9.Phylogenetic Diversity - Communities

Anna Werkowski; Z620: Quantitative Biodiversity, Indiana University

01 March, 2023

OVERVIEW

Complementing taxonomic measures of α- and β-diversity with evolutionary information yields insight into a broad range of biodiversity issues including conservation, biogeography, and community assembly. In this worksheet, you will be introduced to some commonly used methods in phylogenetic community ecology.

After completing this assignment you will know how to:

1. incorporate an evolutionary perspective into your understanding of community ecology
2. quantify and interpret phylogenetic α- and β-diversity
3. evaluate the contribution of phylogeny to spatial patterns of biodiversity

Directions:

1. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) with your name.
2. Complete as much of the worksheet as possible during class.
3. Use the handout as a guide; it contains a more complete description of data sets along with examples of proper scripting needed to carry out the exercises.
4. Answer questions in the worksheet. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom today, it is *imperative* that you **push** this file to your GitHub repo, at whatever stage you are. This will enable you to pull your work onto your own computer.
6. When you have completed the worksheet, **Knit** the text and code into a single PDF file by pressing the `Knit` button in the RStudio scripting panel. This will save the PDF output in your '9.PhyloCom' folder.
7. After Knitting, please submit the worksheet by making a **push** to your GitHub repo and then create a **pull request** via GitHub. Your pull request should include this file *9.PhyloCom_Worksheet.Rmd* and the PDF output of `Knitr` (*9.PhyloCom_Worksheet.pdf*).

The completed exercise is due on **Wednesday, March 1st, 2023 before 12:00 PM (noon)**.

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 `/9.PhyloCom` folder,
4. load all of the required R packages (be sure to install if needed), and
5. load the required R source file.

```
#rm(list = ls())
#getwd()
#setwd("/Users/annawerkowski/Documents/Github/QB2023_Werkowski/2.Worksheets/9.PhyloCom")

#load packages
#package.list <- c('picante', 'ape', 'seqinr', 'vegan', 'fossil',
                  #'reshape', 'devtools', 'BiocManager', 'ineq',
                  #'labdsv', 'matrixStats', 'pROC')
#for (package in package.list){
  #if (!require(package, character.only = TRUE, quietly = TRUE)){
    #install.packages(package, repos = 'https://cran.us.r-project.org')
    #library(package, character.only = TRUE)
  #}
#}

#load source code
#source("./bin/MothurTools.R")
#env <- read.table("data/20130801_PondDataMod.csv", sep = ",", header = TRUE)
#env <- na.omit(env)
```

2) DESCRIPTION OF DATA

need to discuss data set from spatial ecology!

We sampled >50 forested ponds 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., the DNA sequence) and 16S rRNA transcripts (i.e., the RNA transcript of the gene) of bacteria. We used a program called `mothur` to quality-trim our data set and assign sequences to operational taxonomic units (OTUs), which resulted in a site-by-OTU matrix.
In this module we will focus on taxa that were present (i.e., DNA), but there will be a few steps where we need to parse out the transcript (i.e., RNA) samples. See the handout for a further description of this week's dataset.

3) LOAD THE 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 from the source-code file.

```
#load site by species matrix
#comm <- read.otu(shared ="/Users/annawerkowski/Documents/Github/QB2023_Werkowski/2.Worksheets/9.PhyloC
#select DNA data using 'grep()'
```

```
#comm <- comm[grep("*-DNA", rownames(comm)), ]
#perform replacement of all matches with 'gsub()'
#rownames(comm) <- gsub("\\-DNA", "", rownames(comm))
#rownames(comm) <- gsub("\\_", "", rownames(comm))
#remove sites not in the env data set
#comm <- comm[rownames(comm) %in% env$Sample_ID, ]
#remove zero-abundance OTUs from data set
#comm <- comm[ , colSums(comm) > 0]


#import taxonomic information
#tax <- read.tax(taxonomy = "/Users/annawerkowski/Documents/Github/QB2023_Werkowski/2.Worksheets/9.Phyl
```

Next, in the R code chunk below, do the following:
1. load the FASTA alignment for the bacterial operational taxonomic units (OTUs),
2. rename the OTUs by 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 phylogenetic 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.

```
#import the alignment file
#ponds.cons <- read.alignment(file = "./data/INPonds.final.rdp.1.rep.fasta", format = "fasta")

#clean up the data
#ponds.cons$nam <- gsub("\\|.*$", "", gsub("^.*?\t", "", ponds.cons$nam))
#import outgroup sequence
#outgroup <- read.alignment(file = "./data/methanosarcina.fasta", format = "fasta")
#convert alignment file to DNAbin
#DNAbin <- rbind(as.DNAbin(outgroup), as.DNAbin(ponds.cons))
#visualize alignment
#image.DNAbin(DNAbin, show.labels = T, cex.lab = 0.05, las = 1)
```

```
#make a distance matrix
#seq.dist.jc <- dist.dna(DNAbin, model = "JC", pairwise.deletion = FALSE)
#make a neighbor-joining tree file
#phy.all <- bionj(seq.dist.jc)
#drop tips of zero-occurrence OTUs
#phy <- drop.tip(phy.all, phy.all$tip.label[!phy.all$tip.label %in% c(colnames(comm), "Methanosarcina")
#identify outgroup sequence
#outgroup <- match("Methanosarcina", phy$tip.label)
#root the tree
#phy <- root(phy, outgroup, resolve.root = TRUE)
#plot the rooted tree
#par(mar = c(1,1,2,1) + 0.1)
#plot.phylo(phy, main = "Neighbor Joining Tree", "phylogram",
           #show.tip.label = FALSE, use.edge.length = FALSE,
           #direction = "right", cex = 0.6, label.offset = 1)
```

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.

```
#calculate PD and S
#pd <- pd(comm, phy, include.root = FALSE)
```

In the R code chunk below, do the following:
1. plot species richness (S) versus phylogenetic diversity (PD),
2. add the trend line, and
3. calculate the scaling exponent.

```
#biplot of S and PD
#par(mar = c(5,5,4,1) + 0.1)

#plot(log(pd$SR), log(pd$PD),
         #pch = 20, col = "red", las = 1,
         #xlab = "ln(S)", ylab = "ln(PD)", cex.main = 1,
         #main = "Phylodiversity (PD) vs. Taxonomic Richness (SR)")

#test of power-law relationship
#fit <- lm('log(pd$PD) ~ log(pd$SR)')
#abline(fit, col = "red", lw = 2)
#exponent <- round(coefficients(fit)[2], 2)
#legend("topleft", legend = paste("Scaling exponent =", exponent, sep = ""),
       #bty = "n", lw = 2, col = "red")
```

Question 1: Answer the following questions about the PD-S pattern.
a. Based on how PD is calculated, why should this metric be related to taxonomic richness? b. Describe the relationship between taxonomic richness and phylodiversity. c. When would you expect these two estimates of diversity to deviate from one another? d. Interpret the significance of the scaling PD-S scaling exponent.

> ***Answer 1a***: The PD is calculated by analyzing the presence-absence of OTUs and does not weigh common or rare species more than one another. Higher PD values indicate a branch with evolutionary divergent taxa. Lower PD values indicate taxa with a restricted evolutionary history. By using this metrix and finding high PD values, we can determine the richness and subsequent diversity of an assemblage. ***Answer 1b***: The more taxonomically rich an assemblage is, the more phylogentically diverse it becomes. ***Answer 1c***: I would expect these estimates to deviate if there were an extreme speciation event that caused major changes within an assemblage. ***Answer 1d***: The scaling exponent is important because it acts as a line of best fit for the PD-S pattern.

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.

```
#estimate standardized effect with picante
#ses.pd <- ses.pd(comm[1:2,], phy, null.model = "richness", runs = 25, include.root = FALSE)
```

Question 2: Using `help()` and the table above, run the `ses.pd()` function using two other null models and answer the following questions:

a. What are the null and alternative hypotheses you are testing via randomization when calculating `ses.pd`?
b. How did your choice of null model influence your observed ses.pd values? Explain why this choice affected or did not affect the output.

Answer 2a: I was unable to do this portion because of invalid code. ***Answer 2b***: I was unable to do this portion because of invalid code.

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.

```
#create a phylogenetic distance matrix
#phydist <- cophenetic.phylo(phy)
```

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.

```
#estimate standardized effect size of NRI via randomization
#ses.mpd2 <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                    #abundance.weighted = FALSE, runs = 25)

#calculate NRI
#NRI <- as.matrix(-1 * ((ses.mpd2[,2] - ses.mpd2[,3]) / ses.mpd2[,4]))
#rownames(NRI) <- row.names(ses.mpd2)
#colnames(NRI) <- "NRI"
```

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.

```
#estimate standardized effect size of NTI via randomization
#ses.mntd2 <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                      #abundance.weighted = FALSE, runs = 25)

#calculate NTI
#NTI <- as.matrix(-1 * ((ses.mntd2[,2] - ses.mntd2[,3]) / ses.mntd2[,4]))
#rownames(NTI) <- row.names(ses.mntd2)
#colnames(NTI) <- "NTI"
```

Question 3:

a. In your own words describe what you are doing when you calculate the NRI.
b. In your own words describe what you are doing when you calculate the NTI.
c. Interpret the NRI and NTI values you observed for this dataset.

d. 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 calculated using abundance data. How does this affect the interpretation of NRI and NTI?

```
#rerun code for NRI and NTI with abundance data
#estimate standardized effect size of NRI via randomization
#ses.mpd3 <- ses.mpd(comm, phydist, null.model = "taxa.labels",
                     #abundance.weighted = TRUE, runs = 25)

#calculate NRI
#NRI2 <- as.matrix(-1 * ((ses.mpd3[,2] - ses.mpd3[,3]) / ses.mpd3[,4]))
#rownames(NRI2) <- row.names(ses.mpd3)
#colnames(NRI2) <- "NRI"

#estimate standardized effect size of NTI via randomization
#ses.mntd3 <- ses.mntd(comm, phydist, null.model = "taxa.labels",
                       #abundance.weighted = TRUE, runs = 25)

#calculate NTI
#NTI2 <- as.matrix(-1 * ((ses.mntd3[,2] - ses.mntd3[,3]) / ses.mntd3[,4]))
#rownames(NTI2) <- row.names(ses.mntd3)
#colnames(NTI2) <- "NTI"
```

Answer 3a: When we calculate the NRI, we are looking at the phylogenetic clustering and overdispersion that is occuring in an assemblage. ***Answer 3b***: When we calculate the NTI, we are doing a similar test to NRI, but we are looking at the mean nearest phylogenetic neighbor distance which in turn compares the taxons. ***Answer 3c***:
Answer 3d: When you change the code to evaluate abundance, you are having the code look at a larger set of data versus the incidence. The abundance looks at how much, versus the indicidence which looks at how many times something appears.

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.

```
#mean pairwise distance
#distance.mp <- comdist(comm, phydist)

#UniFrac Distance
#dist.uf <- unifrac(comm, phy)
```

In the R code chunk below, do the following:
1. plot Mean Pair Distance versus UniFrac distance and compare.

```
#par(mar = c(5,5,2,1) + 0.1)
#plot(distance.mp, dist.uf, pch = 20, col = "red", las = 1,
     #asp = 1, xlim = c(0.15, 0.5), ylim = c(0.15, 0.5),
     #xlab = "Mean Pair Distance", ylab = "UniFrac Distance")
```

```
#abline(b = 1, a = 0, lty = 2)
#text(0.5, 0.47, "1:1")
```

Question 4:

a. In your own words describe Mean Pair Distance, UniFrac distance, and the difference between them.
b. Using the plot above, describe the relationship between Mean Pair Distance and UniFrac distance. Note: we are calculating unweighted phylogenetic distances (similar to incidence based measures). That means that we are not taking into account the abundance of each taxon in each site.
c. Why might MPD show less variation than UniFrac?

> ***Answer 4a***: Mean pair distance is the average distance between phylogenies of two chosen taxa. Unifrac distance is the total number of unshared branches of chosen samples divided by the total amount of branches (both shared and unshared) in the phylogenetic tree.The mean pair distance tells us basic distance between the taxa but the Unifrac distance tells us how connected those taxa are based on their shared and unshared branches. ***Answer 4b***: The Unifrac distance changes but the mean pair distance stays the same. ***Answer 4c***: The mean pairwise distance is analyzing less things - the Unifrac is looking at the shared and unshared branches, which adds way more complexity to the evaluation. It doesn't surprise me that the Unifrac shows more variation.

B. Visualizing Phylogenetic Beta-Diversity

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

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.

```
#pond.pcoa <- cmdscale(dist.uf, eig = T, k = 3)
#explainvar1 <- round(pond.pcoa$eig[1] / sum(pond.pcoa$eig), 3) * 100
#explainvar2 <- round(pond.pcoa$eig[2] / sum(pond.pcoa$eig), 3) * 100
#explainvar3 <- round(pond.pcoa$eig[3] / sum(pond.pcoa$eig), 3) * 100
#sum.eig <- sum(explainvar1, explainvar2, explainvar3)
```

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 using either the R base package or the `ggplot` package,
2. include the appropriate axes,
3. add and label the points, and
4. customize the plot.

```
#define plot parameters
#par(mar = c(5,5,1,2) + 0.1)
#initiate plot
#plot(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     #xlim = c(-0.2, 0.2), ylim = c(-.16, 0.16),
     #xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
     #ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
     #pch = 16, cex = 2.0, type = "n", cex.lab = 1.5,
```

```
     #cex.axis = 1.2, axes = FALSE)
#add axes
#axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
#axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
#abline(h = 0, v = 0, lty = 3)
#box(lwd = 2)

#add points and labels
#points(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
       #pch = 19, cex = 3, bg = "gray", col = "gray")
#text(pond.pcoa$points[ ,1], pond.pcoa$points[ ,2],
     #labels = row.names(pond.pcoa$points))
```

In the following R code chunk: 1. perform another PCoA on taxonomic data using an appropriate measure of dissimilarity, and 2. calculate the explained variation on the first three PCoA axes.

Question 5: Using a combination of visualization tools and percent variation explained, how does the phylogenetically based ordination compare or contrast with the taxonomic ordination? What does this tell you about the importance of phylogenetic information in this system?

> ***Answer 5***: The phylogenetically based ordination showed more variation percentage than the taxonomic ordination. This tells us that phylogenetic information is important in this system because it explains more of the variation that is occuring.

C. Hypothesis Testing

i. Categorical Approach

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.

```
#define environmental category
#watershed <- env$Location

#run PERMANOVA with adonis
#phylo.adonis <- adonis2(dist.uf ~ watershed, permutations = 999)

#compare to PERMANOVA results based on taxonomy
#tax.adonis <- adonis2(vegdist(decostand(comm, method = "log"), method = "bray") ~ watershed, permutati
```

ii. Continuous Approach

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. calculate environmental distance between ponds based on the Euclidean distance between sites in the environmental data matrix (after transforming and centering using `scale()`).

```
#define environmental variables
#envs <- env[, 5:19]
#remove redundant variables
#envs <- envs[, -which(names(envs) %in% c("TDS", "Salinity", "Cal_Volume"))]
#create distance matrix for environmental variables
#env.dist <- vegdist(scale(envs), method = "euclid")
```

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

```
#conduct Mantel test
#mantel(dist.uf, env.dist)
```

Last, conduct a distance-based Redundancy Analysis (dbRDA).

In the R code chunk below, do the following:
1. conduct a dbRDA to test the hypothesis that environmental variation effects the phylogenetic diversity of bacterial communities,
2. use a permutation test to determine significance, and 3. plot the dbRDA results

```
#conduct dbrda
#ponds.dbrda <- vegan::dbrda(dist.uf ~ ., data = as.data.frame(scale(envs)))
#permutation tests: axes and environmental variables
#anova(ponds.dbrda, by = "axis")
#ponds.fit <- envfit(ponds.dbrda, envs, perm = 999)
#ponds.fit

#calculate explained variation
#dbrda.explainvar1 <- round(ponds.dbrda$CCA$eig[1] /
                      #sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100
#dbrda.explainvar2 <- round(ponds.dbrda$CCA$eig[2] /
                      #sum(c(ponds.dbrda$CCA$eig, ponds.dbrda$CA$eig)), 3) * 100
```

##MAKE A DBRDA PLOT##

```
#define plot parameters
#par(mar = c(5,5,4,4) + 0.1)
#initiate plot
#plot(scores(ponds.dbrda, display = "species"), xlim = c(-2,2), ylim = c(-2,2),
     #xlab = paste("dbRDA 1 (", dbrda.explainvar1, "%)", sep = ""),
     #ylab = paste("dbRDA 2 (", dbrda.explainvar2, "%)", sep = ""),
     #pch = 16, cex = 2.0, type = "n", cex.lab = 1.5, cex.axis = 1.2, axes = FALSE)


#add axes
#axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
#axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
#abline(h = 0, v = 0, lty = 3)
#box(lwd = 2)

#add points and labels
#points(scores(ponds.dbrda, display = "wa"),
       #pch = 19, cex = 3, bg = "gray", col = "gray")
#text(scores(ponds.dbrda, display = "wa"),
     #labels = row.names(scores(ponds.dbrda, display = "wa")), cex = 0.5)

#add environmental vectors
#vectors <- scores(ponds.dbrda, display = "bp")

#arrows(0,0, vectors[,1] * 2, vectors[,2] * 2,
```

```
       #lwd = 2, lty = 1, length = 0.2, col = "red")
#text(vectors[,1] * 2, vectors[,2] * 2, pos = 3,
     #labels = row.names(vectors))
#axis(side = 3, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = "red", lwd = 2.2,
    #at = pretty(range(vectors[,1])) * 2, labels = pretty(range(vectors[,1])))
#axis(side = 4, lwd.ticks = 2, cex.axis = 1.2, las = 1, col = "red", lwd = 2.2,
    #at = pretty(range(vectors[,2])) * 2, labels = pretty(range(vectors[,2])))
```

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

> ***Answer 6***: Variation explained by both of the axes is relatively similar. More of the vector arrows are skewing towards the left-hand side of the plot. Environmental vectors like Temp and pH seemed to have the largest interaction with the data points. There appears to be lots of clustering occuring in this dataset as well.

6) SPATIAL PHYLOGENETIC COMMUNITY ECOLOGY

A. Phylogenetic Distance-Decay (PDD)

A distance decay (DD) relationship reflects the spatial autocorrelation of community similarity. That is, communities located near one another should be more similar to one another in taxonomic composition than distant communities. (This is analagous to the isolation by distance (IBD) pattern that is commonly found when examining genetic similarity of a populations as a function of space.) Historically, the two most common explanations for the taxonomic DD are that it reflects spatially autocorrelated environmental variables and the influence of dispersal limitation. However, if phylogenetic diversity is also spatially autocorrelated, then evolutionary history may also explain some of the taxonomic DD pattern. Here, we will construct the phylogenetic distance-decay (PDD) relationship

First, calculate 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 taxonomic similarity among ponds,
3. calculate the phylogenetic similarity among ponds, and
4. create a dataframe that includes all of the above information.

```
#geographic distances among ponds
#long.lat <- as.matrix(cbind(env$long, env$lat))
#coord.dist <- earth.dist(long.lat, dist = TRUE)

#taxonomic similarity among ponds (Bray-Curtis distance)
#bray.curtis.dist <- 1 - vegdist(comm)

#phylogenetic similarity among ponds
#unifrac.dist <- 1 - dist.uf

#transform all distances into pairwise long format with the melt function
#unifrac.dist.mlt <- melt(as.matrix(unifrac.dist))[melt(upper.tri(as.matrix(unifrac.dist)))$value,]

#bray.curtis.dist.mlt <- melt(as.matrix(bray.curtis.dist))[melt(upper.tri(as.matrix(bray.curtis.dist)))
```

```
#coord.dist.mlt <- melt(as.matrix(coord.dist))[melt(upper.tri(as.matrix(coord.dist)))$value,]

#env.dist.mlt <- melt(as.matrix(env.dist))[melt(upper.tri(as.matrix(env.dist)))$value,]

#create a data frame from the lists of distances
#df.dist <- data.frame(coord.dist.mlt, bray.curtis.dist.mlt[,3], unifrac.dist.mlt[,3],env.dist.mlt[,3])
#names(df.dist)[3:6] <- c("geo.dist", "bray.curtis", "unifrac", "env.dist")
```

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.

```
#set initial plot parameters
#par(mfrow = c(2,1), mar = c(1,5,2,1) + 0.1, oma = c(2,0,0,0))
#make plot for taxonomic DD
#plot(df.dist$geo.dist, df.dist$bray.curtis, xlab = "", xaxt = "n", las = 1, ylim = c(0.1, 0.9), ylab =
#regression for taxonomic DD
#DD.reg.bc <- lm(df.dist$bray.curtis ~ df.dist$geo.dist)
#summary(DD.reg.bc)
#abline(DD.reg.bc, col = "red4", lwd = 2)
#new plot paramters
#par(mar = c(2,5,1,1) + 0.1)
#make plot for phylogenetic DD
#plot(df.dist$geo.dist, df.dist$unifrac, xlab = "", las = 1, ylim = c(0.1, 0.9),
     #ylab = "Unifrac Similarity", col = "darkorchid4")
#regression for phylogenetic DD
#DD.reg.uni <- lm(df.dist$unifrac ~ df.dist$geo.dist)
#summary(DD.reg.uni)
#abline(DD.reg.uni, col = "red4", lwd = 2)
#add x-axis label to plot
#mtext("Geographic Distance (km)", side = 1, adj = 0.55, line = 0.5, outer = TRUE)
```

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

```
#source("/Users/annawerkowski/Documents/Github/QB2023_Werkowski/2.Worksheets/9.PhyloCom/bin/diffslope.R
#diffslope(df.dist$geo.dist, df.dist$unifrac, df.dist$geo.dist, df.dist$bray.curtis)
```

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

> ***Answer 7***: The distance decay plots showed some extreme differences between them. I do not know if those are slope-related as I could not get the code to run, but from the images in the handout, I would like to conclude the following. Both the Bray-Curtis and Unifrac similarities had relatively similar geographic distances. However, when it came to the variation in the similarity values on the Y-axis, they were much more different. The Bray-Curtis similarity had a full spread of data points between the 0.1 - 0.9 range while the Unifrac similarity had a tight clustering between 0.6 - 0.8 range. I believe this to mean that the Unifrac similarity showed more clusters of similar data points while the Bray-Curtis exposed a large range of similar and dissimilar values.

SYNTHESIS

Ignoring technical or methodological constraints, discuss how phylogenetic information could be useful in your own research. Specifically, what kinds of phylogenetic data would you need? How could you use it to answer important questions in your field? In your response, feel free to consider not only phylogenetic approaches related to phylogenetic community ecology, but also those we discussed last week in the PhyloTraits module, or any other concepts that we have not covered in this course.

> As I have been doing more thinking about my research, I have been wanting to evaluate the phylogenetic similarity between native and invasive species occupying the same niches. I wonder if their relatedness has any impact on conspecific or heterospecific growth. To do this, I would need the genetic information of the species I am studying, which would most likely be found in rbcl data. I am sure if I thought about the usage of phylogeny on my research I could come up with better ways to evaluate those items.