Molecular Phylogenetics and Evolution xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Dubious resolution and support from published sparse supermatrices: The importance of thorough tree searches

Mark P. Simmons [a,*], Pablo A. Goloboff [b,c]

[a] Department of Biology, Colorado State University, Fort Collins, CO 80523, USA
[b] Consejo Nacional de Investigaciones Científicas y Técnicas, Miguel Lillo 205, 4000 S.M. de Tucumán, Argentina
[c] Instituto Miguel Lillo, Facultad de Ciencias Naturales, Miguel Lillo 205, 4000 S.M. de Tucumán, Argentina

ARTICLE INFO

Article history:
Received 9 April 2014
Revised 30 May 2014
Accepted 1 June 2014
Available online xxxx

Keywords:
Maximum likelihood
Parsimony
Rapid bootstrap
RAxML
SH-like aLRT
Tree hybridization

ABSTRACT

We re-analyzed 10 sparse supermatrices wherein the original authors relied primarily or entirely upon maximum likelihood phylogenetic analyses implemented in RAxML and quantified branch support using the bootstrap. We compared the RAxML-based topologies and bootstrap values with both superficial- and relatively thorough-tree-search parsimony topologies and bootstrap values. We tested for clades that were resolved by RAxML but properly unsupported by checking if the SH-like aLRT equals zero and/or if the parsimony-optimized minimum branch length equals zero. Four of our conclusions are as follows. (1) Despite sampling nearly 50,000 characters, highly supported branches in a RAxML tree may be entirely unsupported because of missing data. (2) One should not rely entirely upon RAxML SH-like aLRT, RAxML bootstrap, or superficial parsimony bootstrap methods to rigorously quantify branch support for sparse supermatrices. (3) A fundamental factor that favors thorough parsimony analyses of sparse supermatrices is being able to distinguish between clades that are unequivocally supported by the data from those that are not; superficial likelihood analyses that quantify branch support using the bootstrap cannot be relied upon to always make this distinction. (4) The SH-like aLRT and parsimony-optimized-minimum-branch-length tests generally identify the same properly unsupported clades; the latter is a more severe test.

1. Introduction

For over 25 years molecular phylogeneticists have been generating sequence data for numerous species within most macroscopic eukaryotic lineages, and typically sample the same set of gene regions within each lineage (e.g., ITS, *matK*, *rbcL*, and *trnL-F* for vascular plants). This wealth of publicly available data, coupled with genomic studies that cover an increasingly diverse set of taxa, has enabled systematists to create supermatrices (Sanderson et al., 1998) that often contain upwards of 200 species and 10,000 characters without actually generating any novel sequence data. Because of their broad taxonomic reach, numerous species sampled, and the expectation that their inclusion of many thousands of characters will lead to accurate phylogenetic inference, these supermatrix studies are generally highly cited and referenced by numerous scientists outside of the systematics community. The taxonomic breadth and numbers of species and characters are impressive, but so is the enormity of tree space (Felsenstein, 1978a) and the percentage of inapplicable and missing data, which typically constitute the majority (and sometimes >95%; e.g., Peters et al., 2011) of the "sparse" supermatrices.

Missing data in empirical sparse supermatrices that consist largely or entirely of publicly available data are inevitably non-randomly distributed among species and gene regions. The potential for these non-randomly distributed missing data, in the context of other factors such as rate heterogeneity among characters and/or branches, to cause phylogenetic artifacts in maximum likelihood (Felsenstein, 1973) and/or Bayesian MCMC (Yang and Rannala, 1997) analyses has been forcefully argued as either a minor (e.g., Wiens and Morrill, 2011; Roure et al., 2013) or a major (e.g., Lemmon et al., 2009; Simmons 2012a,b; Dell'Ampio et al., 2014) problem in empirical studies.

Even without the non-randomly-distributed-missing-data problem, obtaining optimal trees for empirical matrices with hundreds or thousands of terminals is a difficult problem for which dedicated heuristic techniques have been developed because standard branch-swapping techniques (such as standard subtree pruning and regrafting) are likely to fail (e.g., Goloboff, 1999;

* Corresponding author. Address: Department of Biology, 200 West Lake Street, Colorado State University, Fort Collins, CO 80523-1878, USA. Fax: +1 970 491 0649.
E-mail address: psimmons@lamar.colostate.edu (M.P. Simmons).

http://dx.doi.org/10.1016/j.ympev.2014.06.002
1055-7903/© 2014 Published by Elsevier Inc.

Nixon, 1999; Roshan et al., 2004; Goloboff and Pol, 2007). Even after optimal trees have been identified, there is the problem of sufficiently sampling the breadth of all optimal trees so that systematic inferences are restricted to properly supported clades that are present in the strict consensus (Schuh and Polhemus, 1980; Nixon and Carpenter, 1996; Goloboff and Farris, 2001). Matrices with low phylogenetic signal, whether caused by inclusion of few parsimony-informative characters, character conflict, or the distribution of missing data (as in sparse supermatrices), are particularly liable to have multiple equally optimal trees (Maddison, 1991; Morrison, 2007; Sanderson et al., 2011), which makes accurate identification of the strict consensus especially important.

Given the expected difficulty of finding optimal trees and the reasons to expect multiple optima, it is curious that many prominent sparse-supermatrix studies (e.g., see Section 2.1 below) have relied exclusively upon likelihood analyses implemented in RAxML (Stamatakis, 2006) for phylogenetic inference. RAxML relies upon "lazy" and local subtree pruning and regrafting and only ever presents a single fully resolved optimal tree. Stamatakis et al. (2008, p. 770) asserted that rapid bootstrapping in RAxML "... solves—to a large extent—the computational problems associated with present-day full [maximum likelihood] analyses with a couple of hundred or a few thousand taxa." Peters et al. (2011, p. 10) were equally confident: "Unless one wants to analyze data sets that are significantly larger than ours (i.e., 1146 terminals and 88,626 characters), there is no computational or speed argument left to perform supertree or parsimony methods in favor of ML analyses." With respect to the issue of finding equally optimal trees, Bininda-Emonds and Stamatakis (2007) asserted that presenting a single fully resolved optimal tree is not problematic because the complexity of the likelihood (as opposed to parsimony) surface typically only allows for one or a few equally optimal trees. In contrast to Stamatakis et al. (2008), Siddall (2010) noted that in rapid bootstrapping the results are biased in favor of the original tree topology and Simmons and Norton (2014) showed how rapid bootstrapping with the GTRCAT model in RAxML can provide extremely high support values for simple 4-terminal polytomies and matrices that have no missing data. In contrast to Bininda-Emonds and Stamatakis (2007), Morrison (2007) argued that presenting a single fully resolved optimal tree in many cases constitutes specious precision that is not representative of the data.

Given the strong differences in opinion expressed by the above authors regarding the use of parametric methods to analyze sparse supermatrices as well as the suitability of lazy, local subtree-pruning-regrafting searches in RAxML to conduct those analyses, it is unclear whether the resulting trees should be embraced as "... presenting the state-of-the-art with respect to hypotheses of evolutionary relationships within the group" (Bininda-Emonds, 2011, p. 1; in reference to Peters et al. (2011)) wherein the bootstrap values are conservative estimates of branch support (Pyron and Wiens, 2011; Pyron et al., 2011) or rather an example from bioinformatics wherein, "Overzealous data mining is seen to have replaced carefully performed experimental analyses..." (Morrison, 2013, p. 349).

Two alternative (or perhaps complementary) approaches to test for properly unsupported clades in phylogenetic analyses wherein only a single optimal tree is presented (as in GARLI (Zwickl, 2006), PhyML (Guindon et al., 2010), and RAxML) are to check if the SH-like aLRT (Shimodaira–Haesgawa-like approximate likelihood ratio test; Anisimova and Gascuel, 2006; Guindon et al., 2010) value equals zero and to check if the parsimony-optimized minimum branch length equals zero (Simmons and Norton, 2014; Simmons and Randle, 2014). These two approaches have the advantage of requiring little additional computational power beyond the initial tree search and of being implemented in widely used programs. Therefore, they are readily applicable to supermatrices containing thousands of terminals. Both approaches are capable of identifying properly unsupported clades in simple simulated examples (4-terminal polytomies (Simmons and Norton, 2014); 8-terminal trees with various distributions of missing data or other ambiguous characters (Simmons and Randle, 2014)), but the question remains: how do they perform on large empirical sparse supermatrices? In such cases limiting SH-like aLRT comparisons to alternative topologies that are connected by nearest-neighbor-interchange swaps may grossly overestimate support when other swaps (e.g., subtree-pruning regrafting to a distant node) produce trees of the same likelihood. Identifying properly unsupported clades in sparse supermatrices is arguably the most important context for these two alternative approaches because of the high probability of having numerous properly unsupported clades given the superficial tree searches that are employed relative to the vast number of possible trees and the very high percentage of missing data in the matrix.

In this study we re-analyzed 10 published sparse supermatrices wherein the original authors relied primarily or entirely upon likelihood analyses implemented in RAxML and quantified branch support using the bootstrap. We compared the fully resolved RAxML-based topologies and bootstrap values with both superficial and relatively thorough-tree-search parsimony topologies (either fully resolved or the strict consensus) and bootstrap values. We also tested for properly unsupported clades on the RAxML topologies by checking if the SH-like aLRT value equals zero and checking if the parsimony-optimized minimum branch length equals zero. By making these comparisons among alternative tree-search methods and ways of quantifying branch support, we sought to quantify the extent to which these sparse supermatrices contain properly unsupported clades and inflated branch-support values based on the limitations of both superficial tree searches as well as the non-random distributions of missing data. We found unsupported resolution and inflated branch support in all 10 sparse supermatrices, though the extent to which these problems occurred varies widely.

2. Methods

2.1. Supermatrices sampled

The following 10 prominent recently published supermatrices were selected for inclusion in this study: Fabre et al. (2009; hereafter "Fabre"), Hedtke et al. (2013; hereafter "Hedtke"), Hinchliff and Roalson (2013; hereafter "Hinchliff"), Nyakatura and Bininda-Emonds (2012; hereafter "Bininda"), Peters et al. (2011; hereafter "Peters"), Pyron and Wiens (2011; hereafter "Wiens"), Pyron et al. (2011; hereafter "Pyron"), Soltis et al. (2013; hereafter "Soltis"), Springer et al. (2012; hereafter "Springer"), and van der Linde et al. (2010; hereafter "Linde"). These supermatrices include 180–2872 terminals, 5814–88,626 characters, and 66.7–98.4% missing or inapplicable data (Table 1). All of the matrices are based on sequence characters, all but one of these supermatrices include characters sampled from two or three genomes, and the taxa sampled range from Magnoliophyta (Hinchliff, Soltis) to Insecta (Hedtke, Linde), and Vertebrata (Bininda, Fabre, Peters, Pyron, Springer, Wiens; Table 1). Given the wide breadth of sampling with respect to numbers of terminals and characters, percent missing data or inapplicable entries, and genomes and taxa sampled, we hypothesize that our results will be broadly applicable to contemporary plant and animal sparse supermatrix studies in general.

In the three cases where the authors of the original studies analyzed two or more supermatrices, we selected the one that they focused on in their results and discussion (though Hedtke focused about equally on both of their matrices). For Hedtke we sampled

Please cite this article in press as: Simmons, M.P., Goloboff, P.A. Dubious resolution and support from published sparse supermatrices: The importance of thorough tree searches. Mol. Phylogenet. Evol. (2014), http://dx.doi.org/10.1016/j.ympev.2014.06.002