Introduction to Tree Building

Roderic Page
DEEB, IBLS
University of Glasgow

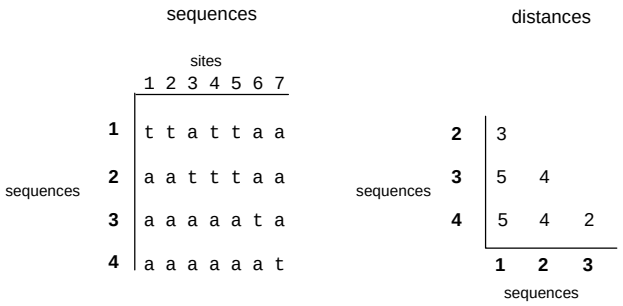
Note: This material is based on Chapter 6 in Page, R. & Holmes, E. 1998 "Molecular Evolution: A Phylogenetic Approach," Blackwell Science ISBN 0-86542-889-1

Tree building methods

This lecture provides a quick overview of different tree building methods. A great (and ever increasing) number of methods have been described for doing this, which raises the inevitable question of how to come to grips with this plethora of possibilities. Two ways which seem useful to us are either to divide the methods by how they handle data, or to divide them by the approach taken when building trees.

Kinds of data: distances versus discrete characters

This division is based on how the data is treated; distance methods first convert aligned sequences into a pairwise distance matrix, then input that matrix into a tree building method, whereas discrete methods consider each nucleotide site (or some function of each site) directly. As an example, consider the following sequences and corresponding (uncorrected) distance matrix:



The trees obtained by parsimony (a discrete method) and minimum evolution (a distance method) are identical in topology and branch lengths (Figure 1).

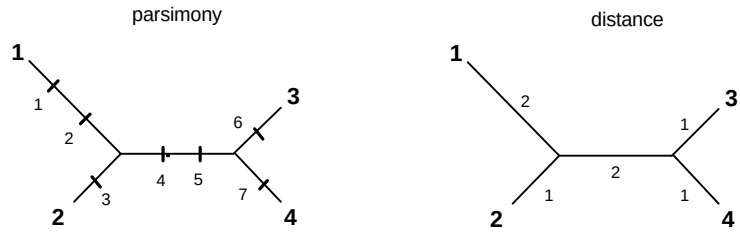


Figure 1 A parsimony tree and a distance tree for the same sequence data. Note that both trees have the same topology and branch lengths, but that the parsimony tree identifies which site contributes to the length of each branch.

Clustering vs. optimality

Another way of dividing tree building methods is by the way they construct trees. **Cluster methods** follow a set of steps (an algorithm) and arrive at a tree. For example, if we have five sequences we might start with three of them (remember that there is only one possible unrooted tree for three sequences) and decide where to place the fourth sequence. Given the resulting tree for four sequences, we the decide where to add the fifth and last sequence to our tree (Figure 2).

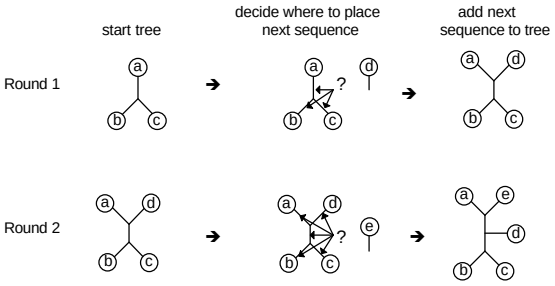
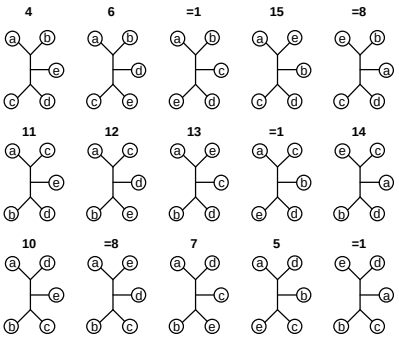


Figure 2 An example of how a clustering method builds a tree. The tree is constructed by starting with the tree for three sequences, then adding each remaining sequence in turn until finally all sequences have been added.



Tree building methods in the second class use **optimality criteria** to choose among the set of all possible trees. This criterion is used to assign to each tree a "score" or rank which is a function of the relationship between tree and data (examples include maximum parsimony and maximum likelihood).

Comparing different methods

Based on the distinction we have made between tree building methods that use distances versus those that use discrete characters, and methods that use a clustering algorithm versus those with an explicit optimality criterion, we can classify some commonly used methods (Figure 3).

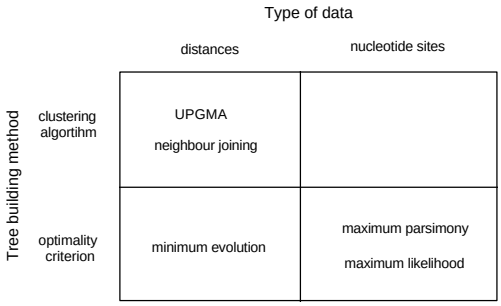


Figure 3 Some common phylogenetic methods classified by the method used to build the tree, and by the type of data used.

Given the range of tree building methods available, how can we decide which ones are better than others? David Penny and colleagues have suggested five desirable properties a tree building method should have:

- **efficiency** (how fast is the method?)
- **power** (how much data does the method need to produce a reasonable result?)
- **consistency** (will it converge on the right answer given enough data?)
- **robustness** (will minor violations of the method's assumptions result in poor estimates of phylogeny?)
- **falsifiability** (will the method tell us when its assumptions are violated, i.e., that we shouldn't be using the method at all).

Some Commonly Used Methods

Distance methods

Distance methods are based on the idea that if we knew the actual evolutionary distance between all members of a set of sequences, then we could easily reconstruct the evolutionary history of those sequences. This follows from the relationships between distances and trees : evolutionary distance is a tree metric and hence defines a tree. In practice however, distances are rarely, if ever, exactly tree metrics, and hence one class of "goodness of fit" methods seek the metric tree that best accounts for

the "observed" distances (i.e., the pairwise distances calculated between the sequences). The second class of method seeks the tree whose sum of branch lengths is the minimum ("minimum evolution").

Goodness of fit

The goodness of fit F between observed distance d_{ij} and tree distances p_{ij} for each pair of sequences i and j is given by

$$F_a = \sum_{1 \le i < j \le n} \left| d_{i,j} - p_{i,j} \right|^a$$

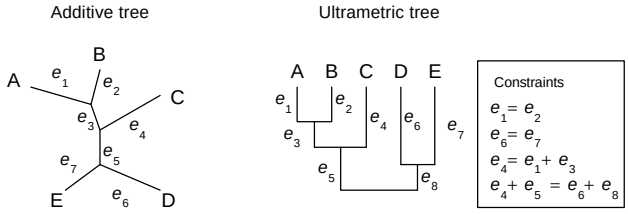
where α can take various values. If $\alpha = 1$ then the criterion is Farris's f statistic, if $\alpha = 2$ then F is the least-squares-fit criterion.

Minimum evolution

Given an unrooted metric tree for n sequences there are $(2n - 3)$ branches, each with length e_i. The sum of these branch lengths is the **length** L of the tree:

$$L = \sum_{i=1}^{2n-3} e_i$$

The minimum evolution tree is the tree which minimises L. This method is similar in spirit to parsimony, which we shall discuss below, however the length in this case is computed from the pairwise distances between the sequences rather than from the fit of individual nucleotide sites to a tree.



Additive and ultrametric trees for the same sequences. Both trees specify the same cladistic relationships among the taxa, but whereas the additive tree has $(2n-3=7)$ independent branches the ultrametric tree has only $(n-1) = 4$ because some branches are constrained to be equal to others, or to combinations of others. This constraint is equivalent to a molecular clock.

Discrete methods

In contrast to distance methods, discrete methods operate directly on the character data, rather than on pairwise distances. Hence they endeavour to avoid the loss of information that occurs when characters are converted into distances. The two major discrete methods are **maximum parsimony** (MP) and **maximum likelihood** (ML). Maximum parsimony chooses the tree or trees that require the fewest evolutionary changes. Maximum likelihood chooses the tree (or trees) that of all trees is the one that is most likely to have produced the observed data.

Parsimony

Among parsimony's advantages are that it is relatively straightforward to understand, it apparently makes few assumptions about the evolutionary process, it has been extensively studied mathematically, and some very powerful software implementations are available. However, the justification for choosing the most parsimonious tree as the best estimate of phylogeny is the subject of considerable controversy. Essentially two main arguments have been presented. The first is that parsimony is a methodological convention that compels us to maximise the amount of evolutionary similarity that we can explain as homologous similarity, that is, we want to maximise the similarity that we can attribute to common ancestry. Any character which does not fit a given tree requires us to postulate that the similarity between two sequences shown by that character arose independently in the two sequences — the similarity is due to homoplasy not homology. Hypotheses of homoplasy (such as convergence or parallel evolution) may be judged "ad hoc" in that they are attempts to explain why data do not fit a particular hypothesis. The most parsimonious tree minimises the number of ad hoc hypotheses required, and for that reason is preferred.

The second view is that parsimony is based on an implicit assumption about evolution, namely that evolutionary change is rare. Rarity of change implies that the tree that minimises change is likely to be the best estimate of the actual phylogeny. Under this view, parsimony may be viewed as an approximation to maximum likelihood methods, and indeed it was in this context that parsimony methods were first proposed by Edwards and Cavalli-Sforza.

Of the two positions, the latter has the advantage that it is possible to explore the circumstances under which parsimony will fail to reconstruct the correct phylogeny, and to develop a framework in which parsimony can be compared to other methods.

Likelihood

Given competing explanations for a particular outcome, which explanation should we choose? The principle of **likelihood** suggests that the explanation that makes the observed outcome the most likely (i.e., the most probable) occurrence is one to be preferred. Put more formally, if given some data *D*, and a hypothesis *H*, the likelihood of that data is given by

$$L_D = \Pr(D|H)$$

which is the probability of obtaining *D* given *H*. In the context of molecular phylogenetics *D* is the set of sequences being compared, and *H* is a phylogenetic tree, hence we want to find the likelihood of obtaining the observed sequences given a particular tree. The tree that makes our data the most probable evolutionary outcome is the **maximum likelihood estimate** of the phylogeny.

Do the methods work?

Given the range of possible methods for inferring phylogeny, we naturally want to know if they work, that is, do they recover the actual evolutionary elationships among nucleotide sequences? Several approaches have been developed to answer this question: analysis, simulation, known phylogenies, and congruence.

1. Analytical approach
   We can work out what assumptions a method makes, and when these are violated.

2. Simulation
   We can use computer generated phylogenies.

3. Known phylogenies
   We can use known or artificially generated phylogenies.

4 Congruence between data sets
   Do independent data sets yield the same trees?