

## Lab 10: Independent Contrasts

Today we're going to use *Mesquite* to derive and analyze independent contrasts for continuous data. We are going to use PDAP:PDTree a Mesquite module that is not installed as a default. It has been developed by Midford, Garland, and W. P. Maddison, and it can be downloaded from [http://mesquiteproject.org/pdap\\_mesquite/index.html](http://mesquiteproject.org/pdap_mesquite/index.html). You can find installation instructions there. It is remarkably easy to install new *Mesquite* modules.

### Testing Assumptions

First we will analyze data on several ungulate and carnivore taxa from Garland, et al. 1993. *Systematic Biology* 42:265-292. The first column is log body mass and the second is log range size. We will do a number of comparisons to make sure that the data fit the assumptions of our model, and then analyze the independent contrasts.

Open *Mesquite* and then open the file **Example1** in the **Lab 10 examples** folder. Open a **new tree window** with the **stored tree**. Set the **branches proportional to lengths** (in **drawing** menu) and make the tree prettier. First let's just do some basic comparisons. Open a **mirror tree window** (in **tree** menu). Trace the two characters on trees to look for an evolutionary relationship (**mirror > left side** (or **right side** as the case may be) **> trace character history**).

Now let's look at the raw relationships between the data on the tips of the tree. In the tree window select **Analysis > New Scattergram for > taxa, same, stored characters**. To look at the regression line for these two characters select **scattergram > analysis > other choices, scattergram regression diagnostics**. A line will appear showing the least squares fit. To view the statistical parameters on this analysis hit the **text** tab. After a list of the points that make up the graph, you will find a summary of the statistics that describe the fit. At the bottom of that summary is a p-value for the significance of the slope. As you can see there is a very significant positive relationship between these two variables.

Now we'll look at the independent contrasts. In the **tree** window select **analysis > new chart for tree > PDAP diagnostic chart, stored characters**. The first plot that will appear is a graph of the absolute value of the standardized independent contrast as a function of branch length. Because we divide the independent contrast by the branchlength to standardize it you may get an inverse relationship between branch length and standardized IC, if you don't pick your branchlengths appropriately. Toggle back and forth between the different characters, as you can see both have an inverse relationship between IC and branch length. To investigate the significance of this relationship click the **text** tab and look for the p-values at the bottom of the statistical parameters.

Uh oh, our second character (log range size) shows a significant relationship. What are we going to do? We have two options either transform the branchlengths or transform the character values. To decrease the effects of an inverse relationship like this we should take the log of either parameter. Since we are already dealing with the log of the characters let's start by taking the log of the branch lengths. In the **tree** window select **tree>alter/transform branchlengths > other choices, natural log transform**. How did that affect the relationship between branchlength and contrasts for the second character? Great, but what about the first

character? There's now a positive relationship between branchlength and the standardized contrast. One way to deal with this problem would be to use a different set of branchlengths for each character. That is an OK thing to do, although the slopes of the comparisons between the contrasts become more difficult to interpret. To do this you would save the contrasts from each set of branchlengths independently (just wait I'll explain how to save contrasts) and then analyze these using statistical software. However, you can not use separate branchlengths for analyses in *Mesquite*, so let's try to transform the first character to make it fit our assumptions. Since we have a positive relationship between these variables let's take the exponent of the character this time. In the **character matrix** window select the first column of characters, then choose **matrix > alter/transform > other choices, exp transform**. Now the slope is insignificant, and we feel better about our branchlengths.

*Mesquite* is capable of evaluating fit in several other ways. One way is to look at the distribution of independent contrasts. The absolute value of independent contrasts should fit a half normal distribution with mean 0. To test this, save the independent contrasts (wait one second) and calculate the variance if the mean is assumed to be 0. Then compare the cumulative distribution of your independent contrasts to the expected cumulative distribution. The contrast should also be independent of the nodal value. To test this go to the PDAP window and select **PDAP chart > Abs. contrast vs absolute node value**. That's not so good. We should really try using different sets of branchlengths for each character, at least when we have more time. To test the independence of the standardized contrasts with the age of the node, select **PDAP chart > Abs. contrast vs. node height**. How does that look?

## Comparing Contrasts

To do a linear comparison of these two standardized contrasts select **PDAP chart > Y contrasts vs. X coordinates positivized**. This graphs the SACs in such a way that the x-coordinate is positive. The y-coordinate is then negative if the change it describes is in the opposite direction of the x-coordinate or positive if it is in the same direction. The black line is the least squares the red line is the reduced major axis and the green line is the major axis. In the **tree** window select a clade using the **clade select tool**. That clade will light up in the **PDAP** window, and you can see which contrasts are from that clade.

This analysis allows for two different types of statistical calculations. Click the **text** tab to see the outcomes of both evaluations. The first set of statistics is for a parametric test; assuming that the contrasts for both characters are normally distributed you can see if the slope is significantly greater than zero. The first set of p-values test this hypothesis. The second set of p-values do not consider the magnitude of change, they only consider whether changes in both characters happen in the same direction. Although this test is less powerful, it works even when the assumptions of normal data are broken. As you can see, under both sets of assumptions this relationship is highly significant.

The fact of the matter is that there are many statistical tests that you may want to do which can not be performed by *Mesquite*. To do these tests you need to save the independent contrasts and use other statistical software to evaluate them. To accomplish this you must be in the **Y contrasts vs. X coordinates positivized** analysis. Select **PDAP chart > Generate file of independent contrasts**. This will only work for two contrasts at a time, so if you have more than two characters that you are comparing you need to switch the characters evaluated in the window and save them to file two at a time. Independent contrasts will be independent of each other, so it doesn't matter which two characters you pick, before you save the data. It is

important to understand the output of this file, so save one of these files and open it in a text editor. The first column shows the two nodes that the contrast is taken between. The second column shows the node at the root of this particular contrast. The next two columns show the independent contrasts and the fifth column shows the expected variance. ***Because the third and fourth columns are not standardized, you must divide them by the fifth column in order to get the standardized contrast for most statistical purposes.*** You can learn about the rest of the columns by exploring the *PDAP* documentation.

You can compare your independent contrasts analysis to your original data by selecting **PDAP chart > conf. and pred. intervals**. The dots are your taxon values, the black line is your original least fit regression on the standardized contrasts, the blue line is your 95% confidence interval, and the green line is your prediction 90% prediction interval. To change the percentage standards for your two intervals go to **PDAP chart > set PI/CI width**. Why do so many of the actual taxon values fall outside of the expected range?

Independent contrasts can be used to deduce the character state of the bottom node and calculate confidence intervals. The results are essentially the same as weighted square change parsimony. Select **PDAP chart > Root Node Reconstruction**. All the blue dots are your actual tip values; the red dot is your reconstructed ancestor with 95% confidence intervals for both characters. To change the confidence intervals select **PDAP chart > Set CI/PI width**. You can also reconstruct ancestral states at other nodes, but they have to be a root. Reroot the tree in your tree window at another node. To be rooted at a node the root should be a polytomy. Go back to your PDAP window and look at the reconstruction for this node.

There is one more test of fit that can be done by comparing the residuals from our least squares analysis to their actual values. Select **PDAP chart > residual vs contrast**. These values should be homoskedastic. That is to say that the variance of your residuals should be independent of your x-value.

## Justification of Independent Contrasts

So now that you see how to do an independent contrast, let's look at what effect that has on your actual data analysis. Independent contrasts assume a Brownian motion model of evolution, so they should draw the correct conclusions when those assumptions hold true. To evaluate this we'll run a series of simulations and compare their results. In the **character matrix** window select **characters > new matrix from > simulated matrices on current tree, evolved continuous characters, Brownian default**, and simulate 4 characters. A new character matrix will appear. Go to the tree window and plot these characters against each other (**Analysis > New scattergram for > taxa, same, stored characters**, and pick your simulated matrix). Now plot a regression line (**scattergram > analysis > other choices, scattergram regression diagnostics**). Investigate the correlations between each of the four characters. Do you see more correlation between these characters than you would expect for a random data set? Why?

Now let's do the standardized contrasts on these characters. In the tree window select **PDAP chart > new chart for tree > PDAP diagnostic chart, stored characters, your simulated matrix**. Look at the contrasts (**PDAP chart > Y contrasts vs X contrasts (positivized)**). Investigate the p-values of these slopes. Do you see any significantly relationships? How do these p-values match your expectations?

We used simulations here to show you where independent contrasts are effective. However, simulations are more useful for evaluating the significance of a test statistic. You could run a number of simulations (say 1000) and see how often you got a statistical value (say R

squared for a linear regression) relative to a null expectation. Often null distributions from complex phylogenetic analyses do not match those calculated from parametric statistics.

## Outliers and Evolutionary Events

Now let's look at a situation where an uncritical analysis of independent contrasts could lead you astray. Open **Example 2** in *Mesquite*. First let's look at the raw correlation between the two characters in a scatter gram (if you can't remember how see the first section). They fall into two distinct groups. Calculate a regression line and look at the p-value. Now compare the two characters in a mirror tree window to see how they are distributed on the tree and if there is any correspondence in their evolution. As you can see, there is one clade that has particularly high values for one character and a particularly low value for the other. Maybe it's just an effect of phylogeny that leads to this correlation. How could we deal with that? Why, independent contrasts of course.

Construct a set of independent contrasts to compare these characters (**Analysis > new chart for tree > PDAP diagnostic, stored characters**). Now compare the contrasts for the two characters (**PDAP chart > y contrasts vs X contrasts**). Look at the p-values for the parametric test. The p-values derived from the parametric test are definitely significant. Why?

Look at the contrasts vs standard deviations chart. As you can see most contrasts fall neatly around a line independently of their branchlength; however, one point is a severe outlier for both characters. This means that the assumptions of the parametric test may not be justified. Go back to the y vs x plot. You can see this outlier in the bottom right corner. Click on it. Now go back to your tree window and look at which branch lit up. Does that correspond to the branch separating the different clade from the rest of the taxa? The answer to this may appear to be no at first. The thing is independent contrasts are taken between sister clades, so you expect the real difference to be calculated on the branches between the different clade and its sister. Thus the contrast refers to the node between these two clades, which will light up the branch beneath that node. So is it the expected contrast? One big evolutionary change is creating correlation throughout our entire data set.

How can we deal with this outlier violating our assumptions? You could save your contrasts and analyze them in another statistics program. There you could exclude your outlier. However, you have to be very careful about excluding data points just cause you don't like them. Another option is to do a non-parametric sign test. PDAP does that for you. It is the second set of p-values that you find in the text tab of the y vs x plot. As you can see, those are not significant.