Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2013, **4,** 688–691 doi: 10.1111/2041-210X.12051

APPLICATION

Congruification: support for time scaling large phylogenetic trees

Jonathan M. Eastman[1,2]*, Luke J. Harmon[1,2] and David C. Tank[2,3]

[1]Department of Biological Sciences, University of Idaho, Moscow, 83844, Russia; [2]Institute for Bioinformatics and Evolutionary Studies (IBEST), University of Idaho, Moscow, 83844, Russia; and [3]College of Natural Resources & Stillinger Herbarium, University of Idaho, Moscow, 83844, Russia

Summary

1. Approaches for efficient statistical estimation of large phylogenies are now available (*Bioinformatics*, 2006, 22, 2688), and yet we lack adequate tools for synthesizing information from previous analyses into large timetrees. Here, we present a cross-platform R tool that integrates with tree of life efforts by mapping divergence times from an existing timetree (a '*reference*') to another uncalibrated phylogeny (a '*target*') that samples from the same lineage. Leveraging existing methods for rate-smoothing phylograms, this tool enables the rapid generation of very large timetrees where direct estimation of the timing of lineage diversification is either impracticable or impossible.

2. The primary output of the tool is to return divergence times for nodes resolved as concordant between the *reference* and *target*. Given the computed set of secondary calibrations, *post hoc* tree transformation can be accomplished using existing resources that assume either a strict or relaxed evolutionary clock.

3. Our software is provided open source in the GEIGER package (http://cran.r-project.org/package = geiger) and is thoroughly demonstrated in the Supporting Information.

Key-words: divergence time, GEIGER, phylogenetics, time scaling, tree of life

Introduction

With the continued development of supertree, supermatrix and megaphylogeny approaches (Smith, Beaulieu & Donoghue 2009), phylogeny estimates of large size are becoming increasingly possible (e.g. Stamatakis 2006; Smith & Donoghue 2008; Smith *et al.* 2011; Hinchliff & Roalson 2013). While these methods enable estimation of phylogenetic structure across a broad swath of biodiversity, lagging in development are explicit model-based methods for the coestimation of topology and divergence times for very large trees (here defined as involving thousands of taxa) and data matrices with large amounts of missing data (i.e. supermatrices). Direct inference of the temporal component of evolution is often missing in these large-scale analyses, where it is possible only to estimate branch lengths in terms of the product *rt*, where *r* is molecular evolutionary rate and *t* is elapsed time. Moreover, it is often difficult to synthesize branch length information from trees estimated for the same lineage, but where the sets of sampled taxa are inconsistent. This issue is especially glaring where trees sample at different levels of a systematic hierarchy (e.g. some are species level, others are exemplar phylogenies of families, genera, etc.).

We use an example from amphibian phylogenetics to illustrate certain difficulties faced by empiricists attempting to estimate credible timetrees and to a possible resolution. A large amphibian phylogeny was published by Pyron & Wiens (2011), for which the underlying data set samples over 2800 species. The phylogeny estimate is clearly unprecedented in sampling density for amphibian molecular phylogenetics, yet the tree is also unscaled to absolute time. This lack of temporal information is limiting in many ways. For instance, one cannot calculate or compare absolute diversification rates of amphibian traits or lineages. For the same lineage, Roelants *et al.* (2007) estimated a timetree (i.e. a phylogenetic tree with branch lengths calibrated to units of time elapsed) with relatively sparse sampling across the vast majority of extant amphibian families. It would seem desirable to use information contained within the Roelants *et al.* (2007) tree to scale the larger, more densely sampled tree of Pyron & Wiens (2011) to absolute time. This would be an overly arduous task without a tool we present and dub 'congruification'. In the Supporting Information, we illustrate our solution and demonstrate congruification in scaling the Pyron & Wiens (2011) tree to time.

Description

Estimating phylogeny for which branch lengths are in units of time is currently a very difficult computational problem for large trees and sparse data matrices. Some of the most sophisticated methods available for coestimation of topology and

*Correspondence author. E-mail: jonathan.eastman@gmail.com time-scaled branch lengths (e.g. BEAST, Drummond *et al.* 2012; MRBAYES, Ronquist & Huelsenbeck 2003) are simply not practical for data sets with thousands of taxa, yet there is tremendous demand for such timetrees. We assume that reasonable time scalings of a phylogram can be achieved by exploiting information contained in existing time-calibrated trees that sample from the same lineage.

We introduce 'congruification' as a novel tool for automating the resolution of all possible secondary calibrations of a tree, requiring at minimum two phylogenetic trees: a *reference* (scaled in units of absolute time) and a *target* (yet unscaled to time, but whose branch lengths are in units of the expected number of character-state changes). Unless the *target* and *reference* correspond perfectly in the particular taxa sampled, congruification also requires an expert-curated linkage table that links tip labels found in the *reference* to those found in the *target* (Fig. 1). This allows concordant nodes to be resolved even where 'levels' of sampling differ between the *reference* (e.g. family level) and *target* (e.g. species level). Indeed, if we are willing to assert the monophyly of lineages exemplified in the *reference* tree, we can time-scale a target even where there is no direct overlap in the sampled tips of the two trees. In order to exploit the *reference* to the fullest extent, the provided linkage table should be consistent with shared ancestry among tips of both trees (e.g. if the linkage table contains taxonomic names, these taxa should be monophyletic).

Algorithm

The following describes the algorithm used to congruify a *target* tree. The algorithm is available from the R-package, GEIGER.

1. Congruification maps leave from the *target* tree to individual samples in the *reference,* based on a linkage table. The linkage table used in Fig. 1 would map samples '*J1*' and '*J2*' from the *target* (Fig. 1a) to the representative tip '*J*' in the *reference* (Fig. 1b), making the implicit assumption that lineage '*J*' is monophyletic. A linkage table (see Fig. 1 caption) is unnecessary if all tips occurring in the *target* also occur in the *reference,* and mappings are thereby one to one.

2. The algorithm determines the set of tips (T_I) that are sampled in the *target* and are present as mapped tips in the *reference*. Tips that occur in only the *target* or *reference* are excluded from T_I. The subset of T_I that descends from each

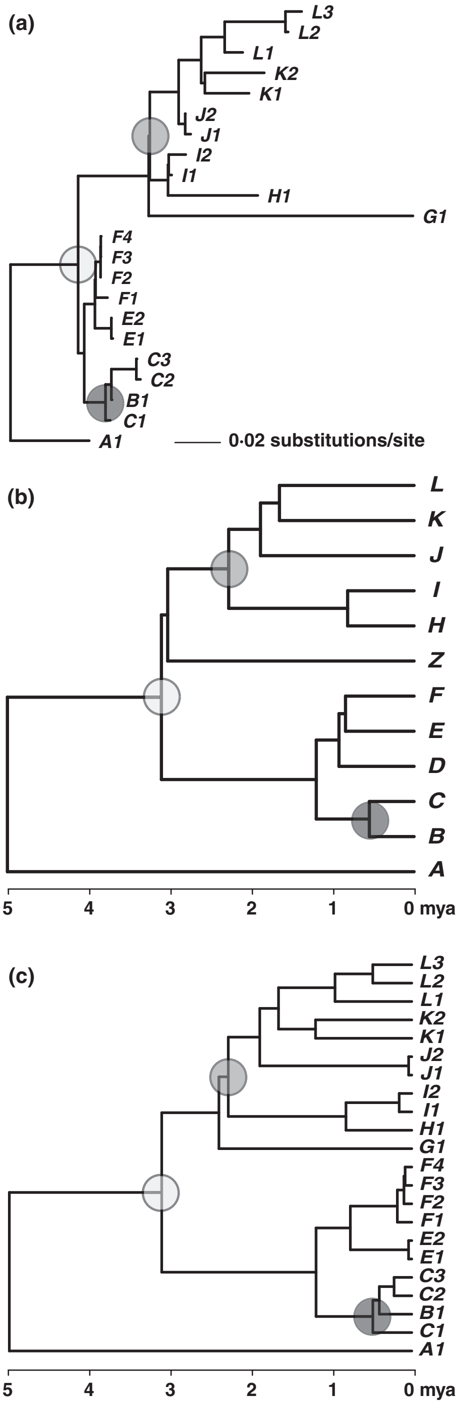

Fig. 1. Illustration of the time scaling, using a *target* tree (panel a) and a *reference* tree (panel b); panel c shows a 'congruified' *target* tree. Tip labels in the middle panel represent exemplars for 12 lineages (e.g. genera). Tip labels in upper- and lower-most panels represent samples (e.g. species) from within many (but in this case not all) of the lineages shown in b. Circles denote three nodes (of several others) that are concordant between the *target* and *reference*. All members of *H, I, J, K* and *L* are subtended by the light grey circle in all trees. Note that despite the incompletely overlapping sets of tips between the *reference* and *target* (i.e. species from lineage *Z* and lineage *G* occur in only the *reference* and *target*, respectively), concordant splits can still be identified (white circle). Furthermore, despite nonmonophyly of lineage *C* in the *target* (panels a and c), the dark grey circle demarcates a concordant split between *reference* and *target* where all members of the *B* and *C* lineages are subtended. Panel c shows the ultrametricized *target*, informed by node heights for concordant splits from the *reference*. The look-up table associated with this instance of congruification might appear as follows:

tip	lineage
A1	A
B1	B
C1	C
C2	C
...	...
L2	L
L3	L

node in the tree is used to erect a binary ‘membership’ vector at that node, the entries of which denote the presence or absence of each species in T_I among the node’s descendants. A membership vector is constructed for each node in the *target* and the *reference*. Only exactly matching vectors are accepted as identifying concordant nodes between the *target* and *reference*. If nonunique membership vectors exist in a given tree (i.e. multiple nodes have an identical array of descendant tips found in T_I), precedence is given to the most tipward node for each set of nonunique vectors.

3. Temporal information is mapped from *reference* to *target* using the resolved set of concordant nodes. Divergence times of nodes in the *reference* are extracted to serve as calibration points for the concordant nodes in the *target*. A node in the *target* that does not have a perfect match in the *reference* is left uncalibrated.

4. The *target* is scaled to absolute time using a separate software using the set of calibrations extracted from the *reference*. The set of calibrations from concordant nodes between the trees is returned to the user in tabular format. This output can be easily manipulated for time scaling using any of several existing algorithms: these include R8S (Sanderson 2002), PATHD8 (Britton *et al.* 2007), and TREEPL (Smith & O’Meara 2012).

Example

In the Supporting Information, we demonstrate a complete example of congruification using the time-calibrated Roelants *et al.* (2007) *reference* and the uncalibrated Pyron & Wiens (2011) *target* amphibian trees. Though similar in phylogenetic breadth, these trees differ markedly in the density of sampling (152 vs. 2871 taxa sampled, respectively). We use several accessory functions to resample the *reference* as an exemplar tree, and we use the NCBI taxonomy database to provide a linkage table between samples in the *target* and *reference*. Final time scaling of the *target* is achieved using the recently introduced tool for rate smoothing, TREEPL (Smith & O’Meara 2012).

Discussion

We note that congruification makes no attempt at reconciliation of topological discordance between the *reference* and *target*. Indeed, the tool only serves to scale the *target* to be as temporally consistent with the *reference* as possible (Fig. 1). While this implementation provides an automated means of resolving secondary calibrations, certain aspects of congruification are less than ideal. Graur & Martin (2004) raise the issue of extra-primary or indirect calibration, wherein molecular-based estimates of divergence time are used in secondary analyses as a calibration *point* (i.e. treating the divergence date as an error-free estimator of the unknown parameter value). The use of congruification is not far removed from such a practice. In cases where direct use of independent calibrations (e.g. from the fossil record) is currently impracticable, we argue that congruification nevertheless has the ability to provide reasonable estimates of divergence timing in the *target*. Much as Graur & Martin (2004) caution, the uncertainty associated with any estimator must not only be acknowledged but should be represented in every *post hoc* analysis. We thus strongly recommend time scaling the *target* using many independent samples in proportion to a target distribution (e.g. a posterior distribution of trees) for both the *reference* and *target*. We caution that the presence of rogue taxa or erroneous placement of taxa in either the *reference* or *target* will prevent a full exploitation of dates in the *reference*. Certainly, the greater the topological consistency between *reference* and *target*, the greater the amount of temporal information that can be mapped from *reference* to *target* and exploited for time scaling.

To maximize correspondence between the *target* and *reference*, it may be preferable to prune the *reference* tree to leave a single representative for broader monophyletic lineages. Tree correspondence will be maximal when all the tips in the *reference* can be matched either to the set of tips present in the *target* or to groups defined by the look-up table. Pruning the *reference* to exemplars will be especially desirable where sampling differs extensively between the two trees (e.g. the same genus is sampled in both the *target* and *reference* but the species sampled are nonoverlapping). We provide a function, *exemplar.phylo( )*, that automates this procedure if the *reference* is supplied with a taxonomy for the sampled lineages. The default behaviour of this function is to use the NCBI taxonomic database. As a simple extension of the congruification algorithm, we provide a convenient tool – *nodelabel.phylo( )* – by which to label internal nodes of phylogenies to correspond with groups defined by a given taxonomic resource. Several functions related to manipulation and use of taxonomic data, including those just described, are demonstrated in the Supporting Information.

Conclusion

By integrating vetted and time-scaled inferences of phylogeny, careful systematic treatments and newly generated estimates of phylogeny, congruification will allow users to rapidly generate large timetrees, a central facet in the study of pattern and process in the diversification of life.

Acknowledgements

We thank M.E. Alfaro, J.W. Beaulieu, J.W. Brown, B.C. O’Meara, M.W. Pennell, K. Roelants, G.J. Slater, S.A. Smith and A.E. Zanne for support provided during development of this project. Two anonymous reviewers, G. Thomas and E. Paradis, provided constructive critiques of the manuscript in its earlier form. Financial support to LJH was provided by NSF DEB 0919499 and DEB 1208912; and the NESCent working group, *Tempo and mode of plant trait evolution* (NSF #EF- 0905606), provided support for JME and DCT.

References

Britton, T., Anderson, C.L., Jacquet, D., Lundqvist, S. & Bremer, K. (2007) Estimating divergence times in large phylogenetic trees. *Systematic Biology*, **56**, 741–752.

Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, **29**, 1969–1973.

Graur, D. & Martin, W. (2004) Reading the entrails of chickens: molecular timescales of evolution and the illusion of precision. *Trends in Genetics*, **20**, 80–86.

Hinchliff, C.E. & Roalson, E.H. (2013) Using supermatrices for phylogenetic inquiry: an example using the sedges. *Systematic Biology*, **62**, 205–219.

Pyron, R.A. & Wiens, J.J. (2011) A large-scale phylogeny of Amphibia including over 2,800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Molecular Phylogenetics and Evolution*, **61**, 543–583.

Roelants, K., Gower, D.J., Wilkinson, M., Loader, S.P., Biju, S.D., Guillaume, K., Moriau, L. & Bossuyt, F. (2007) Global patterns of diversification in the history of modern amphibians. *Proceedings of the National Academy of Sciences*, **104**, 887–892.

Ronquist, F. & Huelsenbeck, J.P. (2003) Mrbayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Sanderson, M.J. (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. *Molecular Biology and Evolution*, **19**, 101–109.

Smith, S.A., Beaulieu, J. & Donoghue, M.J. (2009) Mega-phylogenies for comparative biology: an alternative to supertree and supermatrix approaches. *BMC Evolutionary Biology*, **9**, 37.

Smith, S.A. & Donoghue, M.J. (2008) Rates of molecular evolution are linked to life history in flowering plants. *Science*, **322**, 86–89.

Smith, S.A. & O'Meara, B.C. (2012) Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics*, **28**, 2689–2690.

Smith, S.A., Beaulieu, J., Stamatakis, A. & Donoghue, M.J. (2011) Understanding angiosperm diversification using small and large phylogenetic trees. *American Journal of Botany*, **98**, 404–414.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Received 30 December 2012; accepted 4 March 2013
Handling Editor: Emmanuel Paradis

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Data S1. Vignette demonstrating a worked example of congruification.