

# **tREeBOOT: Assessing phylogenetic uncertainty in supertrees**

Stephen A. Smith and Brian R. Moore\*

*Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA*

R.H. tREeBOOT

*Please address editorial correspondence to:*

Brian R. Moore  
Dept. Ecol. & Evol. Biol.  
Yale University  
New Haven CT 06405  
USA

Phone: 203-432-7168  
FAX: 203-432-2374  
E-mail: [brian.moore@yale.edu](mailto:brian.moore@yale.edu)

## ABSTRACT

**Summary:** tREeBOOT implements methods for assessing uncertainty in phylogenetic supertrees, including source-tree bootstrapping and hierarchical bootstrapping. These algorithms can exploit either bootstrap profiles or posterior probability densities of source trees to provide estimates of bootstrap proportions or posterior probabilities of nodes in the supertree, respectively. The ability to generate supertree bootstrap profiles should increase the utility of these composite phylogenies for evolutionary studies. tREeBOOT enables source-tree and hierarchical bootstrapping algorithms to be used with MRP, MinCut, and modified MinCut supertree estimation methods.

**Availability:** <http://www.blackrim.org/programs/treeboot.html>

**Contact:** stephen.smith@yale.edu or brian.moore@yale.edu

The estimation of ever larger phylogenies demands consideration of alternative inference strategies, including divide-and-conquer approaches that decompose the global inference problem to a set of smaller, more manageable component problems. A prominent focus of research in this area is the development of supertree methods (e.g., Bininda-Emonds, 2004), which infer a composite tree by combining a set of partially overlapping component phylogenies (i.e., source trees). Although promising, supertree estimation has been plagued by two stubborn problems which greatly diminish the utility of supertrees for systematic and evolutionary studies. First, supertrees tend to be dissociated from the primary character data. Conventional phylogenetic analysis provides a topological summary of typically complex patterns of character covariation within the corresponding data matrix. These topological summaries, in turn, provide the raw data for supertree methods, which effectively prevent sub-signals within the component data sets from contributing to the supertree estimate. Consequently, combining a set of source-tree topologies may yield a supertree that is not sanctioned by a simultaneous analysis of the component data matrices (e.g., Gatesy et al., 2002). Second, it is unclear how best to assess uncertainty in supertree estimates (e.g., Wilkinson et al., 2005). Solutions to these problems have recently been pursued via methods that variously incorporate nonparametric bootstrapping into the supertree estimation procedure (Moore et al., 2006).

The first method, *source-tree bootstrapping*, involves repeatedly drawing a random sample of source trees (with replacement) from the set of optimal source-tree estimates. Each bootstrap set of source trees is then combined under the preferred supertree algorithm and the resulting supertree estimate(s) appended to a file. This procedure is repeated an arbitrary number of times to generate a bootstrap profile of supertrees, which is summarized using majority-rule consensus. The frequencies of nodes in the consensus supertree provide an estimate of the degree of congruence among the original set of source-tree topologies that is analogous to standard bootstrap proportions.

Rather than drawing from point estimates of source trees, *hierarchical bootstrapping* samples from a set of source-tree distributions (i.e., the set of source-tree bootstrap profiles generated by conventional bootstrap analysis, or the set of source-tree posterior probability densities estimated by Bayesian analysis). Hierarchical bootstrapping entails repeatedly drawing a random set of source trees (with replacement) from their respective distributions. If based on posterior probability densities, the contribution of each source tree is scaled by its posterior probability; alternatively, source trees sampled from the set of bootstrap profiles contribute equally to the supertree estimate. The bootstrap sample of source trees is then combined using the preferred supertree algorithm and the resulting supertree estimate(s) appended to a file. This procedure is repeated an arbitrary number of times to generate a bootstrap profile of supertrees, which is summarized by majority-rule consensus. The frequencies of nodes in the consensus supertree provide an estimate of the underlying character support that is analogous either to standard bootstrap proportions or posterior probabilities (if based on source-tree bootstrap profiles or posterior probability densities, respectively).

Experiments with these methods indicate that source-tree and hierarchical bootstrapping may provide a means for both improving the accuracy of supertree estimation and for assessing the uncertainty in supertrees (Burleigh et al., 2006; Moore et al., 2006). The ability to gauge uncertainty in supertree estimates should greatly enhance their appeal in evolutionary studies; e.g., of character evolution, historical biogeography, coevolution, diversification rates. For example, estimates of diversification rates could be evaluated over the bootstrap profile of supertrees to accommodate phylogenetic uncertainty on the inference at hand. Furthermore, because it effectively assesses conflict among a set of topologies, source-tree bootstrapping may also find useful for application to other problems, such as evaluating the level of conflict in the gene-tree/species-tree problem (e.g., Burleigh et al., 2006).

The source-tree and hierarchical bootstrapping algorithms are implemented in tREeBOOT, a freely available command-line Java application (that will run on any operating system supported by Java Runtime Environment, including Macintosh OS X, Windows, Unix, and most Linux distributions). The bootstrapping algorithms are inherently modular procedures that could be interpolated with any supertree algorithm. Currently, tREeBOOT allows these algorithms to be used with the MRP supertree algorithm (Baum, 1992; Ragan, 1992) implemented in PAUP\* (Swofford, 2000), and the MinCut and modified MinCut supertree algorithms implemented in the program supertree (Page, 2002). tREeBOOT also implements *source-tree nodal weighting*, in which the contribution of each source-tree node to the supertree estimate is proportionate to its bootstrap proportion or posterior probability. Input and output files use the standard NEXUS format (Maddison et al., 1997). A user manual, command reference, and example data sets are included with the program distribution.

## REFERENCES

- Baum, B.R. (1992) Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, **41**, 3–10.
- Bininda-Emonds, O.R.P. (2004) The evolution of supertrees. *Trends Ecol. Evol.*, **19**, 315–322.
- Burleigh, J.G., Driskell, A. and Sanderson, M.J. (2006) Supertree bootstrapping methods for assessing phylogenetic variation among genes. *Syst. Biol.*, **55**, ###–###.
- Gatesy, J., Matthee, C., DeSalle, R. and Hayashi, C. (2002) Resolution of a supertree/supermatrix paradox. *Syst. Biol.*, **51**, 652–664.
- Maddison, D.R., Swofford, D.L. and Maddison, W.P. (1997) NEXUS: an extensible file format for systematic information. *Syst. Biol.*, **46**, 590–621.
- Moore, B.R., Smith S.A. and Donoghue, M.J. (2006) Increasing data transparency and estimating phylogenetic uncertainty in supertrees: Approaches using nonparametric bootstrapping. *Syst. Biol.*, **55**, ###–###.
- Page, R.D.M. (2002) Modified MinCut supertrees. In Guigó R. and Gusfield D. eds. *WABI 2002*, LNCS, 2452:537–551.
- Ragan, M.A. (1992) Phylogenetic inference based on matrix representation of trees. *Mol. Phylogenet. Evol.*, **1**, 53–58.
- Semple, C. and Steel, M. (2000) A supertree method for rooted trees. *Discrete Appl. Math.*, **105**, 147–158.
- Swofford, D.L. (2000) PAUP\*: Phylogenetic analysis using parsimony (\*and other methods), version 4.0. Sinauer, Sunderland, Massachusetts.
- Wilkinson, M., Pisani, D., Cotton, J.A. and Corfe, I. (2005) Measuring support and finding unsupported relationships in supertrees. *Syst. Biol.*, **54**, 823–831.