Introduction to COMPARE and the Phylogenetic Comparative Method

There are many questions that can be answered using the phylogenetic comparative method. In this lab, we will focus on statistical methods for inferring correlations between traits and ancestral states, discussing also what to do when the phylogeny is only poorly known.

A. **General Instructions**

1. You can run COMPARE directly on the web (**http://compare.bio.indiana.edu**) or download a version to your own computer and run it locally. Source code is also available, so you can make changes, if necessary. Always make sure you have the latest version before using it. Reference it as shown on the web page.
2. COMPARE is a java applet, and runs on any platform within a java applet reader or web browser. So, for example, it can be run on a PC using Netscape or on a Mac using Internet Explorer. If you're having trouble running COMPARE, start by making sure that you have a recent version of your web browser and that you've checked the COMPARE web page for suggestions about your specific platform.
3. For security reasons, COMPARE cannot save to or read from your hard drive. You might want to keep a word processor open while you're running analyses so that you can copy/paste and save your results.

B. Entering Data

1. To begin, click on the "Demo" button to see the basic input data needed. Briefly, you need a phylogeny, list of taxon names, and measurements of those taxa (i.e., the comparative data). Taxon names in the names box and the phylogeny must match exactly, including case. The Demo data set includes both branch lengths and phenotypic data within the phylogeny structure. These are both optional.
2. If creating your own tree from scratch, consider starting by clicking on the "Generate Trees" button. COMPARE can generate a sample tree for you based only on the number of taxa you specify. You can then use the "Draw/Modify Tree" button to make changes on your phylogeny by dragging branches around. Change trait values and/or branch lengths by clicking on nodes. Choose "replace" when you're ready to save your changes in the phylogeny input box on the main page. It is usually a good idea to keep taxon names short to avoid later complications.
3. Alternatively, you can create your tree in MacClade or PAUP and copy/paste the contents of a NEXUS file into COMPARE (Click on the NEXUS button). One potential problem is that COMPARE requires that the phylogeny be a completely resolved, binary tree. Add very short branch lengths between taxa to resolve any polytomies before creating the NEXUS file. (As long as the inserted branches are Very short relative to other distances on the tree, it does not matter which taxa you join in resolving the polytomies.)
4. Trait means for each taxon must also be accompanied by standard errors. If you don't have measures of the standard error, set these equal to zero. In most cases, the methods will automatically assume that the standard errors are zero anyway. COMPARE forces you to write them in to emphasize that this is an assumption.

C. Correlations among traits. There are many ways to estimate relationships among traits in COMPARE. Below, we explore the DEMO data set in several different ways. In each case, start by choosing a method from the options box in the middle left of the main window. Click on Execute and the results should appear shortly (within a few seconds) in the Results box below. The output begins with the input data and tree. You can copy/paste these results into a word processor or other program by using the shortcut keys (Ctrl-C, Ctrl-V in Windows).

1. Let's start by choosing PGLS-Relationships. PGLS can be viewed as an extension of Felsenstein's independent contrasts method that allows for flexibility in the underlying evolutionary assumptions. This flexibility is obtained through the use of a single parameter (alpha), which can be interpreted as a measure of evolutionary constraint acting on the phenotypes. When alpha is very small, the method approximates Felsenstein's contrasts (FIC). When alpha is large, comparative data are independent of phylogeny, and the method gives results similar to a raw data analysis (not-phylogenetic; TIPS). Results include parameter estimates at a range of different alphas, the maximum likelihood estimate of alpha, and parameter estimates given that maximum likelihood, assuming that alpha is very small (FIC) and assuming that alpha is very large (TIPS). Again, the PGLS is intermediate, depending on the estimated value of alpha. (Note that PGLS is one of very few phylogenetic methods that can be used to incorporate within-species variation. The default is Not to include this information [check the switch just below the input data to change this], simply because doing so makes the program run slowly.) Write the correlation between traits 1 and 2 for PGLS, FIC and TIPS below:

2. The Phylogenetic Mixed Model (PMM) might also be viewed as an extension of Felsenstein's method, but applying a very different evolutionary model than does PGLS. PMM starts with Brownian motion evolution along a phylogeny and then adds a burst of non-heritable evolutionary change to each of the terminal taxa (e.g., to account for phenotypic plasticity due to environmental differences). When this non-heritable component is very small, PMM results will approach Felsenstein's method. When the non-heritable component is large, PMM results will approach a non-phylogenetic analysis (TIPS). Write the results for the PMM below. Compare this to the results from the PGLS method. Although these two methods may give similar results, they do so for very different reasons. Applying both gives you an idea of how robust your results are to differences in the underlying evolutionary assumptions of both methods.

3. The Independent Contrasts method can be more useful in some cases, if you want to explore a particular data set in more detail or apply statistics other than correlation coefficients. In COMPARE, "independent contrast" will print out the actual contrast values which can then be used in scatterplots or in additional statistical procedures (e.g., multivariates). Begin by clicking on "Plot Points" at the top of the main COMPARE window and looking at the spread of the data. Choose **Independent Contrasts** and Execute. Use your mouse to copy and paste the results into Excel or some other similar package and save. (Again, COMPARE cannot save to disk because of security reasons associated with running a program over the internet.) Create a scatterplot of the independent contrasts for the two traits against each other and compare to the raw data plot in COMPARE. Do you see any outliers or particularly influential points? Identify these on the phylogeny from the list of contrast node names given by COMPARE. Look at the correlation coefficient also given in the COMPARE results (this will be different than one you calculate in Excel, because the intercept must be forced to zero). COMPARE implements Felsenstein's independent contrasts method (as originally described), which gives results identical to those produced using Grafen's "standard regression". What can you conclude about the evolutionary relationship between the two characters using Felsenstein's interpretation? Using Grafen's? Can you gain any detailed insight into your correlation results from COMPARE based on details of the scatterplots?

4. The Spatial Autocorrelation method can also be used to explore a particular data set in more detail. This model partitions variation in each trait into "phylogenetic" or "specific" effects. We "correct" for phylogeny by estimating the "specific" effects and conducting further statistical analyses on these. COMPARE calculates correlations and regression slopes between pairs of traits using the spatial autocorrelation approach. Focus on the "Specific" results, even though total and "phylogenetic" results are also reported for your information. Results from autocorrelational analyses can sometimes be quite different from those resulting from the evolutionary methods above. Try a run on the demo data set. Compare and contrast your results to those above.

5. Consider whether your evolutionary question is really about adaptation. For example, are you asking whether a trait is related to the environment in which it is found? If so, consider applying Hansen's adaptation model instead. Hansen's model envisions organisms as groups of traits evolving in response to stabilizing selection imposed by a variety of potentially conflicting environmental factors. His PCM uses comparative data and information on how long each taxon has been subjected to a particular environmental state to estimate the relative impact of that specified environmental factor on phenotypic evolution. To apply it, you must have some idea of when on the phylogeny the taxon was in each environmental state. In COMPARE, the first option window asks you to specify where environmental shifts happened, and you can click on one or more nodes to identify each branch of the tree with a particular environment. The result are the slopes and overall fit of a regression estimating the relative importance of the environmental factor you identified on the evolution of the trait. Theta values are regression slopes describing the relative importance of the specified environmental factor on trait evolution. Theta1 is the estimated difference between optimal phenotypes in Environment 1 vs. 2. Theta2 is the optimal phenotype estimated for Environment 2. R^2 values describe the overall fit of the model, and provide measures of how much of the trait variation is explained by this single environmental factor. If the environment explains a lot about the trait, we might conclude that it is an adaptation. What can you conclude from your analysis of the DEMO data? Does this approach match your own thoughts about adaptation? How or how not? Are there other interpretations of the model parameters that might be useful?

D. Questionable Phylogenies. We are often not confident that the phylogeny is correct. One common problem is that branch lengths on the tree are either not available or not trust-worthy. The problem is actually worse in comparative analyses because for most methods, branch lengths are assumed to be in units of the expected amount of phenotypic change. Thus, branch lengths are actually descriptions of the type of phenotypic evolution rather than estimates of time, and unless we actually know what sort of phenotypic evolution is happening along each branch of the tree, we are unlikely ever to have correct branch length information. Because varying the branch lengths on a phylogeny often has a much large impact on the results of a phylogenetic analysis than does changing the phylogenetic topology, the problem is a serious one.

1. Ideally, we would start a phylogenetic analysis with a simple best estimate of the phylogeny and branch lengths (ideally a phylogeny based on independent characters and a good estimate of branch lengths in units of time). If this is available, we must still consider robustness of the results to incorrect branch length information because again, phylogenetic comparative methods require branch lengths in units of character change rather than time. A comparison of PGLS and PMM results provides a good description of the robustness of the results to variation in the underlying evolutionary model (i.e., branch length differences). Plots of independent contrasts vs the branch lengths might also be used as a heuristic tool. Make some general statement about the robustness of your above results to variation in the microevolutionary model based on the results of these tests.

2. If several phylogenies are available (e.g., 5 most parsimonious trees), these can be entered on different lines of the phylogeny input box. Create 5 copies of the DEMO phylogeny in the input box with minor modifications of the branch lengths. Change the “Number of phylogenies” at the top of the main window to 5 and click on the “Results Summary” box. Run “independent contrasts”, and COMPARE will calculate the mean and variance of the parameter estimates, and a confidence interval on the regression slopes based on assuming that all 5 of these phylogenies are equally likely. What can you conclude about the robustness of your results based on this analysis?

3. Another possibility is to generate a large number of computer-generated phylogenies. Use the "Generate Trees" button to switch to the secondary page. Choose “Get initial data from main window” and “Generate branch lengths only” to generate 500 phylogenies with the same topology but different branch lengths. Click the appropriate button to replace the original phylogeny in the main window with these. Run Felsenstein's contrasts analysis again, choosing the Results Summary option. The results will be means and variances for 500 separate analyses run on each branch length combination, and a 95% confidence interval combining the results. Note that if the regression slope is not significantly different from zero, the correlation coefficient will also be zero. What can you conclude about the evolutionary relationship between the two characters given these results? Compare to results from above analyses.

4. COMPARE can also be used to generate a large number of computer-generated topologies. Copy and paste the trait data from the main COMPARE window to a separate word processor to save it. Use the "Generate Trees" button again to switch to the secondary page. Choose “Generate without initial data” and and “With Constraint”. Type “(1,2,3),(4,5,6),7,8,9” into the Constraints box. This specifies that you want phylogenies which group 3 taxa (A, B, and C) into one clade, 3 other taxa (D-F) into a second clade, and has 3 other taxa (G-I) which are not included in either of the other two clades. Taxon names are assigned by starting with A and reading from left to right along the Constraints box. Generate 500 trees with a specific constraint. Click on Replace and then re-enter the taxon data before running the analysis. What can you conclude about the evolutionary relationship between the two characters given these results? Compare to results from above analyses.

5. Other authors have suggested that we can determine if a single set of branch lengths is appropriate by plotting each set of standardized independent contrasts by their standard deviations. The result should be a broad horizontal band with no evidence of a relationship or a "wedge" shape. If you get anything else, this is evidence that the trait may not conform well to the BM assumption. Some people suggest that you should transform your data to remove wedge shapes or other patterns. I don't recommend this because it makes the result very difficult to interpret in evolutionary terms. (Ask me about this, if you're interested.) Nevertheless, you can do something similar in COMPARE by raising the branch lengths to some power or multiplying them by some factor. Try it - how does doing this affect your results?

E. Estimating Ancestral States

1. There are now 3 main methods for estimating ancestral states of continuous traits: 1) sum-of-squared-changes parsimony, 2) Schluter et al.'s maximum likelihood, and 3) Martins & Hansen generalized least squares. Schluter et al.'s method is an improvement on parsimony because it provides measures of the accuracy of those estimates (SEs, you can implement the method using ancml from http://www.zoology.ubc.ca/~schluter/ancml.html). Martins & Hansen's PGLS method also gives SEs, and allows for much greater flexibility in the underlying evolutionary assumptions. Under the simplest conditions, all three give identical estimates of the ancestral states. Unfortunately, simulation results suggest that the SEs are poorly estimated by both ML and GLS methods and that the GLS flexibility in assumptions does almost nothing for estimation accuracy.
2. Clear and restart the Demo data in COMPARE. Generate 10 trees with the different branch lengths but the same topology and replace the single one from the Demo. Instead of "Independent Contrasts", choose "PGLS Ancestor". Click on the Results Summary box and Execute. An options box should open. Leave the options as they are to get ancestral estimates for the simplest conditions, or if doing the analyses under parsimony or ML methods. The results are estimates for each of the nodes on each tree. (Click on Draw/Modify Tree to identify the nodes.) How much do the different branch lengths affect the ancestral estimates or the standard errors?

3. Repeat this analysis, but this time choose the switch that uses standard errors. This applies a version of the GLS method in which we also take within-species variation into account (using measures of the variation among individuals in a species that are provided by the researcher in the input box). Briefly describe what you find.

4. Repeat but apply an exponential rather than a linear model, with a restraining force of 0.1 and then 2.0, with and without including within-species variation. Compare and contrast your results for the 5 separate runs of the GLM method.