Lab 09:
Phylogenetic, Specimen, and Taxonomic
Databases and R Packages

By Will Freyman

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.

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.

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.

I don't know of any journal that requires submitting trees to OTOL prior to publication, but I hope journals will move towards OTOL instead of TreeBASE because it is much easier to submit data to OTOL.

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)
```

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 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=FALSE)
```

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.

3.2.1 GBIF Web Portal

Go to `http://www.gbif.org/`, and click on the `Data` pull down menu. Click on `Explore 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:

```
occ = occ_search(scientificName="Chamerion latifolium", limit=500)
```

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

```
occ
```

Take a look at the first occurence:

```
occ$data[1,]
```

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

```
occ$data[1,3]
occ$data[1,4]
```

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

```
install.packages("ggplot2")
library(ggplot2)
gbifmap(occ$data)
```

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.

Please email me the following:

Use R to access GBIF data and send me a map of your favorite taxon's distribution. Good luck on the quiz and enjoy spring break!