Phylogenetic Diversity - Communities

Student Name; Z620: Quantitative Biodiversity, Indiana University

27 February, 2015

OVERVIEW

Taxonomic measures of α- and β-diversity do not integrate evolutionary information. However, evolutionary information is important for a broad range of biodiversity issues including conservation, biogeography, and community assembly. In this handout, we introduce some commonly used methods in phylogenetic community ecology. These methods will allow us to account for phylogenetic diversity and provide insight into the mechansisms that give rise to the over- and under-dispersion (i.e., clustering) of biological communities.

After completing this exercise you will know how to:

1. measure phylogenetic α diversity
2. measure phylogenetic β diversity
3. evaluate the contribution of phylogeny to geographical patterns of biodiversity

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 *PhyloCom_exercise.Rmd* and the PDF output of `Knitr` (*PhyloCom_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 "*/PhyloCom*" folder,
4. load all of the required R packages (be sure to install if needed), and
5. load the required R source file.

2) DESCRIPTION OF DATA

We will revisit the data that was used in the Geographical Ecology module. As a reminder, in 2013 we sampled ~ 50 forested ponds located in Brown County State Park, Yellowood State Park, and Hoosier National Forest in southern Indiana. In addition to measuring a suite of geographic and environmental variables, we characterized the diversity of bacteria in the ponds using molecular-based approaches. Specifically, we amplified the 16S rRNA gene (i.e., "DNA") and 16S rRNA transcripts (i.e., "RNA") of bacteria using barcoded primers on the Illumina MiSeq platform. We then used a `mothur` pipeline to quality-trim our data set and assign sequences to operational taxonomic units (OTU). In this exercise, we will use the DNA sequence data for making community phylogenetic inference.

3) LOADING OF DATA

In the R code chunk below, do the following:
1. load the environmental data for the Brown County ponds (*20130801_PondDataMod.csv*),
2. load the site-by-species matrix using the `read.otu()` function,
3. subset the data to include only DNA-based identifications of bacteria,
4. rename the sites by removing extra characters,
5. remove unnecessary OTUs in the site-by-species, and
6. load the taxonomic data using the `read.tax()` function.

Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the OTUs,
2. rename the OTUs be removing everything before the tab (\t) and after the bar (|),
3. import the *Methanosarcina* outgroup FASTA file,
4. convert both FASTA files into the DNAbin format and combine using `rbind()`,
5. visualize the sequence alignment,
6. using the alignment (with outgroup), pick a DNA substitution model and create a distance matrix,
7. using the distance matrix above, make a neighbor joining tree,
8. remove any tips (OTUs) that are not in the community data set,
9. plot the rooted tree.

4) PHYLOGENETIC ALPHA DIVERSITY

A. Faith's Phylogenetic Diversity (PD)

In the R code chunk below, do the following:
1. calculate Faith's D using the `pd()` function.

In the R code chunk below, do the following:
1. make a biplot for species richness (SR) and Faith's D (PD),
2. add the trend line, and
3. calculate the scaling exponent.

Question 1: Answer the following questions about the PD-SR pattern.
a. Describe the relationship between taxonomic richness and phylodiversity?
b. Mathematically, why should they be correlated?
c. Under what conditions would you expect these two estimates of diversity to deviate from one another?

Answer 1a:
Answer 1b:
Answer 1c:

i. Randomizations and Null Models

In the R code chunk below, do the following:
1. estimate the standardized effect size of PD using the `richness` randomization method.

Question 2: Using `help()` and the table above, identify two null models that can be used with the `ses.pd()` function. Run `ses.pd()` with these null models and answer the following questions:

a. What hypotheses are being tested with the p-values associated with `ses.pd`?
b. What features might affect the intepretation of the `ses.pd` output?

Answer 2a:
Answer 2b:

B. Phylogenetic Dispersion Within a Sample

Another way to assess phylogenetic α-diversity is to look at dispersion within a sample.

i. Phylogenetic Resemblance Matrix

In the R code chunk below, do the following:
1. calculate the phylogenetic resemblance matrix for taxa in the Indiana ponds data set.

ii. Net Relatedness Index (NRI)

In the R code chunk below, do the following:
1. Calculate the NRI for each site in the Indiana ponds data set.

iii. Nearest Taxon Index (NTI)

In the R code chunk below, do the following: 1. Calculate the NTI for each site in the Indiana ponds data set.

Question 3: In the NRI and NTI examples above, the arguments "abundance.weighted = FALSE" means that the indices were calculated using presence-absence data. Modify and rerun the code so that NRI and NTI are caculated using abundance data. How does this affect the interpretation of NRI and NTI?

Answer 3:

C. Mini-Exercise On Phylogenetic Alpha-Diversity

Conduct an exploratory analysis of the phylogenetic α-diversity for the Indiana ponds. Using techniques from the *Week1_Handout.Rmd* and any other statistical methods at your disposal, identify enviornmental variables that may influence the taxonomic richness (SR), phylodiveristy (PD), and dispersion (NRI and NTI) of the bacterial communities in the sampled ponds. Generate some hypotheses as to what ecological and/or evolutionary processes are giving rise to these patterns of phylogenetic α-diversity.

Mini-Exercise Discussion:

5) PHYLOGENETIC BETA DIVERSITY

A. Phylogenetically Based Community Resemblance Matrix

In the R code chunk below, do the following:
1. calculate the phylogenetically based community resemblance matrix using Mean Pair Distance, and
2. calculate the phylogenetically based community resemblance matrix using UniFrac distance.

In the R code chunk below, do the following:
1. plot mean pair distance versus UniFrac distance and compare.

Question 4: Using the plot above, describe the relationship between Mean Pair Distance and UniFrac distance. Note: we are calcuating unweighted phylogenetic distances (similar to incidence based measures). That means that we are not taking into account the abundance of each taxa in each site.

Answer 4:

B. Visualizing Phylogenetic Beta-Diversity

Now that we have our phylogenetically based community resemblance matrix, we can visualize phyogenetic diversity among samples using the same techniques that we used in the β-diversity module.

In the R code chunk below, do the following:
1. perform a PCoA based on the UniFrac distances, and
2. calculate the explained variation for the first three PCoA axes.

Now that we have calculated our PCoA, we can plot the results.

In the R code chunk below, do the following:
1. plot the PCoA results,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

Question 5: How much variation is explained in the first three PCoA axes if Mean Pair Distance is used as the resemblance matrix instead of UniFrac?

Answer 5:

C. Hypothesis Testing

i. Categorical Approach: Watershed Effect

In the R code chunk below, do the following:
1. test the hypothesis that watershed has an effect on the phylogenetic diversity of bacterial communities.

ii. Continuous Approach: Environmental Gradients

In the R code chunk below, do the following: 1. from the environmental data matrix, subset the variables related to physical and chemical properties of the ponds, and
2. calcuate the euclidean distance between ponds based on these variables (transform and center), this is your environmental gradient.

In the R code chunk below, do the following:
1. conduct a Mantel's test to evaluate whether or not UniFrac distance is correlated with environmental varaition.

Last, let's conduct a Canonical Correspondence Analysis (CCA). You will recall that this constrained ordination technique allows one to test for the effects of an explanatory matrix (e.g., environmental data) on a response matrix (e.g., phylogenetic distance matrix).

In the R code chunk below, do the following:
1. conduct a CCA to test the hypothesis that environmental variation effects the phylogenetic diversity of bacteiral communities,
2. use a permutation test to determine significance,

3. plot the CCA results, and
4. customize the CCA plot.

Question 6: Based on the multivariate procedures conducted above, describe some of the phylogenetic patterns of β-diversity for bacterial communities in the Indiana ponds.

Answer 6:

D. Mini-Exercise On Phylogenetic Beta-Diversity

Generate and test hypotheses about the phylogenetic β-diversity of the Indiana ponds. Using the environmental variables identified in the α-diversity mini-exercise, redo either the PERMANOVA or CCA above using just these variables. Feel free to add any additional approaches we used in *Beta_Handout.Rmd*. Compare your results to the findings from above and discuss any differences.

Mini-Exercise Discussion:

6) GEOGRAPHICAL PHYLOGENETIC COMMUNITY ECOLOGY

A. Phylogenetic Distance-Decay (PDD)

You will recall from the Geographical Ecology module, that the distance decay (DD) relationship reflects the spatial autocorrelation of community similarity. That is, geographically near communities should be more similar than geographically distant communities. Here, we will test to what degree spatial autocorrelation can also affect phylogenetic DD.

First, we will calcualte distances for geographic data, taxonomic data, and phylogenetic data among all unique pair-wise combinations of ponds.

In the R code chunk below, do the following:
1. calculate the geographic distances among ponds,
2. calculate the taxanomic similarity amoung ponds,
3. calculate the phylogenetic similarity among ponds, and
4. create a dataframe that includes all of the above information.

Now, let's plot the DD relationships:
In the R code chunk below, do the following:
1. plot the taxonomic distance decay relationship,
2. plot the phylogenetic distance decay relationship, and
3. add trend lines to each.

In the R code chunk below, do the following:
1. test if the trend lines in the above distance decay relationships are different from one another.

Question 7: Interpret the slopes from the taxonomic and phylogenetic DD relationships. If there are differences, hypothesize why this might be.

Answer 7:

B. Phlogenetic diversity-area relationship (PDAR)

i. Constructing the PDAR

In the R code chunk below, do the following:
1. write a function to generate the PDAR.

ii. Evaluating the PDAR

We will examine the relationship between phylogenetic diversity and area using both Spearman's correlation coefficient (S) and Pearson's correlation coefficient (P). It is informative to use both because while S is computed on ranks and depicts monotonic relationships (the degree to which the relationship is continually increasing or decreasing), P is computed on the observed values and therefore depicts linear relationships.

In the R code chunk below, do the following:
1. calculate the area for each pond,
2. use the `PDAR()` function you just created to calculate the PDAR for each pond,
3. calculate the Pearson's and Spearman's correlation coefficients,
4. plot the PDAR and include the correlation coefficients in the legend, and
5. customize the PDAR plot.

Question 8: For the bacteria in the Indiana ponds, the slope (z) of the SAR was 0.14. This is slightly lower than z-values observed for many macroscopic organisms (e.g., fish, birds, plants), but is higher than what has been reported for other microbial systems. However, what did we observe for the microbial PDAR in the Indiana ponds? How might we explain the differences between the taxonomic (SAR) and phylogenetic (PDAR)?

Answer 8:

7) HOMEWORK

1. In their study of phylogenetic diversity-area relationships (PDARs), Helmus and Ives (2012), increased area by aggregating plots in two ways. First, they aggregated plots with respect to whether the plots were adjacent; what they called 'spatial'. Second, they aggregated plots at random ('non-spatial'). While both spatial and non-spatial sampling methods capture the effect of increasing area, sampling with respect to location (i.e., spatial) also captures the effect of increasing distance or spatial autocorrelation. Explain the general difference that Helmus and Ives (2012) observed between spatial and non-spatial PDARs, and why this difference should be expected in both the species-area relationship (SAR) and the phylogenetic diversity-area relationship.

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 March 4, 2015 at 12:00 PM (noon).