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. © Cambridge University Press 2009.

	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.

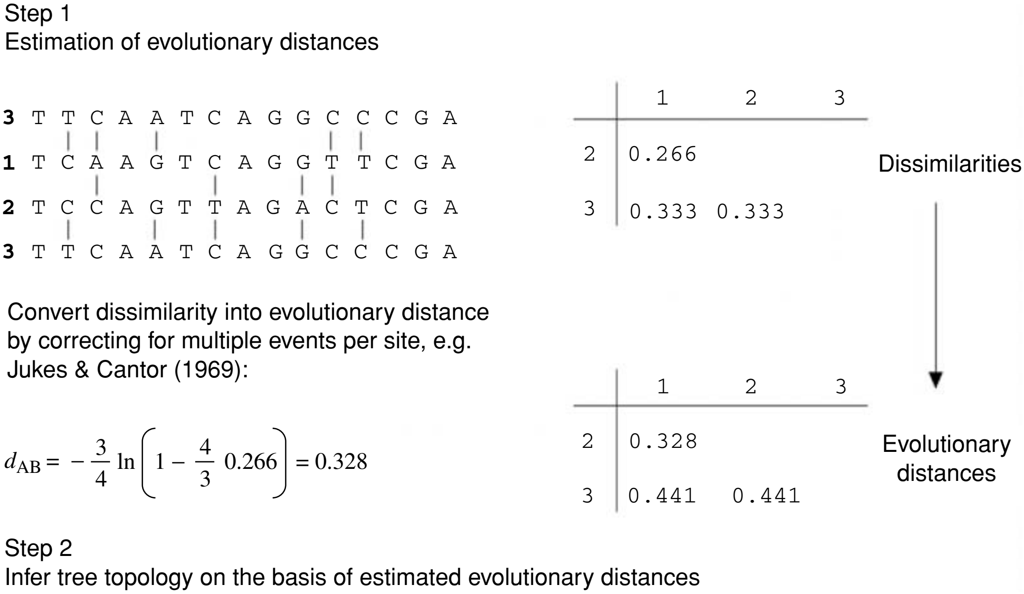

Step 1
Estimation of evolutionary distances
3 T T C A A T C A G G C C C G A
1 T C A A G T C A G G T T C G A
2 T C C A G T T A G A C T C G A
3 T T C A A T C A G G C C C G A
1 2 3
2 0.266
3 0.333 0.333
Dissimilarities
Convert dissimilarity into evolutionary distance by correcting for multiple events per site, e.g. Jukes & Cantor (1969):
$d_{AB} = -\frac{3}{4} \ln\left(1 - \frac{4}{3}\, 0.266\right) = 0.328$
1 2 3
2 0.328
3 0.441 0.441
Evolutionary distances
Step 2
Infer tree topology on the basis of estimated evolutionary distances

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)

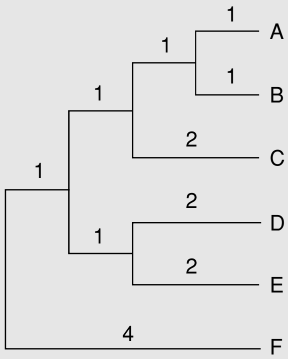
Clustering methods are tree-building methods that were originally developed to construct taxonomic ***phenograms*** (Sokal & Michener, 1958; Sneath & Sokal, 1973); that is, trees based on overall phenotypic similarity. Later, these methods were applied to phylogenetics to construct ***ultrametric trees***.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

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

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:

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

Box 5.1 Cluster analysis (Sneath & Sokal, 1973)

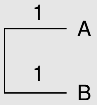


	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

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

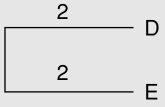
$$d_{(AB)C} = (d_{AC} + d_{BC})/2 = 4$$
$$d_{(AB)D} = (d_{AD} + d_{BD})/2 = 6$$
$$d_{(AB)E} = (d_{AE} + d_{BE})/2 = 6$$
$$d_{(AB)F} = (d_{AF} + d_{BF})/2 = 8$$

	(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.*)

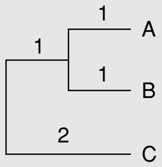
(2) Compute the distance from cluster (D, E) to each other OTU (cluster)

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

	(AB)	C	(DE)
C	$\boxed{4}$		
(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



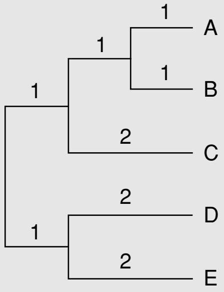
(2) Compute the distance from cluster (A, B, C) to each other OTU (cluster)

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

	(ABC)	(DE)
(DE)	$\boxed{6}$	
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)

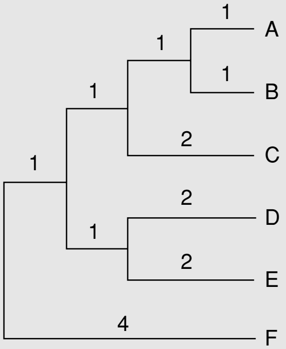


(2) Compute the distance from cluster (A, B, C, D, E) to OTU F

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

	(ABC), (DE)
F	8

$N = N - 1 = 2$



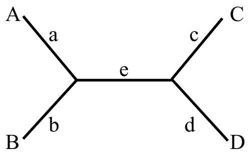
Conversely, in UPGMA, the averaging of the distances is based on the number of OTUs in the different clusters; therefore, the distance between "*u*" and "*k*" is computed as follows:

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

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

$$d_{AB} + d_{CD} \leq \min \quad (d_{AC} + d_{BD}\,,\, d_{AD} + d_{BC})$$



$$\begin{array}{lll} d_{AB} = a + b & d_{AC} = a + e + c & d_{AD} = a + e + d \\ d_{CD} = c + d & d_{BD} = b + e + d & d_{BC} = b + e + a \end{array}$$

$$(a + b + c + d) \leq \min\,[\,(a + b + c + d + 2e)\,,\,(a + b + c + d + 2e)\,]$$

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:

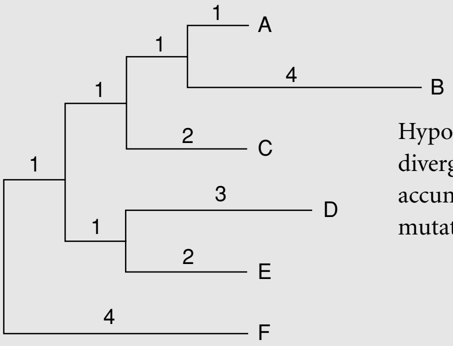
$$S = \sum_{i=1}^{2n-3} v_i \tag{5.5}$$

where n is the number of taxa in the tree and v_i is the *i*th 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)

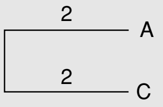


Hypothetical tree topology: since the divergence of sequences A and B, B has accumulated four times as many mutations as sequence A.

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:

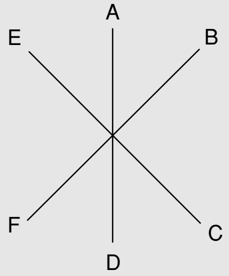
(1) Compute the net divergence r for every endnode ($N = 6$)

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

(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

$$S_{\mathrm{AU}} = d_{\mathrm{AB}}/2 + (r_{\mathrm{A}} - r_{\mathrm{B}})/2(N-2) = 1$$
$$S_{\mathrm{BU}} = d_{\mathrm{AB}} - S_{\mathrm{AU}} = 4$$

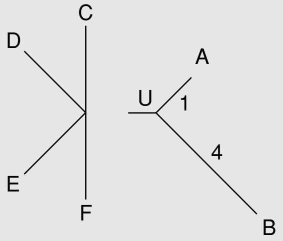
or alternatively

$$S_{\mathrm{BU}} = d_{\mathrm{AB}}/2 + (r_{\mathrm{B}} - r_{\mathrm{A}})/2(N-2) = 4$$
$$S_{\mathrm{AU}} - d_{\mathrm{AB}} - S_{\mathrm{BU}} = 1$$

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

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

	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 ($N = 5$)

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

(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

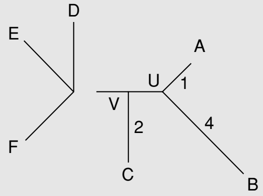
(4) Compute the branch lengths from node V to C and U

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

(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 ($N = 4$)

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

(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

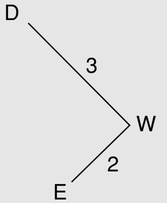
(4) Compute the branch lengths from node W to E and D

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

(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

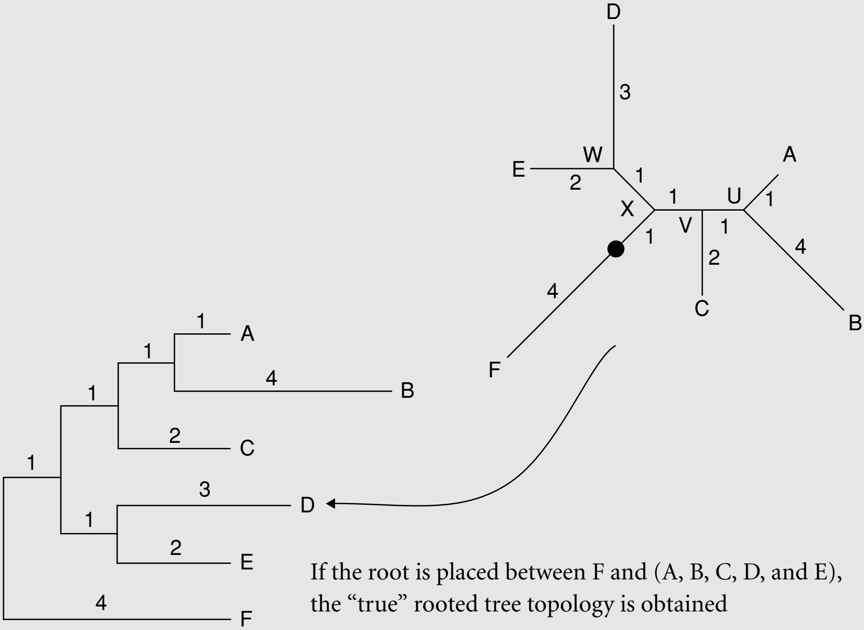
(4) Compute the branch lengths from X to V and F

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

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

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

	W
X	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

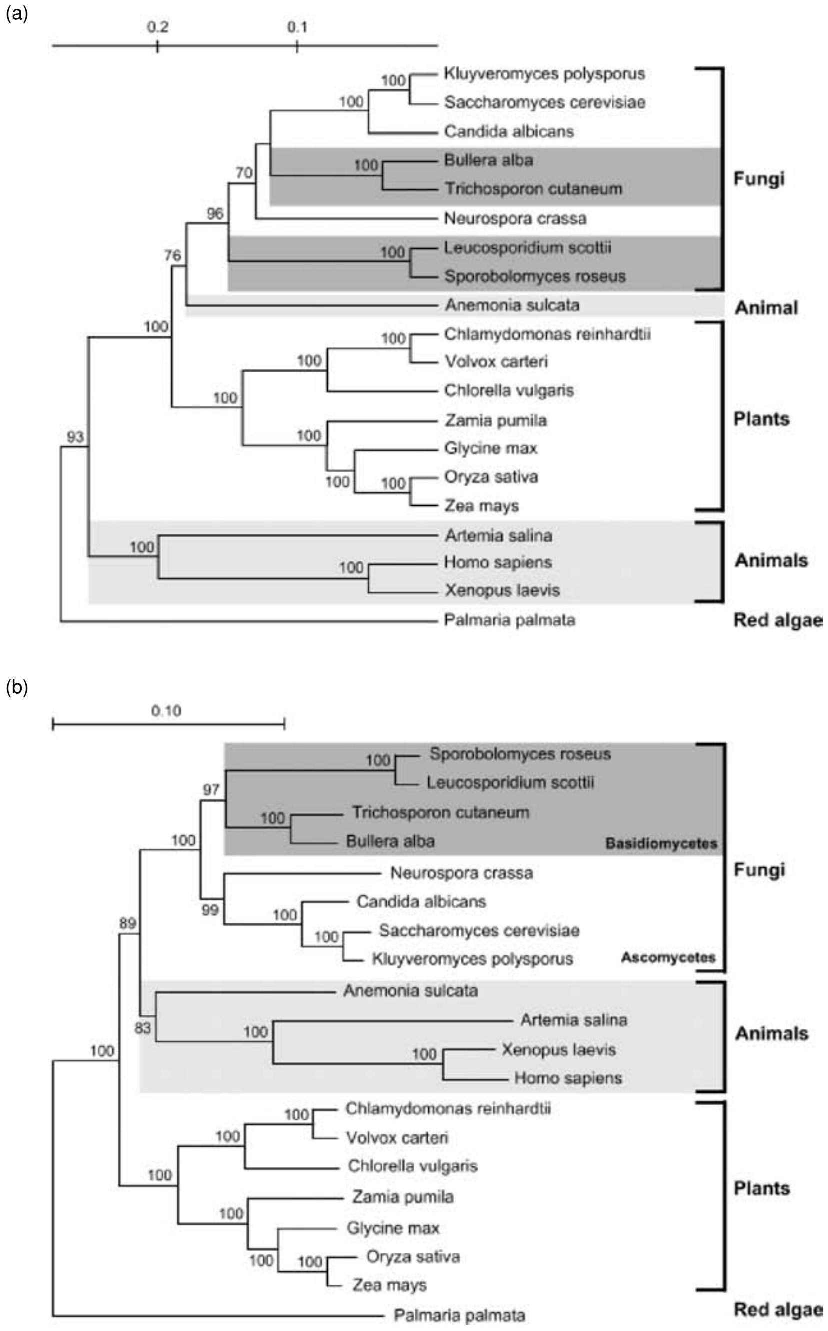


Fig. 5.4 Phylogenetic trees based on the comparison of 20 small subunit ribosomal RNA sequences. Animals are indicated by light gray shading; dark gray shading indicates the basidiomycetes. The scales on top measure evolutionary distance in substitutions per nucleotide. The red alga *Palmaria palmata* 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 $1/\mathrm{D}_{ij}^2$; 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

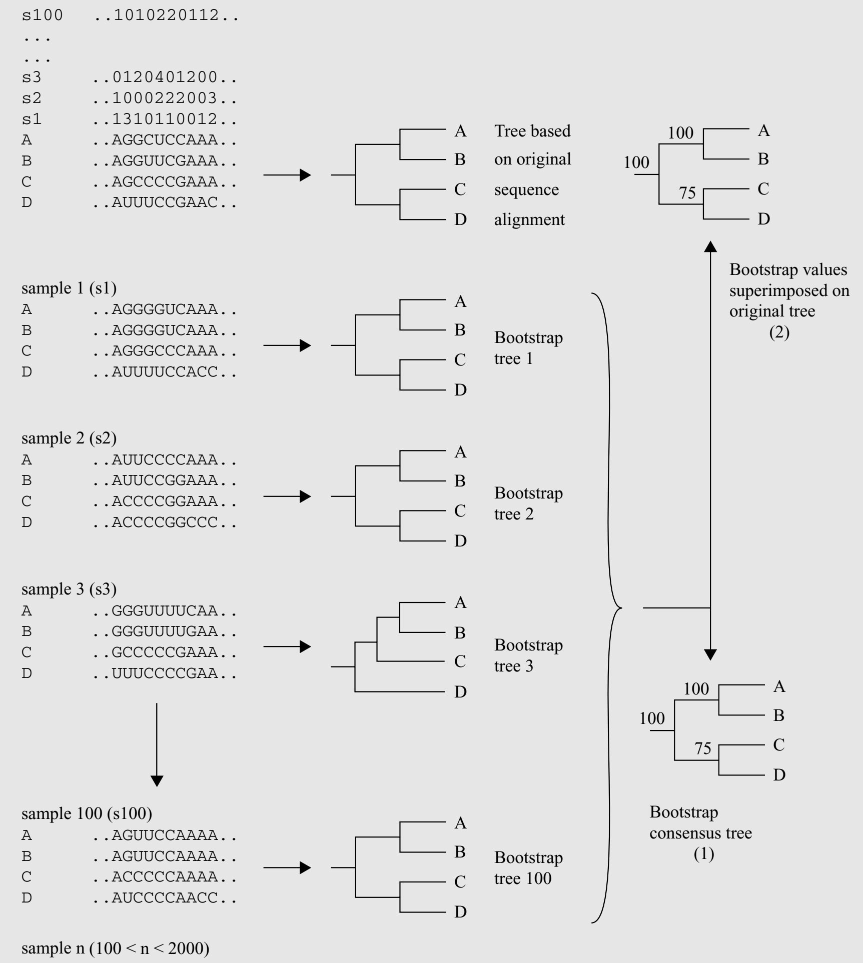
The two techniques used most often to evaluate the reliability of the inferred tree or, more precisely, the reliability of specific clades in the tree are bootstrap analysis (Box 5.3) and ***jackknifing*** (see Section 5.3.2).

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.

PRACTICE

Marco Salemi

5.5 Programs to display and manipulate phylogenetic trees

In the following sections, we will discuss two applications that are useful for displaying, editing, and manipulating phylogenetic trees: TreeView and FigTree. TreeView 1.6.6 (*http://taxonomy.zoology.gla.ac.uk/rod/treeview.html*) is a user-friendly and freely available program for high-quality display of phylogenetic trees, such as the ones reconstructed using Phylip, Mega, Tree-Puzzle (Chapter 6) or Paup* (Chapter 8). The program also implements some tree manipulation functionalities, for example, defining outgroups and re-rooting trees. Program versions available for MacOsX and PC use almost identical interfaces; a manual and installation instructions are available on the website, but note that a printer driver needs to be installed to run the program under Windows. FigTree is a freeware application for visualization and sophisticated editing of phylogenetic trees. Trees can be exported in PDF format for publication quality figures or saved in nexus format with editing information included as a FigTree block. The program is written in JAVA and both MacOSX and Windows executables can be downloaded from *http://tree.bio.ed.ac.uk/software/figtree/*. The program Mega, discussed in the previous chapter, also contains a built-in module for displaying and manipulation of phylogenetic trees.

Phylogenetic trees are almost always saved in one of two formats: NEWICK or NEXUS. The NEWICK standard for a computer-readable tree format makes use of the correspondence between trees and nested parentheses; an example for a four-taxon tree is shown in Fig. 5.5. In this notation, a tree is basically a string of balanced pairs of parenthesis with every two balanced parentheses representing an internal node. Branch lengths for terminal branches and internal nodes are written after a colon. The NEXUS format incorporates NEWICK formatting along with other commands and usually has a separate taxa-definition block (see Box 8.4 for more details on the NEXUS alignment format). The NEXUS equivalent for the tree in Fig. 5.5 with branch lengths is:

```
#NEXUS
Begin trees;
   Translate
      1 A,
      2 B,
      3 C,
```

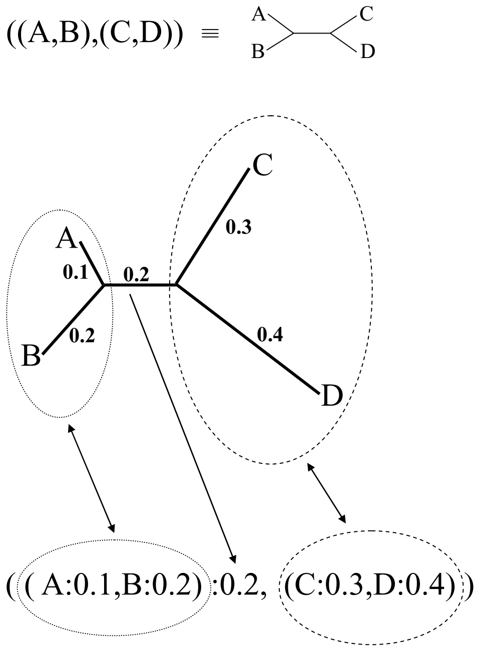


Fig. 5.5 NEWICK representation of phylogenetic trees. A hypothetical unrooted tree of four taxa (A, B, C, and D) with numbers along the branches indicating estimated genetic distances and its description in NEWICK format (see text for more details).

```
      4 D,
      ;
tree PAUP_1 = [&U] ((1:0.1,2:0.2):0.2,(3:0.3,4:0.4));
End;
```

5.6 Distance-based phylogenetic inference in Phylip

The Phylip software package implements four different distance-based tree-building methods: the Neighbour-Joining (NJ) and the UPGMA methods, carried out by the program Neighbor.exe; the *Fitch–Margoliash* method, carried out by the program Fitch.exe; and the Fitch and Margoliash method assuming a molecular clock, carried out by the program Kitch.exe. The UPGMA method and the algorithm implemented in the program Kitch.exe both assume ultrametricity of the sequences in the data set, i.e. that the sequences are contemporaneous and accumulate mutations over time at a more or less constant rate. As discussed above, ultrametric methods tend to produce less reliable phylogenetic trees when mutations occur at significantly different rates in different lineages.

Non-ultrametric methods, such as NJ or Fitch–Margoliash, do not assume a molecular clock, and they are better in recovering the correct phylogeny in situations when different lineages exhibit a strong heterogeneity in evolutionary rates. Chapter 11 will discuss how to test the molecular clock hypothesis for contemporaneously and serially sampled sequences. However, the statistical evaluation of the molecular clock and the substitution model best fitting the data (see previous chapter and Chapter 10) require the knowledge of the tree topology relating the operational taxonomic units (OTUs, i.e. the taxa) under investigation. Therefore, the first step in phylogenetic studies usually consists of constructing trees using a simple evolutionary model, like the JC69 or the Kimura 2-parameter model (Kimura, 1980), and tree-building algorithms not assuming a molecular clock. The reliability of each clade in the tree is then tested with bootstrap analysis or Jackknifing (see Section 5.3). In this way it is possible to infer one or more trees that, although not necessarily the true phylogeny, are reasonable hypotheses for the data. Such "approximate" trees are usually appropriate to test a variety of evolutionary hypotheses using maximum likelihoods methods, including the model of nucleotide substitution and the molecular clock (see Chapters 10 and 11). Moreover, when a "reasonable" tree topology is known, the free parameters of any model, for example, the ***transition/transversion ratio*** or the shape parameter α of the Γ-distribution (see previous chapter), can be estimated from the data set by maximum likelihood methods. These topics will be covered in Chapters 6 and 10.

In what follows, we will use the Windows versions of PHYLIP, MEGA, TREEVIEW and FIGTREE. However, the exercises could also be carried out under MacOSX using the Mac executables of each application (PHYLIP, TREEVIEW and FIGTREE) or using a virtual PC emulator (MEGA). The data sets required for the exercises can be downloaded from *www.thephylogenetichandbook.org*.

5.7 Inferring a Neighbor-Joining tree for the primates data set

To infer a NJ tree using the program NEIGHBOR.EXE from the PHYLIP package for the primates alignment (`primates.phy` file: alignment in sequential PHYLIP format), an input file with pairwise evolutionary distances is required. Therefore, before starting the neighbor-joining program, first calculate the distance matrix using the program DNADIST.EXE, as explained in the previous chapter, employing the F84 model and an empirical transition/transversion ratio of 2. The matrix in the `outfile` is already in the appropriate format, and it can be used directly as input file for NEIGHBOR.EXE. Rename the `outfile` file to `infile` and run NEIGHBOR.EXE by double-clicking the application's icon in the exe folder within the PHYLIP folder; the following menu will appear:

```
Neighbor-Joining/UPGMA method version 3.66

Settings for this run:
N Neighbor-joining or UPGMA tree?        Neighbor-joining
O                  Outgroup root?        No, use as
                                         outgroup species 1
L      Lower-triangular data matrix?     No
R      Upper-triangular data matrix?     No
S                     Subreplicates?     No
J  Randomize input order of species?     No. Use input order
M           Analyze multiple data sets?  No
0   Terminal type (IBM PC, ANSI, none)?  IBM PC
1    Print out the data at start of run  No
2  Print indications of progress of run  Yes
3                        Print out tree  Yes
4       Write out trees onto tree file?  Yes
Y  to accept these or type the letter for one to change
```

Option N allows the user to choose between NJ and UPGMA as tree-building algorithm. Option O asks for an outgroup. Since NJ trees do not assume a molecular clock, the choice of an outgroup merely influences the way the tree is drawn; for now, leave option O unchanged. Options L and R allow the user to use as input a pairwise distance matrix written in lower-triangular or upper-triangular format. The rest of the menu is self-explanatory: enter Y to start the computation. Outputs are written into outfile and outtree, respectively. The content of the outfile can be explored using a text editor and contains a description of the tree topology and a table with the branch lengths. The outtree contains a description of the tree in the so-called NEWICK format (see Fig. 5.5). Run TreeView.exe and select Open from the File menu. Choose All Files in Files of type and open the outtree just created in the exe sub folder of the PHYLIP folder. The tree in Fig. 5.6a will appear. At the top of the TreeView window a bar with four buttons (see Fig. 5.6a) indicates the kind of tree graph being displayed. The highlighted button indicates that the tree is shown as a ***cladogram***, which only displays the phylogenetic relationships among the taxa in the data set. In this kind of graph branch lengths are not drawn proportionally to evolutionary distances so that only the topology of the tree matters. To visualize the ***phylogram*** (i.e. the tree with branch lengths drawn proportionally to the number of nucleotide substitutions per site along each lineage) click the last button to the right of the bar at the top of the window. The tree given in Fig. 5.6b will appear within the TREEVIEW window. This time, branch lengths are drawn proportionally to genetic distances and the

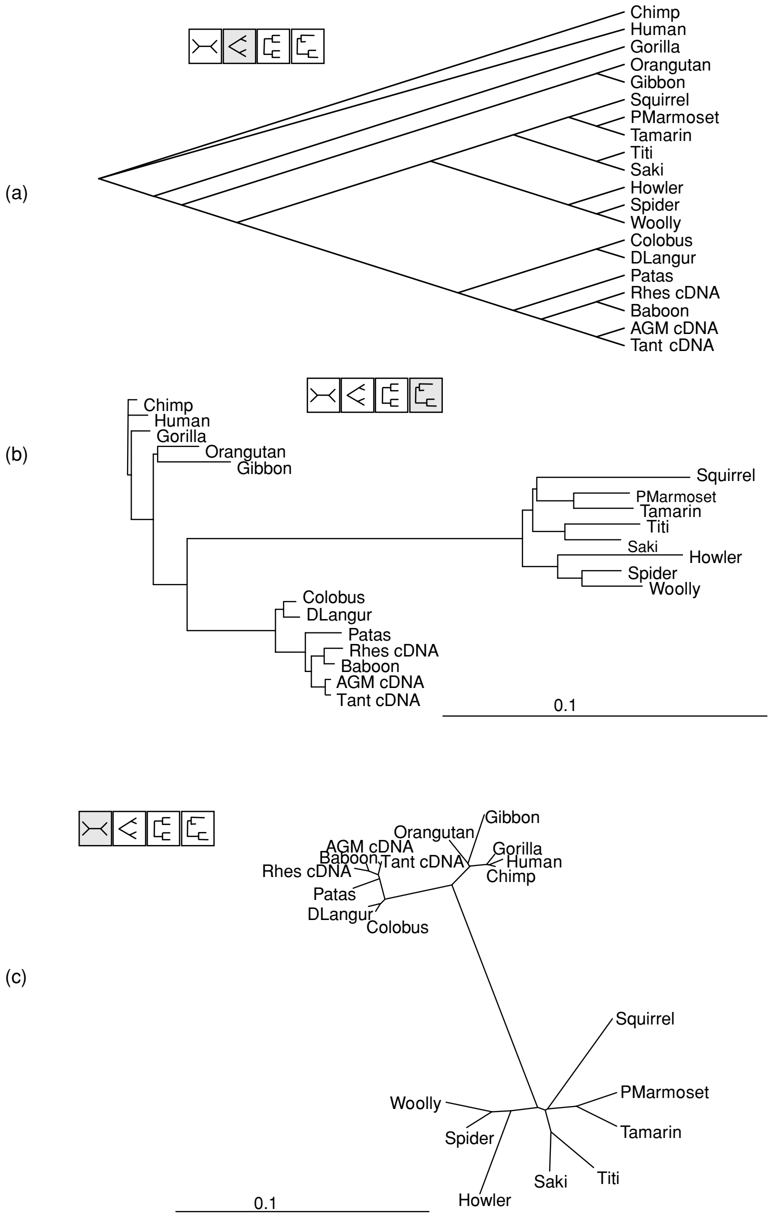
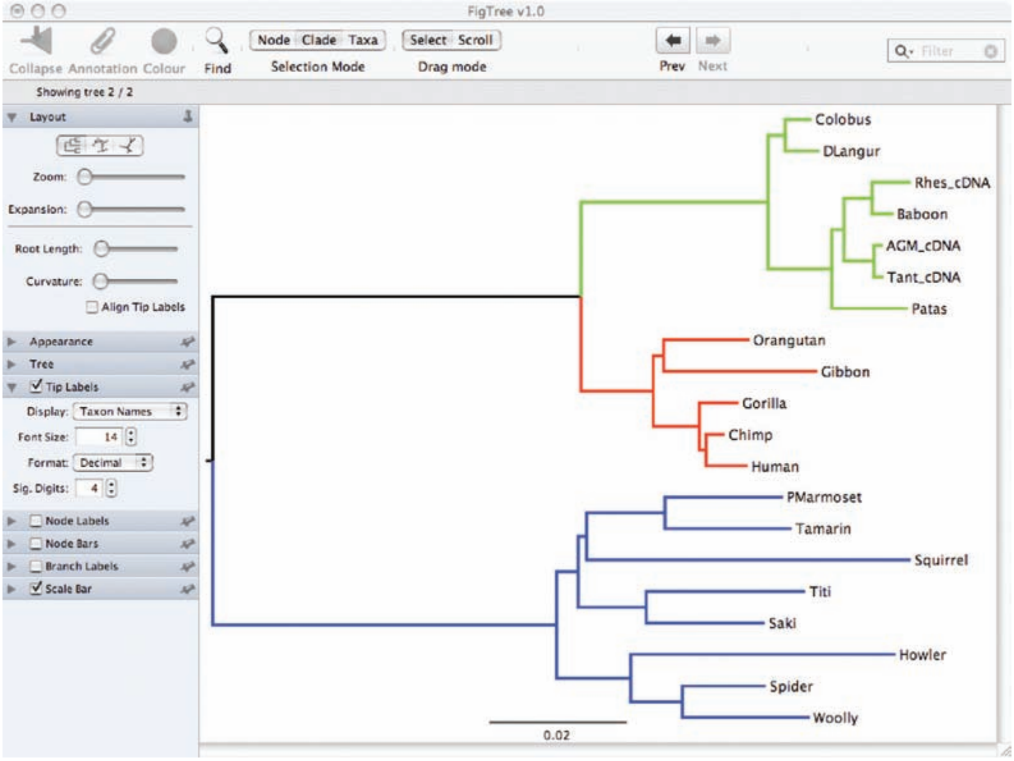


Fig. 5.6 Different TreeView displays for the Neighbor-joining tree of the Primates data set. Genetic distances are calculated with the F84 model with an empirical transition/transversion ratio of 2. (a) Slanted cladogram. (b) Rooted Phylogram. The scale on bottom represents genetic distances in substitutions per nucleotide. (c) Unrooted phylogram. The scale on top represents genetic distances in substitutions per nucleotide. Note that, even if the trees in panel (a) and (b) appear to be rooted, the NJ method actually infers unrooted tree topologies. Therefore, the position of the root is meaningless and the trees shown in A and B should be considered equivalent to the one given in panel (c) (see text for more details).

bar at the bottom of the tree indicates a length corresponding to 0.1 nucleotide substitutions per site.

It is important to keep in mind that a NJ tree is unrooted, and the fact that the Chimp and Human sequences appear to be the first to branch off does not mean that these are in fact the oldest lineages! The unrooted tree can be displayed by clicking on the first button to the left of the bar at the top of the `TreeView` window (see Fig. 5.6c). The unrooted phylogram clearly shows three main monophyletic clades (i.e. group of taxa sharing a common ancestor). The first one, at the bottom of the phylogram (Fig. 5.6c) includes DNA sequences from the so-called New World monkeys: woolly monkey (Woolly), spider monkey (Spider), Bolivian red howler (Howler), white-faced saki (Saki), Bolivian grey titi (Titi), tamarin, pygmy marmoset (PMarmoset), squirrel monkey (Squirrel). A second monophyletic clade on the top left of the phylogram includes sequences from Old World monkeys: African green monkey (AGM), baboon (Baboon), tantalus monkey (Tant), rhesus monkey (Rhes), patas monkey (Patas), dour langur (DLangur), kikuyu colobus (Colobus). The Old World monkeys clade appears to be more closely related to the third monophyletic clade shown on the top right of the tree in Fig. 5.6c, which includes DNA sequences from the Hominids group: Gibbon, Orangutan, Gorilla, Chimp, and Human. If we assume a roughly constant evolutionary rate among these three major clades, the branch lengths in Fig. 5.6b and c suggest that Hominids and Old World monkeys have diverged more recently and the root of the tree should be placed on the lineage leading to the New World monkeys. This observation is in agreement with the estimated divergence time between Old World monkeys and Hominids dating back to about 23 million years ago, and the estimated split between New World monkeys and Old World monkeys dating back to about 33 millions years ago (Sawyer *et al.*, 2005). This confirms that we can place the root of the tree on the branch connecting the New World monkeys and the Old World monkeys/Hominids clades. This will correspond to the root halfway between the two most divergent taxa in the tree. This rooting technique, called ***midpoint rooting*** is useful to display a tree with a meaningful evolutionary direction when an outgroup (see Chapter 1 and below for more details) is not available in the data set under investigation. To obtain the midpoint rooted tree, open the `outtree` file with the program FIGTREE and choose the option `Midpoint Root` from the `Tree` menu. The program creates a midpoint rooted tree that can be displayed in the window by clicking the `Next` button on the top left of the `FigTree` window. To increase the taxa font size click on the little triangle icon on the `Tip Labels` bar (on the left of the window, see Fig. 5.7) and increase the `Font Size` to 12 by using the up and down buttons on the right of the `Font Size` display box. The FIGTREE application also implements other user-friendly tools for editing phylogenetic trees. For example, by selecting the `Appearance` bar on the right



Display: Taxon Names

Font Size: 14

Format: Decimal

Sig. Digits: 4

Fig. 5.7 Editing phylogenetic trees with FigTree. FigTree screenshot (MacosX) of the NJ phylogenetic tree for the primates data set shown in Fig. 5.6. The FigTree command panels on top and left allow various manipulation of the tree (see text). The tree was midpoint rooted and branches colored according to specific lineages: New World monkeys, Old World monkeys, and Hominids.

of the window we can increase the `Line Weight` (for example to 2 or 3), which results in thicker branches. The `Scale Bar` option allows increasing the `Font Size` and the `Line Weight` of the scale bar at the bottom of the tree. We can also display branch lengths by checking the box on the left of the `Branch Labels` bar. For publication purposes it is sometimes useful to display different colors for the branches of a tree. For example, we may want to color the Hominids clade in green, the Old World monkey one in red, and the New World monkey one in blue. Select the `Clade` button on the top of the FigTree window. Click on the internal branch leading to the Hominids clades: that branch and all its descending branches within the clade will be highlighted. Click on the Color icon displayed on the top of the FigTree window and select the color red. Repeat the same procedure for the other two main clades choosing different colors. The edited tree can be saved in NEWICK format by selecting `Save as ...` from the `File` menu, which contains additional information such as the color of the branch lengths in a FIGTREE block. The tree can be re-opened later, but only `FigTree` is capable of displaying the editing information in the FIGTREE block. The tree can also be exported as PDF file by selecting `Export PDF...` from the `File` menu. An example of the primates tree midpoint rooted and edited with FIGTREE is displayed in Fig. 5.7.

5.7.1 Outgroup rooting

Midpoint rooting only works if we have access to independent information (like in the case discussed above), and/or when we can safely assume that the evolutionary rates along different branches of the tree are not dramatically different. When rates are dramatically different, a long branch may represent faster accumulation of mutations rather than an older lineage, and placing the root at the midpoint of the tree may be misleading. Chapter 11 will discuss in detail the molecular clock (constancy of evolutionary rates) hypothesis and the so-called ***local clock*** and ***relaxed clock*** models that can be used to investigate the presence of different evolutionary rates along different branches of a tree.

An alternative rooting technique consists of including an outgroup in the data set and placing the root at the midpoint of the branch that connects the outgroup with the rest (ingroup) of the taxa. As an example, we will use the `mtDNA.phy` data set including mitochondrial DNA sequences from birds, reptiles, several mammals, and three sequences from lungfish. After obtaining the NJ tree with F84-corrected distances as above, open the tree in TREEVIEW. The unrooted phylogram is shown in Fig. 5.8a. It can be seen that birds and crocodiles share a common ancestor, they are called ***sister taxa***. A monophyletic clade can also be distinguished for all the mammalian taxa (platypus, opossum, mouse, rat, human, cow, whale, seal). From systematic studies based on morphological characters it is known that lungfish belongs to a clearly distinct phylogenetic lineage with respect to the amniote vertebrates (such as reptiles, birds, and mammals). Thus, the three lungfish sequences (LngfishAf, LngfishSA, LngfishAu) can be chosen as outgroups and the tree can be rooted to indicate the evolutionary direction.

(i) choose `Define outgroup...` from the `Tree` menu and select the three lungfish sequences (LngFishAu, LngFishSA, LngFishAf)
(ii) choose `Root with outgroup...` from the `Tree` menu

The rooted phylogram displayed by TREEVIEW is shown in Fig. 5.8b.

Choosing an outgroup in order to estimate the root of a tree can be a tricky task. The chosen outgroup must belong to a clearly distinct lineage with respect to the ingroup sequences, i.e. the sequences in the data set under investigation, but it does not have to be so divergent that it cannot be aligned unambiguously

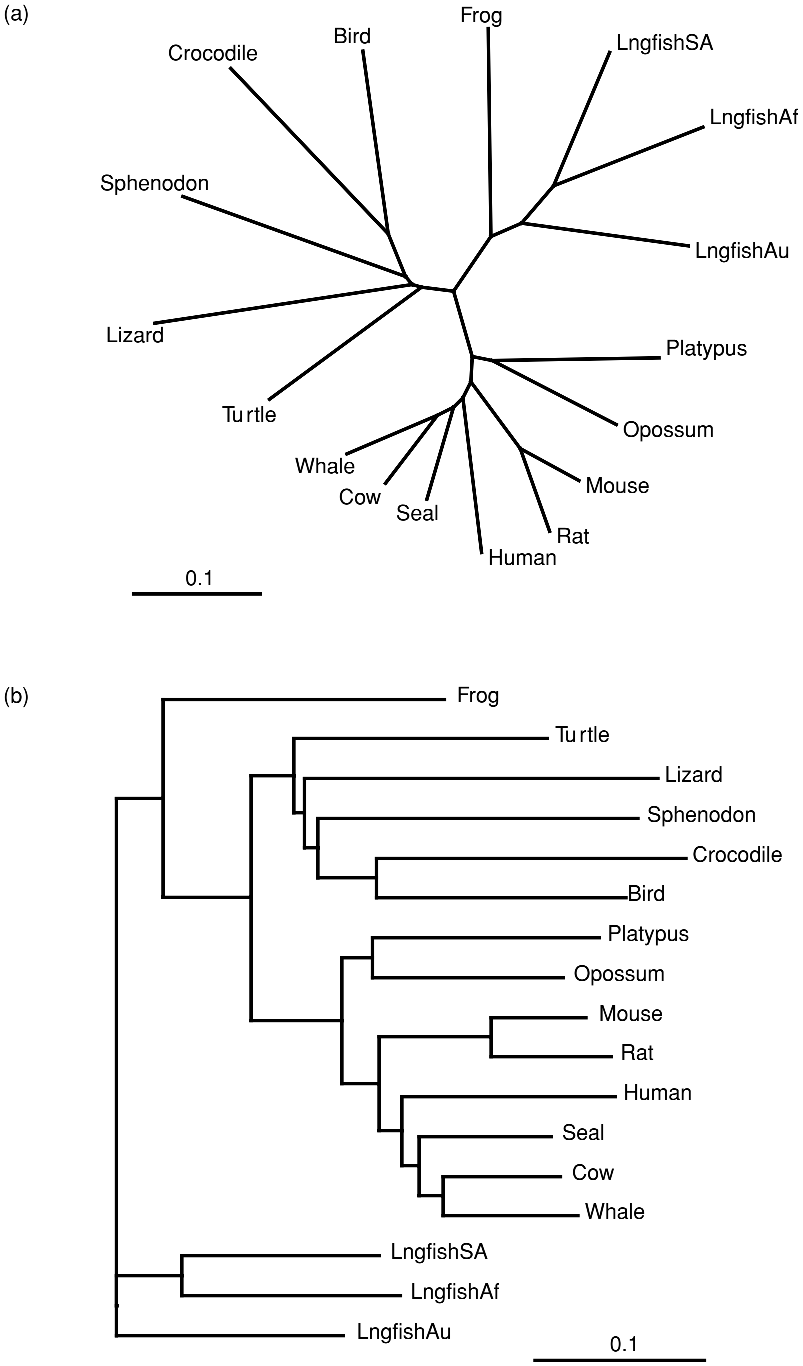


Fig. 5.8 Neighbor-joining tree of the mtDNA data set. Genetic distances were calculated with the F84 model and a transition/transversion ratio of 2. The scale at the bottom represents genetic distances in nucleotide substitutions per site. (a) Unrooted phylogram. (b) Rooted phylogram using the Lungfish sequences as outgroup.

against them. Therefore, before aligning outgroup and ingroup sequences and performing phylogenetic analyses, it can be useful to evaluate the similarity between the potential outgroup and some of the ingroup sequences using Dot Plots (see Section 3.10 in Chapter 3). If the Dot Plot does not show a clear diagonal, it is better to choose a different outgroup or to align outgroup and ingroup sequences only in the genome region where a clear diagonal is visible in the Dot Plot.

5.8 Inferring a *Fitch–Margoliash* tree for the mtDNA data set

The Fitch–Margoliash tree is calculated with the program Fitch.exe by employing the same distance matrix used for estimating the NJ tree in Section 5.7. The only option to be changed is option `G` (select `Yes`), which slows down a little the computation but increases the probability of finding a tree minimizing the difference between estimated pairwise distances and ***patristic distances*** (see Section 5.2.3). Again, the tree written to the `outtree` file can be displayed and edited with the TreeView or FigTree program. The phylogram rooted with DNA sequences from lungfish is shown in Fig. 5.9.

5.9 Bootstrap analysis using Phylip

The mtDNA data set discussed in this book was originally obtained to support the common origin of birds and crocodiles versus an alternative hypothesis proposing mammals as the lineage most closely related to birds (Hedges, 1994). Both the NJ and the Fitch–Margoliash tree show that birds and crocodiles cluster together. However, the two trees differ in clustering of turtles and lizards (compare Figs. 5.8 and 5.9). It is not unusual to obtain slightly different tree topologies using different tree-building algorithms. To evaluate the hypothesis of crocodile and bird monophyly appropriately, the reliability of the clustering in the phylogenetic trees estimated above must be assessed. As discussed in Section 5.3, one of the most widely used methods to evaluate the reliability of specific branches in a tree is bootstrap analysis. Bootstrap analysis can be carried out using Phylip as following:

(i) Run the program `Seqboot.exe` using as `infile` the aligned mtDNA sequences; the following menu will appear:

```
Bootstrapping algorithm, version 3.66

Settings for this run:
D      Sequence, Morph, Rest., Gene Freqs?   Molecular
                                             sequences
J Bootstrap, Jackknife, Permute, Rewrite?   Bootstrap
%   Regular or altered sampling fraction?   regular
```

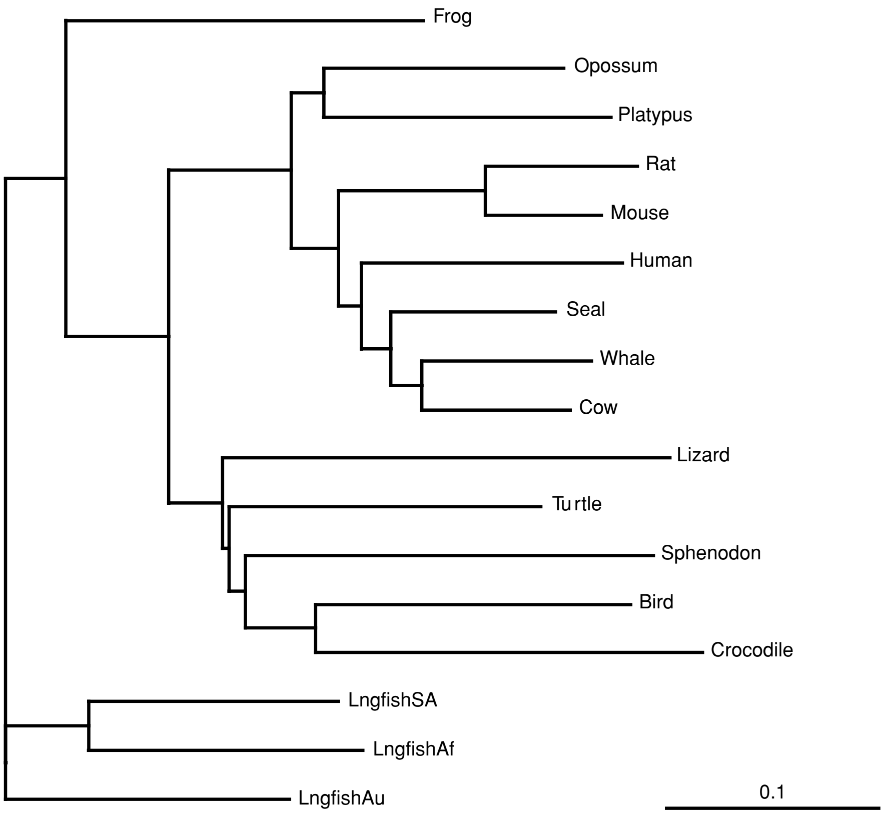


Fig. 5.9 Fitch–Margoliash tree of the mtDNA data set. Genetic distances were calculated with the F84 model and a transition/transversion ratio of 2. The scale at the bottom represents genetic distances in nucleotide substitutions per site. The phylogram was rooted using the Lungfish sequences as outgroup.

```
B     Block size for block-bootstrapping?   1 (regular
                                            bootstrap)
R                      How many replicates?   100
W             Read weights of characters?   No
C               Read categories of sites?   No
S    Write out data sets or just weights?   Data sets
I           Input sequences interleaved?   Yes
0     Terminal type (IBM PC, ANSI, none)?   IBM PC
1      Print out the data at start of run   No
2    Print indications of progress of run   Yes
Y to accept these or type the letter for one to change
```

Option J selects the kind of analysis (bootstrap by default). Option R determines the number of replicates. Type R and enter 1000. When entering Y, the program asks for a random number seed. The number should be odd and it is used to feed a random number generator, which is required to generate replicates of the data set with random resampling of the alignment columns (see Section 5.3). Type 5 and the program will generate the replicates and write them to the `outfile`:

(ii) Rename the `outfile` to `infile` and run `DNAdist.exe`. After selecting option `M` the following option should appear in the menu:

```
Multiple data sets or multiple weights? (type D or W)
```

Type `D` followed by the `enter` key, then type 1000 and press the `enter` key again. Type `Y` to accept the options followed by the `enter` key. `DNAdist.exe` will compute 1000 distance matrices from the 1000 replicates in the original mtDNA alignment and write them to the `outfile`.

(iii) Rename the `outfile` to `infile`. It is now possible to calculate, using the new `infile`, 1000 NJ or Fitch–Margoliash trees with `Neighbor.exe` or `Fitch.exe` by selecting option `M` from their menus. As usual, the trees will be written in the `outtree` file. Since the computation of 1000 Fitch–Margoliash trees can be very slow, especially if the option G is selected, enter 200 in option `M` so that only the first 200 replicates in the `infile` will be analyzed by `Fitch.exe` (keep in mind, however, that for publication purposes 1000 replicates is more appropriate). Enter again a random number and type `Y`.

(iv) The collection of trees in the `outtree` from the previous step is the input data to be used with the program `Consense.exe`. Rename `outtree` to `intree`, run `Consense.exe`, and enter `Y`. Note that there is no indication of the progress of the calculation given by `Consense.exe`, which depending on the size of the data set can take a few seconds to a few minutes. The consensus tree (see Section 5.3), written in the `outtree` in the usual NEWICK format (Fig. 5.5), can be viewed with TREEVIEW. Detailed information about the bootstrap analysis are also contained in the `outfile`.

The bootstrap values can be viewed in TREEVIEW by selecting `Show internal edges labels` from the `Tree` menu. Figure 5.10a shows the NJ bootstrap consensus tree (using 1000 bootstrap replicates) as it would be displayed by TREEVIEW. In 998 out of our 1000 replicates, birds and crocodile cluster together (bootstrap value 99.8%). A slightly different value (e.g. 99.4%) could have been obtained when a different set of 1000 bootstrap replicates were analyzed, e.g. by feeding a different number to the random number generator. This is an excellent support that strengthens our confidence in the monophyletic origin of these two species. The difference in clustering between the NJ and Fitch–Margoliash tree (Lizard vs. Turtle monophyletic with Crocodile–Bird–Sphenodon) is only poorly supported. A similar weak support would be obtained for the Fitch–Margoliash bootstrap analysis indicating that, in both cases, there is considerable uncertainty about the evolutionary relationships of these taxa. Note that the program by default draws rooted trees using an arbitrarily chosen outgroup and that branch lengths in this consensus tree, represented in cladogram style, are meaningless. It is important

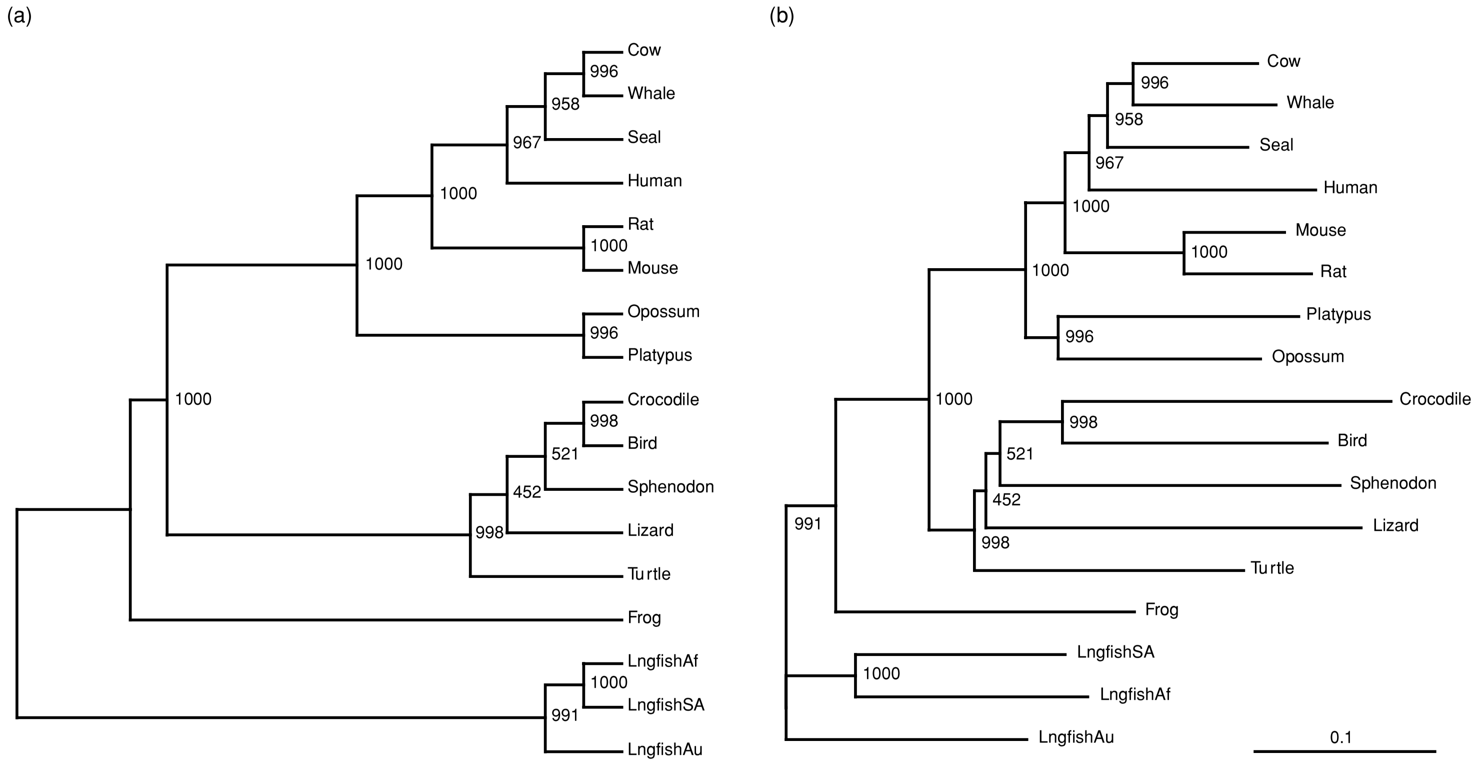


Fig. 5.10 (a) Neighbor-joining consensus tree for 1000 bootstrap replicates of the mtDNA data set as displayed in TreeView. (b) Inferred neighbor-joining tree for the mtDNA data set with bootstrap values. In both cases, the bootstrap values are shown to the right of the node representing the most recent common ancestor of the clade they support.

to distinguish between the consensus tree and the "real" tree constructed using the NJ method. The two trees do not necessarily have the same topology (see Box 5.4). The bootstrap consensus tree is inherently more unresolved since it is a consensus of different clustering hypothesis. When the trees are different, the topology of the consensus tree has to be disregarded since it is not based on the real data set but on bootstrap re-samplings of the original data. For publication the bootstrap values are usually displayed on a tree inferred from the original alignment, which usually includes branch lengths proportional to genetic distances. For example, Fig. 5.10b shows the NJ tree of the mtDNA data set (the same from Fig. 5.8b) where the bootstrap values from the consensus tree in Fig. 5.10a have been added. In this case, the topologies are the same because all the clusters in the inferred tree are also in the bootstrap consensus tree. The tree from TREEVIEW can be exported as a graphic file by selecting `Save as graphic ...` from the `File` menu. This file can then be imported in most graphics programs and bootstrap values can be added on the branches like in Fig. 5.10b. The consensus tree can also be viewed and edited using FIGTREE.

5.10 Impact of genetic distances on tree topology: an example using MEGA4

In the previous chapter we have introduced several nucleotide substitution models based on the ***Markov process*** and we have shown how genetic distance estimates are affected by the underlying assumptions of each model. Since pairwise genetic distances are used in turn by distance-based algorithms to infer topology and branch lengths of phylogenetic trees, it is not surprising that distances estimated according to different models might produce different trees.

As pointed out at the beginning of this chapter, it can be demonstrated that the NJ method is able to infer the true phylogenetic history (the true tree) of a set of taxa when the estimated genetic distances represent, in fact, the true genetic distances, i.e. the actual number of nucleotide substitutions per site between each pair of lineages. Such a condition is rarely met by real data sets. As we have seen in the previous chapter, Markov models, from the simple Jukes and Cantor to the complex General Time Reversible, are at best a reasonable approximation of the evolutionary process and more realistic models can be implemented at the cost of longer computation time and larger variance of the estimates. Even assuming that the selected substitution model accurately represents the evolutionary process for a given set of taxa, random fluctuations of the Poisson process will affect our estimates to some degree. Therefore, it is crucial not only to choose an appropriate model of nucleotide substitution that is complex enough to describe the underlying evolutionary process without increasing excessively the variance of the estimates,

Box 5.4 Bootstrap analysis with Phylip `Consense.exe outfile`

Part of the output file produced by the program `Consense.exe` for the 1000 bootstrap replicates of the ***mtDNA*** data set discussed in Section 5.6 is shown below.

```
Majority-rule and strict consensus tree program, version 3.573c

Species in order:

Opossum
Platypus
Rat
Mouse
Whale
Cow
Seal
Human
Lizard
Bird
Crocodile
Turtle
Sphenodon
Frog
LngfishSA
LngfishAf
LngfishAu

Sets included in the consensus tree

Set (species in order)    How many times out of 1000.00

..........  ...****        1000.00
..**......  .......        1000.00
........**  *******        1000.00
....**....  .......         999.00
..........  ....**.         999.00
........**  ***....         999.00
..******..  .......         998.00
..........  ....***         997.00
.........*  *......         994.00
....***...  .......         903.00
..********  *******         879.00
....****..  .......         757.00
........**  *.*....         669.00
.........*  *.*....         666.00
```

Box 5.4 (*cont.*)

```
Sets NOT included in consensus tree:

Set (species in order)     How many times out of 1000.00
.........*   ***....        245.00
..*****...   .......        170.00
........**   *......        159.00
........*.   ..*....        112.00
.*******..   .......         88.00
.........*   **.....         77.00
....**.*..   .......         75.00
..**...*..   .......         66.00
........**   **.....         46.00
.*......**   *******         33.00
........*.   .*.....         24.00
......**..   .......         12.00
..**..*...   .......          9.00
..........   *.*....          5.00
..........   ...*..*          3.00
..****....   .......          3.00
..****.*..   .......          3.00
..........   .**....          2.00
..**..**..   .......          2.00
........*.   *.....&          1.00
....******   ****"**          1.00
.......***   *******          1.00
.....**...   .......          1.00
.........*   *******          1.00
..........   .....**          1.00
```

The `outfile` first lists the *taxa* included in the tree (`Species in order`). The second section of the file, `Sets included in the consensus tree`, lists the clades which are present in more than 50% of the *bootstrap replicates* and are therefore included in the *consensus tree*. The clades are indicated as follows: each "`.`" represents a *taxon* in the same order as it appears in the list and *taxa* belonging to the same clade are represented by "`*`". For example, "`..** ............. 1000.00`" means that the clade joining rat and mouse (the third and fourth species in the list) is present in all 1000 trees estimated from the *bootstrap replicates*. It can happen that a particular clade present in the original tree is not included in the *consensus tree* because a different topology with other clades was better supported by the bootstrap test. In this case, the *bootstrap value* of that particular clade can be found in the third section, `Sets NOT included in consensus tree`.

but also to realize that, except in a few cases, a phylogenetic tree is at best an educated guess that needs to be carefully evaluated and interpreted.

Chapter 10 will describe a general statistical framework that can be used to select the best fitting nucleotide substitution model for any given data set. In what follows we will show, using the primates data set as an example, how different evolutionary assumptions can affect the tree estimated with a tree-based algorithm like NJ, and can occasionally lead to dramatically "wrong" tree topologies. We will perform tree inference using the MEGA4 program and the file `primates.meg`. The file contains aligned DNA sequences in mega format (see Chapter 4) and can be downloaded from *www.thephylogenetichandbook.org*.

Using MEGA4 it is possible to estimate a NJ tree and perform the bootstrap test in an automated fashion. The program will display the tree in a new window and superimpose bootstrap support values along each branch of the tree. To estimate a NJ tree using Kimura-2P (K2P) corrected distances and perform bootstrap analysis on 1000 replicates, open the `primates.meg` file in MEGA4 and select the submenu `Bootstrap Test of Phylogeny > Neighbor-Joining...` from the `Phylogeny` menu in the MEGA4 main window. The `Analysis Preferences` window will appear. Click on the green square to the right of the `Gaps/missing data` row and select `pairwise deletion` (specifying that for each pair of sequences only gaps in the two sequences being compared should be ignored). Similarly, select in the `Model` row `Nucleotide > Kimura-2-parameter`. To set the number of bootstrap replicates, click on the `Test of Phylogeny` tab on the top of the window and enter 1000 in the `Replications` cell. Select again the `Option Summary` tab and click on the `compute` button at the bottom of the window. After a few seconds (or a few minutes, depending on the speed of your computer processor) the NJ tree with bootstrap values will appear in the `Tree Explorer` window. By default, the tree is midpoint rooted and should look like the one given in Fig. 5.11a. If the location of the root needs to be placed on any other branch of the tree, this can be done by selecting the top button on the left side of the window (the button is indicated by an icon representing a phylogenetic tree with a green triangle on its left), placing the mouse on the branch chosen as the new root and clicking on it: a re-rooted tree will be displayed in the same window. To go back to the midpoint-rooted tree, simply select `Root on Midpoint` from the `View` menu.

In a similar way, NJ trees can be obtained using different nucleotide substitution models, with and without Γ-distributed rates across sites and with a different shape parameter α (see previous chapter) by selecting the appropriate options in the `Analysis Preferences` window. Figure 5.11a, b, c, d show the NJ trees (with 1000 bootstrap replicates each) obtained with the K2P model without Γ-distribution (5.11a), and with Γ-distribution using different α values: $\alpha = 0.5$

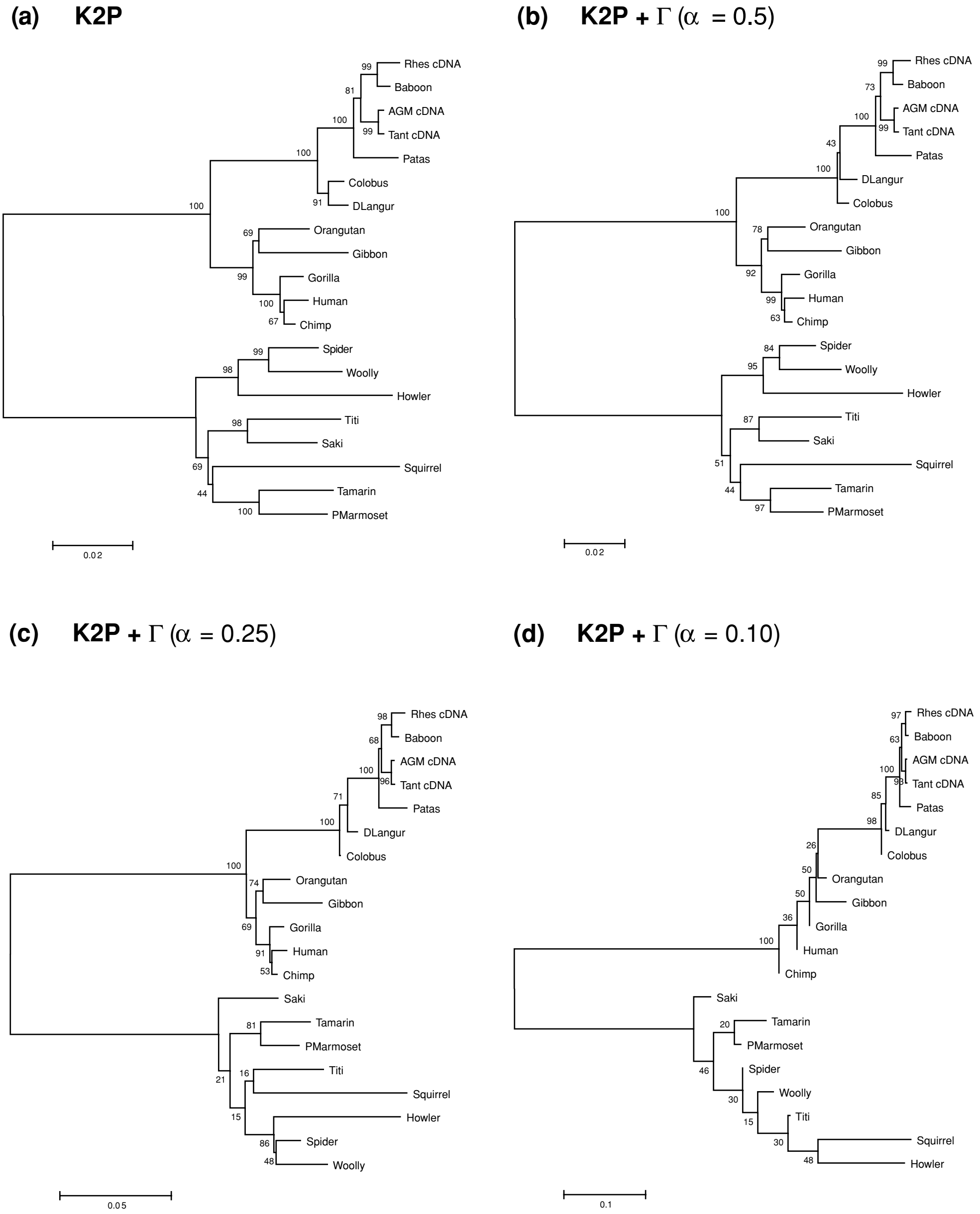


Fig. 5.11 Neighbor-joining trees of the Primates data set inferred by using different nucleotide substitution models. The scale at the bottom measures genetic distances in nucleotide substitutions per site. All phylograms were midpoint rooted and bootstrap values are now shown to the left of the node representing the most recent common ancestor of the clade they support. (a) NJ tree with Kimura 2-parameter estimated distances. (b) NJ tree with Kimura 2-parameter estimated distances using Γ-distributed rates along sites with α-parameter of 0.5. (c) NJ tree with Kimura 2-parameter estimated distances using Γ-distributed rates along sites with α-parameter of 0.25. (d) NJ tree with Kimura 2-parameter estimated distances using Γ-distributed rates along sites with α-parameter $= 0.1$. (e) NJ tree with Tamura–Nei estimated distances. (f) NJ tree with Tamura–Nei estimated distances using Γ-distributed rates along sites with α-parameter of 0.5. (g) NJ tree with Tamura–Nei estimated distances using Γ-distributed rates along sites with α-parameter of 0.25. (h) NJ tree with Tamura–Nei estimated distances using Γ-distributed rates along sites with α-parameter of 0.1.

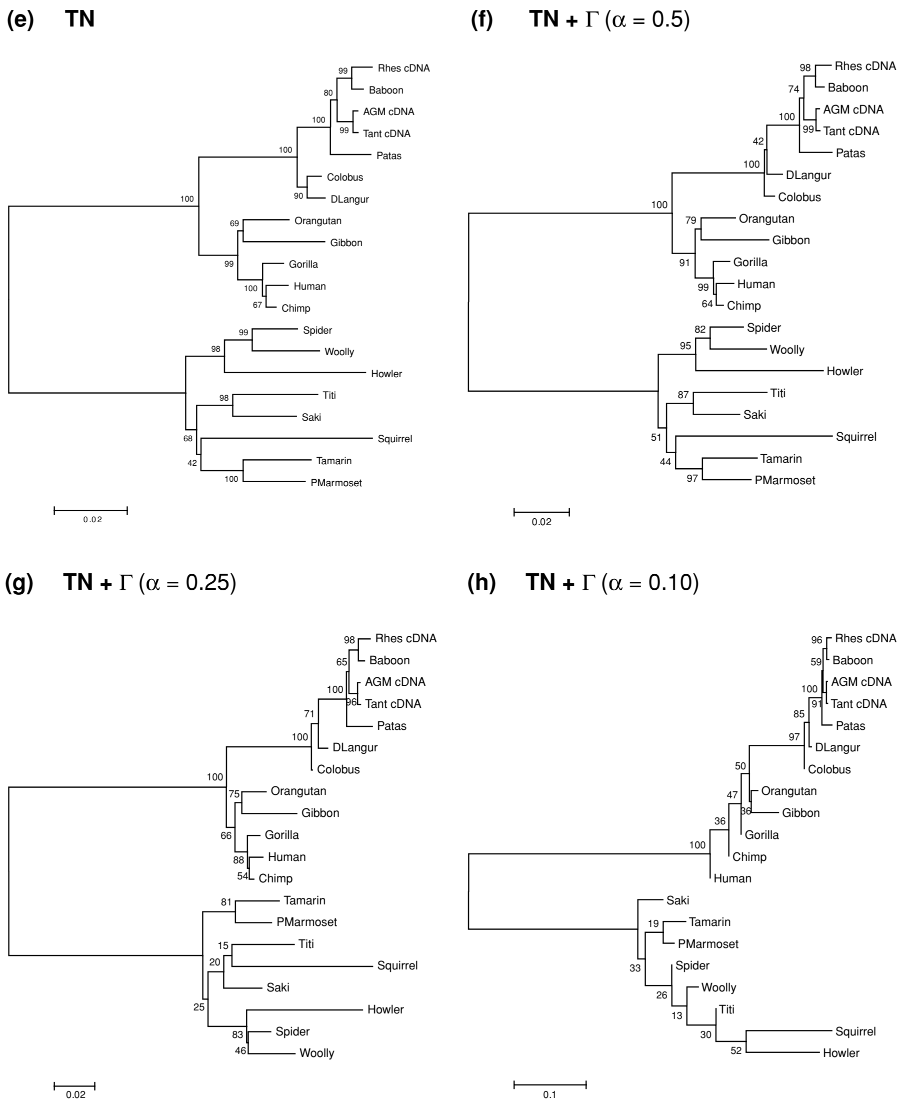


Fig. 5.11 (*cont.*)

(Fig. 5.11b), $\alpha = 0.25$ (Fig. 5.11c), and $\alpha = 0.1$ (Fig. 5.11d). Figure 5.11e, f, g, h show trees obtained with the Tamura–Nei model (TN, see previous chapter) without (Fig. 5.11e), and with (Fig. 5.11f, g, h) different Γ-distributions. The K2P and the TN tree without Γ-distribution and with Γ-distribution but moderate rate heterogeneity ($\alpha = 0.5$) appear very similar except for the position of colobus and langur taxa that are monophyletic in the trees in Fig. 5.11a and 5.11e but ***paraphyletic*** (with the colobus sequence branching off in trees in Fig. 5.11b and Fig.

5.11f). The K2P and TN trees obtained using $\alpha = 0.25$ (strong rate heterogeneity) also show a discordant position for the white-faced saki sequence (Fig. 5.11c and Fig. 5.11g). The K2P and the TN trees with $\alpha = 0.1$ show again a similar (but not identical!) topology. However, such trees are in obvious contradiction to all the known data about Hominids evolutionary relationships. In fact, according to the K2P tree (Fig. 5.11d) the chimp lineage is the oldest lineage at the base of the Hominids clade, while according to the TN tree (Fig. 5.11h), it is the human lineage that branches off at the root of the Hominids clade. In general, a different α parameter of the Γ-distribution tends to have a greater impact on the tree topology than a different evolutionary model. However, the example shows (hopefully, in a convincing way) that the nucleotide substitution model and the parameters of the model (especially α) can significantly impact the phylogenetic reconstruction and, therefore, need to be chosen carefully. The following chapter and Chapter 10 will discuss how maximum likelihood methods can be used to estimate substitution model parameters and to test different evolutionary hypotheses within a rigorous statistical framework.

5.11 Other programs

Several other programs are freely available to compute phylogenetic trees with distance-based methods. Some of them, like PAUP*, will be discussed in the next chapters. A comprehensive list of tree visualization and editing software is provided at *http://bioinfo.unice.fr/biodiv/Tree_editors.html*. A rather complete list of the phylogeny software available is maintained by Joe Felsenstein at *http://evolution.genetics.washington.edu/PHYLIP/software.html* and most packages can be downloaded following the links of his web page. They always contain complete documentation on how to install and run them properly. However, the programmer usually assumes the user has a background in phylogeny and molecular evolution, and for those programs not discussed in this book it may be necessary to read the original papers in order to understand the details or the meaning of the implemented computation.