(The following section of notes on parsimony is based on notes from `http://www.fos.auckland.ac.nz/~biosci742/4_3_2.html#4.3.4`)

20.6.2 Branch and bound

Branch and bound is a method of systematically analysing all possible trees by building up a tree one taxon (leaf) at a time and only continuing to build up a tree if it could potentially lead to the best tree.

Given n taxa, build an initial tree, t^* using some method. The score of that tree is s^*. Now we begin to systematically build up trees one taxon at a time as follows:

Initialise: Choose 3 taxa and form the (unique) unrooted partial tree.
Add this tree to a queue.

Iterate: Choose a taxon and add to previous best partial tree (at front of queue) in each possible position to get a k new partial trees, $t_1, \ldots, t_k$
If $score(t_i) \leq s^*$, add t_i to queue and order the queue by score.
If $score(t_i) > s^*$, discard t_i.
If t_i is complete (all taxa have been added) and $score(t_i) < s^*$, set $s^* = score(t_i)$.

Finish: When queue is empty, return tree with lowest score.

This becomes clearer by looking at an explicit example so refer to Figure 14.

The result is effectively the same as an exhaustive search, without wasting time on topologies that we know will be rejected.

The algorithm can be optimised by having having a good initial tree (try perhaps using a neighbour-joining tree) and by ordering the taxa so that they are added in a way that promoter earlier cutoffs.

This is an improvement over exhaustive search (which is feasible for up to about 10 sequences) and is feasible for around 20-30 sequences.

20.6.3 Heuristic search

Heuristic methods search for the optimal tree but offer no guarantee that it will be (or has been) found. These methods use hill-climbing to seek the optimal tree:

- choose an initial tree
- Iterate:
  - modify the tree and assess it
  - if the modified tree is an improvement, keep it. Else, return to the previous tree
  - stop when no improvement occurs

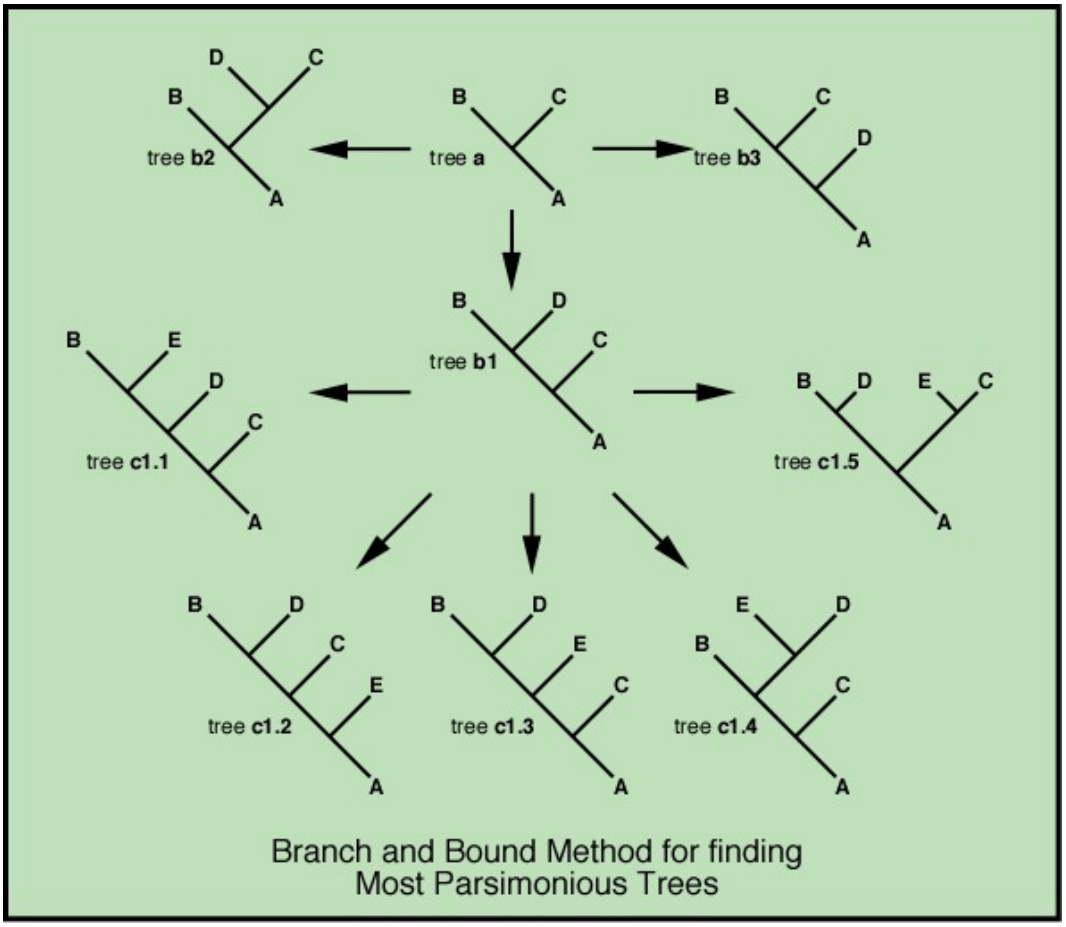

Figure 14: Say we have sequences from 5 taxa. We start by building the single 3-taxon tree using taxa A, B and C (tree a). Next the fourth taxon D is added in all three possible positions to generate trees b1, b2, and b3. One of these trees, say b1, is chosen. Then the fifth taxon E is added in all possible positions to give trees c1.1, c1.2, c1.3, c1.4, and c1.5. The length of each of these five 5-taxon trees is calculated. The shortest of these is the most parsimonious found to this point. Now return to the partial tree b2. If the length of b2 is equal to, or greater than, that of the shortest seen so far, then we know that adding any more taxa will only make the tree longer. If this is the case, then we stop using b2, and don't consider any of the trees built upon it. If b2 is shorter than the best seen so far, then it is used as the basis of further tree building, until the threshold length is reached. As we work through new topologies, we continuously update our record of the shortest seen so far. Once we have exhausted all possibilities, the shortest tree is the most parsimonious for that alignment.

The initial tree can be chosen using *stepwise attachment*, a greedy algorithm that starts by joining 3 taxa into a tree and then progressively adds further taxa by finding the best place to attach a taxon and leaving it there. Since taxa are never moved once they have been attached even if it becomes obvious that something has been attached in the wrong place, this method will almost never find the best tree to start with. It will, however, nearly always give us something better than the worst tree.

Modifications to the tree can be made by various methods of detaching and reattaching branches in different a different place. This is known as branch swapping. An example of one type of re-arrangement method is given in Figure 15.

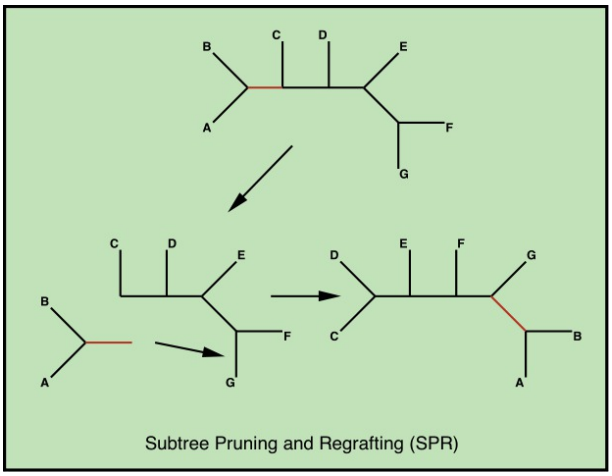

Figure 15: An example of a way of modifying the tree: subtree pruning and regrafting. An edge is chosen and the subtree at that point is removed. Another edge is chosen on the remainder of the tree and the removed subtree is reattached at that point.

The method described above is guaranteed to find a local minimum of the parsimony score, but may not find the global minimum as the starting tree may be too far from the best tree.

To improve the chances of finding the best tree, the method for building the initial tree can be randomised so the the starting point is different in every case. For example, if the initial tree is built by stepwise attachment, the order in which the taxa are added can be randomised. Different starting points may end up finding different local minima.

20.7 Disadvantages of parsimony

Beyond the difficulty of finding the maximum parsimony tree, parsimony has several disadvantages.

Firstly, parsimony does not account for hidden or multiple substitutions at the same site as it explains all substitutions with the minimum possible number of changes. So if we observe a locus in three sequences, one with an A, the other two with a C. Reconstructing

the parsimony tree, we will assume that the ancestral state was a C and a single mutation had occurred to produce the A. There are clearly many other explanations for this data set (for example, there were multiple mutations so that the ancestral state was A and two mutations to C occurred or the was a hidden mutation where the ancestral state was a C, there was a mutation to a G and then an A) which, although each less likely than the single mutation, collectively are quite likely. This effect means that *parsimony tends to underestimate the length of trees.*

The most serious problem with maximum parsimony is *long-branch attraction*, a consequence of the failure of the method to estimate multiple or hidden substitutions. When a tree has some branches with significantly greater length than other branches, MP will underestimate how many substitutions have occurred on the long branches. Homoplasies (parallel or convergent substitutions) will cause MP to underestimate the evolutionary difference between the branch tips. Conversely it will over-estimate the degree to which those tips have shared an evolutionary past. The long branches will be joined erroneously as near- or sister-clades, that is they will "attract" one another. Using longer genomic sequences in the analysis will only increase the number of variable sites exhibiting homoplasies, without improving the phylogenetic signal. As a consequence of this, MP is statistically inconsistent, that is, the chance of obtaining the wrong answer increases as more data are used. It can be positively misleading. See Figure 16.

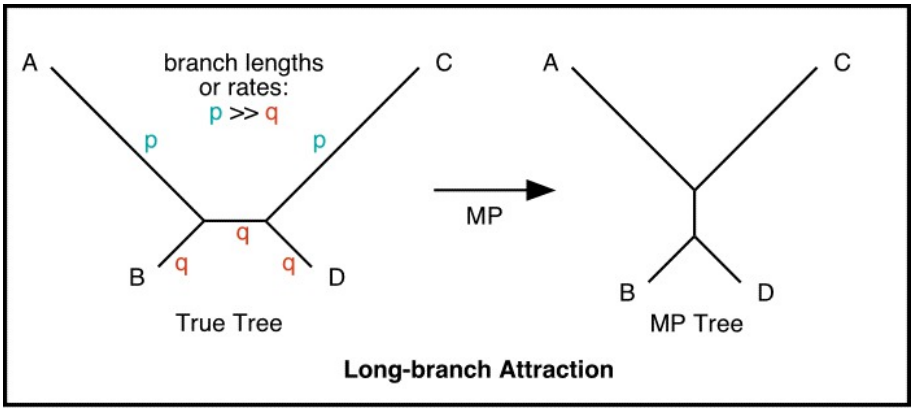

Figure 16: The tree on the left is the true tree. It has a pair of sequences (A, and B) which are highly diverged. Reconstructing the tree from these sequences using Maximum Parsimony (MP) results in the wrong tree (right). The two highly diverged sequences look closer to each other than they should due to chance mutations.

So while parsimony is simple to understand as a heuristic, relatively quick to implement and compute and will do a good job of reconstruction when substitutions are rare, it does not explain the process of sequence evolution well (no physical model of the process) and is prone to failing when there are hidden and multiple substations, especially when there are long branches (highly diverged sequences) in the tree. A further problem is that the maximum parsimony tree is just a single tree that contains no information about

uncertainty — we aren't sure which splits in the tree we are certain about or which splits could be rearranged to produce an equally likely (or very near to equal) tree.

We thus seek a method of reconstructing a tree that is based on statistical principles, one that will find the most likely tree taking into account a model of the process that gave rise to the data. The method should also give us an idea of the uncertainty in the reconstruction. Indeed, we will look at methods that provide us with many different possible trees that all represent feasible reconstructions of the evolutionary relationships between sequences.

21 Statistical approaches to modelling evolution

Distance and parsimony based methods for tree reconstruction are based on a number of assumptions that often do not hold. Distance methods are simple and fast to implement but are only guaranteed to reconstruct the correct tree under very restrictive circumstances. Parsimony is not based on a realistic model of evolution and, as we have seen, is statistically inconsistent (it reconstructs the wrong tree even with infinite data).

Our approach then will be similar to the approach we have taken in other parts of the course: we will model the process of sequence evolution, and based on that model we will write down the likelihood of a tree. We will then seek to find the maximum likelihood tree and finally look at Bayesian approaches to finding the best tree.

We model only the *substitution process* in which one base is replaced by another, for example $A \to T$ or $A \to C$. We will ignore the (more complicated) processes of insertions, deletions, recombination etc.

21.1 Likelihood of a given tree

Consider a tree with four leaves and sequence at each leaf consisting of a single site. An example of such a tree with four sequences, labeled A, B, C, and D is shown in Figure 17. The values of the sequences are C, C, T and T, respectively.

The maximum parsimony tree for these sequences groups A and B together and has a parsimony score of 1. But how likely is it? Inherent in the parsimony idea is that only one mutation occurred on the tree and it must have occurred along the branch between the two ancestral nodes. That mutation was from a C to a T (or vice-versa) implying that the unknown ancestral values at the ancestral nodes were also C and T, as shown on the left in Figure 17.

Ideally, we would account for all possibilities for the ancestral values: they could be any of $(A, A), (A, C), \ldots (T, T)$. All possibilities are shown in Figure 18. That is, if the unknown ancestral states are X and Y, then we could look at the likelihood of the tree for each possible combination of X and Y and sum these together. This is the principle of marginalisation introduced in Section 11.4: we want to know the probability of the tree and the data, but have some other random variables floating about too (the ancestral

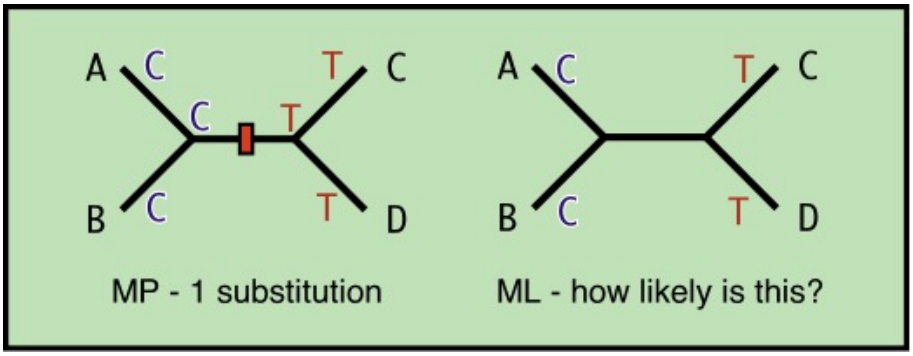

Figure 17: The parsimony tree on the left with the ancestral states reconstructed. Under parsimony, this is considered to be the one true tree. Under likelihood methods, we want to decide how likely the tree is given the data (observed values at the leaves). That requires summing over all possible ancestral values (shown in 18).

states X and Y) which we deal with by simply summing over all possible values to get

$$P(\text{Tree and data}) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(\text{Tree and data}, x, y)$$

It turns out that once we have a tree with values for the site known at all nodes (not just the leaves), we can calculate the likelihood with relative ease. That is, we know how to calculate $P(\text{Tree and data}, x, y)$ for any value of x and y, a sketch of which is given in Figure 19.

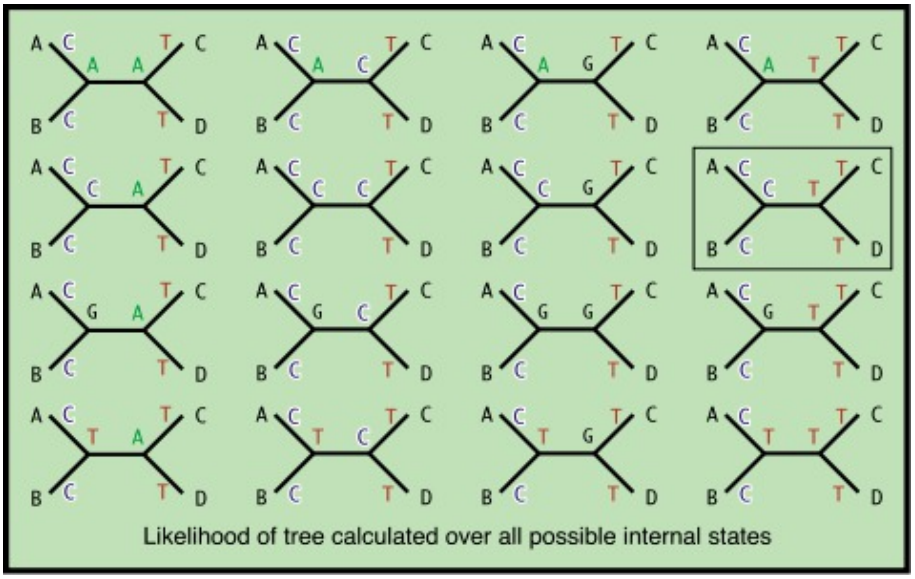

Figure 18: All possible ancestral value for the tree considered in Figure 17. The likelihood of the tree is the sum of the likelihoods of the tree and the ancestral values, where the sum is over all ancestral values. That is, to find the likelihood of the tree in on the right in Figure 17, we need to find the likelihood of each of the trees in this figure and sum them up.

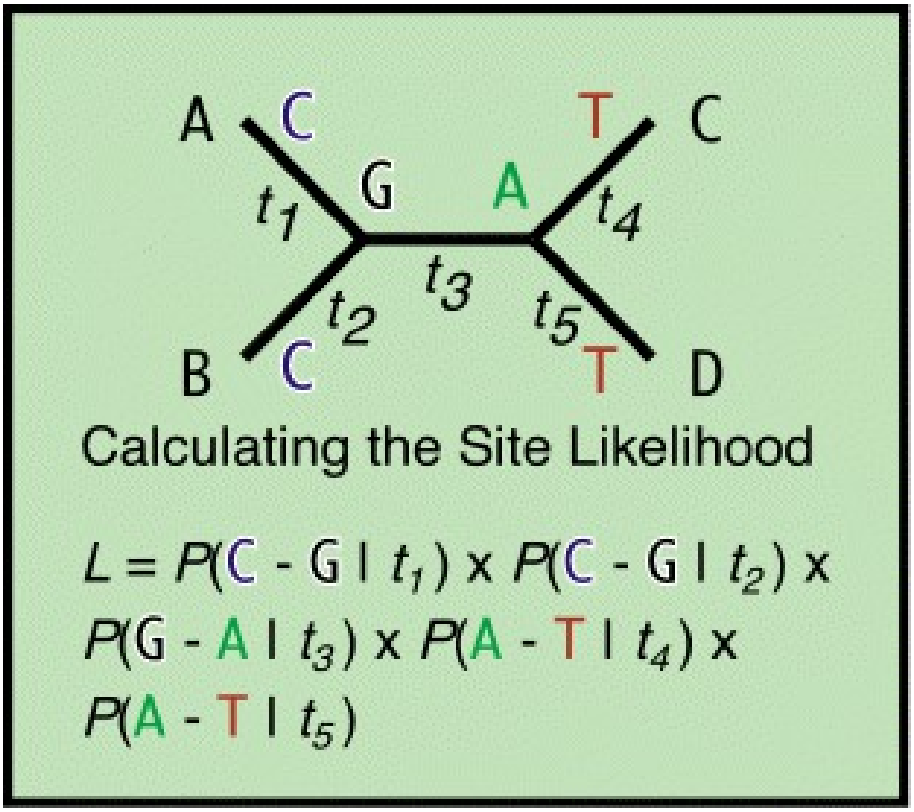

Figure 19: The likelihood for a tree with data at the leaves and imputed ancestral data is given by a product of the probabilities of mutating between the different values along each branch: that is the probability of mutating from C to G along the branch of length t_1 multiplied by the probability of mutating from C to G along the branch of length t_2 multiplied by the probability of mutating from G to A along the branch of length t_3 and so on. How these probabilities are calculated in developed in Section 21.2