Trees and their terms

Roderic Page
DEEB, IBLS
University of Glasgow

Note: This material is based on Chapter 2 in Page, R. & Holmes, E. 1998 "Molecular Evolution: A Phylogenetic Approach," Blackwell Science ISBN 0-86542-889-1

Trees

Figure 1 illustrates some terminology used 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.

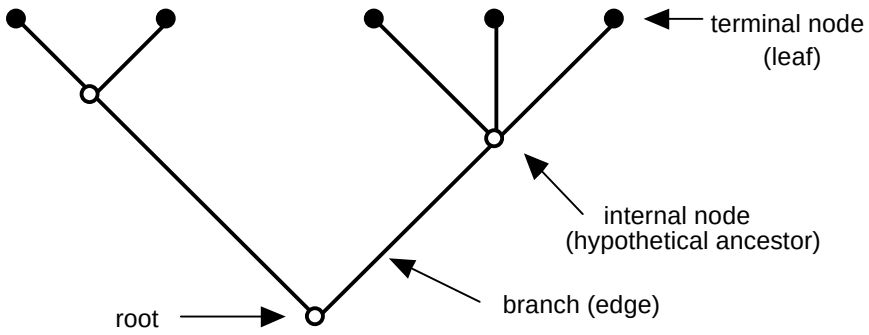


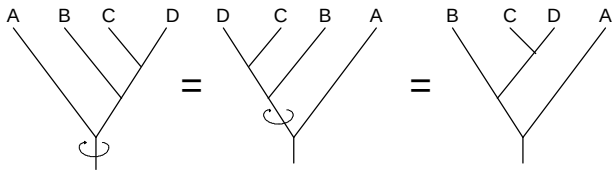
Figure 1 A simple tree and associated terms.

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).

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 represent as **branch lengths** (or **edge lengths**). Trees with branch lengths are sometimes called **weighted trees**.

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.

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 (Figure 2).

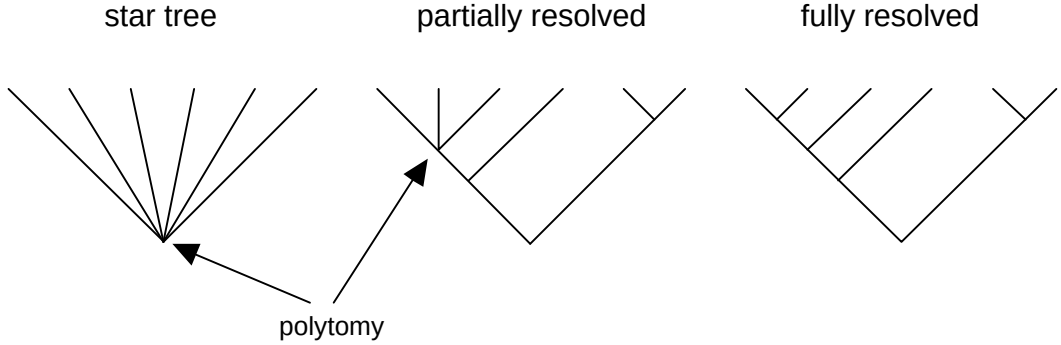


Figure 2 Three trees showing various degrees of resolution, ranging from a complete lack of resolution (star tree) to a fully resolved tree. Any internal node with more than two immediate descendants is a polytomy.

Polytomies can represent two rather different situations (Figure 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."

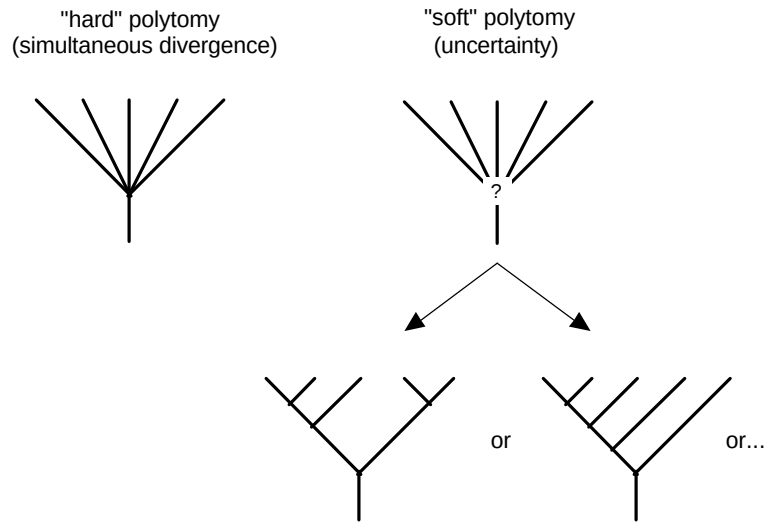


Figure 3 Polytomies can represent either simultaneous divergence of multiple sequences ("hard"), or lack of resolution due to insufficient data or conflicting trees ("soft").

A shorthand for trees

Trees can be represented by a shorthand notation that uses nested parentheses. Each internal node is represented by a pair or 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 4 gives an example of this shorthand.

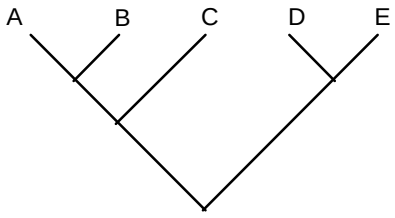


Figure 4 A tree and its shorthand representation using nested parentheses.

Cladograms, additive trees, and ultrametric trees

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

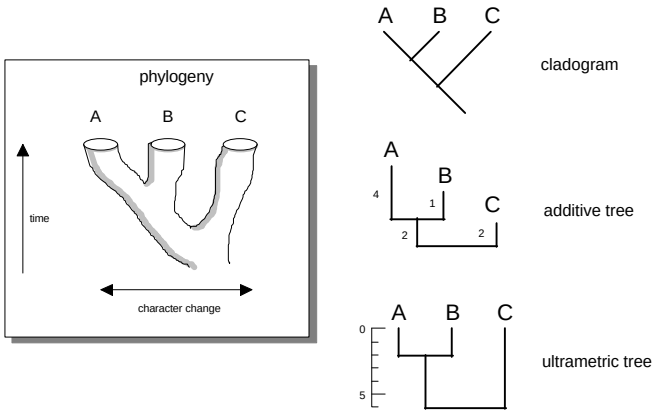


Figure 5. A phylogeny and the three basic kinds of tree used to depict that phylogeny. The cladogram represents relative recency of common ancestry; the additive tree depicts the amount of evolutionary change that has occurred along the different branches, and the ultrametric tree depicts times of divergence.

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 Figure 5, sequence A has acquired 4 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.

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 don't 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 Figure 6, the gibbon (B) and orangutan(O) sequences are neighbours on the tree, yet the orangutan 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 orangutan sequences would be indeed closely related.

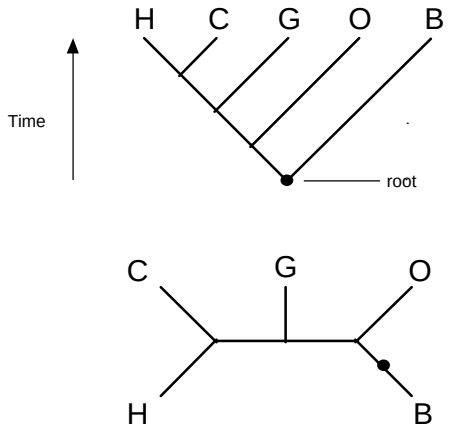


Figure 6. Rooted and unrooted trees for humans (H), chimps (C), gorilla (G), orangutan (O), and gibbon (B). The rooted tree (top) corresponds to the unrooted tree below.

In the unrooted tree for the apes shown in Figure 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 (Figure 7).

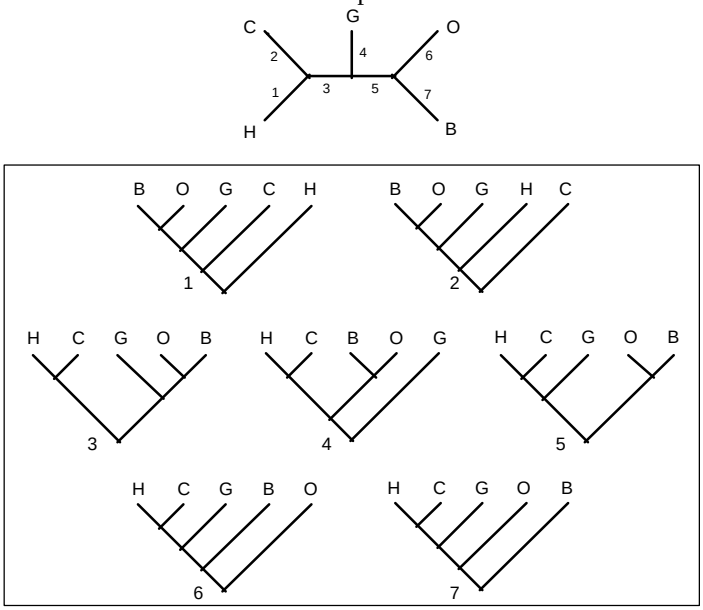


Figure 7. The seven rooted trees that can be derived from an unrooted tree for five sequences. Each rooted tree 1-7 corresponds to placing the root on the corresponding numbered branch of the unrooted tree. (Sequence labels as for Figure 6).

The distinction between rooted and unrooted trees is important because many methods for reconstructing phylogenies reconstruct unrooted trees, and hence cannot distinguish among the seven trees shown in Figure 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. 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

$$U_n = (2n-5)(2n-7)\ldots(3)(1) \qquad (1)$$

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

$$R_n = (2n-3)(2n-5)\cdots(3)(1) = (2n-3)U_n \quad (2)$$

for $n \geq 3$. Table 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 the 135 human mitochondrial DNA sequences used in the "out of Africa" study (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.

Table 1 Numbers of unrooted and rooted trees for 2-10 sequences.

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

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. However, other aspects of the tree may also reflect evolutionary phenomena and hence be of interest. Figure 8 shows the three possible **shapes** (or **topologies**) for a rooted tree for five sequences. All 105 possible trees (Table 1) for five sequences will have one of these three shapes.

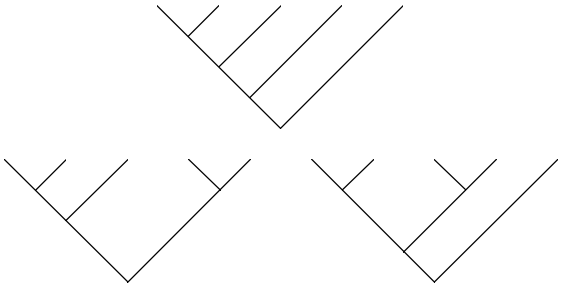


Figure 8 The three possible shapes for a rooted tree for five sequences.

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 Figure 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 resulting from chopping internal branches. The tree shown in Figure 9 has two informative splits: { {C, H}, {G, O, B} } and { {G, C, H}, {O, B}}.

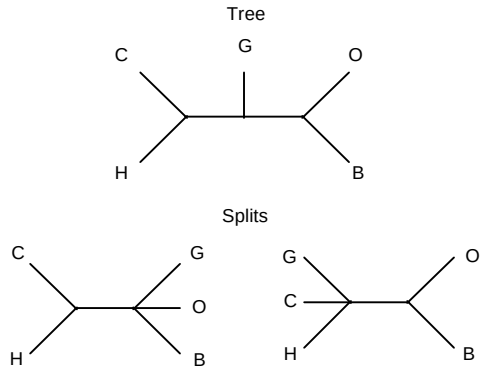


Figure 9. An unrooted tree and its two splits.

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 Figure 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. On Monday 8th you will encounter some methods for reconstructing phylogenies that make use of this relationship between nucleotide sites and splits.

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, a 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 spectacular recovery of ancient DNA from fossils up to 100+ million years old; 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 a termite trapped in Dominican amber an ancestor to modern termites or is it on a 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 (Figure 10). Treating a taxon with autapomorphies as an ancestor would require us to postulate additional evolutionary changes. Note that under this rule a taxon with no autapomorphies need not be an ancestor, rather there is nothing to refute that possibility.

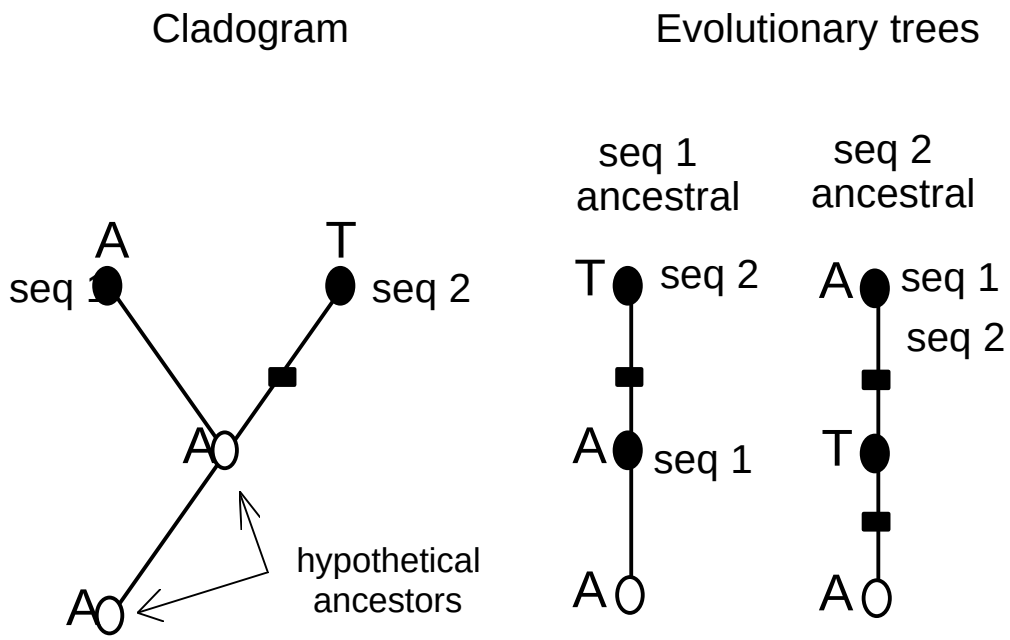


Figure 10 A cladogram for two sequences (seq 1 and seq 2) showing the nucleotide at a single site, and two possible of several evolutionary trees derived from that cladogram. We could postulate that either sequence is ancestral to the other. However, postulating seq 2 to be an ancestor of seq 1 requires 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 2.1).

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, Figure 0.11 shows a cladogram for eight HIV sequences obtained from a single patient over three 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.

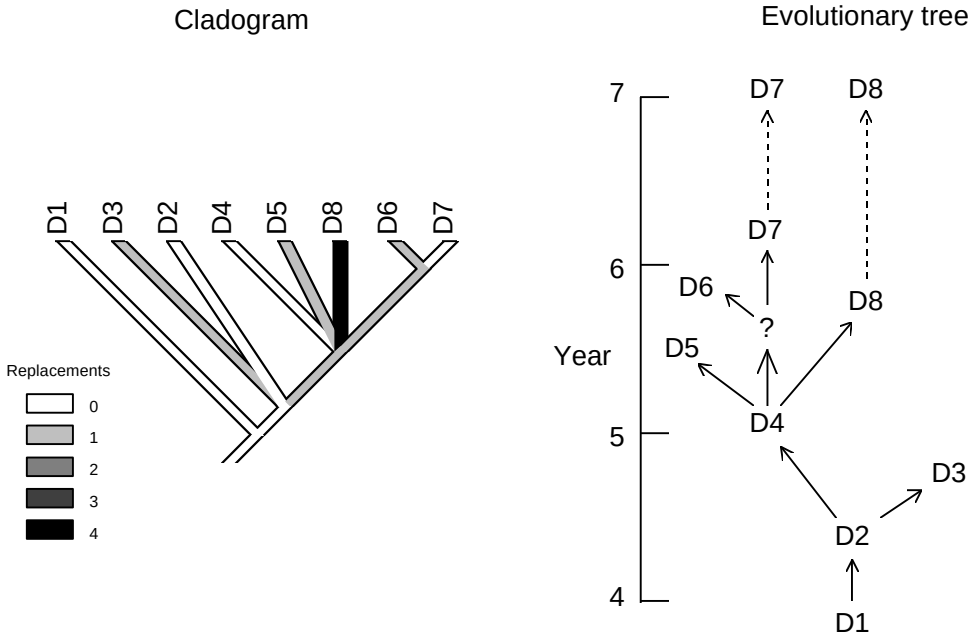
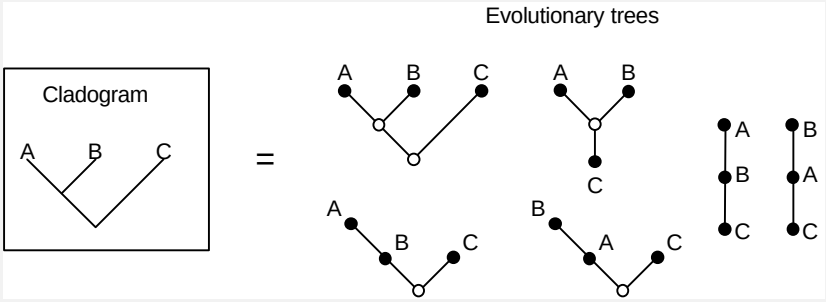


Figure 0.11 Cladogram and corresponding evolutionary tree for eight V3 loop amino acid sequences for HIV samples taken from a single patient over three years. In the cladogram on the left all eight sequences are depicted as terminal nodes; however, four sequences (D1, D2, D4, and D7) have no autapomorphies (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).

Box 1 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., Figure 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 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 (Figure 0.11).

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. 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 later in the course. However, in order for a distance measure to be used to build phylogenies it must satisfy some basic requirements: it must be either a metric, and it must be additive.

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 nonnegative 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 sequence 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 (Figure 12), hence the name.

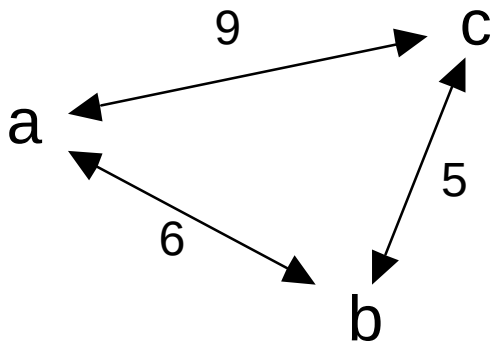


Figure 12 The triangle inequality. The distance between any pair of sequence must be no greater than that between those sequences and a third sequence.

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.

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 (Figure 13).

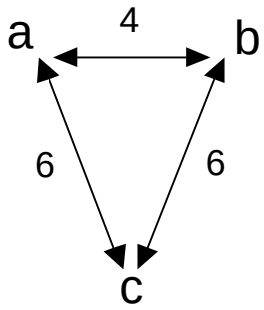


Figure 13 The ultrametric inequality. The two largest pairwise distances, in this case *d*(a, c) and *d*(b, c) are equal, hence the ultrametric defines an isosceles triangle.

Ultrametric distances have the very useful evolutionary property of implying a constant rate of evolution. Indeed the “relative rate” test for a molecular clock 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.

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.

Tree distances

An additive distance measure defines a tree. Perhaps the easiest way to see this is to consider the distances shown in Figure 14.

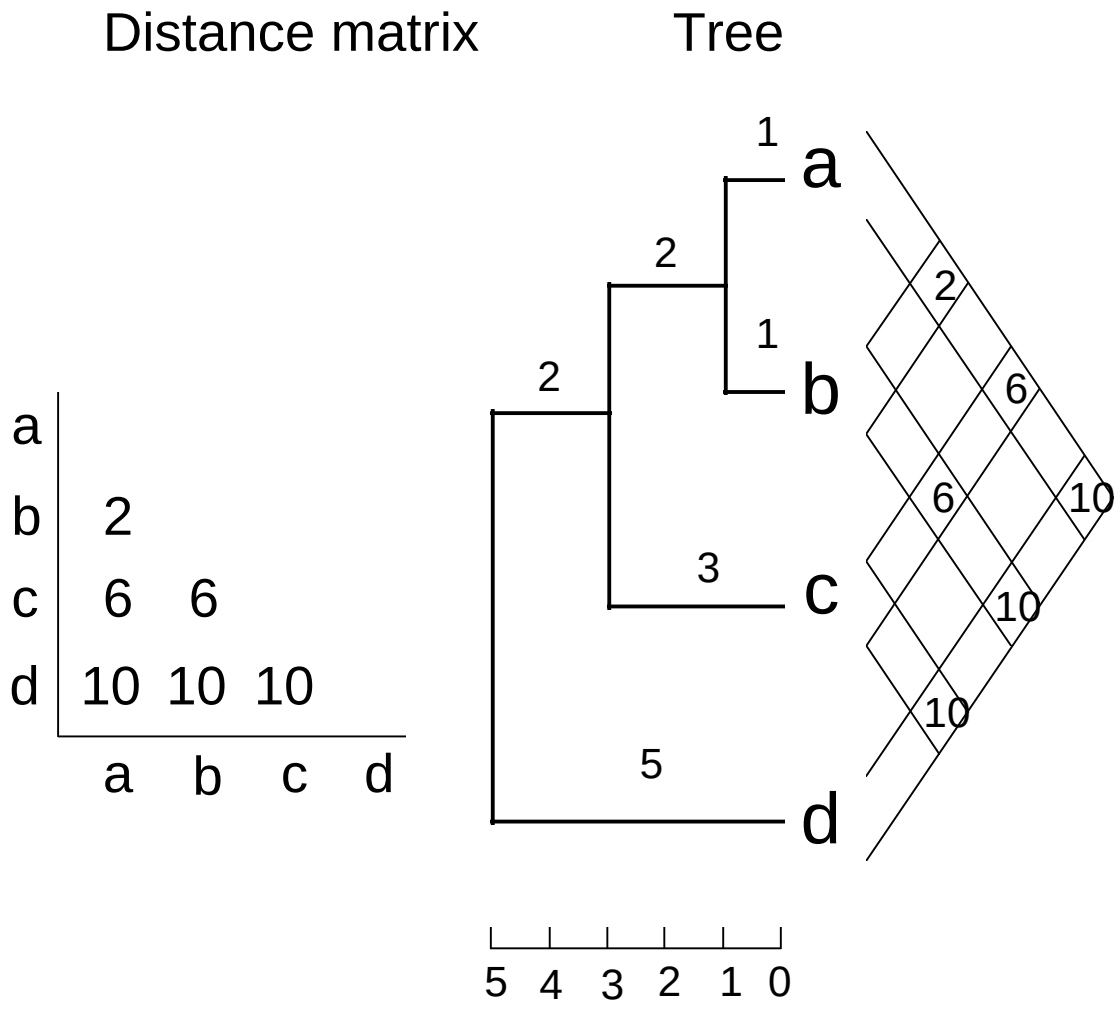


Figure 14 An ultrametric distance matrix between four sequences a-d and the corresponding ultrametric tree. For any two sequences, the value in the distance matrix corresponds to the sum of the branch lengths along the path between the two sequences on the tree.

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 Figure 14 are ultrametric. These same distances can be represented by the ultrametric tree shown in Figure 14. 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 value as that in the distance matrix. For example, travelling from 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 Figure 15.

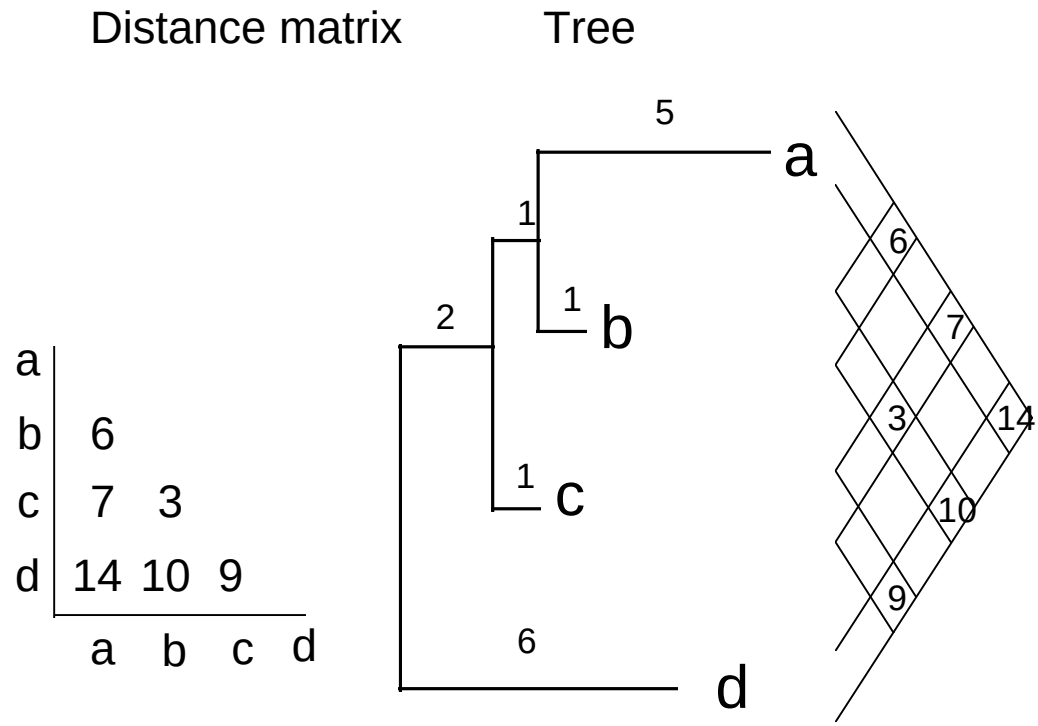


Figure 15 An additive distance matrix between four sequences and the corresponding additive tree. For any two sequences, the value in the distance matrix corresponds to the sum of the branch lengths along the path between the two sequences on the tree.

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.

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 Figure 14 and Figure 15, 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.

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 (1993) for the recovery of DNA from amber, 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).

References

Barthélemy, J.-P. and Guénoche, A. (1991). "Trees and proximity representations", John Wiley & Sons, Chicester.

Holmes, E. C., Zhang, L. Q., Simmonds, P., Ludlam, C. A. and Leigh Brown, A. J. (1992). Convergent and divergent sequence evolution in the surface envelope glycoprotein of humman immunodeficiency virus type I within a single infected patient. *Proceedings of the National Academy of Sciences, USA* **89:** 4835-4839.

Maddison, W. P. (1989). Reconstructing character evolution on polytomous cladograms. *Cladistics* **5:** 365-377.

Maddison, W. P. and Maddison, D. R. (1992). "MacClade: Analysis of phylogeny and character evolution. Version 3.0", Sinauer Associates, Sunderland, Massachusetts.

Smith, A. B. (1994). "Systematics and the fossil record: documenting evolutionary patterns", Blackwell Scientific, Oxford.