Phylogeny methods based on distance matrices

- Clustering algorithms, based on distance matrices, are in the program of the parallel course *582670 Algorithms for bioinformatics.*

- During this course, *Biometry and bioinformatics II,* lecture time will not be devoted to this topic. Instead, you are supposed to read the chapter (20 pages) copied here from the book *Lemey et al., The phylogenetic handbook, 2009, www.cambridge.org/9780521877107*

- In addition to clustering methods, this book chapter includes also other general issues, such as bootstrapping, basics of model choice.

5

Phylogenetic inference based on distance methods

THEORY

Yves Van de Peer

5.1 Introduction

In addition to ***maximum parsimony*** (***MP***) and ***likelihood*** methods (see Chapters 6, 7 and 8), pairwise distance methods form the third large group of methods to infer evolutionary trees from sequence data (Fig. 5.1). In principle, distance methods try to fit a tree to a matrix of pairwise ***genetic distances*** (Felsenstein, 1988). For every two sequences, the distance is a single value based on the fraction of positions in which the two sequences differ, defined as ***p-distance*** (see Chapter 4). The *p*-distance is an underestimation of the true genetic distance because some of the nucleotide positions may have experienced multiple substitution events. Indeed, because mutations are continuously fixed in the genes, there has been an increasing chance of multiple substitutions occurring at the same sequence position as evolutionary time elapses. Therefore, in distance-based methods, one tries to estimate the number of substitutions that have actually occurred by applying a specific ***evolutionary model*** that makes particular assumptions about the nature of evolutionary changes (see Chapter 4). When all the pairwise distances have been computed for a set of sequences, a tree topology can then be inferred by a variety of methods (Fig. 5.2).

Correct estimation of the genetic distance is crucial and, in most cases, more important than the choice of method to infer the tree topology. Using an unrealistic evolutionary model can cause serious artifacts in tree topology, as previously shown

The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme (eds.). Published by Cambridge University Press.

	Character-based methods	Non-character-based methods
Methods based on an explicit model of evolution	Maximum likelihood methods	Pairwise distance methods
Methods not based on an explicit model of evolution	Maximum parsimony methods	

Fig. 5.1 Pairwise distance methods are non-character-based methods that make use of an explicit substitution model.

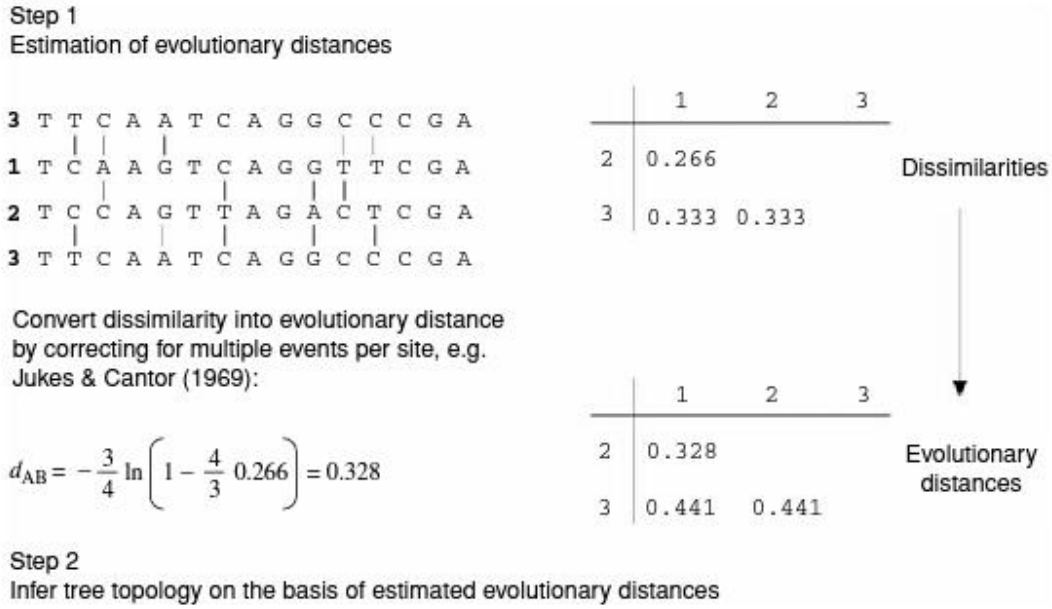


Fig. 5.2 Distance methods proceed in two steps. First, the evolutionary distance is computed for every sequence pair. Usually, this information is stored in a matrix of pairwise distances. Second, a tree topology is inferred on the basis of the specific relationships between the distance values.

in numerous studies (e.g. Olsen, 1987; Lockhart *et al.*, 1994; Van de Peer *et al.*, 1996; see also Chapter 10). However, because the exact historical record of events that occurred in the evolution of sequences is not known, the best method for estimating the genetic distance is not necessarily self-evident.

Substitution models are discussed in Chapters 4 and 10. The latter discusses how to select the best-fitting evolutionary model for a given data set of aligned nucleotide or amino acid sequences in order to get accurate estimates of genetic distances. In the following sections, it is assumed that genetic distances were estimated using

an appropriate evolutionary model, and some of the methods used for inferring tree topologies on the basis of these distances are briefly outlined. However, by no means should this be considered a complete discussion of distance methods; additional discussions are in Felsenstein (1982), Swofford *et al.* (1996), Li (1997), and Page & Holmes (1998).

5.2 Tree-inference methods based on genetic distances

The main distance-based tree-building methods are cluster analysis and minimum evolution. Both rely on a different set of assumptions, and their success or failure in retrieving the correct phylogenetic tree depends on how well any particular data set meets such assumptions.

5.2.1 Cluster analysis (UPGMA and WPGMA)

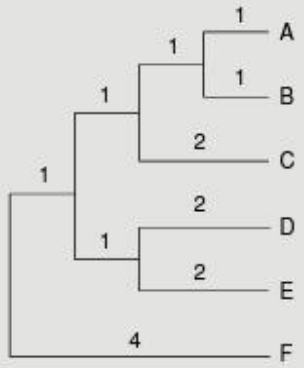
Clustering methods are tree-building methods that were originally developed to construct taxonomic ***phenograms*** (Sokal & Michener, 1958; Sneath & Sokal, 1973); that is, trees based on overall phenotypic similarity. Later, these methods were applied to phylogenetics to construct ***ultrametric trees***. ***Ultrametricity*** is satisfied when, for any three taxa, A, B, and C,

$$d_{AC} \leq \max(d_{AB}, d_{BC}). \tag{5.1}$$

In practice, (5.1) is satisfied when two of the three distances under consideration are equal and as large (or larger) as the third one. Ultrametric trees are *rooted* trees in which all the end nodes are equidistant from the root of the tree, which is only possible by assuming a ***molecular clock*** (see Chapter 11). Clustering methods such as the ***unweighted-pair group method with arithmetic means*** (***UPGMA***) or the **weighted-pair group method with arithmetic means** (***WPGMA***) use a sequential clustering algorithm. A tree is built in a stepwise manner, by grouping sequences or groups of sequences – usually referred to as *operational taxonomic units* (*OTUs*) – that are most similar to each other; that is, for which the genetic distance is the smallest. When two OTUs are grouped, they are treated as a new single OTU (Box 5.1). From the new group of OTUs, the pair for which the similarity is highest is again identified, and so on, until only two OTUs are left. The method applied in Box 5.1 is actually the WPGMA, in which the averaging of the distances is not based on the total number of OTUs in the respective clusters. For example, when OTUs A, B (which have been grouped before), and C are grouped into a new node "u," then the distance from node "u" to any other node "k" (e.g. grouping D and E) is computed as follows:

$$d_{uk} = \frac{d_{(A,B)k} + d_{Ck}}{2} \tag{5.2}$$

Box 5.1 Cluster analysis (Sneath & Sokal, 1973)



	A	B	C	D	E
B	(2)				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

N = 6

Cluster analysis proceeds as follows:

(1) Group together (cluster) these OTUs for which the distance is minimal; in this case group together A and B. The depth of the divergence is the distance between A and B divided by 2.

(2) Compute the distance from cluster (A, B) to each other OTU

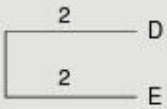
$d_{(AB)C} = (d_{AC} + d_{BC})/2 = 4$
$d_{(AB)D} = (d_{AD} + d_{BD})/2 = 6$
$d_{(AB)E} = (d_{AE} + d_{BE})/2 = 6$
$d_{(AB)F} = (d_{AF} + d_{BF})/2 = 8$

	(AB)	C	D	E
C	4			
D	6	6		
E	6	6	(4)	
F	8	8	8	8

Repeat steps 1 and 2 until all OTUs are clustered (repeat until $N = 2$)

$N = N - 1 = 5$

(1) Group together (cluster) these OTUs for which the distance is minimal, e.g. group together D and E. Alternatively, (AB) could be grouped with C.



Box 5.1 (*cont.*)

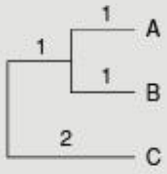
(2) Compute the distance from cluster (D, E) to each other OTU (cluster)

$d_{(DE)(AB)} = (d_{D(AB)} + d_{E(AB)})/2 = 6$
$d_{(DE)C} = (d_{DC} + d_{EC})/2 = 6$
$d_{(DE)F} = (d_{DF} + d_{EF})/2 = 8$

	(AB)	C	(DE)
C	(4)		
(DE)	6	6	
F	8	8	8

$N = N - 1 = 4$

(1) Group together these OTUs for which the distance is minimal, e.g. group (A, B) and C



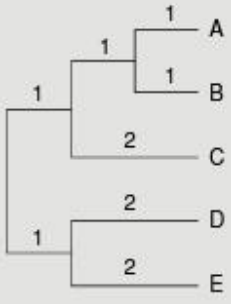
(2) Compute the distance from cluster (A, B, C) to each other OTU (cluster)

$d_{(ABC)(DE)} = (d_{(AB)(DE)} + d_{C(DE)})/2 = 6$
$d_{(ABC)F} = (d_{(AB)F} + d_{CF})/2 = 8$

	(ABC)	(DE)
(DE)	(6)	
F	8	8

$N = N - 1 = 3$

(1) Group together these OTUs for which the distance is minimal, e.g. group (A, B, C) and (D, E)

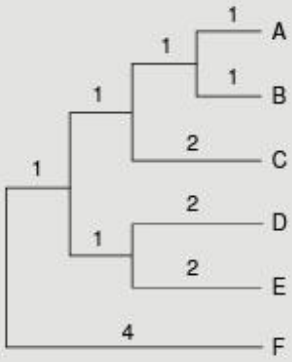


(2) Compute the distance from cluster (A, B, C, D, E) to OTU F

$$d_{(ABCDE)F} = (d_{(ABC)F} + d_{(DE)F})/2 = 8$$

	(ABC), (DE)
F	(8)

$N = N - 1 = 2$



Conversely, in UPGMA, the averaging of the distances is based on the number of OTUs in the different clusters; therefore, the distance between "u" and "k" is computed as follows:

$$d_{uk} = \frac{\left(N_{AB} d_{(A,B)k} + N_C d_{Ck}\right)}{(N_{AB} + N_C)} \tag{5.3}$$

where N_{AB} equals the number of OTUs in cluster AB (i.e. 2) and N_C equals the number of OTUs in cluster C (i.e. 1). When the data are ***ultrametric***, UPGMA and WPGMA have the same result. However, when the data are not ultrametric, they can differ in their inferences.

Until about 15 years ago, clustering was often used to infer evolutionary trees based on sequence data, but this is no longer the case. Many computer-simulation studies have shown that clustering methods such as UPGMA are extremely sensitive to unequal rates in different lineages (e.g. Sourdis & Krimbas, 1987; Huelsenbeck & Hillis, 1993). To overcome this problem, some have proposed methods that convert non-ultrametric distances into ultrametric distances. Usually referred to as *transformed distance methods*, these methods correct for unequal rates among different lineages by comparing the sequences under study to a reference sequence or an ***outgroup*** (Farris, 1977; Klotz *et al.*, 1979; Li, 1981). Once the distances are made ultrametric, a tree is constructed by clustering, as explained previously. Nevertheless, because there are now better and more effective methods to cope with non-ultrametricity and non-clock-like behavior, there is little reason left to

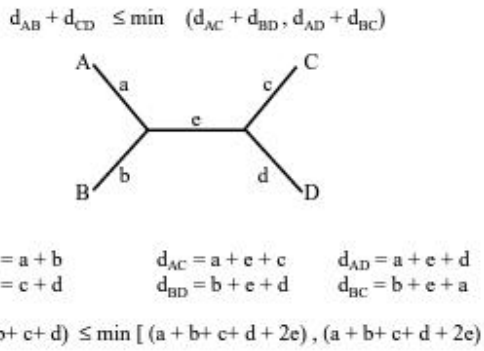


Fig. 5.3 Four-point condition. Letters on the branches of the unrooted tree represent branch lengths. The function min[ ] returns the minimum among a set of values.

use cluster analysis or transformed distance methods to infer distance trees for nucleotide or amino acid sequence data.

5.2.2 Minimum evolution and neighbor-joining

Because of the serious limitations of ordinary clustering methods, algorithms were developed that reconstruct so-called ***additive distance*** trees. Additive distances satisfy the following condition, known as the ***four-point metric condition*** (Buneman, 1971): for any four taxa, A, B, C, and D,

$$d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD}, d_{AD} + d_{BC}) \tag{5.4}$$

Only ***additive distances*** can be fitted precisely into an *unrooted* tree such that the genetic distance between a pair of OTUs equals the sum of the lengths of the branches connecting them, rather than an average, as in the case of cluster analysis. Why (5.4) needs to be satisfied is explained by the example shown in Fig. 5.3. When A, B, C, and D are related by a tree in which the sum of branch lengths connecting two terminal taxa is equal to the genetic distance between them, such as the tree in Fig. 5.3, $d_{AB} + d_{CD}$ is always smaller or equal than the minimum between $d_{AC} + d_{BD}$ and $d_{AD} + d_{BC}$ (see Fig. 5.3). The equality only occurs when the four sequences are related by a star-like tree; that is, only when the internal branch length of the tree in Fig. 5.3 is $e = 0$ (see Fig. 5.3). If (5.4) is not satisfied, A, B, C, and D cannot be represented by an additive distance tree because, to maintain the additivity of the genetic distances, one or more branch lengths of any tree relating them should be negative, which would be biologically meaningless. Real data sets often fail to satisy the four-point condition; this problem is the origin of the discrepancy between *actual* distances (i.e. those estimated from pairwise comparisons among nucleotide or amino acid sequences) and tree distances (i.e. those actually fitted into a tree) (see Section 5.2.3).

If the genetic distances for a certain data set are ultrametric, then both the ultrametric tree and the additive tree will be the same if the additive tree is rooted

at the same point as the *ultrametric tree*. However, if the genetic distances are not ultrametric due to non-clock-like behavior of the sequences, additive trees will almost always be a better fit to the distances than ultrametric trees. However, because of the finite amount of data available when working with real sequences, stochastic errors usually cause deviation of the estimated genetic distances from perfect tree additivity. Therefore, some systematic error is introduced and, as a result, the estimated tree topology may be incorrect.

Minimum evolution (***ME***) is a distance method for constructing additive trees that was first described by Kidd & Sgaramella-Zonta (1971); Rzhetsky & Nei (1992) described a method with only a minor difference. In ME, the tree that minimizes the lengths of the tree, which is the sum of the lengths of the branches, is regarded as the best estimate of the phylogeny:

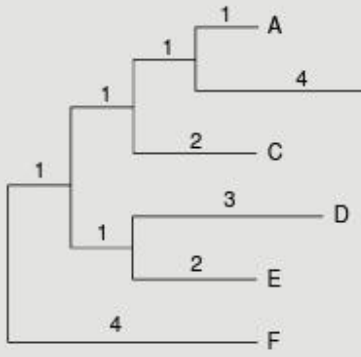
$$S = \sum_{i=1}^{2n-3} v_i \tag{5.5}$$

where n is the number of taxa in the tree and v_i is the ith branch (remember that there are $2n$–3 branches in an unrooted tree of n taxa). For each tree topology, it is possible to estimate the length of each branch from the estimated pairwise distances between all OTUs. In this respect, the method can be compared with the maximum parsimony (MP) approach (see Chapter 8), but in ME, the length of the tree is inferred from the genetic distances rather than from counting individual nucleotide substitutions over the tree (Rzhetsky & Nei, 1992, 1993; Kumar, 1996). The minimum tree is not necessarily the "true" tree. Nei *et al.* (1998) have shown that, particularly when few nucleotides or amino acids are used, the "true" tree may be larger than the minimum tree found by the optimization principle used in ME and MP. A drawback of the ME method is that, in principle, all different tree topologies have to be investigated to find the minimum tree. However, this is impossible in practice because of the explosive increase in the number of tree topologies as the number of OTUs increases (Felsenstein, 1978); an exhaustive search can no longer be applied when more than ten sequences are being used (see Chapter 1).

A good heuristic method for estimating the ME tree is the ***neighbor-joining*** (***NJ***) ***method***, developed by Saitou & Nei (1987) and modified by Studier & Keppler (1988). Because NJ is conceptually related to clustering, but without assuming a clock-like behavior, it combines computational speed with uniqueness of results.

NJ is today the method most commonly used to construct distance trees. Box 5.2 is an example of a tree constructed with the NJ method. The method adopts the ME criterion and combines a pair of sequences by minimizing the S value (see 5.5) in each step of finding a pair of neighboring OTUs. Because the S value is not minimized globally (Saitou & Nei, 1987; Studier & Keppler, 1988),

Box 5.2 The neighbor-joining method (Saitou & Nei, 1987; modified from Studier & Keppler, 1988)

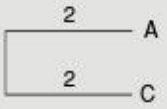


Hypothetical tree topology: since the divergence of sequences A and B, B has accumulated four times as many mutations as sequence A.

Suppose the following matrix of pairwise evolutionary distances:

	A	B	C	D	E
B	5				
C	(4)	7			
D	7	10	7		
E	6	9	6	5	
F	8	11	8	9	8

Clustering methods (discussed in Box 5.1) would erroneously group sequences A and C, since they assume clock-like behavior. Although sequences A and C look more similar, sequences A and B are more closely related.



Neighbor-joining proceeds as follows:

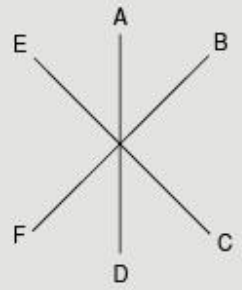
(1) Compute the net divergence r for every endnode ($N = 6$)

$r_A = 5 + 4 + 7 + 6 + 8 = 30$ $\quad r_D = 38$
$r_B = 5 + 7 + 10 + 9 + 11 = 42$ $\quad r_E = 34$
$r_C = 32$ $\quad r_F = 44$

(2) Create a rate-corrected distance matrix; the elements are defined by $M_i = d_{ij} - (r_i + r_i)/(N - 2)$

$M_{AB} = d_{AB} - (r_A + r_B)/(N - 2) = 5 - (30 + 42)/4 = -13$
$M_{AC} = \ldots.$
$\ldots$

	A	B	C	D	E
B	(−13)				
C	−11.5	−11.5			
D	−10	−10	−10.5		
E	−10	−10	−10.5	(−13)	
F	−10.5	−10.5	−11	−11.5	−11.5



(3) Define a new node that groups OTUs i and j for which M_i is minimal For example, sequences A and B are neighbors and form a new node U (but, alternatively, OTUs D and E could have been joined; see further)

(4) Compute the branch lengths from node U to A and B

$$S_{AU} = d_{AB}/2 + (r_A - r_B)/2(N-2) = 1$$
$$S_{BU} = d_{AB} - S_{AU} = 4$$

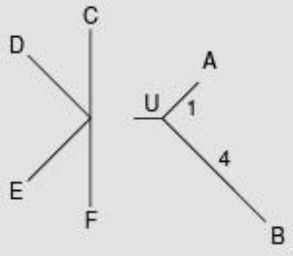
or alternatively

$$S_{BU} = d_{AB}/2 + (r_B - r_A)/2(N-2) = 4$$
$$S_{AU} - d_{AB} - S_{BU} = 1$$

(5) Compute new distances from node U to each other terminal node

$$d_{CU} = (d_{AC} + d_{BC} - d_{AB})/2 = 3$$
$$d_{DU} = (d_{AD} + d_{BD} - d_{AB})/2 = 6$$
$$d_{EU} = (d_{AE} + d_{BE} - d_{AB})/2 = 5$$
$$d_{FU} = (d_{AF} + d_{BF} - d_{AB})/2 = 7$$

	U	C	D	E
C	3			
D	6	7		
E	5	6	5	
F	7	8	9	8



(6) $N = N - 1$; repeat step 1 through 5

Box 5.2 *(cont.)*

(1) Compute the net divergence r for every endnode ($N = 5$)

$r_B = 21 \quad r_E = 24$
$r_C = 24 \quad r_F = 32$
$r_D = 27$

(2) Compute the modified distances:

	U	C	D	E
C	(−12)			
D	−10	−11		
E	−10	−10	(−12)	
F	−10.7	−10.7	−10.7	−10.7

(3) Define a new node: e.g. U and C are neighbors and form a new node V; alternatively, D and E could be joined

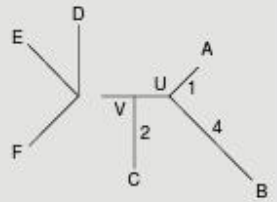
(4) Compute the branch lengths from node V to C and U

$S_{UV} = d_{CU}/2 + (r_U - r_C)/2(N-2) = 1$
$S_{CV} = d_{CU} - S_{UV} = 2$

(5) Compute distances from V to each other terminal node

$d_{DV} = (d_{DU} + d_{CB} - d_{CU})/2 = 5$
$d_{EV} = (d_{EU} + d_{CB} - d_{CU})/2 = 4$
$d_{FV} = (d_{FU} + d_{CF} - d_{CU})/2 = 6$

	V	D	E
D	5		
E	4	5	
F	6	9	8



(6) $N = N - 1$; repeat step 1 through 5

(1) Compute the net divergence r for every endnode ($N = 4$)

$r_V = 15 \quad r_E = 17$
$r_D = 19 \quad r_F = 23$

(2) Compute the modified distances

	V	D	E
D	−12		
E	−12	(−13)	
F	(−13)	−12	−12

(3) Define a new node: e.g. D and E are neighbors and form a new node W; alternatively, F and V could be joined

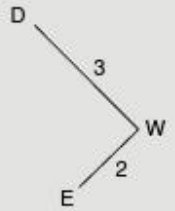
(4) Compute the branch lengths from node W to E and D

$S_{DW} = d_{DE}/2 + (r_D - r_E)/2(N-2) = 3$
$S_{DW} = d_{DE} - S_{DW} = 2$

(5) Compute distances from W to each other terminal node

$d_{VW} = (d_{DV} + d_{EV} - d_{DE})/2 = 2$
$d_{FW} = (d_{Dr} + d_{gr} - d_{DE})/2 = 6$

	W	V
V	2	
F	6	6



(6) $N = N - 1$; repeat step 1 through 5

(1) Compute the net divergence r for every endnode ($N = 3$)

$r_V = 8 \quad r_F = 17 \quad r_W = 8$

(2) Compute the modified distances

	W	V
V	−14	
F	−14	−14

(3) Define a new node: e.g. V and F are neighbors and form a new node X; Alternatively, W and V could be joined, or W and F could be joined

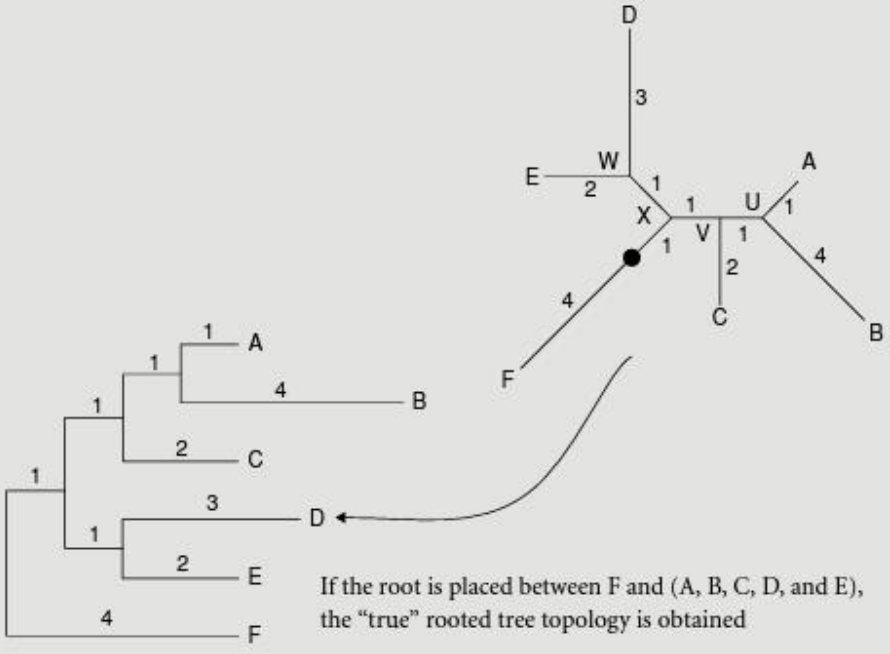
	W
X	1

(4) Compute the branch lengths from X to V and F

$S_{VX} = d_{FV}/2 + (r_V - r_F)/2(N-2) = 1$
$S_{FX} = d_{FV} - S_{VX} = 5$

(5) Compute distances from X to each other terminal node

$d_{WX} = (d_{FW} + d_{VW} - d_{FV})/2 = 1$



If the root is placed between F and (A, B, C, D, and E), the "true" rooted tree topology is obtained

the NJ tree may not be the same as the ME tree if pairwise distances are not additive (Kumar, 1996). However, NJ trees have proven to be the same or similar to the ME tree (Saitou & Imanishi, 1989; Rzhetsky & Nei, 1992, 1993; Russo *et al.*, 1996; Nei *et al.*, 1998). Several methods have been proposed to find ME trees, starting from an NJ tree but evaluating alternative topologies close to the NJ tree

by conducting local rearrangements (e.g. Rzhetsky & Nei, 1992). Nevertheless, it is questionable whether this approach is really worth considering (Saitou & Imanishi, 1989; Kumar, 1996), and it has been suggested that combining NJ and bootstrap analysis (Felsenstein, 1985) might be the best way to evaluate trees using distance methods (Nei *et al.*, 1998).

Recently, alternative versions of the NJ algorithm have been proposed, including **BIONJ** (Gascuel, 1997), **generalized neighbor-joining** (Pearson *et al.*, 1999), **weighted neighbor-joining** or **weighbor** (Bruno *et al.*, 2000), **neighbor-joining maximum-likelihood** (NJML; Ota & Li, 2000), **QuickJoin** (Mailund & Pedersen, 2004), **multi-neighbor-joining** (Silva *et al.*, 2005) and **relaxed neighbor-joining** (Evans *et al.*, 2006). BIONJ and weighbor both consider that long genetic distances present a higher variance than short ones when distances from a newly defined node to all other nodes are estimated (see Box 5.2). This should result in higher accuracy when distantly related sequences are included in the analysis. Furthermore, the weighted neighbor-joining method of Bruno *et al.* (2000) uses a likelihood-based criterion rather than the ME criterion of Saitou & Nei (1987) to decide which pair of OTUs should be joined. NJML divides an initial neighbor-joining tree into subtrees at internal branches having bootstrap values higher than a threshold (Ota & Li, 2000). A topology search is then conducted using the ***maximum-likelihood*** method only re-evaluating branches with a bootstrap value lower than the threshold. The generalized neighbor-joining method of Pearson *et al.* (1999) keeps track of multiple, partial, and potentially good solutions during its execution, thus exploring a greater part of the tree space. As a result, the program is able to discover topologically distinct solutions that are close to the ME tree. Multi-neighbor-joining also keeps various partial solutions resulting in a higher chance to recover the minimum evolution tree (Silva *et al.*, 2005). QuickJoin and relaxed neighbor-joining use heuristics to improve the speed of execution, making them suitable for large-scale applications (Mailund & Pedersen, 2004; Evans *et al.*, 2006).

Figure 5.4 shows two trees based on evolutionary distances inferred from 20 small subunit ribosomal RNA sequences (Van de Peer *et al.*, 2000a). The tree in Fig. 5.4a was constructed by clustering (UPGMA) and shows some unexpected results. For example, the sea anemone, *Anemonia sulcata*, clusters with the fungi rather than the other animals, as would have been expected. Furthermore, neither the basidiomycetes nor the ascomycetes form a clear-cut ***monophyletic*** grouping. In contrast, on the NJ tree all animals form a highly supported monophyletic grouping, and the same is true for basidiomycetes and ascomycetes. The NJ tree also shows why clustering could not resolve the right relationships. Clustering methods are sensitive to unequal rates of evolution in different lineages; as is clearly seen, the branch length of *Anemonia sulcata* differs greatly from that of the

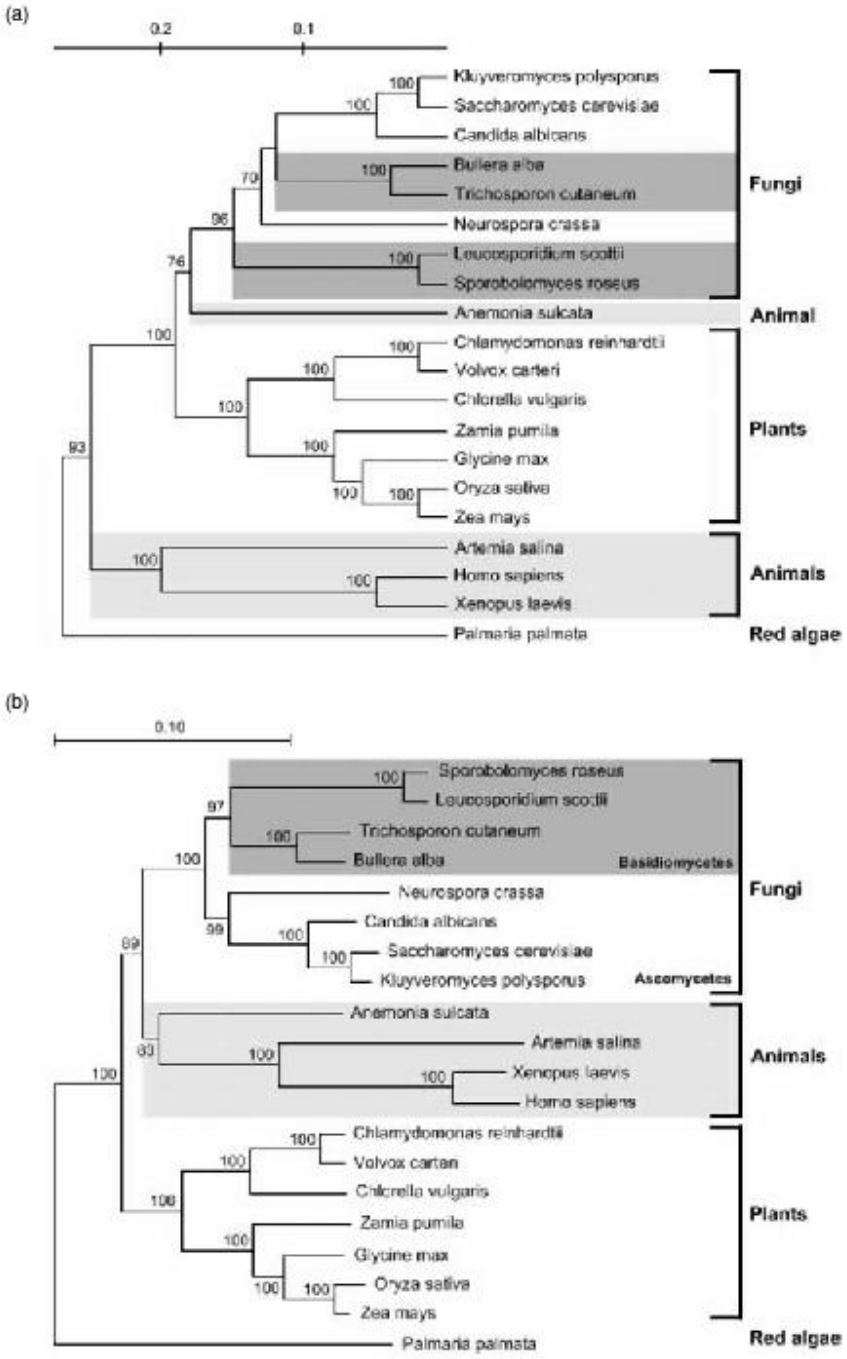


Fig. 5.4 Phylogenetic trees based on the comparison of 20 small subunit ribosomal RNA sequences. Animals are indicated by light gray shading; dark gray shading indicates the basidiomycetes. The scales on top measure evolutionary distance in substitutions per nucleotide. The red alga *Palmaria palmata* was used to root the tree. (a) Ultrametric tree obtained by clustering. (b) Neighbor-joining tree.

other animals. Also, different basidiomycetes have evolved at different rates and, as a result, they are split into two groups in the tree obtained by clustering (see Fig. 5.4a).

5.2.3 Other distance methods

It is possible for every tree topology to estimate the length of all branches from the estimated pairwise distances between all OTUs (e.g. Fitch & Margoliash, 1967; Rzhetsky & Nei, 1993). However, when summing the branch lengths between sequences, there is usually some discrepancy between the distance obtained (referred to as the *tree* distance or ***patristic distance***) and the distance as estimated directly from the sequences themselves (the observed or actual distances) due to deviation from tree additivity (see Section 5.2.2). Whereas ME methods try to find the tree for which the sum of the lengths of branches is minimal, other distance methods have been developed to construct additive trees depending on goodness of fit measures between the actual distances and the tree distances. The best tree, then, is that tree that minimizes the discrepancy between the two distance measures. When the criterion for evaluation is based on a *least-squares fit*, the goodness of fit F is given by the following:

$$F = \sum_{i,j} w_{ij}(D_{ij} - d_{ij})^2 \tag{5.6}$$

where D_{ij} is the observed distance between i and j, d_{ij} is the tree distance between i and j, and w_{ij} is different for different methods. For example, in the Fitch and Margoliash method (1967), w_{ij} equals $1/D_{ij}^2$; in the Cavalli-Sforza and Edwards approach (1967), w_{ij} equals 1. Other values for w_{ij} are also possible (Swofford *et al.*, 1996) and using different values can influence which tree is regarded as the best. To find the tree for which the discrepancy between actual and tree distances is minimal, one has in principle to investigate all different tree topologies. However, as with ME, distance methods that are based on the evaluation of an explicit criterion, such as goodness of fit between observed and tree distances, suffer from the explosive increase in the number of different tree topologies as more OTUs are examined. Therefore, heuristic approaches, such as ***stepwise addition*** of sequences and local and global rearrangements, must be applied when trees are constructed on the basis of ten or more sequences (e.g. Felsenstein, 1993).

5.3 Evaluating the reliability of inferred trees

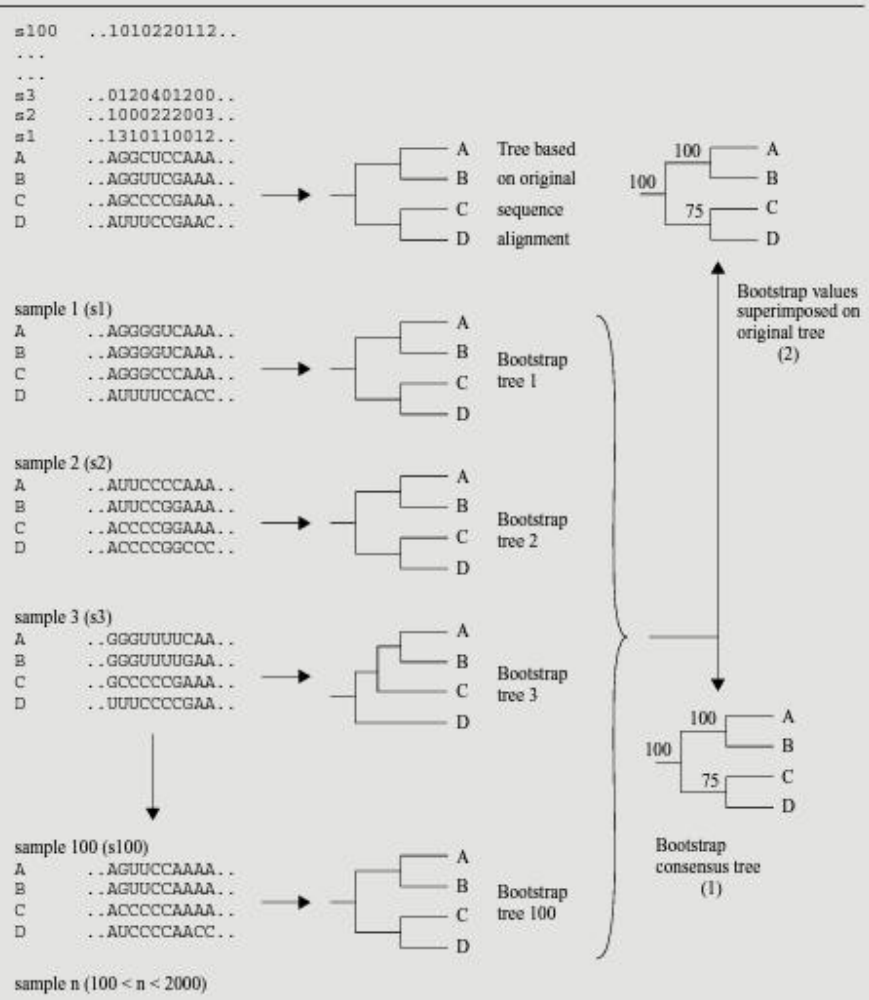
The two techniques used most often to evaluate the reliability of the inferred tree or, more precisely, the reliability of specific clades in the tree are bootstrap analysis (Box 5.3) and ***jackknifing*** (see Section 5.3.2).

5.3.1 Bootstrap analysis

Bootstrap analysis is a widely used sampling technique for estimating the statistical error in situations in which the underlying ***sampling distribution*** is either unknown or difficult to derive analytically (Efron & Gong, 1983). The bootstrap method offers a useful way to approximate the underlying distribution by resampling from the original data set. Felsenstein (1985) first applied this technique to the estimation of confidence intervals for phylogenies inferred from sequence data. First, the sequence data are bootstrapped, which means that a new alignment is obtained from the original by randomly choosing columns from it with replacements. Each column in the alignment can be selected more than once or not at all until a new set of sequences, a *bootstrap replicate*, the same length as the original one has been constructed. Therefore, in this resampling process, some characters will not be included at all in a given bootstrap replicate and others will be included once, twice, or more. Second, for each reproduced (i.e. artificial) data set, a tree is constructed, and the proportion of each clade among all the bootstrap replicates is computed. This proportion is taken as the statistical confidence supporting the monophyly of the subset.

Two approaches can be used to show bootstrap values on phylogenetic trees. The first summarizes the results of bootstrapping in a ***majority-rule consensus*** tree (see Box 5.3, Option 1), as done, for example, in the PHYLIP software package (Felsenstein, 1993). The second approach superimposes the bootstrap values on the tree obtained from the original sequence alignment (see Box 5.3, Option 2). In this case, all bootstrap trees are compared with the tree based on the original alignment and the number of times a cluster (as defined in the original tree) is also found in the bootstrap trees is recorded. Although in terms of general statistics the theoretical foundation of the bootstrap has been well established, the statistical properties of the bootstrap estimation applied to sequence data and evolutionary relationships are less well understood; several studies have reported on this problem (Zharkikh & Li, 1992a, b; Felsenstein & Kishino, 1993; Hillis & Bull, 1993). Bootstrapping itself is a neutral process that only reflects the phylogenetic signal (or noise) in the data as detected by the tree-construction method used. If the tree-construction method makes a bad estimate of the phylogeny due to systematic errors (caused by incorrect assumptions in the tree-construction method), or if the sequence data are not representative of the underlying distribution, the resulting confidence intervals obtained by the bootstrap are not meaningful. Furthermore, if the original sequence data are biased, the bootstrap estimates will be too. For example, if two sequences are clustered together because they both share an unusually high GC content, their artificial clustering will be supported by bootstrap analysis at a high confidence level. Another example is the artificial grouping of sequences with an increased ***evolutionary rate***. Due to the systematic underestimation of

Box 5.3 Bootstrap Analysis (Felsenstein, 1985)



the genetic distances when applying an unrealistically simple substitution model, distant species either will be clustered together or drawn toward the root of the tree. When the bootstrap trees are inferred on the basis of the same incorrect evolutionary model, the early divergence of long branches or the artificial clustering of long branches (the so-called ***long-branch attraction***) will be supported at a high bootstrap level. Therefore, when there is evidence of these types of artifacts, bootstrap results should be interpreted with caution.

In conclusion, bootstrap analysis is a simple and effective technique to test the relative stability of groups within a phylogenetic tree. The major advantage of the bootstrap technique is that it can be applied to basically all tree-construction

methods, although it must be remembered that applying the bootstrap method multiplies the computer time needed by the number of bootstrap samples requested. Between 200 and 2000 resamplings are usually recommended (Hedges, 1992; Zharkikh & Li, 1992a). Overall, under normal circumstances, considerable confidence can be given to branches or groups supported by more than 70% or 75%; conversely, branches supported by less than 70% should be treated with caution (Zharkikh & Li, 1992a; see also Van de Peer *et al.*, 2000b for a discussion about the effect of species sampling on bootstrap values).

5.3.2 Jackknifing

An alternative resampling technique often used to evaluate the reliability of specific clades in the tree is the so-called ***delete-half jackknifing*** or jackknife. Jackknife randomly purges half of the sites from the original sequences so that the new sequences will be half as long as the original. This resampling procedure typically will be repeated many times to generate numerous new samples. Each new sample (i.e. new set of sequences) – no matter whether from bootstrapping or jackknifing – will then be subjected to regular phylogenetic reconstruction. The frequencies of subtrees are counted from reconstructed trees. If a subtree appears in all reconstructed trees, then the jackknifing *value* is 100%; that is, the strongest possible support for the subtree. As for bootstrapping, branches supported by a jackknifing *value* less than 70% should be treated with caution.

5.4 Conclusions

Pairwise distance methods are tree-construction methods that proceed in two steps. First, for all pairs of sequences, the genetic distance is estimated (Swofford *et al.*, 1996) from the observed sequence dissimilarity (*p*-distance) by applying a correction for multiple substitutions. The genetic distance thus reflects the expected mean number of changes per site that have occurred, since two sequences diverged from their common ancestor. Second, a phylogenetic tree is constructed by considering the relationship between these distance values. Because distance methods strongly reduce the phylogenetic information of the sequences (to basically one value per sequence pair), they are often regarded as inferior to character-based methods (see Chapters 6, 7 and 8). However, as shown in many studies, this is not necessarily so, provided that the genetic distances were estimated accurately (see Chapter 10). Moreover, contrary to maximum parsimony, distance methods have the advantage – which they share with maximum-likelihood methods – that an appropriate substitution model can be applied to correct for multiple mutations. Popular distance methods such as the NJ and the Fitch and Margoliash methods have long proven to be quite efficient in finding the "true" tree topologies or those

that are close (Saitou & Imanishi, 1989; Huelsenbeck & Hillis, 1993; Charleston *et al.*, 1994; Kuhner & Felsenstein, 1994; Nei *et al.*, 1998). NJ has the advantage of being very fast, which allows the construction of large trees including hundreds of sequences; this significant difference in speed of execution compared to other distance methods has undoubtedly accounted for the popularity of the method (Kuhner & Felsenstein, 1994; Van de Peer & De Wachter, 1994).

Distance methods are implemented in many different software packages, including **PHYLIP** (Felsenstein, 1993), **MEGA4** (Kumar *et al.*, 1993), **TREECON** (Van de Peer & Dewachter, 1994), **PAUP*** (Swofford, 2002), **DAMBE** (Xia, 2000), and many more.