Molecular Phylogenetics and Evolution 67 (2013) 277–296

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

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

Quantification and relative severity of inflated branch-support values generated by alternative methods: An empirical example

Mark P. Simmons [a,*], Andrew P. Norton [b]

[a] Department of Biology, Colorado State University, Fort Collins, CO 80523-1878, USA
[b] Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, CO 80523, USA

ARTICLE INFO

Article history:
Received 8 November 2012
Revised 18 January 2013
Accepted 31 January 2013
Available online 9 February 2013

Keywords:
Bootstrap
Maximum likelihood artifacts
Missing data
PhyML
RAxML
Supermatrix

ABSTRACT

A supermatrix of 272 terminals from Rubiaceae tribe Spermacoceae that were scored for up to 10 gene regions (two nrDNA, eight plastid) was used as an empirical example to quantify sources of error in heuristic parametric (Bayesian MCMC and maximum likelihood) phylogenetic analyses. The supermatrix includes dramatic disparities in which terminals were sampled for which gene regions. The sources of error examined include poor quality tree searches, requiring a single fully resolved optimal tree, undersampling-within-replicates and frequency-within-replicates bootstrap artifacts, and extrapolation from one character partition to another such that synapomorphies that would only be ambiguously optimized by parsimony are optimized with high probability by parametric methods. Four of our conclusions are as follows. (1) The resolution and support provided by parametric methods for clades that lack unambiguously optimized (by parsimony) synapomorphies are less robust to the addition of terminals and characters than those clades that have unambiguously optimized synapomorphies. (2) Those tree-search methods which can create phylogenetic artifacts (frequency-within-replicates resampling, undersampling-within-replicates resampling, requiring a single fully resolved optimal tree, non-independent resampling among replicates) provide the greatest resolution and support irrespective of whether that resolution or support is corroborated by more conservative and better justified methods. (3) Partitioning data matrices cannot be relied upon to consistently obviate potentially dubious resolution and support caused by missing-data artifacts in likelihood analyses when the models require linked branch lengths among partitions. (4) Undersampling-within-replicates and frequency-within-replicates resampling artifacts are not unique to parsimony and should be accounted for in likelihood analyses by allowing multiple equally likely trees to be saved within each resampling pseudoreplicate and applying the strict-consensus bootstrap rather than the frequency-within-replicates bootstrap.

1. Introduction

Contemporary phylogenetic analyses typically sample multiple sets of characters (e.g., separate gene regions) and include both novel data as well as data taken from previous studies that were often generated by different labs to address separate questions (phylogenetic, genomic, gene-family, or otherwise). Consequently, some character sets and terminals are well sampled whereas others are very poorly sampled. The general willingness to sample character sets with large amounts of missing data (Wiens, 1998, 2003, 2006) combined with recognition of the importance of extensive terminal sampling to subdivide long branches (Hillis, 1996, 1998; Graybeal, 1998) and the desire to maximize terminal sampling for their study lineage leads many systematists to create supermatrices that are dominated by missing and inapplicable data (in the extreme, 4.3% scored cells by McMahon and Sanderson (2006); 1.55% scored orthologs by Peters et al. (2011)). Rigorous tree searches applied to these matrices are computationally difficult (e.g., Felsenstein, 1978a; Rice et al., 1997), particularly in a maximum likelihood (hereafter "likelihood") or Bayesian context (e.g., Sanderson and Kim, 2000; Soltis et al., 2007).

Six artifacts that can occur when quantifying resolution and support provided by a given matrix for the optimal phylogenetic tree(s) are described below. We expect all six of these artifacts to be particularly problematic when analyzing supermatrices with high levels of missing data and low overlap in character sampling among terminals. The first four artifacts apply to both parsimony and likelihood contexts, the fifth is unique to Bayesian inference, and the sixth applies to both Bayesian and likelihood methods.

It is widely recognized that Bayesian posterior probabilities are inflated relative to the ideal as well as both bootstrap (Felsenstein, 1985) and jackknife (Farris et al., 1996) resampling values unless nearly all of their assumptions are met (Suzuki et al., 2002;

* Corresponding author. Fax: +1 970 491 0649.
E-mail address: psimmons@lamar.colostate.edu (M.P. Simmons).

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.01.020

Cummings et al., 2003; Simmons et al., 2004; Huelsenbeck and Rannala, 2004), which is unlikely with empirical data. Hence the discrepancy in levels of support that many authors consider biologically meaningful when analyzing their empirical data (i.e., the widespread citation of Hillis and Bull (1993) for a 70% bootstrap cut-off, while applying a 0.95 posterior probability cut-off; e.g., Fouquet et al., 2012; Rothfels et al., 2012; Schweizer et al., 2012). But inflated bootstrap or jackknife support is another possibility that is not as widely recognized. The first four artifacts listed below can inflate the apparent support inferred from these resampling values.

First, reporting the bootstrap or jackknife majority rule consensus (Margush and McMorris, 1981) rather than plotting those percentages onto the strict consensus (Schuh and Polhemus, 1980) of all optimal trees may lead to inferences of high support for clades that are properly unsupported by the data (Simmons and Freudenstein, 2011). Only those clades present in all optimal trees are unequivocally supported by the data (Nixon and Carpenter, 1996b; Goloboff et al., 2003).

Second, application of frequency-within-(pseudo)replicate (FWR) resampling support, as recommended by Felsenstein (2004) and implemented in PAUP* (Swofford, 2001a), rather than strict-consensus resampling support, as recommended by Davis et al. (1998) and implemented in TNT (Goloboff et al., 2008), can cause inflated support values for both supported (Davis et al., 1998; Freudenstein and Davis, 2010) and properly unsupported (Goloboff and Pol, 2005; Simmons and Freudenstein, 2011) clades. The FWR artifact occurs by assigning partial support to clades that are unsupported (i.e., present in some but not all optimal trees within a given pseudoreplicate; Goloboff et al., 2003) and is particularly misleading when ambiguity is determinate to inferred support (i.e., the more ambiguity in one part of the tree the greater the support in another part of the tree; Sharkey and Leathers, 2001; Sumrall et al., 2001).

Third, saving a non-representative subset of multiple optimal trees, particularly when only a single tree is held (as in GARLI [Zwickl, 2006], PhyML [Guindon and Gascuel, 2003], and RAxML [Stamatakis, 2006]), can create the undersampling-within-(pseudo)replicates (UWRs) artifact (Goloboff and Farris, 2001; Simmons and Freudenstein, 2011). Like the FWR artifact, the UWR artifact can cause inflated support values for both supported and properly unsupported clades. Both the FWR and UWR artifacts are expected to be particularly serious for small clades when there are multiple optimal trees (Simmons and Freudenstein, 2011).

Fourth, bootstrap resampling creates arbitrary character weightings by resampling characters with replacement (Freudenstein and Davis, 2010), which can result in high support values for properly unsupported clades (Simmons and Freudenstein, 2011). This artifact does not apply to jackknife resampling, which relies on character deletion rather than resampling with replacement. Yet several programs (e.g., GARLI, PhyML, and RAxML) only implement bootstrap resampling.

Fifth, Pickett and Randle (2005) and Goloboff and Pol (2005) demonstrated that small clades (i.e., those with few terminals), which have high prior probabilities when all trees are assigned equal prior probabilities (as in MrBayes; Ronquist and Huelsenbeck, 2003), can have inflated posterior probabilities relative to larger clades when there is insufficient data to overcome the priors.

Sixth, extrapolation by parametric (i.e., Bayesian and likelihood) methods of branch lengths and model-parameter values from one character partition to another may provide highly probable synapomorphies for clades that would only be ambiguously optimized without this extrapolation, particularly with non-randomly distributed missing data and model heterogeneity that is not accounted for across partitions (Gatesy et al., 2002; Lemmon et al., 2009; Simmons, 2012a,b). Lemmon et al. (2009) demonstrated that these artifacts may result in both long-branch attraction (Felsenstein, 1978b) and long-branch repulsion (Siddall, 1998) and can be positively misleading.

Ekman and Blaalid (2011) proposed that the missing-data artifacts described by Lemmon et al. (2009), at least in a Bayesian context, are themselves primarily artifacts of simulating trees with all branches of equal length yet applying the default exponential branch-length prior in MrBayes. Alternatively, Wiens and Morrill (2011) dismissed the generality of Lemmon et al.'s (2009) findings, at least in a Bayesian context, by attributing Lemmon et al.'s observed errors to sampling invariant or saturated characters and failing to apply different model parameters to distinct process partitions (Bull et al., 1993). These criticisms of Lemmon et al. (2009) were partially addressed by Simmons (2012a,b), wherein partitioned models were applied in the context of contrived examples, simulations, and empirical data. Furthermore, Simmons (2012a,b) demonstrated that Bayesian analyses are more robust to these artifacts than likelihood analyses. Consequently, just because Wiens and Morrill (2011) did not observe severe artifacts in a Bayesian context does not necessarily mean that their results can be directly extrapolated to likelihood analyses.

Simmons (2012a,b) found that partitioning blocks of characters based on their distributions of missing data helped to decrease, but did not eliminate, missing-data artifacts of dubious resolution and support in likelihood and Bayesian analyses. Yet the decreased resolution and support values could have simply been a consequence of the greater variance caused by adding model parameters (e.g., Wertheim et al., 2010) rather than being advantageous to addressing missing data per se.

In addition to partitioning characters and allowing each partition to have its own rate multiplier and potentially a different model and/or model-parameter values, in MrBayes vers. 3.1 and 3.2 it is also possible to unlink branch-lengths. The latter alternative may dramatically decrease or even eliminate the missing-data artifacts (John Gatesy, pers. comm., 2011). Simmons (2012a,b) only examined the former alternatives, whereas Johnson et al. (2012) also examined the latter alternative. Johnson et al. (2012, p. 147) reported that:

"However, these [unlinked-edge-lengths] models cannot be used when data for some genes is completely missing for some taxa. In this case the software does not have sufficient information to estimate partition-specific branch lengths for all combinations of partitions and taxa. (We found that while software such as RAxML or MrBayes would run without crashing in such cases the results were wildly erratic – results not shown.)"

Chris Randle (pers. comm., 2012) noted that, at least in a Bayesian MCMC context (Yang and Rannala, 1997), just because the software does not have sufficient information to estimate partition-specific branch lengths for all partitions and taxa does not necessarily mean that wildly erratic *topological* results for the analysis as a whole would necessarily be produced. Rather, the topology of the tree regions affected by missing data in some partitions will simply be determined based on the partitions that do have data. Hence partitioned models with unlinked branch lengths are a potentially promising approach to address missing-data artifacts.

1.1. A priori hypotheses to test

In this project we sought to test the following seven a priori hypotheses and quantify the relative severity of five of the six potential artifacts described above (not including the first artifact) by using an empirical supermatrix of 272 terminals that were sampled for up to 10 gene regions. To our knowledge, this is the first study to quantify the relative severity of all five of these artifacts in any context.