Accuracy of Phylogenetic Trees Estimated from DNA Sequence Data[1]

John Sourdis and Costas Krimbas
Department of Genetics, Agricultural College of Athens

The relative merits of four different tree-making methods in obtaining the correct topology were studied by using computer simulation. The methods studied were the unweighted pair-group method with arithmetic mean (UPGMA), Fitch and Margoliash's (FM) method, the distance Wagner (DW) method, and Tateno et al.'s modified Farris (MF) method. An ancestral DNA sequence was assumed to evolve into eight sequences following a given model tree. Both constant and varying rates of nucleotide substitution were considered. Once the DNA sequences for the eight extant species were obtained, phylogenetic trees were constructed by using corrected (d) and uncorrected (p) nucleotide substitutions per site. The topologies of the trees obtained were then compared with that of the model tree. The results obtained can be summarized as follows: (1) The probability of obtaining the correct rooted or unrooted tree is low unless a large number of nucleotide differences exists between different sequences. (2) When the number of nucleotide substitutions per sequence is small or moderately large, the FM, DW, and MF methods show a better performance than UPGMA in recovering the correct topology. The former group of methods is particularly good for obtaining the correct unrooted tree. (3) When the number of substitutions per sequence is large, UPGMA is at least as good as the other methods, particularly for obtaining the correct rooted tree. (4) When the rate of nucleotide substitution varies with evolutionary lineage, the FM, DW, and MF methods show a better performance in obtaining the correct topology than UPGMA, except when a rooted tree is to be produced from data with a large number of nucleotide substitutions per sequence. (5) Data on the proportion of different nucleotides (p) between different DNA sequences tend to produce the correct tree with a slightly higher probability than those on the Jukes and Cantor distance (d) irrespective of the method used, but the difference is usually very small.

Introduction

Although there are many methods of constructing phylogenetic trees from molecular data, the relative merits of the methods are poorly understood. Using computer simulation, Tateno et al. (1982) studied the efficiencies of several tree-making methods in estimating the topology and branch lengths from nucleotide sequence data. The methods studied were the unweighted pair-group method with arithmetic mean (UPGMA; Sneath and Sokal 1973), Fitch and Margoliash's (FM; Fitch and Margoliash 1967) method, Farris's (1972) distance Wagner (DW) method, and Tateno et al.'s (1982) modified Farris (MF) method. They showed that UPGMA and the FM method are slightly better in obtaining the correct rooted tree when the rate of nucleotide

[1] Key words: UPGMA, Fitch and Margoliash's method, distance Wagner method, modified Farris method, rooted trees, unrooted trees.

Address for correspondence and reprints: Dr. John Sourdis, Department of Genetics, Agricultural College of Athens, Iera Odos 75, Athens 11855, Greece.

Mol. Biol. Evol. 4(2):159–166. 1987.

0737-4038/87/0402-4206$02.00

substitution is constant and the number of nucleotide substitutions is relatively large, whereas the DW and MF methods are better than the other two methods in obtaining the correct unrooted tree. The superiority of the DW and MF methods was especially clear when the rate of nucleotide substitution varied with evolutionary lineage.

However, this conclusion is based on a relatively small number of replications with a DNA sequence of 300 nucleotides, so that its generality is not very clear. Particularly, the Tateno et al. study on the effect of varying rate of substitution is not very reliable, because they studied only two different cases. We therefore studied this problem in more detail, conducting extensive computer simulation. The specific aims of this study were to clarify (1) the effect of the length of DNA used, (2) the effect of the number of nucleotide substitutions per site, (3) the effect of varying rate of nucleotide substitutions, and (4) the effect of the use of corrected and uncorrected nucleotide substitutions on the efficiency of recovering the correct tree. The last point is important, because some tree-making methods (e.g., the Farris method) are supposed to require a metric distance whereas others do not. The results obtained will be reported here.

Model and Methods

We used essentially the same method of computer simulation used by Tateno et al. (1982). The model tree used consisted of eight operational taxonomic units (OTUs) and had the topology given in figure 1. In the case of a constant rate of nucleotide substitution, the expected branch lengths were given by multiples of u, the expected number of nucleotide substitutions per site per unit evolutionary time. Thus, the expected length between the ancestor and OTU 1 was 7 u. In the case of varying substitution rate, this rate (u) varied with evolutionary unit time following the gamma distribution, so that the variance of the number of substitutions was twice as large as the mean (see Tateno et al. 1982 for details).

The ancestral sequence for each replicate simulation was generated by using pseudorandom numbers under the assumption that the four nucleotides A, T, C, and G are equally frequent. When nonsense codons appeared, they were replaced by sense codons that were again randomly generated. The ancestral sequence was duplicated

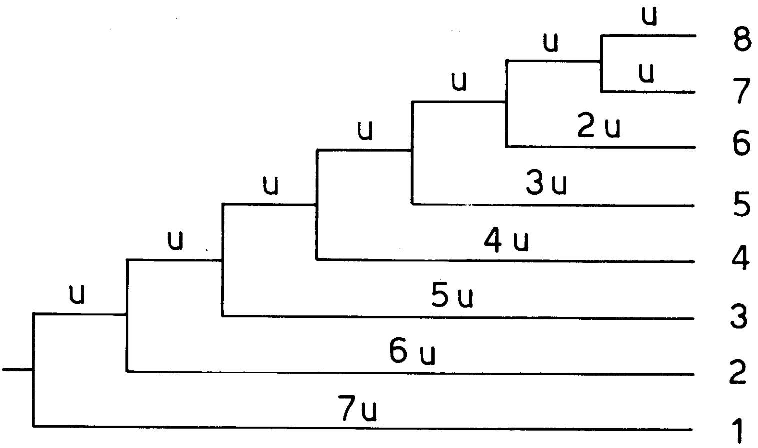

FIG. 1.—Model tree when there is a constant rate of nucleotide substitution. u = The expected number of nucleotide substitutions per site.

at each branching point of the model tree and subjected to random nucleotide substitution. All descendant nucleotide sequences were examined at the terminal points of the tree, and the number of uncorrected nucleotide differences per site (p) was computed for all pairs of OTUs. This p was then converted into the number of corrected nucleotide substitutions per site (d) by using Jukes and Cantor's (1969) formula. Thus we produced two different sets of distance matrices. Note that p is a metric following the triangle inequality whereas d is not. Each of the two sets of distance matrices was used to construct phylogenetic trees by using the UPGMA and the FM, DW, and MF methods. The topology of the trees obtained was then compared with that of the model tree. Following Tateno et al. (1982), we measured the accuracy of the topology in terms of (1) the proportion of the correct topology (P) obtained among all replications (empirical probability of obtaining the correct topology) and (2) the average distortion index ($\bar{d}_T$). The $\bar{d}_T$ is twice the number of branch interchanges required for a topology to be converted into the correct one. However, since $\bar{d}_T$ was highly correlated with P, we shall consider only P in the present paper. P was computed for both rooted and unrooted trees. The number of replications varied from 50 to 450 (see tables 1–3), depending on the accuracy of the results and the availability of computer time.

Table 1
P_R and P_U for Four Tree-making Methods When There Is a Constant Rate of Nucleotide Substitution with d

	P_R				P_U			
U (*N*)	UPGMA	FM	DW	MF	UPGMA	FM	DW	MF
150 bp:								
0.03 (275) . . .	0.000	0.009	0.004	0.004	0.008	0.061	0.036	0.062
0.05 (375) . . .	0.016	0.069	0.034	0.064	0.040	0.168	0.136	0.200
0.10 (275) . . .	0.073	0.153	0.131	0.156	0.160	0.316	0.357	0.393
0.20 (125) . . .	0.232	0.232	0.208	0.216	0.408	0.456	0.552	0.456
0.40 (125) . . .	0.400	0.184	0.104	0.114	0.552	0.424	0.448	0.376
300 bp:								
0.03 (250) . . .	0.040	0.100	0.064	0.096	0.088	0.240	0.208	0.260
0.05 (275) . . .	0.123	0.189	0.153	0.171	0.229	0.487	0.443	0.484
0.10 (275) . . .	0.342	0.422	0.338	0.393	0.509	0.702	0.723	0.713
0.20 (375) . . .	0.557	0.552	0.405	0.519	0.667	0.787	0.821	0.788
0.40 (175) . . .	0.691	0.531	0.400	0.502	0.834	0.760	0.771	0.791
900 bp:								
0.03 (275) . . .	0.378	0.429	0.415	0.437	0.484	0.822	0.796	0.807
0.05 (275) . . .	0.600	0.636	0.607	0.640	0.735	0.935	0.949	0.938
0.10 (275) . . .	0.876	0.844	0.753	0.844	0.916	0.989	0.992	0.992
0.20 (250) . . .	0.960	0.908	0.848	0.912	0.980	0.980	0.996	0.988
0.40 (100) . . .	0.990	0.950	0.820	0.950	1.000	0.990	0.990	0.990
1,500 bp:								
0.03 (275) . . .	0.600	0.663	0.649	0.662	0.738	0.947	0.959	0.976
0.05 (275) . . .	0.835	0.843	0.805	0.843	0.884	0.988	1.000	0.995
0.10 (275) . . .	0.917	0.935	0.900	0.935	0.946	1.000	1.000	1.000
0.20 (275) . . .	0.993	0.987	0.926	0.987	0.993	1.000	1.000	1.000
0.40 (125) . . .	1.000	0.950	0.950	0.990	1.000	1.000	1.000	1.000

NOTE.—U = Expected number of nucleotide substitutions per site between the ancestral sequence and the extant sequence; and *N* = number of replications.

Table 2
P_R and P_U for Four Different Tree-making Methods When There Is a Constant Rate of Nucleotide Substitution with p

SEQUENCE LENGTH AND U (*N*)	P_R				P_U			
	UPGMA	FM	DW	MF	UPGMA	FM	DW	MF
300 bp:								
0.03 (300) . . .	0.037	0.097	0.100	0.093	0.067	0.244	0.304	0.277
0.05 (300) . . .	0.120	0.243	0.203	0.250	0.230	0.553	0.543	0.590
0.10 (300) . . .	0.307	0.403	0.310	0.394	0.420	0.707	0.720	0.747
0.20 (300) . . .	0.590	0.540	0.393	0.527	0.694	0.814	0.810	0.813
0.40 (150) . . .	0.780	0.547	0.320	0.600	0.833	0.773	0.847	0.853
900 bp:								
0.03 (275) . . .	0.353	0.542	0.495	0.556	0.447	0.814	0.825	0.858
0.05 (275) . . .	0.633	0.680	0.637	0.683	0.720	0.953	0.967	0.967
0.10 (275) . . .	0.876	0.818	0.756	0.818	0.924	0.985	0.993	0.989
0.20 (275) . . .	0.953	0.927	0.789	0.927	0.982	0.996	1.000	1.000
0.40 (125) . . .	1.000	0.944	0.800	0.944	1.000	0.976	0.992	0.984
1,500 bp:								
0.03 (200) . . .	0.625	0.710	0.665	0.725	0.765	0.965	0.975	0.980
0.05 (200) . . .	0.875	0.870	0.835	0.870	0.925	0.995	0.995	1.000
0.10 (200) . . .	0.985	0.915	0.880	0.915	0.995	0.995	1.000	1.000
0.20 (200) . . .	0.995	0.980	0.955	0.985	1.000	0.995	1.000	1.000
0.40 (50)	1.000	0.880	0.820	0.960	1.000	0.800	1.000	1.000

NOTE.—See table 1 for notations.

Table 3
P_R and P_U for Four Different Tree-making Methods When There Is a Varying Rate of Nucleotide Substitution with d

SEQUENCE LENGTH AND U (*N*)	P_R				P_U			
	UPGMA	FM	DW	MF	UPGMA	FM	DW	MF
300 pb:								
0.03 (450) . . .	0.009	0.033	0.035	0.033	0.024	0.118	0.149	0.138
0.05 (450) . . .	0.016	0.105	0.089	0.098	0.060	0.300	0.315	0.336
0.10 (450) . . .	0.089	0.182	0.162	0.184	0.187	0.498	0.562	0.516
0.20 (450) . . .	0.227	0.336	0.293	0.327	0.345	0.636	0.709	0.673
0.40 (150) . . .	0.393	0.380	0.260	0.373	0.533	0.660	0.693	0.673
900 bp:								
0.03 (375) . . .	0.069	0.187	0.190	0.197	0.152	0.427	0.557	0.557
0.05 (375) . . .	0.125	0.299	0.296	0.314	0.211	0.672	0.800	0.786
0.10 (375) . . .	0.370	0.576	0.509	0.579	0.501	0.883	0.917	0.925
0.20 (275) . . .	0.725	0.747	0.629	0.720	0.829	0.963	0.976	0.971
0.40 (75)	0.907	0.800	0.600	0.787	0.987	0.969	0.947	0.933
1,500 bp:								
0.03 (250) . . .	0.160	0.328	0.316	0.356	0.300	0.744	0.816	0.836
0.05 (250) . . .	0.296	0.544	0.488	0.532	0.416	0.900	0.936	0.944
0.10 (250) . . .	0.704	0.712	0.676	0.716	0.816	0.996	1.000	1.000
0.20 (100) . . .	0.880	0.900	0.660	0.700	0.910	0.990	1.000	0.990

NOTE.—See table 1 for notations.

To determine the effect of the length of the DNA used, we considered nucleotide sequences of 150, 300, 600, 900, 1,200, and 1,500 bp. We also considered four different magnitudes of sequence divergence—i.e., U ≡ 7 u = 0.03, 0.05, 0.10, 0.20, and 0.40, where U is the expected number of nucleotide substitutions per site between the ancestral sequence and any of the extant sequences. In the case of varying substitution rate, the value of U varied with evolutionary lineage. However, the overall mean of U for all replications was the same as that for the case of constant substitution rate.

Our algorithms for reconstructing the tree by the FM, DW, and FM methods were slightly different from the standard ones. In the case of the FM method, we did not use Fitch and Margoliash's percent SD for choosing the final tree. Instead, we used the following measure:

$$R = \sum_{i<j}^{n} (e_{ij} - o_{ij})^2,$$

where $o_{ij} = d_{ij}/\sum_{i<j} d_{ij}$ and $e_{ij} = d'_{ij}/\sum_{i<j} d'_{ij}$. Here, d_{ij} is (corrected or uncorrected) d between the *i*th and *j*th OTUs whereas d'_{ij} is the corresponding patristic distance (sum of the lengths of all branches connecting the *i*th and *j*th OTUs). From among five reconstructed trees for each set of data (and excluding trees with negative branches), we chose a tree showing the smallest value of *R*. We could not test a large number of trees for each data set because this method requires a large amount of computer time. However, since there were a substantial number of trees with negative branch lengths, the total number of trees examined was considerably larger than five. The root of a tree was placed at the midpoint of a pair of OTUs showing the longest patristic distance.

Trees with negative branches were also excluded in the DW and MF methods. The algorithm we used for these methods terminates the process of tree making whenever a negative branch appears and restarts the process again from the next closest OTUs. (Note that in the standard DW and MF methods the tree-making process starts from a pair of OTUs with the smallest distance.) If this second trial fails to produce a tree with no negative branches, another trial is made starting from the next closest OTUs. This process is continued until a tree with no negative branches is produced. In practice, however, trees with negative branches were produced relatively infrequently with these two methods.

Results

The empirical probability of obtaining the correct topology for the case of constant rate of nucleotide substitution is presented in table 1. The results presented here were obtained by using data on the Jukes and Cantor distance. To save space, the results for the cases in which the total number of base pairs (*N*) was equal to 600 and 1,200 are not included; but the general pattern of the effect of the number of nucleotides used can be seen from this table. It is clear that when n = 150 bp and U ⩽ 0.05, the probability of obtaining the correct unrooted tree (P_U) = <0.20; particularly when U = 0.03, the probability is very low. This low probability is, of course, expected, because there are not many nucleotide differences between different sequences in this case. However, P_U increases as U increases. In the case of UPGMA, P_U is very small when U = 0.03 but becomes as high as 0.552 when U = 0.4. The P_U values for the FM, DW, and MF methods show the same pattern of increase with increasing U;

when U is small, they are much higher than those for UPGMA. However, the difference between the two groups of tree-making methods gradually diminishes with increasing U, and when U = 0.4 the P_{U} for UPGMA is higher. We also note that the values for the FM, DW, and MF methods are smaller when U = 0.4 than when U = 0.2. This apparently occurs because there are more backward and parallel mutations when U = 0.4 than there are when U = 0.2. Table 1 shows that the relationship of P_{U} with U for UPGMA is different from that for the other three methods. This difference apparently occurred because UPGMA is primarily used for constructing a species tree whereas the others are useful for constructing a tree when genes are polymorphic (see Nei 1987, p. 288–289).

The probabilities of obtaining the correct rooted tree (P_{R}) are considerably lower than the probabilities of obtaining P_{U}. This indicates that a substantial amount of error in constructing a rooted tree occurs at the time of rooting. This error is particularly large for the DW method, so that, among the FM, DW, and MF methods, it now shows the smallest values of P_{R}. When U is large, UPGMA shows the highest values of P_{R}.

When sequences with 300 bp are used, both P_{U} and P_{R} are substantially higher than they are when 150-bp sequences are used. UPGMA now produces the correct unrooted tree with a probability of 0.834 when U = 0.4. However, the relative merits of different tree-making methods remain nearly the same. When sequences with 900 nucleotides are used, the P_{U} and P_{R} values increase further. In this case, $P_{\mathrm{U}} > 0.90$ for all methods when U ⩾ 0.1. Even the rooted trees are correct, with $P \geqslant 0.8$—except for the DW method when U = 0.1. The relative merits of the four tree-making methods were again nearly the same when N = 150 bp and N = 300 bp. Essentially the same results also were obtained when N = 600 bp, though the P_{U} and P_{R} values were slightly smaller. When sequences of 1,500 bp with U ≥ 0.05 were used, the FM, DW, and MF methods almost always produced the correct unrooted tree. UPGMA shows a slightly lower P_{R} value, but if U ≥ 0.2, the unrooted tree obtained is expected to be correct most of the time. However, the rooted tree reconstructed may still be incorrect even when U ≥ 0.2, particularly when the FM, DW, and MF methods are used.

In the construction of phylogenetic trees from DNA sequence data, the Jukes and Cantor distance or some other linearized distance is often used. However, these corrected distances are not metrics, so they may not be appropriate for some tree-making methods, such as the DW method. To study this problem, we constructed trees by using data on the proportion of different nucleotides between the sequences compared (p). The results when N = 300, 900, and 1,500 bp are presented in table 2. Comparison of this table with table 1 indicates that, in recovering the true tree, data on p give slightly better results than those on d, not only for the DW method but also for the other methods. However, the average difference in P_{U} or P_{R} between the two sets of data is generally very small. Saitou and Nei (1986), using different model trees in their study of the relative merits of several tree-making methods, had similar results. Therefore, this finding seems to be a quite general one.

Table 3 shows the P_{U} and P_{R} values obtained from data on d when there is a varying rate of nucleotide substitution. These values are considerably lower than those seen when p remains constant (table 1); however, the extent of these differences is not the same for all tree-making methods: it is larger for UPGMA than for the other three methods. This is, of course, expected, because UPGMA depends on the assumption of a constant rate of nucleotide substitution and this assumption is violated in the present case. UPGMA is generally less reliable than the other tree-making methods in obtaining both unrooted and rooted trees, though there are several exceptions. Note also that even when the FM, DW, and MF methods are used, a large number of nucleotides must be used to obtain a reasonably accurate phylogenetic tree. (We also studied the P_U and P_R values for data on the Jukes and Cantor distance. This set of data also showed larger values of P_U and P_R, but the increment was very small [data not shown]).

Discussion

As mentioned earlier, Tateno et al. (1982) studied the relative merits of UPGMA and the FM, DW, and MF methods when N = 300 bp and U = 0.047, 0.093, and 0.187. Even though the number of replications used was only 20, their results are in rough agreement with ours. However, since they did not consider the case when U = 0.4, they did not notice that both the P_U and P_R values for UPGMA can be larger than those for the other methods (table 1). Actually, as long as the rate of nucleotide substitution is constant, UPGMA shows a good performance in obtaining the correct rooted or unrooted tree when p of substitutions per site is large. Furthermore, if we consider rooted trees only, this conclusion seems to be true even for the case of varying substitution rate.

Our study also indicates that the FM, DW, and MF methods show nearly the same performance in obtaining the correct topology, whereas the UPGMA P_U and P_R values are often different. Tateno et al. (1982) did not notice this pattern, apparently because the number of replications they used was too small.

Our results show that when the rate of nucleotide substitution varies from branch to branch, the FM, DW, and MF methods are usually better than UPGMA in constructing an unrooted tree. This result is the same as those of Tateno et al. (1982) and Blanken et al. (1982). This is true even when there is a constant substitution rate, unless the number of substitutions per site is large. In practice, the rate of nucleotide substitution would not be strictly constant (see Nei 1987), so that one may use any one of the FM, DW, and MF methods in constructing an unrooted tree.

We have also shown that, in all the tree-making methods, data on the proportion of different nucleotides between different sequences (p) often show a better performance than do those on d, though the difference is small. Therefore, one may use data on p for topology construction. However, if data on p are used, the branch lengths of a tree would be underestimated. One way to solve this problem would be to use data on d to estimate branch lengths after the topology of a tree is constructed by using data on p.

Acknowledgment

We are greatly indebted to Professor M. Nei for his help in the preparation of the manuscript.

LITERATURE CITED

BLANKEN, R. L., L. C. KLOTZ, and A. G. HINNEBUSCH. 1982. Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. J. Mol. Evol. **19**:9–19.

FARRIS, J. S. 1972. Estimating phylogenetic trees from distance matrices. Am. Nat. **106**:645–668.

FITCH, W. M., and E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **155:** 279–284.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

SAITOU, N., and M. NEI. 1986. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. (submitted).

SNEATH, P. M. A., and R. R. SOKAL. 1973. Numerical taxonomy. Freeman, San Francisco.

TATENO, Y., M. NEI, and F. TAJIMA. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. **18:**387–404.

MASATOSHI NEI, reviewing editor

Received July 11, 1986; revision received September 4, 1986.