C. R. Rao and R. Chakraborty, eds., *Handbook of Statistics, Vol. 8*
© Elsevier Science Publishers B.V. (1991) 317–346

Statistical Methods for Phylogenetic Tree Reconstruction

Naruya Saitou

1. Introduction

Reconstruction of the phylogeny of organisms is one of the most important problems in evolutionary study. A phylogeny is usually illustrated by a tree-like figure. Thus we call it 'phylogenetic tree' or simply 'tree' in this chapter. It may be interesting to note that Darwin (1859) was the first to show such a tree to explain the pattern of divergence of species through evolution, though his tree was an imaginary one.

Previously phylogenetic trees were reconstructed mostly by using morphological data. With the advent of the study of molecular evolution, however, it is now customary to construct phylogenetic trees from molecular data, especially from nucleotide sequences. In this chapter we will therefore be concerned primarily with nucleotide sequence data. Nevertheless, most of the methods discussed in this chapter can also be applied to other types of data, including non-molecular data. I will first discuss theoretical aspects of phylogenetic trees in the next section. Distance matrix methods and character-state methods are explained in Sections 3 and 4, respectively. Lastly, the efficiency of different methods is discussed.

2. Theoretical aspects of phylogenetic trees

2.1. Rooted trees and unrooted trees

Mathematically, a phylogenetic tree is literally a 'tree' in graph theory. (A graph is composed of node(s) and edge(s). A node represents any object and an edge represents the relationship between nodes.) A tree is a special kind of graph: there should be only one path between any two nodes. Thus there is no loop in a tree (e.g., see Figure 1). In evolutionary study, the term 'branch' is used instead of 'edge' and not only the branching pattern (topological relationship between nodes) but the length of each branch is often important.

A tree can be either directed or undirected. In a directed tree there is a

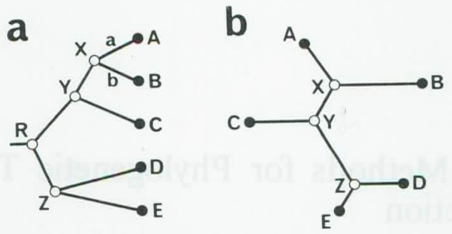

Fig. 1. Examples of a rooted tree (a) and an unrooted tree (b) for five OTUs.

particular node, or root, and there will be a unique path from this node to any other node. Hence a directed tree is also called a rooted tree. Figure 1(a) shows an example of a rooted tree, in which the root is designated as R. In organismal evolution, the direction is of course that of time and the root is the common ancestor. Therefore a phylogenetic tree in an ordinary sense is a rooted tree.

An undirected tree does not have such root, and it is also called an unrooted tree. Although an unrooted tree itself may not be regarded as a phylogenetic tree, it can be converted to a rooted tree if the position of the root is specified. Figure 1(b) is an example of an unrooted tree, and the topological relationship of nodes is identical with that of Figure 1(a) if we ignore the root (R) of Figure 1 and the difference of branch lengths between these two trees. Unrooted trees are sometimes called 'networks', but that has a different meaning in graph theory.

Nodes can be any kind of object, and in evolutionary study, it can be a species, populations, or genes, as will be discussed in Section 2.4. It is useful to distinguish exterior nodes (full circles in Figure 1) and interior nodes (empty circles in Figure 1). Exterior nodes have only one branch but interior nodes have more than one branch. We usually have informations on the exterior nodes only. Exterior nodes are often referred to as operational taxonomic units (OTUs) and interior nodes may be called hypothetical taxonomic units (HTUs).

2.2. *Possible number of trees*

When we consider the phylogenetic relationship of three OTUs, there are three possibilities (Figure 2(a)). The true phylogenetic tree is one of these rooted trees. The number of possible trees rapidly increases with increasing the number of OTUs compared. The general equation for the possible number of bifurcating rooted trees for n ($\geqslant 2$) OTUs is given by

$$(2n-3)!/(2^{n-2}(n-2)!)\,. \tag{2.1}$$

Equation (2.1) was first presented by Cavalli-Sforza and Edwards (1967). The number of bifurcating unrooted trees for n OTUs is given by replacing n by $n-1$ in equation (2.1). Figure 2(b) shows the three possible unrooted trees for four OTUs. Table 1 gives the possible number of rooted and unrooted bifurcating trees up to 10 OTUs. For the number of multifurcating trees, see Felsenstein (1978a).

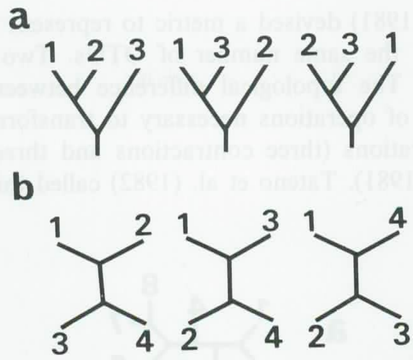

Fig. 2. (a) Three possible rooted trees for three OTUs. (b) Three possible unrooted trees for four OTUs.

Table 1
Possible numbers of rooted and unrooted trees for *i* OTUs

Number of OTUs	Possible number of	
	Rooted trees	Unrooted trees
2	1	1
3	3	1
4	15	3
5	105	15
6	945	105
7	10395	945
8	135135	10395
9	2027025	135135
10	34459425	2027025

It is clear from Table 1 that the search of the true phylogenetic tree for more than 10 OTUs is as if looking for a needle in a haystack, if we examine trees one by one. Unfortunately, the problem of finding the true tree from all the possible trees belongs to a so-called NP-complete problem, and there is no effective algorithm for this: we have to do an exhaustive search of all possible trees. This is why so many heuristic methods have been proposed for reconstruction of phylogenetic trees.

2.3. *Topological differences between trees*

The branching pattern of a tree with a given number of OTUs is called a 'topology' in evolutionary study (here too the word has a different meaning in graph theory). Each tree has its own topology, distinguished from those of other trees. However, the amount of topological difference can vary from tree to tree.

Robinson and Foulds (1981) devised a metric to represent the topological difference between trees for the same number of OTUs. Two trees (a and b) are presented in Figure 3. The topological difference between these two trees is defined by the number of operations necessary to transform each other. In this case, we need six operations (three contractions and three decontractions; see Robinson and Foulds, 1981). Tateno et al. (1982) called this metric a 'distortion index'.

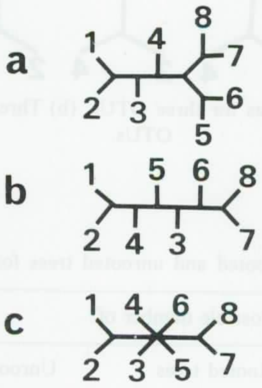

Fig. 3. Two topologically different trees (a and b) for 8 OTUs and their consensus tree (c).

A multifurcating tree c of Figure 3 is sometimes called the 'consensus tree' for trees a and b of Figure 3 (Adams, 1972). When trees a and b are equally likely, one way is to present the consensus tree c. Tree c can be obtained after three contractions of trees a and b.

2.4. Gene trees and species trees

Traditionally a phylogenetic tree automatically means a tree of species. However, genes are usually the units of comparison in molecular evolution, and there are several important differences between the phylogenetic tree of species and that of genes. The former is called 'species tree' and the latter 'gene tree' (Tateno et al., 1982; Nei, 1987). The most prominent difference between these two trees is illustrated in Figure 4. Because a gene duplication occurred before the speciation of species A and B, both species have two homologous genes (1 and 2) in their genomes. In this situation, we should distinguish 'orthology', that is homology of genes reflecting the phylogenetic relationship of species, from 'paralogy', that is homology of genes caused by gene duplication(s) (Fitch, 1970). Thus, genes *A1* and *B1* (or *A2* and *B2*) are 'orthologous', but genes *A1* and *A2*, *B1* and *B2*, *A1* and *B2*, or *A2* and *B2* are 'paralogous'. If one is not aware of the gene duplication event, gene tree for *A1* and *B2* may be misrepresented as the species tree of A and B, and thus a gross overestimation of the divergence time may occur.

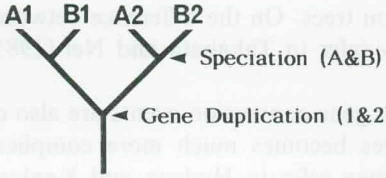

Fig. 4. A gene tree for four genes from two species.

Even when orthologous genes are used, a gene tree may be different from the corresponding species tree. This difference comes from the existence of allelic polymorphism of the ancestral species. A simple example is illustrated in Figure 5. A gene sampled from species A has its direct ancestor in the ancestral species X, and so does a gene sampled from species B. Thus the divergence between two genes sampled from different species always overestimates that of species (see Figure 5(a)). The amount of overestimation is related to the population size of the ancestral species X (see Tajima, 1983, for details). It may be interesting to note that Nei (1972) considered this overestimation and estimated the amount by the average heterozygosity in the present population. Thus Nei's genetic distance is an estimation of the amount of divergence of species or populations, not genes.

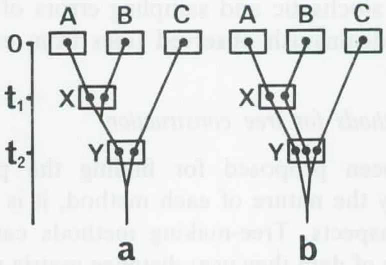

Fig. 5. Two gene trees (a and b) in which the topological relationship of genes is the same as or different from that of the specie trees, respectively.

The difference between species tree and gene tree is not confined to the amount of divergence. If the time $T\ (=t_2-t_1$; see Figure 5) is not long, the two ancestral genes sampled from species A and B may coexist in the ancestral species Y. At this stage we also have a gene that is ancestral to a gene sampled from species C. Thus the topological relationship of these three genes is determined by chance alone, and it is possible to have a gene tree in which the topological relationship is different from that of species tree, as shown in Figure 5(b). For a more detailed discussion, readers may refer to Nei (1987).

When we consider phylogenetic relationships of different populations rather than species, the phylogenetic tree of populations may be called 'population tree'. If the effect of migration between populations is ignored, a population tree is essentially the same as a species tree. Thus the above explanation on species trees is applicable to population trees. On the difference between a population tree and a gene tree, readers may refer to Takahata and Nei (1985) and Pamilo and Nei (1988).

If the recombination or gene conversion events are also considered, the problem of constructing gene trees becomes much more complicated. Readers who are interested in this topic may refer to Hudson and Kaplan (1985).

2.5. Expected trees and observed trees

Branch lengths of a phylogenetic tree is ideally proportional to physical time. Thus the branch a and b of Figure 1(a) should have the same length. We call this type of tree 'expected tree'. In real data, however, we may have different lengths for branches a and b. This is because the amount of genetic change is not always the same between these two branches. A tree reconstructed from observed data is called 'observed tree'. Both species tree and gene tree can be either an expected or an observed tree. The distinction between expected and observed trees is similar to that between expected and realized distance trees defined by Nei (1987).

If the rate of evolution is constant among different lineages, or if the molecular clock is assumed, one may think that an observed tree is the same as the expected tree. However, this may not be the case. Even if the rate is constant, an observed tree usually show some variations on the number of nucleotide substitutions for each branch because of stochastic and sampling errors of nucleotides compared. It is thus important to distinguish observed trees from expected trees.

2.6. Classification of methods for tree construction

Many methods have been proposed for finding the phylogenetic tree from observed data. To clarify the nature of each method, it is useful to classify these methods from various aspects. Tree-making methods can be divided into two types in terms of the type of data they use; distance matrix methods and character-state methods. A distance matrix consists of a set of $\frac{1}{2}n(n-1)$ distance values for n OTUs (see Table 2 as an example), whereas an array of character states is used for the character-state methods. Sections 3 and 4 are thus classified.

Another classification is by the strategy of a method to find the best tree. One way is to examine all or a large number of possible trees and choose the best one in terms of a certain criterion. We call this the 'exhaustive search method'. For example, the maximum parsimony method belongs to this category. The other strategy is to examine a local topological relationship of OTUs and find the best tree. This type of method is called the 'stepwise clustering method' (Saitou and Imanishi, 1989). Most of the distance methods are stepwise clustering methods.

Table 2
An example of distance matrix[a]

	C	P	G	H	O
Common chimpanzee	–	0.0117	0.0415	0.0373	0.0895
Pygmy chimpanzee	0.0118 ±0.0036	–	0.0405	0.0319	0.0863
Gorilla	0.0427 ±0.0069	0.0416 ±0.0068	–	0.0362	0.0905
Human	0.0382 ±0.0065	0.0327 ±0.0060	0.0371 ±0.0064	–	0.0873
Orangutan	0.0953 ±0.0106	0.0916 ±0.0104	0.0965 ±0.0107	0.0928 ±0.0104	–

[a] Data from Hixson and Brown (1986). Figures above the diagonal indicate the proportion of the nucleotide difference, and those below the diagonal are the estimated number of nucleotide substitution per site (with their SEs).

3. Distance methods

3.1. Distance matrices

In distance methods, a phylogenetic tree is constructed by considering the relationship among the distance values of a distance matrix. There are two kinds of distances; metric and non-metric. The former follows the principle of triangle inequality and the latter does not. The triangle inequality is:

$$D_{ij} \leqslant D_{ik} + D_{jk}\,, \tag{3.1}$$

where D_{ij} is the distance between OTUs i and j. In numerical taxonomy, especially in cladistics, use of metric distance is advocated (see Section 4.2 for cladistics). In molecular evolution, however, the estimated number of nucleotide substitutions as an evolutionary distance does not necessarily follow the principle of triangle inequality.

An example of a distance matrix is presented in Table 2. The data are Hixson and Brown's (1986) mitochondrial DNA sequences for human (H), chimpanzee (C), pygmy chimpanzee (P), gorilla (G), and orangutan (O). Gaps in the aligned sequences were excluded, and a total of 939 nucleotides were used for the analysis. Values above the diagonal are the proportions of nucleotide differences, and those below the diagonal are evolutionary distances (numbers of nucleotide substitutions per site).

There are many methods for estimating evolutionary distances (see Nei, 1987, for a review), and we used a simple method of Jukes and Cantor (1969) for this case. Evolutionary distance (d) between two nucleotide sequences is estimated by

$$d = -\tfrac{3}{4}\log[1 - \tfrac{4}{3}p], \tag{3.2}$$

where p is the proportion of different nucleotides between the two sequences. The standard error of d is estimated by

$$\mathrm{SE}(d) = 3/(3 - 4p)[p(1 - p)/L]^{1/2}, \tag{3.3}$$

where L is the number of nucleotides compared (Kimura and Ohta, 1972).

3.2. *Phenetic methods*

A simple way of classification of organisms is to combine phenotypically similar objects first. This approach is called 'phenetics' in numerical taxonomy. A phylogenetic tree constructed by a phenetic approach is called 'phenogram'. There are many ways to obtain a phenogram from a distance matrix (see Sneath and Sokal, 1973 for a review), and all are step wise clustering methods. In this section, two methods (UPGMA and WPGMA) that are frequently used in molecular evolution will be discussed. Original ideas of UPGMA (Unweighted Pair Group Method by average) and WPGMA (Weighted Pair Group Method by Arithmetic average) were first proposed by Sokal and Michener (1958). UPGMA was independently proposed by Nei (1975) for molecular data.

Let us explain the algorithms of UPGMA and WPGMA using the evolutionary distance matrix of Table 2. We first choose the smallest distance, that is, D_{PC} (= 0.0118). Then OTUs P and C are combined and the distances between the combined OTU (PC) and the remaining OTUs are computed as:

$$D_{(\mathrm{PC})i} = \tfrac{1}{2}(D_{\mathrm{P}i} + D_{\mathrm{C}i}), \tag{3.4}$$

where i represents an OTU other than P or C. Hence there are now only six distance values (see Table 3(a)). At the next step, again the smallest distance ($D_{(\mathrm{PC})\mathrm{H}} = 0.0355$) is chosen from the distance matrix of Table 3(a). Then the OTU

Table 3
An example of the procedure used in UPGMA and WPGMA

(a) First step

	PC	G	H
G	0.0422		
H	0.0355	0.0371	
O	0.0935	0.0965	0.0928

(b) Second step[a]

	PCH	G	O
PCH	–	0.0397	0.0932
G	0.0405	–	0.0965
O	0.0932	0.0965	–

[a] Figures above and below the diagonal are obtained by WPGMA and UPGMA, respectively.

(PC) and OTU H are further combined into OTU (PCH). When we compute the average distance between OTU (PCH) and other two OTUs (G and O), the difference between UPGMA and WPGMA arises. In the case of UPGMA, the original distances are always used for obtaining the averaged distances, whereas the current distance matrix is used for WPGMA. Thus, for example,

$$D_{(\mathrm{PCH})\mathrm{G}} = \tfrac{1}{3}(D_{\mathrm{PG}} + D_{\mathrm{CG}} + D_{\mathrm{HG}}) = 0.0405$$

if we apply UPGMA, and

$$D_{(\mathrm{PCH})\mathrm{G}} = \tfrac{1}{2}(D_{(\mathrm{PC})\mathrm{G}} + D_{\mathrm{HG}}) = 0.0397$$

if we apply WPGMA. If the data strictly follows the constancy of evolutionary rate, these two values are identical. In reality, however, the rate may not be the same. Thus there can be a slight difference between distances obtained by UPGMA and that by WPGMA. Table 3(b) shows the distance matrix after the second step.

After two more steps, all five OTUs are clustered into a single OTU. The final tree thus obtained by UPGMA is shown in Figure 6(a). The tree obtained by WPGMA has the same topological relationship with that tree in the present example. Note that an estimated distance between any pair of OTUs can be obtained by summing all branch lengths connecting these two OTUs.

Boxes of Figure 6(a) represent the ranges of one standard error (SE) of the distance of each branching points from the present time, computed by Nei et al.'s (1985) method. Computation of SEs was done as follows. The SE of distance X

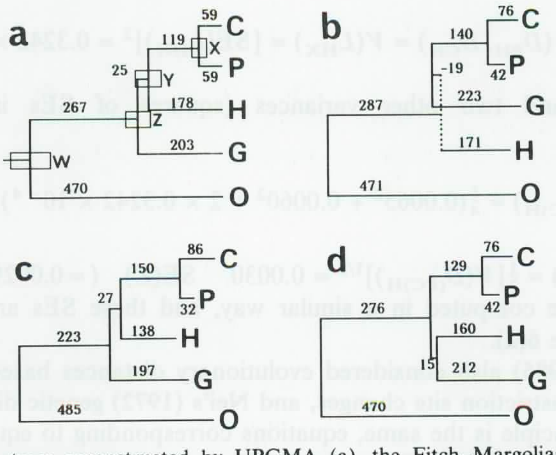

Fig. 6. Phylogenetic trees reconstructed by UPGMA (a), the Fitch–Margoliash method (b), the distance Wagner method (c), and the neighbor-joining method (d). The data are from Table 2. All figures should be multiplied by 10^{-4}. Boxes in tree a denote one SE of the distances between each branching points and the extant species.

($= D_{XC} = D_{XP}$) of Figure 6(a) is given by $SE(X) = \frac{1}{2}SE(D_{PC})$, where $SE(D_{PC})$ is SE of D_{PC}. Thus $SE(X) = \frac{1}{2} \times 0.0036 = 0.0018$ from Table 2.

Estimation of SE(Y) is more complicated. We first note the relation $SE^2(Y) = V(Y) = \frac{1}{4}V(D_{(PC)H})$, where $V(\cdot)$ denotes a variance. $D_{(PC)H}$ has been estimated by $\frac{1}{2}(D_{PH} + D_{CH})$ applying equation (3.4). Thus

$$\begin{aligned} V(D_{(PC)H}) &= V[\tfrac{1}{2}(D_{PH} + D_{CH})] \\ &= \tfrac{1}{4}[V(D_{PH}) + V(D_{CH}) + 2\mathrm{Cov}(D_{PH}, D_{CH})] , \end{aligned} \tag{3.5}$$

where Cov(i, j) is the covariance between distances i and j. Because the lineages of two chimpanzee species evolved independently after the divergence at point X, $\mathrm{Cov}(D_{PH}, D_{CH}) = V(D_{HX})$, where D_{HX} is estimated by

$$D_{HX} = D_{(PC)H} - \tfrac{1}{2}D_{PC} = \tfrac{1}{2}(D_{PH} + D_{CH}) - \tfrac{1}{2}D_{PC} . \tag{3.6}$$

Thus,

$$D_{HX} = \tfrac{1}{2}(0.0327 + 0.0382) - \tfrac{1}{2} \times 0.0118 = 0.02955 ,$$

from Table 2. This value is an estimate of evolutionary distance (d) between nodes X and H of Figure 6(a). However, the proportion (p) of nucleotide difference is used for estimating the variance (or square of SE) of d (see equation 3.3). Therefore we estimate p from d applying equation (3.2) as

$$\hat{p} = \tfrac{3}{4}[1 - e^{-4d/3}] . \tag{3.7}$$

In the present example, $\hat{p}$ becomes 0.02898. Putting this value into equation (3.3),

$$\mathrm{Cov}(D_{PH}, D_{CH}) = V(L_{HX}) = [SE(L_{HX})]^2 = 0.3242 \times 10^{-4} .$$

Putting this and two other variances (squares of SEs in Table 2) into equation (3.5),

$$V(D_{(PC)H}) = \tfrac{1}{4}(0.0065^2 + 0.0060^2 + 2 \times 0.3242 \times 10^{-4}) = 3.577 \times 10^{-5} .$$

Hence, $SE(Y) = \frac{1}{2}[V(D_{(PC)H})]^{1/2} = 0.0030$. SE(Z) ($= 0.0029$) and SE(W) ($= 0.0049$) were computed in a similar way, and these SEs are represented as boxes in Figure 6(a).

Nei et al. (1985) also considered evolutionary distances based on amino acid substitutions, restriction site changes, and Nei's (1972) genetic distance. Although the general principle is the same, equations corresponding to equation (3.7) differ in each distance measure. Recently, R. Chakraborty (personal communication) improved Nei et al.'s method for Nei's genetic distance.

The constancy of the evolutionary rate is implicitly assumed for the phenetic approach of numerical taxonomy. With the discovery of molecular clock, such a phenetic methods, especially UPGMA, has been advocated for reconstructing phylogenetic trees (Nei, 1975). In this connection, it should be noted that UPGMA gives least-squares estimates of branch lengths for the tree obtained (Chakraborty, 1977). That is, UPGMA minimizes the quantity

$$S = \sum (D_{ij} - 2\lambda t_{ij})^2 , \tag{3.8}$$

where λ is the evolutionary rate and t_{ij} is the time since divergence between OTUs i and j.

3.3. *Fitch and Margoliash's method*

Fitch and Margoliash (1967) proposed an exhaustive search method for reconstructing a phylogenetic tree. The first step of this method is to estimate branch lengths of a tree, of which tree topology is expected to be the same or quite similar to that obtained by UPGMA. The principle of the branch length estimation is as follows. Let us designate L_{ij} for the length of branch connecting nodes i and j. Then $D_{ij} = L_{i\mathrm{X}} + L_{j\mathrm{X}}$ in the tree of Figure 7. This is because the additivity of distances is assumed. From this relationship, $L_{i\mathrm{X}}$'s (i = 1, 2, and 3) are estimated by

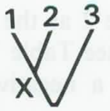

Fig. 7. A relationship of three OTUs.

$$L_{1\mathrm{X}} = \tfrac{1}{2}(D_{12} + D_{13} - D_{23}) , \tag{3.9a}$$

$$L_{2\mathrm{X}} = \tfrac{1}{2}(D_{12} + D_{23} - D_{13}) , \tag{3.9b}$$

$$L_{3\mathrm{X}} = \tfrac{1}{2}(D_{13} + D_{23} - D_{12}) . \tag{3.9c}$$

When we compare n (>3) OTUs, OTU 3 is a composite OTU, which consists of all the remaining OTUs. Then D_{13} and D_{23} are given by taking averages of D_{1j}'s and D_{2j}'s (j = 3, 4, ..., n), respectively.

If D_{12} is found to be the smallest, then OTUs 1 and 2 are combined as in the case of UPGMA and the averaged distances between this combined OTU (12) and other OTUs are computed using equation (3.4). $L_{1\mathrm{X}}$ and $L_{2\mathrm{X}}$ are also computed applying equations (3.9a) and (3.9b) at this step. The same procedure is repeated until all OTUs are clustered to become a single OTU.

At the next step the so-called percent standard deviation (PSD) is used as the criterion. For distance data of n OTUs,

$$\mathrm{PSD} = \left[\frac{2\sum \{(D_{ij} - E_{ij})/D_{ij}\}^2}{n(n-1)}\right]^{1/2} \times 100\,, \tag{3.10}$$

where the summation is for all possible pairs of OTUs and E_{ij} is estimated (patristic) distance between OTUs i and j. E_{ij} is obtained by summing estimated lengths of branches connecting OTUs i and j.

Tateno et al. (1982) proposed a criterion (S_0) similar to PSD:

$$S_0 = \left[\frac{2\sum (D_{ij} - E_{ij})^2}{(n-1)}\right]^{1/2}. \tag{3.11}$$

We can use either PSD or S_0 as the criterion to find the best tree, and the tree that has the smallest PSD or S_0 is chosen as the best tree through an exhaustive search of all possible trees.

Table 7 shows values of PSD and S_0 for four trees obtained from data of Table 2. Tree 1 has been obtained by UPGMA (see Figure 6(a)). Two chimpanzee species are clustered for trees 2 and 3 as in tree 1, but the branching pattern among human (H), chimpanzees (PC), and gorilla (G) is different each other. Human and pygmy chimpanzee are clustered in tree 4. The Fitch–Margoliash (FM) method chose tree 2 as the best tree among these four trees as did Tateno et al.'s (1982) method (see Table 7). The tree thus obtained is shown in Figure 6(b). Note that there is a negative branch (between H and (PCG) cluster) in this tree.

Because the FM method produces an unrooted tree, orangutan was assumed to be the outgroup species among five species compared in this example, and the root was located on the branch going to orangutan, assuming a constancy of evolutionary rate.

Prager and Wilson (1978) and Sourdis and Krimbas (1987) modified the criterion of the FM method for choosing the best tree. They discarded trees in which negative branch lengths were obtained. Tree 2 is discarded if we use this modified criterion, and instead tree 3 will be chosen. However, it is possible that even the true tree may have a branch with negative distance. A negative value for a branch may appear if there are many backward and parallel substitutions.

Other types of modification of the FM method have been proposed by de Soete (1983) and Elwood et al. (1985). Readers who are interested in their modifications may refer to original papers.

3.4. *Distance Wagner method*

Farris (1972) proposed a method that can be considered as an application of the principle of maximum parsimony to distance data. Because the technique of reconstructing unrooted trees from character-state data is called 'Wagner network' in cladistics, Farris named this method 'distance Wagner' (DW) method. In this case, however, a distance measure satisfying triangle inequality (a metric) is supposed to be used. Thus in the following, proportion of nucleotide difference of Table 2, that is a metric, is used.

We first connect two OTUs of which distance is the smallest. This is D_{PC} ($= 0.0117$) from Table 2. Then these two OTUs are combined and the distance between this combined OTU (PC) and the remaining OTUs are computed by equation (3.4). Second, the OTU that has the smallest distance from the OTU (PC) is chosen. The appropriate OTU is H. After this choice, L_{PX}, L_{CX}, and L_{HX} are computed by applying equations (3.9a)–(3.9c).

We now proceed to the next step, where one more OTU is added to the unrooted tree for three OTUs. There are three possibilities for each remaining OTU (either G or O) to be connected to the tree. For example, OTU G may be connected to either branch PX, CX, or HX (see Figure 8). Thus lengths of all possible branches are computed and the branching pattern that gives the shortest length is chosen to be connected to the three-OTU tree. Branch lengths are computed in a similar way as equation (3.9):

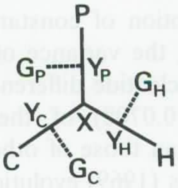

Fig. 8. Three possible additions of OTU G to P–C–H tree.

$$L_{G_P Y_P} = \tfrac{1}{2}(D_{GP} + L_{G_P X} - L_{PX}) , \tag{3.12a}$$

$$L_{G_C Y_C} = \tfrac{1}{2}(D_{GC} + L_{G_C X} - L_{CX}) , \tag{3.12b}$$

$$L_{G_H Y_H} = \tfrac{1}{2}(D_{GH} + L_{G_H X} - L_{HX}) , \tag{3.12c}$$

where subscripts of G and Y designate the positions of branch connecting OTU G (see Figure 8). In equation (3.12), L_{PX}, L_{CX}, and L_{HX} have already been computed at the previous step, whereas $L_{G_P X} = L_1$ or L_2, $L_{G_C X} = L_2$ or L_3, and $L_{G_H X} = L_1$ or L_3, where

$$L_1 = D_{GC} - L_{CX} , \quad L_2 = D_{GH} - L_{HX} , \quad L_3 = D_{GP} \quad L_{PX} . \tag{3.13}$$

Among L_1, L_2, and L_3, the largest value is used for all of the $L_{G_i X}$ (i = P, C, or H) in equation (3.12). Tateno et al. (1982) considered the use of the distance Wagner method for evolutionary distance that are not metric. In this case, a gross overestimation of branch length can happen by this procedure. Thus they used the average of L_1, L_2, and L_3, instead of the largest value. Faith (1985) took a different modification for estimating $L_{G,X}$ of equation (3.12). In this case, equation (3.9) is repeatedly used and the weighted average gives the estimates for these branch lengths.

In the present example, L_3 (= 0.0374) is the largest and putting this and the other values into equations (3.12a)–(3.12c), $L_{G_H Y_H}$ (= 0.0224) turns out to be the smallest. Thus OTU G is connected to the branch HX. The same procedure is repeated until all OTUs are connected. The final tree is presented in Figure 6(c).

As in the case of the FM method, the DW method also produces unrooted trees. Thus we can locate the root at the branch going to orangutan (O). When we have no information on the outgroup species, the location of the root can be estimated as the mid-point of the largest estimated distance, if a rough constancy of evolutionary rate is assumed (Farris, 1972). Estimated (patristic) distance between OTUs C and O (= 0.0970), that is considerably larger than the observed distance (D_{CO} = 0.0895), is the largest in the present example, and the root was placed at the point of which distance is 0.0485 ($= \frac{1}{2} \times 0.0970$) from node O (see Figure 6(c)). Under the assumption of constant evolutionary rate, the root can also be obtained by minimizing the variance of evolutionary rate (Farris, 1972).

Because the proportion of nucleotide difference was used for the DW method, the length (0.0223 + 0.0485 = 0.0708) of the branch going to OTU O of Figure 6(c) is slightly shorter than those of other trees in Figure 6. This is probably because Jukes and Cantor's (1969) evolutionary distances, in which multiple hits were corrected, were used in the latter trees. However, some of the lengths of the other branches of Figure 6(c) are larger than those of the other trees.

3.5. *Neighbor-joining method*

A pair of OTUs are called 'neighbors' when these are connected through a single interior node in an unrooted, bifurcating tree. For example, OTUs A and B in Figure 1(b) are a pair of neighbors. If we combine these OTUs, this combined OTU (AB) and OTU C become a new pair of neighbors. It is thus possible to define the topology of a tree by successively joining pairs of neighbors and producing new pairs of neighbors. For example, the topology of tree a in Figure 3 can be described by the following pairs of neighbors: [1, 2], [5, 6], [7, 8], [1–2, 3], and [1–2–3, 4]. Note that there is another pair of neighbors, [5–6, 7–8], that is complementary to [1–2–3, 4] in defining the topology. In general, $n - 2$ pairs of neighbors can be produced from a bifurcating tree of n OTUs.

The neighbor-joining (NJ) method of Saitou and Nei (1987) produces a unique final tree by sequentially finding pairs of neighbors. The algorithm of the NJ method starts with a starlike tree, as given in Figure 9(a), which is produced under the assumption that there is no clustering of OTUs. In practice, some pairs of OTUs are more closely related to each other than other pairs are. Consider a tree that is of the form given in Figure 9(b). In this tree the neighboring OTUs [1, 2] are separated from the other OTUs (3, 4, ..., 8) by branch XY. Any pair of OTUs can take the positions of 1 and 2 in the tree, and there are $\frac{1}{2}n(n-1)$

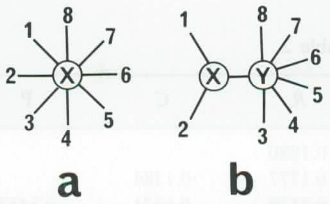

Fig. 9. A star-like tree (a) and a tree (b) that is one step aside from the star-like tree.

ways of choosing them for n OTUs. Among these possible pairs of OTUs, we choose the one that gives the smallest sum of branch lengths. Thus the principle of minimum evolution is used in the NJ method. This pair of OTUs is then regarded as a single OTU, and the next pair of OTUs that gives the smallest sum of branch lengths is again chosen. This procedure is continued until all $n - 2$ neighbors are found.

The sum of the branch lengths is computed as follows. First, the branch length between nodes X and Y in the tree of Figure 9(b) is estimated by

$$L_{XY} = \frac{1}{2(n-2)}\left[\sum_{k=3}^{n}(D_{1k} + D_{2k}) - (n-2)(L_{1X} + L_{2X}) - 2\sum_{i=3}^{n} L_{iY}\right]. \tag{3.14}$$

Noting the relationships $L_{1X} + L_{2X} = D_{12}$ and $\sum_{i=3}^{n} L_{iY} = [\sum_{3\leqslant i<j}^{n} D_{ij}]/(n-3)$, we find that the sum ($S_{12}$) of all branch lengths of the tree in Figure 9(b) becomes

$$\begin{aligned} S_{12} &= L_{1X} + L_{2X} + L_{XY} + \sum L_{iY} \\ &= \frac{1}{2(n-2)}\left[\sum_{k=3}^{n}(D_{1k} + D_{2k})\right] + \tfrac{1}{2}D_{12} + \frac{1}{n-2}\sum_{3\leqslant i<j} D_{ij}\,. \end{aligned} \tag{3.15}$$

It can be shown that equation (3.15) is the sum of least squares estimates of branch lengths for tree 9b (see Appendix A of Saitou and Nei, 1987). In general, we compute all S_{ij} $(1 \leqslant ij \leqslant n)$ and choose the pair of OTUs i and j that shows the smallest S_{ij} value.

Definition of S_{ij} seems complicated, but it can be computed in a simplified form

$$D_{ij} = -(R_i + R_j)/2(n-2) + \tfrac{1}{2}D_{ij} + Q/(n-2)\,, \tag{3.16}$$

where $R_i = \sum_{k=1}^{n} D_{ik}$, $R_j = \sum_{k=1}^{n} D_{jk}$, and $Q = \sum_{k<l}^{n} D_{kl}$. Because R_i $(1 \leqslant i \leqslant n)$ and Q can be computed before computation of S_{ij}'s, computation of S_{ij} is actually quite simple (see also Studier and Keppler, 1988). Note that $D_{ij} = D_{ji}$ and $D_{ii} = 0$ are assumed in the computation of R_i's.

Let us apply the NJ method to the evolutionary distance matrix of Table 2. $Q = 0.5803$, and R_i's are presented at the first column of Table 4. From these, S_{ij}'s were computed as shown in Table 4, and we find that S_{PC} ($= 0.1384$) is the

Table 4
R_i values and S_{ij} matrix for Table 2

	R_i	C	P	G	H
Common chimpanzee	0.1880				
Pygmy chimpanzee	0.1777	0.1384			
Gorilla	0.2179	0.1471	0.1483		
Human	0.2008	0.1477	0.1467	0.1422	
Orangutan	0.3762	0.1470	0.1469	0.1427	0.1437

smallest. Thus OTUs P and C are combined and the distance between the combined OTU (PC) and a remaining OTU *i* is computed by equation (3.4). The same procedure is repeated for the new distance matrix, and finally tree *d* of Figure 6 is obtained.

Algorithm of the NJ method is quite similar to that of UPGMA. Instead of choosing the smallest distance, we choose the smallest S_{ij} value at each step, and the distance averaging follows. Therefore the computation is very rapid.

When a distance matrix is strictly additive (any distance is sum of appropriate branch lengths), the NJ method was proved to reconstruct the true tree (Saitou and Nei, 1987; Studier and Keppler, 1988).

3.6. *Transformed distance methods*

When evolutionary rate varies from lineage to lineage in a phylogenetic tree as in a tree in Figure 10, the following distance transformation may give an improved topology for the average distance method (Farris, 1977)

$$D'_{ij} = \tfrac{1}{2}(D_{ij} - D_{i\mathrm{R}} - D_{j\mathrm{R}}) . \tag{3.17}$$

where R refers to the reference OTU. This property has been independently rediscovered by Klotz et al. (1979) and by Li (1981) (see also Klotz and Blanken, 1981).

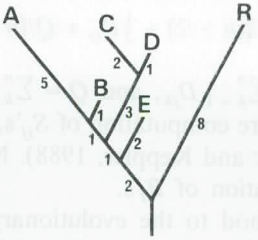

Fig. 10. A phylogenetic tree in which evolutionary rate varies considerably among different lineages. Figures are branch lengths.

The underlying logic of the transformation is as follows. If we change the sign of D'_{ij} of the above equation, $-D'_{ij}$ (a positive value) corresponds to the branch length between the reference OTU R and the interior node connecting OTUs i and j (see equation (3.9)). Thus if we apply UPGMA to the distance matrix, the correct topology should be obtained. The reference OTU can be a composite one that consists of more than one OTU.

Assuming the exact additivity, distances were computed from Figure 10 and they are shown in Table 5(a). If we apply UPGMA to this distance matrix, OTUs C and D are first clustered, but OTUs B and E will be erroneously clustered at the next step. Table 5(b) shows a transformed distance matrix, in which OTU R was treated as a reference. If we apply UPGMA to this matrix, OTUs C and D are first clustered, followed by OTUs A and B and OTUs (CD) and (AB). Thus it is clear that the transformation gives the correct tree topology.

Table 5
Original and transformed distance matrices based on the tree of Figure 8

	(a) Original distances (D)					(b) Transformed distances (D')			
	A	B	C	D	E	A	B	C	D
B	6					− 12			
C	11	7				− 11	− 11		
D	10	6	3			− 11	− 11	− 14	
E	9	5	8	7		− 10	− 10	− 10	− 10
R	17	13	16	15	12				

There is one problem in this method; we usually do not know which OTU is a reference or outgroup. Li (1981) used UPGMA for estimating the root of a tree, and two groups of OTUs separated by this root are alternatively used for transformation of distances of the other group. If the position of the root determined by UPGMA is correct, Li's method is expected to perform efficiently. However, it is possible that UPGMA misdetermines the root.

Let us apply Li's (1981) method to the data of Table 2. UPGMA tree (see Figure 6(a)) is first constructed, and five OTUs are divided into two groups according to the position of the root. Because one group consists of only one OTU (O), transformed distance matrix of Table 6 is computed for OTUs (P, C, H, and G) of the other group. It is clear that tree 3 (see Table 7) will be obtained if we apply UPGMA to this matrix. The tree finally obtained (not shown) is quite similar to tree 6d, which was obtained by the NJ method.

3.7. *Other methods*

Many other distance methods have been proposed for reconstruction of phylogenetic trees, and we briefly discuss some of them.

Table 6
Transformed distance matrix of Table 2

	C	P	G
Pygmy chimpanzee	− 0.0876		
Gorilla	− 0.0746	− 0.0733	
Human	− 0.0750	− 0.0759	− 0.0761

Let us consider an unrooted tree with four OTUs (see Figure 2(b)), and assume that every distance is the sum of relevant branch lengths or that the strict additivity holds. Then we have the relation $D_{12} + D_{34} < D_{13} + D_{24} = D_{14} + D_{23}$ for the leftmost tree of Figure 2(b). We can use a similar relationship for finding the tree,

$$D_{12} + D_{34} < D_{13} + D_{24} \quad \text{and} \quad D_{12} + D_{34} < D_{14} + D_{23} \, . \tag{3.18}$$

This condition is called the four-point metric (Buneman, 1971). The additive condition (Dobson, 1974) or the relaxed additivity condition (Fitch, 1981) is closely related to the four-point metric. It should be noted that the DW, the NJ, and the transformed distance method are all reduced to this condition in the case of four OTUs (Saitou and Nei, 1986, 1987).

Sattath and Tversky (1977) used the four-point metric for inferring tree topology for more than four OTUs. Interestingly, their method has a similarity with the NJ method. Fitch (1981) proposed a method also applying the four-point metric in a somewhat different way. Readers may refer to the original papers.

Edwards and Cavalli-Sforza (1965) proposed a method called 'cluster analysis'. The division of OTUs that gives the largest between-cluster sum of squares (or the smallest within-cluster sum of squares) is sequentially chosen in this method. There are $2^{n-1} - 1$ ways for n OTUs to be divided into two clusters, and all possibilities are examined. The same procedure is applied to each cluster thus found, and finally a rooted tree is obtained after $n - 1$ steps.

Cavalli-Sforza and Edwards (1967) proposed two exhaustive search methods (the additive tree and the minimum evolution methods) and they applied these methods to gene frequency data of human populations. The additive tree method assumes that distances along the tree are additive, and the least square method is used to minimize the errors between observed distances and estimated distances that are obtained by summing estimated branch lengths. This procedure is applied for all possible trees, and the tree that has the smallest sum of squares is chosen. The minimum evolution method for n OTUs is equivalent to the Steiner problem in $n - 1$ dimensions (see Courant and Robbins, 1941, for a review of Steiner problem). Computation of the additive tree and the minimum evolution methods are cumbersome when the number of OTUs is large.

Saitou and Imanishi (1989) proposed a simple method applying the principle of minimum evolution. In this method, branch lengths of a given tree are estimated by the procedure of Fitch and Margoliash (1967), and the tree with the smallest sum of branch lengths (SBL) is chosen as the best tree. It has been shown that the property of this method is similar to that of the NJ method (Saitou and Imanishi, 1989). An example of the minimum evolution (ME) method is shown in Table 7. The method chose tree 3 as the best one, as in the case of the NJ and ML methods. The ME method seems to be closely related to Dayhoff's (1978) method (see Blanken et al., 1982).

Table 7
Results of five exhaustive search methods for data of Table 2 (from Saitou and Imanishi, 1989)[a]

Method	Tree 1: ((PC)H)G	Tree 2: ((PC)G)H	Tree 3: (PC)(HG)	Tree 4: ((PH)C)G
FM	+ 0.60	0	+ 0.47	+ 25.33
TA	+ 0.35	0	+ 0.16	+ 4.45
ME	+ 0.35	+ 1.10	0	+ 6.59
MP	0	+ 1	0	+ 8
ML	− 2.98	− 3.97	0	− 33.86

[a] Values for FM, TA, and ME are PSD, $S_0 \times 1000$, and SBL × 1000, respectively. Values for MP is the required number of nucleotide substitutions, and those for ML is the log-likelihood. Values of the best tree are set to be zero, and the other values represent differences from that of the best tree.

3.8. *Statistical tests*

There are several methods for testing the statistical significance of a tree obtained. In the case of a UPGMA tree, SEs of the distances of branching points can be computed by Nei et al.'s (1985) method, as has been shown in Section 3.2. Thus by a standard *t*-test, the difference of the distances of the branching points Y and Z are shown to be not statistically significant, whereas those between X and Y and between Z and W are statistically significant in Figure 6(a).

Hasegawa et al. (1985) applied a generalized least square method for estimating branch lengths for a given tree, under the assumption of a constant evolutionary rate with their own model of nucleotide substitution, and gave equations for computing variances of estimated branch lengths. Later they also applied the bootstrap method (Felsenstein, 1985; see also Section 4.2) for computing variances (Hasegawa et al., 1987). Readers may refer to the original papers.

When we do not have the assumption of the constant rate of evolution, unrooted trees should be considered. In this case the estimation of SEs for each branch length is not easy. However, an approximate SE for each branch can be obtained by applying equation (3.7). That is, estimated branch length (d) is used for estimating the proportion ($\hat{p}$) of nucleotide difference, and this $\hat{p}$ is used to estimate SE of d using equation (3.3). If we apply this simple procedure to tree *d* of Figure 6, the length (0.0015) of the branch connecting the H–G cluster and the

CP–O cluster is not significantly greater than zero (its SE being 0.0013). Similarly, the length (0.0129) of the branch connecting the C–P cluster and HG–O cluster is significantly greater than zero (its SE being 0.0037). Thus the clustering of chimpanzee (C) and pygmy chimpanzee (P) is supported, whereas that of human (H) and gorilla (G) is not. It should be noted, however, that this method is expected to give a smaller SE than the true value. Thus the test based on this estimation is not conservative.

Templeton (1985) proposed a method (delta Q-test) for a statistical test of different tree topologies. However, Saitou (1986) and Ruvolo and Smith (1986) showed that the delta Q-test is theoretically unjustified. Thus this method is not recommended.

4. Character-state methods

4.1. Character states

Any discrete characters can be used for character-state methods, such as morphological characters, amino acid and nucleotide sequences, and restriction site maps. In principle, each character is considered separately in character-state methods. However, a more essential unit of comparison for the character-state method is 'configuration'. A configuration is a distribution pattern of characters for a given set of OTUs. If there are two characters (ancestral and derived), there are 2^n configurations for n OTUs. For the case of nucleotide sequences, there are four characters (A, G, T, and C) and the number (c) of configuration becomes

$$c = \tfrac{1}{6}(4^{n-1} + 3 \times 2^{n-1} + 2) \tag{4.1}$$

(Saitou and Nei, 1986). For example, there are 5, 15, and 51 configurations for 3, 4, and 5 sequences. Any length of nucleotide sequences for a given set of sequences can be described as an array of configurations, and the distribution pattern of the number of each configuration is essential for the construction of a tree.

The maximum parsimony method and the maximum likelihood method will be discussed in the following.

4.2. Maximum parsimony method

The evolutionary process of morphological characters can be classified into two different aspects. Groups of organisms similar in general levels of organization is called 'grade' and groups of common genetic origin is called 'clade' (Simpson, 1961). Cladistics or cladism is named after clade. However, cladists usually rely on the maximum parsimony method alone for finding the phylogenetic tree. A phylogenetic tree constructed by a cladistic approach is called 'cladogram'.

Camin and Sokal (1965) proposed the principle of maximum parsimony for reconstructing rooted trees from morphological characters. The tree that requires the smallest change of characters is chosen under this principle. When one considers a rooted tree, it is necessary to define the direction of the tree. This is done by determining the state of a character either to a derived one (apomorphy) or to a primitive one (plesiomorphy). A clade is defined by a synapomorphy, or a sharing of a derived state. Whether the state of a character is apomorphous or plesiomorphous depends on the opinion of each researcher. Thus there is a certain level of subjectiveness on the determination of direction of a tree.

On the study of molecular evolution, the direction of a tree is not easy to determine. Thus unrooted trees are usually constructed. An unrooted tree can be converted into a rooted tree by the knowledge of an outgroup OTU or by assuming the constancy of evolutionary rate as discussed earlier. Eck and Dayhoff (1966) proposed the maximum parsimony (MP) method for amino acid sequence data, and Fitch (1971) presented an algorithm for computing the minimum number of nucleotide substitutions for a given tree. A method of estimating the minimum number of nucleotide substitutions from amino acid sequence data was also proposed by Fitch and Farris (1974). Later Fitch (1977) clarified the properties of the MP method for nucleotide sequence data. There are several variations for the maximum parsimony method, and reader may refer to a comprehensive review by Felsenstein (1982).

Let us consider a tree for five nucleotide sequences, and assume that nucleotides A, A, T, G, and G were observed in sequences 1–5 in this order at one nucleotide site (see Figure 11). If tree 11a is considered, interior nodes X and Z should have nucleotides A and G, respectively, from the maximum parsimony principle. However, node Y can have either A, G, or T, because two nucleotide substitutions (denoted by full circles in Figure 11) are required in all three cases.

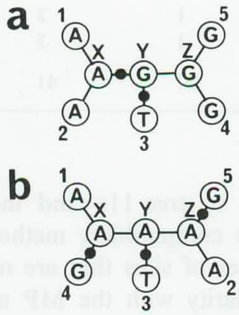

Fig. 11. Two trees (a and b) for five sequences. Nucleotides for each node are shown in circles, and dots denote nucleotide substitutions.

The case of nucleotide G at node Y is shown in tree 11a. If tree 11b is considered, we now need three nucleotide substitutions. Nodes X, Y, and Z can have nucleotides A or G simultaneously. This example shows that the nucleotides of ancestral or interior nodes may not be determined unambiguously, and the estimation of the length of each branch length is often difficult in the MP method.

In the above example, the minimum numbers of required nucleotide substitutions are different in trees 11a and 11b. This kind of nucleotide configuration is informative in choosing the best tree. A nucleotide configuration is 'informative' when there are at least two different kinds of nucleotides, each represented at least two times. Only these informative configurations are used in the MP method and non-informative ones are discarded. Non-informative configurations include invariant sites and singular sites in which only one nucleotides are represented more than one times (Fitch, 1977).

Certain nucleotide configurations can be fitted to a given tree with the minimum number of substitutions (the number of variable nucleotides minus one), whereas the other configurations require more than the minimum. The nucleotide sites with the first group of configurations are called compatible sites, and those with the latter group are called incompatible sites. The nucleotide site considered in the above example is compatible to tree 11a and incompatible to tree 11b. This difference is considered in the compatibility method (LeQuesne, 1969), and the tree that has the largest number of sites that are mutually compatible is chosen. Thus this method has a similarity with the MP method. When the number of OTUs is 4 or 5, these two methods are identical.

Table 8
Maximum parsimony analysis of Hixson and Brown's data

	Configuration[a]					Number of observations	Number of substitutions for			
	C	P	G	H	O		Tree 1	Tree 2	Tree 3	Tree 4
1	*y*	*y*	*x*	*x*	*x*	8	8	8	8	16
2	*x*	*x*	*y*	*x*	*y*	5	5	10	10	5
3	*x*	*x*	*y*	*y*	*x*	5	10	10	5	10
4	*x*	*x*	*x*	*y*	*y*	4	8	4	8	8
5	*x*	*y*	*x*	*y*	*x*	1	2	2	2	1
6	*x*	*y*	*x*	*x*	*y*	1	2	2	2	2
7	*y*	*x*	*y*	*x*	*x*	1	2	2	2	2
8	*y*	*x*	*x*	*x*	*y*	1	2	2	2	2
9	*x*	*x*	*y*	*y*	*z*	1	2	2	2	3
	Total					27	41	42	41	49

[a] *x*, *y* and *z* are different nucleotides.

Table 8 shows an example of the maximum parsimony analysis using Hixson and Brown's (1986) mitochondrial DNA sequence data. Only informative configurations are listed in the table, and the number of nucleotide sites involved is 27 out of 939 sites. Four trees of Table 7 were examined and trees 1 and 3 are equally parsimonious, though only one additional substitution is required for tree 2. Two chimpanzee species are not clustered in tree 4, and this tree requires much larger number of nucleotide substitutions.

Because the maximum parsimony principle is simple and it is philosophically related to Occam's razor, it has become very popular not only in classical taxonomy but in molecular evolution. When the overall amount of divergence is small, the MP method may be appropriate. However, a gross underestimation of branch lengths occurs when the amount of divergence is large (e.g., Saitou, 1989). Furthermore, the MP method is not appropriate for finding the tree topology in some cases. Felsenstein (1978b) showed a condition in which the MP method and the compatibility method is positively misleading. Thus we should be cautious for the use of the MP method and the compatibility method.

In the standard maximum parsimony method, all changes are equally weighted, since the method was originally applied for morphological data in which the probability of change of each character is rarely known. In nucleotide sequence data, however, we have some knowledge on the probability of nucleotide changes. For example, transitional changes have been known to dominate the substitution process in mitochondrial DNA. In this case, it may be more appropriate to apply the MP method only for transversional changes (e.g., Saitou and Nei, 1986). Noting this kind of property in the real data, Tateno (1990) proposed a general method for giving different weights to each change of nucleotides before applying the principle of maximum parsimony. Readers interested in this method may refer to the original paper.

Recently Lake (1987) proposed a method called 'evolutionary parsimony'. While the standard MP method focuses at signals (compatible configurations for a given tree), Lake's method calls attention to noises (incompatible configurations), and transitional changes and transversional changes are distinguished. Unfortunately, the evolutionary parsimony is applicable only for four OTUs at this stage. Readers who are interested in this method may refer to Lake (1987).

Cavender (1978, 1981) proposed a statistical test for the MP method. However, he considered only four OTUs, and the results seems to be not appropriate for real evolutionary data. Felsenstein (1985) introduced the bootstrap method for tree comparison. This method involves resampling characters from one's own data, with replacement, to create a series of artificial samples of the same size as the original data. The MP method is applied to each of these, and the variation among the resulting trees are taken to indicate the size of the error in the original data. For more detail, readers may refer to the original paper.

Templeton (1983) applied the Wilcoxon signed-rank test to the MP method for restriction site data, and concluded that the clustering of chimpanzee and gorilla are statistically significant by analyzing Ferris et al.'s (1981) data. However, Nei and Tajima (1985, 1987) indicated drawbacks of the MP method through a theoretical study, and a more detailed study of Li (1986) showed that Ferris et al.'s (1981) data were not enough for obtaining statistically significant clustering of chimpanzee and gorilla (see also Smouse and Li, 1988). Thus Templeton's (1983) method is not recommended.

When we compare relatively large number of OTUs and character states, it is necessary to use computers. Platnick (1988) reviewed two computer programs for the MP method: PAUP (version 3.0) by D. L. Swofford and Hennig86 (version 1.5) by J. S. Farris. PHYLIP (version 3.1) by J. Felsenstein also contains several programs for the MP method and its variations.

4.3. Maximum likelihood method

The maximum likelihood (ML) method of tree-making was originally proposed by Cavalli-Sforza and Edwards (1967) for gene frequency data. Later, Kashap and Subas (1974) applied the ML method for three amino acid sequences, assuming the constancy of evolutionary rate. Langley and Fitch (1974) also used the ML method for estimating the branch lengths of a given tree, and compared these estimates with those obtained by the MP method.

Felsenstein (1981) developed the ML method for finding an unrooted tree from nucleotide sequence data. Let us explain the principle of his method. Consider tree *b* of Figure 1 as an example. We first restrict our attention to a specific nucleotide site, and assume that nucleotide N_i was observed at exterior node i ($i =$ A, B, C, D, or E). On the other hand, nucleotide N_j at interior node j ($j =$ X, Y, or Z) is unknown, and it can be one of four nucleotides. Then the likelihood (L) of this site becomes

$$L = \sum_{N_Y} \left\{ g_Y P_{YC} \left[\sum_{N_X} P_{YX} P_{XA} P_{XB} \right] \left[\sum_{N_Z} P_{YZ} P_{ZD} P_{ZE} \right] \right\}, \tag{4.2}$$

where g_Y is the probability that node Y has nucleotide N_Y, $P_{ij} \equiv \Pr(N_i, N_j, L_{ij})$ is the probability of observing nucleotide N_i and N_j at nodes i and j, respectively, with the branch length L_{ij}. Summation is for four possible nucleotides, because N_X, N_Y, and N_Z are variables. To obtain $\Pr(N_i, N_j, L_{ij})$, we must specify the pattern of nucleotide substitution. If we use Jukes and Cantor's (1969) random substitution model,

$$\Pr(N_i, N_j, L_{ij}) = \begin{cases} \frac{1}{4} + \frac{3}{4}\exp(-4L_{ij}/3) & \text{if } N_i = N_j\,, & (4.3a) \\ \frac{1}{4} - \frac{1}{4}\exp(-4L_{ij}/3) & \text{if } N_i \neq N_j\,. & (4.3b) \end{cases}$$

It should be noted that the reversibility of time is assumed in the above formulation, a necessary assumption for unrooted trees. When different character-state data such as amino acid sequences or restriction sites are to be used, equation (4.3) should be modified by taking into account the nature of each character state. But the essential nature of the likelihood function of equation (4.2) remains the same.

Likelihood for each nucleotide site defined by equation (4.2) is then multiplied for all sites, and usually log-likelihood is computed for different set of branch lengths for a given tree topology, and the set that shows the highest log-likelihood is numerically searched. Fukami and Tateno (1988) proved that there exists a single ML point in the possible parameter range under the Jukes–Cantor model of nucleotide substitution.

Saitou (1990) showed that the ML estimate of the number of nucleotide substitutions between two nucleotide sequences is identical with that obtained by

Jukes and Cantor's (1969) and by Kimura's (1980) method. In the case of more than two sequences, however, this identity does not hold.

Original formulation of the ML method by Cavalli-Sforza and Edwards (1967) included the probability of tree topology, assuming a Yule process. Felsenstein (1981) took a different procedure, in which the ML value for each tree is compared and the tree with the highest ML value is chosen. Nei (1987) argued that the ML value computed in this way is conditional for each tree. Recently Hasegawa and Kishino (1988) tried to justify Felsenstein's (1981) procedure by applying an information theory. When we consider a gene tree or gene genealogy within a population, however, the probability of observing a specific tree topology should be considered (see Tajima, 1983). Noting this theoretical problem, Saitou (1988) proposed a step-wise tree searching algorithm for the ML method. This is similar to that of the NJ method, in which a star-like tree is first considered. The final tree is nested from a previous tree with a trifurcation, and the difference in the maximum log-likelihood values between the two trees can be used for hypothesis testing. Yet even this procedure has some theoretical problem (see Saitou, 1989b). Thus we should be cautious of conducting a statistical test based on the ML method.

Let us apply the ML method to Hixson and Brown's (1986) data. Program DNAML of Felsenstein's PHYLIP version 3.1 was used to obtain ML values for four trees of Table 7. The transition/transversion ratio was set to be 5.0 and observed frequency of nucleotides were used (Saitou and Imanishi, 1989). Tree 3 showed the highest likelihood value among four trees, and tree 4 was the worst. Interestingly, the rank of these trees in terms of the ML values is identical with that of the minimum evolution method (see Table 7), though the estimated branch lengths (not shown) by the ML method were somewhat different to those of the NJ method (Figure 6(d)).

Felsenstein (1987) developed the 'maximum likelihood' method for DNA–DNA hybridization data. However, he considered several components of experimental errors, and this method is closely related to analysis of variance. Thus it may not be considered as a standard ML method.

5. Relative efficiency of tree-making methods

Many different methods have been proposed for reconstructing phylogenetic trees, as reviewed above. Then which method should we use? It is generally difficult to compare different tree-making methods using actual data, because we rarely know the true phylogenetic tree. Therefore, the relative efficiencies of tree-making methods should be studied through computer simulated data, in which the true tree is known. For example, Peacock and Boulter (1975) simulated amino acid sequence data, Tateno and Nei (1978) simulated nucleotide sequence data, and Nei et al. (1983) simulated gene frequency data. More recently, Fiala and Sokal (1985) simulated morphological data by specifying a transition probability model.

Tateno et al. (1982) did a comprehensive study of tree reconstruction from nucleotide sequences. They considered a phylogenetic tree for eight or more sequences, and a Poisson process was mainly used to simulate nucleotide substitutions for 300 nucleotides. They compared four distance methods (UPGMA, FM, DW, and a modified DW). Their results indicated that the efficiency of each method depended on various conditions. A similar but more extensive studies have been done by Tateno (1985), Tateno and Tajima (1986), and Sourdis and Krimbas (1987). Using a similar scheme of simulation as Tateno et al. (1982) developed, Saitou and Nei (1987) compared six distance methods (UPGMA, DW, a modified DW, Li's (1981) method, Sattath and Tversky's (1977) method, and the NJ method), and they showed that their NJ method and Sattath and Tversky's method were generally better than the other distance methods.

Blanken et al. (1982) considered an addition of one nucleotide sequence to the known phylogenetic tree, thus it is different from ordinary problem of tree construction.

Saitou and Nei (1986) considered trees for relatively small number (up to five) of nucleotide sequences, and derived the expected proportion of each nucleotide configuration for a given tree. Using this information they simulated a multinomial sampling of nucleotides to obtain the simulated sequences. The number of nucleotides required to obtain the correct tree with a probability of 95% has been examined for UPGMA, the FM method, the DW method and the transformed distance method (or the four point metric), and the MP method. Their results for unrooted trees for four sequences show that UPGMA and the FM method are inferior to the other methods. Li (1986) did a similar study for restriction site data.

Hasegawa and Yano (1984) and Saitou (1988) compared the MP and ML methods for the case of four nucleotide sequences, and they showed that the ML method can find the correct tree with an appreciable proportion when the MP method is positively misleading in the sense of Felsenstein (1978b).

Sourdis and Nei (1988) extended Saitou and Nei's (1987) study by including Faith's (1985) modification of the DW method and the MP method for comparison. They showed that the MP method was generally inferior to some distance methods such as the NJ method. For the case of the MP method, they examined trees that are close to the true tree, but this strategy has been shown to be effective by a preliminary study in which all possible trees were examined (Sourdis and Nei, 1988).

Recently, Saitou and Imanishi (1989) compared five exhaustive search methods (MP, ML, FM, FM using S_0, and the minimum evolution (ME) method using SBL) with the NJ method under the model tree for six sequences, and all 105 unrooted trees were examined, except for the ML method in which a limited number of trees were examined. They showed that the NJ, ME, and ML methods performed better than the other three methods. This result was obtained when the evolutionary distance was used for distance methods. When the proportion of nucleotide difference (a metric) was used, all distance methods showed a poor performance. Li et al. (1988) compared the NJ method with Lake's evolutionary parsimony in the case of four OTUs. However, they used only the proportion of nucleotide difference. Therefore, the validity of their conclusion is questionable.

In summary, popular methods such as UPGMA and the FM method have been shown to be generally inferior to the other methods. Considering the computation time when a relatively large number of OTUs is compared, a stepwise clustering method such as the NJ method seems to be the first choice for researchers interested in molecular phylogeny.

Acknowledgments

I thank Dr. Masatoshi Nei for his helpful comments on the manuscript. It should also be mentioned that the overall structure and some details of this review followed chapter 11 of his book (Nei, 1987). This work was partially supported by grants from the Ministry of Education, Science, and Culture of Japan.

References

Adams, E. N. (1972). Consensus techniques and the comparison of taxonomic trees. *Syst. Zool.* **21**, 390–397.

Blanken, R. L., Klotz, L. C. and Hinnebusch, A. G. (1982). Computer comparison of new and existing criteria for constructing evolutionary trees from sequence data. *J. Mol. Evol.* **19**, 9–19.

Buneman, P. (1971). The recovery of trees from measurements of dissimilarity. In: F. R. Hodson, D. G. Kendall and P. Tautu (eds.), *Mathematics in the Archeological and Historical Sciences*, Edinburgh University Press, Edinburgh, 387–395.

Camin, J. H. and Sokal, R. R. (1965). A method for deducing branching sequences in phylogeny. *Evolution* **19**, 311–326.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967). Phylogenetic analysis: Models and estimation procedures. *Am. J. Hum. Genet.* **19**, 233–257.

Cavender, J. A. (1978). Taxonomy with confidence. *Math. Biosci.* **40**, 271–280.

Cavender, J. A. (1981). Tests of phylogenetic hypothesis under generalized models. *Math. Biosci.* **54**, 217–229.

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

Courant, R. and Robbins, H. (1941). *What is mathematics?* Oxford University Press, Oxford.

Darwin, C. (1859). *On the origin of species*. John Murray, London.

Dayhoff, M. O. (1978). Survey of new data and computer methods of analysis. In: M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington DC, 327.

de Soete, G. (1983). A least squares algorithm for fitting additive trees to proximity data. *Psychometrika* **48**, 621–626.

Dobson, A. J. (1974). Unrooted trees for numerical taxonomy. *J. Appl. Prob.* **11**, 32–42.

Eck, R. and Dayhoff, M. O. (1966). In: M. O. Dayhoff (ed.), *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Silver Springs, MD.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). A method for cluster analysis. *Biometrics* **21**, 362–375.

Elwood, H. J., Olsen, G. J. and Sogin, M. L. (1985). The small-subunit ribosomal RNA gene sequences from the hypotrichous ciliates *Oxytricha nova* and *Stylonychia pustulata*. *Mol. Biol. Evol.* **2**, 399–410.

Faith, D. P. (1985). Distance methods and the approximation of most-parsimonious trees. *Syst. Zool.* **34**, 312–325.

Farris, J. S. (1972). Estimating phylogenetic trees from distance matrices. *Am. Natur.* **106**, 645–668.

Farris, J. S. (1977). On the phenetic approach to vertebrate classification. In: M. K. Hecht, P. C. Goody and B. M. Hecht (eds.), *Major Patterns in Vertebrate Evolution*, Plenum Press, New York, 823–850.

Felsenstein, J. (1978a). The number of evolutionary trees. *Syst. Zool.* **27**, 27–33.

Felsenstein, J. (1978b). Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**, 401–410.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376.

Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Quart. Rev. Biol.* **57**, 379–404.

Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* **39**, 783–791.

Felsenstein, J. (1987). Estimation of hominoid phylogeny from a DNA hybridization data set. *J. Mol. Evol.* **26**, 123–131.

Ferris, S. D., Wilson, A. C. and Brown, W. M. (1981). Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* **78**, 2432–2436.

Fiala, K. L. and Sokal, R. R. (1985). Factors determining the accuracy of cladogram estimation: Evaluation using computer simulation. *Evolution* **39**, 609–622.

Fitch, W. M. (1970). Distinguishing homologous from analogous proteins. *Syst. Zool.* **19**, 99–113.

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

Fitch, W. M. (1977). On the problem of discovering the most parsimonious tree. *Am. Natur.* **111**, 223–257.

Fitch, W. M. (1981). A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.* **18**, 30–37.

Fitch, W. M. and Farris, J. S. (1974). Evolutionary tree with minimum nucleotide replacements from amino acid sequences. *J. Mol. Evol.* **3**, 263–278.

Fitch, W. M. and Margoliash, E. (1967). Construction of phylogenetic trees. *Science* **155**, 279–284.

Fukami, K. and Tateno, Y. (1989). On the maximum likelihood method for estimating molecular trees: Uniqueness of the likelihood point. *J. Mol. Evol.* **28**, 460–464.

Hasegawa, M., Kishino, H. and Yano, T. (1985). Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**, 160–174.

Hasegawa, M., Kishino, H. and Yano, T. (1987). Man's place in Hominoidea as inferred from molecular clocks of DNA. *J. Mol. Evol.* **26**, 132–147.

Hasegawa, M. and Yano, T. (1984). Maximum likelihood method of phylogenetic inference from DNA sequence data. *Bull. Biomet. Soc. Jpn.* **5**, 1–7.

Hasegawa, M. and Kishino, T. (1989). Confidence limits on the maximum likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution* **43**, 672–677.

Hixson, J. and Brown, W. M. (1986). A comparison of the small ribosomal RNA genes from the mitochondrial DNA of the great apes and humans: sequence, structure, evolution, and phylogenetic implications. *Mol. Biol. Evol.* **3**, 1–18.

Hudson, R. R. and Kaplan, N. L. (1985). Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* **111**, 147–164.

Jukes, T. H. and Cantor, C. R. (1969). Evolution of protein molecules. In: H. N. Munro (ed.), *Mammalian Protein Metabolism*, Academic Press, New York, 21–132.

Kashyap, R. L. and Subas, S. (1974). Statistical estimation of parameters in a phylogenetic tree using a dynamic model of the substitutional process. *J. Theor. Biol.* **47**, 75–101.

Kimura, M. and Ohta, T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *J. Mol. Evol.* **2**, 87–90.

Klotz, L. C. and Blanken, R. L. (1981). A practical method for calculating evolutionary trees from sequence data. *J. Theor. Biol.* **91**, 261–272.

Klotz, L. C., Komar, N., Blanken, R. L. and Mitchell, R. M. (1979). Calculation of evolutionary trees from sequence data. *Proc. Natl. Acad. Sci. USA* **76**, 4516–4520.

Lake, J. A. (1987). A rate-independent technique for analysis of nucleic acid sequence: Evolutionary parsimony. *Mol. Biol. Evol.* **4**, 167–191.

Langley, C. H. and Fitch, W. M. (1974). An examination of the constancy of the rate of molecular evolution. *J. Mol. Evol.* **3**, 161–177.

LeQuesne, W. J. (1969). A method of selection of characters in numerical taxonomy. *Syst. Zool.* **18**, 201–205.

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

Li, W. H. (1986). Evolutionary change of restriction cleavage sites and phylogenetic inference. *Genetics* **113**, 187–213.

Li, W. H., Wolfe, K. H., Sourdis, J. and Sharp, P. M. (1989). Reconstruction of phylogenetic trees and estimation of divergence times under nonconstant rates of evolution. Cold Spring Harbor Symp. Quant. Biol. (in press).

Nei, M. (1972). Genetic distance between populations. *Am. Natur.* **106**, 283–292.

Nei, M. (1975). *Molecular Population Genetics and Evolution*, North-Holland, Amsterdam.

Nei, M. (1987). *Molecular Evolutionary Genetics*, Columbia University Press, New York.

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

Nei, M. and Tajima, F. (1985). Evolutionary change of restriction cleavage sites and phylogenetic inference for man and apes. *Mol. Biol. Evol.* **2**, 189–205.

Nei, M. and Tajima, F. (1987). Problems arising in phylogenetic inference from restriction-site data. *Mol. Biol. Evol.* **4**, 320–323.

Nei, M., Tajima, F. and Tateno, Y. (1983). Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J. Mol. Evol.* **19**, 153–170.

Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5**, 568–583.

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

Platnick, N. I. (1988). Programs for quicker relationships. *Nature* **335**, 310.

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

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Bioscience* **53**, 131–147.

Ruvolo, M. and Smith, T. F. (1986). Phylogeny and DNA–DNA hybridization. *Mol. Biol. Evol.* **3**, 285–289.

Saitou, N. (1986). On the delta *Q*-test of Templeton. *Mol. Biol. Evol.* **3**, 282–284.

Saitou, N. (1988). Property and efficiency of the maximum likelihood method for molecular phylogeny. *J. Mol. Evol.* **27**, 261–273.

Saitou, N. (1989). A theoretical study of the underestimation of branch lengths by the maximum parsimony principle. *Syst. Zool.* **38**, 1–6.

Saitou, N. (1990). Maximum likelihood methods. *Methods in Enzymology* **183**, 584–598.

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

Saitou, N. and Nei, M. (1986). The number of nucleotides required to determine the branching order of three species with special reference to the human–chimpanzee–gorilla divergence. *J. Mol. Evol.* **24**, 189–204.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425.

Sattath, S. and Tversky, A. (1977). Additive similarity trees. *Psychometrika* **42**, 319–345.

Simpson, G. G. (1961). *Principles of Animal Taxonomy*, Columbia University Press, New York.

Smouse, P. E. and Li, W. H. (1987). Likelihood analysis of mitochondrial restriction-cleavage patterns for the human–chimpanzee–gorilla trichotomy. *Evolution* **41**, 1162–1176.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy*, Freeman, San Francisco, CA.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *Univ. Kans. Sci. Bull.* **28**, 1409–1438.

Sourdis, J. and Krimbas, C. (1987). Accuracy of phylogenetic trees estimated from DNA sequence data. *Mol. Biol. Evol.* **4**, 159–166.

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

Studier, J. A. and Keppler, K. J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* **5**, 729–731.

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

Takahata, N. and Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics* **110**, 325–344.

Tateno, Y. (1985). Theoretical aspects of molecular tree estimation. In: T. Ohta and K. Aoki (eds.), *Population Genetics and Molecular Evolution*, Japan Scientific Society Press, Tokyo, 293–312.

Tateno, Y. (1990). A method for molecular phylogeny construction by direct use of nucleotide sequence data. *J. Mol. Evol.* **30**, 85–90.

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

Tateno, Y., Nei, M. and Tajima, F. (1982). Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. *J. Mol. Evol.* **18**, 387–404.

Tateno, Y. and Tajima, F. (1986). Statistical properties of molecular tree construction methods under the neutral mutation model. *J. Mol. Evol.* **23**, 354–361.

Templeton, A. R. (1983). Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and apes. *Evolution* **37**, 221–244.

Templeton, A. R. (1985). The phylogeny of the hominoid primates: a statistical analysis of the DNA–DNA hybridization data. *Mol. Biol. Evol.* **2**, 420–433.