Chapter 2
Trees

2.1 Introduction to trees

All of life is related by common ancestry. Recovering this pattern, the 'Tree of Life', is one of the prime goals of evolutionary biology. This chapter introduces the fundamentals of trees. You may find it useful to read the chapter through once, then refer back to it as you read the rest of the book. All the concepts introduced in this chapter will be discussed in more detail in subsequent chapters; our goal here is to give you some familiarity with trees so that interpreting them eventually becomes second nature.

2.1.1 Tree terminology

Figure 2.1 illustrates the terminology used in this book to describe trees. Unfortunately tree terminology varies greatly among authors, and among different disciplines, such as mathematics and biology. Where possible we will list the commonly used synonyms that you may encounter in the literature.

A **tree** is a mathematical structure which is used to model the actual evolutionary history of a group of sequences or organisms. This actual pattern of historical relationships is the **phylogeny** or **evolutionary tree** which we try and estimate. A tree consists of **nodes** connected by **branches** (also called **edges**). **Terminal nodes** (also called **leaves**, **OTUs [Operational Taxonomic Units]**, or **terminal taxa**) represent sequences or organisms for which we have data; they may be either extant or extinct. **Internal nodes** represent hypothetical ancestors; the ancestor of all the sequences that comprise the tree is the **root** of the tree (see below).

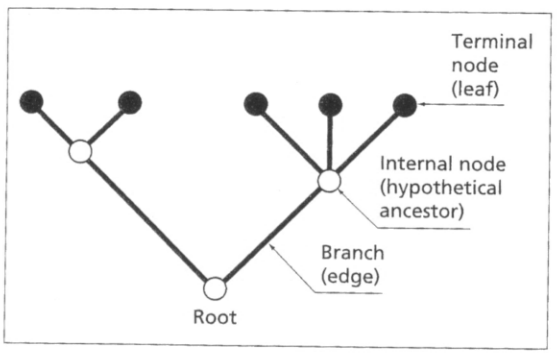
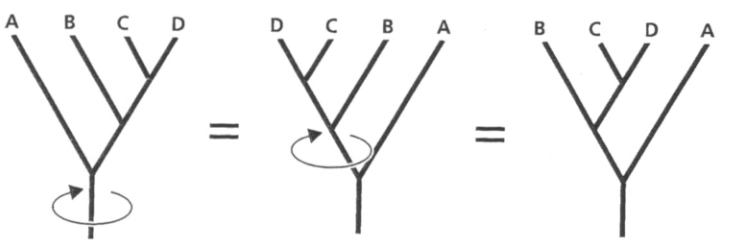


Fig. 2.1 A simple tree and associated terms.

The nodes and branches of a tree may have various kinds of information associated with them. For example some methods of phylogeny reconstruction (e.g. parsimony) endeavour to reconstruct the characters of each hypothetical ancestor; most methods also estimate the amount of evolution that takes place between each node on the tree, which can be represented as **branch lengths** (or **edge lengths**). Trees with branch lengths are sometimes called **weighted trees**.

Box 2.1 Trees are like mobiles

There are many different ways of drawing trees, so it is important to know whether these different ways actually reflect differences in the kind of tree, or whether they are simply stylistic conventions. For instance, the order in which the labels on a tree are drawn on a piece of paper (or computer screen) can differ without changing the meaning of the tree. This is because the edges of a tree can be freely rotated without changing the relationships among the terminal nodes. The diagram below shows the same tree drawn three different ways:



In this sense a tree is just like a mobile; no matter how many times you rotate the 'hanging' objects you do not change how they are connected to one another.

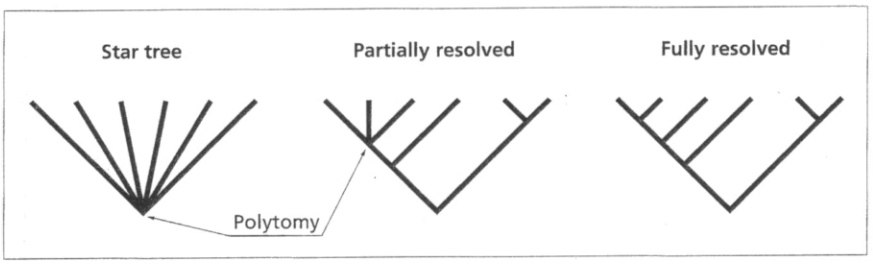


Fig. 2.2 Three trees showing various degrees of resolution, ranging from a complete lack of resolution (star tree) to a fully resolved tree. Any internal node with more than two immediate descendants is a polytomy.

The number of adjacent branches possessed by an internal node is that node's **degree**. If a node has a degree greater than three (i.e. it has one ancestor and more than two immediate descendants) then that node is a **polytomy**. A tree that has no polytomies is fully resolved (Fig. 2.2).

Polytomies can represent two rather different situations (Fig. 2.3); firstly they may represent simultaneous divergence—all the descendants evolved at the same time (a 'hard' polytomy); alternatively, polytomies may indicate uncertainty about phylogenetic relationships—the lineages did not necessarily all diverge at once, but we are unsure as to the actual order of divergence (a 'soft' polytomy). These two interpretations—simultaneous divergence or uncertainty—are obviously quite different. Typically polytomies are treated as 'soft'. It may be thought unlikely that multiple lineages would diverge at exactly the same time; however, if lineages diverge rapidly in time relative to the rate of character evolution then there may be insufficient evidence available to us to ever be able to reconstruct the exact order of splitting, in which case the polytomy is effectively 'hard'.

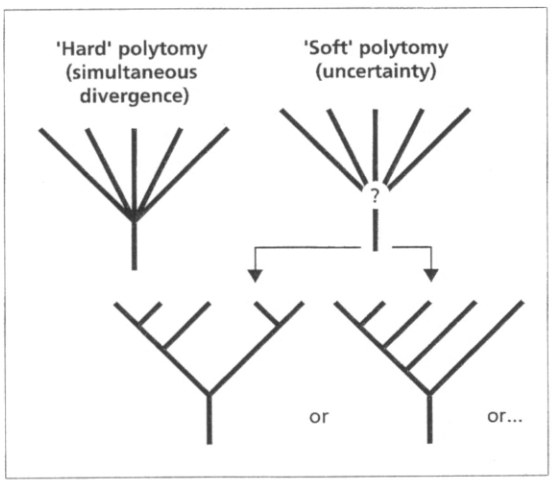


Fig. 2.3 Polytomies can represent either simultaneous divergence of multiple sequences ('hard'), or lack of resolution due to insufficient data or conflicting trees ('soft').

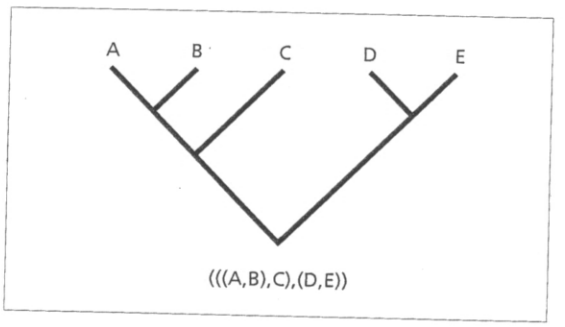


Fig. 2.4 A tree and its shorthand representation using nested parentheses.

2.1.2 A shorthand for trees

Trees can be represented by a shorthand notation that uses nested parentheses. Each internal node is represented by a pair of parentheses that enclose all descendants of that node. This format makes it easy to describe a tree in the body of some text without having to draw it. The format is also used by many computer programs to store representations of trees in data files. Figure 2.4 gives an example of this shorthand.

2.1.3 Cladograms, additive trees and ultrametric trees

Different kinds of tree can be used to depict different aspects of evolutionary

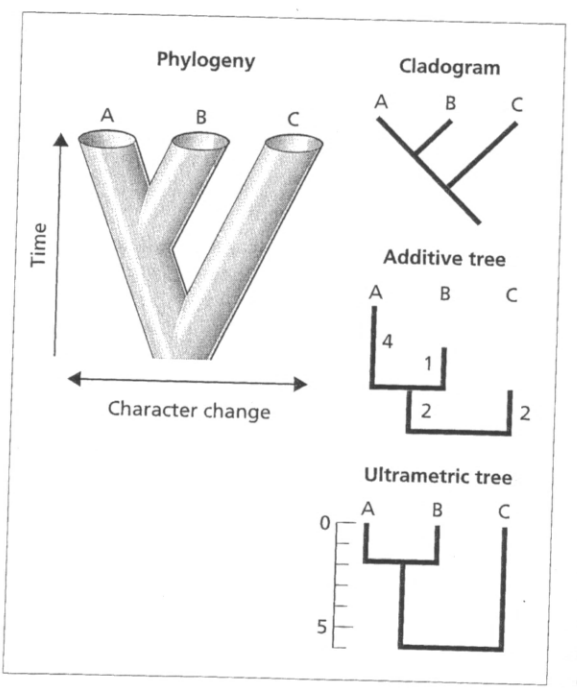


Fig. 2.5 A phylogeny and the three basic kinds of tree used to depict that phylogeny. The cladogram represents relative recency of common ancestry; the additive tree depicts the amount of evolutionary change that has occurred along the different branches, and the ultrametric tree depicts times of divergence.

history. The most basic tree is the **cladogram** which simply shows relative recency of common ancestry, that is, given the three sequences, A, B and C, the cladogram in Fig. 2.5 tells us that sequences A and B share a common ancestor more recently than either does with C. In the biomathematical literature cladograms are often called '*n*-trees'.

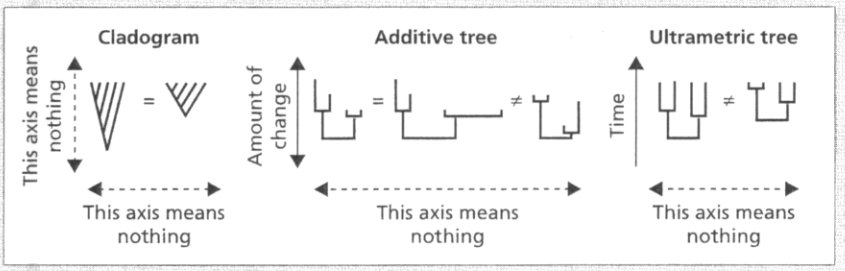
Additive trees contain additional information, namely branch lengths. These are numbers associated with each branch that correspond to some attribute of the sequences, such as amount of evolutionary change. In the example shown in Fig. 2.5, sequence A has acquired four substitutions since it shared a common ancestor with sequence B. Other commonly used terms for additive trees include 'metric trees' and 'phylograms'.

Ultrametric trees (sometimes also called 'dendrograms') are a special kind of additive tree in which the tips of the trees are all equidistant from the root of the tree. This kind of tree can be used to depict evolutionary time, expressed either directly as years or indirectly as amount of sequence divergence using a molecular clock.

Additive and ultrametric trees both contain all the information found in a cladogram—the cladogram is the simplest statement about evolutionary relationships that we can make. For some questions knowledge of relative recency of common ancestry is sufficient. However, there are other evolutionary questions (such as determining relative rates of evolution) which require the additional information contained in additive and ultrametric trees.

Box 2.2 What do the horizontal and vertical axes of a tree represent?

It is tempting to think of a tree as being a graphical plot like a scatter plot, in which case the question arises 'what do the horizontal and vertical axes represent?' For cladograms, which have no branch length information, neither axis has any special meaning; you can squash the tree flatter, or stretch it out without changing the relationships among the terminal nodes. Hence, the two cladograms shown below are the same.



continued on p. 16

Box 2.2 *continued*

For additive trees one of the axes does have meaning; it represents the amount of evolutionary change. In the diagram above if we stretch the tree along the horizontal axis (i.e. left to right) we do not change interpretation of the tree; however, changes in the vertical axis (up and down) change the amount of evolutionary change along the branches, hence the trees are not the same *additive* trees. Similarly, for an ultrametric tree one axis typically represents time whereas the other has no meaning. The two ultrametric trees shown above are different because the two trees specify different divergence times.

A last consideration is that trees can be drawn in a number of orientations, such as 'planted' with the root at the bottom as in the diagram above, 'left-to-right' as in Fig. 2.17, or even 'top-down' with the root at the top. The choice among these representations is entirely arbitrary; in some circumstances it may be more convenient to draw the tree one way rather than another. Just remember that if the tree diagram is rotated then the x- and y-axes in the above diagram need to be rotated as well. Hence, if an additive tree is drawn left to right then the horizontal axis represents evolutionary change and the vertical axis has no meaning.

2.1.4 Rooted and unrooted trees

Cladograms and additive trees can either be rooted or unrooted. A **rooted** tree has a node identified as the root from which ultimately all other nodes descend, hence a rooted tree has direction. This direction corresponds to evolutionary time; the closer a node is to the root of the tree, the older it is in time. Rooted trees allow us to define ancestor–descendant relationships between nodes: given a pair of nodes connected by a branch, the node closest to the root is the ancestor of the node further away from the root (the descendant). **Unrooted** trees lack a root, and hence do not specify evolutionary relationships in quite the same way, and they do not allow us to talk of ancestors and descendants. Furthermore, sequences that may be adjacent on an unrooted tree need not be evolutionarily closely related. For example, given the unrooted tree in Fig. 2.6, the gibbon (B) and orang-utan (O) sequences are neighbours on the tree, yet the orang-utan is more closely related to the other apes (including humans). This is because the root of the tree lies on the branch leading to the gibbon. Had we placed the root elsewhere, say on the branch leading to the gorilla (G), then the gibbon and orang-utan sequences would indeed be closely related.

In the unrooted tree for the apes shown in Fig. 2.6, we could have placed the root on any of the seven branches of the tree. Hence, this unrooted tree corresponds to a set of seven rooted trees (Fig. 2.7).

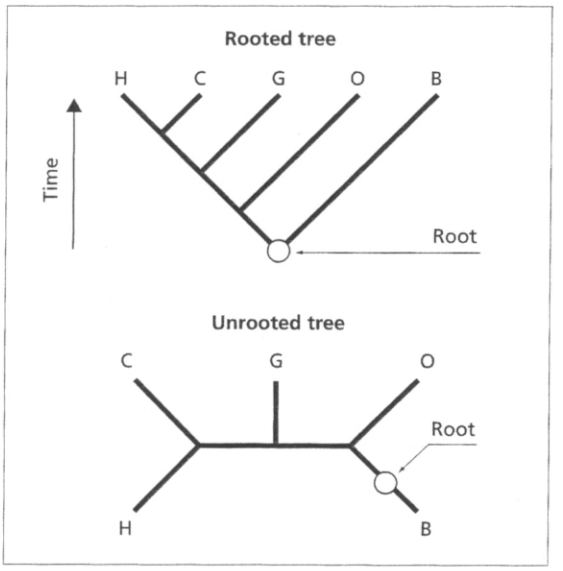


Fig. 2.6 Rooted and unrooted trees for human (H), chimp (C), gorilla (G), orang-utan (O), and gibbon (B). The rooted tree (top) corresponds to the unrooted tree below.

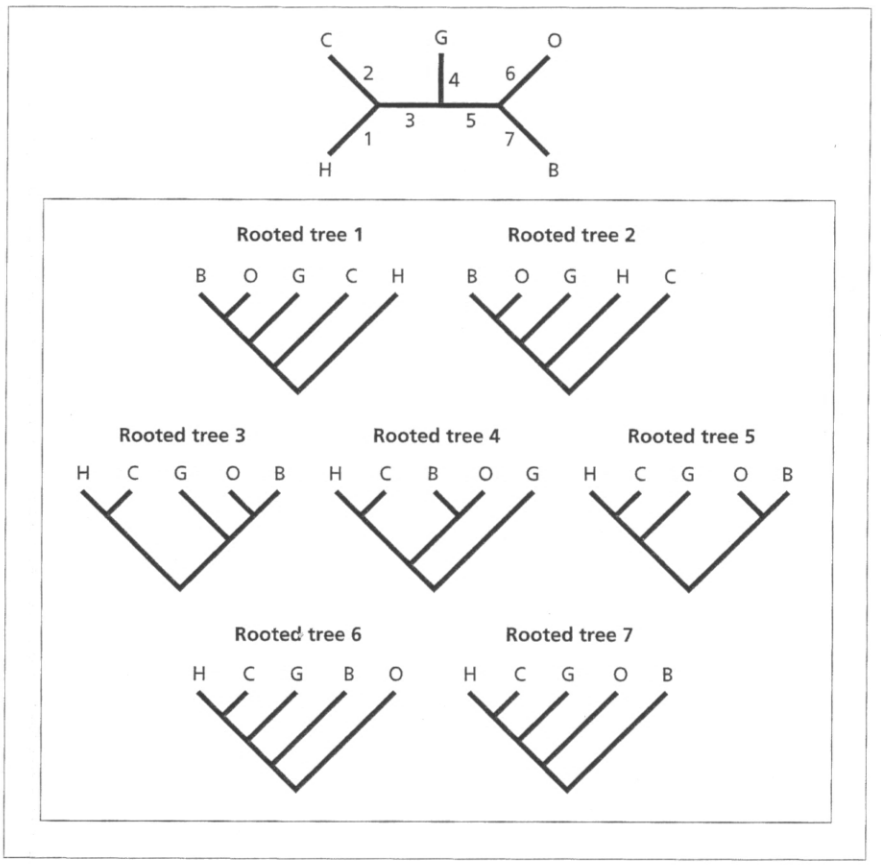


Fig. 2.7 The seven rooted trees that can be derived from an unrooted tree for five sequences. Each rooted tree 1–7 corresponds to placing the root on the corresponding numbered branch of the unrooted tree. (Sequence labels as for Fig. 2.6.)

The distinction between rooted and unrooted trees is important because many methods for reconstructing phylogenies generate unrooted trees, and hence cannot distinguish among the seven trees shown in Fig. 2.7 on the basis of the data alone. In order to root an unrooted tree (i.e. decide which of the seven trees is the actual evolutionary tree) we need some other source of information. Methods of rooting trees are discussed in Chapter 6 (note that this does not apply to ultrametric trees which are rooted by definition).

The number of possible unrooted trees U_n for n sequences is given by

$$U_n = (2n-5)(2n-7) \ldots (3)(1) \tag{2.1}$$

for $n \geq 2$. The number of rooted trees R_n for $n \geq 3$ is given by

$$\begin{aligned} R_n &= (2n-3)(2n-5) \ldots (3)(1) \\ &= (2n-3)U_n \end{aligned} \tag{2.2}$$

Table 2.1 lists the numbers of rooted and unrooted fully resolved trees for 2–10 sequences. Note that the number of unrooted trees for n sequences is equal to the number of rooted trees for $(n-1)$ sequences. Note also that the number of trees rapidly reaches very large numbers: for 10 sequences there are over 34 million possible rooted trees. For a relatively modest 20 sequences there are 8 200 794 532 637 891 559 000 possible trees, whereas the number of different trees for 135 human mitochondrial DNA sequences used in the study of the evolution of modern humans (see Chapter 4), 2.113×10^{267}, exceeds the number of particles in the known universe! This explosion in number of trees is a fundamental problem for phylogeny reconstruction, where the goal is to identify which tree of all the possible trees is the best estimate of the actual phylogeny.

2.1.5 Tree shape

Typically, the information in a tree in which we are most interested is the relationship among the sequences, and perhaps the lengths of the branches.

Number of sequences	Number of unrooted trees	Number of rooted trees
2	1	1
3	1	3
4	3	15
5	15	105
6	105	945
7	945	10 395
8	10 395	135 135
9	135 135	2 027 025
10	2 027 025	34 459 425

Table 2.1 Numbers of unrooted and rooted trees for 2–10 sequences.

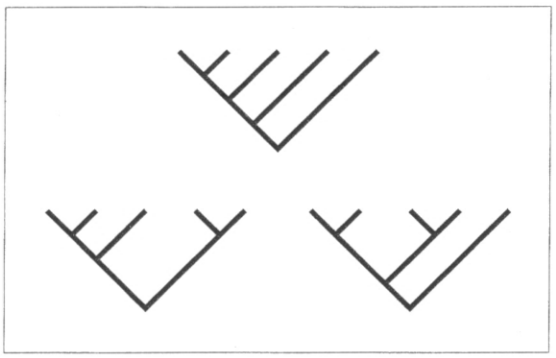


Fig. 2.8 The three possible shapes for a rooted tree for five sequences.

However, other aspects of the tree may also reflect evolutionary phenomena and hence be of interest. Figure 2.8 shows the three possible **shapes** (or **topologies**) for a rooted tree for five sequences. All 105 possible trees (Table 2.1) for five sequences will have one of these three shapes.

2.1.6 Splits

Trees can be represented in a variety of ways other than as graphs. One useful representation is as sets of sets, called **splits** or **partitions**. Each split takes the set of sequences (e.g. {H, C, G, O, B}) and partitions them into two mutually exclusive sets: you can think of a split as the two sets of sequences obtained by chopping ('splitting') the tree at a given branch. For example, the tree shown in Fig. 2.9 has seven branches and hence seven splits. However, all splits comprising a single terminal node on one hand and the rest of the tree on the other are not 'phylogenetically informative' in the sense that all possible trees will contain those splits. Hence, the only informative splits are those

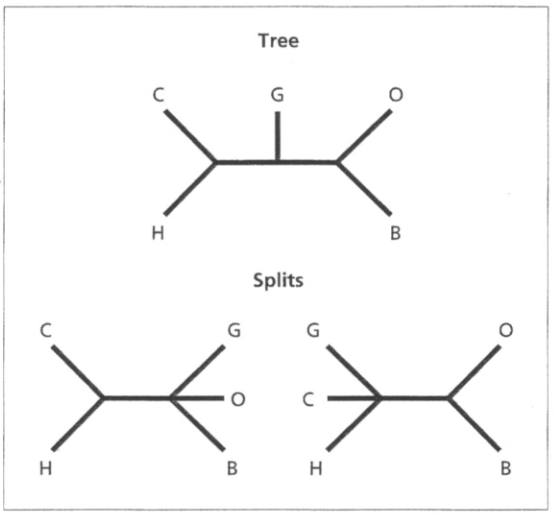


Fig. 2.9 An unrooted tree and its two splits.

resulting from chopping internal branches. The tree shown in Fig. 2.9 has two informative splits: {{C, H}, {G, O, B}} and {{G, C, H}, {O, B}}.

Given these two splits we can combine them to reconstruct the original tree. Notice that there are other possible partitions of the set {H, C, G, O, B}, such as {{H, G}, {C, O, B}}. This split groups humans and gorillas together to the exclusion of the other apes, which is **incompatible** with the split {{C, H}, {G, O, B}}, which groups humans and chimps. Incompatible splits cannot be combined to form a tree.

Another way of representing the splits in Fig. 2.9 is to assign arbitrary letters to each half of a split, such as the letter 'A' to each sequence on the left and the letter 'T' to each sequence on the right. This gives the following table:

Sequence	Split 1	Split 2
H	A	A
C	A	A
G	T	A
O	T	T
B	T	T

Each split now resembles a single nucleotide site with only the bases A and T. In Chapter 6 you will encounter some methods for reconstructing phylogenies that make use of this relationship between nucleotide sites and splits.

2.2 Reconstructing the history of character change

The tree relating a set of sequences tells us only part of what we want to know. The tree alone does not tell us when a particular evolutionary change, such as a nucleotide substitution, took place, or whether the occurrence of the same amino acid in two sequences is the result of inheritance from a common ancestor or independent evolution. To address these questions we need to be able to reconstruct the history of character change. This problem is addressed

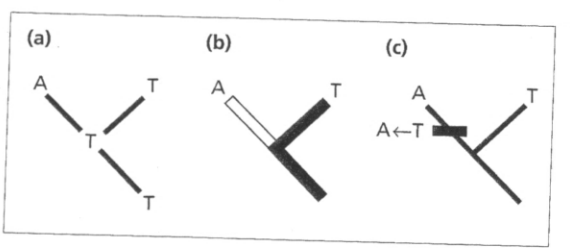


Fig. 2.10 Three equivalent ways of representing the same evolutionary change on the same tree. (a) Each node is labelled by the corresponding nucleotide; (b) each branch is coloured corresponding to the nucleotide at the end of each branch; and (c) indicating on which branch the change took place.

in more detail in Chapter 5, so here we will merely introduce some of the different ways of representing evolutionary change on a tree (Fig. 2.10) and describe some basic terminology.

Given a tree, we can distinguish between ancestral ('primitive') and derived character states. If a sequence has the same base as the common ancestor of all the sequences being studied then it is the primitive or **plesiomorphic** state; otherwise it is a derived or **apomorphic** state. Unique derived character states are **autapomorphies** (*aut* = alone), shared derived states are **synapomorphies** (*syn* = shared) (Fig. 2.11). Given any two character states that are identical (e.g. the same nucleotide base) the similarity between them may be because they have both inherited it directly from their ancestor which also had that state. This is an instance of **homology**. Alternatively, the similarity may have occurred independently in which case it is **homoplasy**. Only homologous similarity directly reflects common ancestry. Homoplasy is

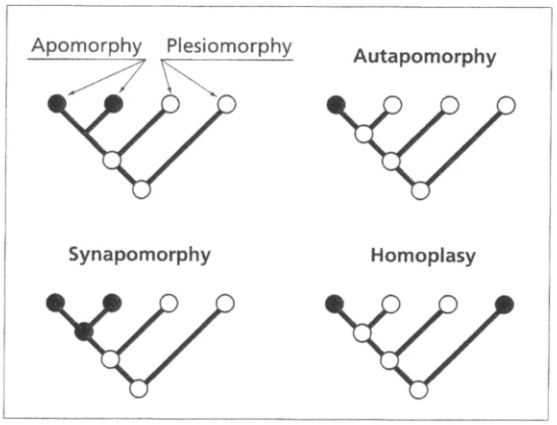


Fig. 2.11 Trees showing the terminology used to describe different patterns of ancestral (○) and derived (●) character states.

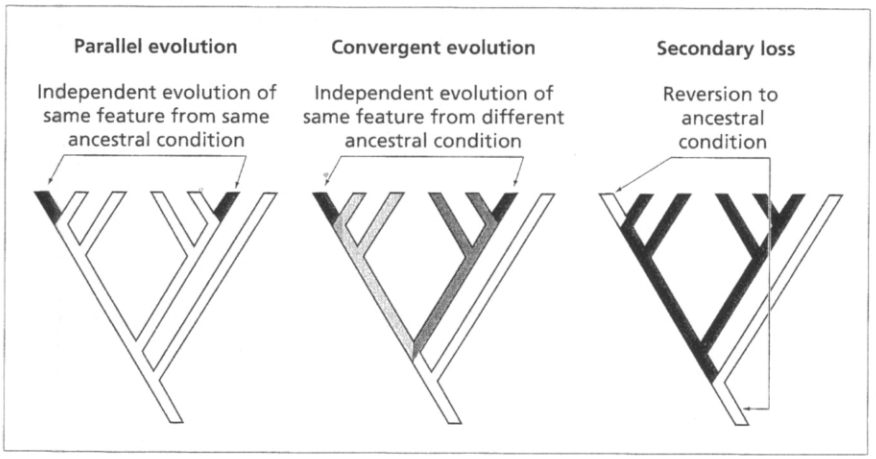


Fig. 2.12 Three different kinds of homoplasy.

a poor indicator of evolutionary relationships because the similarity does not reflect shared ancestry. It is sometimes useful to distinguish between different kinds of homoplasy (Fig. 2.12). **Convergence** and **parallel evolution** both result in the independent evolution of the same feature in two unrelated sequences; the difference between the two lies in whether the similarity was acquired from the same (parallelism) or a different (convergence) ancestral condition. Homoplasy may also be due to the **secondary loss** of a derived feature, which results in the apparent reversion to the ancestral condition, such as the loss of legs in snakes and some amphibia.

2.2.1 Ancestors

Phylogenies presuppose ancestors—previously living organisms that are now extinct but which left descendants which comprise modern species. These ancestors (or their sequences) are represented by the internal nodes of a tree. These ancestors are hypothetical, but some methods of phylogenetic reconstruction allow us to infer what they (or their sequences) may have looked like.

All molecular phylogenies include ancestors, but for the most part these remain hypothetical entities represented by the internal nodes of the tree, and inferred solely on the basis of sequences from extant organisms. It used to be thought that the possibility that a sequence being studied was actually an ancestor could be safely ignored, hence all sequences were placed at the tips of evolutionary trees. However, two recent developments have meant that molecular biologists must deal with a problem previously restricted to palaeontology—namely the recognition of ancestors. The first of these developments is the recovery of DNA from extinct taxa; the second is the increasing number of sequences being obtained from viruses such as human immunodeficiency virus (HIV) which evolve sufficiently fast to be tracked in 'real time'.

If all sequences are from extant organisms, then they can be placed at the tips of the tree. However, if some of the sequences are extinct it is possible, if unlikely, that they may have been ancestral to one or more of the extant sequences: is a sequence extracted from an extinct taxon an ancestor to modern taxa or is it on an evolutionary side branch? Cladists have adopted the convention that extinct taxa that lack autapomorphies are candidates for being ancestral, as it is equally parsimonious to treat them as **sister taxa** (i.e. each other's closest relative) or as ancestors (Fig. 2.13). Treating a taxon with autapomorphies as an ancestor would require us to postulate additional evolutionary changes (this invokes the parsimony principle discussed in Chapter 6). Note that under this rule a taxon with no autapomorphies need not be an ancestor, rather there is nothing to refute that possibility.

We can apply the cladistic convention to viral sequences where the virus is evolving sufficiently rapidly for successive samples to show evolutionary change. For example, Fig. 2.14 shows a cladogram for eight HIV sequences

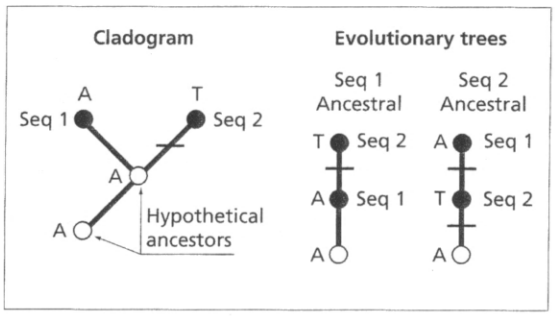


Fig. 2.13 A cladogram for two sequences (Seq 1 and Seq 2) showing the nucleotide at a single site, and two of several possible evolutionary trees derived from that cladogram. We could postulate that either sequence is ancestral to the other. However, postulating Seq 2 to be an ancestor of Seq 1 requires the gain and subsequent loss of T, whereas if Seq 1 is an ancestor no additional substitutions need be postulated. Note that a third phylogeny would be identical to the cladogram (see Box 2.3).

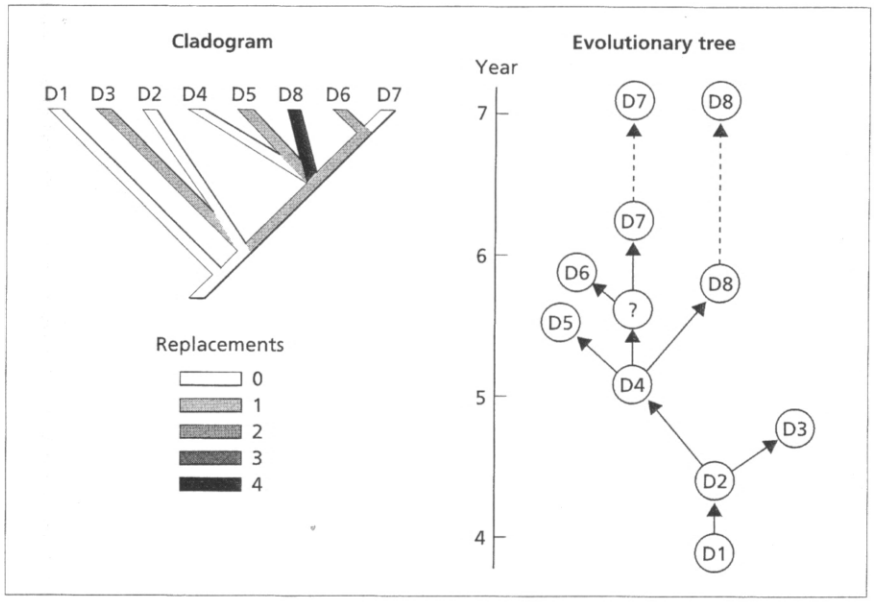
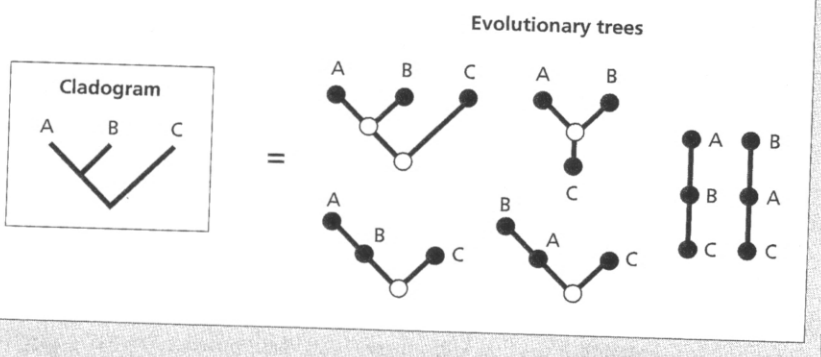


Fig. 2.14 Cladogram and corresponding evolutionary tree for eight V3 loop amino acid sequences for HIV samples taken from a single patient over 3 years. In the cladogram on the left all eight sequences are depicted as terminal nodes; however, four sequences (D1, D2, D4 and D7) have no autapomorphies (i.e. there are no replacements along the branch leading to each sequence) and hence are possible ancestors. The evolutionary tree on the right depicts the same relationships as the cladogram, but the sequences lacking autapomorphies (except D7) are treated as ancestors which is consistent with the order of appearance of the sequences. Modified from Holmes *et al.* (1992).

obtained from a single patient over 3 years. Because the samples were obtained over a period of time it is possible that some of the sequences sampled earlier in time gave rise to later sequences. Indeed, some sequences lack autapomorphies and hence by the cladistic criterion are potential ancestors, a conclusion which is supported by the order of the sequences in time.

Box 2.3 Cladograms and evolutionary trees

In this book we use the term 'cladogram' to refer to an evolutionary tree that has no information on branch lengths (e.g. Fig. 2.5). Within cladistics a distinction is made between a cladogram and an evolutionary tree. In a cladogram the terminal taxa are always at the tips of the tree, no matter if the taxa are extant or extinct, or whether one or more of the taxa are ancestral to any of the others. However, in an evolutionary tree some of the taxa may be ancestral to the others. Given the cladogram ((A, B), C) shown below, there are six different evolutionary trees that are consistent with the cladogram. One of these trees is the cladogram itself; the other five trees have one or more of the taxa A, B and C being ancestral to the others. Note that in all six trees A and B are more closely related to each other than to C.



In the vast majority of cases none of the taxa (or sequences) being studied will be ancestral and hence the cladogram is also an evolutionary tree. Exceptions may occur when fossils are being studied (although the probability that a given fossil is actually part of an ancestral lineage is rather remote) or in the case where samples have been taken over time from a rapidly evolving lineage, such as a virus (Fig. 2.14).

2.3 Trees and distances

Measures of sequence dissimilarity may be used to estimate the number of evolutionary changes that occurred in two sequences since they last shared a

common ancestor (see Chapter 5). These measures quantify the evolutionary distance between the two sequences. Trees themselves can also be represented by distances, and this link has motivated a range of tree-building methods that seek to convert pairwise distances between sequences into evolutionary trees. We shall describe these measures in Chapter 5. However, in order for a distance measure to be used to build phylogenies it must satisfy some basic requirements: it must be a metric, and it must be additive.

2.3.1 Metric distances

Let d (a, b) be the distance between two sequences, a and b. A distance d is a **metric** if it satisfies these properties:

1	$d(a, b) \geq 0$	(non-negativity)
2	$d(a, b) = d(b, a)$	(symmetry)
3	$d(a, c) \leq d(a, b) + d(b, c)$	(triangle inequality)
4	$d(a, b) = 0$ if and only if $a = b$	(distinctness)

The first property is *non-negativity*; two sequences must have a non-negative distance. The second property is *symmetry*; two sequences have the same dissimilarity regardless of the direction in which the dissimilarity is measured. These two properties may seem trivial, but not all measures of similarity meet these seemingly obvious requirements.

The third property is the *triangle inequality*, which states that the dissimilarity between any two sequences cannot exceed the sum of the dissimilarities between each sequence and a third. This condition is equivalent to ensuring that it is possible to represent the distances between the three sequences as a triangle (Fig. 2.15), hence the name. The last condition (*distinctness*) requires that sequences that are different must have a non-zero dissimilarity.

Of these conditions, 1, 2 and 4 are generally true for all measures of sequences dissimilarity calculated directly from sequences. However, indirect measures of sequence dissimilarity such as those obtained from DNA–DNA hybridisation or from immunological measurements need not always obey these conditions, particularly condition 2.

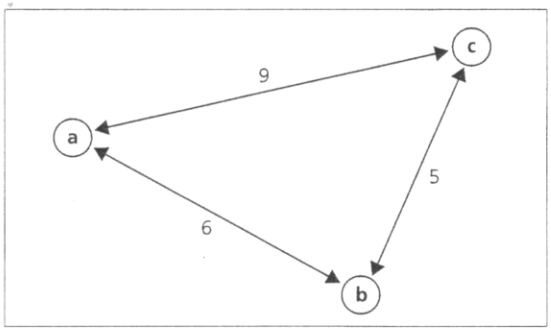


Fig. 2.15 The triangle inequality. The distance between any pair of sequences must be no greater than that between those sequences and a third sequence.

2.3.2 Ultrametric distances

A metric is an ultrametric if it satisfies the additional criterion that:

5 $d(a, b) \leq \text{maximum } [d(a, c), d(b, c)]$

This criterion implies that the two largest distances are equal, so that they define an isosceles triangle (Fig. 2.16).

Ultrametric distances have the very useful evolutionary property of implying a constant rate of evolution. Indeed the 'relative rate' test for a molecular clock (see Box 7.2, Chapter 7) is a test of how far the pairwise distances between three sequences depart from ultrametricity. Furthermore, if distances between sequences are ultrametric then the most similar sequences are also the most closely related.

2.3.3 Additive distances

Being a metric (or ultrametric) is a necessary, but not sufficient condition for being a valid measure of evolutionary change. A measure must also satisfy the *four-point condition*:

6 $d(a, b) + d(c, d) \leq \text{maximum } [d(a, c) + d(b, d), d(a, d) + d(b, c)]$

This is equivalent to requiring that of the three sums $d(a, b) + d(c, d)$, $d(a, c) + d(b, d)$, and $d(a, d) + d(b, c)$, the two largest are equal.

2.3.4 Tree distances

An additive distance measure defines a tree. Perhaps the easiest way to see this is to consider the distances shown in Fig. 2.17. Sequence d is equidistant from all other sequences; sequence c is equidistant from a and b. If we take any three sequences the distances between them define an isosceles triangle (the two largest distances are equal), hence the distances shown in Fig. 2.17

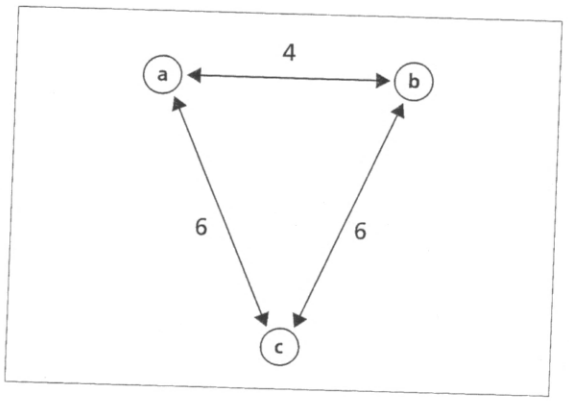


Fig. 2.16 The ultrametric inequality. The two largest pairwise distances, in this case $d(a, c)$ and $d(b, c)$, are equal and hence the ultrametric defines an isosceles triangle.

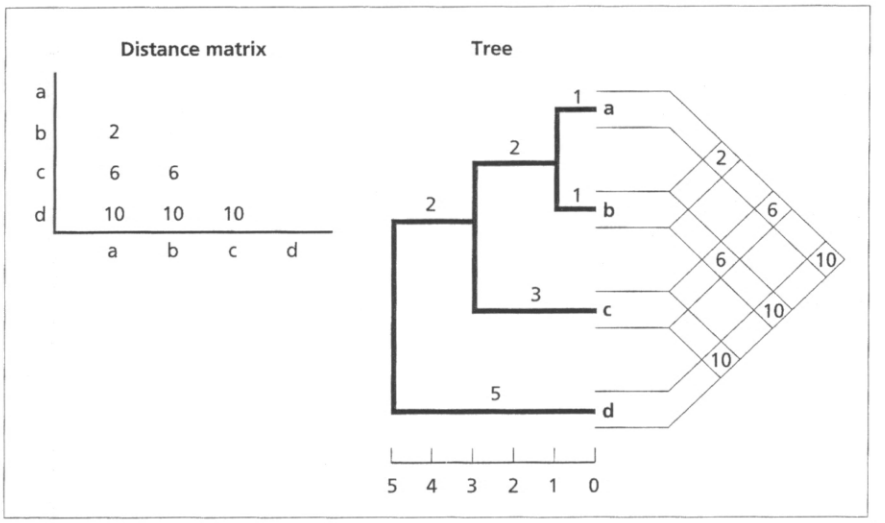


Fig. 2.17 An ultrametric distance matrix between four sequences a–d and the corresponding ultrametric tree. For any two sequences, the value in the distance matrix corresponds to the sum of the branch lengths along the path between the two sequences on the tree.

are ultrametric. These same distances can be represented by the ultrametric tree shown in Fig. 2.17. If we trace the shortest path between any pair of sequences in the tree, and add up the corresponding branch lengths, we obtain the same values as those in the distance matrix. For example, travelling from

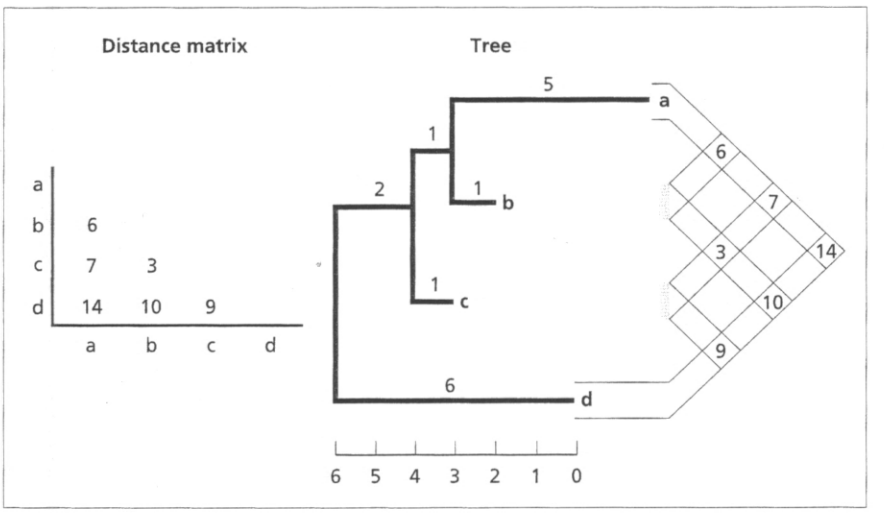


Fig. 2.18 An additive distance matrix between four sequences and the corresponding additive tree. For any two sequences, the value in the distance matrix corresponds to the sum of the branch lengths along the path between the two sequences on the tree.

sequence a to sequence d and adding branch lengths we obtain the value of $1 + 2 + 2 + 5 = 10$, hence $d(a, d) = 10$.

When distances are not ultrametric but only metric they can still be represented by a tree, in this case an additive tree (Fig. 2.18). This additive tree again represents the additive distances exactly. Notice that sequences b and c are the most similar ($d(b, c) = 3$) but are not the most closely related. Similarity and evolutionary relationship will only coincide exactly if the distances are ultrametric. This has important implications for using distances to reconstruct trees (Chapter 6).

The distances obtained from the tree are **tree distances** (also called 'patristic distances'), to distinguish them from **observed distances** which are obtained directly from the sequences themselves. In the examples shown in Fig. 2.17 and Fig. 2.18, the observed and tree distances match exactly. For real data this is rarely the case, indicating that the observed distances cannot be completely accurately represented by a tree. The discrepancy between observed and tree distances can be used to measure how good the fit is between the observed distances and the best tree representation of those distances (see Chapter 6).

2.4 Organismal phylogeny

Although the main subject of this book is molecular evolution, a major use of DNA sequences is the reconstruction of the evolutionary history of the organisms from which those sequences are obtained.

2.4.1 Clades and classification

Phylogenies form the basis of **classification**, which is the formal naming of groups of organisms. Cladistic classifications recognise only **monophyletic** groups or **clades**. A monophyletic group includes all the descendants of an ancestral taxon, whereas a non-monophyletic group omits some of those descendants (Fig. 2.19). A good example of a non-monophyletic group is the 'apes', which is not monophyletic as it excludes humans.

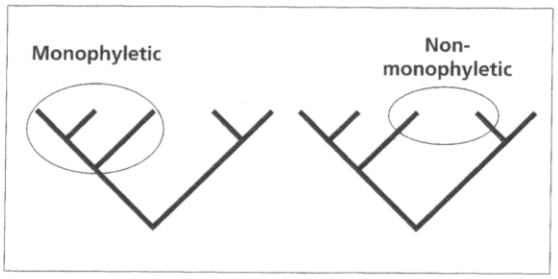


Fig. 2.19 The difference between monophyly and non-monophyly. A monophyletic group includes all descendants of their common ancestor, whereas in a non-monophyletic group one or more descendant is not included.

Many authors distinguish between two kinds of non-monophyly: **paraphyly** and **polyphyly** (Fig. 2.20). Paraphyletic groupings are based on shared primitive characters (plesiomorphies), and hence typically exclude one or more taxa that have autapomorphies. The paradigm example is the 'reptiles' as classically defined, which excludes birds because of their novel anatomy and behaviour, even though crocodiles are more closely related to birds than they are to other reptiles. Polyphyletic groups are typically assemblages of taxa that have been erroneously grouped on the basis of convergent characters, such as 'vultures'. The New and Old World vultures look strikingly similar but have evolved independently from different ancestors (storks and birds of prey, respectively).

Cladistic classifications have often been criticised as being limited in that they tell us little about the organisms themselves beyond who their nearest relatives are. For example, advocates of more traditional approaches to classification like to be able to reflect the evolutionary innovation shown by birds compared to their closest living relatives (the crocodiles) by elevating birds to their own class and consigning their dowdy relatives to the non-descript group the 'reptiles'. However, rigorous and objective alternatives to cladistic classifications have been hard to construct. Cladistic classifications also have the great advantage of being immune to variation in rates of evolution. For

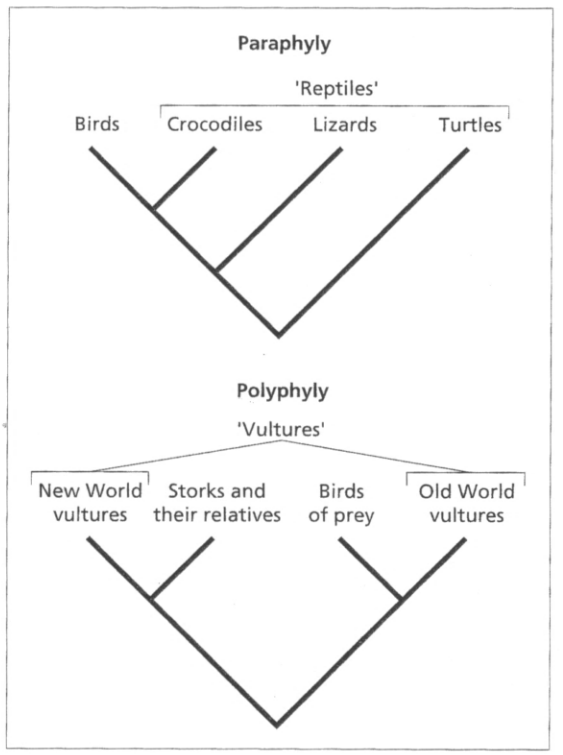


Fig. 2.20 Two kinds of non-monophyletic groups. 'Reptiles' are a paraphyletic grouping that is based on the absence of the apomorphic ('derived') characters possessed by birds. 'Vultures' are a polyphyletic grouping comprising two groups of birds that have independently evolved similar morphology and habits from different ancestors.

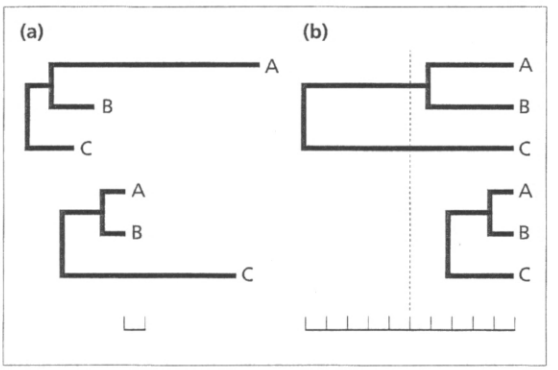


Fig. 2.21 Examples of variation in rate of evolution among genes from the same organisms. For all four trees the cladistic group AB is preserved. Dashed line is an arbitrary threshold for placing species in different higher taxonomic groups.

example, in Fig. 2.21(a) we have additive trees for two different genes from the same organisms A–C. In the first case the gene from species A has evolved much more rapidly than its homologous gene in B and C. Classifying the three species based on similarity would lead us to group B and C together. However, a second gene indicates that A and B are more conservative than C, in which case we might prefer to group A and B to the exclusion of C. If similarity is our criterion for delimiting taxonomic groups then we would have to choose between these two genes, essentially an arbitrary choice. Note however that the cladistic relationship remains the same in both trees. Figure 2.21(b) shows a different case where the rate of evolution for a given gene is constant. This might lead us to base taxonomic groups on amount of genetic divergence. However, using this method another gene evolving at a slower rate might lead to a different classification. Again, the cladistic groupings have not changed.

2.4.2 Gene trees and species trees

The naïve expectation of molecular systematics is that phylogenies for genes match those of the organisms, hence obtaining the first necessarily gives us the second. However, there are a number of reasons why this need not be so. The first is that gene duplications may result in a species containing a number of distinct but related sequences. In the example shown in Fig. 2.22 three species A–C each have two copies of the same gene, α and β. A phylogeny for all six genes allows us to correctly recover the organismal phylogeny ((A, B), C) from either the α or β genes. However, if we were unfortunate enough to sequence only genes 1, 3 and 5, and were unaware that they were part of a larger gene tree, we would infer that the organismal tree was ((A, C), B) because gene 3 from species C is more closely related to gene 1 from species A than is gene 5 from species B, even though species B is actually closer to species A than is species C.

This example illustrates that organismal phylogeny can be correctly inferred only if we have the complete set of genes, or if we restrict ourselves to

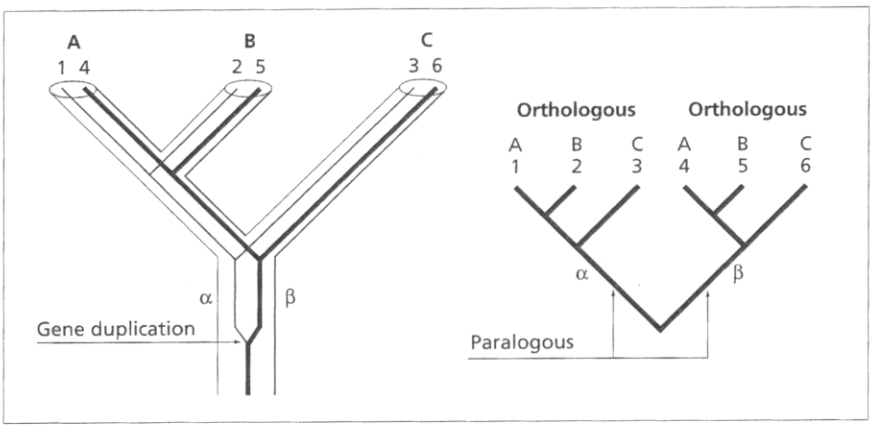


Fig. 2.22 Phylogeny for three species A–C and six genes that stem from a gene duplication resulting in two paralogous clades of genes, α and β. The α genes 1–3 are orthologous with each other, as are the β genes 4–6; however each α gene is paralogous with each β gene as they are separated by a gene duplication event, not a speciation event.

a set of genes that have not themselves undergone a duplication. That is, we require **orthologous** genes. Two homologous genes are orthologous if their most recent common ancestor did not undergo a gene duplication, otherwise they are termed **paralogous**. In Fig. 2.22 genes 1–3 are orthologous, as are genes 4–6, but any pair of α and β genes are paralogous.

2.4.3 Lineage sorting and coalescence

Another process that complicates the relationship between organismal and gene phylogeny is **lineage sorting**. Even if we restrict our attention to orthologous genes for the reason given above, the presence of ancestral polymorphism coupled with the differential survival of those alleles can result in allele phylogeny not matching organismal phylogeny. If we start with a pair of orthologous alleles and travel down the tree (i.e. backwards in time) we will eventually encounter their most recent common ancestor. This is the point at which the two gene lineages **coalesce** (Fig. 2.23) and the time at which this occurs is the **coalescence time**.

In the example shown in Fig. 2.23, alleles 3 and 4 have a recent coalescence point which lies within their organismal lineage B. However, alleles 1 and 2 have a more ancient coalescence time which pre-dates the age of their lineage A, that is, they are older than species A. Furthermore, even though alleles 1 and 2 are both found in the same species, they are not each other's closest relative. The presence of a paraphyletic pair of alleles in lineage A may have consequences later on in evolutionary time. Imagine that shortly after species A and B diverged, and while alleles 1 and 2 were still both extant, species A itself speciated into species A_1 and A_2 (Fig. 2.24). In this case, species A_1 inherited

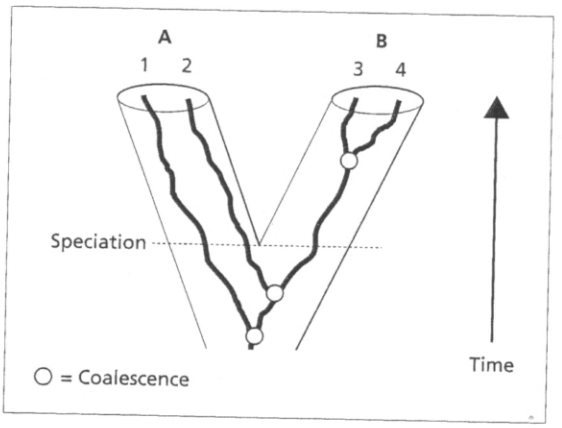


Fig. 2.23 A gene tree for four alleles (1–4) in two organismal lineages, A and B. The points at which pairs of allele lineages join (coalesce) are marked by open circles. Alleles 3 and 4 coalesce within lineage B, but alleles 1 and 2 are older than lineage A. Note especially that alleles 1 and 2 do not form a monophyletic group—2 is more closely related to 3 and 4 than it is to the other allele (1) found in the same species.

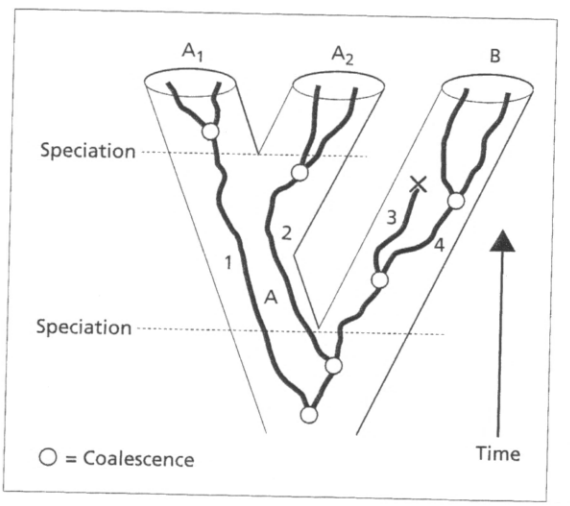


Fig. 2.24 One possible extrapolation into the future of the gene and species trees shown in Fig. 2.23. In this instance species A diverged into species A_1 and A_2. The two alleles (1 and 2) present in A when it speciated were inherited by A_1 and A_2 respectively. Allele 3 has gone extinct.

allele 1 and species A_2 inherited allele 2. Put another way, the two allele lineages 1 and 2 were sorted among the descendants of A. Note that even though all three species have monophyletic suites of alleles, the alleles found in A_2 are actually more closely related to species B than to its sister species A_1. Were we to use the phylogeny of these alleles to infer the phylogeny of the three species A_1, A_2 and B we would incorrectly conclude that the species tree was (A_1, (A_2, B)). This hypothetical example illustrates the problem of lineage sorting. If the alleles present in a lineage prior to that lineage speciating are not monophyletic then the distribution and relationships of these alleles need not accurately reflect the phylogeny of the organisms themselves.

Lineage sorting is likely to be a problem for organismal phylogenetics if the time it takes for alleles within a lineage to coalesce is greater than the interval between successive speciation events.

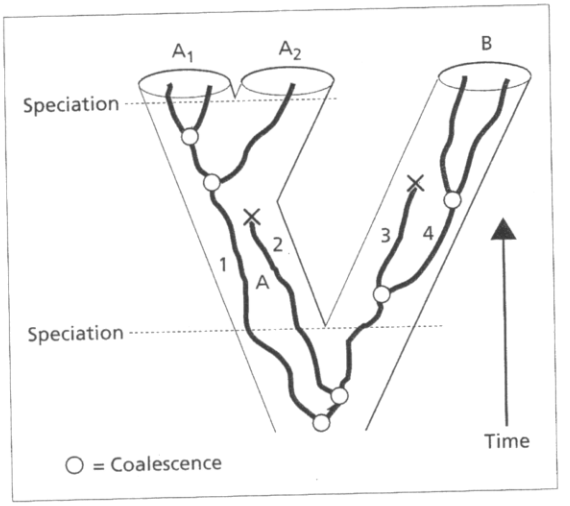


Fig. 2.25 The same situation as in Fig. 2.24 but lineage A speciating later in time, by which time allele 2 has gone extinct. Consequently species A_1 and A_2 inherit a monophyletic set of alleles.

Figure 2.25 shows an alternative extrapolation of Fig. 2.23 in which species A splits into two daughter species later than in Fig. 2.24, after allele 2 has gone extinct. Consequently, when A speciates its descendants receive a monophyletic set of alleles. In this case, allele phylogeny faithfully reflects species phylogeny.

The key difference between Fig. 2.24 and Fig. 2.25 is the length of time between successive speciations of the same lineage. Due to a combination of chance and selection, allele lineages will either persist, radiate or go extinct. The longer the interval between speciation events the greater the chance that these processes will result in lineages with a monophyletic set of alleles. The importance of gene trees and coalescence times for modern population genetics is discussed in more detail in Chapter 4.

2.5 Consensus trees

Often we want to compare trees derived from different sequences, or from the same sequences using different methods. Given two or more different trees we can ask 'what do these trees agree on?' **Consensus trees** are trees that represent the commonality (if any) among a set of trees. For example, consider the two trees for hominoids shown in Fig. 2.26. The two trees are very similar, but tree 1 groups humans and chimps together, whereas tree 2 groups the chimp and gorilla. Both trees agree that humans and African apes are more closely related to each other than each is to the orang-utan, and that the great apes and humans form a clade that excludes the gibbon. We can summarise the agreement between these trees in a consensus tree, which contains a polytomy indicating that there is conflict concerning our relationships with the African apes. This is an example of a soft polytomy (see section 2.1.1); the consensus tree is not indicating that the African apes

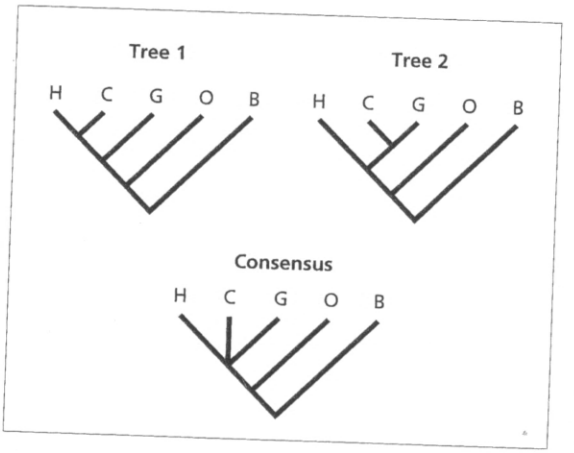
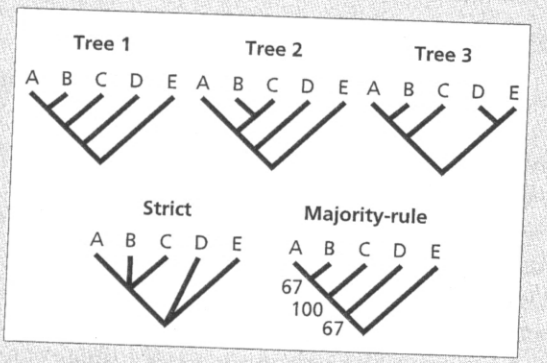


Fig. 2.26 Two different trees for humans (H), chimps (C), gorillas (G), orang-utans (O) and gibbons (B), and their consensus tree.

and humans speciated simultaneously, rather we have insufficient evidence to determine the exact order of speciation. There is a range of different consensus methods, three of which are discussed in Box 2.4.

Box 2.4 Types of consensus tree

A consensus tree summarises information common to two or more trees. There is a range of different methods which differ in what aspect of tree information they use, and how frequently that information must be shared among the trees to be included in the consensus. We discuss three commonly used methods here. The **strict consensus** tree includes only those groups (or splits, see section 2.1.6) that occur in all the trees being considered. Among the three trees below, only the split {{A, B, C}, {D, E}} is common to all three trees, and so the strict consensus of these trees contains just that split.



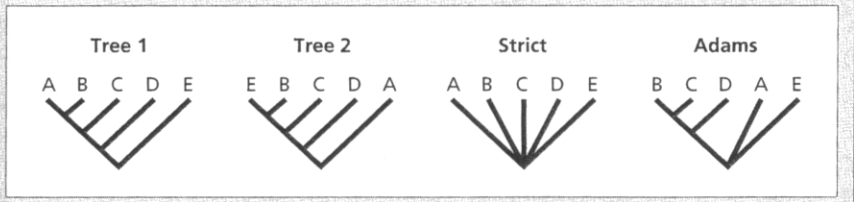
Notice, however, that there are some splits such as {{A, B}, {C, D, E}} that are found in two of the three trees. We can relax the requirement that a split be in

continued

Box 2.4 *continued*

all trees; for example, we could retain those splits found in a majority of trees, and this is exactly what the **majority-rule consensus** does. Any split in more than half the trees is included in the consensus tree, so the two splits shared by two of the three trees are also included. Note that any split in the strict consensus tree will also be in the majority-rule tree. The splits in the majority-rule tree are usually labelled by what percentage of trees that split occurs in.

Strict and majority-rule consensus methods are two examples of methods that use splits. However, trees that have no splits in common (and hence will give completely unresolved strict and majority-rule consensus trees) may still have points of similarity. For instance, the two trees below share no splits, yet both agree that if we consider just sequences B, C and D, B is more closely related to C than either is to D.



The strict consensus tree for these trees is a star tree, but this somewhat overstates the differences between the two trees. The **Adams consensus** tree captures the information that both tree 1 and tree 2 have the subtree ((B, C), D). Although Adams consensus trees can sometimes be a little difficult to interpret, they are very useful in situations where one or more sequences have very different positions on different trees, but there is a subset of sequences upon whose relationships the different trees agree.

2.6 Networks

So far in this chapter we have assumed that the evolutionary relationships among sequences and organisms are best represented by a tree. In other words, we are using the tree to model reality. However, the actual evolutionary history may not be particularly tree-like, in which case analyses that assume a tree may be seriously misleading.

For example, the metaphor of a 'family tree' is itself rather misleading, as anyone will know who has drawn one. A tree has single root and branches outwards such that the branches never meet, whereas in a family tree or pedigree every time a male and female organism mate their branches fuse. Generally the history of each individual gene can be adequately represented by a tree; however, in cases where a gene has undergone recombination a

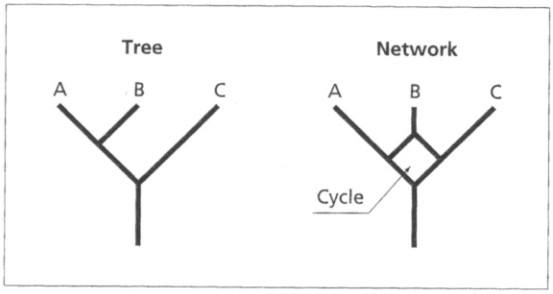


Fig. 2.27 A tree and a network. Networks contain cycles, whereas trees do not.

network may be more appropriate. A network (Fig. 2.27) contains one or more cycles (a set of nodes where it is possible to trace a path starting and ending at the same node without visiting any other node more than once).

2.7 Summary

1 Evolutionary relationships can be represented by a variety of trees. Cladograms depict relative recency of common ancestry, additive trees incorporate branch lengths, ultrametric trees can be used to represent evolutionary time.

2 Trees may be either rooted or unrooted, but only rooted trees have an evolutionary direction.

3 The number of possible trees increases rapidly with increasing number of sequences.

4 Evolutionary trees can depict ancestor–descendant relationships.

5 Distances satisfying the 'four-point condition' define a corresponding tree.

6 Gene trees may differ from species trees.

2.8 Further reading

Maddison and Maddison (1992) give an excellent introduction to trees and phylogenies. Barthélemy and Guénoche (1991) provide a detailed and elegant discussion of the kinds of trees, and the relationship between distances and trees. See Poinar and Poinar (1995) for the recovery of DNA from amber, Austin *et al.* (1997) for a sceptical review of the authenticity of geologically ancient DNA, and Smith (1994) on the problem of ancestors. The HIV example is taken from Holmes *et al.* (1992). For the distinction between hard and soft polytomies, see Maddison (1989). Swofford (1991) provides an excellent review of consensus methods. For a discussion of classification see chapter 14 in Ridley (1996).