Molecular Phylogenetics and Evolution 78 (2014) 66–86

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Disparate parametric branch-support values from ambiguous characters

Mark P. Simmons[a,*], Christopher P. Randle[b]

[a] Department of Biology, Colorado State University, Fort Collins, CO 80523-1878, USA
[b] Department of Biological Sciences, Sam Houston State University, Huntsville, TX 77341, USA

ARTICLE INFO

Article history:
Received 7 December 2013
Revised 23 April 2014
Accepted 29 April 2014
Available online 10 May 2014

Keywords:
Ambiguous optimization
Bootstrap
Decisive taxon coverage
Parsimony
Redundant taxa
SH-like aLRT

ABSTRACT

The greater power of parametric methods over parsimony is frequently observed in empirical phylogenetic analyses by providing greater resolution and higher branch support. This greater power is provided by several different factors, including some that are generally regarded as disadvantageous. In this study we used both empirical and (modified) simulated matrices to examine how Bayesian MCMC, maximum likelihood, and parsimony methods interpret ambiguous optimization of character states. We describe the information content in "redundant" terminals as well as a novel approach to help identify clades that cannot be unequivocally supported by synapomorphies in empirical matrices. Four of our main conclusions are as follows. First, the SH-like approximate likelihood ratio test is a more reliable indicator than the bootstrap of branches that are only ambiguously supported in likelihood analyses wherein only a single fully resolved optimal tree is presented. Second, bootstrap values generated by methods that only ever present a single fully resolved optimal tree are less robust to differences in taxon sampling than are those generated by more conservative methods. Third, PAUP* likelihood is more resilient to producing apparently unambiguous resolution and high support from ambiguous characters than is GARLI collapse 1 and MrBayes, which in turn are more resilient than PhyML. GARLI collapse 0, IQ-TREE, and RAxML are the least resilient bootstrapping methods examined. Fourth, frequent discrepancies with respect to resolution and/or branch support may be obtained by methods that only ever present a single fully resolved optimal tree in different contexts that are apparently unique to the specific program and/or method of quantifying branch support.

1. Introduction

Data that appear ambiguous to equally weighted parsimony (Farris, 1970; Fitch, 1971) may be determinate to parametric methods (e.g., maximum likelihood, Bayesian Markov Chain Monte Carlo (MCMC; Felsenstein, 1973; Yang and Rannala, 1997) in a manner that both increases power (Penny et al., 1992) and is generally regarded as advantageous. Examples include assigning different rates to different character-state transformations in a Q-matrix (e.g., transitions vs. transversions at 2-fold degenerate third codon positions; Kimura, 1980), allowing different character partitions to evolve at different rates and different model-parameter values (e.g., exons vs. introns, stem vs. loop regions of rDNA; Castoe et al., 2004), and using branch lengths to potentially obviate long-branch attraction (Felsenstein, 1978b).

There are other cases that also increase power but are generally less well known and regarded as disadvantageous. Examples include how some parametric methods handle long branches that actually are sister groups (Siddall, 1998; Siddall and Whiting, 1999), non-random distributions of missing data (Lemmon et al., 2009; Simmons, 2012a, 2012b), polytomies caused by lack of synapomorphies or character conflict (Suzuki et al., 2002; Simmons and Norton, 2014), and matrices for which multiple equally optimal trees should be reported but are not presented by the particular implementation (i.e., computer program) of likelihood (Sanderson et al., 2010, 2011; Simmons and Goloboff, 2013). In this study we investigated another set of cases that might be generally considered to be disadvantageous for parametric methods—how they interpret ambiguous optimization of character states.

Swofford et al. (1996, p. 428, their Fig. 9) presented a classic contrived example that includes both branch-length heterogeneity and ambiguous optimization of character states. They focused on a single character with state A in one ingroup lineage, state C in the sister ingroup lineage, and state G in the outgroup. Long branches (as determined by other characters in the data matrix) lead to both the outgroup terminal and the ingroup lineage with state C. In contrast, only a short branch leads to the ingroup lineage of eight terminals that all have state A. Given the discrepancy in branch

[*] Corresponding author. Address: Department of Biology, Colorado State University, 200 West Lake Street, 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.2014.04.029
1055-7903/© 2014 Elsevier Inc. All rights reserved.

lengths, a G:A change probably occurred on the long outgroup branch and an A → G change probably occurred on the long ingroup branch leading to the ingroup terminal with state C. Therefore, the most recent common ancestor of the ingroup probably had state A and if a new terminal is added to the analysis with state C, likelihood would (all else equal) favor placing that new terminal as sister to the existing terminal with state C. In contrast, parsimony, which is a less powerful method that ignores branch-length information derived from other characters (Penny et al., 1992), would be unable to select among three equally parsimonious placements for the new terminal. Swofford et al. (1996, p. 429) favored the "appropriate predisposition" of maximum likelihood over the lack of resolution provided by parsimony.

The advantage of likelihood over parsimony in Swofford et al.'s (1996) example is premised on branch-length heterogeneity, with two of the three branches being long while the third is short. But what happens if the outgroup branch remains long and both ingroup lineages are on equal-length branches aside from the single character in question? If the same result for a new terminal with C is obtained, then should likelihood's predisposition still be regarded as advantageous relative to the lack of resolution provided by parsimony?

In this equal-ingroup-branch-length situation, one of the two changes probably occurred on the long outgroup branch, but under a Jukes–Cantor model (JC; Jukes and Cantor, 1969) the second change is equally probable on both ingroup branches. Therefore, one might expect likelihood to be unable to select which of the two ingroup branches to place the new terminal with state C. Another possibility is to invoke three (or more) changes, but, as Swofford et al. (1996, p. 429) noted, "… we know that at least two changes must have occurred, and since change is rare in this example, histories with three of more changes are less likely than those with only two changes."

We investigated this type of scenario, as well as others that may result in ambiguous optimization of character states, in the context of Bayesian MCMC, likelihood, and equally weighted parsimony analyses. We identified character-state distributions that may cause discrepancies in resolution and branch support between these three phylogenetic-inference methods from empirical matrices that were originally analyzed by Guo et al. (2013) and Simmons and Norton (2013). These matrices are based on 10 plastid and nuclear rDNA gene regions that were sampled for a lineage of several genera within the flowering-plant taxon Rubiaceae tribe Spermacoceae. After these character-state distributions were identified, we used (modified) simulations to test whether each of them can actually cause the observed discrepancy or whether they are merely incidental correlations. We also describe the information content in "redundant" terminals as well as a novel approach to help identify clades that cannot be unequivocally supported by synapomorphies in empirical matrices.

2. Materials and methods

2.1. Empirical matrices

Simmons and Norton (2013) identified 423 clades that were resolved by one or more of the parametric methods they implemented but were unresolved in the strict consensus of all most parsimonious trees identified (i.e., their codes 4 → 9 and 11 from their Table 3). Of these 423 clades we focused on those 156 that met the following three criteria. First, they could be compartmentalized (Maddison et al., 1984) into a sub-matrix (relative to the complete matrices of Guo et al. (2013) that were analyzed by Simmons and Norton (2013)) that includes at least three successive sister groups while including less than 30 terminals. This was done for computational tractability, particularly for the PAUP* ver. 4.0b10 (Swofford, 2001a) likelihood analyses. Second, the focal clade was unresolved in the parsimony-based strict consensus in the compartmentalized sub-matrix. This was done because the focus of this manuscript is on those clades that are resolved by one or more parametric methods but are unsupported (i.e., not present in the strict consensus of all optimal trees; Goloboff et al., 2003) by equally weighted parsimony. Third, the focal clade was well supported (i.e., ⩾70% bootstrap or ⩾0.95 posterior probability) by one or more of the parametric methods in Simmons and Norton's (2013) analyses. This was done because these cutoffs are frequently applied by authors of empirical studies to determine whether or not to make evolutionary inferences and taxonomic changes based on clades resolved in their optimal trees (e.g., see Simmons and Norton's (2014), Table S1).

Compartmentalized phylogenetic analyses were not necessarily performed for all 156 focal clades independently of each other. Rather, when possible, single compartmentalized matrices were created that included two or more of the focal clades for computational efficiency. All 88 compartmentalized matrices are posted as Supplemental online data at http://rydberg.biology.colostate.edu/Research/.

Compartmentalized phylogenetic analyses for the 156 focal clades were performed for two reasons, both of which are expected to improve phylogenetic inference and estimates of branch support relative to the original analyses from Simmons and Norton (2013) on the complete matrices with 74 → 272 terminals each. First, the original resolution and support of the focal clade may have been artifacts of low quality searches of tree space given the vast numbers of alternative topologies (Felsenstein, 1978a) and the reliance of most parametric methods on subtree-pruning–regrafting (with additional restrictions in both PhyML and RAxML) while only ever presenting a single fully resolved optimal tree (see Goloboff (1999), Nixon (1999) and Davis et al. (2005) for the importance of performing more thorough searches for matrices with numerous terminals). Second, by eliminating terminals that are only distantly related to the focal clade it is less likely that ambiguous resolution elsewhere in the tree will artificially inflate bootstrap support for the focal clade (Sharkey and Leathers, 2001; Sumrall et al., 2001; Simmons and Freudenstein, 2011).

Six potentially confounding factors in comparing the resolution and support from the compartmentalized analyses relative to Simmons and Norton's (2013) complete-terminal-sampling analyses are as follows. First, although the model used (GTR + Γ) is identical between both analyses, the estimated model-parameter values are likely to differ and there are fewer sequences with which to estimate the values. This factor is applicable to GARLI, PhyML, RAxML, and MrBayes, but not to PAUP*, wherein the jModeltest (Posada, 2008) estimated values were fixed for both sets of analyses. Second, for those methods that estimate model-parameter values from the data, there are fewer data in the compartmentalized matrices to derive these estimates from. This may increase the variance in parameter-value estimates and decrease support values. Third, it is possible that optimizations of individual characters would change in the compartmentalized analyses if more distantly related terminals to the focal groups were sampled. Fourth, unless the terminals sampled in the compartmentalized analyses were supported as a clade with 100% bootstrap support and 1.0 posterior probability, then the null hypothesis is that support values would increase in the compartmentalized analyses because of the removal of homoplasy (Sanderson and Wojciechowski, 2000). Fifth, in the compartmentalized analyses it is possible for characters that were parsimony-informative within the focal clade or its relatives to be rendered parsimony-uninformative. For example, it is possible that a symplesiomorphy in a complete-terminal-sampling analysis will be re-interpreted as