Reading for Lecture 16

Release v10

Christopher Lee

November 26, 2011

Contents

1 Monophyletic Groups

In a rooted tree, we define the concept of a *monophyletic group* as a subset of sequences in the tree that represents *all descendants* of the most recent common ancestor (MRCA) of that subset of sequences. First, let's consider how a group of sequences could *fail* this test. Say we had a species tree consisting of chimp, human, monkey and mouse. The group (chimp, human, mouse) is not a monophyletic group, because its MRCA is close to the ancestor of all placental mammals, and monkey is a descendant of that MRCA. By contrast, the groups (chimp, human) or (chimp, human, monkey) are valid monophyletic groups on this tree. Note that the validity of a monophyletic group is defined strictly in terms of the set of sequences in that tree; if we were to add even one more sequence to the tree, that could change many of its monophyletic groups.

1.1 Consensus Monophyletic Trees

Monophyletic groups are of interest because they provide a useful way of generalizing phylogenetic tree structure to take into account uncertainty about some details of the tree structure. Say we have three species A, B, C that are more closely related to each other than to the rest of the tree. If we had strong evidence that A and B are neighbors (e.g. the distance from C to A or B is much greater than the distance between A and B) then we can draw them in the usual binary tree format, with an internal node directly connecting A and B, and a more distant internal node connecting that pair to C and the rest of the tree. On the other hand, what if A, B and C were approximately equidistant, so there is no clear "branching order" among them? In this case we need a representation that on the one hand does not imply a branching order among A, B and C (because that is not known) but on the other hand does clearly indicate that they are more closely related to each other than to the rest of tree. In short, we want the tree structure to represent what we know (are confident of) vs. what we do not know (are uncertain about).

One way to do this would be to draw the *monophyletic group structure* of the tree. For example, in this case we are confident that A, B and C are a monophyletic group, but we are not sure of any monophyletic subgroups within that group. So we can draw this monophyletic group as a single internal node M_{ABC} , with edges to all three sequences A, B, C. Note that the only difference versus a standard binary tree is that internal nodes (which represent monophyletic groups) can have *more* than two child nodes. By contrast, if we were also confident that (A,B) formed a monophyletic group M_{AB}, we would draw an edge from $M_{ABC} \rightarrow M_{AB}$, and edges from $M_{AB} \rightarrow A, B$ (and an edge from $M_{ABC} \rightarrow C$).

We can formalize this intuitive idea as follows:

- we define a *consensus monophyletic group* as a monophyletic group whose posterior probability given the observations is greater than 50%. (Of course, we can choose a higher confidence threshold e.g. 95% for what groups we consider "confident enough" to show).
- We define a *consensus monophyletic tree* as a tree whose nodes are consensus monophyletic groups, and whose edges represent nested containment relations among these monophyletic groups. E.g. in the example above, (A,B) is a subset of (A,B,C), so we would draw a directed edge from $M_{ABC} \rightarrow M_{AB}$.
- It is straightforward to prove that the set of consensus monophyletic groups for a given set of sequences themselves form a tree by the above criterion. That is, for any two consensus monophyletic groups in a given set of sequences, it is guaranteed that if they have any overlap (i.e. one or more sequences in common) then one is a strict subset of the other.
- Note that each internal node (consensus monophyletic group) represents the MRCA of that monophyletic group.

1.2 Computing Confidence Probabilities on Trees

In the previous section we took for granted that we could obtain posterior probabilities for individual "branches" in the phylogeny of a set of sequences. We now briefly consider one common way of computing such probabilities.

Bootstrap estimation: sampling with replacement

Bootstrapping is a common method in statistics for estimating the variance in a parameter of interest based on a sample. The idea is that the sample itself provides information about this variance. Specifically, rather than just computing a mean or Maximum Likelihood Estimate of the parameter from the whole sample, we can "re-sample" the sample via a procedure called sampling with replacement, and thus obtain many samples. We estimate our parameter from each sample, and in this way obtain a distribution of values for the parameter. Concretely, for tree construction, we have a set of m columns representing the alignment of our sequences. We treat this as our original sample. We generate a new sample by drawing m columns from the original sample as follows:

- we choose one column at random from the original sample;
- we add it to our new sample;
- we "return" it to our original sample, i.e. so that it is just as likely to be chosen again as it was to be chosen in this draw.
- we repeate this procedure m times to produce a new sample of the same size as (but different composition than) the original sample.

Note that this "sampling with replacement" protocol can choose the same column more than once; equally well, some columns that were present in the original sample may not be chosen at all for inclusion in the new sample. Thus the new sample will have a different composition (and different parameter estimate) than the original sample. For phylogeny analysis, this means computing pairwise distances based on the alignment columns in the new sample, and building a new tree from these distances. Note that both the distances and the tree are likely to be (somewhat) different from what we obtained from the original sample; the question is *how different*.

We can generate as many of these "bootstrap samples" as we wish (limited only by compute time) and in this way generate a bootstrap sample distribution of our parameter of interest. For our example sequences A, B, C, the question is how frequently $\delta(A,B) < \delta(A,C)$ (which implies that A and B are neighbors, whereas $\delta(A,B) > \delta(A,C)$ implies A and C are neighbors). If many columns in the alignment indicate that A and B are more similar to each other than C is, then every bootstrap sample is likely to yield $\delta(A,B) < \delta(A,C)$ (and therefore will place A and B as neighbors). On the other hand, if the actual data are close to a "tie" (that is, only slightly more columns favor A and B as more similar than favor A and C as more similar), then the bootstrap samples will reflect this; i.e. only slightly more of the samples will yield $\delta(A,B) < \delta(A,C)$ than yield $\delta(A,B) > \delta(A,C)$.

We can easily derive the probability that A and B are a monophyletic group from the bootstrap data, by simply counting what fraction of the bootstrap samples yielded trees in which A and B were a monophyletic group. In the former case above, that would give a bootstrap confidence of close to 100%; in the latter case it might be an approximately equal split between grouping (A,B) vs. (A,C) vs. (B,C) (which would give each monophyletic group only about 1/3 probability).

It should be noted that bootstrap probabilities are not true posterior probabilities, i.e. they are not guaranteed to equal what one would obtain from a comprehensive Bayesian posterior probability calculation. However, bootstrapping is commonly used to estimate which features of a tree are robust (confident) versus which are uncertain given the inherent limitations of the data.

2 Unrooted Tree "Groups"

Since monophyletic trees are defined in terms of the MRCA, they depend on knowing the directionality of each edge, and are only applicable to rooted trees. Can a similar idea by applied to unrooted trees? Yes. We make the following analogy. In a rooted tree, a given internal node (MRCA) splits its descendants (its subtree) from the rest of the tree (not its descendants). In an unrooted tree, an internal node provides no such clear "cut" of the tree, because all three of its edges are undirected; we do not know which two to group together as we did in the rooted tree. However, in an unrooted tree, any internal edge creates a similar "cut" of the tree into two disjoint subtrees. For example, for a tree of four sequences there is only one internal edge, and it splits the tree into two pairs of neighbors.

The cut associated with any internal edges guarantees the following property. The edge divides the tree into two groups of sequences. For any pair A,B drawn from the first group, and any pair C,D drawn from the second group, the sum of the distances *within* the two groups is guaranteed to be less than the sum of the distances *between* the two groups, i.e. $\delta(A,B) + \delta(C,D) < \delta(A,C) + \delta(B,D)$. (This is just the four-point condition).

2.1 Consensus Groups on Unrooted Trees

Once again, we can apply posterior probability criteria to filter out tree structures that fall below some specified level of confidence. E.g. following the bootstrapping approach outlined above for monophyletic groups, we could generate a large number of bootstrap samples and for each possible for each possible "cut" (i.e. split of the sequences into two separate groups), assess what fraction of them support that cut (i.e. satisfy the distance inequality above).

- We define a *consensus cut* as a cut whose posterior probability is greater than 50% (e.g. is observed in more than half of the bootstrap samples). Similarly, we can set any desired confidence threshold for what cuts to report e.g. 95% probability.
- We represent such a consensus cut as an undirected edge, whose vertices (two ends) correspond to the two groups of sequences that it separates.
- Note that consensus cuts for a given set of sequences form an unrooted tree (just as the consensus monophyletic groups form a rooted tree). And just as consensus monophyletic trees are no longer guaranteed to be binary trees, a consensus cut edge may have more than two "subtree" consensus cut edges connected to it. (For example, a consensus edge might have a group of five sequences, and if the grouping relationships among those five

sequences are not clear enough to meet our confidence threshold, all five sequences will be directly connected to one end (vertex) of this consensus cut edge.