The Journal of Open Source Education

treesiftr: An R package and server for viewing phylogenetic trees and data

April M Wright[1]

1 Southeastern Louisiana University

DOI: 10.21105/jose.00035

Software

- Review
- Repository
- Archive

Submitted: 31 August 2018
Published: 17 January 2019

Summary

treesiftr is a Shiny (Chang, Cheng, Allaire, Xie, & McPherson, 2018) application for visualizing the relationship between phylogenetic trees and the underlying data used to estimate them. It can also be used in RStudio (RStudio Team, 2015) or at the command line as an R package (R Core Team, 2013). treesiftr works by subsetting a phylogenetic matrix according to user-provided input about which characters to visualize. A maximum parsimony tree is then estimated from each data subset. Maximum parsimony was chosen for speed and analytical simplicity. Under the parsimony optimality criterion, the preferred tree is the one that suggests the fewest evolutionary steps, or character changes over evolutionary history. The tree is scored under both parsimony and Lewis' Mk model (Lewis, 2001), a maximum likelihood model for estimating phylogeny from discrete character data. The data and tree are then visualized using ggtree (Yu, Smith, Zhu, Guan, & Lam, 2017), based upon the ggplot2 (Wickham, 2016) package. Expected outputs are the same whether the learner is interacting via the GUI or the RStudio interface; however, the RStudio interface does have additional options not available in the GUI.

The included Shiny application renders the visual output of the subsetting and estimation process, and can be used to provide further input to the treesiftr functions, such as decorating the trees with scores under different optimality criteria. It can be run locally or used via the web without installing any R packages or having knowledge of R. The web instance can accommodate 150 concurrent users.

Also included in the package are two worksheets and an instructor guide. The worksheet intended for use with the Shiny app, "treesiftr GUI", introduces the application, the underlying data, and the application functions. It includes several questions, and a glossary. The worksheet for the RStudio interface, "treesiftr Advanced", asks the same questions, but also emphasizes general R skills, such as subsetting data, and specific R skills, such as interacting with phylogenetic data. And instructor guide contains answers to the questions, as well as renderings of the outputs students should see. treesiftr is not intended to replace a lecture on phylogeny, but to supplement student understanding.

Phylogenetic trees represent the evolutionary relationships between a set of taxa. These taxa can be species, higher-level biological groupings (e.g., genera or families), lower-level groupings (e.g., populations or strains) or even individuals. Estimating phylogenetic trees is crucial in many areas of biology. Phylogenetic trees are built from homologous characters, characters which are similar in multiple taxa because they were inherited from the most recent common ancestor (Darwin, 1859). For morphological traits, homology is assigned by experts by evaluating the character's evolutionary history and ontology. In molecular characters, it is typically assigned via multiple sequence alignment (Feng & Doolittle, 1987).

Once homology is assigned, a phylogeny is estimated from the data. Several optimality criteria may be used to do this. They broadly fall into parametric methods, such

as maximum likelihood (Felsenstein, 1973) and Bayesian methods (Huelsenbeck & Ronquist, 2001), and non-parametric methods, such as maximum parsimony (Farris, Kluge, & Eckardt, 1970) and neighbor-joining (Gower & Ross, 1969). Parametric methods assume a model of underlying character evolution, while non-paramtric methods do not. Parametric methods have been shown in many instances to be more accurate (Felsenstein, 1978) and, as they enable a rigorous framework of model testing, should be preferred over non-parametric methods. However, they are also compute-intensive, and so treesiftr makes use of the maximum parsimony criterion for analytical simplicity. Under maximum parsimony, the tree that is favored for a set of taxa is the one which implies the fewest evolutionary changes.

Statement of Need

Understanding phylogenetic trees is challenging for students (Baum, Smith, & Donovan, 2005, Meisel (2010)). Some of this challenge comes from inherently misunderstanding evolution in pre-Darwinian terms, such as a Platonic 'great chain of being' or as being ladder-like (Rudolph & Stewart, 1998, Sandvik (2008)). This is often inherently coupled to the idea of humans being the pinnacle of evolution (O'Hara, 1997). treesfitr, therefore, by default uses a non-human dataset to decouple tree-thinking from preconceived notions of organismal hierarchy. Students also have difficulty simply reading phylogenetic trees - often reading along the tips of a tree, as opposed to looking at the internal nodes (Baum et al., 2005, Meisel (2010)). To counter this misconception, treesiftr continuously estimates new phylogenies, moving both tips and nodes. This causes students to have to continually re-evaluate the tree and the information on it (Meir & Kingsolver, 2007).

There are many phylogenetic tree viewers on the market such as FigTree, IcyTree (Vaughan, 2017), and Phylogeny.IO (Jovanovic & Mikheyev, 2016). But visualizing the data that underlie a particular tree is still largely accomplished via the Mesquite software (W. P. Maddison & Maddison, 2008). To use Mesquite, one must perform local installs. The software must also be interacted with via a GUI, without the opportunity for learners to practice programmatic skills while learning about phylogenetic trees and data. I wrote this package to allow students to learn about phylogenetic trees and data without performing local installs, and to give the option for students to practice programmatic skills while doing so.

treesiftr was initially written for use in the Analytical Paleobiology Workshop in summer 2018. This course is an intensive 30-day paleobiological workshop predominantly for graduate students. treesiftr was used in the last week of the course, by which time the learners had been working with R via the RStudio application for three weeks. In this setting, the exercise itself took about 45 minutes, and was embedded in a 3-hour lecture block (see below). I also added a web-based GUI for portability, and use in undergraduate biology classrooms. In undergraduate biology classrooms, learners are largely naive with respect to scientific computing, and performing installations on school or personal computers for a single class period may not be feasible. In particular, I am faculty at Southeastern Louisiana University, which is an institution serving students in a low-income region of a low-income state. Many students do not have reasonable computers to perform local installs of software. For demonstrating phylogeny and evolutionary history in my genetics class, a web-based viewer and activity set is, therefore, preferable. The discussion of phylogeny is one hour and fifteen minute class period, with the hands-on activity taking about 30 minutes, preceded by lecture on the basics of tree thinking (Baum et al., 2005) and tree reading.

The included worksheets and package are not intended to be a replacement for a lecture. They are instead intended to be a hands-on supplement for the lecturer to use in class. By allowing the student to choose subsets of data, and have a tree of those data appear instantly, the relationship between the data and the estimated tree is enforced

visually. Each worksheet does come with a glossary of terms that are required to describe phylogenetic trees, and it is the responsibility of the instructor to help students learn these terms before they start the activity. There is one example slideshow included in the `inst/slides` directory. These slides are written in RMarkdown, with executable R code segments that illustrate skills such as interacting with phylogenetic trees in R, and terminology to describe phylogenies. An overview of different methodologies in phylogeny estimation is also provided. This is an example of a lecture that could be used to explain the terminology presented in the worksheets. It is not necessary to use these particular slides for the activity to function.

References

Baum, D. A., Smith, S. D., & Donovan, S. S. S. (2005). The tree-thinking challenge. *Science*, *310*(5750), 979–980. doi:10.1126/science.1117727

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). *Shiny: Web application framework for R*. Retrieved from https://CRAN.R-project.org/package=shiny

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. Murray, London.

Farris, J. S., Kluge, A. G., & Eckardt, M. J. (1970). A numerical approach to phylogenetic systematics. *Systematic Biology*, *19*(2), 172–189. doi:10.2307/2412452

Felsenstein, J. (1973). Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *Systematic Zoology*, *22*(3), 240–249. doi:10.1093/sysbio/22.3.240

Felsenstein, J. (1978). The number of evolutionary trees. *Systematic Zoology*, *27*(1), 27–33. doi:10.2307/2412810

Feng, D., & Doolittle, R. F. (1987). Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *Journal of Molecular Evolution*, *25*(4), 351–360. doi:10.1007/BF02603120

Gower, J. C., & Ross, G. J. (1969). Minimum spanning trees and single linkage cluster analysis. *Applied Statistics*, 54–64. doi:10.2307/2346439

Huelsenbeck, J., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, *17*(8), 754–755. doi:10.1093/bioinformatics/17.8.754

Jovanovic, N., & Mikheyev, A. S. (2016). Interactive web-based visualization of phylogenetic trees using phylogeny.io. *PeerJ Preprints*, *4*, e2579v1. doi:10.7287/peerj.preprints.2579v1

Lewis, P. O. (2001). A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *sysbio*, *50*(6), 913–925. doi:10.1080/106351501753462876

Maddison, W. P., & Maddison, D. R. (2008). Mesquite: A modular system for evolutionary analysis. Version 2.5. http://mesquiteproject.org. Retrieved from http://mesquiteproject.org

Meir, J., E. Perry, & Kingsolver, J. (2007). College students' misconceptions about evolutionary trees. *The American Biology Teacher*, *69*. doi:10.1662/0002-7685(2007)69[71:csmaet]2.0.co;2

Meisel, R. P. (2010). Teaching tree-thinking to undergraduate biology students. *Evolution: Education and Outreach*, *3*(4), 621–628. doi:10.1007/s12052-010-0254-9

O'Hara, R. J. (1997). Population thinking and tree thinking in systematics. *Zoologica Scripta*, *26*(4), 323–329. doi:10.1111/j.1463-6409.1997.tb00422.x

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

RStudio Team. (2015). *RStudio: Integrated development environment for R*. Boston, MA: RStudio, Inc. Retrieved from http://www.rstudio.com/

Rudolph, J. L., & Stewart, J. (1998). Evolution and the nature of science: On the historical discord and its implications for education. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *35*(10), 1069–1089. doi:10.1002/(SICI)1098-2736(199812)35:10<1069::AID-TEA2>3.0.CO;2-A

Sandvik, H. (2008). Tree thinking cannot taken for granted: Challenges for teaching phylogenetics. *Theory in Biosciences*, *127*(1), 45–51. doi:10.1007/s12064-008-0022-3

Vaughan, T. G. (2017). IcyTree: Rapid browser-based visualization for phlogenetic trees and networks. *Bioinformatics*, 315–328. doi:10.1093/bioinformatics/btx155

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Yu, G., Smith, D., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. doi:10.1111/2041-210X.12628