Methods in Enzymology

Volume 266

Computer Methods for Macromolecular Sequence Analysis

EDITED BY

Russell F. Doolittle

CENTER FOR MOLECULAR GENETICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA

[25] Reconstruction of Gene Trees from Sequence Data

By NARUYA SAITOU

Properties of Gene Tree

Reconstruction of the phylogeny of genes is essential not only for the study of evolution but also for biology in general because replication of nucleotide sequences automatically produces a bifurcating tree of genes. It should be emphasized that the phylogenetic relationship of genes is different from the mutation process. The former always exists, whereas mutations 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 the occurrence of mutational events. In this respect, the extraction

of mutations from genes and their products is important for reconstructing phylogenetic trees of genes. The advancement of molecular biotechnology has made it possible to produce nucleotide sequences routinely. We therefore focus on the analysis of nucleotide sequences. However, a substantial part of this chapter also applies to other molecular data.

Formal Characteristics of Trees

A tree is a kind of graph. A graph is composed of node(s) and branch(es). There should be only one path between any two nodes on a tree. In evolutionary studies, a node represents a gene, species, or population, depending on the purpose, and a branch represents the topological relationship between nodes (often including information on lengths that represent mutational changes or evolutionary time). Nodes are divided into external and internal ones. The former are also called operational taxonomic units (OTUs). There are five OTUs (1–5) and four internal nodes (X, Y, Z, and R1) in the tree in Fig. 1a. Branches are also divided into external and internal ones. An external branch connects an external node and an internal node (e.g., branch 1-X of Fig. 1), whereas an internal branch connects two internal nodes (e.g., branch X-Z of Fig. 1).

A tree can be either rooted or unrooted. A rooted tree has a special node called the root which is defined as the position of the common ancestor

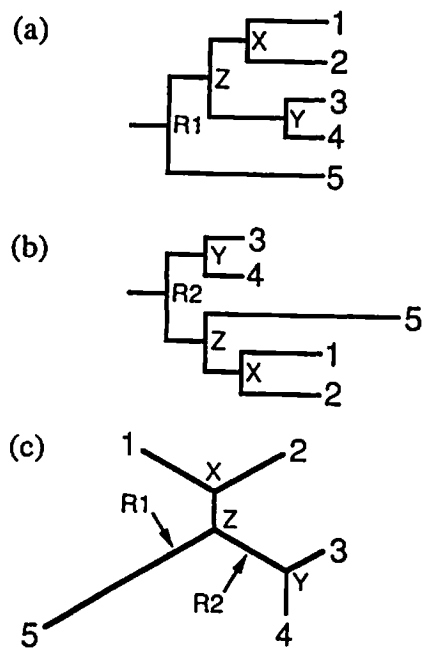

FIG. 1. Examples of rooted trees (a, b) and an unrooted tree (c) for five OTUs.

(see Fig. 1a,b). There will be a unique path from the root to any other node, and the direction of this is, of course, that of time. A phylogenetic tree in an ordinary sense is a rooted tree. Unfortunately, however, many methods for building phylogenetic trees produce unrooted trees, such as the tree of Fig. 1c. An unrooted tree can be converted to a rooted tree if the position of the root is specified. The trees in Fig. 1a,b were produced from the unrooted tree of Fig. 1c.

This relation 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 n sequences, we will add one (or more) sequence that is known to be an outgroup to the n sequences. The unrooted tree obtained for the $n + 1$ sequences can easily be converted to a rooted tree of n sequences. Sequence 5 corresponds to the outgroup in the tree in Fig. 1c when the root is R1, and the tree of Fig. 1a is then obtained. When the root is R2, sequences 3 and 4 are considered to be the outgroup to sequences 1, 2, and 5, and we obtain the tree of Fig. 1b.

The number of possible tree topologies rapidly increases with an increase in the number of OTUs. The general equation for the possible number of topologies for bifurcating unrooted trees (Tn) for n (≥ 3) OTUs is given by[1]

$$Tn = (2n - 5)!/[2^{n-3}(n - 3)!] \quad (1)$$

If we apply Eq. (1), there are 221,643,095,476,699,771,875 possible tree topologies for 20 OTUs. It is clear that the search for the true phylogenetic tree of many sequences is a very difficult problem. This is why so many methods have been proposed for building phylogenetic trees.

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 them. One such difference is illustrated in Fig. 2. Because a gene duplication occurred before the speciation of species A and B in Fig. 2a, both species have two homologous genes 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).[2] Thus, genes 1 and 3 (and 2 and 4) are orthologous, whereas genes 1 and 4 (and 2 and 3) are paralogous. If one is not aware of the gene duplication event, the gene tree for 1 and 4

[1] L. L. Cavalli-Sforza and A. W. F. Edwards, *Am. J. Hum. Genet.* **19**, 233 (1967).
[2] W. M. Fitch and E. Margoliash, *Evol. Biol.* **4**, 67 (1970).

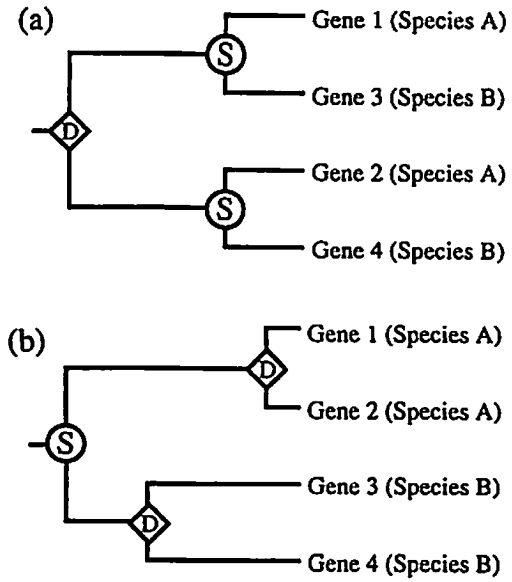

FIG. 2. Two possibilities of a gene tree for four genes sampled from two species. (a) Gene duplication (denoted by D) occurred before speciation (denoted by S). (b) Speciation occurred before two gene duplications.

may be misrepresented as the species tree of A and B, and thus a gross overestimation of the divergence time may occur. It should also be noted that the divergence time between genes 1 and 3 is identical with that between genes 2 and 4, since both times correspond to the same speciation event.

When two homologous gene copies are found in species A and B, another situation is possible, as shown in Fig. 2b. Now two gene duplications occurred after the speciation of species A and B, and two gene copies in the genome of each species are more closely related with each other than the corresponding homologous genes at different species. Because two duplication events occurred independently, the divergence time between genes 1 and 2 is different from that between genes 3 and 4.

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 Fig. 3a. A gene sampled from species A has its direct ancestor at the speciation time T_1 generations ago, and so does a gene sampled from species B. Thus, the divergence time between the two genes sampled from the different species always overestimates that of the species. The amount of overestimation corresponds to the coalescence time in the ancestral species, and its expectation is $2N$ for neutrally evolving nuclear genes of

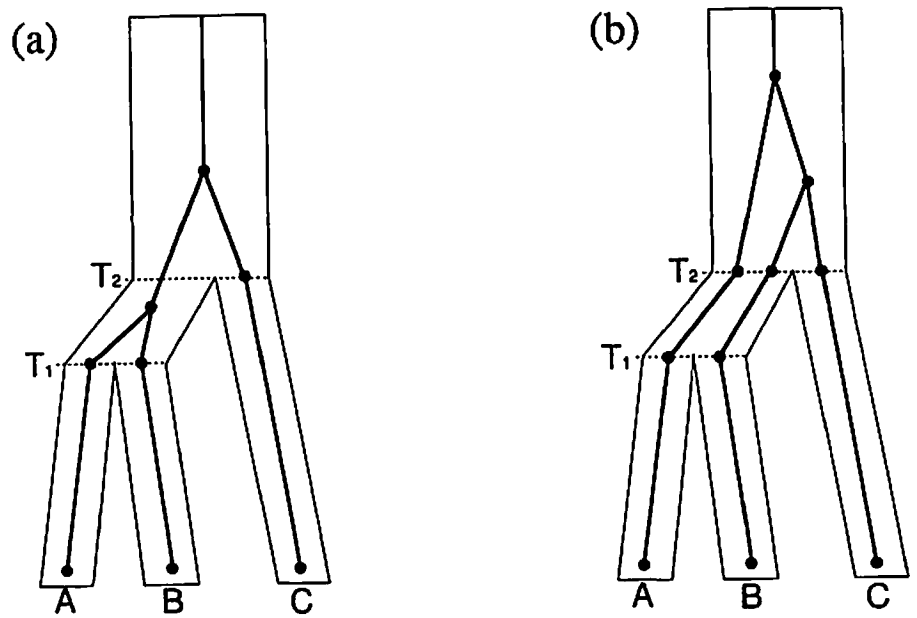

FIG. 3. Difference between a gene genealogy and species tree. (a) Topology of the gene genealogy is the same as that for the species tree. (b) Topology of the gene genealogy is different from that for the species tree. Full circles and thick lines denote the genealogical relationship, whereas thin lines (outlining the gene tree) denote the species tree. A, B, and C denote three genes each sampled from extant species, whereas X and Y denote ancestral genes. T_1 and T_2 denote the two speciation times.

diploid organism, where N is the population size of the ancestral species.[3] Therefore, if the two speciation events (T_1 and T_2) are close enough, the topological relationship of the gene tree may become different from that of the species tree, as shown in Fig. 3b. Although species A and B are more closely related than to C, genes sampled from species B and C happen to be more closely related with each other than to that sampled from species A. The probability (P_{error}) of obtaining an erroneous tree topology is given by[4]

$$P_{error} = (2/3)\, e^{-T/2N} \qquad (2)$$

where $T = T_2 - T_1$ generations. For example, P_{error} is 0.404 when $T = 50{,}000$ and $N = 50{,}000$. Therefore, a species tree estimated from a single gene may not be correct even if the gene tree was correctly estimated. In this case, we should use more than one gene.

[3] F. Tajima, *Genetics* **105**, 437 (1983).
[4] M. Nei, "Molecular Evolutionary Genetics." Columbia Univ. Press, New York, 1987.

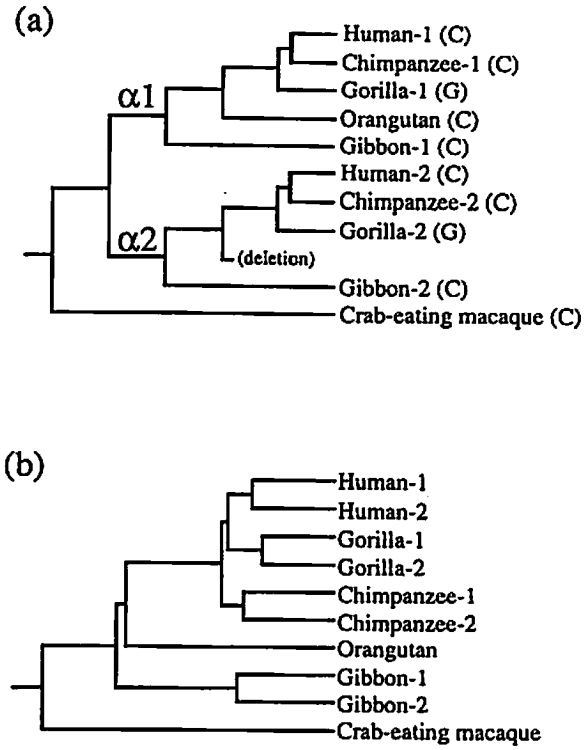

FIG. 4. Alteration of an estimated gene tree caused by gene conversion. (a) The most presumable gene tree for the primate immunoglobulin $\alpha 1$ and $\alpha 2$ genes. C or G in parentheses after species names indicate one nucleotide configuration possibly caused by gene conversion in the gorilla genome. (b) A spurious gene tree (modified from Kawamura *et al.*[5]).

When gene conversion and/or recombination has occurred within the gene region under consideration, a gene tree may be different from the species tree. Kawamura *et al.*[5] examined primate immunoglobulin α genes 1 and 2. Figure 4a shows the plausible gene tree; the gene duplication clearly preceded speciation of hominoids, followed by deletion of the $\alpha 2$ gene from the orangutan genome. However, there are many nucleotide sites that possibly experienced gene conversion. One such example is shown in Fig. 4a; 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. If this kind of nucleotide configuration is contiguous, gene conversion is suspected. The resulting spurious gene tree (Fig. 4b) is distorted from the tree of Fig. 4a because of the strong effect of gene conversion.

Ideally, branch lengths of a phylogenetic tree are proportional to the

[5] S. Kawamura, N. Saitou, and S. Ueda, *J. Biol. Chem.* **267**, 7359 (1992).

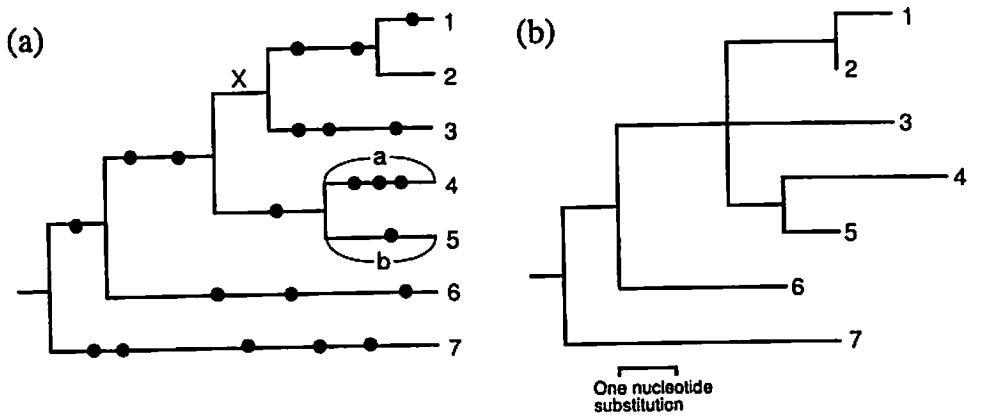

FIG. 5. Examples of the expected gene tree (a) and the corresponding realized gene trees (b). Filled circles on the expected gene tree denote nucleotide substitutions. Because no substitution occurred at branch X of the expected gene tree (a), the corresponding branch does not exist in the realized gene tree (b).

physical time since divergence. Thus the branch a and b of Fig. 5a should be the same length. We call this type of rooted tree the expected tree.[4] 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, whereas an expected species tree is a gross simplification of the course of differentiation of populations. Therefore, the speciation time is always unclear.

As mentioned earlier, the genealogical relationship of genes, or expected gene tree, is independent from 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. We call this ideal reconstruction of the gene tree the realized gene tree (Fig. 5b), whereas the reconstructed one from observed data is called the estimated gene tree.[6] Branch lengths of realized and estimated gene trees are proportional to mutational events. These mutational events are not necessarily proportional to physical time. 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 Fig. 5a.

A species tree reconstructed from observed data is called an estimated species tree, but there is no realized species tree. It should also be noted that both expected and realized trees are rooted, while estimated trees are often unrooted due to the limitations of available information.

[6] N. Saitou, *in* "Molecular Biology: Current Innovations and Future Trends Part 2" (H. G. Griffin and A. M. Griffin, eds.), p. 115. Horizon Scientific Press, Norfolk, England, 1995.

TABLE I
CLASSIFICATION OF TREE-MAKING METHODS

Method	Stepwise clustering	Exhaustive search
Distance matrix	UPGMA	KITCH
	Distance Wagner	Fitch–Margoliash
	Neighbor joining	Minimum evolution
Character state		Maximum parsimony
		Compatibility
		Maximum likelihood

Methods for Building Phylogenetic Trees of Genes

Classification of Tree-Building Methods

Many methods have been proposed for building a 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 all the possible pairwise distances, whereas an array of character states is used for the character-state methods. UPGMA (unweighted pairgroup method using arithmatic mean),[7] the Fitch and Margoliash method,[8] the distance Wagner method,[9] the neighbor-joining method,[10] and the minimum evolution methods[1,11,12] are distance matrix methods, whereas the maximum parsimony method,[13] the compatibility method,[14] and the maximum likelihood method[15] are character-state methods (Table I).

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 other strategy is to examine a local topological relationship of OTUs and find the best tree. These types of methods are called stepwise clustering methods. Both strategies are used for the distance matrix methods, while the exhaustive search strategy is usually used for character-state methods (Table I).

[7] P. H. P. Sneath and R. Sokal, "Numerical Taxonomy." Freeman, San Francisco, 1977.
[8] W. M. Fitch and E. Margoliash, *Science* **155**, 279 (1967).
[9] J. S. Farris, *Am. Nat.* **106**, 645 (1972).
[10] N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
[11] N. Saitou and T. Imanishi, *Mol. Biol. Evol.* **6**, 514 (1989).
[12] A. Rzhetsky and M. Nei, *Mol. Biol. Evol.* **9**, 945 (1992).
[13] W. M. Fitch, *Am. Nat.* **111**, 223 (1977).
[14] W. J. Le Quesne, *Syst. Zool.* **18**, 201 (1969).
[15] J. Felsenstein, *J. Mol. Evol.* **17**, 368 (1981).

TABLE II
ESTIMATED NUMBER OF NUCLEOTIDE SUBSTITUTIONS PER SITE BETWEEN EVERY PAIR OF 10 SEQUENCES[a]

	1	2	3	4	5	6	7	8	9
2	0.0516								
3	0.0550	0.0031							
4	0.0483	0.0221	0.0253						
5	0.0582	0.0651	0.0685	0.0549					
6	0.0094	0.0416	0.0450	0.0384	0.0549				
7	0.0125	0.0584	0.0619	0.0551	0.0651	0.0157			
8	0.0284	0.0687	0.0722	0.0654	0.0754	0.0317	0.0285		
9	0.0925	0.1221	0.1259	0.1185	0.1370	0.0820	0.0786	0.0927	
10	0.1921	0.2183	0.2228	0.2054	0.2309	0.1798	0.1795	0.1833	0.1860

[a] Gaps were eliminated from the comparison, and a total of 323 nucleotide sites were compared. Kimura's two-parameter method was used [M. Kimura, *J. Mol. Evol.* **16**, 111 (1980)]. Sequence identifications: 1, *Mus mus domesticus* functional gene; 2, *M. mus domesticus* pseudogene; 3, *M. mus castaneus* pseudogene; 4, *M. spicilegus* pseudogene; 5, *M. leggada* pseudogene; 6, *M. mus domesticus* cDNA; 7, *M. leggada* functional gene; 8, *M. platythrix* functional gene; 9, *Rattus norvegicus* cDNA; 10, *Homo sapiens* cDNA.

In distance matrix methods, a phylogenetic tree is constructed by considering the relationship among the distance values D_{ij} (distance between OTUs i and j). An example of a distance matrix is presented in Table II. The distances were computed from the nucleotide sequences for ten p53 functional genes and pseudogenes.[16] These sequence data will be used consistently in this chapter for worked-out examples. There are many methods for estimating evolutionary distances from nucleotide sequences.

Because there are already many reviews on tree-building methods,[4,6,17,18] we describe only the following six methods; UPGMA, the neighbor-joining method, the minimum evolution method, the maximum parsimony method, the maximum likelihood method, and network methods.

Methods Assuming Molecular Clock

When constancy of the evolutionary rate, or a molecular clock, is assumed, we can reconstruct rooted trees. This is because sequences should

[16] H. Ohtsuka, M. Oyanagi, Y. Mafune, N. Miyashita, T. Shiroishi, K. Moriwaki, R. Kominami, and N. Saitou, *Mol. Phylogenet. Evol.* in press.

[17] N. Saitou, *in* "Handbook of Statistics, Volume 8: Statistical Methods for Biological and Medical Sciences" (C. R. Rao and R. Chakraburty, eds.), p. 317. Elsevier, Amsterdam, 1990.

[18] D. L. Swofford and G. J. Olsen, *in* "Molecular Systematics" (D. M. Hillis and C. Moritz, eds.), p. 411. Sinauer Associates, Sunderland, Massachusetts, 1990.

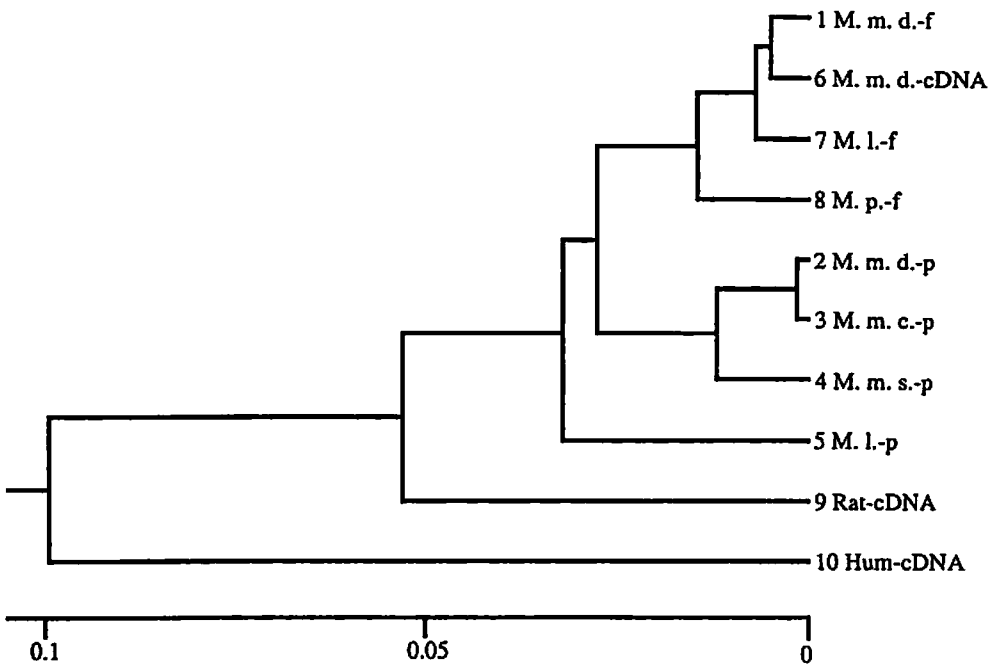

FIG. 6. UPGMA tree for the distance matrix in Table II. Sequence identifiers correspond to those of Table II.

be clustered in the order of their mutational difference if the amount of mutational changes is strictly proportional to evolutionary time. There are many ways to obtain such rooted trees from a distance matrix.[7] In this section, UPGMA and KITCH are discussed.

Let us briefly explain the UPGMA algorithm using the distance matrix shown in Table II. We first choose the smallest distance, D_{23} (=0.0031). Then OTUs 2 and 3 are combined and the distances between the combined OTU [2–3] and the remaining eight OTUs are computed by taking arithmetic means. At the next step, again the smallest distance (D_{16} = 0.0094) is chosen from the distance matrix. Then the OTUs 1 and 6 are combined into OTU [1–6]. This process is continued until all the OTUs are finally clustered into a single one. The resultant rooted tree is shown in Fig. 6. Although *Mus* functional genes (sequences 1, 6, 7, and 8) formed a monophyletic cluster, the corresponding *Mus* pseudogenes (sequences 2–5) did not form a monophyletic one.

Because of the long history of UPGMA (originally proposed by Sokal and Michener[19]), there are many computer programs available for UPGMA, and these are not specified. It should be noted that there are several

[19] R. Sokal and C. D. Michener, *Univ. Kansas Sci. Bull.* **28**, 1409 (1958).

synonyms for UPGMA, such as the simple linkage method, the clustering method, and the nearest neighbor method.

KITCH is a computer program in the PHYLIP package[20] and is related to the Fitch and Margoliash method,[8] but constancy of the evolutionary rate is assumed. Because of this restriction, the result of KITCH is usually quite close to that of UPGMA. In fact, when KITCH was applied to the distance matrix of Table II, a result (not shown) identical to that of UPGMA was obtained. It seems that there is no use for this exhaustive search program if one already has the result using a UPGMA program.

A simulation study[10] has shown that UPGMA is not efficient in reconstructing the true topological relationship when the constancy of evolutionary rate is not assumed. Therefore, it is not advisable to use methods assuming a molecular clock for estimating realized trees. However, those are still useful for estimating expected trees, where all the branch lengths are proportional to physical time.

When we have only an unrooted tree with no outgroup, there is a way of rooting it if we assume a rough constancy of the evolutionary rate. Given the unrooted tree topology, we successively cluster OTU pairs starting from the smallest distance similar to UPGMA.[21] If we apply this algorithm to the unrooted tree of Fig. 1c, we will obtain the tree of Fig. 1a.

Neighbor-Joining Method

A pair of OTUs are called neighbors when these are connected through a single internal node in an unrooted bifurcating tree. For example, OTUs 1 and 2 of Fig. 1c are a pair of neighbors. If we combine these OTUs, this combined OTU [1–2] and OTU 5 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[10] 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 it has been proved that the expected value of the sum of branch lengths is smallest for the tree with the true branching pattern.[22] Because of the simple algorithm, more than 100 OTUs

[20] J. Felsenstein, "PHYLIP: Phylogeny Inference Package, Version 3.5c." Univ. of Washington, Seattle, 1993.

[21] N. Ishida, T. Oyunsuren, S. Mashima, H. Mukoyama, and N. Saitou, *J. Mol. Evol.* **41,** 180 (1995).

[22] A. Rzhetsky and M. Nei, *Mol. Biol. Evol.* **10,** 1073 (1993).

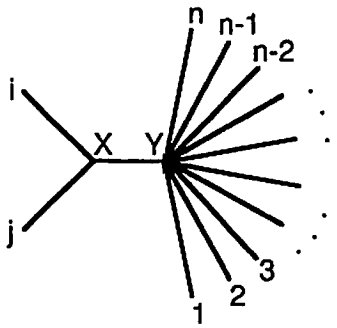

FIG. 7. Tree of N OTUs in which OTUs i and j are neighbors.

can be handled within a relatively short computer time by using the neighbor-joining method. For example, Horai *et al.*[23] produced a neighbor-joining tree for 193 human mitochondrial DNA sequences.

The following explanation of the neighbor-joining algorithm is based on Saitou.[6] 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 (S_0) of n branch lengths can be shown to be

$$S_0 = Q/(n-1) \tag{3}$$

where

$$Q = \Sigma_{i<j} D_{ij} \tag{4}$$

In practice, some pairs of OTUs are more closely related to one another 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 Fig. 7, where OTUs i and j are assumed to be neighbors. The sum of branch lengths is defined by

$$S_{ij} = (B_{iX} + B_{jX}) + B_{XY} + \Sigma_{k\neq i,j} B_{kY} \tag{5}$$

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

$$D_{ij} = B_{iX} + B_{jX} \tag{6a}$$
$$D_{ik} = B_{iX} + B_{XY} + B_{kY} \qquad (k \neq i,j) \tag{6b}$$
$$D_{jk} = B_{jX} + B_{XY} + B_{kY} \qquad (k \neq i,j) \tag{6c}$$
$$D_{kl} = B_{iY} + B_{jY} \qquad (k,l \neq i,j) \tag{6d}$$

[23] S. Horai, R. Kondo, Y. Nakagawa-Hattori, S. Hayashi, S. Sonoda, and K. Tajima, *Mol. Biol. Evol.* **10**, 23 (1993).

With the tree shown in Fig. 7, it can be shown by applying the above relationship that

$$B_{XY} = [Q - (n-1)D_{ij} - (n-1)\,\Sigma_{k,l\neq i,j}\,D_{kl}/(n-3)]/2(n-2) \quad (7)$$

If we neglect OTUs i and j in Fig. 7, the remaining $n-2$ OTUs form a starlike tree, as is clear from Eq. (6d). Thus we apply Eq. (3) and obtain

$$\Sigma_{k\neq i,j}\,B_{kY} = \Sigma_{k,l\neq i,j}\,D_{kl}/(n-3) \quad (8)$$

We also note that

$$\Sigma_{k,l\neq i,j}\,D_{kl} = Q - (R_i + R_j - D_{ij}) \quad (9)$$

where $R_i = \Sigma_j\,D_{ij}$ and $R_j = \Sigma_i\,D_{ij}$. Putting Eqs. (6a), (7), and (8) into Eq. (5) with consideration of Eq. (9), we obtain

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

Equation (10) was first proposed 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 Fig. 7, the branch lengths are estimated using the Fitch and Margoliash procedure[8] as

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

and

$$B_{jX} = D_{ij} - B_{iX} \quad (11b)$$

Therefore, all the branch lengths as well as the tree topology will be determined after $n-2$ steps for n OTUs.

Table III shows the output of the computer program NJNUC, and Fig. 8 shows the neighbor-joining tree. Human p53 cDNA sequence (OTU 10) was assumed to be the outgroup. Branch lengths are estimated numbers of nucleotide substitutions that occurred in this p53 sequence, and all of them are integer values. To obtain those numbers, estimated numbers of nucleotide substitutions per site (numbers in parentheses in Table III) were multiplied with the number of compared nucleotide sites, then the resulting values were rounded. If a branch length turned out to be zero,

[24] J. A. Studier and K. J. Keppler, *Mol. Biol. Evol.* 5, 729 (1988).

TABLE III
OUTPUT OF PROGRAM NJNUC FOR p53 SEQUENCE DATA[a]

Node 11	OTU 9 = 14.632 (4.530E-02)	OTU 10 = 45.440 (1.407E-01)
Node 12	OTU 2 = −0.065 (−2.011E-04)	OTU 3 = 1.069 (3.311E-03)
Node 13	Node 12 = 4.463 (1.382E-02)	OTU 4 = 2.693 (8.339E-03)
Node 14	Node 13 = 4.281 (1.325E-02)	OTU 5 = 11.547 (3.575E-02)
Node 15	OTU 8 = 5.427 (1.680E-02)	Node 11 = 9.108 (2.820E-02)
Node 16	OTU 7 = 1.913 (5.923E-03)	Node 15 = 1.235 (3.822E-03)
Node 17	Node 14 = 5.102 (1.580E-02)	OTU 6 = 0.532 (1.646E-03)
Node 18	(Last node)	
OTU 1	1.675 (5.186E-03)	
Node 17	0.907 (2.808E-03)	
Node 16	1.410 (4.364E-03)	

[a] The distance matrix of Table II was used. Numbers after the OTU or node designation are branch lengths in terms of nucleotide substitutions that occurred at the compared sequence region between that node/OTU and the node written at the top of each row. Numbers in parentheses are branch lengths in terms of nucleotide substitutions per site.

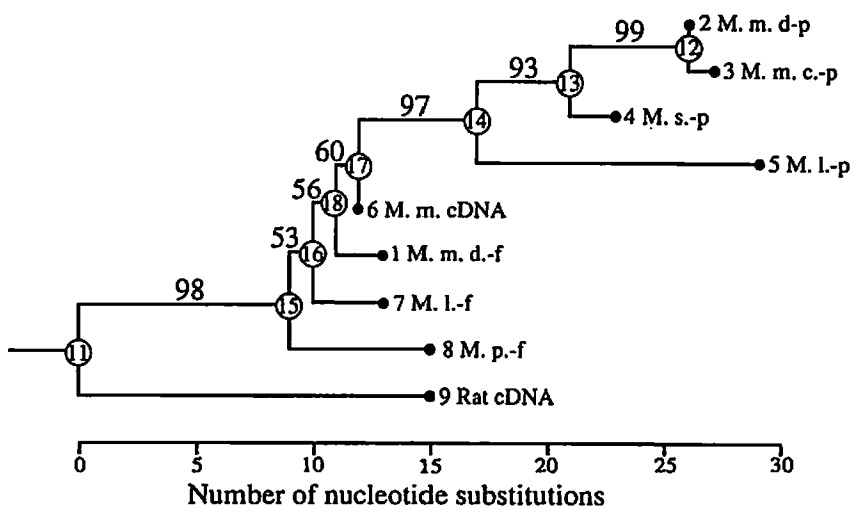

FIG. 8. Neighbor-joining tree constructed from the distance matrix of Table III. This tree was drawn on the basis of the output shown in Table IV. Branch lengths are proportional to the number of nucleotide substitutions per branch. Sequence identifiers correspond to those of Table II, and numbers in circles are internal node identifications. Numbers above internal branches are bootstrap probabilities (%). Human cDNA sequence was assumed to be the outgroup.

such as the branch 2–12, that branch was truncated. A tree obtained by applying this procedure, first proposed by Nerurker *et al.*,[25] is an estimation of the realized tree. The topology of the neighbor-joining tree (Fig. 8) is somewhat different from that of the UPGMA tree (Fig. 6). Now the *Mus* pseudogenes (sequences 2–5) form a monophyletic cluster, while *Mus* functional counterparts (sequences 1, 6, 7, and 8) do not form a monophyletic cluster.

Numbers above internal nodes in the tree of Fig. 8 are bootstrap probabilities (percentages) based on 1000 replications (program NJBOOT2 was used for obtaining the bootstrap probabilities). For example, all the four pseudogene sequences are clustered with a high bootstrap probability (97%) at the internal node C. The bootstrap method was proposed for estimating variances from unknown probability distributions[26] and was introduced into phylogenetic study.[27] Character-state data are necessary to use the bootstrap method, but trees built using any distance matrix method can be tested using this technique. We first randomly resample n nucleotide sites from the given sequence data of n nucleotides with replacement. This resampling is replicated at least 1000 times. For example, one replication may have n nucleotide sites with the positions $1, 2, 2, 4, 5, \ldots, n-2, n$, and n. Resampling is usually done by generating pseudorandom numbers. Each replicated sequence data set is 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 realize 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. The bootstrap method is currently widely used, but its influence on phylogenetic inference is not thoroughly known; theoretical studies are still going on.

A series of programs (NJ, NJNUC, etc.) are included in the TreeTree package developed by the author. This method is also available in packages PHYLIP,[20] MEGA,[28] CLUSTAL W,[29] and MOLPHY.[30] Programs NJBOOT2 and TREEVIEW run on MS-DOS developed by K. Tamura (E-mail: Koichiro-Tamura@c.metro-u.ac.jp) are also available.

[25] V. R. Nerurkar, K.-J. Song, N. Saitou, R. R. Mallan, and R. Yanagihara, *Virology* **196**, 506 (1993).
[26] B. Efron, *Ann. Stat.* **7**, 1 (1979).
[27] J. Felsenstein, *Evolution* **39**, 783 (1985).
[28] S. Kumar, K. Tamura, and M. Nei, "MEGA: Molecular Evolutionary Genetics Analysis, Version 1.0." The Pennsylvania State Univ., University Park, 1993.
[29] J. D. Thompson, D. G. Higgins, and T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
[30] J. Adachi and M. Hasegawa, "MOLPHY: Programs for Molecular Phylogenetics, Version 2.2." Institute of Statistical Mathematics, Tokyo, 1994.

Minimum Evolution Methods

The concept of minimum evolution was used in the neighbor-joining method, and this concept was first used by Cavalli-Sforza and Edwards.[1] Saitou and Imanishi[11] proposed a simple method 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[12] proposed a minimum evolution method in which branch lengths with their standard errors are computed by applying the least squares method. A neighbor-joining tree is first constructed as the candidate tree, and the related trees with only small topological differences are then searched. A simplified algorithm for computing least squares estimates of branch lengths has been proposed to reduce the computation time.[22]

Table IV shows an output of a minimum evolution program ME_TREE. There is one topological difference between the neighbor-joining tree (see Fig. 8) and this minimum evolution tree, regarding the clustering of sequences 1 and 7. The sum of branch lengths are 0.346188 and 0.346526 for the minimum evolution and neighbor-joining trees, respectively. Standard errors of branch lengths are small when the bootstrap values of the corresponding branches (see Fig. 8) are high. For example, the internal branch

TABLE IV
OUTPUT OF PROGRAM ME_TREE FOR p53 SEQUENCE DATA[a]

Branch	Branch length ± SE	Significance level (%)
1 and 16	0.005795 ± 0.003568	89.48
2 and 12	−0.000243 ± 0.000247	67.30
3 and 12	0.003349 ± 0.003354	67.78
4 and 13	0.008107 ± 0.005287	87.14
5 and 14	0.036693 ± 0.011213	99.90
6 and 17	0.001429 ± 0.002244	47.14
7 and 16	0.006695 ± 0.003946	90.90
8 and 15	0.017395 ± 0.006561	99.20
9 and 11	0.045297 ± 0.012698	99.96
10 and 11	0.140679 ± 0.023459	99.96
11 and 15	0.027608 ± 0.011227	98.58
12 and 13	0.014052 ± 0.006792	96.06
13 and 14	0.013044 ± 0.007296	92.50
14 and 17	0.014755 ± 0.006999	96.42
15 and 18	0.005024 ± 0.004819	70.16
16 and 18	0.002454 ± 0.004777	39.00
17 and 18	0.004056 ± 0.002930	83.24

[a] The distance matrix of Table II was used.

TABLE V
CLASSIFICATION OF NUCLEOTIDE CONFIGURATIONS OF p53 SEQUENCE DATA OF 323 NUCLEOTIDES FOR MAXIMUM PARSIMONY METHOD

Category	Observed number	Minimum number of changes
Noninformative configuration		
Invariant	236	0
Variant with 2 nucleotides	50	50
Variant with 3 nucleotides	9	18
Variant with 4 nucleotides	0	0
Informative configuration		
Variant with 2 nucleotides	26	26
Variant with 3 nucleotides	2	4
Variant with 4 nucleotides	0	0
Total	323	98

(0.014755 ± 0.006999) connecting nodes 14 and 18 is significantly larger than zero (significance level of 96.42%), and the corresponding bootstrap probability for the neighbor-joining tree is 97%.

There is a computer program (ME_TREE) run on MS-DOS.[31] Another program run on SUN workstations has been developed by Igor Belyi [WWW (World Wide Web) home page is http://www.cse.psu.edu/~belyi].

Maximum Parsimony Methods

The principle of maximum parsimony was first used for morphological data,[32] but it was independently proposed also for molecular data.[33] There are several kinds of maximum parsimony methods based on various assumptions, but the one that produces unrooted trees as in the case of the neighbor-joining method is mainly used for nucleotide sequence data.[13] The maximum parsimony principle is the minimization of the character-state changes (tree length) on the given tree topology, and is related to the principle used in minimum evolution methods. However, the performance of the two methods in choosing the best topology can be quite different.

Let us consider the example sequence data. We first classify the 323 nucleotide sites into different configurations (Table V). A nucleotide configuration is a distribution pattern of nucleotides for a given number of

[31] A. Rzhetsky and M. Nei, *Comput. Appl. Biosci.* **10**, 409 (1994).
[32] J. H. Camin and R. Sokal, *Evolution* **19**, 311 (1965).
[33] R. V. Eck and M. O. Dayhoff, *in* "Atlas of Protein Sequence and Structure" (M. O. Dayhoff ed.). National Biomedical Research Foundation, Silver Spring, Maryland, 1966.

sequences. The possible number (C_n) of configurations for n sequences is given by[34]

$$C_n = (4^{n-1} + 3 \times 2^{n-1} + 2)/6 \quad (12)$$

For example, there are 51 possible nucleotide configurations for five sequences. It should be noted that the number of possible configuration increases if we distinguish transitional differences from transversional ones.

Those configurations are first divided into noninformative and informative ones (Table V). Configurations that do not contribute to the selection of tree topology are called noninformative for the maximum parsimony method. All the sequences have the same nucleotide at the invariant configuration. There were 236 sites that fell into this category. We do not need any nucleotide substitution for this configuration under the maximum parsimony principle. One and two substitutions are necessary for any topology for variant with two and three nucleotides of the noninformative configuration, respectively. 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.[13] Only 28 of 323 sites had informative configurations. In total, we need at least 98 nucleotide substitutions for this data set.

Because there already exist several descriptions of the maximum parsimony method,[4,6,17,18] we will skip the explanation of the method in this chapter and show only the worked-out example. The result of the maximum parsimony analysis using PAUP is presented in Table VI. Nine equally parsimonious trees that require 112 substitutions were found by using branch-and-bound as well as heuristic options. However, two of them turned out to be identical with each other if we truncate an internal branch with zero length. Thus, the real number of equally parsimonious trees was eight (trees 1–8 of Table VI). Tree 6 had the same topology with the neighbor-joining tree (Fig. 8), and tree 4 was the maximum likelihood tree. Trees 9–12 required 113 substitutions and thus are subparsimonious. Biologically, however, tree 10 or 11 seems to be more reasonable.[16] It is also interesting to note that the minimum evolution tree (tree 12 of Table VI) was not a maximum parsimonious tree.

The principle of maximum parsimony attracted many because of its simplicity and logical clarity. However, there are some problems with this method when molecular data are used. Felsenstein[35] showed analytically that the maximum parsimony method may be positively misleading when the rate of evolution is grossly different among lineages of four sequences.

[34] N. Saitou and M. Nei, *J. Mol. Evol.* **24**, 189 (1986).
[35] J. Felsenstein, *Syst. Zool.* **27**, 401 (1978).

TABLE VI
MAXIMUM PARSIMONY AND MAXIMUM LIKELIHOOD ANALYSES

Tree ID[a]	Tree topology[b]	RNM[c]	Differences of log L[d]
1	(((((((2,3),4),5),8),6),1),7,(9,10))	112	−4.07
2	((((7,8),1),6),(((2,3),4),5),(9,10))	112	−0.29
3	(1,6,(((((2,3),4),5),8),((9,10),7)))	112	−4.09
4	(((((((2,3),4),5),6),1),8),7,(9,10))	112	Best
5	((((((2,3),4),5),(6,1)),8),7,(9,10))	112	−5.39
6	(((((((2,3),4),5),6),1),7),8,(9,10))	112	−2.55
7	(((((2,3),4),5),(6,(1,7))),8,(9,10))	112	−4.31
8	(((((2,3),4),5),6),((1,7),8),(9,10))	112	−6.10
9	(((1,6),7),((((2,3),4),5),8),(9,10))	113	−9.02
10	((((1,6),7),(((2,3),4),5)),8,(9,10))	113	−8.80
11	((((1,6),7),8),(((2,3),4),5),(9,10))	113	−8.99
12	((((((2,3),4),5),6),(1,7)),8,(9,10))	113	−7.37

[a] Trees 6 and 12 are the neighbor-joining tree (Fig. 8) and the minimum evolution tree (Table IV), respectively.
[b] Sequence identifications are the same as those of Table II.
[c] Required number of mutations when the maximum parsimony method was applied.
[d] Differences of log likelihood values from that of the best tree (tree 4; its log likelihood was −1016.75).

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). The same problem was found even when constancy of the evolutionary rate is assumed.[36,37] Saitou[38] showed that the gross underestimation of the branch lengths occurred when the divergence (number of nucleotide substitutions per site) among sequences was larger than 0.2. Therefore, we should be careful when using the maximum parsimony method.

After the tree topology is determined, however, the principle of maximum parsimony can be useful for estimating the location of mutational events. For example, Gojobori *et al.*[39] estimated the direction of nucleotide substitutions, and Saitou and Ueda[40] mapped the insertions and deletions on the assumed phylogenetic tree of primates. Jermann *et al.*[41] have recon-

[36] A. Zharkikh and W.-H. Li, *Syst. Biol.* **42,** 113 (1993).
[37] N. Takezaki and M. Nei, *J. Mol. Evol.* **39,** 210 (1994).
[38] N. Saitou, *Syst. Zool.* **38,** 1 (1989).
[39] T. Gojobori, W.-H. Li, and D. Graur, *J. Mol. Evol.* **18,** 360 (1982).
[40] N. Saitou and S. Ueda, *Mol. Biol. Evol.* **11,** 504 (1994).
[41] T. M. Jermann, J. G. Opitz, J. Stachkouse, and S. A. Benner, *Nature (London)* **374,** 57 (1995).

structed ancestral ribonuclease proteins from the estimated tree for the artiodactyls. It should be noted that the maximum parsimony principle can be applied to any tree irrespective of the methods used for constructing it.

PAUP 3.1.1 (a commercial product distributed from Illinois Natural History Servey) is run on a Macintosh with many user-friendly options. Maximum parsimony analysis is possible also for PHYLIP[20] and MEGA.[28] MacClade[42] has various useful features for molecular data, although it does not search the topology space.

Maximum Likelihood Methods

The maximum likelihood method is often used for parameter estimation in statistics, and it was first applied to building phylogenetic trees for allele frequency data.[1] Later, various maximum likelihood methods and computer programs were developed for sequence data.

The core algorithm of the maximum likelihood method is as follows. We first define the probability $P_{\alpha\beta} \equiv Pr(N_\alpha, N_\beta, B_{\alpha\beta})$ for observing nucleotides 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 $P_{\alpha\beta}$. 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 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. Interested readers may refer to more detailed descriptions of this method.[4,6,15,43]

Table VI shows the result of the DNAML computation (user tree option was used). Tree 4, one of 8 equally parsimonious trees, was found to have the highest likelihood among the 12 trees compared. Likelihood values for subparsimonious trees (with 113 required substitutions) are somewhat lower than those for equally parsimonious trees.

Because the maximum likelihood method requires massive computer time, there are several searching methods other than the exhaustive search. The default method of DNAML[15,20] is the sequential addition of sequences. Saitou[44] proposed a stepwise clustering of sequences for the maximum likelihood method, and this searching method is the same as that of the

[42] W. P. Maddison and D. R. Maddison, "MacClade Version 3." Sinauer Associates, Sunderland, Massachusetts, 1992.

[43] N. Saitou, this series, Vol. 183, p. 584.

[44] N. Saitou, *J. Mol. Evol.* **27**, 261 (1988).

neighbor-joining method. NucML of MOLPHY[30] has several options for topology searches, and one of them (star decomposition) is similar to that of the Saitou[44] method.

DNAML and DNAMLK (molecular clock is assumed) are included in PHYLIP.[20] There is also a modified version of DNAML called fastDNAML.[45] NucML for nucleotide sequences and ProtML for amino acid sequences are included in MOLPHY.[30]

Methods Producing Networks, Not Trees

The evolutionary history of a gene should be presented as a tree. When we analyze real sequence data, however, this tree structure may not be clearly observed. Bandelt and Dress[46] proposed the split decomposition method for distance matrix data. Unlike most tree-building methods, it usually produces a network, not a tree. A relaxed condition is used for estimating splitting patterns among OTUs, and both the signal (suggesting the tree structure) and the noise (suggesting patterns inconsistent with the tree structure) can be presented simultaneously. The resultant network is shown in Fig. 9. Four parallelograms suggest the existence of some parallel nucleotide changes, though the overall structure is quite close to an unrooted tree.

This network construction can also be applied to sequence data directly. When two nucleotide positions show an incongruent partition pattern, a discordancy diagram[13] appears. Bandelt[47] has extended this idea and proposed the phylogenetic network method. A network structure is useful for delineating anomaly in the history of gene trees. For example, when two regions of a gene experienced recombination(s), we may obtain a network, not a tree, if we analyze the sequence data by combining the two regions.

Regarding program availability, 'Daniel Huson (huson@mathematik.uni-bielefeld.de) and Rainer Wetzel have developed a shareware called SplitsTree run on Macintosh. A program for the phylogenetic network method is under development by H.-J. Bandelt.

Freely Distributed Computer Packages

PHYLIP[20] contains many programs in the form of both source code and executable files. Various kinds of maximum likelihood methods, maximum parsimony methods, and distance matrix methods can be used. It can be retrieved from evolution.genetics.washington.edu (128.95.12.41) or from

[45] G. J. Olsen, H. Matsuda, R. Hagstrom, and R. Overbeek, *Comput. Appl. Biosci.* **10**, 41 (1994).
[46] H.-J. Bandelt and A. Dress, *Adv. Math.* **92**, 47 (1992).
[47] H.-J. Bandelt, *Verhandlungen des Naturwissenschaftlichen Vereins in Hambrug* **34**, 51 (1994).

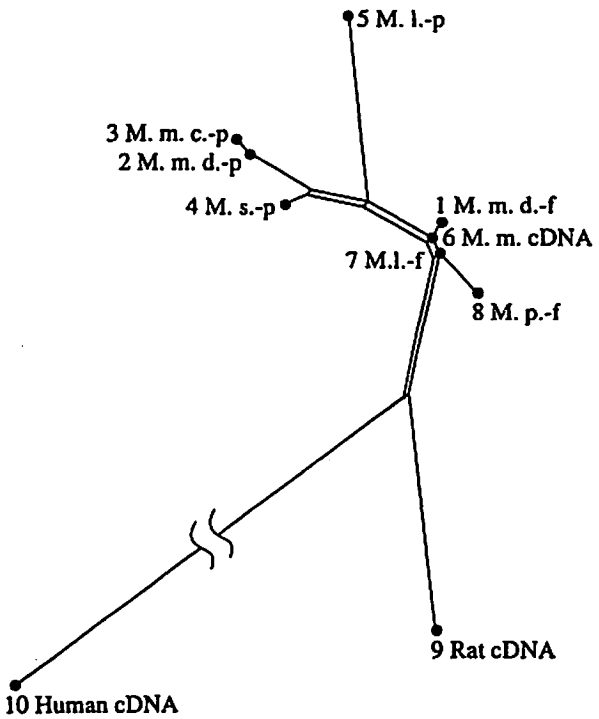

FIG. 9. Network constructed by using the SplitsTree program for the distance matrix of Table II. The length of an external branch to sequence 10 (human cDNA) was truncated, but other branch lengths were drawn proportional to the estimated lengths.

the PHYLIP WWW home page (http://evolution.genetics.washington.edu/phylip.html).

MEGA[28] is run on MS-DOS under 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 E-mail address: imeg@psuvm.psu.edu.

CLUSTAL W[29] is capable of doing multiple sequence alignment. After the alignment, it constructs neighbor-joining trees with bootstrapping. It can be retrieved from ftp.ebi.ac.uk (193.62.196.6) or from the EBI WWW home page (http://www.ebi.ac.uk/software/software.html).

MOLPHY[30] includes programs for maximum likelihood methods for both nucleotide and amino acid sequences. It can be retrieved via ftp from sunmh.ism.ac.jp (133.58.12.20).

Dendro-Maker (developed by Tadashi Imanishi) is run on Macintosh, and draws UPGMA and neighbor-joining trees. It can be retrieved via ftp from ftp.nig.ac.jp:/pub/mac/bio/dendromaker/. Treetool (developed by Mike Maciukenas) is run on Sun Sparc workstations and works with Newick

format tree files for drawing trees. It can be retrieved through the RDP WWW homepage (http://rdp.life.uiuc.edu/).

TreeTree is a package of various programs mainly related to the neighbor-joining method developed by the author (E-mail address: nsaitou@genes.nig.ac.jp). Program NJ requires a distance matrix, whereas NJNUC requires nucleotide sequences. It can be retrieved through the author's WWW home page (http://smiler.nig.ac.jp/).

Acknowledgment

This chapter was partly supported by grants-in-aid for scientific researches from the Ministry of Education, Science and Culture, Japan.

[26] Estimating Evolutionary Distances between DNA Sequences

By WEN-HSIUNG LI and XUN GU

Introduction

Estimation of the evolutionary distance between two DNA sequences requires a stochastic model for nucleotide substitution. Most models for DNA evolution can be regarded as a time-continuous Markovian process, which can be characterized by the rate matrix **R**, or, equivalently, the nucleotide substitution pattern. The most general model for **R** has 12 parameters to be estimated, but its application in practice is difficult. Indeed, one usually uses a simpler model (i.e., a simplified substitution pattern) to derive an analytical formula for estimating the distance (see, e.g., Refs. 1–4).

However, if some assumptions of a simple substitution model are violated, the estimate of a distance will be biased and will not increase linearly with time, that is, it will be nonadditive.[5] Additivity is a highly desirable property for evolutionary distances, because if it does not hold, all distance matrix methods of tree reconstruction may become statistically inconsistent, that is, may lead to an erroneous tree with a probability approaching 1 as

[1] T. H. Jukes and C. R. Cantor, *in* "Mammalian Protein Metabolism" (H. N. Munro, ed.), p. 21. Academic Press, New York, 1969.
[2] M. Kimura, *J. Mol. Evol.* **16**, 111 (1980).
[3] F. Tajima and M. Nei, *Mol. Biol. Evol.* **1**, 269 (1984).
[4] K. Tamura and M. Nei, *Mol. Biol. Evol.* **10**, 512 (1993).
[5] J. Felsenstein, *Annu. Rev. Genet.* **22**, 521 (1988).