Lab 17: Testing for Clade Imbalance

Introduction

Today we're going to look at several different ways of testing clade imbalance. *Mesquite* has two limitations on its ability to test for this property. It can only simulate trees that have the same number of taxa as are in your character matrix and it has no statistic to compare the bottom two nodes. Therefore, we will do what we can in *Mesquite* and then we will turn to R and see what possibilities can be found there.

Today we will explore clade imbalance and diversification using Mesquite and R packages. The goals of the lab are to:

I. Calculate possible clade imbalance of one tree using Mesquite
II. Evaluate differences between sister clades
III. Simulate some data in R

Colless's Imbalance

Colless (1982) proposed a way of measuring imbalance in a tree. It does not compare two clades but instead the overall imbalance throughout the tree. It is calculated as:

$$\frac{\sum_{i=1}^{nodes} |n_{li} - n_{ri}|}{(n-1)(n-2)/2}$$

Where n_{li} and n_{ri} are the number of taxa descended from the left and right branches of node *i* respectively, and n is the total number of taxa in the tree. *(n-1)(n-2)/2* is the maximum possible value of the sum, so that this statistic runs from 0 to 1 with 0 being perfectly even (balanced) and 1 completely lateralized (unbalanced).

Download the *Colless_example.nex* file from the IB200 website. In Mesquite, open the stored tree and calculate Colless's imbalance by hand. Move up the tree and add up the difference between the right and left clades at each node, then divide by the denominator.

Is this value statistically significant? Let's see. To determine significance, we have to use simulations to compare it to values from a null distribution of trees to determine if it is. Select **Analysis > New Bar & Line Chart for > Trees, Simulated Trees**. You will now be offered several different simulation options. We will discuss what these options are below. Essentially each option represents a different null distribution of possible trees. Let's start with **Uniform Speciation (Yule)**. 10 is good for the tree depth. In the **Value to calculate for trees** window select **Show secondary choices** and **Coless's Imbalance** (scan down the list to find it). Let's do **999** simulations for a little power. That's a pretty chart, but we want a p-value. Use the **text** tab to get actual counts and use these numbers to calculate a p-value. The p-value is the number of trees with a Colless's imbalance equal to or higher (if you want to show that the tree

is particularly imbalanced) than your tree. You'll once again need to calculate this by hand - is it significant? Repeat this analysis, but this time use **equiprobable trees** for your null distribution.

Question #1: **What was your p-value using Uniform Speciation for your simulated trees? What was the p-value using Equiprobable trees? What effect did the null distribution have on your p-values? How would you interpret this p-value? (ie. what value are you measuring for significance here?)**

Clade Imbalance

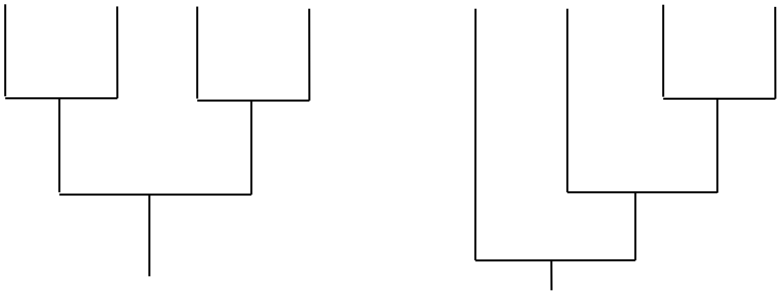
Now we're going to determine whether two sister clades are of statistically different sizes. This will be a comparison for a single tree in which one clade is larger and has n taxa and the other clade has m (<n) taxa. We will be testing weather m is significantly different from n using a number of different null distributions.

Something to keep in mind throughout this lab is whether to do a one-tailed or a two-tailed test. A one-tailed test would be appropriate if you had a hypothesis that clades with a certain character (ie: environment or morphology) should have more taxa than clades which do not have that character, and you are comparing a pair of sister clades in which one has the character and the other does not. However, in most situations you will first have identified that one clade is larger than the other, and after the fact you will hypothesize a reason why. In this case a two-tailed test is appropriate.

Equiprobable Trees

One possibility is that every single labeled topology is equally likely. There is an important distinction between labeled and unlabeled topologies. Unlabeled topologies are just the trees without any taxa. While labeled topologies have the taxa assigned to the branches. Therefore, two unlabeled topologies can have different numbers of labeled topologies associated with them, and thus have different probabilities under an equiprobable trees null distribution.

For example consider the two following unlabeled topologies for four taxa. How many possible labeled topologies can you count on each one? Remember that rotating a branch does not change the topology.



One consequence of this is that imbalanced unlabeled topologies have more possible labeled topologies associated with them than balanced ones do. Thus your p-value for imbalance from a labeled distribution will be lowered relative to one that uses only unlabeled topologies.

Let's make some simulations in *Mesquite* to test for … In the tree window select **Taxa&Trees > Make New Trees Block From>Simulated Trees>Equiprobable**. Let's make 99. You should really make more for statistical power, but you are going to have to go over them by hand. When it offers, open the tree window.

Now open a new Excel spread sheet and for each tree record in column A the number of taxa in the left clade, and record the number of taxa in the right clade in column B. If this takes forever, then screw it, but I think that it should be pretty fast.

To calculate your one-tailed p-value, in the column C subtract your value in column B from your value in column A. Now reorder all your trees according to their value in column C. Go to **Data > Sort, Expand Selection, Sort By : Column C**. You can now calculate your p-values by counting (really just use the row numbers) the number of trees with a difference greater than or equal to the tree you are comparing them to (in this case 7-2=5). Divide this number by 100 (99+1).

You can calculate your two-tailed by adding the trees counted above to those with a difference less than or equal to -5 (=2-7). Record both the one-tailed and two-tailed values. Is your two-tailed p-value twice your one-tailed? In this case any difference arises from random effects on small sample size. Therefore your 2-tailed prediction is probably better, since it is drawn from more trees.

You can now shut down *Mesquite*.

Download ib200_lab17_simulated_trees.R

Open this file in R or R Studio and work through it. There is one question at the end.