

Lab 10: Biodiversity Databases and Community Phylogenetics

Jennifer Lin, GSI: 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> (<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/> (<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/> (<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)
```

```
## Warning: Some names were duplicated: 'pan'.
```

```
apes_resolved
```

```
##   search_string unique_name approximate_match ott_id is_synonym      flags
## 1          pan          Pan             FALSE 417957      FALSE SIBLING_HIGHER
## 2          pongo          Pongo             FALSE 417949      FALSE
## 3          gorilla        Gorilla             FALSE 417969      FALSE SIBLING_HIGHER
## 4        hylobates      Hylobates             FALSE 166552      FALSE
## 5          hoolock        Hoolock             FALSE 712902      FALSE
## 6           homo          Homo             FALSE 770309      FALSE SIBLING_HIGHER
##   number_matches
## 1                2
## 2                2
## 3                1
## 4                1
## 5                1
## 6                1
```

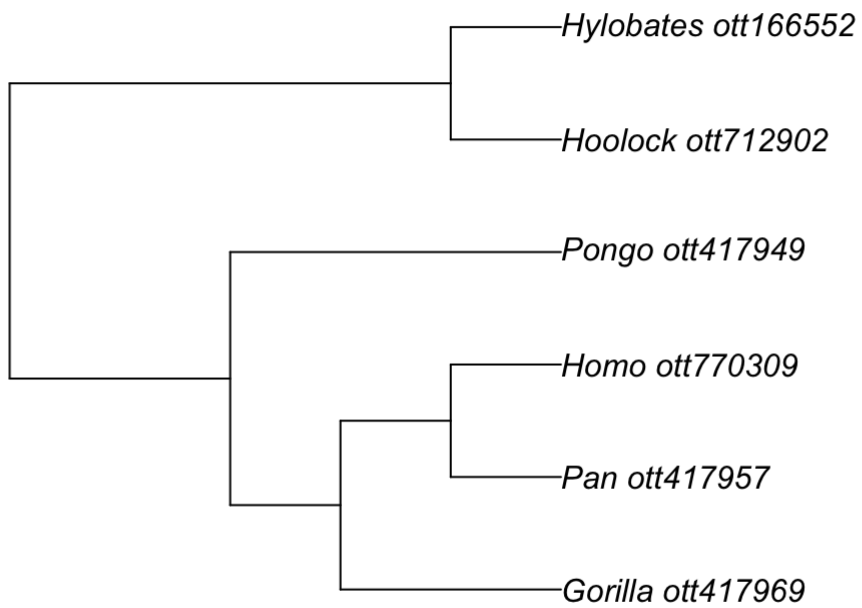
Now we can get the tree with those tips:

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

```
##
Progress [-----] 0/13 ( 0) ?s
Progress [=====] 13/13 (100) 0s
```

```
## Warning in collapse_singles(tr, show_progress): Dropping singleton nodes with
## labels: mrcaott83926ott6145147, mrcaott83926ott3607728, mrcaott83926ott3607876,
## mrcaott83926ott3607873, mrcaott83926ott3607687, mrcaott83926ott3607716,
## mrcaott83926ott3607689, mrcaott83926ott3607732, mrcaott770295ott3607719,
## mrcaott770295ott3607692, Ponginae ott1082538, mrcaott96938ott251578,
## mrcaott96938ott1055751
```

```
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:

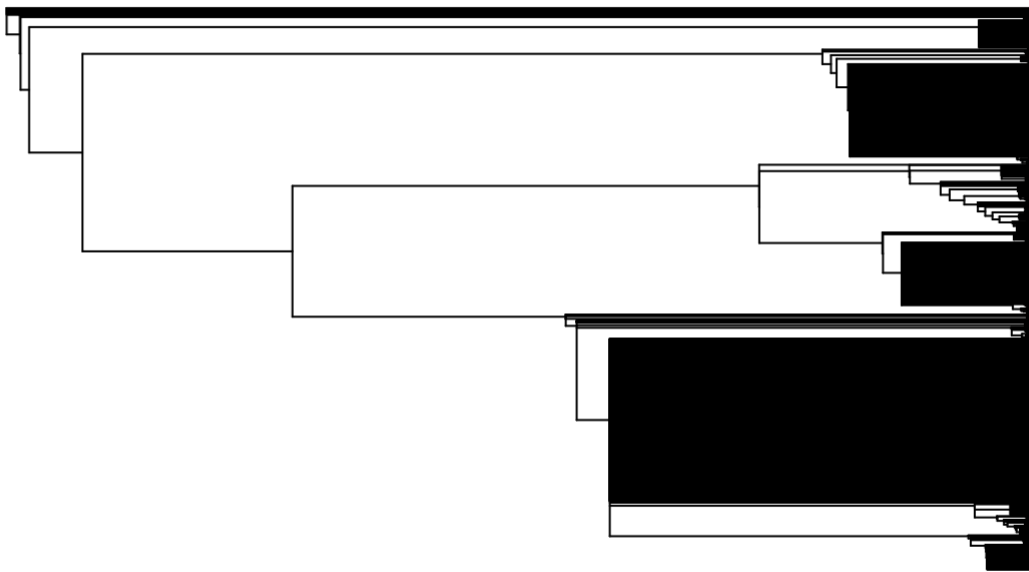
```
tree <- tol_subtree(ott_id = myrtaceae_resolved$ott_id)
```

```
##
Progress [-----] 0/160 ( 0) ?s
Progress [=====] 160/160 (100) 0s
```

Warning in collapse_singles(tr, show_progress): Dropping singleton nodes with
labels: Kjellbergiodendron ott1035839, Welchiodendron ott163307, Xanthostemoneae
ott5747112, Metrosidereae ott5747110, Syzygium masukuense ott730083, Syzygium
guineense ott731937, Syzygium wilsonii ott170431, Syzygium hemilamprum
ott774253, Syzygium eucalyptoides ott723683, Backhousieae ott5747109, Lysicarpus
ott258949, Barongia ott788761, Mitrantia ott283393, Myrceugenia myrcioides
ott376134, Myrceugenia pilotantha ott1004280, Neomyrtus ott326825, Legrandia
ott800355, Psidium cattleyanum ott807037, Myrrhinium ott3938427, Myrrhinium
atropurpureum ott3938428, Hexachlamys ott9535, Curitiba (genus in kingdom
Archaeplastida) ott3938386, Myrtus communis subsp. communis ott963929, Accara
ott3938333, Myrtastrum ott878140, Homalospermum ott959654, Kunzea ericifolia
ott136858, Paragonis ott5762546, Kunzea micrantha ott624366, Kunzea opposita
ott624374, Leptospermum madidum ott1010670, Anticoryne ott7057967, Harmogia
ott5510805, Balaustion ott628694, Astartea aspera ott7057968, Micromyrtus
elobata ott1068286, Actinodium ott299427, Chamelaucium floriferum ott1014112,
Rinzia ericaea ott628692, Syncarpieae ott5747114, Eucalyptus pachycalyx
ott278183, Eucalyptus spathulata ott1022198, Eucalyptus wandoo ott173786,
Eucalyptus loxophleba ott802953, Eucalyptus leucoxylon ott505242, Eucalyptus
leucophloia ott325242, Eucalyptus tereticornis ott173799, Eucalyptus aromaphloia
ott771611, Eucalyptus rubida ott566582, Eucalyptus ovata ott566584, Eucalyptus
dalrympleana ott771612, Eucalyptus polyanthemus ott748053, Eucalyptus lehmannii
ott264276, Eucalyptus piperita ott248674, Eucalyptus delegatensis ott771613,
Eucalyptus pauciflora ott380496, Eucalyptus marginata ott325238, Eucalyptus
rudis ott57674, Eucalyptus eudesmioides ott899588, Eucalyptus racemosa
ott243129, Eucalyptus ligulata ott514309, Eucalyptus radiata ott248689,
Eucalyptus preissiana ott1031154, Eucalyptus cinerea ott21904, Eucalyptus
approximans ott3936017, Eucalyptus persistens ott3936216, Eucalyptus alligatrix
ott3935946, Eucalyptus amplifolia ott3935815, Eucalyptus andrewsii ott3935850,
Eucalyptus apodophylla ott3935710, Eucalyptus arachnaea ott3935791, Eucalyptus
astringens ott3936225, Eucalyptus balladoniensis ott275270, Eucalyptus
brookeriana ott3935952, Eucalyptus caesia ott3935708, Eucalyptus caleyi
ott3936261, Eucalyptus calycogona ott3936083, Eucalyptus camphora ott3936099,
Eucalyptus canescens ott3935834, Eucalyptus celastroides ott3936182, Eucalyptus
conferruminata ott3936064, Eucalyptus conglobata ott3935902, Eucalyptus crucis
ott3936121, Eucalyptus decipiens ott3935737, Eucalyptus densa ott3936105,
Eucalyptus dissimulata ott3936149, Eucalyptus distuberosa ott6121966, Eucalyptus
drummondii ott3935757, Eucalyptus effusa ott3935975, Eucalyptus eremicola
ott3935936, Eucalyptus erythronema ott3935750, Eucalyptus fergusonii ott3935763,
Eucalyptus fibrosa ott3935827, Eucalyptus flocktoniae ott3935889, Eucalyptus
fraseri ott3935835, Eucalyptus gardneri ott3935809, Eucalyptus georgei
ott3935847, Eucalyptus goniantha ott3935716, Eucalyptus goniocalyx ott3936177,
Eucalyptus jutsonii ott3936221, Eucalyptus kessellii ott3936033, Eucalyptus
kingsmillii ott3935790, Eucalyptus kochii ott3936065, Eucalyptus leptocalyx
ott3935915, Eucalyptus leptopoda ott3935775, Eucalyptus macrocarpa ott3936165,
Eucalyptus macrorhyncha ott3935966, Eucalyptus mannensis ott3935934, Eucalyptus
mannifera ott3935872, Eucalyptus melanophloia ott470822, Eucalyptus mimica
ott3935994, Eucalyptus oleosa ott3936200, Eucalyptus parramattensis ott3935920,
Eucalyptus phaenophylla ott3936049, Eucalyptus phenax ott3935707, Eucalyptus
platycorys ott3936202, Eucalyptus pluricaulis ott3936019, Eucalyptus pruinosa
ott3936191, Eucalyptus resinifera ott3935754, Eucalyptus sargentii ott3936264,
Eucalyptus socialis ott3936247, Eucalyptus splendens ott3935881, Eucalyptus
strzeleckii ott3935781, Eucalyptus subangusta ott3935967, Eucalyptus suggrandis
ott3935968, Eucalyptus uncinata ott3935956, Eucalyptus vegrandis ott3935888,
Eucalyptus websteriana ott3936164, Eucalyptus xanthonema ott3936142, Angophora
bakeri ott298682, Corymbia deserticola ott902009, Corymbia cadophora ott1091560,
Corymbia ptychocarpa ott107222, Corymbia ligans ott514438, Corymbia watsoniana
ott1040947, Corymbia candida ott905331, Corymbia ferruginea ott3935595,

```
## Arillastrum ott822824, Stockwellia ott749182, Allosyncarpia ott248149,  
## Lindsayomyrteae ott5747116, Lindsayomyrtus ott484460, Melaleuca acacioides  
## ott190934, Melaleuca cajuputi (species in kingdom Archaeplastida) ott232131,  
## Osbornieae ott5747113, Osbornia (genus in kingdom Archaeplastida) ott833470,  
## Heteropyxideae ott5747123, Psiloxyleae ott5747122, Psiloxylon ott655813,  
## Phymatocarpus ott655536, Conothamnus ott602938, Eremaea (genus in kingdom  
## Archaeplastida) ott602940, Lamarchea ott602950, Pleurocalyptus ott3938345,  
## Purpureostemon ott3938368, Corynanthera ott3938448, Basisperma ott3938459,  
## Stenostegia ott4729275, Mariomella ott7640253, Pseudolaryngodus ott7641189
```

```
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/> (<http://ucjeps.berkeley.edu/consortium/>)
2. Museum of Vertebrate Zoology: VertNet <http://www.vertnet.org/> (<http://www.vertnet.org/>)
3. Essig Museum of Entomology Collections <https://essigdb.berkeley.edu/> (<https://essigdb.berkeley.edu/>)

4. University of California Museum of Paleontology Database <http://ucmpdb.berkeley.edu/>
(<http://ucmpdb.berkeley.edu/>)

3.2 Global Biodiversity Information Facility (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/> (<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 occurrence 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? 540

```
a_californica
```

```
## Records found [540]
## Records returned [500]
## No. unique hierarchies [1]
## No. media records [500]
## No. facets [0]
## Args [limit=500, offset=0, scientificName=Asterella californica, fields=all]
## # A tibble: 500 x 142
##   key scientificName decimalLatitude decimalLongitude issues datasetKey
##   <chr> <chr>                <dbl>          <dbl> <chr> <chr>
## 1 2543... Asterella cal...      32.8        -117. cdrou... 50c9509d-...
## 2 2543... Asterella cal...      32.7        -117. cdrou... 50c9509d-...
## 3 2557... Asterella cal...      37.4        -122. cdrou... 50c9509d-...
## 4 2557... Asterella cal...      37.9        -122. cdrou... 50c9509d-...
## 5 2557... Asterella cal...      37.9        -122. cdrou... 50c9509d-...
## 6 2563... Asterella cal...      32.9        -117. gass84 50c9509d-...
## 7 2563... Asterella cal...      33.1        -117. cdrou... 50c9509d-...
## 8 2563... Asterella cal...      33.6        -118. cdrou... 50c9509d-...
## 9 2563... Asterella cal...      33.7        -118. cdrou... 50c9509d-...
## 10 2563... Asterella cal...      32.8        -117. cdrou... 50c9509d-...
## # ... with 490 more rows, and 136 more variables: publishingOrgKey <chr>,
## #   installationKey <chr>, publishingCountry <chr>, protocol <chr>,
## #   lastCrawled <chr>, lastParsed <chr>, crawlId <int>, extensions <chr>,
## #   basisOfRecord <chr>, taxonKey <int>, kingdomKey <int>, phylumKey <int>,
## #   classKey <int>, orderKey <int>, familyKey <int>, genusKey <int>,
## #   speciesKey <int>, acceptedTaxonKey <int>, acceptedScientificName <chr>,
## #   kingdom <chr>, phylum <chr>, order <chr>, family <chr>, genus <chr>,
## #   species <chr>, genericName <chr>, specificEpithet <chr>, taxonRank <chr>,
## #   taxonomicStatus <chr>, dateIdentified <chr>,
## #   coordinateUncertaintyInMeters <dbl>, stateProvince <chr>, year <int>,
## #   month <int>, day <int>, eventDate <chr>, modified <chr>,
## #   lastInterpreted <chr>, references <chr>, license <chr>, identifiers <chr>,
## #   facts <chr>, relations <chr>, geodeticDatum <chr>, class <chr>,
## #   countryCode <chr>, recordedByIDs <chr>, identifiedByIDs <chr>,
## #   country <chr>, rightsHolder <chr>, identifier <chr>,
## #   http...unknown.org.nick <chr>, verbatimEventDate <chr>, datasetName <chr>,
## #   verbatimLocality <chr>, collectionCode <chr>, gbifID <chr>,
## #   occurrenceID <chr>, taxonID <chr>, recordedBy <chr>, catalogNumber <chr>,
## #   http...unknown.org.occurrenceDetails <chr>, institutionCode <chr>,
## #   rights <chr>, eventTime <chr>, identificationID <chr>, name <chr>,
## #   recordedByIDs.type <chr>, recordedByIDs.value <chr>,
## #   identifiedByIDs.type <chr>, identifiedByIDs.value <chr>,
## #   occurrenceRemarks <chr>, informationWithheld <chr>,
## #   identificationRemarks <chr>, elevation <dbl>, continent <chr>,
## #   nomenclaturalStatus <chr>, recordNumber <chr>, nomenclaturalCode <chr>,
## #   locality <chr>, county <chr>, type <chr>, ownerInstitutionCode <chr>,
## #   bibliographicCitation <chr>, identifiedBy <chr>, collectionID <chr>,
## #   habitat <chr>, institutionID <chr>, endDayOfYear <chr>, language <chr>,
## #   preparations <chr>, startDayOfYear <chr>, verbatimElevation <chr>,
## #   elevationAccuracy <dbl>, individualCount <int>,
## #   http...unknown.org.recordId <chr>, otherCatalogNumbers <chr>,
## #   http...unknown.org.recordEnteredBy <chr>, higherGeography <chr>,
## #   dynamicProperties <chr>, ...
```

Take a look at the first occurrence:

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

```
## # A tibble: 1 x 142
##   key      scientificName decimalLatitude decimalLongitude issues datasetKey
##   <chr> <chr>                <dbl>                <dbl> <chr> <chr>
## 1 2543... Asterella cal...          32.8                -117. cdrou... 50c9509d-...
## # ... with 136 more variables: publishingOrgKey <chr>, installationKey <chr>,
## #   publishingCountry <chr>, protocol <chr>, lastCrawled <chr>,
## #   lastParsed <chr>, crawlId <int>, extensions <chr>, basisOfRecord <chr>,
## #   taxonKey <int>, kingdomKey <int>, phylumKey <int>, classKey <int>,
## #   orderKey <int>, familyKey <int>, genusKey <int>, speciesKey <int>,
## #   acceptedTaxonKey <int>, acceptedScientificName <chr>, kingdom <chr>,
## #   phylum <chr>, order <chr>, family <chr>, genus <chr>, species <chr>,
## #   genericName <chr>, specificEpithet <chr>, taxonRank <chr>,
## #   taxonomicStatus <chr>, dateIdentified <chr>,
## #   coordinateUncertaintyInMeters <dbl>, stateProvince <chr>, year <int>,
## #   month <int>, day <int>, eventDate <chr>, modified <chr>,
## #   lastInterpreted <chr>, references <chr>, license <chr>, identifiers <chr>,
## #   facts <chr>, relations <chr>, geodeticDatum <chr>, class <chr>,
## #   countryCode <chr>, recordedByIDs <chr>, identifiedByIDs <chr>,
## #   country <chr>, rightsHolder <chr>, identifier <chr>,
## #   http...unknown.org.nick <chr>, verbatimEventDate <chr>, datasetName <chr>,
## #   verbatimLocality <chr>, collectionCode <chr>, gbifID <chr>,
## #   occurrenceID <chr>, taxonID <chr>, recordedBy <chr>, catalogNumber <chr>,
## #   http...unknown.org.occurrenceDetails <chr>, institutionCode <chr>,
## #   rights <chr>, eventTime <chr>, identificationID <chr>, name <chr>,
## #   recordedByIDs.type <chr>, recordedByIDs.value <chr>,
## #   identifiedByIDs.type <chr>, identifiedByIDs.value <chr>,
## #   occurrenceRemarks <chr>, informationWithheld <chr>,
## #   identificationRemarks <chr>, elevation <dbl>, continent <chr>,
## #   nomenclaturalStatus <chr>, recordNumber <chr>, nomenclaturalCode <chr>,
## #   locality <chr>, county <chr>, type <chr>, ownerInstitutionCode <chr>,
## #   bibliographicCitation <chr>, identifiedBy <chr>, collectionID <chr>,
## #   habitat <chr>, institutionID <chr>, endDayOfYear <chr>, language <chr>,
## #   preparations <chr>, startDayOfYear <chr>, verbatimElevation <chr>,
## #   elevationAccuracy <dbl>, individualCount <int>,
## #   http...unknown.org.recordId <chr>, otherCatalogNumbers <chr>,
## #   http...unknown.org.recordEnteredBy <chr>, higherGeography <chr>,
## #   dynamicProperties <chr>, ...
```

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

```
a_californica$data[1,3]
```

```
## # A tibble: 1 x 1
##   decimalLatitude
##           <dbl>
## 1             32.8
```

```
a_californica$data[1,4]
```

```
## # A tibble: 1 x 1
##   decimalLongitude
##           <dbl>
## 1            -117.
```

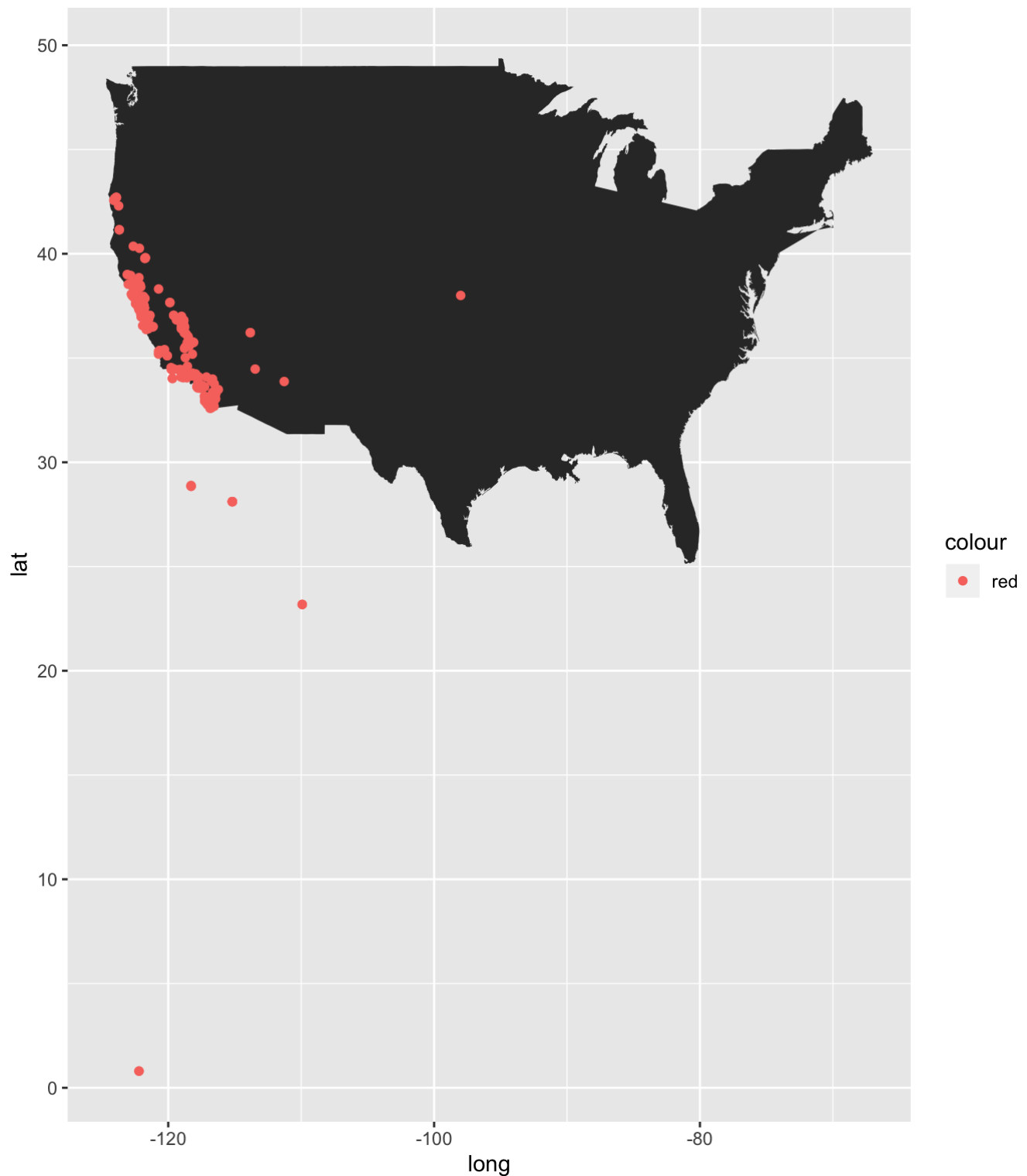

Let's map the data using the R package ggplot2:

```
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"))
```

```
## Warning: Removed 116 rows containing missing values (geom_point).
```



Question 1

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

Answer 1

According to the information I searched online, *Asterella californica* occurs throughout the Coast Ranges, western Klamath Ranges, Sacramento Valley, west slope of the Sierra Nevada, coastal Southern California and the Channel Islands, Transverse and Peninsular Ranges, western Colorado Desert; east to Arizona, north to southwestern Oregon, and south to northern Baja California.

Most of the occurrence records are located near the region mentioned above, which seems to be correct. However, the occurrence record in the middle of US and other occurrence records located on the ocean seem not right to me. The occurrence record in the middle of US is very far away from other occurrence records. In addition, there is only one record in that region. Thus, it may be another species which has similar morphology to *A. californica*. Other records located on the ocean are even more suspicious since there are no lands for *A. californica* to grow. It may be some errors generated when entering these records to the database.

3.3 BIEN Network

The BIEN network provides curated data on neotropical plants, including occurrence data and trait data. If you are working with neotropical plants, check out their online portal (biendata.org) or their R package (https://github.com/bmaitner/RBIEN/blob/master/tutorials/RBIEN_tutorial.Rmd). Click here (<http://bien.nceas.ucsb.edu/>) 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/> (<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)
```

```
##  
## Attaching package: 'taxize'
```

```
## The following objects are masked from 'package:rotl':  
##  
##     synonyms, tax_name, tax_rank
```

Let's check for a taxon name:

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

```
head(mynames)
```

```
## # A tibble: 6 x 5  
##   user_supplied_name submitted_name matched_name data_source_title score  
##   <chr>             <chr>         <chr>         <chr>             <dbl>  
## 1 Helianthos annus Helianthos ann... Helianthus annus uBio NameBank      0.75  
## 2 Helianthos annus Helianthos ann... Helianthus annu... Catalogue of Life   0.75  
## 3 Helianthos annus Helianthos ann... Helianthus annu... ITIS                0.75  
## 4 Helianthos annus Helianthos ann... Helianthus annu... NCBI                0.75  
## 5 Helianthos annus Helianthos ann... Helianthus annu... GRIN Taxonomy for P... 0.75  
## 6 Helianthos annus Helianthos ann... Helianthus annu... Union 4            0.75
```

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)
```

```
## == 3 queries ==
```

```
##  
## Retrieving data for taxon 'Helianthus annuus ssp. jaegeri'
```

```
## ✓ Found: Helianthus annuus ssp. jaegeri
```

```
##  
## Retrieving data for taxon 'Helianthus annuus ssp. lenticularis'
```

```
## ✓ Found: Helianthus annuus ssp. lenticularis
```

```
##  
## Retrieving data for taxon 'Helianthus annuus ssp. texanus'
```

```
## ✓ Found: Helianthus annuus ssp. texanus  
## == Results ==  
##  
## ● Total: 3  
## ● Found: 3  
## ● Not Found: 0
```

Now get the accepted names for each TSN:

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

```
## [[1]]  
## submittedtsn acceptedname acceptedtsn author  
## 1 525928 Helianthus annuus 36616 L.  
##  
## [[2]]  
## submittedtsn acceptedname acceptedtsn author  
## 1 525929 Helianthus annuus 36616 L.  
##  
## [[3]]  
## submittedtsn acceptedname acceptedtsn author  
## 1 525930 Helianthus annuus 36616 L.
```

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

Answer 2

Download occurrence data for *Sciurus carolinensis*.

```
s_carolinensis <- occ_search(scientificName="Sciurus carolinensis", limit=500)
```

We only downloaded the first 500 records, but how many total were found? 162239

```
s_carolinensis
```

```
## Records found [162239]
## Records returned [500]
## No. unique hierarchies [1]
## No. media records [500]
## No. facets [0]
## Args [limit=500, offset=0, scientificName=Sciurus carolinensis, fields=all]
## # A tibble: 500 x 80
##   key    scientificName decimalLatitude decimalLongitude issues datasetKey
##   <chr> <chr>                <dbl>          <dbl> <chr>   <chr>
## 1 2542... Sciurus carol...      40.8          -74.0 cdrou... 50c9509d-...
## 2 2542... Sciurus carol...      44.3          -78.3 cdrou... 50c9509d-...
## 3 2542... Sciurus carol...      39.0          -77.2 cdrou... 50c9509d-...
## 4 2542... Sciurus carol...      28.3          -81.4 cdrou... 50c9509d-...
## 5 2542... Sciurus carol...      38.8          -77.1 cdrou... 50c9509d-...
## 6 2542... Sciurus carol...      43.1          -87.9 cdrou... 50c9509d-...
## 7 2542... Sciurus carol...      39.0          -77.2 cdrou... 50c9509d-...
## 8 2542... Sciurus carol...      31.3          -94.7 gass84 50c9509d-...
## 9 2542... Sciurus carol...      48.8          -93.7 cdrou... 50c9509d-...
## 10 2543... Sciurus carol...      43.1          -87.9 cdrou... 50c9509d-...
## # ... with 490 more rows, and 74 more variables: publishingOrgKey <chr>,
## #   installationKey <chr>, publishingCountry <chr>, protocol <chr>,
## #   lastCrawled <chr>, lastParsed <chr>, crawlId <int>, extensions <chr>,
## #   basisOfRecord <chr>, taxonKey <int>, kingdomKey <int>, phylumKey <int>,
## #   classKey <int>, orderKey <int>, familyKey <int>, genusKey <int>,
## #   speciesKey <int>, acceptedTaxonKey <int>, acceptedScientificName <chr>,
## #   kingdom <chr>, phylum <chr>, order <chr>, family <chr>, genus <chr>,
## #   species <chr>, genericName <chr>, specificEpithet <chr>, taxonRank <chr>,
## #   taxonomicStatus <chr>, dateIdentified <chr>,
## #   coordinateUncertaintyInMeters <dbl>, stateProvince <chr>, year <int>,
## #   month <int>, day <int>, eventDate <chr>, modified <chr>,
## #   lastInterpreted <chr>, references <chr>, license <chr>, identifiers <chr>,
## #   facts <chr>, relations <chr>, geodeticDatum <chr>, class <chr>,
## #   countryCode <chr>, recordedByIDs <chr>, identifiedByIDs <chr>,
## #   country <chr>, rightsHolder <chr>, identifier <chr>,
## #   http...unknown.org.nick <chr>, verbatimEventDate <chr>, datasetName <chr>,
## #   verbatimLocality <chr>, collectionCode <chr>, gbifID <chr>,
## #   occurrenceID <chr>, taxonID <chr>, recordedBy <chr>, catalogNumber <chr>,
## #   http...unknown.org.occurrenceDetails <chr>, institutionCode <chr>,
## #   rights <chr>, eventTime <chr>, identificationID <chr>, name <chr>,
## #   informationWithheld <chr>, occurrenceRemarks <chr>,
## #   infraspecificEpithet <chr>, recordedByIDs.type <chr>,
## #   recordedByIDs.value <chr>, identifiedByIDs.type <chr>,
## #   identifiedByIDs.value <chr>
```

Take a look at the first occurrence:

```
s_carolinensis$data[1,]
```

```
## # A tibble: 1 x 80
##   key      scientificName decimalLatitude decimalLongitude issues datasetKey
##   <chr> <chr>                <dbl>                <dbl> <chr> <chr>
## 1 2542... Sciurus carol...      40.8                -74.0 cdrou... 50c9509d-...
## # ... with 74 more variables: publishingOrgKey <chr>, installationKey <chr>,
## #   publishingCountry <chr>, protocol <chr>, lastCrawled <chr>,
## #   lastParsed <chr>, crawlId <int>, extensions <chr>, basisOfRecord <chr>,
## #   taxonKey <int>, kingdomKey <int>, phylumKey <int>, classKey <int>,
## #   orderKey <int>, familyKey <int>, genusKey <int>, speciesKey <int>,
## #   acceptedTaxonKey <int>, acceptedScientificName <chr>, kingdom <chr>,
## #   phylum <chr>, order <chr>, family <chr>, genus <chr>, species <chr>,
## #   genericName <chr>, specificEpithet <chr>, taxonRank <chr>,
## #   taxonomicStatus <chr>, dateIdentified <chr>,
## #   coordinateUncertaintyInMeters <dbl>, stateProvince <chr>, year <int>,
## #   month <int>, day <int>, eventDate <chr>, modified <chr>,
## #   lastInterpreted <chr>, references <chr>, license <chr>, identifiers <chr>,
## #   facts <chr>, relations <chr>, geodeticDatum <chr>, class <chr>,
## #   countryCode <chr>, recordedByIDs <chr>, identifiedByIDs <chr>,
## #   country <chr>, rightsHolder <chr>, identifier <chr>,
## #   http...unknown.org.nick <chr>, verbatimEventDate <chr>, datasetName <chr>,
## #   verbatimLocality <chr>, collectionCode <chr>, gbifID <chr>,
## #   occurrenceID <chr>, taxonID <chr>, recordedBy <chr>, catalogNumber <chr>,
## #   http...unknown.org.occurrenceDetails <chr>, institutionCode <chr>,
## #   rights <chr>, eventTime <chr>, identificationID <chr>, name <chr>,
## #   informationWithheld <chr>, occurrenceRemarks <chr>,
## #   infraspecificEpithet <chr>, recordedByIDs.type <chr>,
## #   recordedByIDs.value <chr>, identifiedByIDs.type <chr>,
## #   identifiedByIDs.value <chr>
```

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

```
s_carolinensis$data[1,3]
```

```
## # A tibble: 1 x 1
##   decimalLatitude
##           <dbl>
## 1             40.8
```

```
s_carolinensis$data[1,4]
```

```
## # A tibble: 1 x 1
##   decimalLongitude
##           <dbl>
## 1            -74.0
```

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(s_carolinensis$data, mapping = aes( x= decimalLongitude, y = decimalLatitude, col = "red"))
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



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)
```

```
## Loading required package: vegan
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-6
```

```
## Loading required package: nlme
```

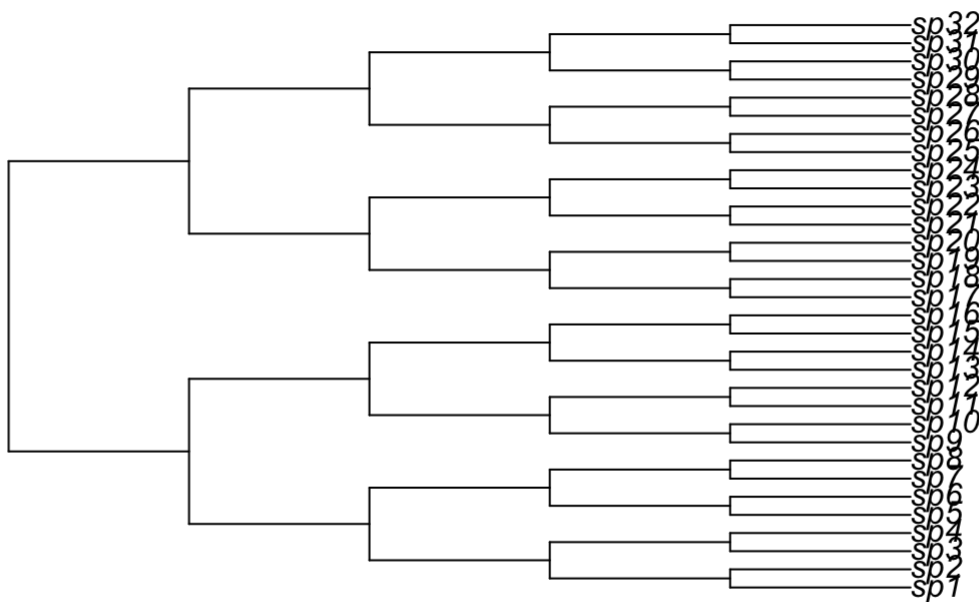
The package *picante* has many of the operations found in the program *phylocom* available from: www.phylodiversity.net/phylocom. The statistic we will focus on is NRI (Net Relatedness Index), which is similar to (Nearest Taxon Index). Both these statistics are defined in your lecture notes (http://ib.berkeley.edu/courses/ib200/lect/mar20_DA.pdf). 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'.

```
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'.

```
samp <- phylocom$samp
```

```
samp
```

```
##          sp1 sp10 sp11 sp12 sp13 sp14 sp15 sp17 sp18 sp19 sp2 sp20 sp21 sp22
## clump1      1   0   0   0   0   0   0   0   0   0   1   0   0   0
## clump2a     1   2   2   2   0   0   0   0   0   0   1   0   0   0
## clump2b     1   0   0   0   0   0   0   2   2   2   1   2   0   0
## clump4      1   1   0   0   0   0   0   2   2   0   1   0   0   0
## even        1   0   0   0   1   0   0   1   0   0   0   0   1   0
## random      0   0   0   1   0   4   2   3   0   0   1   0   0   1
##          sp24 sp25 sp26 sp29 sp3 sp4 sp5 sp6 sp7 sp8 sp9
## clump1      0   0   0   0   1   1   1   1   1   1   0
## clump2a     0   0   0   0   1   1   0   0   0   0   2
## clump2b     0   0   0   0   1   1   0   0   0   0   0
## clump4      0   2   2   0   0   0   0   0   0   0   1
## even        0   1   0   1   0   0   1   0   0   0   1
## random      2   0   0   0   0   0   2   0   0   0   0
```

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

```
traits <- phylocom$traits
```

```
traits
```

##	traitA	traitB	traitC	traitD
## sp1	1	1	1	0
## sp2	1	1	2	0
## sp3	2	1	3	0
## sp4	2	1	4	0
## sp5	2	2	1	0
## sp6	2	2	2	0
## sp7	2	2	3	0
## sp8	2	2	4	0
## sp9	1	3	1	1
## sp10	1	3	2	1
## sp11	2	3	3	1
## sp12	2	3	4	1
## sp13	2	4	1	1
## sp14	2	4	2	1
## sp15	2	4	3	1
## sp16	2	4	4	1
## sp17	1	1	1	1
## sp18	1	1	2	1
## sp19	2	1	3	1
## sp20	2	1	4	1
## sp21	2	2	1	1
## sp22	2	2	2	1
## sp23	2	2	3	1
## sp24	2	2	4	1
## sp25	1	3	1	1
## sp26	1	3	2	1
## sp27	2	3	3	1
## sp28	2	3	4	1
## sp29	2	4	1	1
## sp30	2	4	2	1
## sp31	2	4	3	1
## sp32	2	4	4	1

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).

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

```
prunedphy
```

```
##
## Phylogenetic tree with 25 tips and 24 internal nodes.
##
## Tip labels:
##  sp1, sp2, sp3, sp4, sp5, sp6, ...
## Node labels:
##  A, B, C, D, E, F, ...
##
## Rooted; includes branch lengths.
```

We also need to make sure the species are arranged in the same 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:

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

```
samporder
```

```
##           sp1 sp2 sp3 sp4 sp5 sp6 sp7 sp8 sp9 sp10 sp11 sp12 sp13 sp14 sp15 sp17
## clump1      1  1  1  1  1  1  1  1  0  0  0  0  0  0  0  0
## clump2a     1  1  1  1  0  0  0  0  2  2  2  2  0  0  0  0
## clump2b     1  1  1  1  0  0  0  0  0  0  0  0  0  0  0  2
## clump4      1  1  0  0  0  0  0  0  1  1  0  0  0  0  0  2
## even        1  0  0  0  1  0  0  0  1  0  0  0  1  0  0  1
## random      0  1  0  0  2  0  0  0  0  0  0  1  0  4  2  3
##           sp18 sp19 sp20 sp21 sp22 sp24 sp25 sp26 sp29
## clump1        0  0  0  0  0  0  0  0  0
## clump2a        0  0  0  0  0  0  0  0  0
## clump2b        2  2  2  0  0  0  0  0  0
## clump4         2  0  0  0  0  0  2  2  0
## even           0  0  0  1  0  0  1  0  1
## random         0  0  0  0  1  2  0  0  0
```

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:

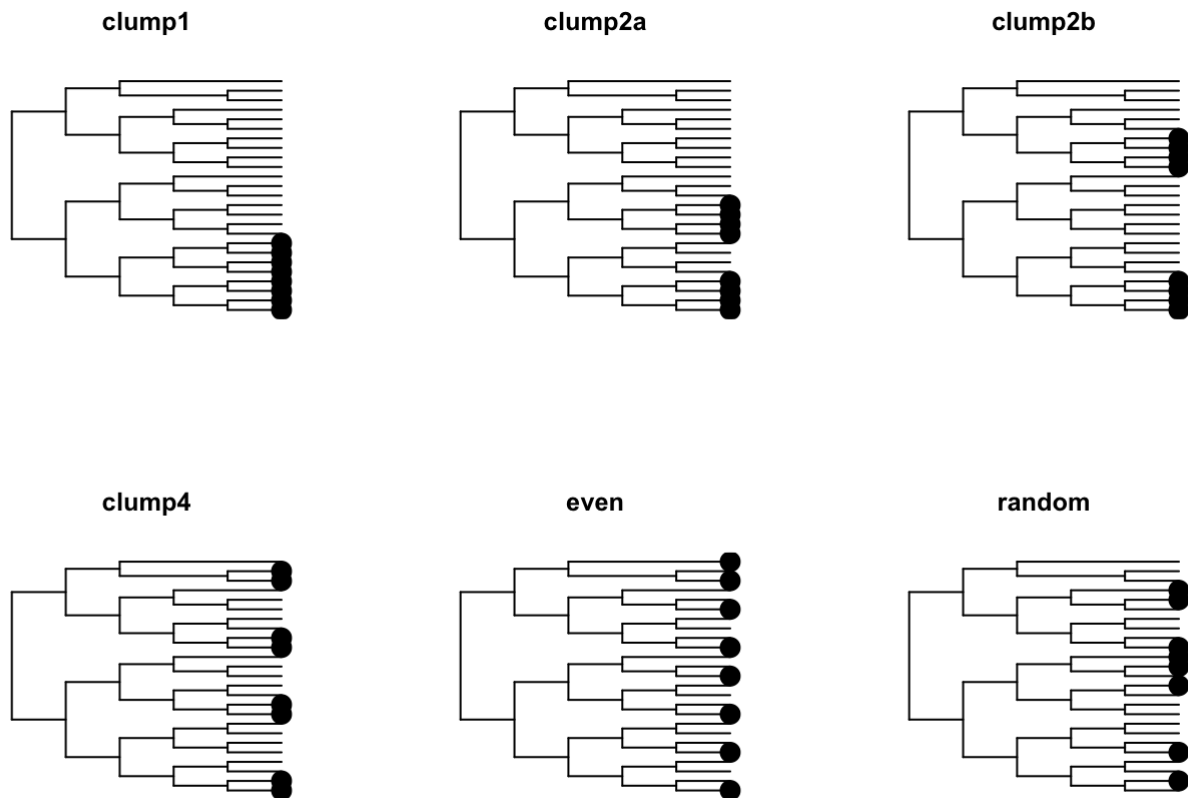
```
par(mfrow = c(2, 3))
```

```
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:

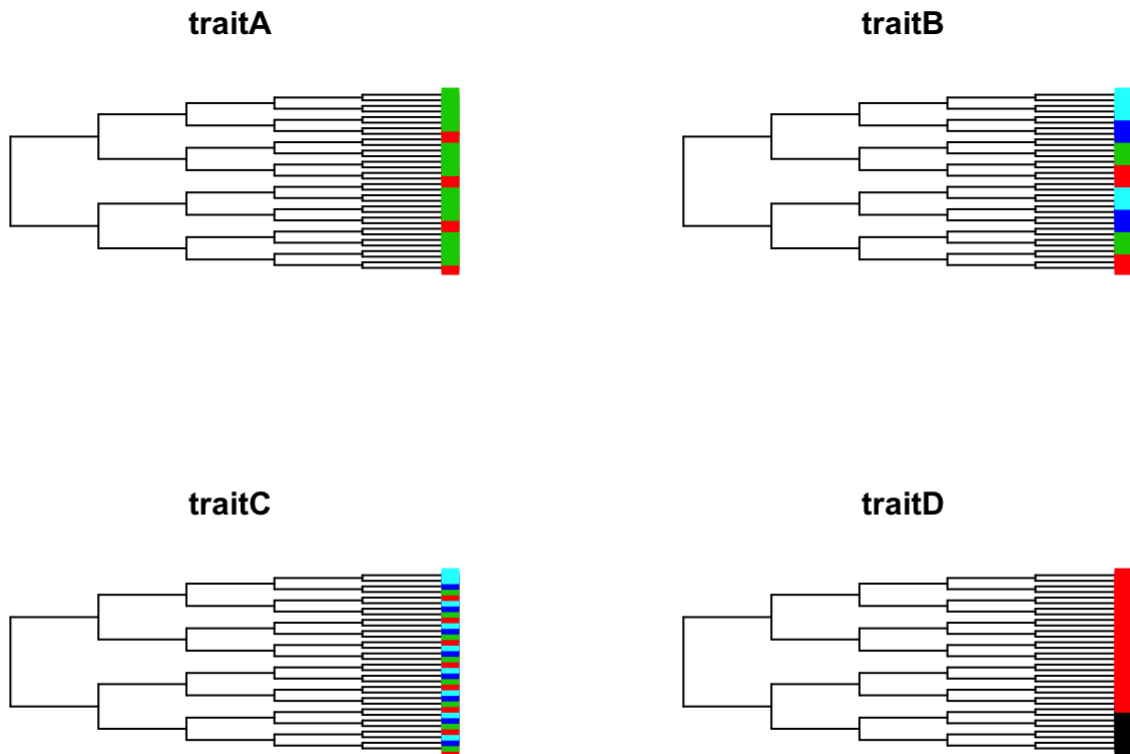
```
par(mfrow = c(2, 2))

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?

Answer 3

Phylogenetic signal is the tendency of related species to resemble each other more than species drawn at random from the same tree. Thus, traitD appears to have the greatest phylogenetic signal.

Phylogenetic signal: traitD > traitA > traitB > traitC

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
```

##	sp1	sp2	sp3	sp4	sp5	sp6	sp7	sp8	sp9	sp10	sp11	sp12	sp13	sp14	sp15	sp16
## sp1	0	2	4	4	6	6	6	6	8	8	8	8	8	8	8	8
## sp2	2	0	4	4	6	6	6	6	8	8	8	8	8	8	8	8
## sp3	4	4	0	2	6	6	6	6	8	8	8	8	8	8	8	8
## sp4	4	4	2	0	6	6	6	6	8	8	8	8	8	8	8	8
## sp5	6	6	6	6	0	2	4	4	8	8	8	8	8	8	8	8
## sp6	6	6	6	6	2	0	4	4	8	8	8	8	8	8	8	8
## sp7	6	6	6	6	4	4	0	2	8	8	8	8	8	8	8	8
## sp8	6	6	6	6	4	4	2	0	8	8	8	8	8	8	8	8
## sp9	8	8	8	8	8	8	8	8	0	2	4	4	6	6	6	6
## sp10	8	8	8	8	8	8	8	8	2	0	4	4	6	6	6	6
## sp11	8	8	8	8	8	8	8	8	4	4	0	2	6	6	6	6
## sp12	8	8	8	8	8	8	8	8	4	4	2	0	6	6	6	6
## sp13	8	8	8	8	8	8	8	8	6	6	6	6	0	2	4	4
## sp14	8	8	8	8	8	8	8	8	6	6	6	6	2	0	4	4
## sp15	8	8	8	8	8	8	8	8	6	6	6	6	4	4	0	2
## sp16	8	8	8	8	8	8	8	8	6	6	6	6	4	4	2	0
## sp17	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp18	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp19	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp20	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp21	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp22	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp23	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp24	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp25	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp26	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp27	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp28	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp29	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp30	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp31	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
## sp32	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
##	sp17	sp18	sp19	sp20	sp21	sp22	sp23	sp24	sp25	sp26	sp27	sp28	sp29	sp30	sp31	
## sp1	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp2	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp3	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp4	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp5	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp6	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp7	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp8	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp9	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp11	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp12	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp13	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp14	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp15	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp16	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	
## sp17	0	2	4	4	6	6	6	6	8	8	8	8	8	8	8	
## sp18	2	0	4	4	6	6	6	6	8	8	8	8	8	8	8	
## sp19	4	4	0	2	6	6	6	6	8	8	8	8	8	8	8	
## sp20	4	4	2	0	6	6	6	6	8	8	8	8	8	8	8	
## sp21	6	6	6	6	0	2	4	4	8	8	8	8	8	8	8	
## sp22	6	6	6	6	2	0	4	4	8	8	8	8	8	8	8	
## sp23	6	6	6	6	4	4	0	2	8	8	8	8	8	8	8	

```

## sp24      6      6      6      6      4      4      2      0      8      8      8      8      8      8      8
## sp25      8      8      8      8      8      8      8      8      0      2      4      4      6      6      6
## sp26      8      8      8      8      8      8      8      8      2      0      4      4      6      6      6
## sp27      8      8      8      8      8      8      8      8      4      4      0      2      6      6      6
## sp28      8      8      8      8      8      8      8      8      4      4      2      0      6      6      6
## sp29      8      8      8      8      8      8      8      8      6      6      6      6      0      2      4
## sp30      8      8      8      8      8      8      8      8      6      6      6      6      2      0      4
## sp31      8      8      8      8      8      8      8      8      6      6      6      6      4      4      0
## sp32      8      8      8      8      8      8      8      8      6      6      6      6      4      4      2
##          sp32
## sp1       10
## sp2       10
## sp3       10
## sp4       10
## sp5       10
## sp6       10
## sp7       10
## sp8       10
## sp9       10
## sp10      10
## sp11      10
## sp12      10
## sp13      10
## sp14      10
## sp15      10
## sp16      10
## sp17       8
## sp18       8
## sp19       8
## sp20       8
## sp21       8
## sp22       8
## sp23       8
## sp24       8
## sp25       6
## sp26       6
## sp27       6
## sp28       6
## sp29       4
## sp30       4
## sp31       2
## sp32       0

```

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")
```

```
##          ntaxa  mpd.obs  mpd.rand.mean  mpd.rand.sd  mpd.obs.rank  mpd.obs.z
## clump1      8  4.857143      8.287502    0.3547470        1.0 -9.6698742
## clump2a     8  6.000000      8.315244    0.3313030        1.0 -6.9882975
## clump2b     8  7.142857      8.317603    0.3245419        8.0 -3.6197063
## clump4      8  8.285714      8.319105    0.3105994       400.0 -0.1075035
## even        8  8.857143      8.326684    0.3208941       996.0  1.6530659
## random      8  8.428571      8.324896    0.3208228       533.5  0.3231538
##          mpd.obs.p  runs
## clump1      0.0010    999
## clump2a     0.0010    999
## clump2b     0.0080    999
## clump4      0.4000    999
## even        0.9960    999
## random      0.5335    999
```

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.

Answer 4

As mentioned before, negative NRI values indicate a high level of phylogenetic overdispersion, and positive NRI values indicate phylogenetic clustering. Because mpd.obs.z is equivalent to -NRI, we can rephrase the sentence above to “positive mpd.obs.z values indicate a high level of phylogenetic overdispersion, and negative mpd.obs.z values indicate phylogenetic clustering”.

We can see that clump1 has the most negative mpd.obs.z value (= most positive NRI value), which indicates high phylogenetic clustering. This is correct because clump1 only have one clump. As for clump2a, clump2b, to clump4, the mpd.obs.z values get larger and larger (NRI value get smaller and smaller), which means the level of phylogenetic overdispersion is getting higher and higher.

To sum up, the NRI values make sense in this scenario.

5.2 Calculating Phylogenetic Distance

We can also calculate measures of phylogenetic distance.


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

multiPhylosignal(traits, phy)
```

```
##           K PIC.variance.obs PIC.variance.rnd.mean PIC.variance.P
## traitA 0.8905609      0.05396825      0.1259181      0.001
## traitB 2.9340184      0.10920635      0.8329285      0.001
## traitC 0.5149502      0.62222222      0.8395371      0.050
## traitD 4.3536696      0.01103943      0.1259887      0.001
##           PIC.variance.Z
## traitA      -3.676127
## traitB      -5.265029
## traitC      -1.633409
## traitD      -5.952393
```

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?

Question 5

What do these numbers mean?

According to the Picante package documentation:

The statistical significance of phylogenetic signal can be evaluated by comparing observed patterns of the variance of independent contrasts of the trait to a null model of shuffling taxa labels across the tips of the phylogeny.

The higher the K statistic, the more phylogenetic signal in a trait.

PIC.variance.P is the quantile of the observed phylogenetically independent contrast variance versus the null distribution, which can be used as a 1-tailed P-value to test for greater phylogenetic signal than expected. Traits with PIC.variance.P < 0.05 have non-random phylogenetic signal.

Which traits have strong phylogenetic signal?

TraitD and traitB have strong phylogenetic signal. According to the column labeled "K", traitD and traitB have K values larger than 1, which indicates that closely related species resemble each other more than expected. As for traitA and traitC, they have K values smaller than 1, which indicates that closely related species differ more than would be expected under Brownian motion.

Do these values agree with your previous predictions?

These results are overall agree with my previous predictions (Phylogenetic signal: traitD > traitA > traitB > traitC).

However, since traitC has a PIC.variance.P larger than 0.05, it may not be a significant non-random phylogenetic signal. Thus, we shouldn't put too much trust on the results of traitC.

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 (mailto:skembel@uoregon.edu)) <http://picante.r-forge.r-project.org/picante-intro.pdf> (<http://picante.r-forge.r-project.org/picante-intro.pdf>)