J Mol Evol (1986) 23:354–361

Journal of
Molecular Evolution

Statistical Properties of Molecular Tree Construction Methods Under the Neutral Mutation Model

Yoshio Tateno[1] and Fumio Tajima[2]

[1] Institute of Physical and Chemical Research (RIKEN), Wako, Saitama 351-01, Japan
[2] Department of Biology, Kyushu University, Hakozaki, Fukuoka 812, Japan

Summary. The statistical properties of three molecular tree construction methods—the unweighted pair-group arithmetic average clustering (UPG), Farris, and modified Farris methods—are examined under the neutral mutation model of evolution. The methods are compared for accuracy in construction of the topology and estimation of the branch lengths, using statistics of these two aspects. The distribution of the statistic concerning topological construction is shown to be as important as its mean and variance for the comparison.

Of the three methods, the UPG method constructs the tree topology with the least variation. The modified Farris method, however, gives the best performance when the two aspects are considered simultaneously. It is also shown that a topology based on two genes is much more accurate than that based on one gene.

There is a tendency to accept published molecular trees, but uncritical acceptance may lead one to spurious conclusions. It should always be kept in mind that a tree is a statistical result that is affected strongly by the stochastic error of nucleotide substitution and the error intrinsic to the tree construction method itself.

Key words: Molecular evolution — Molecular tree — Computer simulation — Unweighted pair-group arithmetic average clustering method — Farris method — Modified Farris method

Introduction

With the rapid accumulation of data on the nucleic acid and amino acid sequences of various genes for a number of species, it has become common practice to construct phylogenetic trees (molecular phylogenetic or molecular trees) based on such molecular data. It may be said that phylogenetic tree construction has entered a second phase, one succeeding the period when morphological and physiological characters were major bases of trees. This transition has made it possible to discuss the evolution of organisms more concretely and to estimate divergence times more objectively even for such organisms as microbes, fossil records of which are seldom obtained (e.g., Hori and Osawa 1979; Dekio et al. 1984). The progress has also enabled insight to be gained into the origin and evolution of organelles (Schwartz and Dayhoff 1978; Küntzel and Köchel 1981) and genes (e.g., Dayhoff 1972; Miyata et al. 1980; Gojobori and Nei 1984; Daniels and Deininger 1985).

This process has elucidated a number of aspects in evolution, but it has also raised new controversies. An example of such a controversy is the issue of primate evolution. From DNA hybridization data on several primate species, Sibley and Ahlquist (1984) concluded that human and chimpanzee split 6–8 million years ago, 2–4 million years after the divergence between the gorilla and the ancestor of those two primates. Templeton (1983) challenged their view, basing his argument on the same data but on a different method. He claimed that there was no statistically significant difference between Sibley and Ahlquist's scheme and an alternative one in which chimpanzee is closer to gorilla than to human. Obviously, much depends on the method of tree construction used. Hasegawa and Yano (1984) reached a similar conclusion to Sibley and Ahlquist's (1984) on the basis of the nucleotide sequences of mitochondrial DNA. Thus, discrepancies are attributable not only to experimental error but also to stochastic error of gene (nucleotide) substitution and to error intrinsic to a molecular tree construction method, or methodological error.

Offprint requests to: Y. Tateno

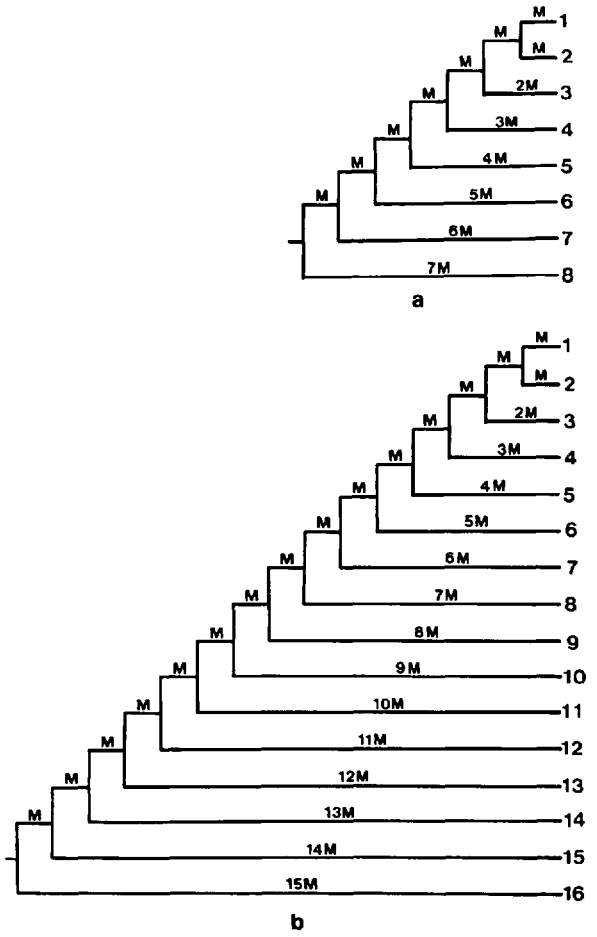

Fig. 1a, b. Model trees used in the simulation experiment: **a** model tree with 8 OTUs; **b** model tree with 16 OTUs. Multiples of M along branches show branch length, where M is the expected number of nucleotide substitutions per gene

While both experimental and stochastic errors can be reduced by the sensible use of more and more data, methodological error is not easily curtailed because it is inherent to the method itself. It is thus meaningful to study statistical properties of methods. Nei et al. (1985) studied the statistical properties of branch points in a tree constructed by the unweighted pair-group arithmetic average clustering (UPG) method (Sokal and Sneath 1963; Nei 1975), and derived a statistical method by which the significance of two branch points can be tested. Although their method is useful for the discussion of the statistical significance of branching in a constructed tree, its extension to other construction methods seems difficult. Moreover, their method does not allow us to treat the methodological and stochastic errors separately.

In this report we shall discuss the statistical properties of three molecular tree construction methods—the UPG, Farris (1972), and modified Farris (Tateno et al. 1983) methods—on the basis of results obtained by computer simulation. The simulation mimics the evolutionary change of a neutral gene (Kimura 1968, 1983) along model molecular trees. Emphasis will be placed particularly on the effect of stochastic and methodological errors on the performances of the three methods.

Model and Method of Computer Simulation

The model and method used in the present study are essentially the same as those of Tateno et al. (1983). An ancestral gene of 100 codons was created in a computer in such a way that the base at each nucleotide site was randomly determined and nonsense codons were avoided. Mutational events on the gene were assumed to be such that (1) they followed the Poisson process, (2) the probability of occurrence of a mutation was the same among the 300 nucleotide sites, and (3) the base at a site chosen to be changed was replaced by each of the remaining three bases with an equal probability of ⅓. Since we are dealing with neutral mutations occurring in the gene, the above assumptions appear reasonable.

It was important in the study to keep every aspect of the evolutionary process at hand. It was thus essential to introduce a model tree whose topology and branch lengths were clearly determined. When one is to study the effects of stochastic and methodological errors on molecular tree construction, it is desirable that the model tree possess branches of various lengths. Thus, two types of model trees were set up, as shown in Fig. 1. In these trees the variability of branch lengths is maximized for a given number of operational taxonomic units (OTUs; Sokal and Sneath 1963). "M" in this figure is the expected number of nucleotide substitutions per gene or the Poisson parameter of the mutational process. The ancestral gene evolves along each of the model trees and diverges into the descendant genes at its terminals.

Since the three methods to be examined each require a distance matrix for tree construction, the genes thus produced were compared pairwise and a matrix of nucleotide differences was computed. This matrix was corrected for multiple substitutions using the Jukes and Cantor (1969) formula. The corrected distance matrix was then fed as the input data to a computer algorithm for each of the three methods, and a molecular tree was constructed. When the constructed tree was compared with the model tree, it was possible to examine the performance of the construction method at this stage. One replicaton of the simulation is completed at this step.

The examination was carried out with respect to two functions: construction of the true topology and estimation of branch lengths. These two functions are not necessarily correlated with each other (Tateno et al. 1983; Tateno 1985). For topological examination a distortion index (d_T) (Robinson and Foulds 1981) was used that measures quantitatively the topological difference between the model and constructed trees [see Tateno (1985) for details]. As the topological difference becomes larger, d_T increases in increments of 2 over the range from 0 to $2(n - 1)$, where n is the number of OTUs in the tree.

For examination of the branch-length estimation, two indices were employed: One (S_E) is the square root of the average squared deviation of the estimated branch lengths from the branch lengths of the model tree, and the other (S_O) is the square root of the average squared deviation of the estimated branch lengths from the branch lengths in the distance matrix. As mentioned by Tateno et al. (1983) there are two types of molecular trees. One is the species tree, which is supposed to depict the divergence among taxa of organisms, and the other is the gene tree, which is intended to present the evolution of genes themselves. The former naturally possesses the property that the lengths of two branches originating from a common ancestor are identical, whereas the

latter does not necessarily do so. S_E could be interpreted as a measure of the deviation of the branch lengths in the constructed tree from those in the species tree and S_O as measure of the deviation of the constructed tree's branch lengths from those in the gene tree.

Topological Construction

It should be mentioned, first of all, that the UPG method constructs a tree with a root, the common ancestor of all OTUs involved, whereas the Farris and modified Farris methods cannot place a root in a constructed tree. Farris (1972) suggested as a rule independent of his method that the root of a tree constructed by his method (and the modified Farris method) be placed at the midpoint between the two OTUs separated by the largest distance. Although Farris's rule was shown to be inappropriate by Tateno et al. (1983), both rooted and unrooted trees were incorporated in the present study. One more point to be noted here is the meaning of the stochastic error: Since no gene-sampling procedure is involved in the present simulation, it means solely the intrinsic error of the rate of nucleotide substitution in the present context (see Tajima 1983).

The computer simulation was carried out first on model tree a in Fig. 1. When M is small, say 1, the three methods do not show significant differences in topological construction (Tateno 1985). This can be considered to be the case where the stochastic error of nucleotide substitution is so large that it overrides the methodological error. The M value was thus set at 2. Miyata et al. (1980) estimated the average evolutionary rate of synonymous substitution for various genes to be 5.1×10^{-9} per site per year. If this estimate is regarded as the rate of neutral mutation, then the model tree with M = 2 gives a divergence time between two OTUs ranging from 2.6×10^7 to 1.8×10^8 years. The number of replications was 500.

Single-Gene Case

The results are shown as histograms of d_T for the three methods in Fig. 2. Figure 2a shows the result for the rooted tree. It is seen from the three distributions that the variance of the UPG method is smaller than that of either of the other two, whereas the modes are not different from one another. This does not necessarily mean that the UPG method is superior to the two methods. Actually, the *t*-test shows no significant difference in topological construction between the UPG and Farris methods. It should, however, be noted that the frequency class of $d_T = 0$ for the Farris and modified Farris methods is about two times as large as that of the UPG method. This indicates that not only the mean and variance of d_T but also the shape of the distribution is important in the examination of the topological construction. Figure 2a also shows that the distributions of the Farris and modified Farris methods have rather long tails toward the right, revealing that the two methods may construct trees drastically different from the model tree, though not often. This is expected from the nature of the two methods. For an OTU to be connected, one must compute the likelihood when the OTU is joined to each branch of the tree so far constructed. The OTU is connected to the branch at which the likelihood is maximum. If the likelihood happens to be largest at an incorrect branch, owing to the stochastic error of nucleotide substitution, then the OTU is connected there, resulting in a tree quite different from the correct one. This is the case where the stochastic error creates or exaggerates the methodological error. The UPG method does not work in this way: Instead, the average distance from the OTU to the tree is taken into account when a new OTU is added to the tree under construction.

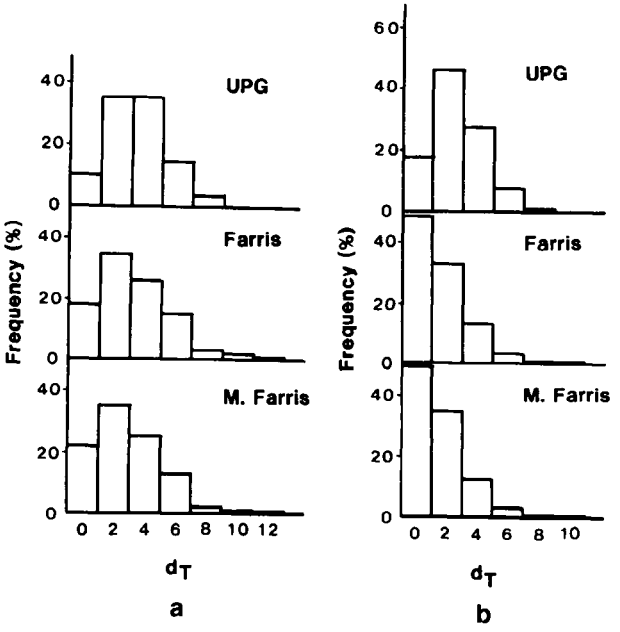

Fig. 2a, b. Distribution of d_T for 8 OTUs: **a** for the rooted tree; **b** for the unrooted tree. M = 2; the number of replications is 500. UPG, unweighted pair-group arithmetic average clustering method; Farris, Farris method; M. Farris, modified Farris method

Figure 2b shows the results for the unrooted tree. As mentioned above, Farris's rule for placing the root does not work well. This is due to a discrepancy of logic between his method and his rule. His method ignores the constancy of the nucleotide substitution rate, whereas his rule incorporates it. The discrepancy causes the difference in distribution between Figs. 2a and 2b. As shown in Fig. 2b, the distributions for the Farris and modified Farris methods become L-shaped for the unrooted tree, indicating a remarkable improvement in compari-

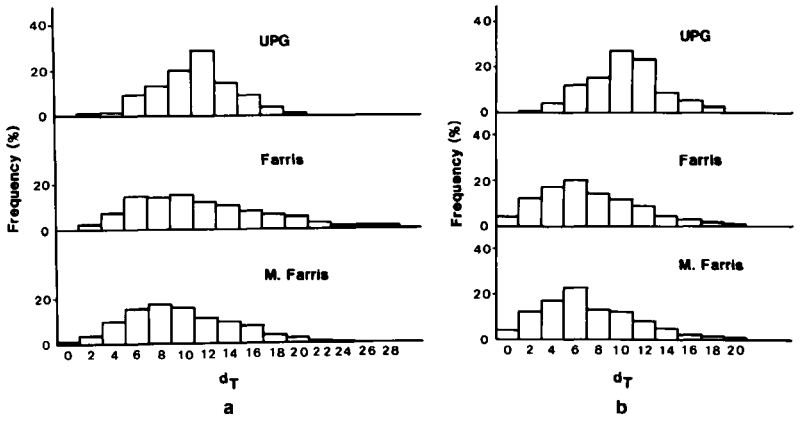

Fig. 3a, b. Distribution of d_T for 16 OTUs: **a** for the rooted tree; **b** for the unrooted tree. M = 2; the number of replications is 500. Abbreviations as in Fig. 2

son with Fig. 2a, in which the corresponding distributions are bell shaped. It is evident that an L-shaped distribution is desirable for d_T, and the two Farris methods show this desirable property if Farris's rule is neglected. This claim should be slightly moderated, however, because considerable improvement is observed also in the UPG method after removal of the root, at least for the case with 8 OTUs. There is a problem again in placing the root in the UPG method, though not as serious as in the previous case. Leaving the matter of rooting aside, the UPG method still gives a bell-shaped distribution, which makes it inferior to the other two methods as regards this point. The notable difference between Figs. 2a and 2b strongly suggests that some improved means of placing the root should be devised, at least for the Farris and modified Farris methods. One way might be to determine the root by the UPG method. Although this device still depends on the constancy of the nucleotide substitution rate, it is expected to be better than Farris's rule, since the root is determined by taking into account all the branch lengths in the tree.

To study the effect of the number of OTUs on tree construction, another set of computer simulations was carried out, on model tree b in Fig. 1. This time the number of OTUs was doubled, but the M value and the number of replications were unchanged. In model tree b the divergence time between OTUs ranges from 2.6×10^7 to 3.9×10^8 years. The result for the rooted tree is presented in Fig. 3a and that for the unrooted tree in Fig. 3b. In this case the distributions in both parts of the figure are bell shaped, and both the mean and variance of d_T are increased over the case with fewer OTUs. These unfavorable outcomes arise mainly from the fact that there are twice as many branches with M = 2 as there are in model tree a. Since the coefficient of variation of branch length is largest for branches with M = 2 in the model tree, the stochastic error of nucleotide substitution disturbs most seriously the topological construction of such branches. It has been shown that the coefficient of variation of the number of nucleotide substitutions affects the topological construction more strongly than does the mean of the number of nucleotide substitutions (Tateno et al. 1983). Another factor that should be considered is the relative consequence of error in the construction of these branches. In model tree a the error does not lead to d_T values larger than 14, but d_T can be as large as 30 in model tree b.

The UPG method again gives the smallest variance among the three methods for both rooted and unrooted trees. It can generally be said that the effect of averaging distances in the UPG method is to reduce the variance in the topological construction. Nevertheless, this effect does not extend to reducing d_T. As judged by the *t*-test, there is no significant difference in mean d_T value between any two of the methods. The modes of the distributions for the Farris and modified Farris methods are, in this case, shifted one or two classes toward the class of $d_T = 0$ compared with the UPG method. It is also seen that the distribution for the modified Farris method extends to the class of $d_T = 0$. In contrast, the distribution of the Farris method has a long tail toward the right, contributing to its large variance. This undesirable property of the Farris method now becomes conspicuous, because when the number of OTUs or the M value is large, the method has a tendency to overestimate branch length under the influence of the stochastic error of nucleotide substitution (see below). Although the correlations between d_T and S_E and between d_T and S_O are not very high (Tateno et al. 1983; Tateno 1985), it is still expected that the overestimation will disturb the topological construction to some extent.

Effect of the Number of Genes

It has been shown above that the stochastic error of nucleotide substitution has a considerable influence on the topological construction of a molecular tree. This implies that a tree could be wrongly constructed owing to this error even under the situation of constant evolutionary rate. The stochastic error

can be reduced by incorporating more than one gene into the computation of a distance matrix. Thus, to see the effect of reduction of the stochastic error, we recomputed the distance matrix taking the average of the distances of 2, 5, and 10 genes. Since the same data as in the one-gene case were used to do this, the numbers of replications were accordingly reduced to 250, 100, and 50, respectively. Trees were then reconstructed using the matrices thus obtained. The results for the case of 16 OTUs are presented in Fig. 4, which plots the mean d_T values against the number of genes for 1, 2, 5, and 10 genes.

As seen in the figure, the mean d_T decreases as the number of genes increases for both rooted and unrooted trees. This is, of course, due to the fact that the variance of the number of nucleotide substitutions becomes smaller as the number of genes increases. It is easily shown that in the present study the stochastic error measured by the variance reduces to 1/m when the number of genes increases to m. The figure shows that the mean d_T decreases roughly proportionally to the inverse of the number of genes. In particular, a steep decline is observed when the number of genes changes from 1 to 2. This indicates that if one can use 2 genes, one will obtain a much better topology than when using just 1 gene.

It is of interest to compare the effectiveness of increasing the number of genes in improving topological construction among the three methods. For this purpose, the measure $G_{1/2}$ was introduced, which is defined as the number of genes for which the mean d_T takes half the value it has for 1 gene. Relatively speaking, the method with the smallest value of $G_{1/2}$ has the largest power of reducing the stochastic error for an increasing number of genes. The $G_{1/2}$ values obtained from Fig. 4a (rooted tree) are 3.0, 2.5, and 2.2 for the UPG, Farris, and modified Farris method, respectively. Those for the unrooted tree are 3.0, 1.7, and 1.8. These results imply that increasing the number of genes improves the topological construction more for the Farris and modified Farris methods than for the UPG method. The difference could be interpreted as reflecting the difference in methodological error between the two Farris methods and the UPG method. That is, the methodological error is larger in the UPG method than in the other two methods when the stochastic error is reduced evenly for the three methods. A similar observation and interpretation hold for the effect of the number of genes in the 8-OTU case (data not shown).

Comparison of Figs. 4a and 4b reveals that the mean d_T of the UPG method does not change very much after removal of the root. This is different from what happens in the 8-OTU case mentioned above. There are at least two reasons for this: One is that the placement of the root has a more serious effect on the final topology in the 8-OTU case than

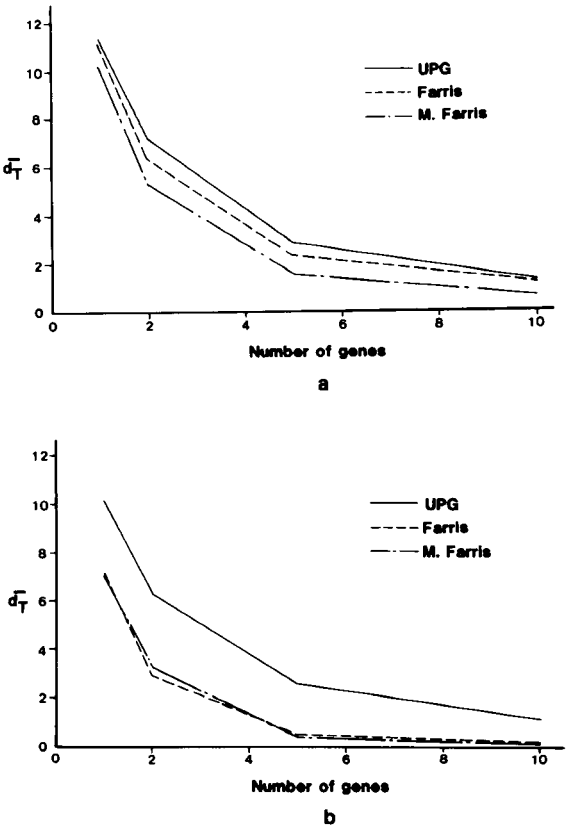

Fig. 4a, b. Change in average d_T value with increasing number of genes for 16 OTUs. The d_T values are plotted against the numbers of genes (1, 2, 5, and 10 genes were considered) for the rooted tree (**a**) and the unrooted tree (**b**). Abbreviations as in Fig. 2

it does in the 16-OTU case, because the number of branch points, including the root, is half as many in the former case as in the latter. The other is that as the number of OTUs increases, the effect of averaging distances becomes more manifest, resulting in a higher probability that the root will be placed at the correct position. That is, in the 16-OTU case, the error in the topological construction occurs mainly in procedures other than the placement of the root.

Estimation of Branch Lengths

Deviation from the Expected Distances

The mean S_E values and their standard deviations for the three methods are presented in the upper half of Table 1. The values are given for the cases of 8 and 16 OTUs and for 1, 2, 5, and 10 genes. When the number of OTUs is 8, the UPG and modified Farris methods give results quite similar to each other for the 1-, 2-, 5-, and 10-gene cases. The result for the Farris method is slightly larger in both mean and standard deviation of S_E than that for either the UPG or modified Farris method, though the difference is not significant as judged by the

Table 1. Mean S_E and S_O values for the three methods

No. of genes	Method		
	UPG	Farris	Modified Farris
		S_E	
		No. of OTUs = 8	
1	4.47 (1.35)	4.74 (1.51)	4.46 (1.35)
2	3.25 (1.00)	3.38 (1.13)	3.24 (1.00)
5	2.00 (0.61)	2.06 (0.67)	1.99 (0.62)
10	1.44 (0.41)	1.46 (0.43)	1.43 (0.41)
		No. of OTUs = 16	
1	6.83 (1.51)	8.90 (2.43)	6.77 (1.53)
2	4.85 (1.00)	6.44 (1.89)	4.80 (1.00)
5	3.06 (0.63)	4.07 (1.32)	3.01 (0.65)
10	2.18 (0.45)	2.87 (0.88)	2.14 (0.46)
		S_O	
		No. of OTUs = 8	
1	2.16 (0.67)	1.16 (0.58)	0.66 (0.26)
2	1.56 (0.48)	0.92 (0.38)	0.49 (0.16)
5	1.01 (0.30)	0.60 (0.23)	0.30 (0.07)
10	0.74 (0.22)	0.45 (0.16)	0.22 (0.05)
		No. of OTUs = 16	
1	3.83 (0.78)	5.29 (1.64)	1.97 (0.34)
2	2.74 (0.53)	4.23 (1.25)	1.39 (0.21)
5	1.73 (0.33)	2.88 (0.80)	0.86 (0.11)
10	1.24 (0.27)	2.09 (0.55)	0.60 (0.07)

Values in parentheses are standard deviations. M = 2; the number of replications was 500. UPG, unweighted pair-group arithmetic average clustering

t-test. In the case with 16 OTUs, however, the values for the Farris method become significantly larger than those for the UPG and modified Farris methods. As mentioned above, the reason for this is that the Farris method has a tendency to overestimate branch length when the number of OTUs is large. To illustrate this tendency, we consider the sample tree in Fig. 5.

Let us suppose that OTU I is to be connected to the tree in this figure, where A, B, and C are OTUs, and X and Y are branch points. Then the distance between I and X, D(I, X), must be estimated in the Farris method. To do this, Farris (1972) uses the following triangle inequality:

$$D(I, q) \leq D(q, X) + D(I, X) \qquad (1)$$

where q is A, B, or C. IF D*(I, q) is the greatest lower bound of D(I, q) and P(q, X) is the least upper bound of D(q, X), then the greatest lower bound of D(I, X), D*(I, X), is given by

$$D^*(I, X) \geq \sup[D^*(I, q) - P(q, X)] \qquad (2)$$

where sup refers to the limes superior. Farris applies formula (2) in such a way that

$$\begin{aligned} D^*(I, X) = \max[&(D(I, A) - D'(A, X)), \\ &(D(I, B) - D'(B, X)), \\ &(D(I, C) - D'(C, X))] \end{aligned} \qquad (3)$$

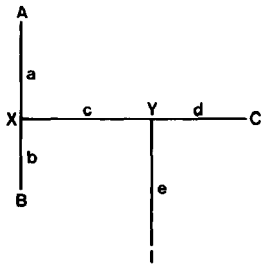

Fig. 5. A tree shown as a network. A, B, and C are OTUs, and X and Y are branch points. OTU I is being connected to the tree. The lower-case letters along the branches represent the respective branch lengths. See text for discussion

where D(*, *) on the right is obtained directly from the distance matrix and D'(*, *) is estimated in advance in his method. Farris states that the motivation for obtaining D*(I, X) is to assess the numbers of multiple, inverse, and parallel mutations occurring in the lineage between I and X, which cannot be obtained from the distance matrix. For example, the number between OTUs A and B is given by D*(A, B) − D(A, B), which he calls the "homoplasy."

There are confusing and fallacious aspects to the aforementioned argument of Farris. First of all he confuses D*(*, *) with D(*, *), and P(*, *) with D'(*, *), in his application. This confusion occurs because he cannot obtain D*(*, *) and P(*, *) as they are. Actually, his definition of these distances is not clear at all. Leaving his definition aside, if one of the quantities D(I, A), D(I, B), or D(I, C) in formula (3) becomes considerably larger than its expected value owing to the stochastic error, then D(I, X) is overestimated even if D'(A, X), D'(B, X), and D'(C, X) are estimated properly. When the number of OTUs is small, such overestimation is not so serious, as seen in the 8-OTU case. If the number becomes larger, however, the cumulative error is no longer negligible and leads to a gross overestimation of branch length in the final result. That is what Table 1 shows for the 16-OTU case. Farris's motivation for obtaining D*(*, *) is also fallacious. Let a, b, c, d, and e be the branch lengths defined in Fig. 5. Then formula (3) becomes

$$D^*(I, X) = \max(c + e, c + e, e - c) = c + e \qquad (4)$$

That is, taking the maximum value is absolutely necessary to obtain the correct distance, c + e; otherwise a wrong distance, e − c, might be chosen.

In the modified Farris method D(I, X) is estimated as the average of D(I, A) − D(A, X) and D(I, B) − D(B, X), and D(I, C) is never involved (Tateno et al. 1983). The modified Farris method is thus superior to the Farris method in logic and in reducing the stochastic error of nucleotide substitution. In this sense, the former method shares a property with the UPG method. This could be the reason for the observation in Table 1 that the modified

Farris method gives results for S_e similar to the UPG method's. Note also that in the modified Farris method no fallacious argument such as homoplasy is involved.

Deviation from the Observed Distances

The S_O results for the three methods are shown in the lower half of Table 1. It is clear that the modified Farris method gives S_O values significantly smaller than do the UPG and Farris methods for both the 8- and 16-OTU cases. In the 8-OTU case, the UPG method gives larger values than the Farris method for all numbers of genes considered. The reason for this is inherent in the former method. In the UPG method the estimated branch lengths of two lineages sharing the same amount of evolutionary time are assumed to be identical, since this method is intended to construct a species tree. The Farris and modified Farris methods do not contain such an assumption, and the estimated branch lengths are almost always different from each other. Thus, since the number of nucleotide substitutions in two lineages can be different owing to the stochastic error in real evolution (and in the present simulation), the Farris and modified Farris methods are expected to give smaller S_O values than the UPG method. That is, the aforementioned assumption contributes to the large methodological error of the UPG method as far as S_O is concerned.

In the 16-OTU case, however, the difference between the UPG and Farris methods is reversed. The reverse relationship arises from the drawback of the Farris method that it tends to overestimate branch length when the number of OTUs is large, as mentioned above. The drawback is so serious as to diminish the advantage over the UPG method that was observed in the 8-OTU case. One should thus be cautious when using the Farris method for a large number of OTUs. The caution is made with regard not only to the estimation of branch length but also to the topological construction.

Discussion

One tends to regard a molecular tree, once it is depicted, as the true tree without reflecting much on its objectivity. Although this is often unavoidable due to lack of independent evidence, it may lead one to spurious conclusions. As shown above, no tree construction method is perfect, and all methods often give a wrong tree even under such a simple evolutionary process as one driven by neutral mutation. It is thus advisable to be careful when relying on a molecular tree constructed by any of the three methods discussed above. This advice is, of course, applicable to other tree construction methods, because no method is free from disturbance by the stochastic error of nucleotide substitution. As long as the number of OTUs is less than or equal to 16, the modified Farris method gives a better tree than either the UPG or Farris method in terms of both topological construction and branch length estimation. The modified Farris method is the most resistant of the three methods to disturbance by the stochastic error. Note also that the performance of the modified Farris method is better than that of either of the other methods when the number of genes increases.

Although the three methods are not error free, it may be of interest to ask how efficient each is when a tree constructed by it is compared with a tree randomly chosen from the set of all possible trees for a given number of OTUs. This question can be answered, at least with respect to topological construction, for the rooted tree. We first define T(k), the number of groups with k OTUs in the correct tree of n OTUs. If the correct tree is model tree a in Fig. 1, then T(i) = 1 for i = 2–8. Next, let R(k) be the expected number of groups with k OTUs in a randomly chosen tree from the set of trees with n OTUs. Tajima (1983) has shown that the probability of dividing n OTUs into n_1 and n_2 OTUs is given by $2/(n-1)$. Thus,

$$R(n-1) = 2/(n-1) \tag{5}$$

A group with $n-2$ OTUs occurs when n OTUs split into 2 and $n-2$ OTUs or when $n-1$ OTUs split into 1 and $n-2$ OTUs. Thus,

$$R(n-2) = 2/(n-1) + 2R(n-1)/(n-2) \tag{6}$$

Similarly, the expected number of groups with k OTUs is given by

$$R(k) = 2/(n-1) + 2R(n-1)/(n-2) + \ldots + 2R(k+1)/k \tag{7}$$

Using formulas (5) and (6), formula (7) can be reduced to

$$R(k) = (k+2)R(k+1)/k = 2n/[k(k+1)] \tag{8}$$

One more necessary factor is the probability, Q(k), that a group of k OTUs randomly sampled from n OTUs is made up of a unique combination of OTUs. Q(k), of course, is given by

$$Q(k) = 1/{}_nC_k \tag{9}$$

where ${}_nC_k$ is the binomial coefficient. Then, using T(k), R(k), and Q(k), it is possible to obtain the expected number of correct groups in a tree of n OTUs that is randomly chosen from the set. This is given by

$$E_n = \sum_{k=2}^{n-1} T(k)R(k)Q(k) \tag{10}$$

Since the total number of groups for n OTUs is n − 2, the mean distortion index d_T^* of a randomly sampled tree is given by

$$d_T^* = 2(n - 2 - E_n) \quad (11)$$

The efficiency of the method may be measured against d_T^* as the standard, that is, as $d_T^*/\overline{d_T}$ where $\overline{d_T}$ is the mean distortion index of the method. If the correct trees are the ones shown in Fig. 1, the value of d_T^* is 11.62 for 8 OTUs and 27.88 for 16 OTUs. Thus, in the case of one gene, the efficiencies are 3.56, 3.67, and 4.05 for the UPG, Farris, and modified Farris methods, respectively, for 8 OTUs, and 2.46, 2.51, and 2.74, respectively, for 16 OTUs. One problem with the efficiency measure is that it depends not only on the method used but also on the model tree, because d_T^* is independent of branch length whereas $\overline{d_T}$ is not. Notwithstanding this problem, it can be said that the efficiency declines with increasing number of OTUs, warning us that the error in molecular tree construction increases with increasing number of OTUs.

Faith (1985) has commented on the modified Farris method under the name "the Tateno, Nei, and Tajima method." He criticized us, saying that we misrepresented the distance Wagner algorithm in our paper (Tateno et al. 1983). His point is that we confused his Eq. (6) with his Eq. (7). That Eq. (7), however, does not completely express what we (Tateno et al. 1983) stated in our introduction of the modified Farris method. We compute D[B, (X, Y)] in Eq. (7) after (not before) OTU A joins the tree in his Fig. 4. Contrary to Eq. (7), he correctly follows the algorithm of the modified Farris method up to his Eqs. (8) and (9), which are involved in the computation of D[B, (X, Y)]. It should be noted that his Eqs. (10) and (11) do not mathematically represent the modified Farris method. Tateno et al. (1983) clearly state that D(X, Y) is computed when OTU A (not B) is joined to the tree. Faith (1985) also says that the modified Farris method is inferior to his modification on the basis of just one example using unspecified data. As the present study shows, his comparison of the methods is far from adequate. We are interested in such molecular data as nucleotide sequences and amino acid sequences in the construction of a phylogenetic tree, because such data are considered to reflect the evolution of organisms more directly than any other characters.

Acknowledgment. We thank Dr. K. Aoki for reading the first draft and making valuable comments that improved the presentation of the manuscript.

References

Daniels GR, Deininger PL (1985) Repeated sequence families derived from mammalian tRNA genes. Nature 317:819–822

Dayhoff MO (ed) (1972) Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, D.C.

Dekio S, Yamasaki R, Jidoi J, Hori H, Osawa S (1984) Secondary structure and phylogeny of *Staphylococcus* and *Micrococcus* 5S rRNAs. J Bacteriol 159:233–237

Faith DP (1985) Distance methods and approximation of most-parsimonious trees. Syst Zool 34:312–325

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

Gojobori T, Nei M (1984) Concerted evolution of the immunoglobulin V_H gene family. Mol Biol Evol 1:195–212

Hasegawa M, Yano T (1984) Phylogeny and classification of *Hominoidea* as inferred from DNA sequence data. Proc Japan Acad 60:389–392

Hori H, Osawa S (1979) Evolutionary change in 5S RNA secondary structure and a phylogenetic tree of 54 5S RNA species. Proc Natl Acad Sci USA 76:381–385

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

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England

Küntzel H, Köchel HG (1981) Evolution of rRNA and origin of mitochondria. Nature 293:751–755

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

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

Nei M, Stephens JC, Saitou N (1985) Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol Biol Evol 2:66–85

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

Schwartz RM, Dayhoff MO (1978) Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts. Science 199:359–403

Sibley CG, Ahlquist JE (1984) The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J Mol Evol 20:2–15

Sokal RR, Sneath PHA (1963) Principles of numerical taxonomy. Freeman, San Francisco

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Tateno Y (1985) Theoretical aspects of molecular tree estimation. In: Ohta T, Aoki K (eds) Population genetics and molecular evolution. Japan Sci Soc Press, Tokyo/Springer-Verlag, Berlin, pp 293–312

Tateno Y, Nei M, Tajima F (1983) Accuracy of estimated phylogenetic trees from molecular data, I. Distantly related species. J Mol Evol 18:387–404

Templeton AR (1983) Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of human and the apes. Evolution 37:221–244

Received February 13, 1986