the only quantity that can be computed from actual data. Thus, it is interesting to see whether this is correlated with any other unobservable quantity or not. We have already seen that the correlation of this quantity with the accuracy of branch lengths is quite low, though it can be used as a rough measure. Table 13 shows the correlations of S_0 with d_T among 20 replications for each value of M for the case of Poisson distribution. It is seen that the correlation is generally low in both rooted and unrooted trees, and particularly in the Farris and modified Farris method there seems to be no real correlation. In UPGMA and the F/M method the correlation tends to be larger, and in two cases it is significant at the one percent level. However, the correlation coefficient is not very high, so that S_0 would not be useful as an indicator of the accuracy of the topology constructed. Table 13 also includes the correlation between S_E and d_T, but the magnitude of the correlation is similar to that of the correlation between S_0 and d_T. The results for the cases of negative binomial distribution and varying substitution rates were nearly the same as those for the case of Poisson distribution.

Table 13. Correlations of d_T with S_0 and S_E for each value of M. The actual number of nucleotide substitutions followed the Poisson distribution. The least significant correlations for the 5% and 1% significant levels are 0.44 and 0.56, respectively

	UPGMA		F/M method		Farris method		Modified Farris method	
	RT	URT	RT	URT	RT	URT	RT	URT
				d_T and S_0				
M = 2	0.61	0.68	0.36	0.35	-0.07	-0.15	-0.20	-0.16
M = 4	0.36	0.20	0.47	0.14	0.26	0.30	0.40	0.48
M = 8	0.44	0.22	0.10	0.10	-0.12	-0.18	0.24	0.20
				d_T and S_E				
M = 2	0.38	0.37	0.17	0.15	0.05	-0.22	0.21	0.02
M = 4	0.36	0.36	0.38	0.30	0.24	0.29	0.21	0.37
M = 8	0.32	0.12	0.19	0.20	0.07	0.05	-0.01	0.02

RT: Rooted tree URT: Unrooted tree

Table 14. Correlations between S_0 and S_E for each value of M. PS and NB stand for the cases of Poisson and negative binomial distributions. The least significant correlations for the 5% and 1% significant levels are 0.44 and 0.56, respectively

	UPGMA		F/M method		Farris method		Modified Farris method	
	PS	NB	PS	NB	PS	NB	PS	NB
M = 2	0.19	0.12	0.07	0.13	0.13	0.22	0.30	0.30
M = 4	-0.08	0.51	0.44	0.81	0.54	0.52	0.33	0.36
M = 8	0.32	-0.03	0.23	-0.08	0.20	0.36	-0.04	-0.04

Correlation between S_E and S_0

As mentioned earlier, S_E cannot be computed from actual data, though it is certainly a good measure of the deviation of estimated branch lengths from true lengths for a species-tree. We have therefore examined the correlation between S_0 and S_E. The results obtained are presented in Table 14. It is clear that the correlation is generally low for both cases of Poisson and negative binomial distributions, so that there is no way to estimate the value of S_E from actual data. A similar result was obtained for the case of varying substitution rates.

Discussion

A number of authors (e.g., Fitch and Margoliash 1967; Prager and Wilson 1978; Farris 1972, 1979) have used quantities similar to our S_0 for knowing the accuracy of the tree reconstructed. Particularly, Prager and Wilson compared the accuracies of the UPGMA, F/M and Farris methods by using the percent standard deviation (PSD) similar to our S_0, and concluded that in general the F/M method is superior to the other two methods. This conclusion is similar to ours for S_0 in the case of Poisson distribution (Table 8). However, this does not mean that the F/M method gives a good tree. Furthermore, our study shows that the modified Farris method is much better than the F/M method even in obtaining a small value of S_0.

It is unfortunate that S_0 is not highly correlated to any of d_T, S_{0B}, and S_E; it can be used only as a very rough criterion of a good tree. In the past a number of authors (e.g., Farris 1979) have argued as though even a small difference in a quantity equivalent to this is very important for evaluating the efficiencies of different tree-making methods. The present study shows that such an argument is trifling. However, this puts us into a difficult situation in judging the goodness of a reconstructed tree, since there are no other observable quantities for real data. How can we judge the superiority of a tree-making method compared with others? One way to circumvent this difficulty is to conduct a simulation study and decide the advantages and disadvantages of each tree-making method statistically, as we did in this paper. Once a method proves to be superior to others, we can use it for all data sets. Of course, this type of study can be done only when we know the evolutionary changes of the characters used. Fortunately, in the case of genetic data we know the approximate pattern of evolutionary changes of genes, so that this can be done. Our computer simulation is certainly dependent on a number of assumptions about the pattern of evolutionary changes of genes, but it is encouraging to see that the three different patterns of evolutionary changes of genes examined gave essentially the same result about the relative merits of tree-making methods.

We have seen that for constructing a topology the Farris and modified Farris methods show on the average a slightly better performance than the other two methods when the coefficient of variation of branch length is large, whereas for estimating the branch lengths of gene-trees, the modified Farris method gives the best result when S_0 is used as the criterion. When S_{0B} is used, the performance of the modified Farris method is as good as that of the F/M and better than that of Farris' method. This suggests that when the coefficient of variation is large the modified Farris method is the best for obtaining a gene-tree. When the coefficient of variation is small, however, this is no longer true and UPGMA seems to be better than the modified Farris method for making a rooted tree (Table 12).

The primary objective of molecular taxonomy or phylogenetics is to construct a species-tree rather than a gene-tree. For this purpose UPGMA shows a good performance in estimating branch lengths. When the coefficient of variation is large, however, this method is not as good as the Farris or modified Farris method for obtaining the correct topology, though the difference in efficiency is not very large except in special cases. If one wants to avoid this problem, he can use the Farris or modified Farris method to construct a topology and then use the distance-averaging method similar to that for UPGMA for the given topology to estimate the branch lengths. The species-tree thus obtained will be useful for estimating the times at which various species or species groups split in the evolutionary process. When the coefficient of variation of branch length is small, however, UPGMA seems to be as good as the Farris and modified Farris methods even in obtaining a good topology. Furthermore, in this case UPGMA gives smaller values of both S_0 and S_E than the Farris and modified

Farris methods. It is interesting to see that this simple method, which was originally proposed for phenetic classification, shows the best performance when the coefficient of variation is small.

Recently, Li (1981) modified UPGMA to take into account the effects of stochastic errors or unequal rates of gene substitution. His method is intended to construct a gene-tree rather than a species-tree. He claimed that when nucleotide substitution data are used his method is substantially better than UPGMA. However, his conclusion is based solely on Tateno's (1978) distortion index ($\overline{DI}$), neglecting S_0 and S_E, and the $\overline{DI}$ value for the trees obtained by Li's method is nearly the same as that of our modified Farris method (Tateno 1978). Therefore, a more careful study should be made about the efficiency of his method.

Our study has shown that the topology of a reconstructed tree is often wrong, whatever the tree-making method is used, unless all branch lengths (number of nucleotide substitutions) of a true tree are sufficiently long. Furthermore, even if the topology obtained is correct, the estimates of branch lengths are not necessarily close to the true values. This result is disturbing, but we must accept it, since it is due to the stochastic nature of gene substitution as well as to the backward and parallel mutations that cannot generally be detected. Probably the only way to reduce the errors involved in an estimated tree is to increase the number of genes used. At any rate, this study gives the warning that one cannot be overconfident about the tree reconstructed, whichever the method is used.

This result might prompt some numerical taxonomists to believe that morphological characters are better than molecular data for constructing phylogenetic trees. We disagree. The evolutionary change of morphological characters is generally much more complex than DNA or proteins. Even such a simple character as fingerprint pattern in man and apes seems to be subject to a complicated genetic change at the phenotypic level (Chakraborty and Nei, unpublished). It seems that at the phenotypic level "backward" and "parallel" mutations are much more common than those at the nucleotide or amino acid level. If this is true, it would be very difficult to reconstruct a correct phylogenetic tree from morphological characters except in special cases.

A number of authors (e.g., Farris 1972; Prager and Wilson 1978) have expressed concern about negative values of the estimates of branch lengths that often occur in a reconstructed tree. In some tree-making methods (e.g., UPGMA) no negative branches are obtained. However, this does not mean that the tree reconstructed is close to the true form, as is clear from Fig. 4b. Actually, branch lengths are statistically estimated in such methods as the modified Farris method, so that the estimates obtained can be negative with a certain probability particularly when the true distances are small. Therefore, we do not have to worry very much about the negative estimates unless they are significantly different from 0. In our simulation study we have seen a number of cases in which the topology was correctly obtained but some of the estimated branch lengths were negative. Of course, when the absolute value of a negative branch was extremely large, the topology was generally incorrect.

In theoretical studies of tree-making methods a distance matrix is often computed by summing up the relevant branch lengths of a hypothetical tree like Fig. 4a, and the estimates of branch lengths of a reconstructed tree are compared with the true branch lengths. If this procedure is used, the Farris method (and modified Farris method) often gives a better result than others (Swofford 1981). Indeed, if we apply this procedure to the trees in Figs. 4a and 5a, we obtain the same result. Unfortunately, however, molecular data do not really represent the exact number of gene (nucleotide) substitutions for most pairs of OTUs, so that the conclusion obtained from this type of study is not reliable. This is obvious from the examples in Figs. 4 and 5.

One important factor in the comparison of the efficiency of different methods of tree-making is the time required for constructing a tree. Prager and Wilson (1978) considered this problem and stated that the time required for constructing a tree for a sizable number of OTUs is much longer for the Farris method than for UPGMA and the F/M method. They apparently studied the time required for manual computation. In our study we used a computer for all the four methods. The computer program for the F/M method was kindly provided by Walter Fitch, whereas the programs for the other methods were developed by the senior author. Our experience indicates that the F/M method is most time-consuming, whereas the remaining three methods require nearly the same amount of computer time, which is far less than that for the F/M method.

In this paper we used nucleotide sequence data for studying the accuracies and efficiencies of different tree-making methods. In practice, other types of data such as amino acid sequences and immunological distances are often used for molecular taxonomy. The genetic distances based on these data are known to be roughly proportional to the evolutionary time. Therefore, our conclusion obtained from the study of nucleotide substitution will directly apply to the trees constructed from these data, though the errors associated with the distance estimates are possibly larger in this case than those for nucleotide substitution.

Farris et al. (1979) recently criticized the use of immunological distances for the reason that it is not a metric and violates the principle of triangle inequality. They even denied Prager and Wilson's (1971) experimental result that their immunological distance is approximately proportional to the number of amino acid differences. They stated: "Amino acid sequence

differences must be metric. If immunological distances are strongly nonmetric, then their correlation with sequence differences cannot be very close." This statement cannot be true. As mentioned earlier, our estimate of nucleotide substitutions, δ, is not a metric, but, as will be seen from Table 7, it is highly correlated to the actual number of nucleotide substitutions. The important property required for a distance measure in tree-making is its linearity with evolutionary time. Immunologcial distance has been shown to increase approximately linearly with evolutionary time, and this property is very valuable. If we start to blame data because they are not manageable for a certain statistical method, there will be no progress in statistics. Statistics has been invented to extract a maximum information from a given set of data, and not *vice versa.*

Acknowledgement. We thank David Swofford, Walter Fitch, and anonymous reviewers for their comments on an earlier draft of this manuscript. This study was supported by research grants from the National Science Foundation and the National Institutes of Health.

References

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Am J Hum Gen 19:233–257

Chakraborty R (1977) Estimation of time of divergence from phylogenetic studies. Can J Genet Cytol 19:217–223

Dayhoff MO (ed) (1969) Atlas of protein sequence and structure, Vol. 4. Natl Biomed Res Found, Silver Spring, MD

Doolittle RF, Blombäck B (1964) Amino-acid sequence investigations of fibrinopeptides from various mammals: evolutionary implications. Nature 202:147–152

Edwards AWF, Cavalli-Sforza LL (1965) A method for cluster analysis. Biometrics 21:362–375

Farris JS (1972) Estimating phylogenetic trees from distance matrices. Am Nat 106:645–668

Farris JS (1979) On the naturalness of phylogenetic classification. Syst Zool 28:200–214

Farris JS, Kluge AG, Mickevich MF (1979) Paraphyly of the *Rana boylii* species group. Syst Zool 28:627–634

Fitch WM, Margoliash E (1967) Construction of phylogenetic trees. Science 155:279–284

Goodman M, Moore GW, Barnabas J, Matsuda G (1974) The phylogeny of human globin genes investigated by the maximum parsimony method. J Mol Evol 3:1–48

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

Kimura M (1969) The rate of molecular evolution considered from the standpoint of population genetics. Proc Natl Acad Sci USA 63:1181–1188

King, JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788–798

Langley CH, Fitch WM (1974) An examination of the constancy of the rate of molecular evolution. J Mol Evol 3:161–177

Li W (1981) Simple method for constructing phylogenetic trees from distance matrices. Proc Natl Acad Sci USA 78:1085–1089

Margoliash E, Smith EL (1965) Structural and functional aspects of cytochrome c in relation to evolution. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 221–242

Moore GW, Barnabas J, Goodman M (1973a) A method for constructing maximum parsimony ancestral amino acid sequences on a given network. J Theor Biol 38:459–485

Moore GW, Goodman M, Barnabas J (1973b) An iterative approach from the standpoint of the additive hypothesis to the dendrogram problem posed by molecular data sets. J Theor Biol 38:423–457

Nei M (1972) Genetic distance between populations. Am Nat 106:283–292

Nei M (1975) Molecular population genetics and evolution. North Holland, Amsterdam and New York

Nei M (1977) Standard error of immunological dating of evolutionary time. J Mol Evol 9:203–211

Nei M (1978) Genetic distance and molecular taxonomy. Abstract in: Proc XIV Intl Cong Genet. Nauka Publishing Office, Moscow, pp 84–85

Nei M, Tateno Y (1978) Nonrandom amino acid substitution and estimation of the number of nucleotide substitutions in evolution. J Mol Evol 11:333–347

Ohta T (1976) Simulation studies on the evolution of amino acid sequences. J Mol Evol 8:1–12

Ohta T, Kimura M (1971) On the constancy of the evolutionary rate of cistrons. J Mol Evol 1:18–25

Peacock D, Boulter D (1975) Use of amino acid sequence data in phylogeny and evaluation of methods using computer simulation. J Mol Biol 95:513–527

Prager EM, Wilson AC (1971) The dependence of immunological cross-reactivity upon sequence resemblance among lysozymes. J Biol Chem 246:5978–5989

Prager EM, Wilson AC (1978) Construction of phylogenetic trees for proteins and nucleic acids: empirical evaluation of alternative matrix methods. J Mol Evol 11:129–142

Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53:131–147

Sarich VM, Wilson AC (1966) Quantitative immunochemistry and the evolution of primate albumins: micro-complement fixation. Science 154:1563–1566

Sneath PHA, Sokal RR (1973) Numerical taxonomy. WH Freeman, San Francisco

Sokal RR, Michener CD (1958) A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull 28:1409–1438

Swofford DL (1981) On the utility of the distance Wagner procedure. In: Funk VA, Brooks DR (eds) Advances in cladistics. Cladistics Publications, Bronx, New York

Tateno Y (1978) Statistical studies on the evolutionary changes of macromolecules. Ph.D. Dissertation, University of Texas at Houston

Tateno Y, Nei M (1978) Goodman et al.'s method for augmenting the number of nucleotide substitutions. J Mol Evol 11:67–73

Waterman MS, Smith TF (1978) On the similarity of dendrograms. J Theor Biol 73:789–800

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Ann Rev Biochem 46:573–639

Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genetic heterogeneity. In: Kasha M, Pullman B (eds) Horizons in biochemistry. Academic Press, New York, pp 189–225

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166

Received May 28, 1981 / Revised April 28, 1982.