From: *Molecular Biology: Current Innovations and Future Trends Part 2.*
ISBN 1-898486-03-4 ©1995 Horizon Scientific Press, Wymondham, U.K.

8

METHODS FOR BUILDING PHYLOGENETIC TREES OF GENES AND SPECIES

Naruya Saitou

Abstract

This chapter deals with phylogenetic trees of genes and species, that are fundamental not only for evolutionary studies but for molecular biology in general. The mathematical properties of phylogenetic trees such as the difference between rooted and unrooted trees and the number of possible tree topologies are first explained. Then the biological properties of phylogenetic trees in general is discussed with special reference to the difference between gene trees and species trees. The next section gives description of various tree-building methods such as the unweighted pair group method with arithmetic mean (UPGMA), the neighbor-joining, the maximum parsimony, and the maximum likelihood methods with worked-out examples. Results from computer simulation studies and statistical tests of estimated phylogenetic trees follow. Introduction of various computer packages for tree-building analyses and future trends are given at the end.

Introduction

The supply of mutations to the continuous flow of self replication of genetic materials (DNA or RNA) is fundamental for organismal evolution. This process is most faithfully described in the phylogenetic relationship of genes. In fact, the semiconservative replication of the DNA double helix automatically produces a bifurcating genealogy of genes. Because every organism is the product of eons of evolution, we are unable to grasp the full characteristics of living beings without understanding the evolutionary history of genes and organisms. It is thus clear that the reconstruction of the phylogeny of genes is essential not only for the study of evolution but also for biology in general.

It should be emphasized that the genealogical relationship of genes is independent of the mutation process, especially when neutral evolution (1) is considered. The former is a direct product of DNA replication and always exists, while the latter, including any kind of mutational event, may or may not happen within a certain time period and DNA region. Therefore, even if several nucleotide sequences happen to be identical, there must be a genealogical relationship for those sequences. However, it is impossible to reconstruct the genealogical relationship without mutational events. In this respect,

the extraction of mutations from genes and their products is also important for reconstructing phylogenetic trees. The advancement of molecular biotechnology has made it possible to routinely produce nucleotide sequences. We will therefore focus on the analysis of nucleotide sequences, however other molecular data can also be used.

Because of the limitations of space, this chapter has focused on basic concepts and recent developments. Interested readers are advised to read more extensive reviews such as Nei (2), Felsenstein (3), Swofford and Olsen (4), and Saitou (5).

General Properties of Phylogenetic Trees

Some Formal Characteristics of a Tree

A phylogenetic tree is literally a 'tree' in graph theory. A graph is composed of node(s) and branch(s). There should be only one path between any two nodes on a tree (see Figure 1). In evolutionary studies, a node represents a gene, species, or population depending on the purpose, a branch represents the topological relationship between nodes, and branch length represents mutational changes or evolutionary time. Nodes are divided into external and internal ones (see Figure 1); the former are often referred to as 'operational taxonomic units' (OTUs). Branches are also divided into external and internal ones. An external branch connects an external node and an internal node, for example branch AX of Figure 1, while an internal branch connects two internal nodes such as branch XY of Figure 1.

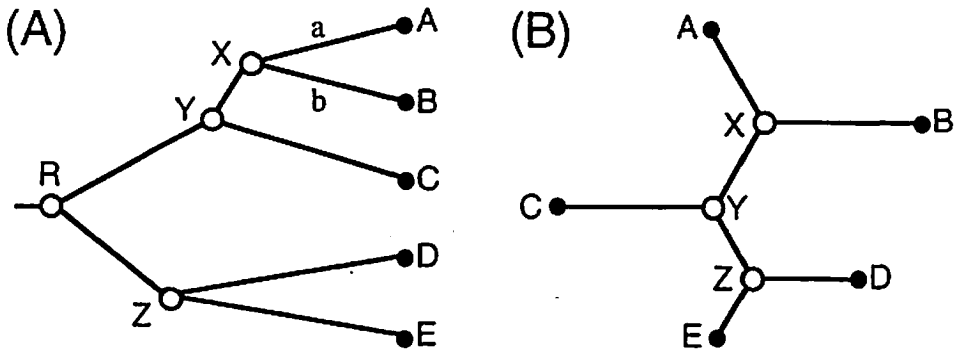

Figure 1. Examples of a rooted tree (A) and an unrooted tree (B) for five OTUs or external nodes. Full circles represent external nodes while empty circles represent internal nodes.

A tree can be either rooted or unrooted. A rooted tree has a special node called a root which is defined as the position of the common ancestor. There will be a unique path from the root to any other node, and the direction of this is of course that of time. Figure 1A shows an example of a rooted tree, in which the root is designated as R. A phylogenetic tree in an ordinary sense is a rooted tree. Unfortunately, however, many methods for building phylogenetic trees produce unrooted trees. An unrooted tree does not have a root, but it can be converted to a rooted tree if the position of the root is specified. Figure 1B is an example of an unrooted tree, and the topological relationship of nodes is identical to that of Figure 1A if we ignore its root (R).

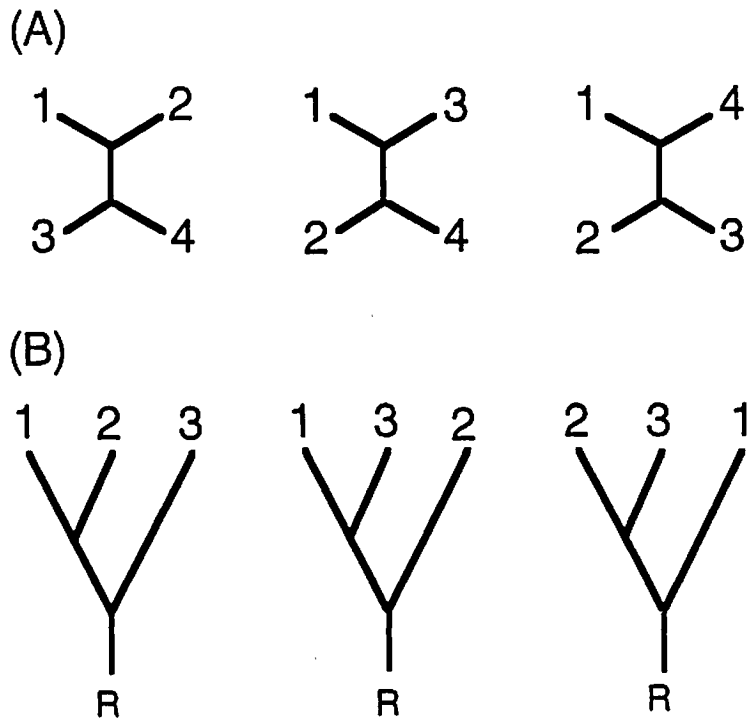

Figure 2. (A) Three possible unrooted trees for four OTUs. (B) Three possible rooted trees for three OTUs.

Figure 2A shows the three possible unrooted tree topologies for four OTUs, and these three unrooted trees and the rooted trees in Figure 2B have a one-to-one correspondence. If we designate the root of each tree in Figure 2B as node R, this is topologically identical with node 4 of Figure 2A. This relationship between rooted and unrooted trees is used for the "outgroup" method of rooting as follows. When we are interested in determining the phylogenetic relationship among the three sequences (or species) 1-3, we will add another one (sequence 4), that is known to be the outgroup to 1-3. The unrooted tree thus built can easily be converted to a rooted tree.

The number of possible tree topologies rapidly increases with an increasing number of OTUs. The general equation for the possible number of topology for bifurcating rooted trees [Nr(n)] and for unrooted trees [Nu(n)] for n (≥3) OTUs is given by

$$Nr(n) = 1\times3\times5\times \ldots \times(2n-3), \qquad \text{Equation 1a}$$

$$Nu(n) = 1\times3\times5\times \ldots \times(2n-5). \qquad \text{Equation 1b}$$

Table 1 gives the possible number of unrooted bifurcating tree topologies for up to 20 OTUs. It is clear that the search for the true phylogenetic tree of many OTUs is a very difficult problem. This is why so many methods have been proposed for building phylogenetic trees.

Table 1. Possible number of unrooted bifurcating tree topology

No. OTUs	No. of topology
3	1
4	3
5	15
6	105
7	945
8	10,395
9	135,135
10	2,027,025
11	34,459,425
12	654,729,705
13	13,749,310,575
14	316,234,143,225
15	7,905,853,580,625
16	213,458,046,676,875
17	6,190,283,353,629,375
18	191,898,783,962,510,625
19	6,332,659,870,762,850,625
20	221,643,095,476,699,771,875

Gene Trees and Species Trees

Phylogenetic trees of genes and species are called 'gene trees' and 'species trees', respectively, and there are several important differences between these. One such difference is illustrated in Figure 3. 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', which is homology of genes reflecting the phylogenetic relationship of species, from 'paralogy', which is homology of genes caused by gene duplication(s). Thus, genes 1A and 1B (and 2A and 2B) are 'orthologous', while genes 1A and 2B (and 1B and 2A) are 'paralogous'. If one is not aware of the gene duplication event, the gene tree for 1A and 2B may be misrepresented as the species tree of A and B, and thus a gross overestimation of the divergence time may occur.

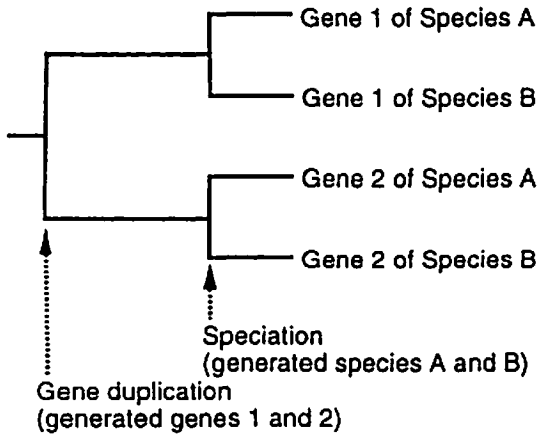

Figure 3. A gene tree for four genes sampled from two species.

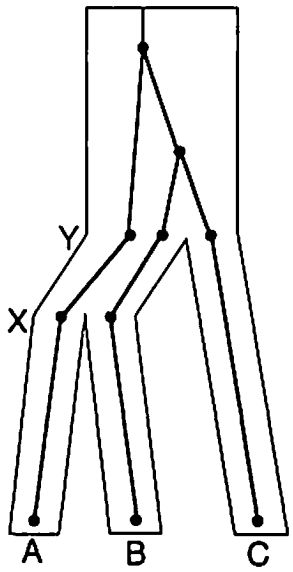

Figure 4. A possibility for topological difference between a gene tree and species tree is represented. Full circles and thick lines denote a gene tree, while thin lines (outlining the gene tree) denote the species tree. A, B, and C denote extant species, while X and Y denote two speciation times.

Even when orthologous genes are used, a gene tree may be different from the corresponding species tree. This difference comes from the existence of gene genealogy in the ancestral species. A simple example is illustrated in Figure 4. A gene sampled from species A has its direct ancestor at the speciation time X, and so does a gene sampled from species B. Thus the divergence between the two genes sampled from the different species always overestimates that of species. The amount of overestimation is related to the population size of the ancestral species X. If two speciation events between X and Y are close enough, the topological relationship of the gene tree may become different from that of species tree, as shown in Figure 4. Although species A and B are more closely related to each other than to C, the genes from species B and C are more closely related than to that from species A; see (2) for details.

When gene conversion and/or recombination has occurred within the gene region under consideration, the gene tree may be different from the species tree. Kawamura *et al.*(6) examined primate immunoglobulin alpha genes 1 and 2 (see Figure 5). Two gorilla genes were both G at a particular nucleotide site, while the remaining genes were C. This suggests either parallel substitution in the gorilla lineage or gene conversion between two gorilla genes occured. If this kind of nucleotide configuration is contiguous, gene conversion is suspected. The resulting gene tree may be distorted if the effect of gene conversion and/or recombination is strong, as was observed by Kawamura *et al.*

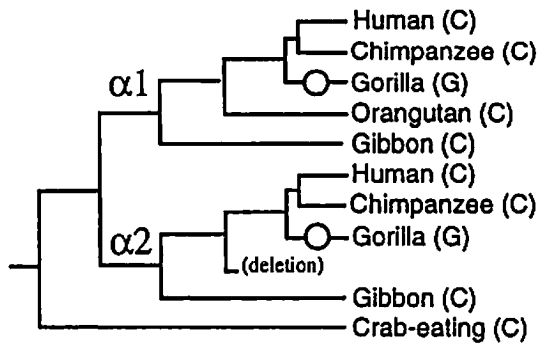

Figure 5. A nucleotide configuration possibly caused by gene conversion observed in primate immunoglobulin α1 and α2 genes (modified from reference 6).

Ideally the branch lengths of a phylogenetic tree are proportional to the physical time since divergence. Thus the branch a and b of Figure 1A should be the same length. We call this type of tree the 'expected tree'; this is a rooted tree. Both species and gene trees have their expected trees, but their properties are somewhat different from each other. An expected gene tree directly reflects the history of DNA replications, while an expected species tree is a gross simplification of the course of differentiation of populations. Therefore, the speciation time is not always clear.

As emphasized in the "Introduction", the genealogical relationship of genes, or the expected gene tree, is independent of the mutation process. However, mutation events are essential for the reconstruction of phylogenetic trees. Thus we can at best estimate a gene tree according to the mutation events realized on its expected gene tree (Figure 6A). We call this ideal reconstruction of the gene tree as the "realized" gene tree (Figure 6B), while the reconstructed one from observed data is called "estimated" gene tree. Branch lengths of realized and estimated genes tree are proportional to mutational events. These mutational events are not necessarily proportional to physical time. Due to limitations of available information, estimated gene trees are often unrooted trees. By definition, expected gene trees are strictly bifurcating, while realized and estimated gene trees may be multifurcating. This is because of the possibility of no mutation at a certain branch, such as branch X of Figure 6A.

A species tree reconstructed from observed data is called an "estimated" species tree, while there is no realized species tree.

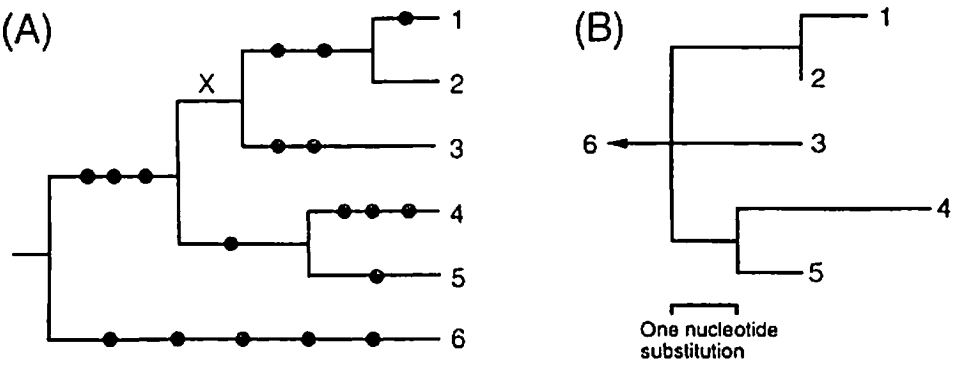

Figure 6. Expected (A) and realized (B) gene trees. Full circles on the expected gene tree denote nucleotide substitutions. Because no substitution occurred at branch X of the expected gene tree (A), the corresponding branch does not exist in the realized gene tree (B).

Methods for Building Phylogenetic Trees

Classification of Tree Building Methods

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-building 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 $n(n-1)/2$ distance values for n OTUs, whereas an array of character states is used for the character-state methods. The UPGMA (7), the Fitch and

Margoliash's method (8), the distance Wagner method (9) and its modification (10), the neighbor-joining method (11), the minimum evolution methods (12, 13, 14), and the split-decomposition method (15) are all distance matrix methods, whereas the maximum parsimony method (7, 16) and the maximum likelihood method (17, 18) are character-state methods.

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 tree topologies and choose the best one according to a certain criterion. We call this the 'exhaustive search method'. The Fitch and Margoliash's method, the minimum evolution methods, the maximum parsimony method, and the maximum likelihood method belong 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' (13). Most of the distance matrix methods, except Fitch and Margoliash's method and minimum evolution methods, are stepwise clustering methods.

In distance matrix methods, a phylogenetic tree is constructed by considering the relationship among the distance values of a distance matrix. An example of a distance matrix is presented in Table 2 (19). The data are mitochondrial DNA sequences for seven primate species. There are many methods for estimating evolutionary distances from molecular data such as amino acid and nucleotide sequences. Due to the limitation of space, this large area of study is omitted from this chapter. Reviews on this matter can be found elsewhere (2, 20, 21).

Table 2. Evolutionary distance (number of nucleotide substitutions per nucleotide site) matrix for seven primate species (from reference 19)

1	Human						
2	Chimpanzee	0.097					
3	Gorilla	0.114	0.118				
4	Orangutan	0.188	0.204	0.196			
5	Gibbon	0.215	0.228	0.227	0.226		
6	Rhesus macaque	0.292	0.323	0.293	0.315	0.296	
7	Squirrel monkey	0.364	0.380	0.354	0.368	0.347	0.396
		1	2	3	4	5	6

Methods Assuming the Molecular Clock

When the constancy of the evolutionary rate, or molecular clock, is assumed, we can reconstruct rooted trees. There are many ways to obtain such rooted trees from a distance matrix (see [7] for a review). In this section, only the UPGMA which is frequently used in molecular evolution is discussed.

Let us briefly explain the UPGMA algorithm using the distance matrix of Table 2. We first choose the smallest distance, D_{12} (= 0.097). Then OTUs 1 (human) and 2 (chimpanzee) are combined and the distances between the combined OTU [12] and the remaining five OTUs are computed by taking arithmetic means. At the next step, again the smallest distance ($D_{[12]3} = [0.114+0.118]/2 = 0.116$) is chosen from the distance matrix. Then the OTU [12] and OTU 3 are further combined into OTU [123]. This process is continued until all the OTUs are finally clustered into a single one. The resultant tree topology (not shown) is identical with that of Figure 7 in which the neighbor-joining method was used. This is because an approximate constancy of

evolutionary rate was satisfied for the distance matrix of Table 2. However, there are many cases in which an UPGMA tree and a neighbor-joining tree (and other trees assuming no constancy of evolutionary rate) are different.

There are two programs (KITCH and DNAMLK) in the PHYLIP computer package (18) that assume constancy of the evolutionary rate. KITCH is related to the Fitch and Margoliash's method, while DNAMLK is related to the maximum likelihood method. Interested readers may refer to the documentation of the PHYLIP package.

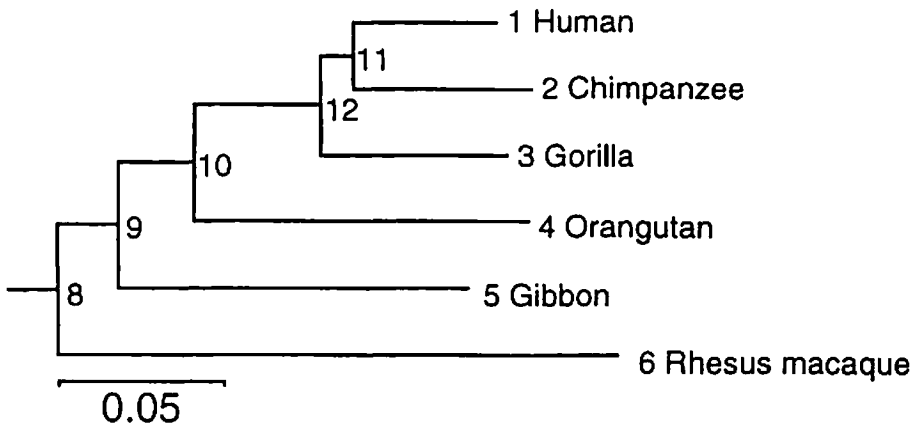

Figure 7. A neighbor-joining tree constructed from the distance matrix of Table 2. This tree was drawn based on the output shown in Table 3. Branch lengths are proportional to the number of nucleotide substitutions per branch.

Neighbor-joining and Minimum-evolution Methods

A pair of OTUs are called 'neighbors' when these are connected through a single internal node in an unrooted bifurcating tree. For example, OTUs A and B of Figure 1B 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. In general, n - 3 pairs of neighbors are necessary to define the topology of an unrooted tree with n OTUs.

The neighbor-joining method (11) produces a unique final unrooted tree by sequentially finding pairs of neighbors by examining a distance matrix. Thus the neighbor-joining method is a distance matrix method as well as a stepwise clustering method. The principle of minimum evolution is used in the neighbor-joining method, and recently Rzhetsky and Nei (22) proved that the expected value of the sum of branch lengths is smallest for the tree with the true branching pattern. Because of the simple algorithm, more than 100 OTUs can be handled within a relatively short computer time by using the neighbor-joining method. For example, Neefs *et al.*(23) produced the neighbor-joining tree for 1,348 rRNA sequences. This may be the current world record for the number of OTUs used for the neighbor-joining method.

The algorithm of the neighbor-joining method is as follows. We start from a starlike tree, which is produced under the assumption of no clustering among all the n OTUs compared. Under this tree, the sum (So) of n branch lengths can be shown to be

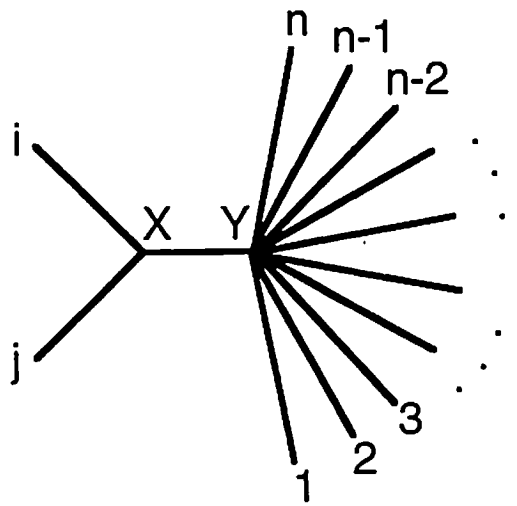

Figure 8. A tree of N OTUs in which OTUs i and j are neighbors.

$So = Q / (n - 1)$, Equation 2

where

$Q = \Sigma_{i<j} D_{ij}$. Equation 3

(Note: Σ repersents the summation)

In practice, some pairs of OTUs are more closely related to each other than other pairs are. Among all the possible pairs of OTUs (n[n-1]/2 pairs for n OTUs), we choose the one that gives the smallest sum of branch lengths. Let us consider the tree of Figure 8, where OTUs i and j are assumed to be neighbors. The sum of branch lengths is defined by

$$S_{ij} = (B_{iX} + B_{jX}) + B_{XY} + \Sigma_{k \neq i,j} B_{kY}, \quad \text{Equation 4}$$

where $B_{\alpha\beta}$ is branch length between nodes α and β. There are the following relationships between distances and branch lengths.

$D_{ij} = B_{iX} + B_{jX}$, Equation 5a

$D_{ik} = B_{iX} + B_{XY} + B_{kY}$ $(k \neq i,j)$, Equation 5b

$D_{jk} = B_{jX} + B_{XY} + B_{kY}$ $(k \neq i,j)$, Equation 5c

$D_{kl} = B_{iY} + B_{jY}$ $(k,l \neq i,j)$. Equation 5d

With the tree of Figure 8, it can be shown applying the above relationship,

$$B_{XY} = [Q - (n-1)D_{ij} - (n-1)\Sigma_{k,l \neq i,j} D_{kl} / (n-3)] / 2(n-2). \quad \text{Equation 6}$$

If we neglect OTUs i and j in Figure 8, the remaining n-2 OTUs form a star-like tree, as it is clear from equation 5d. Thus we apply equation 2 and obtain

$$\Sigma_{k \neq i,j} B_{kY} = \Sigma_{k,l \neq i,j} D_{kl} / (n - 3). \quad \text{Equation 7}$$

We also note that

$\Sigma_{k,l\neq i,j} D_{kl} = Q - (R_i + R_j - D_{ij})$, Equation 8

where

$R_i = \Sigma_j D_{ij}$, Equation 9a

$R_j = \Sigma_i D_{ij}$. Equation 9b

Putting equations 5a, 6, and 7 into equation 4 with considering equation 8, we obtain

$S_{ij} = D_{ij} / 2 + [2Q - R_i - R_j] / 2(n - 2)$. Equation 10

Equation 10 was first shown by Studier and Keppler (24).

This S_{ij} value is computed for all n(n-1)/2 pairs of OTUs, and the pair that has the smallest S_{ij} value is chosen as neighbors. This pair of OTUs is then regarded as a single OTU, and the new distances between the combined OTU and the remaining ones are computed by averaging. This procedure is continued until all pairs of neighbors are found.

If OTUs i and j are chosen as neighbors as shown in Figure 8, the branch lengths are estimated as

$B_{iX} = D_{ij} / 2 + (R_i - R_j) / 2(n - 2)$ Equation 11

and $B_{jX} = D_{ij} - B_{iX}$. Therefore, all the branch lengths as well as the tree topology will be determined after n-2 steps for n OTUs.

Table 3 shows the output of the computer program NJ when the distance matrix of Table 2 was used, and Figure 7 shows the neighbor-joining tree. Squirrel monkey (OTU 7) was assumed to be the outgroup.

Table 3. An output example of the program NJ (written by Saitou) for the evolutionary distance matrix of Table 2

Node 8	OTU 6 (0.169)	OTU 7 (0.227)
Node 9	OTU 5 (0.104)	Node 8 (0.018)
Node 10	OTU 4 (0.100)	Node 9 (0.022)
Node 11	OTU 1 (0.043)	OTU 2 (0.054)
Node 12 (Last node)		
Node 11 (0.010)	OTU 3 (0.056)	Node 10 (0.037)

The concept of minimum evolution was used in the neighbor-joining method, and this concept was first used by Cavalli-Sforza and Edwards (12). Saitou and Imanishi (13) proposed a simple method of applying the principle of minimum evolution. In this method, branch lengths of a given tree are estimated by applying the procedure of Fitch and Margoliash (8), and the tree with the smallest sum of branch lengths is chosen as the best tree. Rzhetsky and Nei (14) recently proposed a minimum evolution method in which branch lengths with their standard errors (SE) are computed by applying the least square method.

Maximum Parsimony Methods

There are several kinds of maximum parsimony methods based on various assumptions, but the maximum parsimony principle is used for all of these. We will only discuss the parsimony method that are frequently used for molecular data. This type of parsimony method produces unrooted trees, as in the case of the neighbor-joining method. The maximum parsimony principle is the minimization of the character-state changes on the given tree topology, and is related to the principle used in minimum evolution methods. However, the performance of these two methods in choosing the best topology can be quite different.

Let us consider an imaginary data set consisting of five sequences A-E, each 100 nucleotide-long. We first classify the 100 nucleotide sites into different configurations (see Table 4). A "nucleotide configuration" is a distribution pattern of nucleotides for a given number of sequences. The possible number (C_n) of configurations for n sequences is given by

$$C_n = (4^{n-1} + 3\times2^{n-1} + 2) / 6 \qquad \text{Equation 12}$$

(25). For example, there are 51 possible nucleotide configurations for 5 sequences.

Table 4. Application of the maximum parsimony method to an imaginary data set of 100 nucleotides

	Sequence						Number of substitutions for tree				
i	A	B	C	D	E	m_i[a]	1	2	3	4	5
Noninformative configuration:											
1	x	x	x	x	x	60	0	0	0	0	0
2	x	x	x	x	y	10	10	10	10	10	10
3	x	x	x	y	x	7	7	7	7	7	7
4	x	x	y	x	x	5	5	5	5	5	5
5	x	y	x	x	x	3	3	3	3	3	3
6	y	x	x	x	x	2	2	2	2	2	2
7	x	x	x	y	z	2	4	4	4	4	4
8	x	y	x	z	w	1	3	3	3	3	3
Informative configuration:											
9	x	x	x	y	y	5	5	5	5	10	10
10	x	x	y	y	y	2	2	4	4	2	4
11	x	y	x	y	y	2	4	2	4	4	4
12	x	x	y	z	z	1	2	3	3	3	3
			Total[b]			10	13	14	16	19	21

Note -- Topology of tree 1 = [AB]C[DE] (same as tree of Figure 1B), tree 2 = [AC]B[DE], tree 3 = [BC]A[DE], tree 4 = [AB]D[CE], and tree 5 = [BC]D[AE]. i = configuration.
a) Observed number of configuration i.
b) Informative configurations only.

Configuration 1 of Table 4 is an invariant one in which all of the five sequences have the same nucleotide x. The observed number of this configuration was 60. It is obvious that we do not need to assume any nucleotide substitution for this configuration, irrespective of the tree topology, under the maximum parsimony principle. In the case of nucleotide configuration 2 of Table 4, one substitution is necessary for any topology,

since only sequence E is different from the remaining sequences. This difference can be explained by assuming a substitution at the external branch going to sequence E. Thus the same 10 nucleotide substitutions are required for every topology. A similar situation holds for configurations 3-6.

There are three different nucleotides in configuration 7; two nucleotide substitutions on the external branches going to sequences D and E are necessary for this configuration. One may wonder that configuration 7 suggests a close phylogenetic relationship between sequences A-C. However, any tree topology requires the same number of substitutions under the maximum parsimony principle. Therefore, there is no discriminatory power for this configuration. The same is true for configuration 8.

Configurations that do not contribute to the selection of the tree topology are called "noninformative" for the maximum parsimony method. There are 100 nucleotides in the data set of Table 4, but 90 of them turned out to be noninformative and only the remaining 10 are "informative" configurations. An informative nucleotide configuration should have more than one kind of nucleotide and at least two of these should be observed in more than one of the sequences (16). There are four informative configurations in the data set of Table 4. Only one nucleotide substitution is required for configuration 9 when trees 1, 2, and 3 are assumed, while two substitutions are required for trees 4 and 5.

Let us consider tree 1 (see Figure 1B). One nucleotide substitution is required at the internal branches YZ and XY for configurations 9 and 10, respectively, while two substitutions are required for configuration 11. In the latter case, there are two possibilities for the location of substitutions. If all the three internal nodes (X-Z) are assumed to be nucleotide y, then the two substitutions must be located at external branches AX and CY. If internal nodes X and Y are assumed to be nucleotide x while node Z remains as y, substitutions are located at branches BY and YZ. It is thus clear that the branch lengths may not be determined unambiguously under the maximum parsimony method. When branch lengths are shown for a maximum parsimony tree, these are estimated under certain assumptions. These assumptions may not be realistic especially when largely diverged sequences are compared.

Because noninformative configurations do not contribute to the determination of the best topology, we consider only informative configurations. The total numbers of required nucleotide substitutions for each topology are given at the bottom of Table 4, and topology 1 requires the smallest number of substitutions. Although only five topologies out of 15 possible topologies are presented, topology 1 (see Figure 1B) is indeed the maximum parsimony tree.

The principle of maximum parsimony attracted many people because of its simplicity and logical clarity. However, there are some problems with this method when molecular data are used. Saitou (26) showed that gross underestimation of the branch lengths occured when the divergence (number of nucleotide substitutions per site) among sequences was larger than 0.2. This problem can be avoided if we use the maximum parsimony method only for determining tree topology. A more serious problem is its efficiency. Felsenstein (27) analytically showed that the maximum parsimony method may be positively misleading when the rate of evolution is grossly different among lineages of four sequences. When the expected number of required substitutions for the true tree is larger than that for a wrong one, the maximum parsimony method will give more and more wrong answers as the number of compared nucleotides is increased (problem of efficiency). Recently, the same problem was found even when

the constancy of the evolutionary rate is assumed (28, 29). Therefore, we should be careful when using the maximum parsimony method.

Maximum Likelihood Methods

The maximum likelihood method is often used for parameter estimation in statistics, and it was first applied to building phylogenetic trees by Cavalli-Sforza and Edwards (12) for allele frequency data. Later, various maximum likelihood methods and computer programs were developed for sequence data. The most frequently used one is Felsenstein's DNAML program (17, 18) for nucleotide sequences.

Let us explain the core algorithm of the maximum likelihood method. We first define the probability $P_{\alpha\beta} \equiv Pr(N_\alpha, N_\beta, B_{\alpha\beta})$ for observing nucleotide N_α and N_β at a particular nucleotide site at nodes α and β, respectively, when branch length is $B_{\alpha\beta}$. It is necessary to define the nucleotide transition matrix to compute $P_{\alpha\beta}$, but it is out of the scope of this chapter. In any case, we then compute the likelihood of a particular tree. For example, the likelihood (L_i) at nucleotide site i for the tree of Figure 1B under the given nucleotides and branch lengths becomes

$$L_i = \Sigma_{NY}\{g_Y P_{YC}[\Sigma_{NX} P_{YX} P_{XA} P_{XB}][\Sigma_{NZ} P_{YZ} P_{ZD} P_{ZE}]\}, \qquad \text{Equation 13}$$

where g_Y is the probability that node Y has nucleotide N_Y, and summation is for four possible nucleotides, for nucleotides at internal nodes X-Z are unknown. Because each nucleotide site is assumed to evolve independently, the likelihood values for all the nucleotide sites are multiplied to obtain the overall likelihood. As is usually done in maximum likelihood techniques, the logarithm of the likelihood (Log-L) is computed. Thus,

$$\text{Log-L} = \log [\Pi_i L_i] = \Sigma_i \log[L_i]. \qquad \text{Equation 14}$$

(Note: Π represents the multiplication of values)

This Log-L is computed by changing branch lengths, and the maximum likelihood solution is determined for this tree topology. This maximum likelihood solution is ideally obtained for all the possible topologies and the one that shows the highest value is chosen.

Table 5 is an example of the DNAML computation (user tree option was used). Using the data set of Table 4, topology 1 was found to have the highest likelihood value among the five topologies compared. It took about three minutes of computation when a Macintosh Centris 650 was used. The order of the likelihood values was the same as that of the required number of substitutions for the maximum parsimony method (see Table 4).

Table 5. Application of the maximum likelihood method for the data set of Table 4. DNAML of PHYLIP (ver 3.5) was used

Tree topology	Log-likelihood[a]
1: [AB]C[DE]	0
2: [AC]B[DE]	-1.87
3: [BC]A[DE]	-3.49
4: [AB]D[CE]	-6.87
5: [BC]D[AE]	-13.18

a) The log-likelihood value for the best tree topology (-336.11) was set to be zero, and differences with the best one are presented.

Because the maximum likelihood method requires massive computer time, there are several searching methods other than the exhaustive search. The default method of Felsenstein's DNAML program (17, 18) is the sequential addition of sequences. Saitou (30) proposed a stepwise clustering of sequences for the maximum likelihood method, and this searching method is the same as that of the neighbor-joining method. The NucML program of the MOLPHY package (31) has several options for topology searches, and one of them (star decomposition) is similar to that of Saitou's method (30).

Recently, DNAML was modified to speedup the computation, and the modified version is called fastDNAML (32). The computation speed of fastDNAML can be more than 100 times higher than DNAML, but NucML of MOLPHY (31) may be slightly faster than fastDNAML (J. Adachi, personal communication).

Other Methods

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

Fitch and Margoliash (8) proposed an exhaustive search method for distance matrix data. The criterion of choosing the best topology is "percent standard deviation" (PSD), defined as

$$PSD = [2\Sigma_{i<j}\{(D_{ij} - E_{ij})/D_{ij}\}^2 / n(n-1)]^{1/2} \times 100, \qquad \text{Equation 15}$$

where E_{ij} is the estimated distance between OTUs i and j. The algorithm of the program FITCH of PHYLIP (18) is based on this method, though the estimated distance is obtained after several cycles of optimization that were not included in the original Fitch and Margoliash method (8). There are many other variations of this method.

Farris (9) proposed a stepwise clustering method for distance matrix data and named it the "distance Wagner" method. Though the principle of minimum evolution is used for this method, its algorithm is quite different from that of the neighbor-joining method. A slight modification (the "modified Farris" method) of its algorithm was proposed by Tateno *et al.* (10).

Bandelt and Dress (15) proposed the "split decomposition" method for distance matrix data. Unlike most tree-building methods, it usually produces a network, not a tree. This is because a relaxed condition is used for estimating the splitting patterns

among OTUs. For example, let us consider the distance matrix data of human (H), chimpanzee (C), gorilla (G), and orangutan (O) sequences in Table 2. If we apply the neighbor-joining method, neighbors [H, C] (or [G, O]) are chosen, because $D_{HC} + D_{GO}$ (= 0.293) is smaller than either $D_{CG} + D_{HO}$ (= 0.306) or $D_{HG} + D_{CO}$ (0.318). In contrast, neighbors [C, G] (or [H, O]) are also kept when the split-decomposition method is applied. The resultant network (not tree) is shown in Figure 9. A short branch separating the human-orangutan pair from the chimpanzee-gorilla pair suggests the existence of some parallel nucleotide changes.

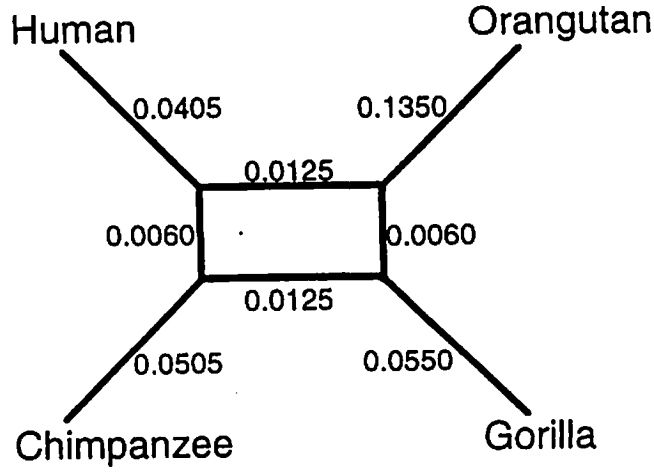

Figure 9. A four-OTU network constructed by using the split decomposition method. Numbers of nucleotide substitutions are given on each branch. Branches are not drawn to be proportional to their lengths.

Comparison of Methods

It is generally difficult to compare different tree-building methods using actual data, because we rarely know the true phylogenetic tree. Therefore, the relative efficiencies of various tree-making methods are usually studied through computer simulated data, in which the true tree is known. A considerable number of simulation studies have been conducted (5, 33), and we will only discuss some of recent developments.

DeBry (34) examined the UPGMA, the neighbor-joining, the modified Farris, and the maximum parsimony methods, and showed that the neighbor-joining method was consistent when perfect correction for evolutionary distances was made. Tateno *et al.* (35) compared the maximum-likelihood, the neighbor-joining, and the maximum parsimony methods using a simple four sequence tree with various assumptions. When the amount of divergence is small (nucleotide substitution of less than 0.05 per site), all the methods gave high efficiencies in estimating topologies and branch lengths. However the efficiency can be low when sequences with large divergences are compared.

Kuhner and Felsenstein (36) compared the Fitch and Margoliash, the maximum-likelihood, the neighbor-joining, the compatibility, and the maximum parsimony methods for 10 sequence data. The maximum likelihood method performed best among those five methods. The efficiency of the Fitch and Margoliash and the neighbor-joining methods were more or less the same as that of the maximum likelihood method, while the compatibility and the maximum parsimony methods had low efficiencies when substitution rates varied among branches.

Because any method may reconstruct erroneous trees when sequences with large divergence are compared, it is better to use various tree-building methods with different assumptions such as the pattern of nucleotide substitution.

Statistical Tests

There are several methods for statistically testing the validity of an estimated phylogenetic tree, and two of these are briefly explained in this section. One is a direct application of a standard statistic using variances of each branch lengths. Figure 10 shows an example of the minimum-evolution tree (14). The standard errors of branch lengths are shown, and this result suggests that birds and mammals are monophyletic, for the internal branch (0.80±0.29) clustering the two groups is significantly larger than zero. This clustering is, however, under controversy, and it is possible that the molecular data used to produce that tree may be biased. Therefore, we should be careful in applying a statistical test to phylogenetic trees.

Another test is the bootstrap method. This method was proposed for estimating variances from unknown probability distribution (37), and was introduced into the phylogeny (38). Character-state data are necessary to use the bootstrap method, but trees built by using any distance matrix method can be tested using this technique. We first randomly re-sample n nucleotide sites from the given sequence data of n nucleotides with replacement. This re-sampling is replicated at least 1,000 times. For example, one replication may have nucleotide sites 1, 1, 2, 4, 5, 7, 7, ... This re-sampling is usually done by generating pseudo-random numbers. Each replicated sequence data are then used as the input data to build phylogenetic trees. A bootstrap probability of a certain internal branch is simply the number of trees that realized this branch divided by the total number of replications. These probabilities are often summarized on the phylogenetic tree estimated by using the original sequence data.

Figure 11 shows an application of the bootstrap method to the neighbor-joining method (39). This tree was built by using the original sequence data, and numbers below each branch are the estimated number of nucleotide substitutions which occurred in this sequence. To obtain those numbers, estimated numbers of nucleotide substitution per site were multiplied with the number of compared nucleotide sites, then the resulting values were rounded. If a branch length turned out to be zero, that branch was neglected. Numbers above internal nodes are bootstrap probabilities (in %) based on 1,000 replications. For example, two HTLV-I sequences of Melanesia (Papua New Guinea and Solomon Islands) are clustered with a high bootstrap probability (91%). The bootstrap method is currently widely used, but its property on phylogenetic inference is not thoroughly known, and theoretical studies are still going on (e.g. 40, 41).

Computer Packages

MEGA (42) is a comprehensive package run on MS-DOS, and it is used to compute and draw UPGMA, neighbor-joining, and maximum parsimony trees in a user-friendly environment. Many kinds of evolutionary distance estimation methods can be used, including synonymous and nonsynonymous substitutions. For further information contact the following Email address: imeg@psuvm.psu.edu.

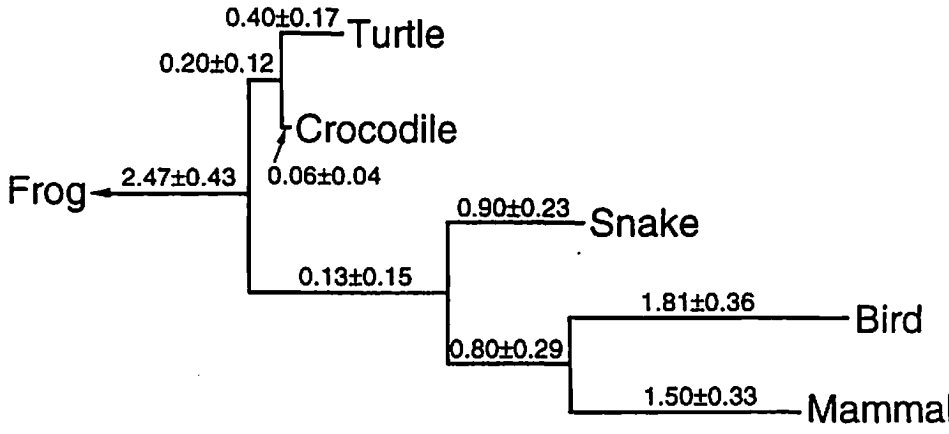

Figure 10. A minimum evolution tree of six land vertebrates (modified from reference 14). The numbers are branch lengths ±SEs. Frog was assumed to be the outgroup.

PHYLIP (18) contains many programs in the form of both source cord and executable files, and can be implemented into any kind of computer. Various kinds of maximum likelihood methods, maximum parsimony methods, and distance matrix methods can be used. It can be ftp-retrieved from the following IP address: evolution.genetics.washington.edu (128.95.12.41).

CLUSTAL V (43) is capable of doing multiple sequence alignment. After the alignment, it can construct neighbor-joining trees with bootstrapping. It can be ftp-retrieved from the following IP address: ftp.ebi.ac.uk (193.62.196.6). A revised version (CLUSTAL W) is completed and can also be ftp-retrieved.

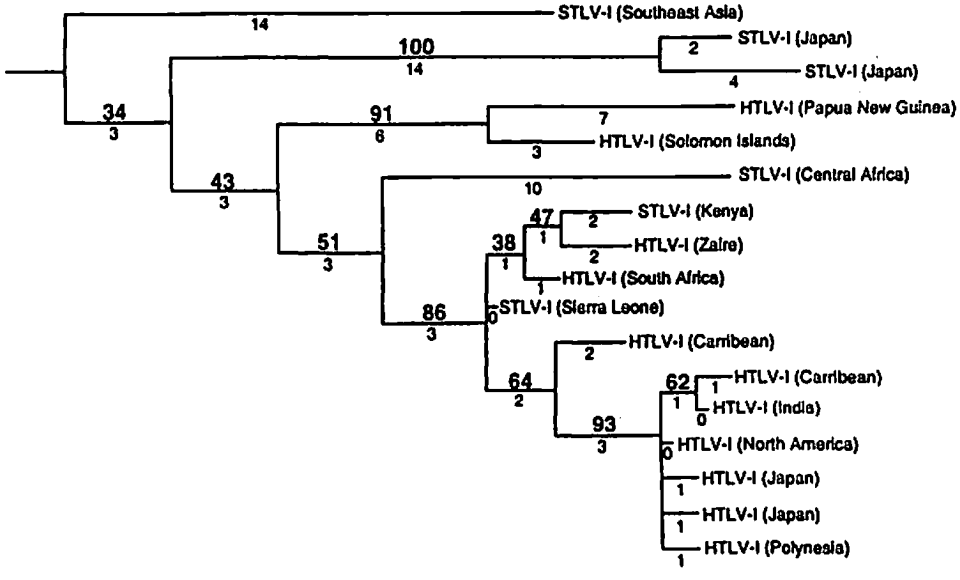

Figure 11. A neighbor-joining tree of HTLV-I (Human T lymphotropic virus type I) and STLV-I (Simian T lymphotropic virus type I) sequences [modified from (39)]. The tree is rooted by including an HTLV-II sequence. Numbers below branches are estimated numbers of nucleotide substitutions at corresponding branches, and those above internal branches are bootstrap probabilities (%).

PAUP (Phylogenetic Analysis Using Parsimony version 3.1) is run on the Macintosh, and does many kinds of maximum parsimony analysis under a user-friendly environment. It is a commercial product. Further information is available from the following Email address: paup@onyx.si.edu.

The author (email address: nsaitou@genes.nig.ac.jp) has developed two program packages, NJ and NJNUC. NJ requires distance matrices for input, while NJNUC requires nucleotide sequences. Adachi and Hasegawa (31) developed a computer package called MOLPHY. MOLPHY includes programs for maximum likelihood methods both for nucleotide and amino acid sequences. It can be ftp-retrieved from the following IP address: sunmh.ism.ac.jp (133.58.12.20).

Future Trends

Building phylogenetic trees from nucleotide and amino acid sequences starts with the collection of homologous sequences. Use of sequence databases is often involved in this process, and an "homology search" is essential for that. When one is dealing with closely related sequences, searching for homologous sequences is easy, while homology among remotely related sequences may be difficult to find. In this sense, the development of new algorithms as well as the improvement of the currently available homology searching algorithms consists one big study field.

After collecting homologous sequences, they need to be aligned. This "multiple alignment" problem is another vast field requiring improvement of algorithms (see [20] for review). Data analysis and experiments are also necessary to derive appropriate gap penalty parameters. Recently, Saitou and Ueda (44) estimated evolutionary rates of insertions and deletions for the non-coding region of DNA sequences from primates. More study on this problem will contribute to the production of better sequence alignments.

Elucidation of the pattern of nucleotide substitution is also essential for any phylogenetic tree-building method. There are at least two factors involved in the substitution pattern; nucleotide transition probability and variation of the substitution rate among sites. The former has been extensively studied, while the latter has gained researchers' interest only recently. Theoretical studies as well as data analysis are conducted, and this trend will continue further.

The above three aspects were not discussed in this chapter due to the limitations of space, but all are important for phylogenetic analysis. In fact, building phylogenetic trees using various algorithms is only the last aspect of the phylogenetic analysis, and this cannot be separated from the previous three aspects. Therefore, a comprehensive method that simultaneously estimates alignment, nucleotide substitution pattern, rate heterogeneity as well as tree topology and branch lengths will be required in the future.

Acknowledgment

This paper is partly supported by Grants-in-Aid for Scientific Research on Priority Areas (New Developments of Molecular Evolution) of Ministry of Education, Science and Culture, Japan.

References

1. Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.
2. Nei, M. 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.
3. Felsenstein, J. 1988. Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 39: 783-791.
4. Swofford, D. L. and Olsen, G. J. 1990. Phylogeny reconstruction. In: Molecular Systematics. D. M. Hillis and C. Moritz, eds. Sinauer, Sunderland, Mass. p. 411-502.
5. Saitou, N. 1991. Statistical methods for phylogenetic tree reconstruction. In: Handbook of Statistics, Volume 8: Statistical Methods for Biological and Medical Sciences. C. R. Rao and R. Chakraborty, eds. Elsevier Science Publishers B.V., Amsterdam. p. 317-346.
6. Kawamura, S., Saitou, N., and Ueda, S. 1992. Concerted evolution of the primate immunoglobulin alpha gene through gene conversion. J. Biol. Chem. 267: 7359-7367.
7. Sneath, P. H. P. and Sokal, R. 1973. Numerical Taxonomy. W. H. Freeman, San Francisco.
8. Fitch, W. M. and Margoliash, E. 1967. Construction of phylogenetic trees. Science, 155: 279-284.
9. Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. Amer. Natur. 106: 645-668.
10. Tateno, Y., Nei, M., and Tajima, F. 1982. Accuracy of estimated phylogenetic trees from molecular data. I. Distantly related species. J. Mol. Evol. 18: 354-361.
11. Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406-425.
12. Cavalli-Sforza, L.L. and Edwards, A.W.F. 1967. Phylogenetic analysis: Models and estimation procedures. Amer. J. Hum. Genet. 19: 233-257.
13. Saitou, N. and Imanishi, T. 1989. Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree reconstruction in obtaining the correct tree. Mol. Biol. Evol. 6: 514-525.
14. Rzhetsky, A. and Nei, M. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9: 945-967.
15. Bandelt, H.-J. and Dress, A. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. Mol. Phylogenet. Evol. 1: 242-252.
16. Fitch, W. M. 1977. On the problem of discovering the most parsimonious tree. Amer. Nat. 111: 223-257.
17. Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17: 368-376.
18. Felsenstein, J. 1993. PHYLIP: Phylogeny Inference Package, Ver. 3.5c. Univ. of Washington, Seattle.
19. Saitou, N. 1991. Reconstruction of molecular phylogeny of extant hominoids from DNA sequence data. Amer. J. Phys. Anthropol. 84: 75-85.
20. Doolittle R. ed. 1991. Methods in Enzymology, Volume 183: Computer Analysis of Protein and Nucleic Acid Sequences. Academic Press, Orlando.
21. Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. J. Mol. Evol. 39: 315-329.

22. Rzhetsky, A. and Nei, M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. 10: 1073-1095.
23. Neefs, J.-M., de Peer, Y. V., De Rijk, P., Chapelle, S. and De Wachter, R. 1993. Compilation of small ribosomal subunit RNA structures. Nuc. Acids Res. 21: 3025-3049.
24. 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.
25. 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.
26. Saitou, N. 1989. A theoretical study of the underestimation of branch lengths by the maximum parsimony principle. Syst. Zool. 38: 1-6.
27. Felsenstein, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. 27: 401-410.
28. Zharkikh, A. and Li, W.-H. 1993. Inconsistency of the maximum parsimony method: The case of five taxa with a molecular clock. Syst. Biol. 42: 113-125.
29. Takezaki, N. and Nei, M. 1994. Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. J. Mol. Evol. 39: 210-218.
30. Saitou, N. 1988. Property and efficiency of the maximum likelihood method for molecular phylogeny. J. Mol. Evol. 27: 261-273.
31. Adachi, J., and Hasegawa, M. 1994. MOLPHY: Programs for Molecular Phylogenetics, ver. 2.2. Institute of Statistical Mathematics, Tokyo.
32. Olsen, G.J., Matsuda, H., Hagstrom, R., and Overbeek, R. 1994. fastDNAml: A tool for construction of phylogenetic trees of DNA sequences using Maximum likelihood. Comp. Appl. Biosci. 10: 41-48.
33. Nei, M. 1991. Relative efficiencies of different tree-making methods for molecular data. In Miyamoto, M. M. and Cracraft, J. eds. Phylogenetic analysis of DNA sequence, Oxford University Press, New York, pp. 90-128.
34. DeBry, R. W. 1992. The consistency of several phylogeny-inference methods under varying evolutionary rates. Mol. Biol. Evol. 9: 537-551.
35. Tateno, Y., Takezaki, N., and Nei, M. 1994. Relative efficiencies of the maximum likelihood, neighbor-joining, and maximum parsimony methods when substitution rate varies with site. Mol. Biol. Evol. 11: 261-277.
36. Kuhner, M. and Felsenstein, J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11: 459-468.
37. Efron, B. 1979. Bootstrap methods: another look at the jackknife. Ann. Statist. 7: 1-26.
38. Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783-791.
39. Song K.-J., Nerurkar, V. R., Saitou, N., Lazo, A., Blakeslee ,J. R., Miyoshi, M., and Yanagihara, R. 1994. Genetic analysis and molecular phylogeny of simian T-cell lymphotropic virus type I: evidence for independent virus evolution in Asia and Africa. Virology 199: 56-66.
40. Zharkikh, A. and Li, W.-H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa with a molecular clock. Mol. Biol. Evol. 9: 1119-1147.

41. Zharkikh, A. and Li, W.-H. 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock. J. Mol. Evol. 35: 356-366.
42. Kumar, S, Tamura ,K, and Nei, M 1993. MEGA: Molecular Evolutionary Genetics Analysis, version 1.0. The Pennsylvania State University, University Park, PA 16802.
43. Higgins, D. G., Bleasby, A. J., and Fuchs, R. 1992. CLUSTAL V: improved software for multiple sequence alignment. Comput. Appl. Biosci. 8:189-191.
44. Saitou, N. and Ueda, S. 1994. Evolutionary rate of insertions and deletions in non-coding nucleotide sequences of primates. Mol. Biol. Evol. 11: 504-512.