Introduction to Phylogenies and the Comparative Method

Sharon Jansa, Jesús N. Pinto-Ledezma and Jeannine Cavender-Bares

In this lab, you will learn basic tools in R for visualizing phylogenies, optimizing ancestral states for a discrete and continuous characters, testing models of character evolution, and performing phylogenetic correction of a regression model. This lab is based in part on one designed by Luke Harmon for a workshop that he and others ran; the original can be seen here: http://lukejharmon.github.io/ilhabela/instruction/2015/07/03/PGLS/ There are many other useful labs in comprative analysis from that workshop that you can peruse at your leisure. http://lukejharmon.github.io/ilhabela/ You will need two datasets, that will be provided for you: 1. anolisDataAppended.csv 2. anolis.phy These are a datamatrix of trait data for a set of Anolis lizards, and a phylogeny for those species.

Set up your data and your working directory

You will need to have a set of R packages to do this lab. Install the following packages:

```
packages <- c("ape", "geiger", "nlme", "phytools", "rr2")
# Package vector names
```

```
# Install packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())

if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}
```

Load installed packages

```
library(ape)
library(geiger)
library(nlme)
library(phytools)
library(rr2)
```

Set up a working directory and put the two data files in that directory. Tell R that this is the directory you will be using, and read in your data:

```
setwd("path/for/your/directory")
```

You can download the data directly on your computer by clicking **HERE** and **HERE** and store them in a folder named **Data"**.

```
anoleData <- read.csv("Data/anolisDataAppended.csv", row.names = 1)
anoleTree <- read.tree("Data/anolis.phy")
```

OK. You should be ready to go.

Working with trees

Let's start by looking at the phylogeny of these lizards and learning a bit about how to work with trees in R. What does your tree look like?

```
plot(anoleTree)
```

Whoa. That's ugly. Let's clean it up.

```
plot.phylo(anoleTree, no.margin = T, cex = 0.5)
```

Better. You can mess around with tree plotting functions in plot.phylo() as much as you'd like. Try this for example:

```
plot.phylo(anoleTree, type = "fan", no.margin = TRUE, cex = 0.7)
```

Pretty.

It may be useful to understand how trees are encoded in R. Typing in just the name of the tree file like this:

```
anoleTree
```

will give you basic information about the phylogeny: the number of tips and nodes; what the tips are called; whether the tree is rooted; and if it has branch lengths.

```
str(anoleTree)
```

will tell you more about tree structure. Trees consist of tips connected by edges (AKA branches)

```
anoleTree$tip.label
```

gives you a list of all your terminal taxa, which are by default numbered 1-n, where n is the number of taxa.

```
anoleTree$Nnode
```

gives you the number of nodes. This is a fully bifurcating rooted tree, so it has 1 fewer node than the number of taxa.

```
anoleTree$edge
```

This tells you the beginning and ending node for all edges. Put that all together with the following

```
plot.phylo(anoleTree, type = "fan", no.margin = TRUE, cex = 0.7, label.offset = 0.1)
nodelabels(cex = 0.5)
tiplabels(cex = 0.5)
```

There are many ways to manipulate trees in R using Ape, Phytools, and other packages. This just gives you a bare-bones introduction.

Working with a data matrix and testing hypotheses in a phylogenetically informed way

Let's ask some questions using the trait data that were measured for these lizards. First, explore the data in the "anoleData" matrix. Here are some options for visualizing data matrices:

```
head(anoleData) # this will show you the first few rows of your data matrix and its header
dimnames(anoleData) # this will show you the row and column headers for your matrix
View(anoleData) # this will let you visualize the entire matrix
```

After looking at the data, please answer the next questions

What variables were measured for each of these species of Anolis? How many species of Anolis were used?

Is that the same number that were in your phylogeny?

Awesomeness is one of your variables. Let's isolate it so we can work with it easily:

```
awe <- anoleData[, "awesomeness"]
names(awe) <- rownames(anoleData)
# data vectors have to be labelled with tip names for the associated tree.
# This is how to do that.
```

In lecture, we talked about one model of character evolution, called a Brownian Motion model. This model assumes that a trait evolves from a starting state (**z0**) according to a random walk with the variance specified by the rate parameter σ^2 (**sigma-squared**). In short, Brownian motion describes a process in which tip states are modelled under the assumption of a multi-variate normal distribution. On a phylogeny, the multi-variate mean of tip states is equal to the root state estimate, and variance accummulates linearly through time.

What does Brownian Motion evolution of Awesomeness in lizards look like?

```
brownianModel <- fitContinuous(anoleTree, awe)
brownianModel # this will show you the fit statistics and parameter values
```

Here, you can see the estimates for ancestral state (z0), and the rate parameter (σ^2), as well as some measures of model fit. The fit of the model is determined using maximum likelihood, and expressed as a log likelihood. The higher the lnL, the more probable the data given the model. However, when comparing different models, we can't use the lnL, because it does not account for the difference in the number of parameters among models. Models with more parameters will always fit better, but do they fit significantly better? For example an OU model has 4 parameters (alpha [α], theta [θ], z0, and sigma-squared [σ^2]), so it should fit better than a BM model, which includes only z0 and sigsq. To account for this, statisticians have developed another measure of fit called the AIC (Akaike Information Criterion): **AIC = (2xN)-2xlnL**, where **N** is the number of parameters. This penalizes the likelihood score for adding parameters. When selecting among a set of models, the one with the lowest AIC is preferred. We will use this information later on in this lab.

In addition to assessing model fit, we can use the Brownian Motion model to reconstruct ancestral states of a character on a tree. To visualize what BM evolution of this trait looks like on a tree. The *contMap()* command in phytools estimates ancestral states and plots them on a tree.

```
contMap(anoleTree, awe, fsize = 0.5, lwd = 3)
```

Describe the evolution of awesomeness on this tree. How many times have exteremely high and extremely low awesomeness evolved on this tree?

What does this say about our ability to test hypotheses about the evolution of awesomeness?

Let's go ahead and test some hypotheses. Hostility is another trait in your data matrix. Let's assess whether there is a correlation between how hostile a lizard is and how awesome it is? We will extract the column "hostility" from the datamatrix and assign it species names, just as we did for "awesomeness" above.

```
host <- anoleData[, "hostility"]
names(host) <- rownames(anoleData)
```

Let's look at a plot of awesomeness as a function of hostility

```
plot(host, awe, xlab = "Hostility", ylab = "Awesomeness")
```

Hm. looks promising. *How would you describe the relationship between these two variables?* Let's be more quantitative in describing that realtionship with a linear model.

```
lm_awehost <- lm(awe ~ host)
summary(lm_awehost)
```

```r
plot(awe ~ host, xlab = "Hostility", ylab = "Awesomeness")
abline(lm_awehost)
```

The coefficients table from the *summary()* command shows the slope and intercept for the linear model describing awesomeness as a function of hostility. Each line shows the estimated coefficient (Estimate), the standard error (Std. Error) of that estimate, as well as a t-statistic and associated p-value, testing whether those parameters are equal to 0. The Multiple R-squared is an estimate of how much variance in the response variable can be explained by the predictor variable.

Write the linear model for this relationship. Are the parameters significantly different from 0?

What is the R ˆ2 value for this data?

How do you feel about that?

Nice. But. We have not considered the fact that these lizards are related to each other. As such, they may share their hostility and awesomeness simply due to the fact that their ancestors were hostile or awesome. In other words, we need to account for non-independence of residuals due to phylogeny. One way to do that is to use phylogenetic-generalized-least-squares regression (PGLS)

```r
pglsModel <- gls(awesomeness ~ hostility, data = anoleData,
                 correlation = corBrownian(phy = anoleTree), method = "ML")
```

Let's break this command down. This command infers a linear model for awesomeness as a function of hostility (gls(awesomness ~ hostility, data = anoleData)), but it specifies existing correlation structure in the data (correlation =) as the covariance of these traits assuming a Brownian motion model (corBrowinan()) based on the anolis tree (phy = anoleTree). The model is fit using maximum likelihood (method = "ML"). To see the results:

```r
summary(pglsModel)
coef(pglsModel)
R2(pglsModel)
```

```r
plot(awe ~ host, xlab = "Hostility", ylab = "Awesomeness")
abline(a = coef(pglsModel)[1], b = coef(pglsModel)[2])
# will plot the pgls regression line on your biplot.
```

Write the linear model for this relationship. Are the parameters significantly different from 0?

What is the R ˆ2 value for this data?

How do you feel about that?

Compare results from the pgls analysis with those that you got from the regular linear model you ran earlier.

Model Fitting

Brownian Motion is only one model of evolution for a continuous variable. Another model is the Ornstein-Uhlenbeck (OU) model, which allows the trait mean to evolve towards a new state (theta), with a selective force (alpha). These two new parameters, plus the starting state (z0) and the rate of evolution (sigsq) parameters from the BM model, make for a 4-parameter model. The Early Burst model (EB) model allows the rate of evolution to change across the tree, where the early rate of evolution is high and declines over time (presumably as niches are filled during an adaptive radiation. The rate of evolution changes exponentially over time and is specified under the model r[t] = r[0] x exp(a x t), where r[0] is the initial rate, a is the rate change parameter, and t is time. The maximum bound is set to -0.000001, representing a decelerating rate of evolution. The minimum bound is set to $log(10^{-5})$/depth of the tree.

Let's evaluate the relative fit of these three models to the Awesomeness trait.

```
brownianModel <- fitContinuous(anoleTree, awe)
```

```
OUModel <- fitContinuous(anoleTree, awe, model = "OU")
```

```
EBModel <- fitContinuous(anoleTree, awe, model = "EB")
```

And recover the parameter values and fit estimates.

```
brownianModel
OUModel
EBModel
```

```
aicw(c(brownianModel$opt$aicc, OUModel$opt$aicc, EBModel$opt$aicc))
```

Make a table with the AIC and lnL values for each model. Which model provides the best fit for awesomeness?

Now, add the results for a model fitting analysis of the Hostility trait to this table.

So, we were wrong. An OU model fits these data better (and you should be able to explain how we know that). Unfortunately, a PGLS analysis with an OU model specified is currently computationally difficult. The best we can do is report the results from our model fitting analysis, and realize that the parameters from BM might not be the best fit.

However, we can still test our hypothesis that hostile lizards are less awesome, and account for phylogeney when we do. First, we should compare the uncorrected linear model of awesomeness as a fuction of hostility vs the PGLS that uses the covariance structure of the residuals under a Browian Motion model.

```
plot(host, awe, xlab = "hostility", ylab = "awesomeness",
     main = "Awesomeness as a Function of Hostility")
abline(lm_awehost, lty = 2) #uncorrected LM
abline(a = coef(pglsModel)[1], b = coef(pglsModel)[2]) #BM
legend("topright", lty = c(1, 2), legend = c("PGLS", "uncorrected"))
```

You might want to know if these regressions really differ in their ability to predict awesomeness from hostility. Asked in another way, are the slopes from these two regressions significantly different from each other? You need to know that a 95% confidence interval for the slope parameter is b (the slope) plus/minus 1.96 standard errors (this is derived from a normal distribution). To calculate your 95% confidence intervals:

```
awehost.sum <- summary(lm_awehost)
awehost.sum$coef[2, 1]+c(-1, 1)*awehost.sum$coef[2, 2]
#for the uncorrected linear model
coef(pglsModel)[2]+c(-1, 1)*sqrt(pglsModel$varBeta[2, 2])
#for Brownian Motion, the 95% CI
```

Did phylogenetic correction make a difference in this case?

What do you conclude about the evolution of awesomeness as a function of hostility?

Discrete Character Mapping

So far, we've been dealing with continuous characters, those that take values along some continuum. Things like height, weight, length, temperatue, humidity, etc. are continuous variables. There is another type of variable called a discrete variable, that takes, well, discrete values. Color (e.g. red, blue, green); Locomotory type (e.g. scansoial, terrestrial, fossorial) are examples of discrete variables.

Look at the data in the anoleData matrix. *Which of these variables are discrete and which are continuous?*

In your data matrix, Island has been coded for each of these species. Let's examine some biogeography for these lizards by reconstructing the ancestral island (i.e. area of origin) and dispersal history. First, isolate your variable.

```
island <- anoleData$island
names(island) <- rownames(anoleData)
```

We can simultaneously fit a model of discrete character evolution and create a set of plausible character histories using a method called stochastic character mapping:

```
island_anc <- make.simmap(anoleTree, island, model = "SYM", nsim = 100)
```

This analysis results in a "Q" matrix showing the relative probabilities of change from state to state. For this character, this would represent dispersal events between islands. The higher the value, the higher the probability of that type of change.

Which pair of islands shows the highest probability of interchange?

Which pair shows the least?

Does this make sense geographically?

Now, you can plot a random simulation of change in this character, that is based on the values inferred above.

```
plotSimmap(island_anc[[1]], fsize = 0.5) # this plots the first of 100 simulations
```

How many transitions are there between black and red?

Are there any reversals?

Are there any branches with more than one change?

Using the command above, but changing the index from 1 to other values, look at a number of reconstructions. How much variation do you see?

We can summarize these simulations and estimate the relative probability of each island as an ancestor for each node on our tree:

```
island_summary <- summary(island_anc)
plot(island_summary, cex = c(0.5, 0.2), fsize = 0.5, offset = 90)
legend("bottomleft", fill = c("black", "red", "green", "blue"),
       legend = c("Cuba", "Hispanola", "Jamaica", "PR"))
```

Where did Anoles likely originate?

How many transitions from Cuba to Hispanola?

Are there any reversals?

Does a Jamaican ancestor ever move to Puerto Rico?

What is the ancestral island for the ancestor of A. distichus and A. evermanni?