Relative Efficiencies of the Maximum-Parsimony and Distance-Matrix Methods of Phylogeny Construction for Restriction Data[1]

Li Jin and Masatoshi Nei†*

*Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston; and †Institute of Molecular Evolutionary Genetics, Pennsylvania State University

The relative efficiencies of the maximum-parsimony (MP), UPGMA, and neighbor-joining (NJ) methods in obtaining the correct tree (topology) for restriction-site and restriction-fragment data were studied by computer simulation. In this simulation, six DNA sequences of 16,000 nucleotides were assumed to evolve following a given model tree. The recognition sequences of 20 different six-base restriction enzymes were used to identify the restriction sites of the DNA sequences generated. The restriction-site data and restriction-fragment data thus obtained were used to reconstruct a phylogenetic tree, and the tree obtained was compared with the model tree. This process was repeated 300 times. The results obtained indicate that when the rate of nucleotide substitution is constant the probability of obtaining the correct tree (P_c) is generally higher in the NJ method than in the MP method. However, if we use the average topological deviation from the model tree ($\bar{d}_T$) as the criterion of comparison, the NJ and MP methods are nearly equally efficient. When the rate of nucleotide substitution varies with evolutionary lineage, the NJ method is better than the MP method, whether P_c or $\bar{d}_T$ is used as the criterion of comparison. With 500 nucleotides and when the number of nucleotide substitutions per site was very small, restriction-site data were, contrary to our expectation, more useful than sequence data. Restriction-fragment data were less useful than restriction-site data, except when the sequence divergence was very small. UPGMA seems to be useful only when the rate of nucleotide substitution is constant and sequence divergence is high.

Introduction

Although DNA sequencing has become much easier since the introduction of the polymerase-chain-reaction method, the restriction-enzyme technique remains useful for constructing a phylogenetic tree for closely related populations or species (Avise and Lansman 1983; Wilson et al. 1985; Nei 1987, pp. 97–107). One of the most widely used methods of constructing a phylogenetic tree from restriction-site data is the maximum-parsimony (MP) method (Eck and Dayhoff 1966, pp. 162–168; Fitch 1971), though Nei and Tajima (1985) and Li (1986*a*) have shown that, when the number of nucleotide substitutions per site between DNA sequences is relatively large, the MP method is likely to introduce serious errors in tree construction. Sourdis

1. Key words: maximum parsimony, UPGMA, neighbor-joining, restriction-site data, restriction-fragment data.

Address for correspondence and reprints: Dr. Masatoshi Nei, Institute of Molecular Evolutionary Genetics, 328 Mueller Laboratory, Pennsylvania State University, University Park, Pennsylvania 16802.

Mol. Biol. Evol. 8(3):356–365. 1991.

0737-4038/91/0803-0007$02.00

and Nei (1988) and Saitou and Imanishi (1989) have shown, for DNA sequence data, that the MP method is less efficient in obtaining the correct tree than is either the minimum-evolution method (Cavalli-Sforza and Edwards 1967) or the neighbor-joining method (Saitou and Nei 1987), except when the number of nucleotide substitutions per site is very small and the number of nucleotides examined is very large.

Nevertheless, no studies have been done about the efficiency of the MP method relative to that of distance-matrix methods for restriction data. Since estimates of the number of nucleotide substitutions from restriction data are often subject to large sampling errors, Sourdis and Nei's and Saitou and Imanishi's conclusion may not apply to this case. We have therefore examined this problem by conducting a computer simulation. In this simulation, we have also examined the relative efficiencies, for phylogenetic construction, of restriction-site data and DNA sequence data. In the present paper, we consider only two distance-matrix methods, i.e., the unweighted pair-group method with arithmetic means (UPGMA) and the neighbor-joining (NJ) method. UPGMA is known to work well when the rate of nucleotide substitution is constant and the extent of sequence divergence is large (Tateno et al. 1982; Nei, accepted). By contrast, the NJ method is known to be as efficient as either the minimum-evolution method or the neighborliness method of Sattath and Tversky (1977) and Fitch (1981).

Method of Simulation

Our computer simulation, similar to that of Sourdis and Nei (1988), was to set up a model tree, simulate the evolutionary change of nucleotide sequences following this model tree, and compare the tree reconstructed from the simulated sequence data with the model tree. The model tree consisted of six DNA sequences (fig. 1). We considered both the case of constant and the case of varying rate of nucleotide substitution. In the case of constant rate of substitution (fig. 1A), the expected number of per-site substitutions from the ancestral sequence to the extant sequences was denoted by U, whereas the length of each branch was a multiple of a, which was one-fifth of U. In the present study we considered the cases of U = 0.0125, 0.025, 0.05, and 0.10. Note that U is half the expected distance (number of nucleotide substitutions) between two most distant DNA sequences. Thus, in the case of U = 0.1, the expected distance

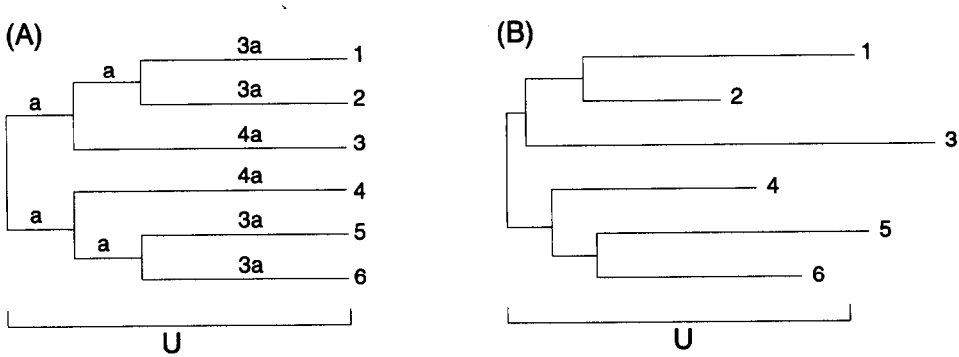

FIG. 1.—Model trees used for computer simulation for six DNA sequences. (A) and (B) represent the cases of constant and varying rate of substitution, respectively. In tree (A), the expected number of nucleotide substitutions per site for a branch is expressed as multiples of a, and the expected number of nucleotide substitutions per site from the ancestral sequence to an extant sequence is U (=5a). In tree (B), the length of each branch was generated randomly according to the gamma distribution described in the text.

between sequences 1 and 4 is 0.2. Since estimates of the number of substitutions obtained from restriction-site data are unreliable when $2U > 0.2$ (Brown et al. 1979; Nei and Li 1979; Li 1981), we did not consider the cases of $2U > 0.2$. In the case of varying rates of nucleotide substitution, the *expected* number of substitutions for a branch was determined under the assumption that it follows a gamma distribution the mean of which is the same as that of the corresponding case of constant rate of substitution [for details of the procedure, see Tateno et al. (1982)]. Figure 1B is the model tree determined by this procedure. Once the expected length of a particular branch was determined, the *actual* number of nucleotide substitutions for that branch was obtained by generating Poisson random numbers in each replicate simulation.

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. Each DNA sequence was assumed to be circular and composed of 16,000 nucleotides, as in the case of mammalian mitochondrial DNAs. The ancestral sequence was duplicated at each branching point of the model tree and subjected to nucleotide substitution. Nucleotide substitution was assumed to follow either the one-parameter model (random substitution among the four nucleotides; Nei 1987, p. 65) or Kimura's (1980) two-parameter model. In the latter model, the rate of transitional change was assumed to be 20 times greater than that of transversional change. At the end of the evolutionary process considered, all nucleotide sequences were recorded. This was repeated 300 times.

The actual pattern of nucleotide substitution is, of course, much more complicated than the one-parameter or two-parameter model (see Nei 1987, chap. 5). In the case of restriction-site data, however, these models seem to be sufficient for obtaining a general idea of the relative efficiencies of different tree-making methods, as will be mentioned in the Discussion.

To obtain restriction-site data, the recognition sequences of 20 six-base restriction enzymes were used to identify the recognition sites of each extant DNA sequence. The enzymes used were *Bal*I, *Bam*HI, *Bcl*I, *Bgl*II, *Cla*I, *Dra*I, *Eco*RI, *Eco*RV, *Hin*dIII, *Kpn*I, *Nco*I, *Nrv*I, *Pst*I, *Pvu*II, *Sac*I, *Sal*I, *Sma*I, *Sst*II, *Xba*I, and *Xho*I. A restriction-site map was then constructed for each DNA sequence.

In phylogenetic construction, not only restriction-site data but also restriction-fragment data are used (for details of the procedure, see Nei 1987, pp. 96–107). In practice, however, small restriction fragments can not be identified by the standard electrophoresis used. Two restriction fragments of which the sizes are similar to each other are also indistinguishable. We therefore assumed that restriction fragments <100 nucleotides long are undetectable, as is usually the case with mammalian mitochondrial DNAs. We also assumed that two restriction fragments of which the sizes differ by $\leq 1\%$ of the average of the two fragments compared are indistinguishable.

To see the effect that undetectability of either small restriction fragments or fragment differences had on the efficiency of phylogenetic reconstruction, we also considered the case where every fragment or every fragment difference is detectable. We call the data obtained under this assumption "high-resolution fragment" (HRF) data.

The nucleotide sequences, restriction-site data, and restriction-fragment data thus obtained were used for constructing phylogenetic trees by using the MP, NJ, and UPGMA methods. The MP tree was constructed by using Saitou and Imanishi's (1989) computer program for DNA sequences and restriction-site data. This program examined all 105 possible topologies for six DNA sequences. In the cases of NJ and UPGMA, the number of nucleotide substitutions between sequences (sequence di-

vergence) was estimated by Jukes and Cantor's (1969), Nei and Tajima's (1983), and Nei and Li's (1979) (see Nei 1987, p. 106) methods for DNA sequences, restriction-site data, and restriction-fragment data, respectively. For DNA sequence data, sequence divergences were estimated by Jukes and Cantor's method even when nucleotide substitution followed the two-parameter model. We did this because, when 2U is small, Jukes and Cantor's method gives fairly reliable estimates even if there is a transition/transversion bias (Tajima and Nei 1984).

Following Tateno et al. (1982), we measured the efficiency of a tree-making method by two quantities, i.e., the probability of obtaining the correct tree (P_c) and the average topological deviation ($\bar{d}_T$) of estimated trees from the model tree. P_c was actually estimated by the proportion, among 300 replications, of those replications for which only the correct tree was obtained. $\bar{d}_T$ was computed by Robinson and Foulds' (1981) method. In the range of values observed here, $\bar{d}_T$ is roughly equal to twice the mean number of branch interchanges required to convert one topology to the other and takes a value of zero when the topology of an estimated tree is identical with that of the model tree.

Results

Constant Rate of Nucleotide Substitution

Table 1 shows the P_c and $\bar{d}_T$ values for the case of a constant rate of nucleotide substitution with the one-parameter model. For restriction-site data, P_c is highest for the NJ method and lowest for UPGMA, for all U values considered except for U = 0.1. For U = 0.1, P_c is nearly the same (no statistical difference) for all the tree-making methods. The standard error of P_c is given by $[P_c(1-P_c)/300]^{1/2}$. A good performance of UPGMA when the distance value is high is similar to that observed

Table 1
P_c (×100) and $\bar{d}_T$ (in Parentheses) for Model Tree A

U and Method	Sequence Length (bp)			Restriction-Site Data	Restriction-Fragment Data	HRF Data
	16,000	1,000	500			
0.0125:						
NJ	100 (0.00)	86 (0.29)	55 (1.02)	69 (0.75)	54 (1.12)	54 (1.12)
UPGMA ...	100 (0.00)	57 (0.99)	31 (2.01)	40 (1.65)	42 (1.63)	43 (1.59)
MP	. . .	79 (0.31)	41 (0.97)	58 (0.76)	. . .	. . .
0.025:						
NJ	100 (0.00)	98 (0.05)	83 (0.35)	82 (0.39)	62 (0.91)	66 (0.79)
UPGMA ...	100 (0.00)	81 (0.43)	58 (0.99)	63 (0.86)	53 (1.21)	55 (1.13)
MP	. . .	96 (0.06)	72 (0.40)	73 (0.42)	. . .	. . .
0.05:						
NJ	100 (0.00)	100 (0.01)	95 (0.09)	84 (0.34)	58 (1.02)	58 (1.00)
UPGMA ...	100 (0.00)	96 (0.07)	80 (0.42)	70 (0.63)	55 (1.06)	56 (1.01)
MP	. . .	100 (0.01)	92 (0.11)	77 (0.37)	. . .	. . .
0.10:						
NJ	100 (0.00)	100 (0.00)	98 (0.04)	81 (0.39)	39 (1.57)	45 (1.38)
UPGMA ...	100 (0.00)	99 (0.02)	90 (0.19)	83 (0.38)	51 (1.24)	59 (0.96)
MP	. . .	100 (0.00)	99 (0.02)	80 (0.36)	. . .	. . .

NOTE.—A one-parameter model is used. The number of replications is 300.

with DNA sequence data, although, in the latter data, U must be even higher (Tateno et al. 1982; Sourdis and Krimbas 1987).

With the criterion of P_c, the NJ method is always better than the MP method. However, this is partly because the MP method often produces several equally parsimonious trees including the correct one, and this case was not included in the computation of P_c. In the NJ method, this type of tied trees was rarely observed. Therefore, one can argue that P_c is not an appropriate criterion to compare the MP method with other methods. A better criterion for this purpose is obviously $\bar{d}_T$, because this measures the average topological distance between the model tree and a bifurcation tree, over all replications and over all equally parsimonious trees obtained (Nei, accepted). (One can also construct a consensus tree for both the MP and NJ methods in this case, but the comparison of such consensus trees has several problems; see Nei, accepted). If we use $\bar{d}_T$ as the criterion, the MP and NJ methods are nearly equally efficient in obtaining the correct tree. (The distribution of $\bar{d}_T$ is highly skewed, so that it is difficult to conduct a statistical test of the difference between $\bar{d}_T$'s. We therefore make our judgment by looking both at the face values and at the general trend of $\bar{d}_T$.)

Distance estimates obtained from restriction-fragment data are known to be less reliable than those obtained from restriction-site data, particularly when $2U > 0.05$ (Nei 1987, p. 107). For this reason, P_c is significantly lower in restriction-fragment data than in restriction-site data, for both the NJ and the UPGMA methods. (The MP method was not used for restriction-fragment data because it cannot be used unless a number of unrealistic assumptions are made.) As in the case of restriction-site data, the NJ method is better than UPGMA, except for the case of $U = 0.1$, whether P_c or $\bar{d}_T$ is used as the criterion. For the NJ method, P_c declines and $\bar{d}_T$ increases as 2U increases over 0.05. This is because restriction-fragment data do not give a reliable estimate of sequence divergence when $2U > 0.05$.

As mentioned earlier, in the restriction-fragment method, small fragments are usually undetected and two fragments of similar sizes are often indistinguishable. Therefore, one might think that restriction-fragment data are less efficient in obtaining the correct tree than are HRF data. Table 1, however, shows that both P_c and $\bar{d}_T$ are very similar for both types of data. This indicates that there is no use in making great efforts to detect small restriction fragments or fragment differences in experiments.

Table 1 also includes the P_c and $\bar{d}_T$ values for nucleotide sequence data. When all 16,000 nucleotides are used, both NJ and UPGMA show $P_c = 100\%$ and $\bar{d}_T = 0$. The MP method is also expected to show the same P_c and $\bar{d}_T$ values, though the simulation was not conducted in this case because of the large computer time required. The relative efficiencies of the MP, UPGMA, and NJ methods for the cases of 1,000 or 500 nucleotides examined are nearly the same as those of the methods of Sourdis and Krimbas (1987), Sourdis and Nei (1988), Saitou and Imanishi (1989), and Jin and Nei (1990).

In our simulation of restriction-site data, we used 20 six-base enzymes under the assumption of equal nucleotide frequencies. In this case the expected number of restriction sites for a sequence of 16,000 nucleotides is $20 \times 16{,}000 \times (1/4)^6 = 78$. Since each restriction site consists of six bases, this type of experiment is expected to survey 468 nucleotides per sequence. This suggests that, with 500 nucleotides, restriction-site data are slightly less efficient in recovering the true tree than are DNA sequence data. This is particularly so if we note that the restriction-site method surveys DNA sequence changes only indirectly.

Table 1, however, shows that, for U = 0.0125 and with 500 nucleotides, P_c is higher and $\bar{d}_T$ is smaller in restriction-site data than in sequence data, the difference in P_c between them being statistically significant at the 1% level. This is somewhat surprising, but it is understandable if we note that the average number of restriction sites (104.9) that could be used for phylogenetic analysis was considerably larger than 78. This difference occurred because many new restriction sites were created during the evolutionary time considered. Thus, the actual number of nucleotides that were used for sequence comparison was effectively 630, which was larger than the number (500) used for direct sequence comparison. This indicates that the restriction-site method is quite efficient when U is small.

However, with 500 nucleotides, the advantage of restriction-site data over sequence data gradually declines as U increases. When U = 0.025, P_c and $\bar{d}_T$ are nearly the same for the two sets of data, but, when $U \geq 0.05$, restriction-site data are less efficient than sequence data, for obtaining the correct tree. This occurs even though the number of restriction sites available for phylogenetic analysis increases with increasing U. The reason for this seems to be that, as U increases, the probability of occurrence of parallel losses and reacquisition of restriction sites at the same position increases and thus the utility of restriction-site data for phylogenetic construction declines (Nei and Tajima 1985; Li 1986*a*).

Table 2 shows the P_c and $\bar{d}_T$ values for the case of constant rate of nucleotide substitution with the two-parameter model. The results obtained are very similar to those in table 1, except that the absolute values of P_c's are slightly smaller in this case than in the case of the one-parameter model. Therefore, the same conclusions discussed above apply to the case of the two-parameter model.

Table 2
P_c ($\times$100) and $\bar{d}_T$ (in Parentheses) for Model Tree A

U and Method	Sequence Length (bp) 16,000	1,000	500	Restriction-Site Data	Restriction-Fragment Data	HRF Data
0.0125:						
NJ	100 (0.00)	83 (0.35)	51 (1.16)	67 (0.79)	54 (1.13)	58 (1.00)
UPGMA ...	100 (0.00)	57 (1.01)	32 (2.02)	37 (1.71)	34 (1.93)	35 (1.87)
MP	. . .	75 (0.36)	39 (1.06)	55 (0.79)	. . .	. . .
0.025:						
NJ	100 (0.00)	95 (0.09)	77 (0.48)	76 (0.56)	59 (0.98)	62 (0.90)
UPGMA ...	100 (0.00)	79 (0.45)	54 (1.11)	60 (0.94)	53 (1.19)	55 (1.13)
MP	. . .	89 (0.14)	67 (0.52)	67 (0.56)	. . .	. . .
0.05:						
NJ	100 (0.00)	99 (0.03)	87 (0.25)	79 (0.45)	54 (1.11)	60 (0.93)
UPGMA ...	100 (0.00)	96 (0.00)	77 (0.51)	69 (0.68)	53 (1.16)	57 (1.03)
MP	. . .	99 (0.01)	85 (0.22)	73 (0.46)	. . .	. . .
0.10:						
NJ	100 (0.00)	98 (0.04)	93 (0.15)	73 (0.57)	33 (1.77)	35 (1.65)
UPGMA ...	100 (0.00)	98 (0.03)	87 (0.25)	76 (0.50)	44 (1.55)	46 (1.37)
MP	. . .	99 (0.01)	91 (0.16)	64 (0.66)	. . .	. . .

NOTE.—A two-parameter model is used. The number of replications is 300.

Varying Rate of Nucleotide Substitution

The P_c and $\bar{d}_T$ values for model tree B in figure 1 are presented in table 3 (for the one-parameter model) and in table 4 (for the two-parameter model). Table 3 shows that, when the rate of nucleotide substitution varies with evolutionary lineage, the P_c values are considerably lower compared with those of the case of constant rate (table 1). Thus, the P_c values for restriction-site data are approximately one-half those in table 1. UPGMA generally shows a small P_c value, indicating that it is not a good method for the case of varying rate, as has already been established with DNA sequence data (e.g., see Saitou and Nei 1987). The P_c value for the MP method is always smaller than that for the NJ method, as in the case of DNA sequence data (Sourdis and Nei 1988).

Table 3 shows that, for U = 0.0125 and with 500 nucleotides, restriction-site data are again more useful for tree-making than are sequence data but that the superiority of restriction-site data declines as U increases. This conclusion is the same as that from tables 1 and 2. For U = 0.0125 and with 500 nucleotides, restriction-fragment data are nearly as useful as restriction-site data and sequence data, but, when $U \geq 0.025$, they are generally less useful. The P_c value for restriction-site data is always <0.5. This indicates that, with the type of true tree given in figure 1B, more than 20 six-base restriction enzymes are necessary to obtain a reliable phylogenetic tree.

When the two-parameter model of nucleotide substitution is used, P_c generally becomes even smaller. However, the relative efficiencies of the MP, UPGMA, and NJ methods remain essentially the same. Therefore, the conclusions obtained above apply regardless of whether the transition/transversion bias exists.

Table 3
P_c (×100) and $\bar{d}_T$ (in Parentheses) for Model Tree B

U and Method	Sequence Data (bp)			Restriction-Site Data	Restriction-Fragment Data	HRF Data
	16,000	1,000	500			
0.0125:						
NJ	100 (0.00)	55 (1.08)	26 (2.07)	39 (1.63)	30 (2.06)	33 (1.95)
UPGMA ...	20 (1.99)	16 (3.23)	9 (3.69)	13 (3.49)	12 (3.55)	13 (3.63)
MP	. . .	38 (1.06)	12 (1.86)	24 (1.71)	. . .	. . .
0.025:						
NJ	100 (0.00)	80 (0.41)	55 (1.05)	49 (1.21)	28 (2.08)	27 (2.06)
UPGMA ...	16 (1.83)	17 (3.18)	11 (3.58)	15 (3.36)	11 (3.53)	10 (3.57)
MP	. . .	69 (0.43)	36 (1.14)	29 (1.41)	. . .	. . .
0.05:						
NJ	100 (0.00)	90 (0.20)	75 (0.53)	45 (1.31)	22 (2.36)	26 (2.09)
UPGMA ...	4 (1.92)	20 (2.77)	15 (3.18)	18 (3.15)	10 (3.48)	9 (3.51)
MP	. . .	88 (0.17)	57 (0.66)	32 (1.47)	. . .	. . .
0.10:						
NJ	100 (0.00)	91 (0.19)	83 (0.35)	41 (1.39)	18 (2.90)	15 (2.87)
UPGMA ...	0 (1.99)	16 (2.61)	20 (2.97)	13 (2.97)	7 (3.45)	10 (3.28)
MP	. . .	89 (0.18)	71 (0.44)	31 (1.66)	. . .	. . .

NOTE.—A one-parameter model is used. The number of replications is 300.

Discussion

One of the main conclusions in the present paper is that, as long as the rate of nucleotide substitution remains constant and $U \leq 0.1$, the NJ and MP methods are equally efficient for restriction-site data when $\bar{d}_T$ is used as the criterion of comparison, whereas the former method is better than the latter when P_c is used as the criterion. This conclusion is essentially the same as that of Sourdis and Nei (1988) and Nei (accepted) for nucleotide sequences. However, when the rate of nucleotide substitution varies with evolutionary lineages, the MP method is less efficient than the NJ method except when U is very small. Therefore, the present study indicates that the NJ method is generally superior to the MP method. Furthermore, there are two additional advantages of the former. First, the computational time required for the NJ method is much shorter than that for the MP method. Second, the former gives branch length estimates as well as the topology, whereas the latter is designed primarily to generate the topology though branch lengths can be estimated under some restricted conditions (Fitch 1971). In the case of restriction-site data, Nei and Tajima (1985) have shown that the MP method seriously underestimates branch lengths even if U is fairly small.

In the present study we have considered only six-base restriction enzymes. In practice, four-base enzymes are also often used. If all four-base restriction sites are identifiable, then, for phylogenetic construction, these enzymes are more useful than six-base enzymes. In practice, however, it is not easy to identify either all restriction sites or a high proportion of restriction fragments for these enzymes. For this reason, researchers often prefer to use six-base enzymes. Of course, if there is a complete DNA sequence available for one individual or one species, four-base enzymes can be used

Table 4
P_c (×100) and $\bar{d}_T$ (in Parentheses) for Model Tree B

U and Method	Sequence Data (bp) 16,000	1,000	500	Restriction-Site Data	Restriction-Fragment Data	HRF Data
0.0125:						
NJ	100 (0.00)	51 (1.19)	25 (2.17)	36 (1.76)	28 (2.14)	29 (2.02)
UPGMA ...	21 (2.00)	16 (3.25)	9 (3.68)	8 (3.83)	8 (3.80)	9 (3.68)
MP	. . .	36 (1.14)	13 (2.02)	20 (1.88)	. . .	. . .
0.025:						
NJ	100 (0.00)	73 (0.58)	48 (1.25)	40 (1.47)	25 (2.35)	25 (2.21)
UPGMA ...	16 (1.85)	17 (3.17)	13 (3.51)	14 (3.44)	11 (3.65)	12 (3.57)
MP	. . .	56 (0.69)	32 (3.51)	29 (1.53)	. . .	. . .
0.05:						
NJ	100 (0.00)	80 (0.40)	63 (0.86)	46 (1.31)	23 (2.25)	27 (2.05)
UPGMA ...	3 (1.93)	19 (2.82)	14 (3.16)	20 (3.06)	15 (3.40)	15 (3.32)
MP	. . .	68 (0.52)	45 (1.01)	30 (1.62)	. . .	. . .
0.10:						
NJ	100 (0.00)	82 (0.38)	62 (0.83)	33 (1.70)	17 (2.89)	17 (2.81)
UPGMA ...	1 (1.98)	15 (2.70)	20 (2.85)	17 (3.07)	11 (3.35)	13 (3.29)
MP	. . .	65 (0.63)	44 (1.05)	20 (1.97)	. . .	. . .

NOTE.—A two-parameter model is used. The number of replications is 300.

more effectively by comparing all restriction sites with the DNA sequence (Cann et al. 1984). Although we have not studied the P_c and $\bar{d}_T$ values for four-base restriction enzymes, we believe that our conclusions will apply to restriction data from these enzymes as well.

As mentioned earlier, we used the one-parameter and two-parameter models of nucleotide substitution to simulate the evolutionary change of DNA sequences, though the actual pattern of nucleotide substitution is usually much more complicated. We did this because, when $U \leq 0.1$, the effects of various disturbing factors are relatively small. One of the most serious factors is variation in the rate of nucleotide substitution among different nucleotide sites. For example, in the coding region of a gene, the third nucleotide position of a codon generally shows a higher rate of substitution than do the first and second positions. This factor, however, introduces a relatively minor effect on the estimate of nucleotide divergence, when six-base restriction enzymes are used. This is because every restriction site consists of two each of the first, second, and third positions. When four-base restriction enzymes are used, a restriction site may include one or two third positions. Yet, Li (1986*b*) has shown that Nei and Li's (1979) method gives a fairly good estimate of nucleotide divergence as long as $U < 0.1$. Considering DNA sequences, Jin and Nei (1990) studied the effect of various types of substitution-rate variation among different nucleotide sites on the P_c values for the NJ and MP methods. Their results show that this effect is important only when U is large. Jin and Nei also showed that nucleotide-substitution patterns different from those of the one-parameter and two-parameter models do not seriously affect the P_c values unless U is large. Since the restriction-enzyme method is supposed to be used only for the case of a small U value, the conclusions obtained in the present paper are expected to apply to a wide variety of situations. When U is large, one should use DNA sequence data rather than restriction-site data.

Acknowledgments

This study was supported by research grants from the National Institute of Health and the National Science Foundation.

LITERATURE CITED

AVISE, J. C., and R. A. LANSMAN. 1983. Polymorphism of mitochondrial DNA in populations of higher animals. Pp. 147–164 *in* M. NEI and R. K. KOHEN, eds. Evolution of genes and proteins. Sinauer, Sunderland, Mass.

BROWN, W. M., M. GEORGE, JR., and A. C. WILSON. 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA **76**:1967–1971.

CANN, R. L., W. M. BROWN, and A. C. WILSON. 1984. Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. Genetics **106**:479–499.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS. 1967. Phylogenetic analysis: models and estimation procedures. Am. J. Hum. Genet. **19**:233–257.

ECK, R. V., and M. O. DAYHOFF. 1966. Atlas of protein sequence and structure. National Biomedical Research Foundation, Silver Springs, Md.

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

———. 1981. A non-sequential method for constructing trees and hierarchical classification. J. Mol. Evol. **18**:30–37.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

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.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

LI, W.-H. 1981. A simulation study of Nei and Li's model for estimating DNA divergence from restriction enzyme maps. J. Mol. Evol. **17**:251–255.

———. 1986*a*. Evolutionary change of restriction cleavage sites and phylogenetic inference. Genetics **113**:187–213.

———. 1986*b*. Evolutionary change of restriction sites under unequal rates of nucleotide substitution among the three positions of codons. J. Mol. Evol. **23**:205–210.

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

———. Relative efficiencies of different tree-making methods for molecular data. *In* M. M. MIYAMOTO and J. CRACRAFT, eds. Recent advances in phylogenetic studies of DNA sequences. Oxford University Press, Oxford (*accepted*).

NEI, M., and W.-H. LI. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76**:5269–5273.

NEI, M., and F. TAJIMA. 1983. Maximum likelihood estimation of the number of nucleotide substitutions from restriction sites data. Genetics **105**:207–217.

———. 1985. Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. Mol. Biol. Evol. **2**:189–205.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

SAITOU, N., and T. IMANISHI. 1989. Relative efficiencies of the Fitch-Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol. Biol. Evol. **6**:514–525.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SATTATH, S., and A. TVERSKY. 1977. Additive similarity trees. Psychometrika **42**:319–345.

SOURDIS, J., and C. KRIMBAS. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. Mol. Biol. Evol. **4**:159–166.

SOURDIS, J., and M. NEI. 1988. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. **5**:298–311.

TAJIMA, F., and M. NEI. 1984. Estimation of evolutionary distance between nucleotide sequences. Mol. Biol. Evol. **1**:269–285.

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.

WILSON, A. C., R. L. CANN, S. M. CARR, M. GEORGE, U. B. GYLLENSTEN, K. M. HELM-BYCHOWSKI, R. G. HIGUCHI, S. R. PALUMBI, E. M. PRAGER, R. D. SAGE, and M. STONEKING. 1985. Mitochondrial DNA and two perspectives on evolutionary genetics. Biol. J. Linnaean Soc. **26**:375–400.

WALTER M. FITCH, reviewing editor

Received August 7, 1990; revision received December 14, 1990

Accepted December 17, 1990