Integrative Biology 200B University of California, Berkeley *Spring 2011*
"Ecology and Evolution" *DD Ackerly and NM Hallinan*
Updated by Nick Matzke

Lab 2a: Null Models of Character Distribution

This lab is divided into two sections. In the first part we will learn how to detect phylogenetic signal in a discrete character using *Mesquite*. The basic principle will be to see if there are significantly less steps in the evolution of our character on our tree, than you would get for some random character that is not associated with the phylogeny. We will explore several different null models of character distribution, and try to get a feel for how tree shape effects the number of steps in the evolution of a character. If there is no signal, then the value of doing phylogenetic tests is questionable.

In the second part of the lab we will look at ways of detecting correlation between two discrete characters.

Open the **Dess200B.nx** file in *Mesquite.* Examine the first tree and the reconstruction of the character desiccation tolerance – make sure you understand why the parsimony reconstruction comes out as it does. Open the tree window. To keep things simple, resolve the polytomy in the seed plants (pick any resolution)

Null 2 - Calculate the distribution of steps under null trees.

The first thing that we need to do is calculate the number of steps in this character on our tree. Select **Analysis>Trace character history** and then select **Parsimony ancestral states**. Record the number of steps in our character. Now we will compare our character to characters randomly distributed over random trees. Select **Analysis > New Bar & Line Chart For > Trees.** This will create a chart where the y-axis is number of trees.

Now you will be given several different options for how to generate trees, we want to simulate trees. We will deal with the different random tree null distributions more in a few weeks. The **Uniform speciation (Yule)** uses a random branching process, in which every branch has an equal probability of duplicating at any time. It only produces trees with the same number of taxa as you have in your character matrix. Pick the **Yule distribution** to start. 10.0 is cool for the tree depth.

Select parsimony **steps in character** (this will be your x-axis). Let's make 999 trees to start. Choose the only available character. You will see a histogram of how many steps in your character the simulated trees have. That looks nice and all, but you want real numbers. Click the **text** tab. Now you can see exactly how many simulations had each number of steps.

Question 1. What is your p-value for the number of steps on your actual tree? Don't forget to include your original tree in the calculation.

Save the outputs to file so you can compare the three different random tree generation algorithms. Select **File > Save window as text** and save the results to a new folder on your desktop. You can change your simulation method by going to **Chart > Tree Simulator**. Try **Equiprobable Trees** (all trees have the same probability) and **Uniform Speciation with sampling** (same as the yule process, except you create a tree with more taxa and then randomly sample from those), save your results as text and calculate a p-value. How does your simulation method affect your p-value? Which null distribution makes the most sense to use?

What is the minimum and maximum number of steps for this character? Create trees by hand showing the two extremes.

Null 1 – Calculate the distribution of steps for shuffled character

Now we will compare our character to a character with its states randomly assigned to the tips of the same tree. Remember that a particular tree topology may inherently tend to generate more or less steps. Therefore a significant signal over random trees may reflect an effect of our phylogeny, rather than an effect of our character. Randomizing the data over the same tree will eliminate this potential complication.

First let's just create one character and look at it. In the **Character matrix** window create a new character using the **add character** tool. Now switch to the **select** tool, copy your original row and paste it onto your new character. Select your new character and then go to **matrix > alter/transform > shuffle states among taxa**. Trace your new character on the tree. You can do that with just the arrow. Make sure that you use your original tree and not one of the trees that you just created. How many steps does this character have?

So, we could do that a thousand times, or we could have *Mesquite* do it for us all at once. First delete the new character that you just made (**Edit > cut**), so that it doesn't confuse things. Go to **analysis > new bar and line chart for > characters**. What will this do? . Select **character value with current tree- parsimony character steps.** So, that's not really what we wanted, it just plotted the two characters from our matrix. To create a bunch of randomized characters select **Chart>Source of characters>Reshuffle character**. Choose your original character (not that it matters in this case). Let's do 999 characters again (well it was trees last time). Save this as text and calculate a p-value. How does this p-value compare to the random trees?

How does tree shape affect the distribution of steps for randomized traits?

Manually create an extreme tree shape. The number of steps in the shuffled characters of your histogram will change automatically. Compare the parsimony steps for shuffled characters on this tree and calculate a p-value using the steps on our original tree. Do it again with a tree shape from the opposite extreme. How does this influence the distribution under the null? Do you have an intuition for the effect of tree shape?

Question 2. Describe one of your extreme shapes in words and provide the p-value.

Integrative Biology 200B *University of California, Berkeley*
"Ecology and Evolution" *Spring 2011*
Updated by Nick Matzke

Lab 2b: Correlated Changes in Discrete Characters

Today we are going to look at several different statistical methods for determining if an apparent correlation between two discrete characters on a phylogeny is significant. All of these methods are available in *Mesquite*. These analyses can be done with other programs and there are other types of analyses that seek to answer the same question, but for now we'll stick with the programs that we will use on a regular basis.

The essential question at hand here is weather a character is more likely to change into a particular state in lineages that are in a particular state for another character, but they all take different approaches. Pagel's test looks at changes across the whole phylogeny, while Pairwise Comparisons breaks a phylogeny down into a number of comparisons between pairs of taxa. On the other hand, Pairwise Comparisons rely on looking for significance by counting changes in characters, while Pagel's test uses comparison of likelihood scores for the same purpose. Both these methods will be described in more detail below. A third method, Maddison's test, uses changes in number of characters, but makes the calculation over the whole phylogeny. This test is available in *MacClade*, but not *Mesquite*, so we won't be using it today.

We will do all these analyses on the **Concentrated Changes** example data set from the the *Mesquite* examples. Open this file in *Mesquite*. Go to the **Show State Names Editor Window** and rename each of the two characters and states as follows: Character 1 = **dependent character, Fruit Type**, state 0 = **ancestral, fleshy**, state 1 = **derived, dry**; Character 2 = **independent character, Ecology**, state 0 = **tropical rainforest**, state 1 = **open savanna**. Save your work (but only this one time).

Comparing Reconstructions Visually

The first thing that we will do is look at the distribution of the two characters against each other in one window, so that you might more clearly see their differences. Open the **Tree Window**. Go to **Tree>mirror tree window**. A new window will appear with the trees facing each other. Expand the window so that you can see what's going on.

First let's compare parsimony reconstructions of our two characters. Select **Mirror> left side >trace character history**, and select **Parsimony ancestral states.** Do the same for the right tree. You will see character 1 traced in both trees. Use the **Trace character** box for the right tree to switch to character 2. Now you can clearly see that Character 1 only switches from state 0 to state 1 in lineages that are in state 1 for character 2.

You can also modify the reconstructions for each of these trees. For example let's look at the likelihood reconstructions. You will see separate trace menus for each tree. For both trees go to **trace>Reconstruction Method>Likelihood Ancestral States**, and select **Current probability models**. It may be helpful if you set the **Tree Form** to **Balls and Sticks**.

You can set the Likelihood model to Mk2 by going to **trace>Probability Model>assym. 2 param.** Don't for get to do this for both trees. You can also use this method to compare different reconstructions of a single character. Why don't you compare the 2 parameter reconstruction of character 1 to the mk1 reconstruction?

Pairwise Comparisons

One possible approach for detecting correlation between two characters is to break the tree down into a number of non-overlapping series of branches that connect two taxa. You then ask how often the taxon with the higher state of the independent character also has the higher (positive) or lower (negative) state of the dependent character. Because the branches don't overlap you can just compare this to a standard nonparametric distribution and yet be certain that any correlations are independent of the phylogeny. This can also work for discrete characters by looking at the number of times one particular state for one character will be associated with one particular state of another character while the other taxon of each pair has the opposite set of character states.

There are a number of different possible pairs for any given tree. *Mesquite* gives you three ways to select your pairs. It can either pick sets of pairs that maximize the total number of pairs, those that maximize the total pairs with differences in one character or those that maximize the total pairs with differences in both characters. In any case it can come up with multiple possible pairs and you can pan through multiple options.

First open the original **tree window** and select **analysis > pairwise comparisons**. For our first set of comparisons let's just take the maximum possible number: select **most pairs** and hit **OK.**

Now the tree will appear greatly changed. All those colored lines are the different pairs of taxa. How many pairs of taxa are there? The numbers up above the tips are the character states for the characters you're comparing. Down in the bottom right are three boxes indicating the characters being compared and the choice of pairs. As you can see right now you are comparing character 1 as the independent character and character 2 as the dependent. That is the opposite of what you want, since it appears that the state of character 2 effects the state of character 1, not the other way around. Switch the characters to make the appropriate comparison.

The third box has the details of the comparison. There is only one possible set of pairs for the **most trees** criterion of pair choice. Below that is a summary of the comparisons: **Positive (Green)**: cases in which one of the taxa has a 1 for both characters and the other taxon has a 0 for both. (00 vs 11); **Negative(Red)**: cases in which one of the taxa has a 1 for one character and a 0 for the other and the other taxon has the opposite. (01 vs 10); **Neutral(Grey)**: cases in which the taxa disagree in the independent character, but have the same dependent character. (01 vs 11) or (00 vs 10); **Remainder(Blue)**: cases in which the independent character is the same for both taxa. The last thing in the box is the p-value, which is very low for this comparison.

Now let's try another set of pairs. Go to **Pairs > pair selector > Pairs contrasting in state for one character**. A **number of pairs** box will appear. This time the program found a bunch of different possible sets of pairs. In fact it found five which was its maximum and is now asking you if you want to find more. Set it to **20** or whatever you like and hit **OK**. How many pairs do you get this time? This set up is pretty much the same as before, except now you can toggle through the different sets of pairs and see their p-values (they are not all the same). At the bottom of this box you can see the range of p-values for all of the pairings that it is currently considering. Try increasing the max number of pairings more. Go to **pairs>max number of pairings** and up it to 100. Do some of these pairings have a lower p-value?

Check out **Pairs > pair selector > Pairs contrasting in state of two characters**. How does this analysis compare? As you can see your statistical significance depends a lot on what set of pairs you pick. How would that affect the reliability of your estimate of significance?

Pagel's Test

Pagel (1994) describes a method of looking for correlated changes in discreet characters using likelihood. Remember an assumption of most likelihood models is that the rate of change (relative to the branch lengths) is constant over the whole tree. Pagel takes as the null hypothesis that each character has a separate rate of change for forward and backward changes, like the independent asymmetric model we used last time. Thus there are four rates in the null hypothesis a forwards rate and a backwards rate for each character. This model is tested against one in which the rate of change for each character depends on the state that the other character is in. In this more complex model there are eight different rates, as each rate from the null hypothesis has been split into two.

To compare the two models all the rate parameters are fit to the tree and the data by maximum likelihood. The overall maximum likelihoods for each model are then compared. The null hypothesis is **nested** within the more complex eight rate model. This means that it is a specific case of the more complicated model, one in which the rates are the same regardless of the state for the other character (ie: rate for character 1 changing from state 0 to 1 when character 2 is in state 0 = rate for character 1 changing from state 0 to 1 when character 2 is in state 1). A model can never have a higher maximum likelihood than a model which it is nested within, because the more complicated model could always be fit to the specific case that defines the simpler model. Therefore, the question is not "which model has the highest likelihood," but "is the more complicated model's likelihood larger than the simpler model by enough to justify its use?"

To compare likelihoods for nested models we take the ratio between them. In many cases we can just compare two times the log of this ratio to a Chi-squared distribution to get a p-value. This is called a **likelihood ratio test (LRT)**. That will not work in this case, so instead data is repeatedly simulated on the tree using the parameter values derived from the simpler model. Maximum likelihoods are then generated from the simulated data using both models and a likelihood ratio is calculated. The likelihood ratio from the actual data is then compared to the distribution of likelihood ratios from the simulated data to generate a p-value.

Go back to the tree window select **Analysis > Correlation Analysis**. A box will open up asking for the number of iterations and the number of simulations. The number of iterations has to do with how many times it restarts the maximum likelihood search. The number of simulations is for calculating the p-value. Check the box marked **present p-value** and hit **OK**. In reality we would like to do many more simulations to accurately estimate the p-value, but we're keeping it short today. A new window will appear describing the programs progress through the simulation.

A **Correlation Analysis** box will appear to the right of your **tree window**. It will display the rate parameters calculated for both models as well as their log likelihoods. At the bottom you will find a p-value.

Question 3. What is this p-value?

There is also a **Correlation** menu that will allow you to change several things about the analysis. The most important thing is the ability to select character X and character Y, in case you are using a data set with multiple characters. Pagel's test could be set up with an independent and a dependent character, but the way it is set up in *Mesquite* the dependence of both characters is tested, so it doues not matter which is X and which is Y.

Change the character state in one of the taxa, so that there are now four changes of the dependent variable in clades that live on the Savanna. Rerun the test. How does effect the p-value? How does it effect the p-value of a Pairwise Comparison? Change that character state back and change the character state of the sister taxon of one of the dried fruit taxa to dry fruit, so that there are now four taxa with the dry fruit in the savanna, but still only three changes. Does this affect the likelihoods in the Pagel94 analysis? What effect would that change have on Pairwise Comparisons test? Why is there a difference? Which do you think is more realistic?