PhyloMeth: Getting Trees and Data

Brian C. O'Meara

24 January, 2022

Let's get some trees from Open Tree of Life. This is in some ways a successor to TreeBASE: another repository of trees (though it has only partial overlap with the trees in TreeBASE (or in another repository, Dryad). Unlike TreeBASE, anyone, not just the author, can add trees to Open Tree's repository. Another important distinction is that Open Tree seeks to create a single tree for all life by creating essentially a supreme super tree.

Now let's get ya tree, replacing the `eval=FALSE` with `eval=TRUE` and replacing the missing info.

```
GetTreeFromOpenTree <- function() {
    library(rotl)
    library(ape)

 formica.id <- rotl::tnrs_match_names("__________")$ott_id

    # Now get Open Tree's current best estimate of the phylogeny for the group
    # They call this the tree of life; we can get the subtree for just this group.
    formica.tree <- rotl::tol_subtree(ott_id=formica.id)

    # Let's plot the tree:
    ape::plot.phylo("__________", type="fan", cex=0.2)
}
```

It has a lot of polytomies, representing uncertainty. A maximally resolved tree (if rooted) will have one fewer internal nodes than terminal nodes: think of a tree with three taxa, ((A,B),C): it will have the MRCA of A and B and the MRCA of A, B, and C: three terminals, two internal nodes. If it had no information, it would only have one node. So we can look at the ratio of number of internal nodes to number of possible internal nodes to figure out how resolved a tree is (subtracting 1 from each to account for the root node that must always exist)

```
    print(paste("The formica tree has ", ape::Ntip(formica.tree), " terminals and ",
    Nnode(formica.tree), " internal nodes out of ",ape::Ntip(formica.tree)-2,
    " possible, which means it is ",
    round(100*(ape::Nnode(formica.tree)-1)/(ape::Ntip(formica.tree)-3), 2),
    "% resolved", sep=""))

    # Open Tree can also return the original studies with the source trees.
    formica.trees <- studies_find_trees(property="ot:ottTaxonName", value="Formica", detailed=FALSE)
    formica.studies.ids <- unlist(formica.trees$study_ids)

    # Let's get info on the first study
    formica.study1.metadata <- rotl::get_study_meta(formica.studies.ids[1])
    print(rotl::get_publication(formica.study1.metadata))

    # And let's get the tree from this study
```

```
  formica.study1.tree1 <- get_study(formica.studies.ids[1])[[1]]

    # And plot it
    ape::plot.phylo(formica.study1.tree1, type="fan", cex=0.2)
}
```

Another question is where to get data. One important way is to collect your own: go out and measure seed size, count insect hairs, measure polar bear weight, etc. However, another way is to gather data already published by others. As with trees, it is important to *cite your sources*. People have put a lot of work into gathering data, and citation is the main way they get credit. It is also an essential way to reward people who choose to share data for others to build upon, correct, and check for reproducibility (some researchers still choose not to share data). Having citations also lets you and future scientists check for problems.

There are many places to get data. Perhaps the most convenient is to use rOpenSci's packages, which have interfaces to places like Barcode of Life, GBIF, Encyclopedia of Life, Neotoma paleoecological database, Fishbase, and much more. A major source for datasets and supplemental from particular papers is the Dryad site. For plant traits, the TRY database, which has many different datasets, can be useful, though policies on sharing can differ by dataset. Morphbank has biological images while Morphobank has images and other phenotypic data. Katherine Bannar-Martin has a list of databases for biological anthropology, mammals, fish, other vertebrates, plants, and more.

For homework, consider what biological question you're curious about and gather data for it. Importantly, **look at the data** before using it. Data are messy. For example, take latitude and longitude. Nice, continuous numbers: not "smooth vs hairy" but something you just read from your GPS – sure, there may be some imprecision, but it's not that bad, right? However, such data are full of errors: incorrect taxonomy, dropping signs or direction labels (140.5 W longitude is -140.5, not 140.5), entering 0,0 for missing data rather than leaving blank, recording the location of the collection where the specimen is housed rather than where it was collected, recording location of specimens under cultivation or in captivity (a polar bear in the San Diego Zoo), or just simple errors. Other kinds of data have their own problems: fish length – is it snout to end of tail fin, snout to vent? Is it for the biggest adult (beware indeterminate growth), average adult, whatever fish, adult or juvenile, was caught? For plant growth habit, how are woody vines counted: as woody (since they have woody structure) or herbaceous (not self supporting or tall)? Any time continuous variation is put into discrete bins, there are weird corner cases – after all, it's very rare for a trait to change instantly in one generation, even seemingly discrete ones like presence of eyes (what about cave fish?) or number of limbs (skinks? male boa constrictors with pelvic spurs?).

```
    # Get data from an external source
    # Load the data in
    # Plot the data, summarize the data, etc. to make sure there are no weird values.
```