Measuring phylogenetic diversity within communities

Jesús N. Pinto-Ledezma and Jeannine Cavender-Bares

The main goal of this practice is to present basic understanding about measuring phylogenetic diversity within communities or best known as the analysis of community phylogenetics. The community phylogenetics integrates ecological and evolutionary concepts and explores the mechanisms (e.g., biotic interactions or environmental filters) governing the assembly of ecological communities.

There are different sources of information and web pages with a lot of information about this field. The most common and useful are the web pages of the books: Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology and Phylogenies in Ecology. Among the most influential papers in this field are Phylogenies and Community Ecology and The merging of community ecology and phylogenetic biology.

Install and load packages

Check if you are in the correct working directory.

```
getwd()
```

Now install the load the necessary packages.

```
packages <- c("picante", "dplyr", "tidyr", "picante", "lubridate",
              "Taxonstand", "ape", "neonUtilities", "phytools", "vegan", "car")

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

if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages], dependencies = TRUE)
}

if ( ! ("V.PhyloMaker" %in% installed.packages())) {remotes::install_github("jinyizju/V.PhyloMaker")}
```

```
lapply(packages, library, character.only = TRUE)
require(V.PhyloMaker)
```

For today's practice we will use data from the **National Ecological Observatory Network**-NEON. We strongly recommend to look at the NEON website to get a deeper understanding of the work at NEON. Also you can read these papers: A continental strategy for the National Ecological Observatory Network and The plant diversity sampling design for The National Ecological Observatory Network.

Prepare data

Download community data from NEON

We will download data of plant communities directly from NEON and to do that we will use the **R** package {**neonUtilities**} and the function '*loadByProduct*'. Before proceed let's take a look at the core information

required in this function. To do that you can type **?loadByProduct** in your console and the documentation for this function will appear in the Help window of RStudio.

The core information we need to inform in the '*loadByProduct*' function are:

1. **dpID** = The identifier of the NEON data product to pull, in the form DPL.PRNUM.REV, e.g. DP1.10023.001.
2. **site** = Either the string 'all', meaning all available sites, or a character vector of 4-letter NEON site codes, e.g. c('ONAQ','RMNP'). Defaults to all.
3. **package** = Either 'basic' or 'expanded', indicating which data package to download. Defaults to basic.

As you can see, the **dpID** and **site** correspond to the kind of data we want to download and the site to the location where the data were collected. We can open the NEON website to find the information required for downloading the data for **plant presence and percent cover**. Also, you can look at the map of NEON sites to see the distribution of sites across The United States.

For this practice we will use the next information: **dpID** = **"DP1.10058.001"** and **site** = **c("HARV", "CPER","ABBY")** that correspond to **plant presence and percent cover** and three sites: **HARV**(Harvard Forest & Quabbin Watershed NEON, MA), **CPER**(Central Plains Experimental Range NEON, CO) and **ABBY**(Abby Road NEON, WA). Note that you can download data only for one site but for the sake of getting more practice of data management in R we will download the full data of plant community data for three sites (you can download data for more sites if you prefer).

```
# Set global option to NOT convert all character variables to factors
options(stringsAsFactors = F)

NEON_data <- loadByProduct(dpID = "DP1.10058.001",
                           site = c("HARV", "CPER","ABBY"),
                           package = "expanded", check.size = TRUE)

# type "y" (with no quotes) in your console to start downloading the data from NEON
```

Let's inspect the downloaded data.

```
names(NEON_data)

View(NEON_data$div_10m2Data100m2Data)

View(NEON_data$div_1m2Data)

# save raw data in your hard drive - this is a common practice that allow reproducibility
# and also save you a lot of time.

dir.create("Data") # You can skip this line if you already have this folder
dir.create("Data/NEON")

save(NEON_data, file = "Data/NEON/RawData_NEON_lab5.RData")
```

The data of abundance of plants correspond to the object **"div_1m2Data"**, thus we will isolate that data from the raw data.

```
sel <- NEON_data$div_1m2Data %>%
  select(namedLocation, domainID, siteID, plotType, plotID, subplotID, endDate,
         taxonID, taxonRank, family, scientificName, nativeStatusCode,
         percentCover, heightPlantSpecies)

unique(sel$namedLocation)
```

```
unique(sel$siteID)
unique(sel$endDate)
```

The isolated data is a data.frame of 77692 rows and 14 columns. Some information is not required so let's clean the data a little bit and select information for only one site and a single period of time.

```
sel <- sel %>%
  drop_na(scientificName) %>% # Removing NAs in the column of species
  mutate(Date = endDate) %>%
  separate(endDate, sep = "-", into = c("Year", "Month", "Day"))

unique(sel$Year)
unique(sel$siteID)
```

Select the site **HARV** and the year **2018**.

```
HARV <- sel %>%
  filter(siteID == "HARV" & Year == 2018)

unique(HARV$Year)
unique(HARV$siteID)

head(HARV)
```

The data corresponding to HARV for the 2018 contains 2161 rows and 17 columns. If we look at the data, specifically to the column **scientificName** we can see that the taxonomy used in NEON correspond to the taxonomy used by the USDA, however this taxonomy is not necessarily used by scientist [:(] that rely for example on the taxonomy of the APG Angiosperm Phylogeny Group for Angiosperms and WCSP World Checklist of Conifers for Gymnosperms.

```
View(HARV)
```

In any case, we need to standardize the species names in order to proced with the calculation of metrics.

To do that, we will use the package {**Taxonstand**} that allow matching the taxonomy among different sources using the working list of all known plant species produced by the botanical community or **The Plant List** (TPL).

Note - this process will take some time (~120 seconds) and return several warning messages but please do not pay attention to those messages.

```
spp <- unique(HARV$scientificName) # vector with scientific names

# Perform taxonomic standardization on plant names (TPL table)
spp_check <- TPL(spp, infra = TRUE, corr = TRUE)

head(spp_check)
```

Check the updated taxonomy.

```
View(spp_check)
```

Select the necessary information and combine with the HARV data.

```
taxonomy <- spp_check %>%
  drop_na(New.Genus, New.Species) %>%
  select(Taxon, Family, New.Genus, New.Species, Tax_res)

HARV_data <- full_join(HARV, taxonomy, by = c("scientificName" = "Taxon"))
```

Finally we can select some specific columns that we will use from now on.

```
HARV_data <- HARV_data %>%
  mutate(sciName = paste0(HARV_data$New.Genus, "_", HARV_data$New.Species)) %>%
  select(siteID, plotID, subplotID, Tax_res, Family, sciName, percentCover) %>%
  filter(Tax_res == "Species")

head(HARV_data)
```

We can save the cleaned data and remove the unnecessary information from the environment.

```
save(HARV, HARV_data, taxonomy, file = "Data/NEON/CleanData_NEON_lab5.RData")

rm(HARV, NEON_data, sel, spp_check, taxonomy, installed_packages, packages, spp)
```

As a last step the cleaned data will be transformed from the long format to a wide format.

```
HARV_mat <- sample2matrix(HARV_data[, c(2, 7, 6)])
nrow(HARV_mat)
ncol(HARV_mat)
```

Alright, until here we have downloaded, cleaned and prepared plant community data for the NEON site HARV. The next step is to prepare the phylogeny for those communities or community level phylogeny. To do that we will use the most up to date phylogeny of vascular plants Constructing a broadly inclusive seed plant phylogeny and the R package {**V.PhyloMaker**}.

```
sppPhylo <- HARV_data[, c(5, 6)]

# Prepare the taxonomy data to extract the phylogeny
sppPhylo <- sppPhylo %>%
  mutate(family = Family) %>%
  mutate(species = gsub("_" , " ", sciName)) %>%
  separate(sciName, sep = "_", into = c("genus", "ephitet")) %>%
  select(species, genus, family)

sppPhylo <- unique(sppPhylo[c("species", "genus", "family")])

head(sppPhylo)
```

Prepare the phylogeny and plot it.

```
result <- phylo.maker(sppPhylo, scenarios = "S3")

phylo <- multi2di(result$scenario.3)
is.binary.phylo(phylo)
is.ultrametric(phylo)

plot(phylo, show.tip.label = FALSE)
```

Before continuing with the lab let's check again if our data (phylogeny and community) match each other. To do this we will use the awesome function **match.phylo.com()** from the package {**picante**}.

```
matched <- picante::match.phylo.comm(phy = phylo, comm = HARV_mat)
matched$phy
matched$comm
```

Alright, we have all data necessary for calculating different phylogenetic diversity metrics, Yay!

Save the phylogenetic data and clean the environment.

```
save(phylo, sppPhylo, result, file = "Data/NEON/Phylo_NEON_lab5.RData")

rm(sppPhylo, result)
```

Phylogenetic diversity metrics

Ok, let's inspect the data that were stored in the object matched.

```
matched$comm[1:10, 1:10]
```

```
plot(matched$phy, show.tip.label = FALSE)
```

Explore diversity metrics

Awesome, we are now ready to explore some the of the **jungle** of metrics for the evaluation of phenotypic and phylogenetic structure of communities (Pausas and Verdú 2010).

Phylogenetic diversity

Let's calculate some metrics manually and then using the package {**picante**} we will calculate the same metrics but for all communities at once.

Phylogenetic diversity is just the sum of the total branch lengths in the community. In this case we are calculating PD using all species in the phylogeny, in other words, assuming that a single community contain the same amount of species as the phylogeny.

```
sum(matched$phy$edge.length) #  sum of the total branch lengths in the community
```

Here using the package {**picante**} we can calculate the same metric for each community or plots at HARV.

```
HARV_PD <- pd(matched$comm, matched$phy, include.root = FALSE) # Faith's PD
head(HARV_PD)
```

```
cor.test(HARV_PD$SR, HARV_PD$PD)
```

```
plot(HARV_PD$SR, HARV_PD$PD, xlab = "Species richness",
     ylab = "PD (millions of years)", pch = 16)
```

Mean pairwise distance (MPD) and mean nearest-pairwise distance (MNTD)

Other common metrics are MPD and MNTD. As in PD, let's calculate MPD and MNTD manually.

```
# MPD
dist.trMB <- cophenetic(matched$phy)
dist.trMB <- dist.trMB[lower.tri(dist.trMB, diag = FALSE)]

mean(dist.trMB)
```

```
# MNTD
dist.trMB2 <- cophenetic(matched$phy)
diag(dist.trMB2) <- NA
apply(dist.trMB2, 2, min, na.rm = TRUE)

mean(apply(dist.trMB2, 2, min, na.rm = TRUE))
```

And now using the package **picante**

```
HARV_MPD <- mpd(matched$comm, cophenetic(matched$phy)) # MPD
head(HARV_MPD)
```

```
HARV_MNTD <- mntd(matched$comm, cophenetic(matched$phy)) # MNTD
head(HARV_MNTD)
```

Community diversity metrics

The analyses of community phylogenetic started making inferences about the mechanisms structuring the local communities through the evaluation of phylogenetic arrangements in local communities (see Cavender-Bares et al. 2009 for an initial criticism). However, new methods are now available, such that more complex balance between ecological and historical processes at local and regional scales can be incorporated into the analyses (Pigot and Etienne 2015, Pinto-Ledezma et al. 2019).

Now, let's calculate some of the most common metrics.

PD - phylogenetic diversity is the sum of the total phylogenetic branch length for one or multiple samples.

Note - we will use the object **HARV_CDM** to store all the results.

Phylogenetic diversity in a community - PD

```
# We can also calculate the standardized effect size of PD in each community
HARV_CDM <- ses.pd(matched$comm, matched$phy, runs = 99)
HARV_CDM <- HARV_CDM[, c(1, 2, 6, 7)]

head(HARV_CDM)
```

Phylogenetic Rao's quadratic entropy - RaoD

Rao's quadratic entropy (Rao 1982) is a measure of diversity in ecological communities that can optionally take species differences (e.g. phylogenetic dissimilarity) into account.

```
HARV_CDM$RaoD <- raoD(matched$comm, force.ultrametric(matched$phy))$Dkk
```

Mean pairwise distance separating taxa in a community - MPD

```
# SES-MPD
HARVsesmpd <- ses.mpd(matched$comm, cophenetic(matched$phy),
                      null.model = "taxa.labels", runs = 99)
```

```
HARV_CDM$mpd <- HARVsesmpd[, c(2)]
HARV_CDM$mpd.obs.z <- HARVsesmpd[, c(6)]
HARV_CDM$mpd.obs.p <- HARVsesmpd[, c(7)]
```

Mean nearest taxon distance for taxa in a community - MNTD

```
# SES-MNTD
HARVsesmntd <- ses.mntd(matched$comm, cophenetic(matched$phy),
                        null.model = "taxa.labels", runs = 99)
```

```
HARV_CDM$mntd <- HARVsesmntd[, c(2)]
HARV_CDM$mntd.obs.z <- HARVsesmntd[, c(6)]
HARV_CDM$mntd.obs.p <- HARVsesmntd[, c(7)]
```

Phylogenetic species variability - PSV

Phylogenetic species variability quantifies how phylogenetic relatedness decreases the variance of a hypothetical unselected/neutral trait shared by all species in a community.

```
# PSV or phylogenetic species variability
HARVpsv <- psv(matched$comm, matched$phy, compute.var = TRUE)
```

```
HARV_CDM$PSV <- HARVpsv[, 1]
```

Phylogenetic species richness - PSR

Phylogenetic species richness is the number of species in a sample multiplied by PSV.

```
# PSR or phylogenetic species richness
HARVpsr <- psr(matched$comm, matched$phy, compute.var = TRUE)
```

```
HARV_CDM$PSR <- HARVpsr[, 1]
```

Phylogenetic species evenness - PSE

Phylogenetic species evenness is the metric PSV modified to incorporate relative species abundances.

```
# PSR or phylogenetic species evenness
HARVpse <- pse(matched$comm, matched$phy)
```

```
HARV_CDM$PSE <- HARVpse[, 1]
```

qDp

qD(p) is a metric that measure the variation in species' divergences within communities. This metric is a modification of the Hill index, weighting a species' proportional abundance by its relative share of phylogenetic information.

```
# Scheiner 2012 qD(p)
source("https://raw.githubusercontent.com/jesusNPL/BiodiversityScience/master/Spring2021/R-Functions/qDp

HARVqDp <- qDp(matched$phy, matched$comm, q = 2)
```

```
HARV_CDM$qDP <- HARVqDp
```

```
head(HARV_CDM, 10)
```

We have calculated several metrics that describe the phylogenetic structure of communities at Harvard Forest & Quabbin Watershed NEON (HARV) site.

Now, let's select some columns to explore if some metrics are related to each other.

```
HARV_CDM_sel <- HARV_CDM %>%
  select(ntaxa, pd.obs, RaoD, mpd, mntd, PSV, PSE, PSR, qDP)

head(HARV_CDM_sel)
```

Compare the metrics

```
scatterplotMatrix(HARV_CDM_sel)
```

Explore the correlation among metrics.

```
cor.table(na.omit(HARV_CDM_sel))
```

You can also plot the relationship.

```
plot(HARV_CDM_sel$mpd, HARV_CDM_sel$PSV, xlab = "MPD", ylab = "PSV", pch = 17)
```

```
HARV_mds <- metaMDS(na.omit(HARV_CDM_sel), trace = FALSE)

ordiplot(HARV_mds, type = "t", display = "species")
```

What do you think?

Which metric would you use for your paper?

The challenge

The challenge for this assignment is:

- **Option 1**

Repeat all process but select two NEON sites across the United States, compare the results of both sites and discuss the difference in the results if there any. For example, you can use the other two sites we have downloaded at the beginning of the lab, i.e., "CPER" and "ABBY".

- **Option 2**

Another option is to calculate the phylogenetic diversity metrics for each year on a NEON site. For example, you can use the data from HARV but instead of doing the analysis only for the 2018 (as we did here) you can repeat the process for all years from 2013 to 2020.

References

Barnett, D. T., Adler, P. B., Chemel, B. R., Duffy, P. A., Enquist, B. J., Grace, J. B., ... Vellend, M. (2019). The plant diversity sampling design for The National Ecological Observatory Network. Ecosphere, 10(2), e02603. doi:10.1002/ecs2.2603

Cadotte, M. W. and Davies, T. J. (2016). Phylogenies in Ecology: A Guide to Concepts and Methods. Princeton: Princeton University Press.

Cavender-Bares, J., Kozak, K. H., Fine, P. V. A. and Kembel, S. W. (2009). The merging of community ecology and phylogenetic biology. Ecology letters 12, 693–715.

Garamszegi, L. Z. (2014). Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology. (ed. Garamszegi, L. Z.) Berlin: Springer-Verlag.

Helmus, M. R. 2007. Phylogenetic measures of biodiversity. American Naturalist 169:E68-E83.

Keller, M., Schimel, D. S., Hargrove, W. W., & Hoffman, F. M. (2008). A continental strategy for the National Ecological Observatory Network. Frontiers in Ecology and the Environment, 6(5), 282–284. doi: 10.1890/1540-9295(2008)6%5B282:acsftn%5D2.0.co;2

Pausas, J. G. and Verdú, M. (2010). The Jungle of Methods for Evaluating Phenotypic and Phylogenetic Structure of Communities. BioScience 60, 614–625.

Pinto-Ledezma, J. N., Jahn, A. E., Cueto, V. R., Diniz-Filho, J. A. F., & Villalobos, F. (2019). Drivers of Phylogenetic Assemblage Structure of the Furnariides, a Widespread Clade of Lowland Neotropical Birds. The American Naturalist, E000–E000.

Rao, C. R. 1982. Diversity and dissimilarity coefficients: a unified approach. Theoretical Population Biology 21:2443.

Smith, S. A., & Brown, J. W. (2018). Constructing a broadly inclusive seed plant phylogeny. American Journal of Botany, 105(3), 302–314.

Scheiner, S. M. (2012). A metric of biodiversity that integrates abundance, phylogeny, and function. Oikos, 121(8), 1191–1202.