Lab 10: Biodiversity Databases and Community Phylogenetics

Carrie Tribble, updated by Ixchel González

04/01/2020

Integrative Biology 200

Principles of Phylogenetics

University of California, Berkeley

1 Introduction

This week we'll briefly check out several different online databases that contain phylogenetic, specimen, and taxonomic information. We'll also look at a few R packages that you can use to automate querying these resources. After, we will explore some methods for community phylogenetic analyses.

2 Phylogenetic Online Databases and Tools

2.1 TreeBASE

TreeBASE https://treebase.org is a repository of phylogenetic information, specifically user-submitted phylogenetic trees and the data used to generate them. Many studies upload their phylogenies and sequence matrices here, so they can be used or reanalyzed in future studies. Many journals require trees to be deposited in TreeBASE before publication.

Try searching TreeBASE for a phylogeny of a taxon that interests you.

2.2 Open Tree of Life

The Open Tree of Life (OTOL) http://opentreeoflife.org/ is a newer system that stores published phylogenies (like TreeBASE) but also synthesizes a constantly updated version of the entire tree of life that you can explore here: https://tree.opentreeoflife.org/.

Navigate the tree by clicking the nodes. Now, try searching OTOL for a phylogeny of the same taxon you searched TreeBASE for. Unlike TreeBASE, OTOL will deliver a taxonomic tree if a molecular phylogeny has not been uploaded for your group of interest.

2.2.1 rotl R Package

To programmatically query and access OTOL data install this R package:

```
install.packages("rotl")

library(rotl)

library(ape)
```

Now we can query a small part of the tree of life as it is currently known. To extract a portion of the tree, we first need to get the ott ids (Open Tree Taxonomy Identifiers) of the taxa we're interested in:

```
apes <- c("Pan", "Pongo", "Pan", "Gorilla", "Hylobates", "Hoolock", "Homo")

apes_resolved <- tnrs_match_names(apes)

apes_resolved
```

Now we can get the tree with those tips:

```
tree <- tol_induced_subtree(ott_ids=apes_resolved$ott_id)

plot(tree)
```

Let's download a published tree by a former member of this class! Andrew Thornhill published a Myrtaceae tree that has been uploaded to the OTOL. First, get the ott id of Myrtaceae:

```
myrtaceae_resolved <- tnrs_match_names("Myrtaceae")
```

Now get the subtree under the Myrtaceae node. It's a big tree, so we'll plot it without tip labels:

```
tree <- tol_subtree(ott_id = myrtaceae_resolved$ott_id)

plot(ladderize(tree), show.tip.label = F)
```

The more authors deposit their published phylogenies in the OTOL, the easier it will get for other researchers to access up-to-date phylogenies!

3 Specimen Online Databases and Tools

3.1 Berkeley Natural History Museums (BNHM)

The BNHM is a consortium of six natural history museums located here at UC Berkeley that house over 12 million specimens. If you are studying anything in California you will likely want to use BNHM resources. These are awesome resources, so please visit each website and learn what is available!

1. University and Jepson Herbaria: Consortium of California Herbaria http://ucjeps.berkeley.edu/consortium/
2. Museum of Vertebrate Zoology: VertNet http://www.vertnet.org/
3. Essig Museum of Entomology Collections https://essigdb.berkeley.edu/
4. University of California Museum of Paleontology Database http://ucmpdb.berkeley.edu/

3.2 Global Biodiverity Information Facilty (GBIF)

GBIF is an incredibly important resource that aggregates biodiversity data from institutions around the world and makes it all available through the internet. GBIF is useful for georeferenced distribution data, and contains both specimen and observation based data. Many of the BNHM resources listed above share their data in GBIF.

If you use GBIF data you should try to double check the quality of your data, as GBIF aggregates data from multiple sources, some of which have lower quality data than others. There are several sources with suggestions on how to best curate GBIF data.

3.2.1 GBIF Web Portal

Go to http://www.gbif.org/, and click on the Get Data pull down menu. Click on Species. Search for your taxon of interest. You should be able to view a map of all the georeferenced data for your taxon. How many georeferenced records are available? You can download all the records as a CSV or Darwin Core file.

3.2.2. rgbif R Package

What if we want to automate downloading GBIF data? Here's a handy R package to programmatically access GBIF:

```
install.packages("rgbif")

library(rgbif)
```

Now let's download occurence data for a taxon. This will take a minute or so:

```
a_californica <- occ_search(scientificName="Asterella californica", limit=500)
```

We only downloaded the first 500 records, but how many total were found?

```
a_californica
```

Take a look at the first occurence:

```
a_californica$data[1,]
```

We can get the latitude and longitude of the first record:

```
a_californica$data[1,3]

a_californica$data[1,4]
```

Let's map the data using the R package ggplot2:

```
library(ggplot2)

map <- map_data("usa")

ggplot() + geom_polygon(data = map, aes( x=long, y = lat)) +
  geom_point(a_californica$data, mapping = aes( x= decimalLongitude, y = decimalLatitude, col = "red"))
```

Question 1

Do you see any problems with this distribution map? Do you trust the data? Why or why not?

3.3 BIEN Network

The BIEN network provides curated data on neotropical plants, including occurence data and trait data. If you are working with neotropical plants, check out their online portal or their R package. Click here to learn more about the services they provide.

4 Taxonomic Databases and Tools

Taxonomy is crucial when studying biodiversity because all biological data is linked through the names we use. However taxon names and concepts change, and systems to resolve synonyms are necessary.

4.1 Integrated Taxonomic Information System (ITIS)

ITIS is a partnership of US, Mexican, and Canadian government agencies that provides a database that standardizes taxonomic names. For each scientific name, ITIS includes the authority (author and date), taxonomic rank, associated synonyms and vernacular names where available, a unique taxonomic serial number, data source information (publications, experts, etc.) and data quality indicators. ITIS is often used as the absolute source of taxonomic data for large-scale biodiversity projects. Browse some of the data here: http: //www.itis.gov/

4.2 Global Names Resolver (GNR)

Often researchers have a list of taxon names, and they simply want to check the spelling and get the most up-to-date synonyms of each name. Services like the GNR can help: http://resolver.globalnames.org/

4.3 taxize R Package

The websites above are immensely helpful tools, but often we would like to use a script to check taxon names instead of copying and pasting names into the website. The R package taxize uses the GNR (and many other taxonomic databases) to do this:

```
install.packages("taxize")

library(taxize)
```

Let's check for a taxon name:

```
mynames <- gnr_resolve(names="Helianthos annus")

head(mynames)
```

Here we see that the name was misspelled, and the GNR recommended Helianthus annus instead. We can also get an accepted name from a synonym. First, get the taxonomic serial numbers (TSN) of the taxa from ITIS:

```
mynames <- c("Helianthus annuus ssp. jaegeri",
             "Helianthus annuus ssp. lenticularis",
             "Helianthus annuus ssp. texanus")

tsn <- get_tsn(mynames, accepted = FALSE)
```

Now get the accepted names for each TSN:

```
lapply(tsn, itis_acceptname)
```

The taxize package will do a lot of other handy taxonomic data wrangling, check out https://github.com/ropensci/taxize for more.

Question 2

Use R to access GBIF data and send me a map of your favorite taxon's distribution.

5 Community Phylogenetics: PICANTE

Community phylogenetics incorporates phylogenetic structure into community ecology, and tries to understand how specific ecological communities are assembled. We'll use the R package PICANTE to calculate community phylogenetic statistics and visualize the results on the phylogeny.

You need the PICANTE R package

```
install.packages("picante")

library(picante)
```

The package picante has many of the operations found in the program phylocom available from: www.phylodiversity.net/phylocom. The statistic we will focus on is NRI (Net Relatedness Index), which is similar to (Nearest Taxon Index). Both these statistics are defined in your lecture notes. There are many other functions in the picante package for phylogenetic analysis of community ecology. If you are interested in these types of analyses then I would suggest that you read over the picante manual.

We will use some of the data that is already found in the picante package. This is the made up phylogenetic tree for our example. Let's rename it 'phy'.

```r
data(phylocom)

phy <- phylocom$phylo

plot(phy)
```

This is a matrix of abundance data for the taxa in this tree from 6 different locations (communities). The names of species should match names in the phylogeny. Let's rename this 'samp'.

```r
samp <- phylocom$sample

samp
```

This is another matrix with discrete character data. Let's rename this 'traits'.

```r
traits <- phylocom$traits

traits
```

5.1 Plotting Community Data

Let's look at how the taxa found in those communities are distributed on our tree. First, let's prune any taxa from our tree that are not also represented in our sample matrix (ie. also represented in our community).

```r
prunedphy <- prune.sample(samp, phy)

prunedphy
```

We also need to make sure the species are arranged in the some order in the community data and the phylogeny. This is an important step - several functions in picante assume that the community or trait data and phylogeny data have species arranged in the same order, so it's good to always make sure we've done so before running any analysis. The following command sorts the columns of samp to be in the same order as the tip labels of the phylogeny:

```r
samporder <- samp[, prunedphy$tip.label]

samporder
```

Let's visualize our data. Now let's see how taxa from the six communities in the Phylocom example data set are arranged on the tree. The following commands set up the layout of the plot to have 2 rows and 3 columns, and then plot a black dot for the species present in each of the six communities:

```r
par(mfrow = c(2, 3))

for (i in row.names(samporder)) {

  plot(prunedphy, show.tip.label = FALSE, main = i)

  tiplabels(tip = which(samporder[i, ] > 0), pch = 19, cex = 2)
}
```

Let's also visualize the trait data that we have. We'll plot the traits with a different color for each trait value:

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

for (i in names(traits)) {

  plot(phy, show.tip.label = FALSE, main = i)

  tiplabels(pch = 22, col = traits[, i] + 1, bg = traits[,i] + 1, cex = 1.5)

}
```

Question 3

Which of the traits **appear** to have the greatest phylogenetic signal?

5.2 Calculating NRI

Now we will calculate NRI (Net Relatedness Index) for our different communities. Calculating NTI (Nearest Taxon Index) is very similar - see the PICANTE manual for instructions if you are interested. Negative NRI and NTI values indicate a high level of phylogenetic overdispersion, and positive NRI and NTI values indicate phylogenetic clustering.

First we need to make a phylogenetic distance matrix.

```
phydist <- cophenetic(phy)

phydist
```

Take a look at phydist. This is a matrix where the rows and columns are the taxa and the elements of the matrix are the phylogenetic distance between those pairs of taxa.

To calculate the NRI:

```
ses.mpd(samporder, phydist,null.model="taxa.labels")
```

The rows are the communities. The columns are:

- ntaxa: Number of taxa in community
- mpd.obs: Observed mean pairwise distance (MPD) in community
- mpd.rand.mean: Mean MPD in null communities
- mpd.rand.sd: Standard deviation of MPD in null communities
- mpd.obs.rank: Rank of observed MPD vs. null communities
- mpd.obs.z: Standardized effect size of MPD vs. null communities (equivalent to -NRI)

The fifth column is the rank of the score against randomized communities; the sixth column is the negative NRI; and the seventh column is the one tailed p-value for significantly high NRI. Communities 1, 2 and 3 are significantly clustered and community 5 is significantly spread out.

Question 4

Compare the results to the figures we made for each community – do the NRI values make sense? Explain how. You can also use different null models to calculate these values. Here, we have indicated that our null model is the distance matrix shuffled across the taxa in the community. Feel free to explore other null models if you like.

5.2 Calculating Phylogenetic Distance

We can also calculate measures of phylogenetic distance.

```r
traits <- traits[phy$tip.label, ]

multiPhylosignal(traits, phy)
```

Question 5

> Look at the column labeled 'K'. This is Bloomberg's K that we have previously discussed in lecture. K measures the phylogenetic signal of a trait – when K is greater than 1.0 there is more phylogenetic signal than expected under a Brownian motion model – closely related species resemble each other more than expected. When K is less than 1.0 it means closely related species differ more than would be expected under Brownian motion. Look at the results. What do these numbers mean? Which traits have strong phylogenetic signal? Do these values agree with your previous predictions?

Send me your answers to questions 1-5, including any relevant figures.

This lab was written by Will Freyman, David Ackerly, Nat Hallinan, Traci Grzymala and Carrie Tribble. Much of this was also taken from Community phylogenetic analysis with picante by Steven Kembel (skembel@uoregon.edu) http://picante.r-forge.r-project.org/picante-intro.pdf