An introduction to the `treebase` Package

Carl Boettiger[a,*], Duncan Temple Lang[b]

[a] *Center for Population Biology, University of California, Davis, United States*
[b] *Department of Statistics, University of California, Davis, United States*

Abstract

This paper describes the functionality provided by the software `treebase`, which provides an interface between the TreeBASE repository and the R programming language. We illustrate how the package can be used to search and retrieve phylogenetic trees from the repository, replicate existing studies that have deposited their phylogenies on TreeBASE, and perform automatically-updatable meta-analyses across the repostory.

Keywords: R, vignette, API, TreeBASE

1. Introduction

Applications that use phylogenetic information as part of their analyses are becoming increasingly central to both evolutionary and ecological research. The explosion of genomic data, followed by methodological advances for inferring phylogenies from this molecular information, have helped spur this increase. The R statistical environment has become the dominant platform for researchers performing applied phylogenetic methods (27 are listed on the taskview on Phylogenetic Methods of the Comprehensive R Archive Network) and new methods are added every year.

TreeBASE is an online repository of phylogenies [12]. With the advent of mandatory data archiving in many of the major journals publishing phylognetics-based research over the past year [3, 9, 14], such resources can expect to become increasingly prominent. TreeBASE provides an application programming interface, or API, that lets applications make queries to the database. We have implemented an R package that plugs into this interface to provide direct access to this data.

2. Examples & Results

The basic functions of the TreeBASE API allow search queries of the phylogenetic data in the repository (using the phylo-ws interface) and the metadata of publications associated with the phylogenies (using the OAI-MPH interface). These interfaces are well-documented on the TreeBASE website. The `treebase` package allows these queries to be made directly from R, just as a user would make them from the browser. The real advantage of the capacity to automate these tasks in R is shown in the later examples on replicating results and performing meta-analyses, but first, we introduce the package with some simple examples.

2.1. Basic Queries

Any of the basic queries avaiable on the web interface can now be made directly from R, including downloading and importing the phylogeny into the R interface. For instance, the following command searches for phylogenies containing dolphins, or all phylogenies submitted by a given author

```
search_treebase("\"Delphinus\"", by="taxon")
search_treebase("Huelsenbeck", by="author")
```

*Corresponding author.
Email address: `cboettig@ucdavis.edu` (Carl Boettiger)

```
dates <- sapply(metadata, function(x) as.numeric(x$date))
hist(dates, main="TreeBase growth", xlab="Year")
```

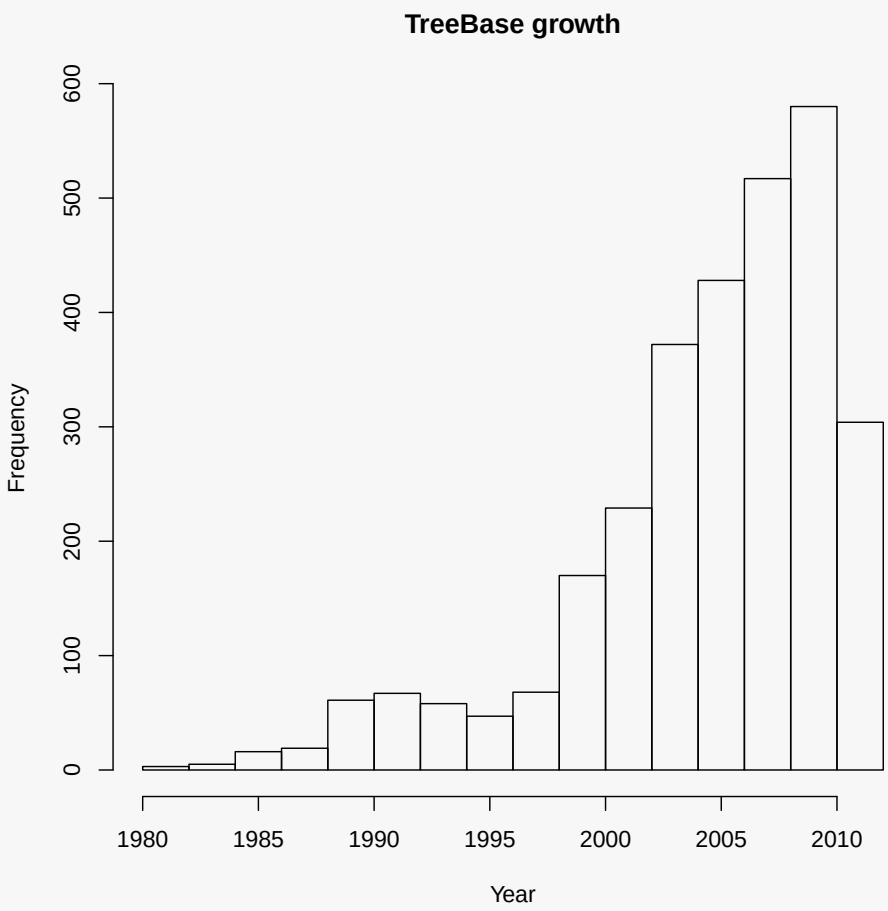

Figure 1: Histogram of publication dates by year, with the code required to generate the figure.

which loads the matching phylogenies into R, ready for analysis. The package and online documentation provide many examples of other possible queries. The package also provides access to the metadata of all publications containing trees deposited in TreeBASE. This can help the user discover phylogenies of interest and also allows the user to perform statistical analyses on the data deposition itself. While the metadata can be searched within a given date range, it is most straight-forward to import the full meta-data records first, with the command

```
metadata <- search_metadata("", by="all")
```

We can then look at the distribution of publication years for phylogenies depositied in TreeBASE and see if there is any trend, Fig 1. We can also look for patterns among publishers, such as comparing the number of submissions of two different publishers:

```
nature <- sapply(metadata, function(x) length(grep("Nature", x$publisher))>0)
science <- sapply(metadata, function(x) length(grep("^Science$", x$publisher))>0)
```

Summing the returned lists in each we find 15 phylogenies in *Nature* and 14 in *Science*. Many more such exercises are possible, and several examples are included in the demo files provided with the package.

2.2. Reproducible research & education

Reproducible research has become a topic of increasing concern in recent years [13, 4, 7]. Access to data and executable scripts that reproduce the results presented are two central elements of this process which are addressed by the `treebase` package.

```
derryberry <- search_treebase("Derryberry", "author")[[1]]
```

```
metadata(derryberry$S.id)
plot(derryberry)
```

For instance, we can replicate the model choice methods they perform to select the most likely branching model candidate.

```
require(laser)
tt <- branching.times(derryberry)
models <-  list(pb = pureBirth(tt), # yule model
                bdfit = bd(tt), # birth-death model
                y2r = yule2rate(tt), # yule model with single shift pt
                ddl = DDL(tt), # linear, diversity-dependent
                ddx = DDX(tt), #exponential diversity-dendent
                sv = fitSPVAR(tt), # vary speciation in time
                ev = fitEXVAR(tt), # vary extinction in time
                bv = fitBOTHVAR(tt) # vary both
                )
aics <- sapply(models, function(x) x$aic)
# name of the winning model
names(models[which.min(aics)])

## [1] "y2r"
```

In this fast-moving field, new methods often become available within the timeframe that another manuscript is submitted by its authors and the time at which if first appears in print. For instance, the more sophisticated methods avaialable in the more recent package, `TreePar`, were not used in this study.

We can easily re-analyze these results using the TreePar software,

```
    require(TreePar)
    x<-sort(getx(derryberry),decreasing=TRUE)
    # look for shifts occuring at times between 0 & 60 with interval of 5
    # consider yule models, allow up to 3 shifts (4 different rates)
    invisible(capture.output( # don't print messages
    yule_4rate <- bd.shifts.optim(x,sampling=c(1,1,1,1),
                                  grid=5,start=0,end=60,ME=FALSE, yule=TRUE)))
    # yule_4rate[[2]][[i]] contains the -loglik and parameters
    # for "best soln on grid" for the model with i shifts
    aic <- sapply(yule_4rate[[2]], function(x) 2 * (length(x) - 1) + 2 * x[1] )
    which.min(aic)

##
## 2
```

and confirm that the Yule 2-rate model is the best fitting by AIC score, even when allowing for up to four different shifts.

2.3. The self-updating meta-analysis

Meta-analyses are becoming increasingly common in phylogenetic methods [*e.g* 6, 8, 5, 11, 2]. Accessing many phylogenies can be difficult, and researchers have gone through heroic efforts to extract phylogenetic information from the literature, such as McPeek and Brown [6], who report that "All branch lengths were measured by hand from these printed sheets using dial calipers" on 163 phylogenies printed out from pdfs in which they appeared in the literature. Researchers may focus their meta-analysis on particular taxa [8], leaving us to wonder if the conclusions hold in other groups. Other studies may use a single empirical tree accompanied by a a collection of simulated phylogenies Cusimano and Renner [1]. Some already turn to TreeBASE to provide a more extensive collection of phylogenies [2]. The analysis in many of these studies already uses or can use methods available in R.

By building such meta-analyses around the `treebase` package, one can not only take advantage of the existing data with substantially less effort, but also provide a script that can be automatically updated as more phylogenies are deposited in the TreeBASE repository.

```
all_trees <- search_treebase("Consensus", "type.tree", branch_lengths=TRUE)
```

For convience and testing purposes, we could just load the cached set of all phylogenetic trees with branch-lengths in TreeBASE at the time the package was built:

```
data(branchlengths)
```

Many meta-analyses require ultrametric trees (branch lengths proportional to time, rather than to mutational steps). This function is just an elementary example to illustrate the process of time-calibrating a tree; more sophisticated methods could be chosen instead.

```
timetree <- function(tree){
    check.na <- try(sum(is.na(tree$edge.length))>0)
    if(is(check.na, "try-error") | check.na)
      NULL
    else
    try( chronoMPL(multi2di(tree)) )
}
tt <- drop_errors(sapply(branchlengths, timetree))

## [1] "dropped 11 trees"
```

The central question of many of meta-analyses mentioned above has been whether or not phylogenies show a changing rates of evolution. A standard test of this is the γ statistic of Pybus and Harvey [10] which tests the null hypothesis that the rates of speciation and extinction are constant. Applying this test to all of the currently available trees in TreeBASE,

```
# gamma statistic
gammas <- sapply(tt, function(phy)  gammaStat(phy))
# associated p-value
p_gammas <- sapply(gammas, function(x)  2*(1-pnorm(abs(x))))
non_const <- sum(p_gammas < 0.025)/length(gammas)
```

we find that 54% of the trees can reject the constant-rates model at the 95% confidence level. Because `treebase` makes it possible to perform this analysis entirely by scripts using the latest treebase data, it is not only easier to perform this analysis but also to update it to reflect the latest data. For instance, this paper is written using R's Sweave tool, where the results and figures are generated on the fly as the paper is compiled. Consequently the analyses presented here can be updated to reflect the latest information in TreeBASE by the click of a button.

3. Acknowledgements

CB wishes to thank the TreeBASE developer team for building and supporting the repository, and all contributers to TreeBASE. CB is supported by a Computational Sciences Graduate Fellowship from the Department of Energy under grant number DE-FG02-97ER25308.

[1] Cusimano, N., Renner, S. S., Jun. 2010. Slowdowns in Diversification Rates from Real Phylogenies May Not be Real. Systematic Biology 59 (4), 458–464.
URL http://sysbio.oxfordjournals.org/cgi/doi/10.1093/sysbio/syq032

[2] Davies, T. J., Allen, A. P., Borda-de Água, L., Regetz, J., Melián, C. J., Jul. 2011. NEUTRAL BIODIVERSITY THEORY CAN EXPLAIN THE IMBALANCE OF PHYLOGENETIC TREES BUT NOT THE TEMPO OF THEIR DIVERSIFICATION. Evolution; international journal of organic evolution 65 (7), 1841–1850.
URL http://doi.wiley.com/10.1111/j.1558-5646.2011.01265.xhttp://www.ncbi.nlm.nih.gov/pubmed/21729042

[3] Fairbairn, D. J., Nov. 2010. the Advent of Mandatory Data Archiving. Evolution; international journal of organic evolution, no–no.
URL http://doi.wiley.com/10.1111/j.1558-5646.2010.01182.x

[4] Keiding, N., Jul. 2010. Reproducible research and the substantive context. Biostatistics (Oxford, England) 11 (3), 376–8.
URL http://www.ncbi.nlm.nih.gov/pubmed/20498225

[5] McPeek, M. a., Dec. 2008. The ecological dynamics of clade diversification and community assembly. The American naturalist 172 (6), E270–84.
URL http://www.ncbi.nlm.nih.gov/pubmed/18851684

[6] McPeek, M. a., Brown, J. M., Apr. 2007. Clade age and not diversification rate explains species richness among animal taxa. The American naturalist 169 (4), E97–106.
URL http://www.ncbi.nlm.nih.gov/pubmed/17427118

[7] Peng, R. D., Dec. 2011. Reproducible Research in Computational Science. Science 334 (6060), 1226–1227.
URL http://www.sciencemag.org/cgi/doi/10.1126/science.1213847

[8] Phillimore, A. B., Price, T. D., Mar. 2008. Density-dependent cladogenesis in birds. PLoS biology 6 (3), e71.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2270327&tool=pmcentrez&rendertype=abstract

[9] Piwowar, H. A., Vision, T. J., Whitlock, M. C., May 2011. Data archiving is a good investment. Nature 473 (7347), 285–285.
URL http://www.nature.com/doifinder/10.1038/473285a

[10] Pybus, O. G., Harvey, P. H., Nov. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proceedings. Biological sciences / The Royal Society 267 (1459), 2267–72.
URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1690817&tool=pmcentrez&rendertype=abstract

[11] Quental, T. B., Marshall, C. R., Jun. 2010. Diversity dynamics: molecular phylogenies need the fossil record. Trends in Ecology & Evolution, 1–8.
URL http://linkinghub.elsevier.com/retrieve/pii/S0169534710001011

[12] Sanderson, M. J., Donoghue, M. J., Piel, W., Eriksson, T., 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of Botany 81 (6), 183.

[13] Schwab, M., Karrenbach, N., Claerbout, J., 2000. Making scientific computations reproducible. Computing in Science & Engineering 2 (6), 61–67.
URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=881708

[14] Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L., Moore, A. J., Mar. 2010. Data archiving. The American Naturalist 175 (2), 145–6.
URL http://www.ncbi.nlm.nih.gov/pubmed/20073990