6

Relative Efficiencies of Different Tree-Making Methods for Molecular Data

MASATOSHI NEI

There are many different tree-making methods that can be used for molecular data (Nei, 1987a; Felsenstein, 1988). Each of these methods has some advantages and disadvantages, and the overall relative efficiencies of the methods in recovering the correct phylogenetic tree are still controversial. The major problem in studying the relative efficiencies is that the true tree is usually unknown for any set of real organisms or any set of real DNA sequences, so that it is difficult to judge which tree is the correct one. However, this problem can be avoided if we use computer simulation. In the case of molecular data, particularly DNA sequences, the pattern of evolutionary change is well-understood so that it is possible to simulate it for any given phylogenetic tree. Sequences of DNA or other molecular data thus generated can be used for reconstructing the tree, and the efficiency of a tree-making method can be studied by examining how often it recovers the correct tree.

This type of study was first conducted by Peacock and Boulter (1975) who considered the evolution of amino acid sequences. However, they compared only two different tree-making methods considering a constant rate of evolution. Furthermore, they did not consider the property of the genetic code in their simulation of amino acid changes. Therefore, their conclusion has limited applicability to actual data. In the late 1970s we initiated a comprehensive study of this problem, considering DNA sequences (Tateno et al., 1982). Initially, we studied relatively simple tree-making methods but later extended our analysis to other methods. Similar studies have also been done by a number of other authors. In this chapter, a summary of the results of these studies is presented. Before the discussion of these results, however, the theoretical basis of each tree-making method that is used for molecular data will be presented.
TREE-MAKING METHODS

As mentioned above, there are many methods for constructing trees from molecular data. According to the type of data used, they can be divided into two categories; that is, distance methods and discrete-character methods. In distance methods, evolutionary distance is computed for all pairs of operational taxonomic units (OTUs; species or populations) or DNA (or amino acid) sequences, and a phylogenetic tree is constructed by considering the relationship among these distances. Once distance values are obtained, there are several ways of obtaining a tree. In discrete-character methods, data with discrete character-states such as nucleotide states in DNA sequences are used, and a tree is constructed by considering the evolutionary relationships of OTUs or DNA sequences at each character or nucleotide position.

It should be noted that some types of molecular data (e.g., DNA hybridization data) exist only as distance data. Therefore, phylogenetic trees for these data can be constructed only by distance methods. By contrast, discrete-character data can usually be converted into distance data, so that they can be analyzed either by distance methods or by discrete-character methods. Some authors (e.g., Farris, 1981; Penny, 1982) have argued that distance methods are inherently inferior to discrete-character methods (e.g., parsimony methods), but their arguments are apparently based on misconceptions of distance methods (Felsenstein, 1986; Nei, 1987b). Actually, some distance methods are often superior to parsimony methods in obtaining the correct tree, as will be mentioned below.

For both distance and discrete-character data, there are several different tree-making methods that are based on different principles. The major ones are as follows.

Distance Data

**Average Distance (Linkage) Method [Unweighted Pair Group Method with Arithmetic Means (UPGMA)]**

The original idea of this method was presented by Sokal and Michener (1958), but their procedure is different from the one currently in use. A formal presentation of this method is given in Sneath and Sokal (1973). To construct a phylogenetic tree by this method, it is necessary to assume a constant rate of evolution.

**Transformed Distance (TD) Method**

Farris (1977) showed that if a distance matrix satisfies the condition of an additive tree (all substitutions being counted) the UPGMA procedure produces the correct tree topology by using the following transformed distance for species \(i\) and \(j\).

\[
d_{ij}^* = (d_{ij} - d_{ir} - d_{rj})/2 + c,
\]

(1)
where \( r \) stands for a reference species (an outgroup or any given ingroup species) and \( c \) is a constant to make \( d_{ij} \) positive. The TD method works well whether the evolutionary rate varies from branch to branch. In practice, of course, the condition of an additive tree is not usually satisfied, so this method may produce a wrong topology. Branch lengths cannot be estimated by this method. However, once a topology is obtained, one can easily estimate branch lengths by using Fitch and Margoliash's (1967) method mentioned below, or some other method. There are several different algorithms for this method (Klotz et al., 1979; Klotz and Blanken, 1981; Li, 1981).

**Fitch and Margoliash's (FM) Method**

Let \( d_{ij} \) be the observed distance between species \( i \) and \( j \) and \( e_{ij} \) be the corresponding patristic distance, which is the sum of the lengths of all branches connecting the two species in a reconstructed tree. Fitch and Margoliash's (1967) method chooses a tree that minimizes the following quantity.

\[
S_{FM} = \left[ \frac{2 \sum_{i<j} (d_{ij} - e_{ij})/d_{ij}^2}{n(n - 1)} \right]^{1/2} \times 100, \tag{2}
\]

where \( n \) is the number of species studied. A similar least-squares method was also proposed by Cavalli-Sforza and Edwards (1967). In this method the \( S_{FM} \) value must be computed for all or a large number of topologies.

**Minimum Evolution (ME) Method**

In an unrooted bifurcating tree of \( n \) species, there are \( 2n - 3 \) possible branches. Let \( \ell_i \) be the branch length of the \( i \)-th branch for a given topology. One can then compute the sum of all branches. That is,

\[
L = \sum \ell_i. \tag{3}
\]

In the ME method, this quantity is computed for all possible topologies, and the topology that shows the minimum \( L \) value is chosen as the final tree. This method was originally presented by Cavalli-Sforza and Edwards (1967), but their method of computing \( \ell_i \) was very complicated. Recently, Saitou and Imanishi (1989) simplified the computation considerably by using Fitch and Margoliash's (1967) approach.

One might think that this method is essentially the same as the maximum parsimony method mentioned below. This is not true. Unlike the parsimony method, this method is not affected by parallel or backward mutations, as long as the evolutionary distance is properly measured. It recovers the correct tree if the distance matrix satisfies the condition of an additive tree.

**Distance Wagner (DW) Method**

This method was originally presented as a distance version of the Wagner parsimony method (Farris, 1972). Farris, therefore, suggested that a meas-
ure called "metric" (a distance measure that satisfies the principle of the triangle inequality) be used for this method. However, this requirement does not necessarily improve the performance of this method (Nei et al., 1983; Sourdís and Krimbas, 1987). Tateno et al. (1982) modified this method to make it more appropriate for molecular data. Swofford (1981) and Faith (1985) also proposed modified versions of this method. Comparison of different topologies is built in the method, and the final tree that is supposed to be the best one is automatically produced. This method gives both topology and branch lengths of the final tree.

**Neighborliness (ST) Method**

The principle of the neighborliness (ST) method (Sattath and Tversky, 1977; Fitch, 1981) is to use Buneman's (1971) four-point condition for an additive tree. Consider a tree with \( n \) (\( \geq 4 \)) OTUs and assume that OTUs 1 and 2 are a pair of neighbors, that is, two OTUs connected through a single interior node in a bifurcating tree (see Saitou and Nei, 1987). Let \( d_{ij} \) be the distance between OTUs \( i \) and \( j \). We then have the following inequalities for an additive tree.

\[
d_{12} + d_{ij} < d_{1i} + d_{2j}, \quad d_{12} + d_{ij} < d_{1j} + d_{2i}, \tag{4}
\]

where \( i \) and \( j \) are any OTUs (3 \( \leq i \leq j \leq n \)). For actual data, the above condition may not hold because of disturbance of additivity of distances. Sattath and Tversky (1977), and Fitch (1981) then proposed algorithms to construct a tree by maximizing the number of cases in which condition (4) holds. In practice, it is easier to use the ST algorithm of Sattath and Tversky (1977).

**Neighbor-Joining (NJ) Method**

The principle of this method is the same as that of the ME method, but the computational process is much simpler (Saitou and Nei, 1987). Comparison of different topologies is built in the algorithm, and the final tree with both topology and branch lengths is automatically produced. Swofford and Olson (1990) recently stated that this method is for estimating an additive tree, but this statement is incorrect.

**Discrete-Character Data**

**Maximum Parsimony (MP) Method**

In this method, the DNA (or amino acid) sequences of ancestral species are inferred from those of extant species, considering a particular tree topology, and the minimum number of evolutionary changes that are required to explain all the observed differences among the sequences is computed. This number is obtained for all possible topologies, and the topology which shows the smallest number of evolutionary changes is chosen as the final tree. This method is used mainly for finding the topology of a tree, but branch lengths can be estimated under certain assumptions (Fitch, 1971).
When the MP method is applied to morphological characters, it is customary to assume that the primitive and derived character-states are known. In the case of molecular data, this assumption generally does not hold, and different character-states are often reversible. It is, therefore, important to use the MP method, which permits reversible mutations (Eck and Dayhoff, 1966; Fitch, 1971). In numerical taxonomy, this type of MP method is sometimes called the Wagner parsimony method (Farris, 1970).

**Evolutionary Parsimony (EP) Method**

This method is primarily applied to four DNA (or RNA) sequences and utilizes information on the transition/transversion bias in nucleotide substitution (Lake, 1987). The actual procedure is to compute three quantities, $X$, $Y$, and $Z$, which are functions of the numbers of certain nucleotide configurations among the four DNA sequences, and to determine which of the three quantities is significantly different from 0. If only one of them is significant, the tree topology corresponding to the quantity is regarded as the correct one. If two or all of $X$, $Y$, and $Z$ are significant, the splitting pattern of DNA sequences is considered unresolvable. A procedure for extending this method to the case of five or more sequences is presented by Lake (1988).

**Maximum Likelihood (ML) Method**

In this method, the nucleotides of all DNA sequences at each nucleotide site are considered separately, and the log-likelihood of having these nucleotides are computed for a given topology by using a particular probability model (Felsenstein, 1981a). This log-likelihood is added for all nucleotide sites, and the sum of the log-likelihood is maximized to estimate the branch length of the tree. This procedure is repeated for all possible topologies, and the topology that shows the highest likelihood is chosen as the final one.

**METHODS OF COMPUTER SIMULATION**

**Generation of Simulated Sequences**

In the case of DNA sequences, we know the basic rules of their evolutionary change (Nei, 1987a: 79–88). There are two types of changes (i.e., nucleotide substitution and deletion/insertion). The latter type of change occurs haphazardly, and it is difficult to quantify the amount of change for a given evolutionary time. Therefore, deletions and insertions are often neglected in the reconstruction of phylogenetic trees. By contrast, the pattern of nucleotide substitution has been studied extensively, and there are several mathematical models that predict the evolutionary change of DNA sequences. These models are, of course, approximations to real changes of DNA sequences, but their predictions are quite accurate unless the number
of nucleotide substitutions per site is very high. Therefore, it is possible to simulate the evolutionary change of DNA sequences.

The basic information for modeling the evolutionary change of DNA sequences is the probability of change of a nucleotide (A, T, C, or G) to another nucleotide during a short period of time (e.g., one year, one generation, 1,000 years, etc.). Let $\lambda_{ij}$ be the probability that the $i$-th nucleotide changes to the $j$-th nucleotide during the unit evolutionary time. We then have the following transition matrix:

$$
M = \begin{bmatrix}
1 - \lambda_1 & \lambda_{12} & \lambda_{13} & \lambda_{14} \\
\lambda_{12} & 1 - \lambda_2 & \lambda_{23} & \lambda_{24} \\
\lambda_{13} & \lambda_{23} & 1 - \lambda_3 & \lambda_{34} \\
\lambda_{14} & \lambda_{24} & \lambda_{34} & 1 - \lambda_4
\end{bmatrix}
$$

(5)

where $\lambda_1 = \lambda_{12} + \lambda_{13} + \lambda_{14}$, $\lambda_2 = \lambda_{21} + \lambda_{23} + \lambda_{24}$, $\lambda_3 = \lambda_{31} + \lambda_{32} + \lambda_{34}$, and $\lambda_4 = \lambda_{41} + \lambda_{42} + \lambda_{43}$.

Let $g_t$ be the probability that a given nucleotide site in a DNA sequence is occupied by the $i$-th nucleotide at a given evolutionary time, and $g$ be the column vector of $g_1$, $g_2$, $g_3$, and $g_4$. We designate $g$ at time $t$ by $g_t$. It is then possible to compute $g_t$ by $g_t = M^t g_0$, where $M$ is the matrix given by (5) and $g_0$ is the vector $g$ at time 0. Therefore, if $g_0$ is given for all nucleotide sites of the DNA sequence under consideration, one can generate a DNA sequence at time $t$. Note that all nucleotide changes are probabilistic. Therefore, the sequence generated for time $t$ may vary from replication to replication.

When $\lambda_{ij}$'s are all the same and are equal to $\lambda$, and the initial sequence is a random sequence, the evolutionary change of DNA sequence becomes much simpler. Consider two DNA sequences that diverged $t$ time-units ago. The probability that the two sequences have different nucleotides at a given site is given by

$$
P = \frac{3}{4}(1 - e^{-\lambda t/3})
$$

(6)

(Jukes and Cantor, 1969). Similarly, the probability that a descendant sequence is different from the ancestral sequence at a site becomes

$$
P_A = \frac{3}{4}(1 - e^{-\lambda t/3}).
$$

(7)

Therefore, using pseudorandom numbers, one can easily generate a DNA sequence at any time. We call this the one-parameter model.

Incidentally, equation (6) is useful for deriving an estimator of the total number of nucleotide substitutions per site ($d = 2\lambda t$) between two sequences that diverged $t$ time-units ago. It is given by

$$
d = -\frac{3}{4}\log_e (1 - \frac{4}{3}p).
$$

(8)
where $p$ is the proportion of different nucleotides between the two sequences compared.

Actual data on nucleotide substitution indicate that transitional changes (A $\Rightarrow$ G and T $\Rightarrow$ C) are more frequent than transversional changes (all other changes). Considering this difference, Kimura (1980) developed a two-parameter model in which transitional and transversional changes can be treated separately. If we use this model, the probability that two DNA sequences that diverged $t$ time-units ago show a transitional difference at a given site becomes

$$
P = \frac{1}{4} - \frac{1}{2} e^{-\alpha t} + \frac{1}{4} e^{-\beta t},
$$

whereas the probability that they show a transversional difference is

$$
Q = \frac{1}{2} - \frac{1}{2} e^{-\beta t}.
$$

Here $\alpha$ and $\beta$ denote the rates per unit of evolutionary time of transitional and transversional changes, respectively. These equations can be used for simulating the evolutionary change of DNA sequences. The two-parameter model can also be used for estimating the number of nucleotide substitutions between two sequences. It can be estimated by

$$
\hat{d} = 2(\alpha + \beta) t
$$

$$
= -\frac{1}{2} \log_e [(1 - 2P - Q)\sqrt{1 - 2Q}]
$$

(Kimura, 1980).

The transition probabilities $\lambda_{ij}$'s in equation (5) have been estimated for a number of groups of genes (e.g., Brown et al., 1982; Gojobori et al., 1982; Aquadro and Greenberg, 1983; Graur, 1985; Saitou, 1987). Generally speaking, transitional changes are more frequent than transversional changes, but actual substitution patterns are more complicated than the two-parameter model. For this reason, several more complicated models have been developed (see Nei, 1987a). However, both the one- and two-parameter models are known to be quite robust. For estimating the number of nucleotide substitutions ($d$), equation (8) gives a good estimate, provided that $d$ is less than 0.5 and the transition/transversion bias is not extreme, whereas the two-parameter model seems to work even for a higher value of $d$ if the number of nucleotides examined is large. For this reason, the one- or two-parameter model is often used for computer simulation. If this is not satisfactory, the matrix method presented earlier can be used (Jin and Nei, 1990).

Recent data (Britten, 1986; Li et al., 1987b) suggest that the rate of nucleotide substitution varies with evolutionary lineage. This can easily be simulated by changing $\lambda_{ij}$'s (or $\alpha$ and $\beta$) with evolutionary lineage. Mathematically, however, the same result is obtained by changing $t$ rather than $\lambda_{ij}$'s, and this is usually much simpler.
Another important aspect of nucleotide substitution is that the substitution rate varies among nucleotide sites (Shoemaker and Fitch, 1989). The first and second nucleotide positions of codons are usually less variable than the third positions. Nucleotides in the active center regions of genes are also usually less variable (Kimura, 1983). In some exceptional genes such as major histocompatibility complex genes, the active center evolves faster (Nei and Hughes, 1991). This problem can also be resolved by considering spatially varying models of nucleotide substitution (Gojobori et al., 1982; Tateno et al., 1982; Olson, 1987; Jin and Nei, 1990). A commonly used method is to assume that $\lambda_y$'s (or $\alpha$ and $\beta$) vary according to a gamma-distribution or a lognormal distribution.

**Model Trees and Estimation of the Trees**

As mentioned earlier, if a model tree is given, we can simulate the evolutionary changes of DNA sequences following the tree and generate the present-day sequences. Once these sequences are obtained, a tree is reconstructed by each tree-making method under consideration. In the case of distance methods, the evolutionary distance between each pair of sequences must be computed. This distance can be computed by various methods. When the number of nucleotide differences per site ($p$) is small for all pairs of sequences, this number can be used as a distance measure. We call this the $p$ distance. When $p$ is large, however, $p$ is not a good estimate of the number of nucleotide substitutions ($d$) because of backward and parallel substitutions. In this case, $d$, is often estimated by equation (8). We call this the $d$ distance in this chapter, whereas the $d$ value estimated by equation (11) will be called the Kimura distance. However, there are many other ways to estimate $d$, as mentioned earlier. Obviously, the most desirable distance measure is the one that gives the best estimate of the total number of nucleotide substitutions. In the case of discrete-character models, the configuration of nucleotides among all sequences is considered at each nucleotide site.

Since the evolutionary change of nucleotides is stochastic, the DNA sequence generated for a given species varies from replication to replication. Therefore, computer simulation is repeated many times for a given model tree, and the agreement between reconstructed trees and the model tree is statistically determined.

There are two types of deviations of a reconstructed tree from the model tree. One is topological error, and the other is the deviation of estimated branch lengths from the true lengths. Topological error can be evaluated by two measures. One is the proportion of replications in which the correct topology is obtained. This is called the probability of obtaining the correct topology ($P_2$). The other measure is the topological distance between a reconstructed tree and the model tree. This distance, Tateno et al.'s (1982) distortion index ($d_P$), is measured by Robinson and Foulds' (1981) index. Roughly speaking, this index is twice the number of branch interchanges.
that are required to transform the topology of a tree to that of another. When simulation is replicated, the mean topological distance \( \bar{d}_T \) for all replications is used as a criterion. In general, \( \bar{d}_T \) is highly negatively correlated to \( P_e \). Therefore, many authors have used \( P_e \) only.

Branch-length errors can also be measured by several different methods (Tateno et al., 1982). In this chapter, however, this problem will not be considered.

RESULTS FROM COMPUTER SIMULATION

As mentioned earlier, relative efficiencies of tree-making methods have been studied by many authors. Each author or group of authors has studied a different set of tree-making methods. However, several tree-making methods were examined repeatedly by different groups of authors; in these cases the results obtained were usually consistent. In this review, the general conclusions obtained from these studies (without going into details) are presented.

Computer simulation of the reconstruction (estimation) of phylogenetic trees is quite time consuming, particularly when the number of DNA sequences is large. Therefore, the number of DNA sequences used is usually six to eight, though some authors (e.g., Tateno et al., 1982) examined up to 32 sequences. The shape (topology and branch lengths) of the model tree examined varies among authors, but they can be divided roughly into the two types (A and B) given in Figure 6-1. Some methods show a better performance in topology A, and others in topology B. In general, however,

![Figure 6-1](image)

Figure 6-1 Model trees (A) and (B) under the assumption of constant rate of nucleotide substitution. \( U \) is the expected number of substitutions per site from the common ancestral sequence to the extant sequences.
the relative merits of different tree-making methods are similar for both topologies (A and B).

In the following discussion, the cases of constant and varying rates of nucleotide substitution will be considered separately.

Constant Rate of Substitution

Number of Nucleotide Substitutions—Effect on Recovery of Correct Topology

When the expected number of nucleotide substitutions per site from the common ancestral sequence to the extant sequence (U) is small (see Fig. 6-1), the FM, DW, and Tateno et al.'s modified Farris (MF) methods are more efficient than UPGMA in recovering the correct topology. However, when U is large and each branch length is large, UPGMA is better than the former methods.

This was first shown by Tateno et al. (1982), but more general results were obtained by Sourdis and Krimbas (1987). Table 6-1 shows the probabilities of obtaining the correct tree (Pc) for various tree-making methods when model tree A in Figure 6-1 is used. When the number of nucleotides used is small and U is small, any method gives a very small Pc value, as expected. This is true whether rooted or unrooted trees are constructed. However, as long as U is equal to 0.1 or less, the FM, DW, and MF methods are better than UPGMA in obtaining the correct rooted or unrooted tree. (The root for the trees obtained by the FM, DW, and MF methods was given as the midpoint of the longest route connecting two sequences.) The former three methods show more or less the same Pc value. When U is large (e.g., U = 0.4), however, UPGMA often shows a higher Pc value than the other methods. Somewhat similar results were also obtained by Blanken et al. (1982). Table 6-1 shows that Pc's for rooted trees are considerably smaller than those for unrooted trees, particularly when the FM, DW, and MF methods are used. This indicates that a substantial amount of error in constructing a rooted tree occurs at the time of rooting.

Comparison of Efficiency of ST and NJ Methods with Other Methods

The ST and NJ methods are almost always better than the UPGMA, DW, MF, and TD methods whether the U value is small or large.

Table 6-2 shows the results of another computer simulation for a different set of tree-making methods. This represents the case where U is relatively small (U = 6a + b = 0.1 in model tree A, and U = 1.5a + c = 0.085 in model tree B of Figure 6-1). In this case, the MF and DW methods are again better than UPGMA in obtaining the correct tree. When tree A is used, they are as efficient as the TD (Li’s algorithm), ST, and NJ methods. However, when tree B is used, they are less efficient than the latter methods. Table 6-3 shows the Pc values for the case of a larger U value (U = 0.52 in tree A, and U = 0.465 in tree B) and a larger number of nucleotides
Table 6-1  Probabilities of Obtaining the Correct Topology \( (P_c \times 100) \) for Rooted and Unrooted Trees for Four Different Tree-Making Methods in the Case of \( a = b \) in Model Tree A of Figure 6-1*

<table>
<thead>
<tr>
<th>U</th>
<th>R</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>UPGMA</td>
<td>FM</td>
<td>DW</td>
<td>MF</td>
<td>UPGMA</td>
<td>FM</td>
</tr>
<tr>
<td>0.03</td>
<td>250</td>
<td>4</td>
<td>10</td>
<td>6</td>
<td>10</td>
<td>9</td>
<td>24</td>
</tr>
<tr>
<td>0.05</td>
<td>275</td>
<td>12</td>
<td>19</td>
<td>15</td>
<td>17</td>
<td>23</td>
<td>41</td>
</tr>
<tr>
<td>0.10</td>
<td>275</td>
<td>34</td>
<td>42</td>
<td>24</td>
<td>39</td>
<td>51</td>
<td>70</td>
</tr>
<tr>
<td>0.20</td>
<td>375</td>
<td>56</td>
<td>55</td>
<td>41</td>
<td>52</td>
<td>67</td>
<td>79</td>
</tr>
<tr>
<td>0.40</td>
<td>175</td>
<td>69</td>
<td>53</td>
<td>40</td>
<td>50</td>
<td>83</td>
<td>76</td>
</tr>
<tr>
<td>m = 300</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td>275</td>
<td>38</td>
<td>43</td>
<td>42</td>
<td>44</td>
<td>48</td>
<td>82</td>
</tr>
<tr>
<td>0.05</td>
<td>275</td>
<td>60</td>
<td>64</td>
<td>61</td>
<td>64</td>
<td>74</td>
<td>94</td>
</tr>
<tr>
<td>0.10</td>
<td>275</td>
<td>88</td>
<td>84</td>
<td>75</td>
<td>84</td>
<td>92</td>
<td>99</td>
</tr>
<tr>
<td>0.20</td>
<td>250</td>
<td>96</td>
<td>91</td>
<td>85</td>
<td>91</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>0.40</td>
<td>100</td>
<td>99</td>
<td>95</td>
<td>83</td>
<td>95</td>
<td>100</td>
<td>99</td>
</tr>
<tr>
<td>m = 900</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.03</td>
<td>275</td>
<td>60</td>
<td>66</td>
<td>65</td>
<td>66</td>
<td>74</td>
<td>95</td>
</tr>
<tr>
<td>0.05</td>
<td>275</td>
<td>84</td>
<td>84</td>
<td>81</td>
<td>84</td>
<td>88</td>
<td>99</td>
</tr>
<tr>
<td>0.10</td>
<td>275</td>
<td>92</td>
<td>94</td>
<td>90</td>
<td>94</td>
<td>95</td>
<td>100</td>
</tr>
<tr>
<td>0.20</td>
<td>275</td>
<td>99</td>
<td>99</td>
<td>93</td>
<td>99</td>
<td>99</td>
<td>100</td>
</tr>
<tr>
<td>0.40</td>
<td>125</td>
<td>100</td>
<td>100</td>
<td>95</td>
<td>95</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>m = 1,500</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*The abbreviations used are: U, expected number of nucleotide substitutions per site from the ancestral sequence to the extant sequence; m, number of nucleotides used; R, number of replications; FM, Fitch and Margoliash's method; DW, distance Wagner method; MF, modified Farris method. Trees were constructed by using the d distance. (Adapted from Souris and Krumbach, 1987.)
<table>
<thead>
<tr>
<th></th>
<th>Model Tree A</th>
<th></th>
<th>Model Tree B</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>300</td>
<td>600</td>
<td>900</td>
<td>300</td>
</tr>
<tr>
<td>m</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UP: p</td>
<td>14 (3.18)</td>
<td>36 (1.72)</td>
<td>58 (0.98)</td>
<td>14 (4.54)</td>
</tr>
<tr>
<td>UP: d</td>
<td>15 (3.18)</td>
<td>34 (1.74)</td>
<td>56 (1.04)</td>
<td>13 (4.56)</td>
</tr>
<tr>
<td>MF: p</td>
<td>39 (1.76)</td>
<td>73 (0.58)</td>
<td>95 (0.10)</td>
<td>24 (2.86)</td>
</tr>
<tr>
<td>MF: d</td>
<td>38 (1.92)</td>
<td>72 (0.62)</td>
<td>95 (0.10)</td>
<td>19 (2.94)</td>
</tr>
<tr>
<td>DW: p</td>
<td>42 (1.70)</td>
<td>75 (0.54)</td>
<td>96 (0.08)</td>
<td>26 (2.36)</td>
</tr>
<tr>
<td>DW: d</td>
<td>37 (1.74)</td>
<td>74 (0.58)</td>
<td>95 (0.10)</td>
<td>28 (2.36)</td>
</tr>
<tr>
<td>TD: p</td>
<td>41 (1.58)</td>
<td>71 (0.70)</td>
<td>94 (0.12)</td>
<td>40 (2.04)</td>
</tr>
<tr>
<td>TD: d</td>
<td>36 (1.84)</td>
<td>66 (0.82)</td>
<td>89 (0.24)</td>
<td>39 (2.10)</td>
</tr>
<tr>
<td>ST: p</td>
<td>48 (1.26)</td>
<td>75 (0.54)</td>
<td>97 (0.06)</td>
<td>45 (1.66)</td>
</tr>
<tr>
<td>ST: d</td>
<td>44 (1.48)</td>
<td>70 (0.62)</td>
<td>96 (0.08)</td>
<td>43 (1.62)</td>
</tr>
<tr>
<td>NJ: p</td>
<td>48 (1.36)</td>
<td>76 (0.54)</td>
<td>97 (0.06)</td>
<td>46 (1.64)</td>
</tr>
<tr>
<td>NJ: d</td>
<td>41 (1.60)</td>
<td>70 (0.62)</td>
<td>96 (0.08)</td>
<td>45 (1.62)</td>
</tr>
</tbody>
</table>

*The abbreviations used are: UP, UPGMA; MF, modified Farris method; DW, distance Wagner method; TD, transformed distance method; LJ's algorithm was used; ST, Saitou and Nei's method; NJ, neighbor-joining method. p, trees reconstructed from p distances; d, trees reconstructed from d distances; m, number of nucleotides used. Number of replications (R) is 100. (Adapted from Saitou and Nei, 1987.)
<table>
<thead>
<tr>
<th>m</th>
<th>500</th>
<th>1,000</th>
<th>2,000</th>
<th>500</th>
<th>1,000</th>
<th>2,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>UP: p</td>
<td>9 (3.78)</td>
<td>27 (2.10)</td>
<td>62 (0.86)</td>
<td>10 (5.20)</td>
<td>18 (3.76)</td>
<td>54 (1.32)</td>
</tr>
<tr>
<td>UP: d</td>
<td>9 (3.78)</td>
<td>27 (2.10)</td>
<td>62 (0.88)</td>
<td>11 (5.30)</td>
<td>18 (3.74)</td>
<td>55 (1.26)</td>
</tr>
<tr>
<td>MF: p</td>
<td>15 (4.02)</td>
<td>41 (1.82)</td>
<td>62 (0.92)</td>
<td>3 (5.68)</td>
<td>17 (3.64)</td>
<td>28 (2.40)</td>
</tr>
<tr>
<td>MF: d</td>
<td>13 (4.42)</td>
<td>34 (2.14)</td>
<td>55 (1.14)</td>
<td>3 (5.72)</td>
<td>13 (3.80)</td>
<td>26 (2.48)</td>
</tr>
<tr>
<td>DW: p</td>
<td>16 (3.78)</td>
<td>46 (1.54)</td>
<td>63 (0.82)</td>
<td>4 (5.42)</td>
<td>18 (3.28)</td>
<td>41 (1.72)</td>
</tr>
<tr>
<td>DW: d</td>
<td>15 (4.22)</td>
<td>40 (1.96)</td>
<td>58 (0.98)</td>
<td>5 (5.50)</td>
<td>18 (3.48)</td>
<td>35 (1.82)</td>
</tr>
<tr>
<td>TD: p</td>
<td>3 (4.26)</td>
<td>37 (2.00)</td>
<td>53 (1.18)</td>
<td>15 (4.48)</td>
<td>28 (2.98)</td>
<td>70 (0.90)</td>
</tr>
<tr>
<td>TD: d</td>
<td>3 (4.84)</td>
<td>25 (2.60)</td>
<td>39 (1.66)</td>
<td>12 (4.72)</td>
<td>27 (3.06)</td>
<td>66 (1.02)</td>
</tr>
<tr>
<td>ST: p</td>
<td>10 (3.56)</td>
<td>44 (1.62)</td>
<td>68 (0.76)</td>
<td>13 (4.00)</td>
<td>36 (2.34)</td>
<td>74 (0.62)</td>
</tr>
<tr>
<td>ST: d</td>
<td>6 (4.06)</td>
<td>40 (1.82)</td>
<td>56 (1.04)</td>
<td>10 (4.32)</td>
<td>34 (2.34)</td>
<td>71 (0.72)</td>
</tr>
<tr>
<td>NJ: p</td>
<td>11 (3.70)</td>
<td>44 (1.68)</td>
<td>67 (0.80)</td>
<td>13 (4.46)</td>
<td>34 (2.38)</td>
<td>75 (0.62)</td>
</tr>
<tr>
<td>NJ: d</td>
<td>5 (4.24)</td>
<td>38 (2.00)</td>
<td>57 (1.06)</td>
<td>14 (4.44)</td>
<td>32 (2.42)</td>
<td>73 (0.72)</td>
</tr>
</tbody>
</table>

*Abbreviations used are the same as those in Table 6-2. R = 100. (Adapted from Saitou and Nei, 1987.)
examined. In this case, the MF and DW methods are as efficient as the ST and NJ methods when tree A is used, but are less efficient than the latter when tree B is used. The TD method is slightly less efficient than the ST and NJ methods for both trees A and B.

It is known that the principle of triangle inequality applies to the $p$ distance but not to the $d$ distance. Therefore it is interesting to compare the $P_c$ values for the trees obtained by using the $p$ and $d$ distances. Tables 6-2 and 6-3 show that there is no real difference in $P_c$ between the $p$ and $d$ distances when UPGMA is used. In other tree-making methods, the $p$ distance tends to give a slightly higher $P_c$ than the $d$ distance. However, the difference between them is generally small, except when the number of nucleotides examined ($m$) is small. As will be shown later, when the rate of nucleotide substitution varies with evolutionary lineage, and $p$ or $d$ is large, the $d$ distance shows a better performance than the $p$ distance.

**Comparison of the MP Method with the ST, NJ, DW, and TD Methods**

The MP method generally has a smaller $P_c$ value than the ST and NJ methods. However, when $U$ is small and the number of nucleotides examined is very large, the MP method is as good as or slightly better than the latter.

As mentioned earlier, the MP method requires examination of all possible topologies, and since the number of possible topologies rapidly increases as the number of OTUs increases, computer simulation has been done for a relatively small number of OTUs. For example, the number of possible topologies is 105 when the number of OTUs is six, whereas it is 10,395 when the number is eight. Table 6-4 shows the $P_c$ values for the model trees A and B in Figure 6-2. When model tree A is used and the number of nucleotide substitutions is small ($U = 0.05$), the NJ and TD methods are better than the MP and DW methods provided that the number of nucleotides examined is $\leq 600$. When the number of nucleotides is large

<table>
<thead>
<tr>
<th>$m$</th>
<th>300</th>
<th>600</th>
<th>1,200</th>
<th>300</th>
<th>600</th>
<th>1,200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constant Rate (Tree A)</td>
<td>Varying Rate (Tree B)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>600</td>
<td>1,200</td>
<td>300</td>
<td>600</td>
<td>1,200</td>
</tr>
<tr>
<td>U = 0.05</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MP</td>
<td>46.3</td>
<td>79.0</td>
<td>97.0</td>
<td>16.7</td>
<td>39.3</td>
<td>58.0</td>
</tr>
<tr>
<td>NJ</td>
<td>58.7</td>
<td>82.0</td>
<td>93.5</td>
<td>30.0</td>
<td>46.7</td>
<td>63.3</td>
</tr>
<tr>
<td>DW</td>
<td>47.7</td>
<td>72.0</td>
<td>92.0</td>
<td>19.2</td>
<td>35.0</td>
<td>52.0</td>
</tr>
<tr>
<td>TD</td>
<td>54.7</td>
<td>82.0</td>
<td>95.0</td>
<td>25.3</td>
<td>42.3</td>
<td>62.3</td>
</tr>
<tr>
<td>U = 0.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MP</td>
<td>44.7</td>
<td>69.3</td>
<td>84.0</td>
<td>16.7</td>
<td>22.3</td>
<td>31.0</td>
</tr>
<tr>
<td>NJ</td>
<td>62.7</td>
<td>84.3</td>
<td>97.0</td>
<td>25.0</td>
<td>34.7</td>
<td>49.3</td>
</tr>
<tr>
<td>DW</td>
<td>49.0</td>
<td>73.0</td>
<td>91.3</td>
<td>16.0</td>
<td>30.7</td>
<td>40.0</td>
</tr>
<tr>
<td>TD</td>
<td>64.3</td>
<td>84.3</td>
<td>96.0</td>
<td>25.0</td>
<td>33.3</td>
<td>48.0</td>
</tr>
</tbody>
</table>

*Abbreviations used are the same as those in Table 6-2. MP, maximum parsimony. $R = 300$. (Adapted from Soursis and Nei, 1986)*
(m = 1.200), however, the MP method seems to be as good as or slightly better than the other methods.

When the number of substitutions per site is large (i.e., U = 0.5), the story is different. In this case, even for m = 1,200, the MP method shows a smaller \( P_e \) value than the NJ and DW methods. This is apparently because there are many backward and parallel substitutions in this case. Distance methods are not affected by these substitutions as long as the distances are correctly estimated.

The low performance of the MP method when m and U are both small is due to the fact that in this method only so-called "informative sites" are used, whereas in distance methods information for all variable sites is used. (Note that noninformative variable sites in parsimony methods are actually informative for constructing a tree in distance methods). When m and U are small, the number of "informative sites" is also small. Therefore, the MP method tends to produce many equally parsimonious trees, which often (but not always) include the correct tree (Soursis and Nei, 1988). In our study, the case where the correct tree and other tied trees were obtained was not included in the probability of obtaining the correct tree. This is part of the reason why \( P_e \) became smaller in the MP method than in the NJ and other methods. When m is large, however, the number of "informative sites" becomes large, thus \( P_e \) increases.

In this connection, one might argue that \( P_e \) is not an appropriate measure for comparing the MP method with others because many tied trees, including the correct one, are often produced in this method. When there are many tied trees, it is customary to construct a consensus tree with multifurcating branches in parsimony methods. A multifurcating consensus tree can also be constructed with distance methods (e.g., Tajima, 1990). However, a multifurcating tree is clearly wrong in the present case, and it is not clear whether the comparison of consensus trees obtained by parsimony and distance methods is actually meaningful.

When there are many tied trees obtained, a more reasonable way of comparison is to compare \( \bar{d}_r \) for the MP and other methods. If many tied
RELATIVE EFFICIENCIES OF DIFFERENT TREE-MAKING METHODS

trees are close to the correct one, $d_r$ is expected to be close to 0; otherwise it will be high. Since Sourdies and Nei did not compute $d_r$, Li Jin and I performed a small-scale computer simulation to evaluate this value using model tree A in Figure 6-2. The results obtained for the MP and NJ methods are presented in Table 6-5. The $P_e$ values in this table are essentially the same as those for the corresponding cases of Table 6-4. Table 6-5 also includes the probability of obtaining the correct tree and other tied trees ($P_r$) and the probability of obtaining only incorrect tree(s) ($P_s$). As in Sourdies and Neis's study, $P_r$ for the MP method is large when $m$ is small but decreases as $m$ increases. For the NJ method, it is always 0. This result is slightly different from that of Sourdies and Nei (1988), because in their simulation some numbers with many digits were rounded to save computer time.

Table 6-5 shows that even if we use $d_r$ as the criterion of comparison, the MP method is generally inferior to the NJ method. The only exceptions are the cases of $m = 600$ and $m = 1,200$ with $U = 0.05$. Therefore, our conclusion remains the same, whether we use $P_e$ or $d_r$ as the criterion.

Comparison of ML Method with the MP, FM, ME, and NJ Methods

The ML method is generally more efficient than the MP and FM methods but is slightly less efficient than the ME and NJ methods.

The $P_e$ values for the ML method for model trees A and B in Figure 6-3 are presented in Table 6-6 in comparison with those for the MP, FM, ME, and NJ methods. The trees for the distance methods were constructed by using both $p$ and $d$ distances. It is clear that the FM method is poorest in obtaining the correct tree among all the methods examined here. The ML method is better than the MP method except for the case of $U = 0.5$ in tree A, but it is slightly less efficient than the ME and NJ methods when $m = 300$. When $m = 600$, it has a $P_e$ value similar to that of the latter two methods.

Table 6-5  Probabilities of Obtaining the Correct Unrooted Tree ($P_e \times 100$), the Correct Tree and Other Tied Trees ($P_r \times 100$) and Incorrect Trees ($P_s \times 100$), and Average Indices of Topological Errors ($\bar{d}_r$) for the MP and NJ Methods*

<table>
<thead>
<tr>
<th>$m$</th>
<th>300</th>
<th>600</th>
<th>1,200</th>
<th>300</th>
<th>600</th>
<th>1,200</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$U = 0.05$</td>
<td></td>
<td></td>
<td>$U = 0.5$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$P_e$</td>
<td>44</td>
<td>79</td>
<td>96</td>
<td>40</td>
<td>63</td>
<td>88</td>
</tr>
<tr>
<td>$P_r$</td>
<td>41</td>
<td>14</td>
<td>3</td>
<td>13</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>$P_s$</td>
<td>15</td>
<td>7</td>
<td>1</td>
<td>47</td>
<td>29</td>
<td>10</td>
</tr>
<tr>
<td>$d_r$</td>
<td>0.96</td>
<td>0.31</td>
<td>0.05</td>
<td>1.30</td>
<td>0.72</td>
<td>0.22</td>
</tr>
</tbody>
</table>

* $P_e$ is the same as $P_e$ of Sourdies and Nei (1988). Model tree A in Figure 6-2 was used. $R = 300$. 
It should be noted that this computer simulation was conducted under the conditions which are favorable for the ML method, but do not completely satisfy the assumptions of this method (Saitou and Imanishi, 1989; N. Saitou, personal communication). Therefore, it seems that a slightly lower performance of the ML method compared with the ME and NJ methods is due to a small violation of the assumptions of this method. This suggests that the ML method is sensitive to violations of its assumptions, at least in this case.

Incidentally, Table 6-6 shows that the ME and NJ methods give essentially the same $P_*$ value, as mentioned earlier. Indeed, Saitou and Imanishi (1989) showed that these two methods usually reconstruct the same tree from a given set of data. This indicates that the NJ method usually finds the minimum evolution tree, though its procedure is very simple.

Recently, Rohlf and Wooten (1988) reported results of their computer simulations on the gene frequency maximum likelihood (GFML) method (Felsenstein, 1981b). They considered situations where the assumptions of the GFML method hold as closely as possible, yet their results showed that UPGMA is generally better than the GFML method for gene-frequency data. Curiously, however, these authors concluded otherwise, considering the possibility that the performance of the ML method would increase as the number of genetic loci used increases. This conclusion is not warranted because the performance of distance methods also increases as the number of loci increases. A similar computer simulation was also conducted by Kim and Burgman (1988). However, since this simulation was conducted under the same assumptions (Gaussian process without
Table 6-6: Probabilities of Obtaining the Correct Tree ($P_c \times 100$) for the Maximum Likelihood (ML) and Other Tree-Making Methods for Model Trees A and B of Figure 6-3*

<table>
<thead>
<tr>
<th></th>
<th>Model Tree A</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th>Model Tree B</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MP</td>
<td>ML</td>
<td>FM</td>
<td>ME</td>
<td>NJ</td>
<td>FM</td>
<td>ME</td>
<td>NJ</td>
<td>FM</td>
<td>ME</td>
<td>NJ</td>
</tr>
<tr>
<td>$U = 0.05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 300$</td>
<td>34</td>
<td>38</td>
<td>26</td>
<td>36</td>
<td>42</td>
<td>26</td>
<td>40</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 600$</td>
<td>76</td>
<td>80</td>
<td>58</td>
<td>80</td>
<td>78</td>
<td>58</td>
<td>80</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U = 0.50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 300$</td>
<td>60</td>
<td>48</td>
<td>30</td>
<td>56</td>
<td>58</td>
<td>22</td>
<td>42</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 600$</td>
<td>84</td>
<td>70</td>
<td>54</td>
<td>92</td>
<td>92</td>
<td>40</td>
<td>82</td>
<td>82</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U = 0.05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 300$</td>
<td>54</td>
<td>62</td>
<td>42</td>
<td>68</td>
<td>68</td>
<td>40</td>
<td>68</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 600$</td>
<td>84</td>
<td>88</td>
<td>56</td>
<td>82</td>
<td>84</td>
<td>56</td>
<td>80</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$U = 0.50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 300$</td>
<td>48</td>
<td>56</td>
<td>34</td>
<td>70</td>
<td>68</td>
<td>30</td>
<td>60</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$m = 600$</td>
<td>58</td>
<td>76</td>
<td>36</td>
<td>76</td>
<td>76</td>
<td>36</td>
<td>70</td>
<td>70</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Abbreviations used are the same as those of previous tables; ML: maximum likelihood; ME: minimum evolution method. $R = 100$. (Adapted from Saitou and Imanishi, 1989.)

Varying Rate of Substitution

Comparison of UPGMA with Other Tree-Making Methods

When the rate of nucleotide substitution varies with evolutionary lineage, UPGMA is worse than most other tree-making methods.

Since UPGMA requires the assumption of a constant rate of evolution, it is expected to have a poor performance in obtaining the correct tree when the rate varies with evolutionary lineage. Indeed, computer simulations have shown that it is very poor compared with other methods in this case (Table 6-7). The $P_c$ values in this table were obtained by using model trees given in Figure 6-4.

Comparison Among Distance Methods

Among the distance methods available now, the ST, NJ, and ME methods generally show a better performance than other methods. The ST, NJ, and ME methods are nearly equally efficient.

This can be seen from Table 6-7 and Table 6-8. The $P_c$ values in Table 6-8 were obtained by using model trees C and D in Figure 6-3. Note that the total number of nucleotide substitutions between the two most distantly
Table 6-7  Probabilities of Obtaining the Correct Unrooted Trees ($P_e \times 100$) and Average Indices ($d_r$) of Topological Errors (in Parentheses) for the Case of Varying Rate of Nucleotide Substitution*

<table>
<thead>
<tr>
<th></th>
<th>Model Tree A</th>
<th>Model Tree B</th>
</tr>
</thead>
<tbody>
<tr>
<td>UP</td>
<td>0 (8.06)</td>
<td>0 (9.74)</td>
</tr>
<tr>
<td>MF</td>
<td>77 (0.50)</td>
<td>57 (1.46)</td>
</tr>
<tr>
<td>DW</td>
<td>69 (0.72)</td>
<td>59 (1.26)</td>
</tr>
<tr>
<td>TD</td>
<td>46 (1.30)</td>
<td>45 (1.68)</td>
</tr>
<tr>
<td>ST</td>
<td>77 (0.50)</td>
<td>69 (0.82)</td>
</tr>
<tr>
<td>NJ</td>
<td>75 (0.56)</td>
<td>72 (0.78)</td>
</tr>
</tbody>
</table>

*Abbreviations used are the same as those in Table 6-2. Model trees A and B are given in Figure 6-4. $m = 600; R = 100$. $p$ distances were used. (Adapted from Saitou and Nei, 1987.)

related sequences ($d_m$) is relatively small in Table 6-7, whereas Table 6-8 includes both cases of a small $d'_m$ and a large $d_m$. When $d_m$ is small, the difference in $P_e$ between the $p$ and $d$ distances is relatively small, though in the case of model trees C and D of Figure 6-3, the $d$ distance gives a higher $P_e$ value than the $p$ distance (Table 6-8). At any rate, Tables 6-7 and 6-8 show that the $P_e$ values for the ST, NJ, and ME methods are generally higher than other distance methods.

Table 6-8 shows that when $d_m$ is large, the $d$ distance gives a higher $P_e$ value than the $p$ distance. This is because the $p$ distance is seriously affected by backward and parallel substitutions, whereas the $d$ distance gives an appropriate estimate of the total number of substitutions. As mentioned earlier, many distance methods are capable of recovering the correct tree as long as the total number of substitutions is correctly estimated.

Comparison of the MP Method with Other Methods

The MP method is worse than several distance methods such as the NJ, ST, and ME methods and the ML method.

---

**Figure 6-4** Model trees used for the simulations in Table 6-7. The value given to each branch represents the expected number of nucleotide substitutions per site for that branch.
Table 6-8  Probabilities of Obtaining the Correct Unrooted Tree ($P_e \times 100$) for the MP, ML, and Other Methods in the Case of Varying Substitution Rate

<table>
<thead>
<tr>
<th></th>
<th>p Distance</th>
<th></th>
<th>d Distance</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MP</td>
<td>ML</td>
<td>FM</td>
<td>ME</td>
<td>NJ</td>
<td>FM</td>
</tr>
<tr>
<td>Model Tree C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300 bp</td>
<td>64</td>
<td>78</td>
<td>26</td>
<td>56</td>
<td>56</td>
<td>34</td>
</tr>
<tr>
<td>600 bp</td>
<td>90</td>
<td>98</td>
<td>42</td>
<td>80</td>
<td>80</td>
<td>68</td>
</tr>
<tr>
<td>$a = 0.05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300 bp</td>
<td>24</td>
<td>92</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>600 bp</td>
<td>20</td>
<td>100</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>Model Tree D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$a = 0.01$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300 bp</td>
<td>68</td>
<td>80</td>
<td>44</td>
<td>64</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>600 bp</td>
<td>94</td>
<td>96</td>
<td>56</td>
<td>90</td>
<td>90</td>
<td>88</td>
</tr>
<tr>
<td>$a = 0.05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>300 bp</td>
<td>25</td>
<td>96</td>
<td>0</td>
<td>6</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>600 bp</td>
<td>46</td>
<td>100</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>64</td>
</tr>
</tbody>
</table>

*Abbreviations used are the same as those in previous tables; bp, base pairs. $R = 50$. Model trees C and D are those in Figure 6-3. (Adapted from Saitou and Imanishi, 1989.)

This statement is supported by the "varying rate" part of Table 6-4 and Table 6-8. In all cases examined here, the MP method always gives a smaller $P_e$ value than the NJ, ME, and ML methods. However, it is better than the FM method when $d_e$ is small ($a = 0.01$). The inferiority of the MP method is explained by Felsenstein's (1978) finding that when evolutionary rate varies with lineage, two OTUs that have long branches tend to form a cluster in the MP method even if this is not the correct branching pattern. This does not happen in the NJ and ME methods as long as the distances between sequences are correctly estimated. This also generally does not occur with the ML method (Hasegawa and Yano, 1984; Saitou, 1988; Saitou and Imanishi, 1989).

Some authors (e.g., Sober, 1985) claimed that the MP method is better than other methods because no assumption is necessary about the process of evolutionary change. This claim is not supported by simulation studies. As mentioned earlier, for the MP method to work well, approximate constancy of evolutionary rate and a small number of nucleotide substitutions per site are necessary.

Comparison of the ML Method with the MP, ME, and NJ Methods

The ML method is generally better than the MP method and is nearly the same as or slightly better than the ME and NJ methods in obtaining the correct tree. However, the performance of this method should be investigated more carefully by considering the cases where the underlying assumptions are violated.

Table 6-8 clearly shows that the ML method is better than the MP method when the evolutionary rate varies extensively. A similar result was also obtained by Hasegawa and Yano (1984), and Saitou (1988) for the case of four sequences. Table 6-8 also shows that the ML method is slightly better
than the ME and NJ methods. This is so despite the fact that the pattern of nucleotide substitution simulated by Saitou and Imanishi (1989) was slightly different from the assumptions of the ML method used.

Nevertheless, it should be noted that the actual pattern of nucleotide substitution is much more complicated than the simple model used in the ML method (Gojobori et al., 1982b; Jin and Nei, 1990), so the general applicability of the above conclusion is questionable. Since the ML method is expected to be more sensitive to violation of the assumptions than are some distance methods, it is hoped that this method will be examined more carefully, considering realistic patterns of nucleotide substitution, particularly variation in substitution rate among different nucleotide sites. Previously, Saitou (1988) showed that in the case of $a = c = 1.0$ and $b = d = 0.1$ in model tree G of Figure 6-5, the NJ method ($P_e = 0.74$) is significantly better than the ML method ($P_e = 0.43$). Recently, however, Hasegawa and Saitou (personal communication) improved the performance of the ML method using a slightly different probability model. This indicates that it is very important to use a proper probability model in the ML method.

In the case of distance methods, the problem of substitution pattern becomes important when one estimates the distances between different sequences rather than when a tree is estimated from distance data. As long as the total number of nucleotide substitutions is correctly estimated, many distance methods are capable of producing correct trees. There are several statistical methods of estimating distance that are quite robust unless the distances are very large (see Nei, 1987a: 72). This problem will be discussed later.

**Comparison of the EP Method with the MP and NJ Methods**

The EP method is inferior to the MP method when nucleotide substitution occurs at random among the four nucleotides (one-parameter model) and

![Figure 6-5](image)

Figure 6-5 Unrooted model trees for four DNA sequences. G: general case. $S_1$, $S_2$, $S_3$: model trees used by Jin and Nei's (1990) computer simulation. The value given to each branch represents the expected number of nucleotide substitutions per site for that branch.
the rate of nucleotide substitution is nearly constant. However, if there is a transition/transversion bias and the rate of substitution varies extensively with evolutionary lineage, the EP method is better than the MP method. The NJ method is almost always superior to the EP method if a proper distance measure is used.

As mentioned earlier, the EP method was developed primarily to be applied to the case of four sequences. Therefore, all computer simulations have been performed for this case. The model tree used is generally of the form given in Figure 6-5. Table 6-9 shows the $P_e$ values for the MP, NJ, and EP methods for the case of Kimura's two-parameter model with model trees $S_1$, $S_2$, and $S_3$ in Figure 6-5. Kimura's model satisfies the assumption of the EP method (Cavender, 1989; Jin and Nei, 1990). Therefore, the EP method is supposed to show a good performance in this case. There are three different sets of $P_e$'s for the NJ method: NJP, NJD, and NJK represent the cases where the $p$ distance, $d$ distance, and Kimura distance [equation (11)], respectively, are used. Since Kimura's two-parameter model was used in generating DNA sequences, NJK is expected to show a better performance than NJP and NJD except when the proportion of transitional changes $B = \alpha / (\alpha + 2\beta)$ is $1/2$. This is indeed the case for almost all parameter sets examined in Table 6-9.

Comparison of the MP and EP methods in Table 6-9 shows that when all branch lengths of the model trees are relatively short ($S_1$), or when the ratio of branch $b$ to branch $a$ is relatively small ($S_2$), the former is better than the latter even for the case of a high proportion of transitional changes ($B = 0.9$). Otherwise ($S_3$), the latter is better than the former. Essentially the same results were obtained for various other sets of parameters (Li et al., 1987a; Jin and Nei, 1990). The performance of the NJ method is expected to depend on the distance measure used, as mentioned earlier. When the appropriate distance (Kimura distance for this case) is used, the NJ method is better than both the MP and EP methods. The same results were obtained for several other sets of parameters as well as for the set of substitution rates estimated from actual data (not the two-parameter model) (Jin and Nei, 1990).

Recently, Sidorow and Wilson (1990) modified the EP method, taking into account the unequal frequencies of the four nucleotides (A, T, C, and G) in the sequences. However, since the basic principle of the EP method remains unchanged, our conclusion is expected to apply to this modified version as well.

Effects of Variation in Substitution Rate Among Different Nucleotide Sites

So far we have assumed that every nucleotide site evolves independently and that the pattern of nucleotide substitution is the same for all sites. This assumption is obviously not satisfied with actual data. In the case of protein-coding genes, the third nucleotide positions of codons usually evolve faster
<table>
<thead>
<tr>
<th>Model Trees</th>
<th>B</th>
<th>MP</th>
<th>NJP</th>
<th>NJD</th>
<th>NJK</th>
<th>EP</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>S_1^t</td>
<td>0.33</td>
<td>100</td>
<td>98</td>
<td>100</td>
<td>100</td>
<td>99</td>
<td>68</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0.60</td>
<td>100</td>
<td>97</td>
<td>100</td>
<td>100</td>
<td>91</td>
<td>45</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>88</td>
<td>85</td>
<td>99</td>
<td>100</td>
<td>61</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>S_1^t</td>
<td>0.33</td>
<td>91</td>
<td>88</td>
<td>100</td>
<td>100</td>
<td>63</td>
<td>12</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>0.60</td>
<td>86</td>
<td>84</td>
<td>98</td>
<td>97</td>
<td>51</td>
<td>6</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>67</td>
<td>69</td>
<td>84</td>
<td>86</td>
<td>40</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>S_1^t</td>
<td>0.33</td>
<td>4</td>
<td>0</td>
<td>98</td>
<td>98</td>
<td>72</td>
<td>26</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>0.60</td>
<td>2</td>
<td>0</td>
<td>92</td>
<td>95</td>
<td>57</td>
<td>13</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>0</td>
<td>0</td>
<td>22</td>
<td>87</td>
<td>42</td>
<td>6</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

*Abbreviations used are the same as those of previous Tables. NJP and NJD, neighbor-joining method with p and d distances, respectively; NJK, neighbor-joining with Kimura distance; EP, evolutionary parsimony; B, proportion of transitional changes \(\alpha(\alpha + 2d)\). EP(\(\chi^2\)) shows the proportion of replications in which only the \(\chi^2\) for X, Y, or Z was greater than 3.84. m = 1,000; R = 400.

\(^a\)a = c = e = 0.05 and b = d = 0.25.

\(^b\)a = c = 0.20, b = d = 0.40, and e = 0.05.

\(^c\)a = c = e = 0.05 and b = d = 0.5.

(Adapted from Jin and Nci, 1990.)
than the first and second positions, and codons encoding active centers of proteins evolve at a slower rate than other codons except in certain special genes (see Nei, 1987a; Hughes and Nei, 1988). Examining the pattern of amino acid substitution in cytochrome c, Uzzell and Corbin (1971) showed that the substitution rate varies from site to site, roughly following a gamma-distribution. This distribution is close to the lognormal distribution for certain parameter values (see Fig. 6-6).

Therefore it is important to examine the effects of variation in substitution rate among different sites on the efficiencies of different tree-making methods. Table 6-10 shows the results of one such study. In this case, substitution rate was assumed to vary according to the lognormal distribution with parameter $\alpha = 8$ (Olsen, 1987). The expected branch lengths used were the same as those for model trees $S_1$, $S_2$, and $S_3$ of Figure 6-5. Table 6-10 indicates that the MP method is better than the EP method for trees $S_1$ and $S_2$, but is worse than the latter for tree $S_3$. Therefore, the conclusion remains the same. The NJ method with the $d$ distance and Kimura distance also shows a better performance than either the MP or EP method for $S_1$ and $S_2$. For $S_3$, however, the EP method now shows a better performance.

The poor performance of the NJ method for $S_3$ is caused by the fact that neither the $d$ distance nor the Kimura distance is an additive measure of the number of nucleotide substitutions in this case. However, an approx-

---

**Figure 6-6** Gamma-distribution with $\alpha = 1$ (---) and $\alpha = 2$ (-----) and the lognormal distribution with $\alpha = 8$ (——). $X =$ substitution rate per site.
Table 6-10  Probabilities of Obtaining the Correct Unrooted Tree ($P_c \times 100$) for the Case of Varying Substitution Rate Among Different Nucleotide Sites With the Two-Parameter Model*

<table>
<thead>
<tr>
<th>Model Trees</th>
<th>$B$</th>
<th>MP</th>
<th>NJD</th>
<th>NJK</th>
<th>NJG1</th>
<th>NJG2</th>
<th>EP</th>
<th>$X$</th>
<th>$Y$</th>
<th>$Z$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.33</td>
<td>97</td>
<td>98</td>
<td>98</td>
<td>100</td>
<td>99</td>
<td>89</td>
<td>44</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>0.60</td>
<td>94</td>
<td>99</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>77</td>
<td>22</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>74</td>
<td>90</td>
<td>92</td>
<td>98</td>
<td>96</td>
<td>53</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.33</td>
<td>83</td>
<td>90</td>
<td>90</td>
<td>94</td>
<td>93</td>
<td>55</td>
<td>12</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>0.60</td>
<td>83</td>
<td>86</td>
<td>86</td>
<td>90</td>
<td>89</td>
<td>50</td>
<td>11</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>65</td>
<td>77</td>
<td>75</td>
<td>80</td>
<td>78</td>
<td>41</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.33</td>
<td>7</td>
<td>30</td>
<td>31</td>
<td>96</td>
<td>75</td>
<td>64</td>
<td>21</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>0.60</td>
<td>5</td>
<td>25</td>
<td>29</td>
<td>94</td>
<td>72</td>
<td>55</td>
<td>9</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>0.90</td>
<td>1</td>
<td>10</td>
<td>27</td>
<td>85</td>
<td>64</td>
<td>41</td>
<td>5</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

*Abbreviations used are the same as those in previous tables. NJG1, NJ method with the gamma-distance of $a = 1$. NJG2, NJ method with the gamma-distance of $a = 2$; $m = 1,000$; $R = 400$. Model trees are presented in Figure 6-5. (Adapted from Jin and Nei, 1990.)
imatively additive measure can be attained by considering the gamma-
distribution (Nei and Gojobori, 1986; Jin and Nei, 1990). Since the log-
normal distribution with \( \alpha = 8 \) is rather close to the gamma-distribution
with parameter \( \alpha = 1 \) or \( 2 \) (Fig. 6-6), we developed the following gamma-
distance for the two-parameter model (Jin and Nei, 1990).

\[
\hat{d} = \frac{a}{2} \left[ (1 - 2\hat{P} - \hat{Q})^{-1/a} + \frac{1}{2} (1 - 2\hat{Q})^{-1/a} - \frac{3}{2} \right],
\]

(12)

where \( \hat{P} \) and \( \hat{Q} \) are the estimates of \( P \) and \( Q \) in equations (9) and (10),
respectively, and \( a \) is the square of the inverse of the coefficient of variation
of substitution rate (Nei, 1987a: 234–235). In the case of \( a = 1 \), \( \hat{d} \) becomes

\[
\hat{d} = \frac{2\hat{P} + \hat{Q}}{2(1 - 2\hat{P} - \hat{Q})} + \frac{\hat{Q}}{2(1 - 2\hat{Q})},
\]

(13)

whereas for \( a = 2 \) it is

\[
\hat{d} = (1 - 2\hat{P} - \hat{Q})^{-1/2} + \frac{1}{2} (1 - 2\hat{Q})^{-1/2} - \frac{3}{2}.
\]

(14)

Table 6-10 shows that when the gamma-distance with \( a = 1 \) is used, the
NJ method outperforms the EP method, even for \( S_3 \). This indicates that
as long as a distance measure that gives an approximate additivity of nu-
cleotide substitutions is used, the NJ method performs very well. The same
can be said for the ST and ME methods, since these methods are expected
to give the correct tree provided that additive distance measures are used.

No study has been done on the effect of intersite variation of substitution
rate on the performance of the ML method. However, since it is difficult
for this method to take this variation into account, the method will be
affected considerably. It is desirable to study this effect quantitatively in
the near future.

STATISTICAL TESTS OF THE TREES OBTAINED

Test of Topological Differences

When a tree (topology) is obtained by a tree-making method, one is nat-
urally interested in the accuracy of the tree (i.e., how good the tree is,
comparied with other alternative trees). A number of authors (e.g., Cav-
ender, 1981; Templeton, 1983; Felsenstein, 1985a,b; Lake, 1987; Prager
and Wilson, 1988; Kishino and Hasegawa, 1989) have proposed statistical
tests for topological differences. However, all these methods depend on
assumptions which in reality do not necessarily hold. Therefore, one has
to be cautious in using these methods. The main problem in these tests is
that the evaluation of statistical significance of the difference between two
topologies depends on the true tree (both topology and branch lengths),
which is usually unknown, and the data sets used, which do not necessarily
satisfy the assumptions required for a given tree-making method.
For example, Prager and Wilson's (1988) test is for comparing two topologies obtained by parsimony methods. In this test, the significance of the difference between the number of sites in which one topology wins, and the number of sites in which the other topology wins is examined by a binomial test. When the number of substitutions per site is small for all branches, and the trees to be compared are bifurcating, this test seems to be acceptable. In other cases, however, it may identify a wrong tree as the correct one and regard it as statistically established. This is obvious from the computer simulation given for $S_3$ in Tables 6-9 and 6-10, where a wrong tree was chosen as the correct one in most replications. Templeton's (1983) test for restriction-site data has the same problem as Wilson and Prager's. In this case, even if the rate of nucleotide substitution is constant, it may choose a wrong tree and regard it as statistically established (Nei and Tajima, 1985, 1989; Pamilo, 1990).

Lake (1987) proposed a statistical test in association with his evolutionary parsimony method. As mentioned earlier, this test examines whether or not quantity $X$, $Y$, or $Z$ is significantly different from 0. In practice, all of $X$, $Y$, and $Z$ may become nonzero under certain patterns of nucleotide substitution. Thus, this is not a test for comparing different topologies (Jin and Nei, 1990). Indeed, this test may identify a wrong tree as the correct one, as is clear from Tables 6-9 and 6-10. Therefore, this test should not be used.

There is one test that is statistically sound if the rate of nucleotide substitution is constant (Felsenstein, 1985a). It applies to a rooted tree for three DNA sequences (Fig. 6-7). In this case there are three possible trees: (A), (B), and (C) in Figure 6-7. If tree (A) is correct, one would expect that the number of nucleotide differences ($n_{12}$) between sequences 1 and 2 to be smaller than that ($n_{13}$) between 1 and 3 or that ($n_{23}$) between 2 and 3. Therefore, if $n_{12}$ is significantly smaller than $n_{13}$ and $n_{23}$ by a binomial test, one can conclude that it is statistically established. In this case, the tree corresponding to the null hypothesis ($n_{12} = n_{13} = n_{23}$) is the trifurcation tree given in Figure 6-7D. Extension of this method to the case of four or more sequences is complicated, and no studies have been done. Note also that even with three DNA sequences, the above test breaks down if the rate of nucleotide substitution is not constant.

```
1 2 3 1 2 3 1 2 3
(A)  (B)  (C)  (D)
```

Figure 6-7  Three different rooted trees (A, B, C) for three DNA sequences (1, 2, and 3). (D): null hypothesis tree.
RELATIVE EFFICIENCIES OF DIFFERENT TREE-MAKING METHODS

Williams and Goodman (1989) recently extended this approach to the case where the molecular clock does not apply. In this method, however, one must infer the ancestral nucleotide at each "informative site" by parsimony analysis. If this inference is wrong, their test is not reliable. Nevertheless, this test is a promising approach when there are useful outgroup species. This test also applies only to the case of three species of which the branching order is to be determined.

At any rate, it is a very difficult task to assess the significance level for the differences between topologies. Since all statistical methods depend on various assumptions, one should be prepared for the possibility that even a tree statistically established at the 5% or 1% level by a certain method may later turn out to be incorrect. One such example is the phylogeny for humans, chimpanzees, gorillas, orangutans, and gibbons. Templeton (1983) conducted a nonparametric test for restriction-site data of mitochondrial DNAs (mtDNAs) from these organisms and concluded that humans and gorillas are evolutionarily closer than humans and chimpanzees. However, later studies of DNA hybridization data for single-copy genomic DNA have not supported this conclusion (Sibley and Ahlquist, 1984; Caccone and Powell, 1989).

In general, one should not be too confident about a tree obtained from any tree-making method unless the number of nucleotides examined is very large. If there is any doubt about the tree obtained, the first thing to do is to increase the amount of data. Unless the amount of data is large, any sophisticated statistical method may lead to an erroneous conclusion.

Accuracy of the Topology Estimated

Although it is very difficult to test topological differences under realistic conditions, it is easier to test the accuracy of branch lengths estimated for a given topology. This test is primarily for examining the statistical significance of a given branch length, but if a branch length is not significantly different from 0, it casts doubt on the clustering pattern associated with the branch. For example, Figure 6-8 shows a phylogenetic tree for humans, chimpanzees, gorillas, orangutans, and gibbons. If branch length a in this tree is not significantly different from 0, the branching order among humans, chimpanzees, and gorillas becomes questionable.

This type of test was initiated by Nei et al. (1985) for a UPGMA tree. Application of this method to Brown et al.'s (1982) mtDNA data, from which the phylogeny in Figure 6-8 was constructed, indicates that the branch length a is not statistically significant. Therefore, Brown et al.'s data are not sufficient for resolving the branching pattern of humans, chimpanzees, and gorillas. Li (1989) extended this method to the case of other distance methods, though his primary interest was to discriminate among the three possible unrooted trees for four DNA sequences.

Theoretically, the standard error of any branch length can be computed for a UPGMA or any other distance method tree. However, the computation becomes quite complicated when the number of sequences is large.
Figure 6-8  UPGMA tree for humans, chimpanzees, gorillas, orangutans, and gibbons. This tree was constructed by using mtDNA sequence data from Brown et al. (1982). Branch length estimates ± SE are given for all internal branches, whereas only branch-length estimates are given for exterior branches. Branch length b is significantly larger than 0, whereas branch lengths a and c are not significant.

In this case, a simpler method would be to use the jackknife or bootstrap method (Efron, 1982). The jackknife method is particularly useful for evaluating the standard error of a given branch length when the number of nucleotides or restriction sites examined is relatively small. Mueller and Ayala (1982) used this method for a tree obtained from gene-frequency data (see also Pamilo, 1990). When this number is large, however, it would be easier to use bootstrapping.

Felsenstein (1985b) used bootstrapping to evaluate the accuracy of a tree obtained by a parsimony method. His method is not to evaluate the standard error of any particular branch length, but to examine how often a particular cluster in a tree appears when nucleotide sites are resampled with replacement many times. This method can be used for any other method. This test can easily be applied, particularly for the NJ method, because the computational time for reconstructing a tree is usually very short.

Nevertheless, one should be cautious about the outcome of this test, because if the data set used does not satisfy the assumption underlying a tree-making method, one may identify a wrong cluster as a correct one. This is particularly so for a maximum-parsimony tree when the rate of substitution varies extensively with evolutionary lineage. It should also be noted that if the original data set is biased for some reason, a cluster may be regarded as statistically significant even if it is a wrong one. This is because the original bias cannot be corrected by the resampling process.

NUMBER OF NUCLEOTIDES TO BE EXAMINED

From the computer simulations mentioned earlier, it is clear that a large number of nucleotides must be examined to obtain the correct phylogenetic tree. The number of nucleotides required to establish a tree statistically
depends on the topology and branch lengths of the true tree, the pattern
of nucleotide substitution, the tree-making method, and other factors.
Saitou and Nei (1986) studied this problem considering mtDNAs from
humans, chimpanzees, gorillas, orangutans, and gibbons.

Mammalian mtDNA is known to evolve about ten times faster than
nuclear DNA (Brown et al., 1979), and the phylogenetic tree for humans,
chimpanzees, gorillas, orangutans, and gibbons (Fig. 6-8) is close to those
given in Fig. 6-9. Suppose that tree A or B in Fig. 6-9 is the correct tree.
How many nucleotides then must be examined to obtain the correct tree
with probability P? The mathematical technique for computing the mini-
mum number of nucleotides required (m*) for the UPGMA, TD, DW,
FM, and MP methods is provided by Saitou and Nei (1986), and it can be
shown that m* for the NJ and ME methods is identical with that of the
TD and DW methods for the case of n ≤ 5 (Saitou and Nei, 1987). One
can therefore determine m* for all these methods.

Some of the results obtained are presented in Table 6-11 for the one-
and two-parameter models of nucleotide substitution under the assumption
that (i) only orangutans are used as an outgroup (four species) and (ii)
both orangutans and gibbons are used as outgroups. In the case of the
two-parameter model, a transition/transversion ratio (α/β) appropriate to
mtDNA is assumed to be 20. Table 6-11 shows that m* for P = 0.95 is
smallest for the NJ, ME, TD, and DW methods, and largest for the FM
method (m* for the MP method being usually close to that of the former).
However, if tree A is correct and α/β = 20, 2,600 ~ 3,100 nucleotides
must be examined even in the NJ method, depending on the number of
outgroup species used. This number is much larger than the number of
nucleotides examined (895) by Brown et al. (1982). Therefore, Brown et
al.’s data are unlikely to resolve the branching pattern of humans, chimp-
panzees, and gorillas. This conclusion is the same as that obtained from
the statistical test of branch length a in Figure 6-8 mentioned earlier.

Saitou and Nei (1986) have also examined the number of nucleotides
required when orangutans and gibbons are not available as outgroups. In

![Figure 6-9](attachment:image.png)

**Figure 6-9** Two model trees for the case of four or five species. A, B, C, any of
humans, chimpanzees, and gorillas; D, orangutans; E, gibbons. Numbers refer to
the expected number of nucleotide substitutions per site per branch.
Table 6-11  Numbers ($m^*$) of Nucleotides Required for Obtaining the Correct
Tree with a Probability of 0.95*

<table>
<thead>
<tr>
<th>Tree-Making Method</th>
<th>Four Species</th>
<th></th>
<th></th>
<th>Five Species</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-p</td>
<td>2-p</td>
<td>1-p</td>
<td>2-p</td>
<td></td>
</tr>
<tr>
<td>NJ, ME, TD, DW</td>
<td>2,100</td>
<td>3,100</td>
<td>1,700</td>
<td>2,600</td>
<td></td>
</tr>
<tr>
<td>MP, CP</td>
<td>2,100</td>
<td>3,200</td>
<td>1,700</td>
<td>2,700</td>
<td></td>
</tr>
<tr>
<td>UPGMA</td>
<td>4,200</td>
<td>4,700</td>
<td>4,200</td>
<td>4,700</td>
<td></td>
</tr>
<tr>
<td>FM</td>
<td>5,000</td>
<td>8,200</td>
<td>3,000</td>
<td>4,900</td>
<td></td>
</tr>
</tbody>
</table>

*Tree A of Figure 6-9

<table>
<thead>
<tr>
<th></th>
<th>Four Species</th>
<th></th>
<th></th>
<th>Five Species</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TD, DW</td>
<td>760</td>
<td>1,200</td>
<td>640</td>
<td>890</td>
<td></td>
</tr>
<tr>
<td>MP, CP</td>
<td>790</td>
<td>1,300</td>
<td>680</td>
<td>980</td>
<td></td>
</tr>
<tr>
<td>UPGMA</td>
<td>1,400</td>
<td>1,500</td>
<td>1,400</td>
<td>1,500</td>
<td></td>
</tr>
<tr>
<td>FM</td>
<td>1,700</td>
<td>2,800</td>
<td>900</td>
<td>1,700</td>
<td></td>
</tr>
</tbody>
</table>

*Abbreviations used are the same as those used in previous tables.

1-p and 2-p indicate the one- and two-parameter models, respectively. CP, compatibility method. (Adapted from Saitou and Nei, 1986.)

In this case, $m^*$ becomes about 4,700 under the assumption of the molecular clock.

Recently, Miyamoto et al. (1987) and Maeda et al. (1988) used about 7,100 and 3,100 nucleotides, respectively, from the nuclear genome, to study the branching pattern of humans, chimpanzees, and gorillas. The total number of nucleotides available is now more than 10,000. However, since the rate of nucleotide substitution in the nuclear genome is about one-tenth that of mtDNA, this number still does not seem to be sufficient for establishing the branching order unquestionably (Maeda et al., 1988).

In this connection, it is interesting that when all cladistically informative sites from these data are used, Williams and Goodman’s (1989) test leads to the conclusion that humans and chimpanzees are closer to each other than to gorillas at the 95% level. If we consider the possibility that the ancestral nucleotide inferred is incorrect, however, this conclusion does not seem to be sufficient for establishing the branching order of the three species involved. Using a different statistical method, Li (1989) could not reach the same conclusion from the same set of data. Of course, if we consider all data from DNA hybridization, mtDNA, and nuclear DNA sequences, humans and chimpanzees seem to be genetically closer to each other than to gorillas.

DISCUSSION

As mentioned above, many studies have been conducted on the relative efficiencies of tree-making methods. They indicate that the relative efficiencies depend on various factors such as the shape of the true tree, the numbers of nucleotide substitutions, transition/transversion ratios, and varying rate of nucleotide substitution. However, we can make some general conclusions.
One of the general conclusions is that the maximum parsimony method, which is currently very popular among numerical taxonomists, is as efficient as the ME, NJ, and ML methods only when the number of nucleotide substitutions per site is very small and the number of nucleotides examined is very large. Otherwise, the latter methods are better than the MP method. This seems to be true whether or not the transition/transversion bias is taken into account in the parsimony method (Jin and Nei, 1990). So far, no studies have been conducted on the relative efficiency of such recent parsimony methods as dynamic weighting parsimony (e.g., Williams and Fitch, 1990). Therefore, more simulation work is necessary.

The NJ method is an approximate method of obtaining the minimum evolution tree, but produces the same tree as that obtained by the ME method in most cases; if it does not, the tree obtained is usually very close to the ME tree. The advantage of the NJ method over the ME method is the short computer time required. The ST method depends on a procedure that is similar to that of the NJ method, and these two methods seem to be equally efficient in obtaining the correct tree. The NJ method, however, has one advantage over the ST method; it gives not only the topology but also the branch-length estimates of the tree obtained.

The ML method is also very efficient in obtaining the correct tree when the underlying assumptions are satisfied. When these assumptions are not satisfied, it seems to be less efficient than the NJ or ME method because the former depends on the details of the probability model, whereas the latter are known to be quite insensitive to various aspects of nucleotide substitution (see Table 6-10). Another disadvantage of the ML method is that it requires enormous computer time, even for a small number of DNA sequences examined.

Some investigators (e.g., Czelusniak et al., 1990) seem to be dissatisfied with the fact that the NJ method gives only one topology. They are interested in knowing how good the NJ tree is compared with alternative topologies. Since the NJ method is intended to obtain the minimum evolution tree, comparison of alternative trees can be made by using the total sum of branch lengths, i.e., \( L \) in equation (3). In the case of the ME method, \( L \) is computed for all topologies, so that one can choose a topology that gives the smallest \( L \) value, though it usually takes a large amount of computer time. In practice, however, it is sufficient to examine the \( L \) values for several alternative topologies that are close to that of the NJ tree. A simple way to find alternative topologies is to use Robinson and Foulds' (1981) measure \( (d_T) \) of computing topological differences. This measure takes a value of 2, 4, 6, 8, etc. I suggest that all topologies which are different from the NJ tree by \( d_T = 2 \) be examined for this purpose. Since the number of such topologies is not large (Sourdis and Nei, 1988), it would be much simpler to examine alternative topologies in this way. Once the topologies are identified, one can easily compute \( L \) for each of the topologies by using Fitch and Margoliash's (1967) method of estimating branch lengths. If this procedure finds a tree which has a smaller \( L \) value than that of the NJ tree, then this tree should be used as the final tree. Otherwise,
the NJ tree will be used as the final tree. (Note that the ME tree is not necessarily the correct tree).

Although the statistical test of topological differences is besieged with many problems as mentioned earlier, a rough test of the difference in \( L \) between two topologies can be made by computing the standard error of the differences by using either the jackknife or bootstrap method. Conceptually, this test is similar to Hasegawa et al.’s (1987) test of maximum-likelihood values, but the computation is much simpler.

When the ME or NJ method is used, one must use an appropriate distance measure. As mentioned earlier, provided that the distance measure gives the correct number of nucleotide substitutions, these methods produce the correct tree. In practice, the pattern of nucleotide substitution is quite complicated, and there is no universally accurate distance measure for this purpose. However, I recommend the following guidelines for measuring nucleotide substitutions (Jin and Nei, 1990).

1. When the Jukes-Cantor estimate of the number of nucleotide substitutions per site \( (d) \) between different sequences is about 0.1 or less, use the Jukes-Cantor distance whether there is a transition/transversion bias or the substitution rate \( (\lambda) \) varies with nucleotide site. In this case, the Kimura distance or the gamma-distance gives essentially the same value as the Jukes-Cantor distance. One may also use the \( p \) distance (proportion of different nucleotides) for constructing a topology.

2. When \( d \) is greater than 0.1 but less than about 0.3, use the Jukes-Cantor distance unless the transition bias is high (e.g., \( B > 0.5 \)). When this bias is high, use the Kimura distance.

3. When \( 1.0 > d > 0.3 \) and there is evidence that \( \lambda \) varies extensively with site, use the gamma-distance. In general, we suggest that the gamma-distance with \( a = 1 \) be used. However, one may choose a different gamma-distance, estimating \( a \) from data. Wilson et al. (1989) recently used a distance with \( a = \frac{1}{2} \) for restriction-site data of mtDNA in hominoids.

4. When \( 1.0 > d > 0.3 \) and the frequencies of the four nucleotides (A, T, C, and G) deviate substantially from equality, use Tajima and Nei’s (1984) distance.

5. When \( d > 1.0 \) for many pairs of sequences, the phylogenetic tree estimated is not reliable for a number of reasons (e.g., large standard errors of \( d \)'s, and sequence alignment errors). We therefore suggest that these sets of data should not be used. In this case, one may eliminate the portion of the gene that evolves very fast and use only the remaining region as is often done in studies of the evolution of different kingdoms or phyla using ribosomal RNA genes (e.g., Gouy and Li, 1989). If a coding region of DNA is examined, amino acid sequences rather than DNA sequences should be used. One may also use a different gene which evolves more slowly.

6. When a phylogenetic tree is constructed from the coding regions of a gene, the distinction between synonymous \( (d_s) \) and nonsynonymous
(\(d_N\)) substitutions (Miyata and Yasunaga, 1980; Li et al., 1985; Nei and Gojobori, 1986) will be helpful, because the rate of synonymous substitution is usually much higher than that of nonsynonymous substitution. When relatively closely-related species with \(d_S < 1.0\) are studied for a large number of codons, one may use \(d_S\) for constructing a tree. This procedure is expected to reduce the effect of variation in substitution rate among different sites, because synonymous substitutions are apparently largely neutral in higher organisms (Nei, 1987a: 79–86). However, when relatively distantly related species are studied, the use of \(d_N\) is recommended.

Finally, it should be mentioned that the simulation study of tree-making methods is far from complete. In most studies, the number of DNA sequences considered is quite small (mostly four to eight) to save computer time. In some tree-making methods (e.g., MP method), the increase in the number of sequences considered may enhance the probability of obtaining the correct tree. Therefore, a more detailed study is necessary to solve this problem. The effects of different patterns of nucleotide substitution, particularly varying rate among different sites, should also be studied more carefully. Another problem that should be studied by computer simulation is the statistical methods for testing topological differences or the accuracy of a tree obtained. Since these methods depend on a number of assumptions about complex evolutionary processes, it would be important to establish the validity of the method by computer simulation.

In this chapter we considered only gene trees. A gene tree is not necessarily the same as the phylogeny of the species or populations from which the gene sequences are sampled (Nei, 1987a; Pamilo and Nei, 1988). The chance that the gene and population trees are different is high when the gene used is polymorphic or when the populations considered split relatively recently. In these cases, one must examine many independently evolving genes from the genome to estimate the correct population tree (Saitou and Nei, 1986; Pamilo and Nei, 1988). In this connection, it is important to note that a phylogenetic tree constructed from mitochondrial or chloroplast DNAs is a gene tree though they include many genes. This is because the genes in the mitochondrial or chloroplast DNA are inherited as a single entity without recombination. Therefore, some caution should be exercised in inferring population phylogenies from mitochondrial or chloroplast DNA. Of course, gene trees are not always studied just for inferring a population phylogeny. In such a case as the study of the evolution of multigene families, a gene tree provides important information.

**COMPUTER PROGRAMS**

Computer programs for computing various distances for DNA sequences (numbers of nucleotide substitutions per site, synonymous substitutions per synonymous site, nonsynonymous substitutions per nonsynonymous site, etc.) and for constructing (NJDRAW) and testing (NJBOOT) a neigh-
bor-joining tree are available upon request from the Institute of Molecular Evolutionary Genetics, Penn State University, 328 Mueller Laboratory, University Park, PA 16802-5303. Please send IBM compatible 3.5 or 5.25 inch floppy diskettes. Each program requires one 360 Kb diskette.

ACKNOWLEDGMENTS

I thank Li Jin for his comments. This study is supported by research grants from the National Institutes of Health and the National Science Foundation.

REFERENCES


Nei, M., and T. Gojobori. (1986) Simple methods for estimating the numbers of


invariable sites should be considered when sequence divergence is calculated.  


*Evolution* 37:221–244.

Uzzell, T., and K. W. Corbin. (1971) Fitting discrete probability distributions to evolutionary events.  

