Phylogenetic Diversity - Traits

Student Name; Z620: Quantitative Biodiversity, Indiana University

20 February, 2015

OVERVIEW

Up to this point, we have been focusing on patterns taxonomic diversity in Quantitative Biodiversity. Although taxonomic diversity is an important dimension of biodiversity, it does not consider the evolutionary history or relatedness of species. The goal of this exercise is to introduce basic concepts of phylogenetic diversity.

After completing this exercise you will know how to:

1. create phylogenetic trees
2. map traits onto phylogenetic trees
3. test for phylogenetic signal of traits on a phylogenetic tree

Directions:

1. Change "Student Name" on line 3 (above) with your name.
2. Complete as much of the exercise as possible during class; what you do not complete in class will need to be done on your own outside of class.
3. Use the handout as a guide; it contains a more complete description of data sets along with the proper scripting needed to carry out the exercise.
4. Be sure to **answer the questions** in this exercise document; they also correspond to the handout. Space for your answer is provided in this document and indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">".
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For homework, follow the directions at the bottom of this file.
7. When you are done, **Knit** the text and code into a PDF file.
8. After Knitting, please submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file *PhyloTraits_exercise.Rmd* and the PDF output of `Knitr` (*PhyloTraits_exercise.pdf*).

1) SETUP

Typically, the first thing you will do in either an R script or an RMarkdown file is setup your environment. This includes things such as setting the working directory and loading any packages that you will need.

In the R code chunk below, provide the code to:
1. clear your R environment,
2. print your current working directory,
3. set your working directory to your "*/PhyloTraits*" folder, and
4. load all of the required R packages (be sure to install if needed).

2) DESCRIPTION OF DATA

The maintenance of biodiversity is thought to be influenced by **trade-offs** among species in certain functional traits. One such trade-off involves the ability of a highly specialized species to perform exceptionally well on a particular resource. In this exercise, we will take a phylogenetic approach to mapping phosphorus resource use onto a phylogenetic tree while testing for specialist-generalist trade-offs.

3) SEQUENCE ALIGNMENT

A. Examining a FASTA-File

Using nano, open the *p.isolates.fasta* file by typing "nano p.isolates.fasta" at the command line (you need to navigate to that file first). This will open the FASTA file in nano, which is a text editing program. You can use [Ctrl+V] and [Ctrl+Y] to scroll towards the bottom and top of the file, respectively. Use these commands to look at the formatting on the FASTA file. When you are done, you can close the file by hitting [Ctrl+X]. Do not save any inadvertent changes to *p.isolates.fasta*.

B. Performing an Alignment

In the R code chunk below, do the following:
1. use `muscle` to align the p.isolates.fasta sequences, and
2. save the output as p.isolates.afa in the */data* folder.

Question 1: Using nano or your favorite text editor, compare the `p.isolates.fasta` file and the `p.isolates.afa` file. Describe the differences that you observe between the files.

Answer 1:

C. Visualizing the Alignment

In the R code chunk below, do the following:
1. read your alignment file,
2. convert the agignment to a DNAbin object,
3. select a region of the gene to visualize (try various regions), and
4. plot the alignment.

Question 2: Make some observations about the `muscle` alingmnet of the 16S rRNA gene sequences for our bacterial isolates and the outgroup, *Methanosarcina*, which belongs to the archaeal domain of life. Move along the alignment by changing the values in the `window` object.

a. Approximatley how long were our reads?

b. What regions do you think would be most appropriate for phylogenetic inference?

Answer 2a:
Answer 2b:

4) MAKING A PHYLOGENETIC TREE

Once you have aligned your sequences, the next step is to construct a phylogenetic tree. Not only is a phylogenetic tree effective for visualizing the evolutionary relationship among taxa, but as you will see later, the information that goes into a phylogenetic tree is needed for downstream analysis.

A. Neighbor Joining Trees

In the R code chunk below, do the following:
1. calcualte the distance matrix using `model = "raw"`,
2. create a Neighbor Joining tree based on these distances,
3. define "Methanosarcina" as the outgroup and root the tree, and
4. plot the rooted tree.

B) SUBSTITUTION MODELS OF DNA EVOLUTION

In the R code chunk below, do the following:
1. make a second distance matrix based on the Fellenstein substitution model,
2. create a saturation plot to compare the *raw* and *Fellenstein* substitution models,
3. make Neighbor Joining trees for both, and
4. create a cophylogenetic plot to compare the topologies of the trees.

In the R code chunk below, do the following:
1. pick another substitution model,
2. create and distance matrix and tree for this model,
2. make a saturation plot that compares that model to the *Fellenstein* model,
3. make a cophylogenetic plot that compares the topologies of both models, and
4. be sure to format, add appropriate labels, and customize each plot.

Question 3: Using the saturation plot and cophylogenetic plots from above, describe the effect that the F84 substitution model has on our phylogenetic reconstruction. If the plots seem to be inconsistent with one another, explain what it giving rise to the differences.

Answer 3:

5) INTEGRATING TRAITS AND PHYLOGENY

A. Loading Trait Database

In the R code chunk below, do the following:
1. import the raw phosphorus growth data, and
2. stardardize the data for each strain.

B. Trait Manipulations

In the R code chunk below, do the following:
1. calculate the maximum growth rate (μ_{max}) of each isoalte across all phosphorus types,
2. create a function that calculates niche breadth (nb), and
3. use this function to calcuate nb for each isolate.

C. Visualzing Traits on Trees

In the R code chunk below, do the following:
1. pick your favorite substitution model and make a nj tree,
2. define your outgroup and root the tree, and
3. remove the outgroup branch.

In the R code chunk below, do the following:
1. define a color palette (use something other than "YlOrRd"),
2. map the phosphorus traits onto your phylogeny,
3. map the nb trait on to your phologeny, and
4. customize the plots as desired (use `help(table.phylo4d)` to learn about the options).

Question 4: Based on the distribution of traits on the neighbor joining tree that we created, what might you predict about the degree of specialization of bacteria on diverse phosphorus resources?

Answer 4:

6) HYPOTHESIS TESTING

A) Phylogenetic Signal: Pagel's Lambda

In the R code chunk below, do the following:
1. create two rescaled phylogenetic trees using lambda values of 0.5 and 0,
2. plot your original tree and the two scaled trees, and
3. label and customize the trees are desired.

In the R code chunk below, do the following:
1. use the `fitContinuous()` function to compare your original tree to the transformed trees.

Question 5: There are two important outputs from the `fitContinuous()` function that can help you interpret the phylogenetic signal in trait data sets. First, compare the lambda values of the untransformed tree to the transformed (lambda = 0). Second, compare the Akaike information criterion (AIC) scores of the two models. Differences in AIC scores provide maximum likelihood inference regarding the quality of model fit to a given data set. Models with lower AIC values are better. However, if the difference in AIC values between two models (Δ AIC) isn't greater than at least 2, then then the models are considered equivalent. With this information in hand, what does Pagel's lambda say about the phylogenetic signal of niche breadth in our data set?

Answer 5:

B) Phylogenetic Signal: Blomberg's K

In the R code chunk below, do the following:
1. correct tree branch-lengths to fix any zeros,
2. calculate Blomberg's K for each phosphorus resource using the `phylosignal()` function,
3. use the Benjamin-Hochberg method to correct for false discovery rate, and
4. calculate Blomberg's K for nich breadt using the `phylosignal()` function.

Question 6: Using the K-values and associated p-values (i.e., "PIC.var.P"") from the `phylosignal` output, answer the following questions:

a. Is there significant phylogenetic signal for niche breadth or standardized growth on any of the phosphorus resources?

b. If there is significant phylogenetic signal, are the results suggestive of clustering or overdispersion?

Answer 6a:
Answer 6b:

C. Calculate Dispersion of a Trait

In the R code chunk below, do the following:
1. turn the continuous growth data into categorical data,
2. add a column to the data with the isolate name,
3. combine the tree and trait data using the `comparative.data()` function in `caper`, and
4. use `phylo.d()` to calculate D on at least three phosphorus traits.

Question 7: Using the estimates for D and the probabilities of each phylogenetic model, answer the following questions:

a. Choose three phosphorus growth traits and test whether they are significantly clustered or overdispersed?

b. How do these results compare the results from the Blomberg's K analysis?

c. Discuss what factors might give rise to differences between the metrics.

Answer 7a:
Answer 7b:
Answer 7c:

7) HOMEWORK

1. Your handout includes the output of a multiple regression model depicting the relationship between the maximum growth rate (μ_{max}) of each bacterial isolate and the niche breadth of that isoalte on the 18 different sources of phosphorus. One feature of the study which we did not reveal earlier in the handout is that the isolates came from two different lakes. One of the lakes is an very oligotrophic (i.e., low phosphorus) ecosystem named Little Long (LL) Lake. The other lake is an extremely eutrophic (i.e., high phosphorus) ecosystem named Wintergreen (WG) Lake. We included a "dummy variable" (D) in the multiple regression model (0 = WG, 1 = LL) to account for the environment from which the bacteria were obtained.
   Based on your knowledge of the traits and their phylogenetic distributions, what conclusions would you draw about our data and the evidence for a generalist-specialist tradeoff?

   Answer:

2. Use Knitr to create a pdf of your completed PhyloTraits_handout.Rmd document, push it to GitHub, and create a pull request. The due date for this assignment is February 25, 2015 at 12:00 PM (noon).