Lab 14: Tree Comparisons

Today we are going to look at ways of comparing trees that may have a different topology. First we're going to mathematically compare a bunch of trees with the same taxa but different topology using *Mesquite* and then generate several types of consensus trees. Next we're going to generate a consensus tree from trees that have overlapping but not identical taxa using Matrix Representation with Parsimony. Last we're going to use Type I Brook's Parsimony to generate a tree of geographic areas from cladograms of taxa living in those areas.

Tree Distances

First we're going to use *Mesquite* to generate a number of different tree distance measures on a bunch of trees with the same taxa, like we talked about in class. The **MrBayes Cephalopod Tprbs** file contains the first 13 trees from the tprobs file of a cephalopod dataset. This has the highest 50% of trees that I found during stationarity. It also has the estimated posterior probabilities of those trees.

Open the **MrBayes Cephalopod Tprbs** file in *Mesquite*. Open the tree window and page through the different trees. Do you see the differences in topology? Which trees look most similar?

To compare these trees to themselves we have to open a second tree window with the same set of trees. I know that seams crazy, but that's just the way it is. Go to **Taxa&Trees> New Tree Window**, then select **Use Trees from Separate NEXUS file**, and choose the **MrBayes Cephalopod Tprbs** file again. You should now have two tree windows with the same set of trees.

From **Tree Window 2** (make sure that you are in **Tree Window 2**) select **Analysis > Values for Current Tree**, and then choose **Compare with other trees**. This is the third option from the bottom. Compare these to **Current tree**. We'll start with **Shared partitions**. Now a box will appear saying that it wants to use the tree in tree window 2 as the current tree for comparison. We don't want that. This would make the comparison between the tree and itself, so select **No**. We want to compare it to the trees in T**ree window 1.**

A box will appear showing the number of partitions that the trees in our two tree windows share. A branch can be viewed as a partition, because it separates the taxa into two groups. So if branches in both trees separate the taxa into the same two groups, then that partition is shared. These are the same as shared clades on an unrooted tree. Thus more similar trees will disagree on fewer partitions and so have more shared partitions. These trees have 12 internal branches, so if two trees are identical, then they will have 12 shared partitions.

Use the swishy arrow at the top of the tree window tab to separate these trees out into different windows, so that you can see both trees at once. Page through the different trees in both windows, so that you can look at every possible pair of trees. Are the trees with higher posterior probabilities more like the tree with the highest posterior

probability? (Remember these trees are listed in the order from the highest posterior probability to the lowest.) Is the tree with the highest posterior probability more similar to the other trees than they are in general to each other? Why would this be? What trees have the biggest differences?

Pull down the little arrow in the box showing the number of shared partitions and select **Tree-Tree value > Patristic Distance correlation**. The patristic distance between two taxa is the sum of the branch lengths between them. This value is the correlation coefficient of the patristic distances between the two different trees. Thus more similar trees should have a higher correlation, with 1 being the maximum for identical trees. How do the trees compare under this measure of difference? Do the two metrics agree on the relative distance between trees? Which metric is more informative?

Question 1. What two trees are the most different under the Patristic distance correlation? Are these two trees also among the most different for the shared partitions?

If you are interested in comparing a large number of trees, I would recommend that you look into the ***TSV package*** for *Mesquite*. It has dome useful ways of visualizing different trees.

Consensus Trees

Now we're going to generate several different consensus trees from that same set of trees using *Mesquite*. I want to emphasize that this is not the appropriate way to generate a consensus tree from *MrBayes*. It is much better to use the **sumt** command in *MrBayes*, because that will consider the trees based on their estimated posterior probabilities and will also calculate branch lengths. However, there are many other situations when you would want to use this method, such as if you generate several most parsimonious trees. I'm just using this tree file, because it is convenient.

Pull down **Taxa&Trees >New Tree Window** and select **Consensus Tree**. Use **Stored Trees**. First let's generate a **Strict Consensus**. Select it then hit **OK**. This will output a tree that only contains the nodes that are present in all your input trees. If you want to save the consensus trees or any tree for that matter, you have to store the trees first by using **Tree>Store Tree** and then save the file. There is no need to do that right now, but it may be important for you in the future.

Now let's generate a **Majority Rule** tree with a cut off at 50%. This will output a tree with all the nodes that appear in more than 50% of the tree. It will also tell you in what percentage of those trees the nodes occurred. Does this have the same topology as the strict consensus? Which tree is better resolved? Why? Are the two trees compatible?

You can generate **Majority Rule** trees with cut offs greater than 50%, so that more clades are eliminated. The cut off point is always kind of arbitrary, but can not be less than 50%. If it were less than 50%, then you couldn't be sure that all the clades are compatible. How high would the cut off have to be to guarantee that you are going to get the same tree as strict consensus?

Matrix Representation with Parsimony (MRP)

So it's easy to generate consensus trees if they all have exactly the same taxa, but what do you do if all the trees have different taxa? For example how would you put together a bunch of trees from different studies with overlapping but not identical taxa? Well, it is a matter of big debate. Maybe you shouldn't even do it at all. Maybe it is best to take the data matrices from all those studies and combine them into one supermatrix for analysis. If you do decide to combine trees, it is not at all clear what the best method is. The mostly commonly used method is Matrix Representation with Parsimony (MRP). Here we're going to do a made up simple example of it.

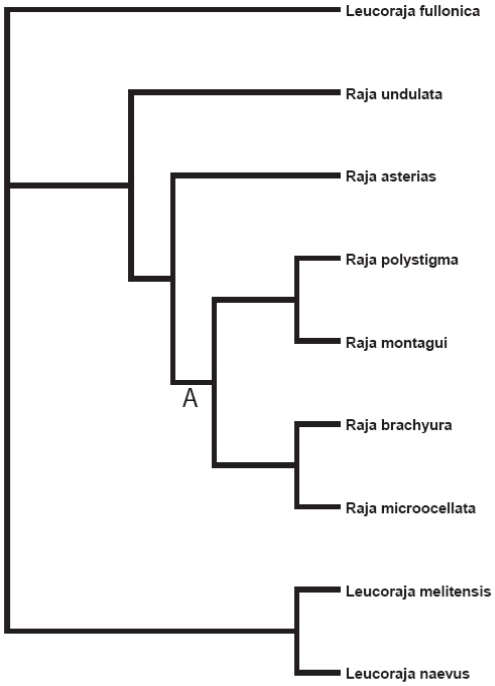
We are going to do MRP on the three trees of rays that you will find on the next page. I just took a single data set of rays, randomly deleted two taxa from it three times, and used those reduced data sets to make trees by parsimony. This is a totally unrealistic situation for several reasons. The taxa have a lot of overlap, whereas three trees picked from the literature would have very little overlap. Thus a real MRP matrix would have a lot of question marks. Also the trees are all generated from the same data set, so that you know that you won't have a contradictory signal from two different data sets, which you are likely to have in reality. However, I didn't have time to find a more realistic set of trees, and these will make filling out the matrix easier.

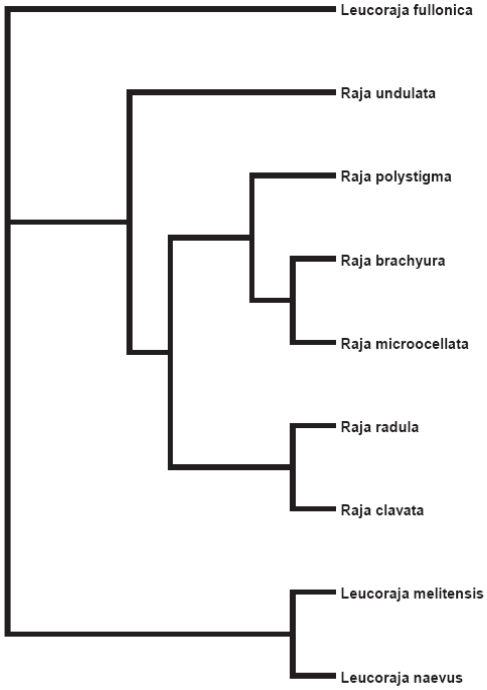
Open the **Ray matrix** file in *Mesquite*.. This is just an empty matrix with the taxa names on it (and apparently lots of spelling mistakes), so that you don't have to bother filling them all in. You are going to have to fill in the data.

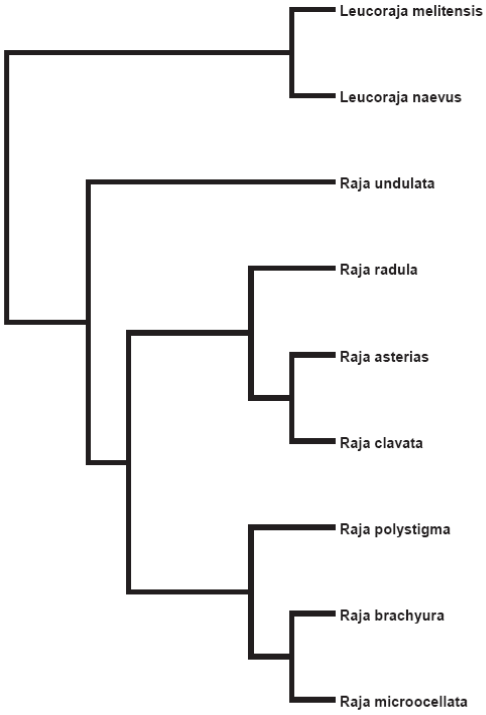
Now code the three trees into the matrix. You do this by treating each interior branch from each tree as a separate character. Remember that every branch separates the taxa into two groups, one on each side of the branch. You can assign each of these partitions a separate character state, so that all the taxa on one side of a branch get a **0** and the other side a **1** for that character. All the taxa that don't appear in that tree should get a **?**. Every tree should have 6 internal branches (There is no node at the root). For example the branch that I marked as **A** in the first tree should be coded:

Raja polystigma	**1**
Raja montagui	**1**
Raja brachyura	**1**
Raja microocellata	**1**
Raja asterias	**0**
Raja undulata	**0**
Raja radula	**?**
Raja clavata	**?**
Leucoraja meitensis	**0**
Leucoraja naevus	**0**
Leucoraja fullonica	**0**

When you're done with the matrix save it. We will now use *Mesquite* to find a most parsimonious tree from this matrix. *Mesquite* is not the best program to do this in. I would recommend using *PAUP** or *phylip*, which is free, but nobody wants to bother with downloading a new program now.. Select **Taxa&Trees >Make New Trees Block**







From>Tree Search>Heuristic, then choose **Tree Length**. Both the next two defaults are fine; run the search on a separate thread. You probably got multiple trees. Is there any difference between these trees? To find out root all the trees in the same place and compare them. They should all be the same. The fact that *Mesquite* made so many trees is an example of why you shouldn't use *Mesquite* for tree searching.

Question 2. Trace the characters on this tree using parsimony. Is this tree homoplasious for any of these characters? Why or why not? What would homoplasy in an MRP analysis represent?

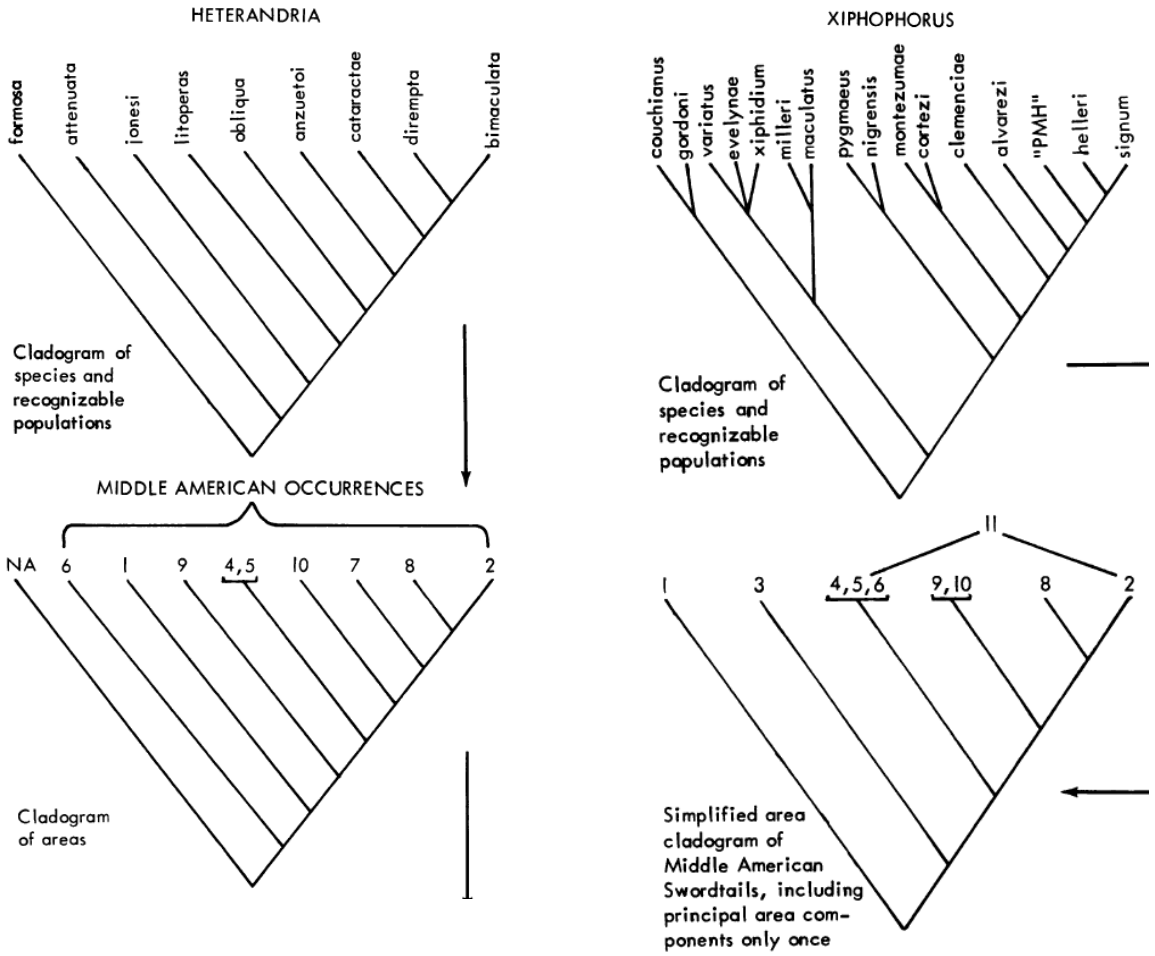
Brook's Parsimony Type I

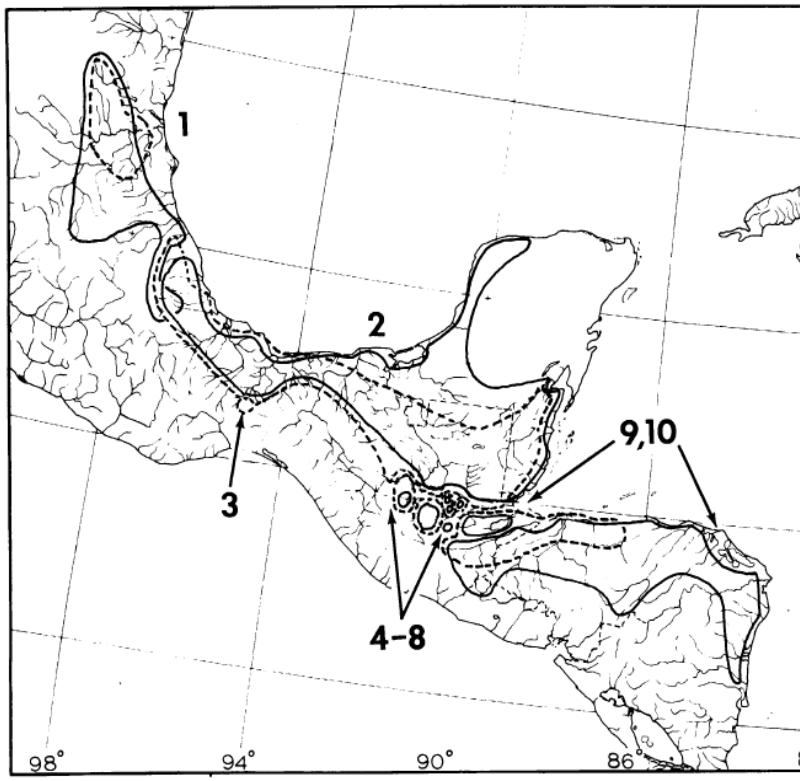
It is also possible to derive a tree from trees of a number of different associated taxa. This falls into a general category of problems in which you have a container tree (for example of geographic areas or hosts) and one or many associated trees (for example taxa living in those areas or parasites). We will learn more about comparing associated trees in the biogeography lab on Thursday.

Brook's Parsimony is one method for making these comparisons. It basically codes the trees for the associated taxa in the same manner as MRP, while taking the root into consideration. Thus it assumes that any two associated taxa found in two different containers, had a common ancestor found in the ancestor of the two containers. In Type I Brook's Parsimony a tree for the container taxa is derived using this coding of an associated taxon tree as characters in a parsimony analysis with other types of characters and/or codings of trees from other associated taxa. In Type II Brook's parsimony a known container tree is compared to a known associate tree by using by using parsimony to reconstruct ancestral states. We will do type I today and type II on Thursday.

We are going to use a real example from Rosen, 1979. Fishes from the uplands of intermontane basins of Guatemala: revisionary studies and comparative biogeography. Bulletin of the American Museum of Natural History 162: 267-376. The cladograms on the next page already have their branches labeled and show the geographical distributions of the terminal taxa. Use this information to create another data matrix in *Mesquite*. The taxa for the data matrix should be the geographical areas, and the characters the branches on the cladogram (including the root). If a branch from one of the cladograms has any descendants that live in a geographic area, then that geographic area should have state 1 for that branch. If a branch has no descendants living in that area, then it should be a 0. Note that there is no need to count autapomorphic characters.

Question 3. Analyze this data set in *Mesquite* using parsimony. What kind of biogeographic conclusions can you draw?


HETERANDRIA
formosa
attenuata
jonesi
litoperas
obliqua
anzuetoi
cataractae
dirempta
bimaculata
Cladogram of species and recognizable populations
MIDDLE AMERICAN OCCURRENCES
NA
6
1
9
4,5
10
7
8
2
Cladogram of areas
XIPHOPHORUS
couchianus
gordani
variatus
evelynae
xiphidium
milleri
maculatus
pygmaeus
nigrensis
montezumae
cortezi
clemenciae
alvarezi
"PMH"
helleri
signum
Cladogram of species and recognizable populations
II
1
3
4,5,6
9,10
8
2
Simplified area cladogram of Middle American Swordtails, including principal area components only once


1
2
3
9,10
4-8
98°
94°
90°
86°