Lab 08: Phylogenetics in R: Trait Evolution in a phylogenetic context

Edited by Ixchel Gonzalez-Ramirez

3/10/2020

Integrative Biology 200

Principles of Phylogenetics

University of California, Berkeley

Introduction

In the last lab we learnt the basics of working with phylogenetic data in R, and mapped traits on the tips of a phylogeny. Today we will take a step further. When we study traits, we often want to reconstruct the history of the trait based on observations. The process of inferring the states of a trait in internal nodes from observations made on tips of a given phylogeny is called Ancestral State Reconstruction.

First of all make sure you have installed and loaded the libraries we will need

```
#installing packages, commented since you should already have them

#install.packages("ape")

library(ape)
```

Part I: Maximum Likelihood Ancestral State Reconstruction of Discrete Characters

In R we have used a number of functions. Almost every command you have run is a function that takes certain arguments as the inputs and produces some output, such as a value, a datatable, or a plot. We can also define our own functions. Here's an example of a silly function: Taking the mean of a vector with some element NA will produce NA as the answer unless you specify to remove NA values when calculating the mean. To get around that, let's create our own function that removes the NA elements automatically.

```
mean_na <- function(vec) { # Create a function called "mean_na", taking as input a vector

  m <- mean(vec, na.rm = T) # m is an object containing the mean omitting NAs

  return(m) #when the function is used, the value m should be returned/printed

}
```

You can test out our new function:

```
test <- c(1,3,4,NA, 8, 1)
```

```r
mean(test)

mean_na(test)
```

Exercise A

Today's lecture describes a continuous-time Markov model for a binary character that has two rates of change: alpha is the rate of transitioning from state 0 to 1, and beta is the rate of transitioning from state 1 to 0. See: http://ib.berkeley.edu/courses/ib200/lect/lect17.pdf.

Write two functions to calculate the probability of each possible transition over a branch of length t. Each function should take as input alpha, beta, and t. As a reminder, alpha and beta are the instantaneous rates of change - your function should calculate the probabilites. Insert your code below.

```r
#insert your code here
```

Great! Now, let's use those functions to estimate the probabilities of states on a simple tree.

```r
t = read.tree(text="((A:0.39,B:0.39):0.93,C:1.32);")

plot(t, show.tip.label=FALSE, main = "Test Tree")

nodelabels()

tiplabels()

edgelabels(t$edge.length)
```

Exercise B

Assume alpha = 2 and beta = 3. Now estimate the probabilities that the tree has the following states:

Node 1 = 0
Node 2 = 0
Node 3 = 1
Node 4 = 1
Node 5 = 0

Hint: the probability of the tree is the product of the probability of each branch. Insert your code below.

```r
#insert your code here
```

It would be a pain if we had to do this manually every time! Thankfully, there are packages with functions that can do these calculations for us. We'll use an example from Liam J. Revell's Phytools package. Phytools is one of the most commonly used R packages for phylogenetic comparative methods, and the Phytools blog can be incredibly helpful: http://blog.phytools.org/

We will get some sample data from the Phytools package. The anoletree dataset. If you are curios to learn more, type ?anoletree in your console.

```
library(phytools)

data(anoletree)

x <- getStates(anoletree, "tips")

tree <- anoletree
```

Plot the data on the tree!

```
plotTree(tree, type = "fan", fsize = 0.9, ftype = "i")

cols <- setNames(palette()[1:length(unique(x))], sort(unique(x)))

tiplabels(pie = to.matrix(x, sort(unique(x))), piecol = cols, cex = 0.2)

add.simmap.legend(colors = cols, prompt = FALSE,
                  x = 0.9 * par()$usr[1], y = -max(nodeHeights(tree)))
```

Here, we will fit a single-rate continuous time Markov model to the data and estimate the ancestral nodes of our tree. In the above exercise, we asked you to calculate the probabilities using a two-rate model. Since there are 6 character states, the instantaneous rate matrix looks like this:

```
-  a  a  a  a  a
a  -  a  a  a  a
a  a  -  a  a  a
a  a  a  -  a  a
a  a  a  a  -  a
a  a  a  a  a  -
```

where the diagonals are -5a

Fit the model using the following code. "ER" means equal rates! Then calculate the log likelihood of the model.

```
fitSR <- rerootingMethod(tree, x, model = "ER")

fitSR$loglik #the likelihood of the model

fitSR$Q #Qmatrix
```

We can plot the results of our ancestral state reconstruction, now showing the ancestral states at the nodes.

```
plotTree(tree, type = "fan", fsize = 0.9, ftype = "i")

nodelabels(node = as.numeric(rownames(fitSR$marginal.anc)),
           pie = fitSR$marginal.anc, piecol = cols, cex = 0.5)

tiplabels(pie = to.matrix(x, sort(unique(x))), piecol = cols, cex = 0.2)

add.simmap.legend(colors = cols, prompt = FALSE,
                  x = 0.9 * par()$usr[1], y = -max(nodeHeights(tree)))
```

Now let's fit a symmetrical-rates model, that allows different (but symmetrical) rates for each character state transition. For a symmetrical-rate 6 state model, the transition rate matrix looks like:

```
-   a   b   c   d   e
a   -   f   g   h   i
b   f   -   j   k   l
c   g   j   -   m   n
d   h   k   m   -   o
e   i   l   n   o   -
```

where the diagonal elements are defined as -1 times the sum of the other row elements, for example, row 1's diagonal element is -(a + b + c + d + e).

We fit the symmetrical rates model using the same code as above, but specifying "SYM" for symmetrical instead of "ER" for equal rates. Also calculate the log likelihood of this model.

```
fitSYM <- rerootingMethod(tree, x, model = "SYM")

fitSYM$loglik #the likelihood of the model
fitSYM$Q #prints Q matrix
```

Exercise C

Plot the tree again, but show the ancestral states inferred using the symmetrical-rates model. Send me screen shots of both reconstructed ancestral states. Even though the single-rate model is simply a special case of the symmetrical-rates model, are the ancestral state reconstructions the same?

Which model fits the data better? Calculate the likelihood ratio test: D = 2 * (loglike of alternative model - loglike of null model). There are 15 parameters in the alternative (SYM) model and 1 parameter in the SR model, so we have 15 - 1 = 14 degrees of freedom. Look up D in a chi-squared distribution table and report the p-value. Is the SYM model supported over the SR model?

```
#insert your code here
```

Part 2: Correlated Evolution of Discrete Traits

Pagel (1994) described an elegant model to test for correlated evolution of discrete traits. Here, we can model two binary traits (with states 0 and 1) as one combined multistate trait (with states 00, 01, 10, 11), illustrated in the table below.

We can then perform an ancestral state reconstruction and estimate the rates of transition between all four 'states'. Figure 1 illustrates the basic set-up of the model. As you can see, there are separate parameters for transitioning between all 4 states. Let's think about a biological example. Suppose we are interested in testing for a correlation between two binary traits: presence/ absence of tubers and presence/ absence of rhizomes. Are plants that have tubers more likely to also have rhizomes? We can rephrase this question as, does the rate of evolving rhizomes depend on if the plant already has tubers? If the traits are correlated, we expect that the rate of evolving tubers (trait 1) depends on the presence of rhizomes (trait 2).

We can compare the statistical fit of the fully correlated model (Fig. 1) to the statistical fit of an alternative model. In the alternative model, we force some rates to be equal such that the rate of transitioning between states of one character is the same, regardless of the state of the other character. In our biological example above, we constrain the rates of evolving tubers with and without rhizomes to be the same.

Let's simulate some fake data and give it a try.

First, let's simulate a tree with 200 tips.

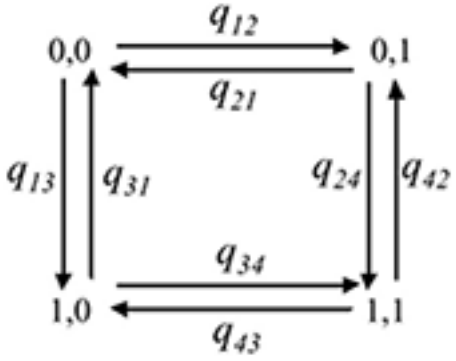

Figure 1: Image from Pagel and Meade (2006). The arrows indicate transition rates between states.

```
tree <- pbtree(n = 200, scale = 1)

plotTree(tree, type = "fan")
```

Now, let's simulate two traits evolving independently. We do this by building a rate matrix and simulating the evolution of a binary trait, twice. Keep in mind that even though we use the same rate matrix, the traits are still evolving independently.

```
Q <- matrix(c(-1,1,1,-1),2,2)

rownames(Q) <- colnames(Q) <- letters[1:2]

Q

binary1 <- sim.history(tree,Q)

binary2 <- sim.history(tree,Q)
```

Let's plot the results of our simulation to see if they look correlated or not. Remember, we simulated the data separately, so they really shouldn't be correlated!

```
par(mfrow=c(1,2));

plotSimmap(binary1, setNames(c("blue","red"),
                             letters[1:2]), ftype="off", lwd=1)

plotSimmap(binary2, setNames(c("blue","red"),
                             letters[1:2]), ftype="off", lwd=1, direction="leftwards")
```

And now, let's fit Pagel's test for correlated evolution to the data:

```
x <- binary1$states

y <- binary2$states

fit <- fitPagel(tree, x, y)

fit
```

Exercise D

Take a close look at this output. Does your test indicate that traits are correlated or not? Which model was favored, and was the difference statistically significant?

Part 3: Simulating Continuous Character Evolution Under Brownian Motion

We often model the evolution of continuous characters using Brownian Motion. Here, we will model the evolution of a continuous character using Brownian Motion. Then, we will show how to use Brownian Motion to study the evolution of continuous characters.

First, we will initialize our simulations by setting the length of time for the simulation (t). This can also be thought of as the number of generations.

```
t <- 0:100
```

Next, let's initialize the instantaneous rate of change. This determines the relative size of jumps that tend to occur. When sig2 (sigma squared) is small, character state changes tend to be smaller. when its big (maximum of 1), individual state changes tend to be larger.

```
sig2 <- 0.01
```

Now, simulate a set of random deviates. In other words, this is the series of character state changes through time (t). Look at the values of x to get an idea of this.

```
x <- rnorm(n = length(t) - 1, sd = sqrt(sig2))

x
```

The character state starts at 0 then changes by the amount (x). Imagine this is the size of some trait, like leaf length. when x is negative, it means it got smaller, and when positive, it got bigger.

Now compute their cumulative sum. We want the cumulative sum because the state a character is in at any time (t) is the sum of all its past transitions. In other words, the character started at "0", got bigger by this much, smaller by that much, bigger again, etc. You add those all up to see the state at the end! If we plot all of the cumulative sums over the values of t, we can track the changes in the trait over time.

We can build up to this by starting to generate trait values over time. Each trait value at time t is calculated by adding a change in trait value given in the X vector to the trait value at t-1.

```
stepwise_values <- numeric()

stepwise_values[1] <- 0

stepwise_values[2] <- stepwise_values[1] + x[1]

stepwise_values[3] <- stepwise_values[2] + x[2]

stepwise_values[4] <- stepwise_values[3] + x[3]

stepwise_values[5] <- stepwise_values[4] + x[4]
```

Now that we've generated the trait values for the first 5 points in time, we can plot those points to see how the trait values vary as time progresses.

```
abs_max <- max(abs(stepwise_values))
```

```
plot(1, stepwise_values[1],
     xlim = c(0,5), ylim = c(-abs_max,abs_max),
     xlab = "t", ylab = "trait value",
     pch = 20, type = "o")

points(c(2:5), stepwise_values[2:5], pch = 20)

lines(1:5, stepwise_values)
```

We can do this for all of our values of X using the cumsum() function.

```
x <- c(0, cumsum(x))

plot(t, x, type = "l", ylim = c(-3, 3))
```

Repeat these simulations several times by rerunning the previous code., writen in the following chunk You'll be drawing new random deviates, so the simulations should look slightly different from each other even though we haven't changed the parameters at all.

```
x <- rnorm(n = length(t) - 1, sd = sqrt(sig2))

x <- c(0, cumsum(x))

plot(t, x, type = "l", ylim = c(-2, 2))
```

The next section of code runs several independent simulations and plot all at once. This can be taught as several lineages evolving independently at the same time.

First, we generate a matrix of simulations, with each simulation as a row and each unit time (t) as a column.

```
t <- 0:100

sig2 <- 0.01

nsim <- 100

X <- matrix(rnorm(n = nsim * (length(t) - 1), sd = sqrt(sig2)), nsim, length(t) - 1)

sim_matrix <- cbind(rep(0, nsim), t(apply(X, 1, cumsum))) #matrix of simulations
```

Then, we plot the first simulation in red and the rest of the simulations in black.

```
plot(t, sim_matrix[1, ], xlab = "time", ylab = "phenotype", ylim = c(-3, 3), type = "l")

apply(sim_matrix[2:nsim, ], 1, function(x, t) lines(t, x), t = t)

lines(t, sim_matrix[1, ], xlab = "time", ylab = "phenotype", ylim = c(-3, 3), col = "red")
```

Exercise D

Try this a couple of times, again changing sig2 and t. You might need to change the y axis to fit your character state range! (change the numbers in 'ylim = c(-3, 3)' to whatever you want. Include one example of your changed code in a code block below, along with the answers to the following questions:

1) How does the size of t affect how close the final character state (phenotype) is to the initial character state (which is always 0)?

2) How does the size of sig2 affect how close the final character state (phenotype) is to the initial character state?

```
#insert your code here
```

How does all of this connect to Brownian motion in the context of phylogenetics? We are never given the full distribution of thousands of Brownian motion simulations. Instead, we are given a couple of trait values in the present, and are often interested in reconstruction the ancestral value - the start point of the simulation. Here's a visualization of that.

```
plot(t, sim_matrix[1, ], xlab = "time", ylab = "phenotype",
     ylim = c(-2, 2), type = "l", col = "grey")

apply(sim_matrix[2:nsim, ], 1, function(x, t) lines(t, x, col = "grey"), t = t)

points(0,0, col = "blue", pch = 20)

text(10,.5, "Ancestral \nValue", col = "blue")

points(x = rep(100, times = 5),
       y = sim_matrix[c(1:5),100],
       col = "red", pch = 20)

text(90, mean(sim_matrix[c(1:5),100]), "Extant\nTrait\nValues", col = "red")
```

So, how does Brownian motion help us understand trait evolution? How can we calculate the ancestral value given the extant trait values? Let's go through the following questions to relate this all back to evolution.

Exercise E

Here's another simulation under Brownian motion. Add 4 additional features to the plot below:

- A horizontal line (in blue) showing the starting trait value. This is also known as the expected value for this Brownian motion simulation.
- 3 verticle lines (in red) illustrating the variance of the distribution at t = 20, t = 60, and t = 100. For example, in the plot below, I've included the variance of the distribution at t = 10. Add on the additional lines to this plot.

```
nsim <- 1000

X <- matrix(rnorm(n = nsim * (length(t) - 1), sd = sqrt(sig2)), nsim, length(t) - 1)

sim_matrix <- cbind(rep(0, nsim), t(apply(X, 1, cumsum)))

plot(t, sim_matrix[1, ], xlab = "time", ylab = "phenotype",
     ylim = c(-5, 5), type = "l", col = "grey")

apply(sim_matrix[2:nsim, ], 1, function(x, t) lines(t, x, col = "grey"), t = t)

segments(x0 = 10, y0 = min(sim_matrix[,10]), y1 = max(sim_matrix[,10]), col = "red" )
```

```
#insert your code here for the completed plot
```

Here's a graphical illustration of the relationship between t and the variance among simulations.

```
v <- apply(sim_matrix, 2, var)

plot(t, v, type = "l", xlab = "time", ylab = "variance among simulations")
```

> **In these simulations, we have discussed the relationships between time, trait values/ phenotype, and variance among simulations. For each of these 3 terms, describe how they would relate to a phylogenetic ancestral state reconstruction of a quantitative trait such as leaf size.**

There's a function for simulating Brownian Motion over a phylogeny in Phytools. Given a tree and your Brownian Motion parameters, you can use this function to simulate under the model. Here, we'll simulate a tree, simulate a character evolving over the tree, and plot the tree and associated character values as a 'traitgram' (Ackerly, 2009).

```
tree <- rcoal(n = 30)

x <- fastBM(tree, a=0, sig2=1.0, internal = TRUE)

phenogram(tree, x, spread.labels = TRUE)
```

Let's now simulate a character evolving under the Ornstein-Uhlenbeck process. The OU model is Brownian motion but with two extra parameters: the optimum (theta) and the strength of selection (alpha). We simulate under this model using the same fastBM model as above.

When alpha is close to 0, the OU model collapses down to Brownian motion:

```
x <- fastBM(tree, a=0, sig2=1.0, alpha=0.01, theta=4.0, internal = TRUE)

phenogram(tree, x, spread.labels = TRUE)
```

As alpha increase, we see more of a trend towards the optimum:

```
x <- fastBM(tree, a=0, sig2=1.0, alpha=0.5, theta=4.0, internal = TRUE)

phenogram(tree, x, spread.labels = TRUE)
```

Alpha acts as a 'rubber band', pulling the trait to the optimum:

```
x <- fastBM(tree, a=0, sig2=1.0, alpha=2.0, theta=4.0, internal = TRUE)

phenogram(tree, x, spread.labels = TRUE)
```

The above examples of Brownian motion and Ornstein-Uhlenbeck assume that the same model applies to the entire tree. It is also possible to have time or branch heterogenous models, in which the parameter values or model changes over the tree. If you are interested in these models, check out the R packages OUwie, ouch, and bayou.

Now that we have a good understanding of Brownian Motion, let's apply the model to reconstruction ancestral states for continous traits. We will use sample data on Anoles again.

```
anole_tree<-read.tree("http://www.phytools.org/eqg2015/data/anole.tre")

svl <- read.csv("http://www.phytools.org/eqg2015/data/svl.csv",
    row.names=1)
```

```
svl <- as.matrix(svl)[,1] #convert dataframe to a vector
```

We've read in the tree and character data. Feel free to take a look at those data files to be sure you're comfortable with how they are formatted, etc. Now, we will use the fastAnc funciton in Phytools to reconstruct the ancestral states using maximum likelihood. You can read more about the specifics of the function by looking at the help page. contMap uses the same method as fastAnc but projects those values along the branches of the tree to produce the figure showing changes along branches. Again, read more in the help page for the function if you are interested.

```
fit <- fastAnc(anole_tree,svl,vars=TRUE,CI=TRUE)

fit_obj <- contMap(anole_tree, svl, plot=FALSE)

plot(fit_obj, type="fan", legend=0.7*max(nodeHeights(anole_tree)),
    fsize=c(0.7,0.9))
```

There are several functions that reconstruct ancestral states in R. Another commonly used function is ace from the ape package. Feel free to explore these additional functions if you like.

Part 4: Correlated Evolution of Continuous Characters

How do we test if two traits are evolving in a correlated manner? Let's simulate a tree and two traits independently evolving on the tree:

```
tree <- rcoal(n = 100)

x <- fastBM(tree)

y = fastBM(tree)
```

Are the traits correlated if we ignore the phylogeny?

```
plot(x, y)

abline(lm(y ~ x))

cor.test(x, y)
```

I just got p-value = 0.02679, but we know these two traits evolved independently (because we simulated them that way! You can see how easy it is to get a type I error (false positive). As an alternative, let's use phylogenetically independent contrasts (PIC; Felsenstein 1985) to estimate the evolutionary correlation between characters:

```
x_c <- pic(x, tree)

y_c <- pic(y, tree)

cor.test(x_c, y_c)
```

Now I got a p-value = 0.4457, so we correctly reject the hypothesis that the two characters were correlated.

Content in this lab is drawn from the IB200 2016 lab by Will Freyman, the Phytools blog, and an exercise co-written by Jenna Baughman and Carrie Tribble.