

Lab 5: 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 either *MacClade* or *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 whether a character is more likely to change into a particular state in lineages with a particular state for another character, but they all take different approaches. Maddison's test and Pagel's test both look 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, Maddison's test and Pairwise Comparisons rely on looking for significance by counting changes in characters, but Pagel's test uses comparison of likelihood scores for the same purpose. All these methods will be described in more detail below.

Maddison's Test

The concentrated changes test of Maddison (1990) is designed to test the association of changes in a binary character with some other binary variable within a clade of interest. It can test whether changes (from 0 to 1) in one (dependent) character are more concentrated than expected by chance on branches having a shared character state for another (independent) character. It does so by assuming that the number of changes is the same and randomizing those changes on the tree. The p value is the fraction of those randomizations in which the changes are at least as correlated with the independent character as they are in your data set.

Maddison's test can be run in *MacClade* but not *Mesquite*, so open *MacClade* and then open the file **Concentrated Changes** in the **Lab 5 examples** folder. Note that there are two characters included. The cladogram upon which these characters will be mapped has already been created, and we'll assume it is robust.

Go to the **data editor** and using **State Names and Symbols** option in the **characters** menu, 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).

Count the number of times dry fruits have evolved from fleshy according to this cladogram. Now examine their ecology and note that all instances of dry fruits occur in clades that inhabit the open savanna. This may suggest something about the causal influences of ecology or fruit type, but we must test this relationship for statistical significance. Are the changes in fruit type from fleshy to dry significantly concentrated in clades inhabiting the savanna? For Maddison's test the question now becomes: What is the probability, given that three gains of dry fruits from fleshy fruits occur on this tree, that they would all occur in clades inhabiting savannas?

Go to the **tools window** (in the bottom corner of the **tree window**) and click on the **test correlation tool** (the tree with a 'c' at its base). Make sure that you are tracing the Ecology character, as the character currently traced will be viewed as the independent character. Then click on the lowest branch on the cladogram. This runs the test over the whole cladogram by clicking on an interior branch you could limit the test to just the specified subclade.

In the dialog box that comes up (called the Correlation test Parameters dialog box) add in the number of gains and losses of dry fruits Leave **distinguished branches** as 1, because it is when the ecology character is in state 1 that you expect the changes to happen. Now click on **exact count** and **calculate**. This will use the formulae presented by Maddison (1990). It will generate every possible distribution of three gains on the tree, and count the number of times that those gains occur in rain forest or savanna. The exact count algorithms become computationally challenging as one increases the number of changes in the dependent variable of interest (in fact, changes in the dependent character of more than 5-6 can take a long time). Doing simulations is often a necessary option. The simulations generate changes randomly on the clade selected and count the number of gains and losses on the branches with the specified dependent and independent variables across the entire cladogram. Only those that correspond to the previously specified number of gains and losses will be examined for the branch distributions of gains and losses

The next dialog box that will appear is the correlation test results dialog box, which will allow you to ask what the probability of having more, as many, or fewer than the indicated number of gains and losses along the distinguished branches (independent variable branches). Find out what the probability of having as many or more than three gains and zero or less losses. Write down the p-value that you get.

Because you did an exact count the computer has stored the counts from every possible arrangement of three gains on the tree. This means that you can ask about the probability of other possible arrangements on the tree. For example, what if only two of the changes occurred in the savanna? If you had used a simulation instead of an exact count, then you could only ask about the exact situation that you asked *MacClade* to simulate in the first place.

What if there were four changes instead of three changes? How much would that affect p-value? Run another analysis with four changes this time. If it takes too long on this computer you may need to run a simulation.

Pagel's Test

Pagel (1994) describes a method of looking for correlated changes in discrete 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.

Pagel’s test can be run in *Mesquite* but not *MacClade*, so shut down *MacClade*, open *Mesquite* and open the same data file. Open a **tree window**. Open a **mirror tree** and reconstruct the two different characters using parsimony. Isn’t that a nice way to look at the correlation? Compare the reconstructions using likelihood also, since that is what we will really be doing in this situation.

Go back to the tree window select **Analysis > Pagel94 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. Select **stored characters** and then hit **OK**. A new window will appear describing the programs progress through the simulation.

A **Pagel94 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. How does this value compare to your p-value from Maddison’s test. There is also a **Pagel94 menu** that will allow you to change several things about the analysis.

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 this compare to the four step p-value from Maddison’s? 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 Maddison’s test? Why is there a difference? Which do you think is more realistic?

Pairwise Comparisons

Another possible approach 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.

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 switch back the character state that you just changed, so that there are again three changes on the tree. In the **tree window** select **analysis > pairwise comparisons**. For are first set of comparisons let's just take the maximum possible number: select **most pairs** and hit **OK**. We want to use the **stored characters**.

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 a character to itself, and that doesn't tell you much; although, it does give you some indication about the power of the test. Switch the characters so that you are making 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**: 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**: 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**: cases in which the taxa disagree in the independent character, but have the same dependent character. (01 vs 11) or (00 vs 10); **Remainder**: 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 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 for 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?

All three of these tests may all be useful under different circumstances, depending on what your exact question is and the nature of your data. One thing to keep in mind is power, how likely it is that a difference which really exists will come up as significantly different. Which one of these methods was the most powerful for the data that we examined? If you have time and you want to play around why don't you see how many extra changes or taxa you have to add to this tree to get a p-value below .01 for Maddison's test and Pagel's test.