Applied Phylogenetics: Phylogenetic Comparative Methods

In-Class Assignment

Matt Johnson

November 11, 2015

1 Objectives

- Working with trait data in a phylogenetic context.
- Observe how phylogenetic history affects trait correlations.
- Conduct tests for phylogenetic signal.
- Investigate the effect of tree shape on phylogenetic comparative methods.

2 Phylogenetic Comparative Methods

In lecture, we learned about the importance of Phylogenetic Comparative Methods (PCMs) for understanding the evolution of traits in an evolutionary context. In this lesson, we will use real data to explore the relationship between observed traits and reconstructed phylogenies. You will use several standard methods for characterizing the effect of the phylogeny on trait correlations and calculate the amount of phylogenetic signal present.

2.1 Example Dataset

We will be working with a data set from a real study:

Evans ME, Hearn DJ, Hahn WJ, Spangle JM, and Venable DL. 2007. Climate and life-history evolution in evening primroses (Oenothera, Onagraceae): a phylogenetic comparative analysis. Evolution 59(9): 1914-1927.

In the study, Evans et al. explore the relationship between life history (annual vs. perrenial) and the preferred (realized) climate observed for several species and subspecies of *Oenothera*. They reconstructed a phylogeny of 30 species from three loci (trnL, trnH, and ITS) using Bayesian Inference. Because the statistical support for several clades in

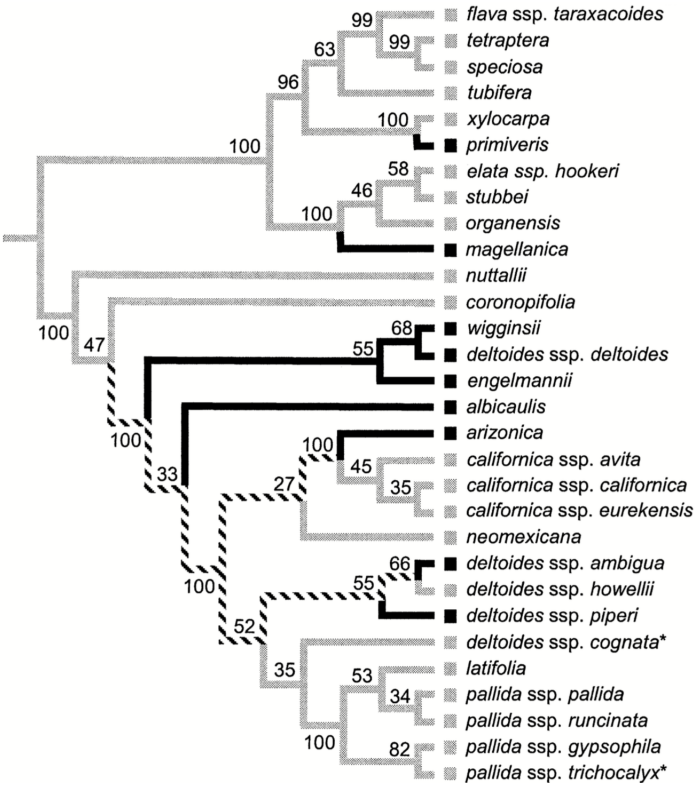

Figure 1: Figure 1 from Evans et al. 2005, showing the phylogeny of 30 Oenothera species. Dark branches represent annual lineages, gray branches represent perennial linages, and hashed branches have unknown life history. Values on each branch represent statistical support (posterior probability from Bayesian inference).

the phylogeny was low, they repeated their of their statistical tests on many trees from the Bayesian analysis.

Their trait data consists of five climatic variables measured in 20 species: Annual Aridity, Summer Precipitation, Maximum Average High Temperature, Winter Precipitation, and Average Winter High Temperature.

2.2 Shiny App

We will explore the Oenothera phylogeny and climate dataset from Evans et al. using a web application. The application runs R, which is a popular programming language for phylogenetics and comparative studies. The packages we are using today include:

ape `http://ape-package.ird.fr/`

phytools `http://www.phytools.org/`

picante `http://picante.r-forge.r-project.org/`

To launch the web app, go to this link: `https://mossmatters.shinyapps.io/Phylo_Comp_Methods_Shiny`

3 Trait Data

In the leftmost tab of the web app, you can explore the dataset of climatic variables for 20 *Oenothera* species. You will notice that several species have many subspecies listed. Do you notice any patterns in the data related to the species boundaries?

3.1 Simple Regression

In the next tab, you can explore how the climatic variables are related to each other *without* considering phylogeny. Use the two drop-down menus to select the dependent (X-axis) and independent (Y-axis) variables. The scatterplot on the right will update automatically, including the linear regression line. The text box will show you the correlation coefficient (r) and the significance of the linear regression (P-value).

3.2 Questions

1. Which two variables have the highest correlation (r)?
2. Which correlation has the lowest P-value?
3. How many pairs of traits are significantly correlated ($P < 0.05$)?
4. What is a non-phylogenetic explanation for correlation of these variables?

4 Phylogenetic Regression

The principle of the phylogenetic comparative method is that organisms have related traits because they have inherited them from common ancestors along a phylogeny. Using this principle, statistical predictions can be made about how two species should resemble each other given their phylogenetic distance. Another way to look at it is to consider the equation for linear regression between two variables X_1 and Y:

$$Y = b_0 + b_1 X_1 + \epsilon$$

where b_1 refers to the correlation coefficient, and the final term is residual error. The phylogenetic comparative method predicts that a percentage of that residual error can be assigned to phylogenetic relatedness.

In this section you will explore how the phylogeny plays a role in the relationships between variables measured among species.

4.1 Choosing a Tree

In the "Tree Plot" tab, use the slider to select one of the 330 trees available. By moving the slider and selecting alternative trees, you may observe different topologies. We will explore how this affects the conclusions of phylogenetic comparative methods in section 6.

4.2 Phylogenetic independent contrasts

Originally described by Felsenstein in 1985, the phylogenetic independent contrast (PIC) has a similar goal to the phylogenetic regression discussed above. In this method, a linear regression between two variables is done *not* with the observed trait values, but with values that represent the contrasts among species and clades on the tree. Because a fully bifurcating tree with N species has N-1 clades, there will be slightly fewer comparisons.

The "Independent Contrasts" tab is similar to the "Simple Regression" tab, but will present the PIC values for each trait. Select two traits from the dropdown menu to see the scatterplot and linear regression line for the contrasts. As before, the correlation coefficient and P-value are shown for the linear regression. Compare the scatter plots, regression lines, and p-values between the simple and phylogenetic regressions.

4.3 Questions

1. For which pairs of traits did the correlation get stronger or weaker?

2. Are there any pairs of traits where an insignificant correlation became significant using independent contrasts? Or vice versa?

3. Do you think the PIC method has the same power to detect correlations as the simple regression method? Why or why not?

5 Phylogenetic Signal

In the previous section, you explored phylogenetic regressions and independent contrasts. Both of these methods assume a particular method of trait evolution along a phylogeny: Brownian motion. The model of Brownian motion uses only a few parameters: the "initial" value of the trait (at the root of the phylogeny) and the rate of evolution. A variety of factors may cause the evolution of a trait to deviate from the expected model. For example, a trait may have one rate of evolution in one part of the tree, and a different rate in another part of the tree. Brownian motion also assumes that the trait can evolve with no boundaries; if there are theoretical limits to the value of the trait, this may constrain evolution from fitting a Brownian motion model.

In many phylogenetic comparative studies, researchers ask: to what extent does the evolution of a trait deviate from expectations under Brownian motion? This is the essence of *phylogenetic signal*: the extent to which closely related species tend to resemble each other.

Two metrics are often used to quantify phylogenetic signal: K and lambda.

5.1 Blomberg's K

Described in Blomberg et al. 2003 *Evolution*, this metric is a ratio of ratios that is inspired by phylogenetic regression. It is a ratio of ratios comparing the percentage of the observed trait variance attributable to phylogeny to the same ratio expected under a perfect Brownian motion process. Therefore, K can be seen as a "partitioning of variance" metric.

A value of 1.0 suggests "perfect" phylogenetic signal the trait evolves exactly as expected under Brownian motion. Values less than one suggest the species are more related than expected under Brownian motion. Values greater than one suggest the species are less related than expected.

A significance test of K involves comparing the observed value of K to values calculated when the names of the taxa are shuffled on the phylogeny (permutation test).

5.2 Pagel's lambda

Originally described in Pagel 1999 *Nature*, the lambda metric is a very different statistical concept. The statistical description involves the "variance-covariance" matrix used in phylogenetic comparative studies: the covariance between phylogenetic distance and trait distance. Lambda is used as a multiplier, to transform the matrix so that it better "fits" the expectations under the Brownian motion model.

A value of lambda equal to 1.0 suggests the trait is evolving under Brownian motion. A value of lambda equal to 0.0 suggests zero phylogenetic signal. Intermediate values suggest intermediate phylogenetic signal.

Unlike Blomberg's K, the value of lambda is not calculated from the data, but is estimated using Maximum likelihood. This makes it a good choice to use in hypothesis testing, because significance tests can use the likelihood ratio test (LRT) or Akaike's Information Criterion (AIC), rather than permutation tests.

One test for significance involves calculating the likelihood of a model where lambda is estimated, compared to a model where it is not. The ratio of these likelihoods can be used as a Chi-squared statistic to generate a P-value.

5.3 Visualizing Phylogenetic Signal

Returning to the web app, in the "Phylogenetic Signal" tab you will see several drop-down menus and a plot of the phylogeny on the right. This plot is known as a "phenogram" or "traitgram" and represents the phylogeny (X-axis) plotted together with the trait data (Y-axis). In this case, the trait data observed at the present time (at the tips of the tree) are used to reconstruct the ancestral states of the traits at ancestral nodes.

Traits with high phylogenetic signal will be distinguishable from traits with low phylogenetic signal, based on the density of the branches.

Use the first dropdown menu to explore the phylogenetic signal in each trait. The second dropdown menu will allow you to select a method for estimating signal, and the checkbox indicates whether a hypothesis test should be conducted. The results are displayed in the third box.

5.4 Questions

1. Based solely on the traitgram, try to predict which trait has the most phylogenetic signal.
2. Does your prediction agree with either of the phylogenetic signal metrics?
3. How many of the traits have significant phylogenetic signal?
4. What does "significant" mean for each metric?

6 Phylogenetic Uncertainty

There are many potential pitfalls in studying phylogenetic comparative methods, in a philosophical or statistical sense. Exploring all of these would take an entire course of its own, so we will focus on one potential source of uncertainty in phylogenetic comparative methods.

If you refer to Figure 1, you will notice that the support values for the phylogeny are very low. This means that most of the clades on the tree do not have statistical support (in this case posterior probability from Bayesian inference).

What does this mean for the phylogenetic comparative methods? Each of the methods we have investigated so far use one single tree as an input. Is the tree selected the "true" tree? If we have low confidence in the tree, how can we explore the effect alternative trees may have on our conclusions about the evolution of traits?

At the bottom of each tab in the web app, there is a slider used to select one phylogeny from a set of trees generated by the Evans et al. phylogenetic data.

Use the slider to explore how the results of the Phylogenetic Regression and Phylogenetic Signal analyses may change given different phylogenies.

6.1 Questions

1. As you change the phylogeny, which aspect appears to change more: the toploogy, or the branch lengths?

2. Are there any cases where the phylogeny changes the interpretation of the analysis?

3. How does the inference of phylogenetic signal change with different phylogenies?

4. If this were your study, how would you summarize the support for phylogenetic signal or regression using all of the trees?

7 Further Information

This tutorial and the web app owe a lot to the book: Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology, edited by Laszlo Zsolt Garamszegi. The chapters feature authors who are actively developing phylogenetic comparative methods and software packages to use them, frequently in R.

The R-sig-phylo mailing list is a typically friendly source of information. You can subscribe at: `https://stat.ethz.ch/mailman/listinfo/r-sig-phylo`

The group also maintains a wiki for learning how to do analyses in R: `http://www.r-phylo.org/wiki/Main_Page`

Liam Revell maintains a blog describing the development of his R package at: `phytools.blogspot.com`