Molecular Phylogenetics and Evolution 73 (2014) 87–96

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Divergent maximum-likelihood-branch-support values for polytomies

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 October 2013
Revised 9 January 2014
Accepted 21 January 2014
Available online 4 February 2014

Keywords:
Bootstrap
False positive
GTRCAT model
Hard polytomy
Random seed
SH-like aLRT

ABSTRACT

We applied simple 4-taxon simulations with 3-way character conflict or a hard polytomy to check for false positive branch support, with a focus on the bootstrap and recently introduced likelihood-based phylogenetic-inference programs. Given that there are only three possible bifurcating topologies, discrepancies among methods identified in this study should generally be restricted to factors other than topological search heuristics. Our four major conclusions are as follows. First, Bayesian MCMCMC posterior probabilities are not the only means of quantifying support that can produce dramatically inflated values when applied to cases of strong character conflict. Rapid bootstrapping with the GTRCAT model in RAxML can provide still greater support values for polytomies and we suggest that it generally be avoided. Second, the SH-like approximate likelihood-ratio test outperforms the bootstrap when applied to polytomies. We suggest that the SH-like aLRT be widely applied to likelihood-based empirical studies to complement the bootstrap by collapsing those branches with an SH-like aLRT percentage of $\leqslant$10, irrespective of how high the likelihood bootstrap support is. Third, the 70% bootstrap cutoff does not equate to a 5% error rate and we suggest that the idea that $\geqslant$70% bootstrap generally equates to 95% probability of accuracy in empirical analyses finally be abandoned. Fourth, rapid bootstrapping with the GTRCAT model in RAxML can generate values with very low precision, which reinforces our assertion that this method should be avoided, let alone be entirely relied upon for phylogenetic inference.

1. Introduction

Some means of quantifying branch support is incorporated into almost every published phylogenetic analysis. The most common way of quantifying branch support for frequentist phylogenetic analyses is to use the bootstrap (Felsenstein, 1985). The bootstrap can be implemented in multiple different ways. Hence bootstrap values generated by one program and one set of tree-search search settings can be dramatically different from those generated by a different program and/or different set of search settings – even when the same optimality criterion and model (when applicable) is applied. Artificially inflated bootstrap values may be caused by use of the frequency-within-replicates rather than the strict-consensus bootstrap (Davis et al., 1998; De Laet et al., 2004; Goloboff and Pol, 2005), saving only a single optimal tree per pseudoreplicate when there are multiple equally optimal trees (Goloboff and Farris, 2001; Simmons and Freudenstein, 2011; Simmons and Goloboff, 2013), and/or extrapolation of branch lengths among character partitions in the presence of non-randomly distributed missing or inapplicable data (Lemmon et al., 2009; Simmons, 2012a, 2012b). Alternatively, artificially deflated bootstrap values may be caused by performing low quality tree searches that do not find the optimal trees in each pseudoreplicate (Freudenstein et al., 2004; Freudenstein and Davis, 2010).

Irrespective of how the bootstrap is implemented and how inflated or deflated those support values may be, numerous pseudoreplicates are required to obtain precise bootstrap values. Between 100 and 1,000 bootstrap pseudoreplicates are typically implemented in contemporary empirical phylogenetic analyses (e.g., Bacon et al., 2013; Ceccarelli and Zaldivar-Riveron, 2013; Guo et al., 2013; Kwek et al., 2013). Hedges (1992) noted that, based on the binomial distribution (also see Efron et al. (1996)), still more bootstrap replicates (1825) are required to have a 95% confidence interval within 1% for bootstrap supports of 95%. Furthermore, Hedges (1992) noted that the breadth of the confidence interval for any given number of bootstrap pseudoreplicates varies depending on the bootstrap support, with lower bootstrap supports having wider confidence intervals.

Pattengale et al. (2009, 2010) criticized Hedges' (1992) and Efron et al.'s (1996) approach to estimating the "accuracy" (properly precision; Hillis and Bull, 1993) of bootstrap values based strictly on the binomial distribution. They asserted that other

* 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.2014.01.018
1055-7903/

factors (e.g., alignment quality, percentage of gaps, and strength of phylogenetic signal) also affect the precision of bootstrap values and hence the number of bootstrap pseudoreplicates performed on a given matrix should be determined using an adaptive stopping criterion for each individual matrix rather than an a priori defined number as advocated by Hedges (1992). Pattengale et al. (2010) asserted that Hedges' (1992) estimate of the precision of bootstrap values be used as an upper bound and that many fewer pseudoreplicates are necessary in practice.

In addition to considering whether or not the bootstrap values are inflated and precisely estimated, one needs to determine how to interpret the values. Hillis and Bull (1993, p. 191, 187) reported that, "bootstrap proportions [>50%] provide conservative estimates of accuracy under many conditions" and "that estimated internal branches with bootstrap proportions above 70% represent true clades over 95% of the time (for the conditions tested in these simulations)." Hillis and Bull (1993) appropriately qualified the generality of their results and noted that their simulation conditions that produced this result are favorable to accurate phylogenetic inference.

Hillis and Bull (1993) is the basis for the widespread approach of using 70% as a meaningful threshold for bootstrap values. Based on a search conducted in Web of Science® on 30 September 2014, Hillis and Bull (1993) has been cited 2123 times (the second most highly cited paper from *Systematic Biology*). The large majority of these citations are for using ⩾70% bootstrap as an indication of strong support (or "well supported, "good support," "significant support," etc.). For example, of the 41 empirically focused (as opposed to conceptually focused) papers published in *Molecular Phylogenetics and Evolution* from 2010 – August 2013, 40 of them cited Hillis and Bull (1993) in this manner (Table S1).

Although rarely cited in this manner in *Molecular Phylogenetics and Evolution* in recent years (but see Ribeiro et al. (2012)), Hillis and Bull (1993) is also still cited as the basis for extrapolating ⩾70% bootstrap to being an indication of 95% probability of accuracy in empirically focused papers (e.g., Antiabong et al., 2013; Engelbrecht et al., 2013; Guz et al., 2013; Keskin et al., 2013; Kumar et al., 2013). This continues to occur despite Hillis and Bull's (1993) original qualifications as well as those made by other authors on both the interpretation of bootstrap values as a measure of accuracy (Brown, 1994; Holmes, 2003; Anisimova et al., 2011) and the cases wherein ⩾70% bootstrap does not equate to 95% probability of accuracy in empirical analyses (Felsenstein and Kishino, 1993; Soltis and Soltis, 2003; Huelsenbeck and Rannala, 2004).

The standard bootstrap is time-consuming and faster alternatives that have been proposed include rapid bootstrapping (RBS; Stamatakis et al., 2008), the ultrafast bootstrap approximation (UFBoot; Minh et al., 2013), and the approximate likelihood-ratio test (aLRT; Anisimova and Gascuel, 2006). Rapid bootstrapping, as implemented in RAxML (Stamatakis et al., 2005), is generally applied with the CAT-based approximation of rate heterogeneity (e.g., GTRCAT). Rapid bootstrapping with the GTRCAT model has been criticized as providing inflated bootstrap values, which are likely caused by biased starting trees and low quality tree searches (Siddall, 2010; Anisimova et al., 2011; Simmons and Norton, 2013). Nonetheless, these methods are not only used as a computational shortcut to regular bootstrapping for supermatrices with several hundred terminals wherein drastic heuristic shortcuts are necessary for likelihood-based analyses (e.g., Hinchliff and Roalson, 2013; Soltis et al., 2013) – they are also periodically applied to much smaller matrices (e.g., Colston et al., 2013; Miner et al., 2013; Olsson et al., 2013).

Minh et al. (2013) introduced UFBoot as an alternative to standard bootstrapping and rapid bootstrapping to quickly quantify branch support in likelihood-based analyses of large matrices. UFBoot applies resampling-estimated log-likelihoods (Hasegawa and Kishino, 1994) wherein likelihood scores for each character from the original matrix are used to estimate likelihood scores for the bootstrap pseudoreplicates. Important quartet puzzling with NNI branch swapping (Vinh and von Haeseler, 2004) is used to explore tree space for each bootstrap pseudoreplicate. Minh et al. (2013, p. 1189) reported that, unlike the conservatively biased standard bootstrap, "UFBoot is unbiased for support values higher than 70%" such that "... a split with support of 95% will have a probability of 0.95 to be correct." They implemented UFBoot, together with standard bootstrapping and the SH-like aLRT, in the program IQ-TREE.

Anisimova and Gascuel (2006) introduced the aLRT as a faster alternative to the bootstrap to quantify branch support. The aLRT tests whether branches have a positive length as opposed to being zero-length and makes nearest-neighbor-interchange (NNI) comparisons at every internal branch in the optimal tree. The aLRT was later modified by Guindon et al. (2010) into the SH-like aLRT (named after Shimodaira and Hasegawa (1999)) by changing the null hypothesis to all three NNI resolutions of a given internal branch being equally likely. Guindon et al. (2010) reported that the SH-like aLRT may perform particularly well relative to the bootstrap when applied to polytomies and very short branches such that the support values assigned to these branches are very low or even zero.

Anisimova and Gascuel (2006, p. 550) asserted that, "The main advantage of the aLRT is that it is much faster than either the [maximum likelihood] bootstrap or the Bayesian inference." Indeed, the SH-like aLRT has been applied to a very large supermatrix for which the stand likelihood bootstrap or Bayesian MCMCMC (Yang and Rannala, 1997) inference may not have been computationally tractable (Pyron et al., 2013).

The simplest way to quantify and compare branch-support values is to restrict phylogenetic analyses to 4-taxon statements. Suzuki et al. (2002) used simulated unrooted 4-taxon statements to test for the fraction of false positives (i.e., type I errors) generated by Bayesian MCMCMC, neighbor joining, and parsimony methods of quantifying branch support using posterior probabilities or the bootstrap. In doing so they generated matrices of 15,000 characters using simple nucleotide-substitution models, and all terminal branches of short and equal length. Taken together, these three factors facilitate accurate phylogenetic inference. Two main types of matrices were generated: those with a zero-length internal branch (hereafter a "star matrix") and those with 5000 characters generated from all three alternative topologies and then concatenated to create the matrix of 15,000 characters (hereafter a "conflict matrix"). Any resolution generated by one of the phylogenetic-inference methods for either matrix is based on stochastic character variation. Hence high branch-support values should be regarded as false-positives. Suzuki et al. (2002) found that Bayesian posterior probabilities, but not neighbor-joining or parsimony bootstrap values, have a high false-positive rate – even when the true nucleotide-substitution model was used for phylogenetic inference. This behavior of Bayesian posterior probabilities has been attributed to the "star-tree paradox" wherein arbitrary resolutions of polytomies are frequently assigned high posterior probabilities by matrices that include numerous characters because of how the priors are set (Lewis et al., 2005; Yang, 2007a).

In this study we applied the simple 4-taxon simulations used by Suzuki et al. (2002) for detection of false positive branch support to four novel programs (GARLI (Zwickl, 2006), IQ-TREE, PhyML (Guindon and Gascuel, 2003), RAxML) and three methods of quantifying branch support (rapid bootstrapping, UFBoot, aLRT) that have been introduced since 2002. All of these programs and methods represent speed-ups relative to standard likelihood-based bootstrapping as implemented in PAUP* (Swofford, 2001). The

speed-ups are obtained through shortcuts during the heuristic searches and these shortcuts can be negatively determinate to the results by, for example, artificially inflating or deflating support values.

In a 4-taxon statement there are only three possible unrooted bifurcating topologies. Hence even programs that rely entirely upon NNI branch swapping (in this case only IQ-TREE) should have no difficulty in sampling all alternative bifurcating topologies, even when polytomies may cause NNI-based programs difficulty in $\geqslant$5-taxon statements (Whelan and Money, 2010). Our expectation is that, aside from those methods that always produce fully resolved trees (in this case GARLI with the "collapse 0" option, IQ-TREE, MrBayes [when not collapsing branches with posterior probability < 0.5], PhyML, and RAxML), heuristic tree-topology searches should not be a limitation in this study. We expect this to hold both during searches for the optimal tree topology for the matrix as a whole as well as within every resampling pseudoreplicate. Therefore, discrepancies among methods identified in this study should generally be restricted to factors other than topological search heuristics.

2. Methods

2.1. Simulated data matrices

Two types of matrices were simulated following Suzuki et al. (2002) – star matrices and conflict matrices. Matrices were simulated by using the Evolver program with MCbase.dat from the PAML ver. 4.1 package (Yang, 2007b). In both cases 15,000 characters were simulated for each of 100 replicate matrices using the Jukes and Cantor (1969) model, no rate heterogeneity among characters, and the four terminal branch lengths of 0.05. Characters were simulated using the JC model with no rate heterogeneity so as to minimize any potential confounding factors of model fit or differences between the parametric methods and parsimony.

The star matrices were simulated with a zero-length internal branch whereas the conflict matrices were simulated with a 0.005-length internal branch. Three separate simulations (one for each of the three unrooted topologies) of 5000 characters each were performed for the conflict matrices and then concatenated together. There are no missing or inapplicable cells in any matrix. The expected average number of substitutions simulated in the star matrices is 3000, and that for the conflict matrices is 3075. The actual average numbers of variable and parsimony-informative characters in the star matrices are 2701 and 65, while those for the conflict matrices are 2771 and 125.

2.2. Tree searches

Both the JC and GTR + Γ (Tavaré, 1986; Yang, 1993) models were applied for all Bayesian and likelihood methods to the degree possible (i.e., not including JC for the RAxML methods because the simplest nucleotide-substitution model implemented in RAxML is GTR + Γ). The rationale for using the JC model is that this is the true nucleotide-substitution model used to simulate the sequences. The rationale for using the GTR + Γ model for non-RAxML methods is to ensure that their results are directly comparable to the RAxML results. When the GTR + Γ model was applied, four discrete rate categories were used to approximate the estimated gamma distribution.

Unpartitioned analyses were performed for all Bayesian and likelihood methods. One thousand random-addition searches were performed to find the optimal tree(s) for the matrix as a whole and 1000 resampling pseudoreplicates were performed for all likelihood and parsimony methods. For those programs in which multiple tree searches could be performed in a given bootstrap pseudoreplicate (i.e., GARLI, PAUP*), only a single search was implemented to emulate those programs that only allow a single search. Two different seeds were used for all likelihood and parsimony methods: 123456 and 654321. Seeds are used as starting points for (pseudo-) random-number generators and different seeds may result in different starting trees for subsequent branch swapping (e.g., Swofford, 2001).

Bayesian analyses were performed using MrBayes ver. 3.1.2 (Ronquist and Huelsenbeck, 2003) rather than ver. 3.2.1 because of problems with the latter crashing when running on PCs. Four independent searches were performed that each consisted of 4 chains and 10,000,000 generations, with the first 5,000,000 generations discarded as burn-in. Convergence on selected matrices was confirmed by examining the potential scale reduction factor, which equaled or approached one for all parameters.

GARLI ver. 2.0.1019 analyses were alternately performed by requiring trees to be fully resolved ("collapsebranches = 0") and collapsing branches with a length of 1×10^{-8} (i.e., effectively zero; GARLI never presents branches with a length of zero; Zwickl, 2012) in the same manner that PAUP* collapses these branches (Swofford, 2001). In both cases the GARLI analyses were performed by using the lowest settings that Zwickl (2009) recommended for an intensive search (streefname = stepwise, attachmentspertaxon = 50, genthreshfortopoterm = 20,000, numberofprecreductions = 20, treerejectionthreshold = 100) for both the optimal-tree searches and every bootstrap pseudoreplicate.

IQ-TREE ver. 0.9.5 analyses were performed using "-n 1000", which applies 1000 iterations to both the optimal-tree search as well as every bootstrap pseudoreplicate. This is the only program for which multiple searches were performed within a given bootstrap pseudoreplicate.

PAUP* ver. 4.0b10 analyses were alternatively performed using likelihood and parsimony optimality criteria. In both cases tree-bisection-reconnection searches were applied, multiple equally optimal trees were allowed, and the strict consensus (Schuh and Polhemus, 1980) of the optimal trees (when applicable) was reported. PAUP* only implements the frequency-within-replicates rather than the strict-consensus bootstrap, so the bootstrap values are artificially inflated when multiple equally optimal trees for a given pseudoreplicate were found. The frequency-within-replicates bootstrap always provides at least as high a percentage for a given resolution as the strict-consensus bootstrap when equal-quality tree searches are applied (Davis et al., 1998).

PhyML ver. 3 – 20120412 (Guindon et al., 2010) analyses were performed using nucleotide frequencies estimated using likelihood (for GTR + Γ only), and subtree-pruning-regrafting branch swapping.

RAxML analyses were performed using four different versions of the program: 7.0.3, 7.2.6, 7.5.3, and 7.7.2. Multiple versions of the program were applied to check for consistency in the GTRCAT results among versions as well as the radically divergent bootstrap results obtained from ver. 7.5.3 with the GTRGAMMA model (not presented here but see raw data in supplemental online data posted at http://rydberg.biology.colostate.edu/Research/ as well as posts from June 21, July 1, 31, and August 6, 2013 at https://groups.google.com/forum/#!forum/raxml).

Both the GTRGAMMA and GTRCAT models were applied in alternative RAxML analyses. Most GTRGAMMA analyses were performed using the standard bootstrap with the "-b" command, whereas most GTRCAT analyses were performed using the rapid bootstrap with the "-f a" command. An additional set of analyses were performed in ver. 7.2.6 using GTRGAMMA with the rapid bootstrap and GTRCAT with the standard bootstrap.

In addition to bootstrap resampling, the SH-like aLRT was also applied in RAxML ver. 7.7.2 using both the GTRCAT and

GTRGAMMA models with the "-f J" command and uploading the optimal tree found from the 123456-seed tree search.

2.3. 100-seed analyses

In addition to the 100-replicate-matrix analyses described in Sections 2.1 and 2.2, a separate set of analyses were performed on the first five replicate matrices for selected methods (GARLI collapse 0 and collapse 1, PAUP* likelihood and parsimony, PhyML, RAxML 7.2.6 and 7.7.2 for both GTRCAT and GTRGAMMA) that applied bootstrap resampling. Each of these five replicate matrices was analyzed by the selected methods using 100 different seeds between 1 and 1 billion generated by http://www.random.org/ (Appendix S1). In addition to testing Hedges' (1992) and Pattengale et al.'s (2009, 2010) assertions regarding bootstrap confidence intervals, these analyses also served to test whether the two arbitrarily chosen seeds (123456 and 654321) used in the analyses described in Section 2.2 were determinate to those results.

2.4. Hypotheses and statistical analyses

We performed six sets of statistical analyses. In all except the sixth set the analyses were performed identically but independently for the conflict and star matrices. As a conservative approach the statistical analyses did not take into account topological differences, when applicable. As described in Section 3.1 below, only a limited number of incongruent topologies were identified across seeds or among methods. Given the general congruence in results between the 123456 and 654321 seeds within methods (the exception being RAxML ver. 7.5.3), only results from the 123456 seed were used when comparing support values among methods.

For the first set of analyses we tested for significant differences in bootstrap values and posterior probabilities across the different methods, using the GTR + Γ (and GTRCAT for RAxML) results. We focused on these results (rather than the JC results) because they are universal across all Bayesian and likelihood methods, whereas there are no JC results for RAxML. Our three a priori hypotheses for this first set of analyses were as follows. First, Bayesian posterior probabilities, as previously documented by Suzuki et al. (2002) for these 4-taxon, 15,000-character, polytomy simulations, would, on average, be significantly higher than branch supports generated by any of the bootstrapping methods. Second, those bootstrapping methods that are only capable of outputting fully resolved trees (GARLI collapse 0, IQ-TREE, PhyML, RAxML) would, on average, provide significantly higher support values than those methods that do not (GARLI collapse 1, PAUP* likelihood, PAUP* parsimony). Third, we hypothesized that there would be no significant differences, on average, between bootstrap values generated by methods with different tree-search implementations (i.e., RAxML rapid bootstrap vs. RAxML standard bootstrap; IQ-TREE, which only implements NNI branch swapping, vs. methods from other programs that implement SPR or TBR swapping) because there are only three possible bifurcating tree topologies for these unrooted four-taxon statements and every method should examine all three topologies.

For the second set of analyses we tested for significant differences between bootstrap values and SH-like aLRT values within each program (IQ-TREE, PhyML, and RAxML) and model (GTR + Γ and JC). We sought to test Guindon et al.'s (2010) theory that the SH-like aLRT would outperform the bootstrap when applied to polytomies. Hence the SH-like aLRT support values should, on average, be lower than the directly comparable bootstrap values.

Third, we tested for significant differences between the GTR + Γ and the JC results within each method. These tests are only applicable for IQ-TREE, MrBayes, PAUP* likelihood, and PhyML. Our a priori hypothesis was that overparameterization with GTR + Γ would not cause any significant differences relative to JC given that the GTR + Γ model includes the JC model and that our simulations include large numbers of characters such that power should not be a limiting factor. Of any minor differences that might be observed we expected the GTR + Γ model to provide slightly lower support values because of the overparameterization (Buckley and Cunningham, 2002).

The first three sets of analyses were performed by examining the effect of method on support levels for the 100-replicate matrices using a mixed-model analysis of variance. In this model, method was treated as a fixed effect and matrix replicate was included as a random effect. Means were separated using a Student's *t*-test. These analyses were performed in JMP® ver. 9.0.2 (SAS Institute, 2007).

Fourth, we tested for significant differences in branch-support values generated using different seeds by the same method. We performed these tests to identify any methods (and programs) that are particularly sensitive to which seed is specified. Our a priori hypothesis was that only those methods that implement particularly low quality heuristic searches (i.e., IQ-TREE UFBoot, which applies resampling-estimated log-likelihoods, and RAxML rapid bootstrapping with GTRCAT, in which the bootstrap pseudoreplicates are not independent) would show any significant differences.

For this fourth set of analyses we first examined the effect of the two alternative seeds (123456 and 654321) in the 100-replicate matrices using a mixed-model analysis of variance. For each method, we performed an ANOVA in JMP® to determine the influence of seed (as a fixed effect) and matrix (as a random effect) on support levels. These tests examined whether one seed consistently produced higher or lower support levels than the other seed across the replicate matrices. We also examined the effect of seed on support levels for the five replicate matrices for which 100 different random seeds were applied by using the same approach.

The fifth set of analyses are focused on false positives. Suzuki et al. (2002) used their 4-taxon polytomy simulations to test whether the null hypothesis for polytomies (i.e., all three topologies should occur with equal frequency when only fully resolved trees are allowed) is rejected more often than expected (i.e., an inflated false-positive rate) based on Bayesian posterior probabilities. We use the same approach here but instead focus on likelihood-based bootstrap values in the context of Hillis and Bull's (1993) widely cited assertion that $\geqslant$70% bootstrap support can be interpreted as $a \geqslant 0.95$ probability that the true clade is resolved. We tested whether false positives at the $P = 0.05$ level are better accounted for by each method using a 70% or 95% cutoff. Both Zharkikh and Li's (1992) and Felsenstein and Kishino's (1993) investigations into polytomies indicate that the 95% cutoff should be more appropriate.

For each method we recorded the number of false positives at the $P = 0.05$ level from the 100 matrix replicates (averaged across the two seeds). We then calculated the Wilson-score confidence interval (based on 100 matrix replicates) for the recorded frequency. If the true false positive rate for a given method is outside of this interval then a frequency as extreme as that observed will occur only 2.5% of the time (Wilson, 1927). A qualification is that we did not adjust the 95% confidence intervals to account for multiple tests.

Sixth, we tested Hedges' (1992), Pattengale et al.'s (2009, 2010) theories regarding precision of bootstrap values (i.e., based strictly on the binomial distribution as asserted by Hedges (1992), or greater precision that predicted by the binomial distribution as asserted by Pattengale et al. (2009, 2010)). We were also interested in whether or not the precision would vary among different methods given the same number of bootstrap pseudoreplicates (1000). Our a priori hypothesis was that those methods that implement

particularly low quality heuristic searches (i.e., IQ-TREE UFBoot [not tested] and RAxML rapid bootstrapping with GTRCAT) would have significantly lower precision than the other methods sampled.

We performed an analysis of covariance from both the conflict and star simulations using the 100-seed results to examine whether the relationship between Hedges' (1992) predicted values of the standard error and those empirically derived from the 100 different seeds was (a) the same across methods and (b) significantly different than unity (i.e., expected = observed). For each method there are five matrix replicates for both the star and conflict simulations. This model contained method, empirical standard deviation, and their interaction as fixed effects with the response variable of observed standard deviation. Empirically derived standard deviations were calculated as $\sqrt{(pq/n)}$, where p = the average bootstrap value and n = 1000 bootstrap pseudoreplicates. Observed standard deviations were calculated over the 100-seed analyses for each method and simulated data set.

3. Results

3.1. Topological incongruence and unresolved trees

Only a relatively few cases of topological incongruence among the methods were observed, and in all but one case the topological outlier received $\leqslant$ 52% support (Table S2). There was widespread incongruence among methods for two replicates from both the conflict and star matrices. The only cases of topological incongruence among different seeds (generally just 123456 and 654321) were for IQ-TREE, for which two cases were identified with the GTR + Γ model and four cases were identified with the JC model for the conflict matrices, and three cases with the GTR + Γ model and one case with the JC model for the star matrices.

Limited incongruence was also observed between the same method with different models (i.e., GTR + Γ and JC) for the Bayesian (one replicate from both the conflict and star matrices), PAUP* likelihood (one replicate from the conflict matrix; also three replicates from the star matrices for which only the GTR + Γ-based trees were resolved), and PhyML (one replicate from the conflict matrices and three replicates from the star matrices) methods.

No cases of incongruence were observed between all methods from the different optimality criteria (i.e., Bayesian posterior probability, likelihood, and parsimony). But there were two cases among the conflict matrices and four cases among the star matrices for which parsimony produced polytomies whereas other methods produced resolution. The only other methods that produced polytomies were PAUP* likelihood and GARLI collapse 1, and then only for some star matrices. Of the 11 replicates for which GARLI collapse 1 produced a polytomy, GARLI collapse 0 provided an average of 51.6% bootstrap, and a maximum of 94.5% (averaged between the two seeds).

Seven cases of incongruence were observed among the different versions of RAxML as well as the different methods implemented therein. All of them were restricted to versions 7.0.3 and/or 7.2.6 and four of the seven were limited to ver. 7.0.3 with the GTRGAMMA model.

The remaining cases of topological incongruence did not have a clear pattern among the methods (Table S2). No cases of incongruence were observed among the different seeds used by a method for a given matrix in the 100-seed analyses.

3.2. Results of statistical analyses

Our first set of analyses focused on testing for significant differences in bootstrap values and posterior probabilities across the different methods, using the GTR + Γ (and GTRCAT for RAxML) models. The results for the conflict matrices are presented in Table 1 and the results for the star matrices are presented in Table 2. For the conflict matrices, RAxML rapid bootstrapping with GTRCAT provided significantly higher support than did all other methods, followed by MrBayes and then all other applicable methods. No significant within-program differences were identified for anything other than RAxML.

RAxML rapid bootstrapping with GTRCAT also provided significantly higher support than did all other applicable methods (including RAxML rapid bootstrapping with GTR + Γ and RAxML standard bootstrapping with GTRCAT) for the star matrices. But in contrast to the conflict matrices, MrBayes provided significantly lower support than did GARLI collapse 0, IQ-TREE bootstrap, PhyML, and RAxML GTR + Γ. Only GARLI collapse 1, IQ-TREE UFBoot, and PAUP* bootstrap GTR + Γ provided significantly lower support than did MrBayes. Significant within-program differences were identified for GARLI (collapse 0 > collapse 1), IQ-TREE (bootstrap > UFBoot), and PAUP* (parsimony > GTR + Γ).

Our second and third sets of analyses focused on testing for significant differences between bootstrap values and SH-like aLRT values within each program (IQ-TREE, PhyML, and RAxML) and model (GTR + Γ and JC) as well as between the two models In all cases, SH-like aLRT values were, on average, significantly lower than bootstrap values within all three programs, both models, and both sets of matrices. Alternatively, in no cases were GTR + Γ bootstrap, posterior probability, or SH-like aLRT values significantly different from JC-based values within the same program.

Our fourth set of analyses tested for significant differences in branch-support values generated using different seeds by the same method. A total of 58 ANOVAs were performed and therefore it is not surprising that four tests produced marginally significant results (Table S3). The only significant results were for IQ-TREE bootstrap (with both the GTR + Γ and JC models) when applied to the star matrices. But different seeds do not appear to be a general problem for IQ-TREE because none of the eight tests applied to UFBoot or the aLRT were even close to significant.

Our fifth set of analyses focused on whether false positives at the P = 0.05 level are better accounted for by each method using a 70% or 95% bootstrap cutoff. For the sake of thoroughness we also extended these tests to include Bayesian posterior probabilities and SH-like aLRTs (Tables 1 and 2). The 95% confidence interval for the number of replicates with $\geqslant$95% support overlapped the 5% error rate for 21 of the 27 methods for the conflict matrices and 15 of the 27 methods for the star matrices. In contrast, the 95% confidence interval for the number of replicates with $\geqslant$70% support did not overlap with the 5% error rate for any of the 27 methods for either the conflict or star matrices.

Those methods with a liberal frequency of false positives at the P = 0.05 level are MrBayes (conflict matrices only) and RAxML rapid bootstrapping with GTRCAT (both conflict and star matrices). Alternatively, those methods with a conservative frequency of false positives at the P = 0.05 level are IQTREE-UFBoot and IQ-TREE aLRT (star matrices only), PhyML bootstrap and SH-like aLRT (three of the four cases for star matrices only), and RAxML SH-like aLRT with the GTRCAT model (and one case for the GTRGAMMA model).

Our sixth set of analyses focused on whether the bootstrap procedures implemented in each method performed as expected and are consistent with a binomial error distribution as asserted by Hedges (1992). With the exception of rapid bootstrapping with the GTRCAT model in RAxML ver. 7.2.6 and 7.7.2, the slope of the relationship between predicted standard errors for the binomial distribution and those generated by the methods was not significantly different from unity (Table 3). There was no evidence that any of these methods performed any differently than expected from the binomial error distribution. But RAxML ver. 7.2.6 and

Table 1
Results for the conflict matrices averaged across the two seeds (when applicable).

Method	Avg. %	Stat. group[a]	Min. %	Max. %	#⩾95% (95% CI)[b]	#⩾70% (95% CI)[b]
GARLI collapse 0 GTR + Γ	61.8	C	35	100	4 (1.6–9.8)	25.5 (18.0–34.8)
GARLI collapse 1 GTR + Γ	61.8	C	35	100	4 (1.6–9.8)	25.5 (18.0–34.8)
IQ-TREE bootstrap GTR + Γ	61.9	C	37	100	4 (1.6–9.8)	26.5 (18.8–35.9)
IQ-TREE bootstrap JC	61.7	C	37	100	4 (1.6–9.8)	28 (20.1–37.5)
IQ-TREE UFBoot GTR + Γ	62.8	C	37.5	100	4 (1.6–9.8)	26 (18.4–35.4)
IQ-TREE UFBoot JC	62.6	C	37.5	99.5	4 (1.6–9.8)	28 (20.1–37.5)
IQ-TREE aLRT GTR + Γ	49.3	D, E	2	100	7 (3.4–13.7)	26.5 (18.8–35.9)
IQ-TREE aLRT JC	48.5	D, E	1.5	100	7 (3.4–13.7)	26 (18.4–35.4)
PhyML bootstrap GTR + Γ	60.9	C	30	100	4 (1.6–9.8)	24 (16.7–33.2)
PhyML bootstrap JC	60.9	C	35	100	3.5 (1.3–9.2)	25.5 (18.0–34.8)
PhyML aLRT GTR + Γ	48.1	D, E	1	100	8 (4.1–15.0)	25 (17.5–34.3)
PhyML aLRT JC	47.7	E	0	100	7 (3.4–13.7)	24 (16.7–33.2)
RAxML 7.7.2 SBS GTR + Γ	61.8	C	35	100	4 (1.6–9.8)	27.5 (19.7–37.0)
RAxML 7.2.6 SBS GTR + Γ	61.8	C	35	100	4 (1.6–9.8)	27.5 (19.7–37.0)
RAxML 7.0.3 SBS GTR + Γ	61.9	C	32.5	100	3.5 (1.3–9.2)	27 (19.3–36.4)
RAxML 7.7.2 RBS GTRCAT	98	A	0	100	97 (91.5–99.0)	97 (91.5–99.0)
RAxML 7.2.6 RBS GTRCAT	98.4	A	0	100	97.5 (92.3–99.2)	98 (93.0–99.4)
RAxML 7.0.3 RBS GTRCAT	98	A	0	100	97 (91.5–99.0)	97 (91.5–99.0)
RAxML 7.2.6 RBS GTR + Γ	61.9	C	33.5	100	3.5 (1.3–9.2)	27.5 (19.7–37.0)
RAxML 7.2.6 SBS GTRCAT	62	C	32	100	4 (1.6–9.8)	27 (19.3–36.4)
RAxML 7.7.2 aLRT GTR + Γ	50.3	D	0	100	5 (2.2–11.2)	30 (21.9–39.6)
RAxML 7.7.2 aLRT GTRCAT	0	F	0	0	0 (0–3.7)	0 (0–3.7)
PAUP* bootstrap GTR + Γ	62	C	36	100	4 (1.6–9.8)	27 (19.3–36.4)
PAUP* bootstrap JC	61.9	C	36.5	100	4.5 (1.9–10.5)	27 (19.3–36.4)
PAUP* bootstrap parsimony	61.1	C	0	100	4 (1.6–9.8)	28.5 (20.6–38.0)
MrBayes GTR + Γ	82.9	B	44	100	32 (23.7–41.7)	76 (66.8–83.3)
MrBayes JC	85.5	B	45	100	42 (32.8–51.8)	84 (75.6–89.9)

[a] Least squared means not connected by the same letter are significantly different.
[b] 95% confidence interval for the 100 replicate matrices.

Table 2
Results for the star matrices averaged across the two seeds (when applicable).

Method	Avg. %	Stat. group[a]	Min. %	Max. %	#⩾95% (95% CI)[b]	#⩾70% (95% CI)[b]
GARLI collapse 0 GTR + Γ	64.8	B, C	33.5	95.5	3.5 (1.3–9.2)	36.5 (27.7–46.3)
GARLI collapse 1 GTR + Γ	54.1	G	0	95	1 (0.2–5.4)	24 (16.7–33.2)
IQ-TREE bootstrap GTR + Γ	64.9	B, C	36	96	5 (2.2–11.2)	37.5 (28.6–47.3)
IQ-TREE bootstrap JC	65.1	B, C	36	95.5	5 (2.2–11.2)	37.5 (28.6–47.3)
IQ-TREE UFBoot GTR + Γ	55.7	G	36.5	86.5	0 (0–3.7)	12 (7.0–19.8)
IQ-TREE UFBoot JC	56.6	F, G	35	87.5	0 (0–3.7)	14.5 (8.9–22.7)
IQ-TREE aLRT GTR + Γ	41	H, I	0	90.5	0 (0–3.7)	23.5 (16.3–32.7)
IQ-TREE aLRT JC	43	H	0	91	0 (0–3.7)	26 (18.4–35.4)
PhyML bootstrap GTR + Γ	62.8	C, D, E	35.5	94	0 (0–3.7)	30 (21.9–39.6)
PhyML bootstrap JC	63.5	C, D	34	95	1 (0.2–5.4)	31 (22.8–40.6)
PhyML aLRT GTR + Γ	35.9	J	0	94	0 (0–3.7)	21 (14.2–30.0)
PhyML aLRT JC	37.7	I, J	0	94	0 (0–3.7)	24 (16.7–33.2)
RAxML 7.7.2 SBS GTR + Γ	65	B, C	32.5	96	3 (1.0–8.5)	37 (28.2–46.8)
RAxML 7.2.6 SBS GTR + Γ	65	B, C	32.5	96	3 (1.0–8.5)	37 (28.2–46.8)
RAxML 7.0.3 SBS GTR + Γ	65	B, C	32.5	96	3.5 (1.3–9.2)	37.5 (28.6–47.3)
RAxML 7.7.2 RBS GTRCAT	95.9	A	25	100	93 (86.3–96.6)	93 (86.3–96.6)
RAxML 7.2.6 RBS GTRCAT	94.9	A	0	100	92 (85.0–95.9)	92 (85.0–95.9)
RAxML 7.0.3 RBS GTRCAT	95.1	A	0	100	92 (85.0–95.9)	92 (85.0–95.9)
RAxML 7.2.6 RBS GTR + Γ	68.1	B	34.5	100	5.5 (2.5–11.8)	45.5 (36.1–55.2)
RAxML 7.2.6 SBS GTRCAT	65.1	B, C	32.5	96.5	2.5 (0.8–7.7)	38.5 (29.6–48.3)
RAxML 7.7.2 aLRT GTR + Γ	39.6	H, I	0	91	0 (0–3.7)	26 (18.4–35.4)
RAxML 7.7.2 aLRT GTRCAT	0	K	0	0	0 (0–3.7)	0 (0–3.7)
PAUP* bootstrap GTR + Γ	54.1	G	0	95.5	1 (0.2–5.4)	24.5 (17.1–33.8)
PAUP* bootstrap JC	56.7	F, G	0	95.5	1.5 (0.3–6.2)	26.5 (18.8–35.9)
PAUP* bootstrap parsimony	62.3	C, D, E	0	96.5	3 (1.0–8.5)	38 (29.1–47.8)
MrBayes GTR + Γ	59.6	E, F	34	98	4 (1.6–9.8)	25 (17.5–34.3)
MrBayes JC	60.3	D, E	35	98	4 (1.6–9.8)	26 (18.4–35.4)

[a] Least squared means not connected by the same letter are significantly different.
[b] 95% confidence interval for the 100 replicate matrices.

Table 3
Regression statistics for the relationship between the theoretical and empirical standard deviations for the sixth set of analyses using the 100-seed analyses of the first five conflict matrices and the first five star matrices.

Method	Term [a]	Estimate (95% CI)
GARLI collapse 0 GTR + Γ	Intercept	0.10 (−0.07–0.26)
GARLI collapse 0 GTR + Γ	Slope	0.95 (0.83–1.07)
GARLI collapse 1 GTR + Γ	Intercept	0.02 (−0.08–0.12)
GARLI collapse 1 GTR + Γ	Slope	1.01 (0.93–1.08)
PhyML bootstrap GTR + Γ	Intercept	0.18 (−0.27–0.62)
PhyML bootstrap GTR + Γ	Slope	0.88 (0.57–1.20)
RAxML 7.7.2 SBS GTR + Γ	Intercept	0.27 (−0.16–0.70)
RAxML 7.7.2 SBS GTR + Γ	Slope	0.88 (0.57–1.18)
RAxML 7.2.6 SBS GTR + Γ	Intercept	0.26 (−0.18–0.70)
RAxML 7.2.6 SBS GTR + Γ	Slope	0.88 (0.56–1.19)
RAxML 7.7.2 RBS GTRCAT	Intercept	0.00 (0.00–0.00)
RAxML 7.7.2 RBS GTRCAT	Slope	**2.97 (2.97–2.97)**
RAxML 7.2.6 RBS GTRCAT	Intercept	0.00 (0.00–0.00)
RAxML 7.2.6 RBS GTRCAT	Slope	**2.97 (2.97–2.97)**
PAUP* bootstrap GTR + Γ	Intercept	0.03 (−0.17–0.23)
PAUP* bootstrap GTR + Γ	Slope	1.00 (0.85–1.16)
PAUP* bootstrap parsimony	Intercept	0.01 (−0.14–0.16)
PAUP* bootstrap parsimony	Slope	1.01 (0.89–1.12)

[a] A perfect correspondence between theoretical (calculated from a binomial distribution with 1000 trials) and empirical standard deviation (calculated from the results of 100 random-seed trials for each of 10 data matrices) would yield an intercept of 0 and a slope of 1. Entries in bold have parameter values whose confidence interval does not overlap these values.

7.7.2 rapid bootstrapping with GTRCAT analyses both reported 100% bootstrap for nine of the 10 matrix replicates (Tables 4 and 5) and just 38% bootstrap for one matrix replicate (the second matrix replicate from the star simulations). The observed standard error for that last matrix replicate was about three times that expected from the binomial error distribution (Table 5).

4. Discussion

4.1. Bootstrap values and posterior probabilities among methods

With respect to our three a priori hypotheses regarding bootstrap values and posterior probabilities using the GTR + Γ (and GTRCAT) results, our first set of statistical analyses supported Suzuki et al.'s (2002) demonstration that Bayesian posterior probabilities are inflated relative to bootstrap values – but only for the conflict matrices. These analyses also supported our second hypothesis that bootstrapping methods which are only capable of outputting fully resolved trees would, on average, provide significantly higher support values than those methods that do not – but only for the star matrices and with two exceptions (for IQ-TREE UFBoot GTR + Γ and PAUP* bootstrap parsimony). Our third hypothesis that there would be no significant differences, on average, between bootstrap values generated by methods with different tree-search implementations was clearly refuted for RAxML rapid bootstrapping with the GTRCAT model, but not for IQ-TREE, which only implements NNI swapping.

It is clear that Bayesian posterior probabilities are not the only means of quantifying support that can produce dramatically inflated values when applied to cases of strong character conflict. RAxML rapid bootstrapping with the GTRCAT model produced significantly higher support values than all other methods when applied to the conflict matrices, providing an average of $\geqslant$98% support for a single bifurcating tree compared to an average posterior probability of 82.9 and an average of $\leqslant$62.8% support by all other methods (Table 1). A similar result was obtained for the star matrices, wherein RAxML rapid bootstrapping with the GTRCAT model provided significantly higher support values than all other

Table 4
Averages and standard deviations for the 100-seed analyses of the first five conflict matrices. The second standard deviation (in bold) is the value predicted from 1,000 bootstrap pseudoreplicates based on the binomial distribution.

Method	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5
GARLI collapse 0 GTR + Γ	53.4 (1.62; **1.58**)	55.2 (1.60; **1.57**)	81.9 (1.32; **1.22**)	92.9 (0.82; **0.81**)	43.7 (1.60; **1.57**)
GARLI collapse 1 GTR + Γ	53.4 (1.62; **1.58**)	55.2 (1.60; **1.57**)	81.9 (1.32; **1.22**)	92.9 (0.82; **0.81**)	43.7 (1.60; **1.57**)
PhyML bootstrap GTR + Γ	53.9 (1.35; **1.58**)	50.1 (1.58; **1.58**)	81.2 (1.33; **1.23**)	92.2 (0.83; **0.85**)	43.0 (1.76; **1.57**)
RAxML 7.7.2 SBS GTR + Γ	52.1 (1.78; **1.58**)	57.5 (1.67; **1.56**)	82.2 (1.21; **1.21**)	92.8 (1.03; **0.82**)	41.5 (1.63; **1.56**)
RAxML 7.2.6 SBS GTR + Γ	52.1 (1.77; **1.58**)	57.5 (1.67; **1.56**)	82.2 (1.21; **1.21**)	92.8 (1.03; **0.82**)	41.5 (1.63; **1.56**)
RAxML 7.7.2 RBS GTRCAT	100 (0; **0**)	100 (0; **0**)	100 (0; **0**)	100 (0; **0**)	100 (0; **0**)
RAxML 7.2.6 RBS GTRCAT	100 (0; **0**)	100 (0; **0**)	100 (0; **0**)	100 (0; **0**)	100 (0; **0**)
PAUP* bootstrap GTR + Γ	53.1 (1.60; **1.58**)	55.1 (1.43; **1.57**)	81.9 (1.20; **1.22**)	92.9 (0.88; **0.81**)	43.0 (1.84; **1.57**)
PAUP* bootstrap parsimony	50.7 (1.55; **1.58**)	52.3 (1.52; **1.58**)	86.9 (1.06; **1.07**)	90.7 (0.93; **0.92**)	37.9 (1.73; **1.53**)

Table 5
Averages and standard deviations for the 100-seed analyses of the first five star matrices. The second standard deviation (in bold) is the value predicted from 1,000 bootstrap pseudoreplicates based on the binomial distribution.

Method	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5
GARLI collapse 0 GTR + Γ	49.8 (1.59; **1.58**)	56.9 (1.48; **1.57**)	88.6 (1.06; **1.01**)	78.7 (1.36; **1.30**)	70.3 (1.48; **1.44**)
GARLI collapse 1 GTR + Γ	0 (0; **0**)	42.4 (1.49; **1.56**)	88.4 (1.08; **1.01**)	70.9 (1.55; **1.44**)	68.0 (1.47; **1.47**)
PhyML bootstrap GTR + Γ	45.4 (1.53; **1.57**)	62.0 (1.53; **1.53**)	86.5 (1.20; **1.08**)	73.3 (1.49; **1.40**)	67.5 (1.41; **1.48**)
RAxML 7.7.2 SBS GTR + Γ	51.6 (1.81; **1.58**)	51.4 (1.62; **1.58**)	88.7 (1.27; **1.00**)	77.7 (1.29; **1.32**)	71.5 (1.31; **1.43**)
RAxML 7.2.6 SBS GTR + Γ	51.6 (1.81; **1.58**)	51.4 (1.62; **1.58**)	88.7 (1.26; **1.00**)	77.7 (1.29; **1.32**)	71.5 (1.28; **1.43**)
RAxML 7.7.2 RBS GTRCAT	100 (0; **0**)	38.0 (4.55; **1.53**)	100 (0; **0**)	100 (0; **0**)	100 (0; **0**)
RAxML 7.2.6 RBS GTRCAT	100 (0; **0**)	38.0 (4.55; **1.53**)	100 (0; **0**)	100 (0; **0**)	100 (0; **0**)
PAUP* bootstrap GTR + Γ	0 (0; **0**)	42.6 (1.53; **1.56**)	88.1 (1.13; **1.02**)	70.9 (1.41; **1.44**)	68.0 (1.60; **1.48**)
PAUP* bootstrap parsimony	52.2 (1.63; **1.58**)	0 (0; **0**)	87.3 (1.16; **1.05**)	78.1 (1.26; **1.31**)	72.7 (1.35; **1.41**)

methods, providing an average of ⩾94.9% support for a bifurcating tree compared to an average posterior probability of just 0.596 and an average of ⩽68.1% support by all other methods (Table 2).

Strikingly, for the conflict matrices and to a lesser degree the star matrices, the other methods of quantifying bootstrap support implemented in RAxML produced roughly similar values to those generated by GARLI, IQ-TREE, PhyML, and PAUP* (Tables 1 and 2). The dramatic increase in support values generated by RAxML rapid bootstrapping with the GTRCAT model cannot be attributed entirely to either rapid bootstrapping or the GTRCAT model alone, as demonstrated by the RAxML rapid-bootstrapping-with-the-GTR + Γ-model and the standard-bootstrapping-with-the-GTRCAT-model results.

Inflated bootstrap values generated by RAxML rapid bootstrapping with the GTRCAT model have previously been pointed out by Siddall (2010) and Simmons and Norton (2013). Based on these reports, which have been generated from both empirical and now simulated data, we recommend avoidance of rapid bootstrapping with the GTRCAT model in RAxML. If these results can be attributed to a coding error within RAxML, then that coding error has affected hundreds of empirical studies based on the following two factors. First, our RAxML GTRCAT results were consistent across RAxML vers. 7.03, 7.2.6, and 7.7.2. Second, Stamatakis et al. (2008), wherein rapid bootstrapping with the GTRCAT model was introduced, has been cited 1,856 times in Web of Science® as of 9 January 2014.

Aside from the RAxML-rapid-bootstrapping-with-the-GTRCAT-model results, our first set of analyses also demonstrate the advantage of allowing (effectively) zero-length branches to be collapsed when applied to lineages with polytomies. Indeed, the two likelihood methods that allowed effectively zero-length branches to be collapsed (GARLI collapse 1 and PAUP* bootstrap GTR + Γ) provided significantly lower bootstrap support than did all but one likelihood method (IQ-TREE UFBoot GTR + Γ) that only ever produce a single fully resolved optimal tree for the star matrices (Table 2). Based on this result, which reinforces the empirically derived result of Simmons and Norton (2013), we recommend that the default approach of collapsing effectively zero-length branches, as in GARLI and PAUP*, be extended to all other likelihood-inference programs as well. Even if doing so will not provide the same improvement in performance as allowing a strict consensus of all optimal trees to be generated (Simmons and Goloboff, 2013), it will at least be an improvement that requires only minimal additional computational time.

4.2. Bootstrap vs. SH-like aLRT

Our results clearly support Guindon et al.'s (2010) theory that the SH-like aLRT outperforms the bootstrap when applied to polytomies because in all cases, SH-like aLRT values were, on average, significantly lower than bootstrap values within all three programs (IQ-TREE, PhyML, and RAxML), both models (GTR + Γ and JC), and both sets of matrices (conflict and star; Tables 1 and 2). However, with the exception of RAxML rapid bootstrapping with GTRCAT, none of the bootstrap methods had an elevated false-positive rate based on the ⩾70% and ⩾95% cutoffs. For many replicates (particularly for the star matrices), the SH-like aLRT values are either zero or in the single digits (see raw data in supplemental online data posted at http://rydberg.biology.colostate.edu/Research/), which is a good indication that robust evolutionary and taxonomic inferences cannot be made from those branches.

Two additional advantages of the SH-like aLRT over the bootstrap is its quick running time and that it is already implemented in programs that only ever hold a single fully resolved tree (i.e., IQ-TREE, PhyML, and RAxML), for which artificially inflated bootstrap values are a concern (particularly in supermatrices with high percentages of non-randomly distributed missing data; Lemmon et al., 2009; Simmons, 2012a, 2012b). Based on those two advantages and Guindon et al.'s (2010) and our results, we suggest that the SH-like aLRT be widely applied to likelihood-based empirical studies to complement the bootstrap by collapsing those branches with an SH-like aLRT percentage of ⩽10, irrespective of how high the likelihood bootstrap support is. But note that this is not a universal endorsement of SH-like aLRT support values over those generated by rigorously applied resampling methods, and SH-like aLRTs cannot necessarily be relied upon to collapse all properly unsupported branches. In particular, the NNI-based comparisons performed by the SH-like aLRT are unproblematic in the 4-terminal simulations performed here, but may be a limitation in trees with many more terminals.

4.3. Models and seeds

Our a priori hypothesis that overparameterization with GTR + Γ would not cause any significant differences relative to JC for our matrices was corroborated given that no cases were identified wherein GTR + Γ bootstrap, posterior probability, or SH-like aLRT values were significantly different from JC-based values within the same program (IQ-TREE, MrBayes, PAUP* likelihood, and PhyML). From this we conclude that the divergent RAxML-rapid-bootstrapping-with-GTRCAT results identified in our first and sixth set of statistical analyses cannot be ascribed to overparameterization relative to the JC model used to simulate the data.

With respect to the two different seeds (123456 and 654321) used for the conflict and star matrices, our a priori hypothesis that only IQ-TREE UFBoot and RAxML rapid bootstrapping with GTRCAT would show any significant differences was not supported. Instead the only significant differences identified between these two seeds were for IQ-TREE with regular bootstrapping when applied to the star matrices. Dramatic differences were identified for RAxML ver. 7.5.3 results (see raw data in supplemental online data posted at http://rydberg.biology.colostate.edu/Research/), but that coding error has since been fixed in RAxML ver. 7.7.2.

4.4. 70% bootstrap ≠ ⩾0.95 probability of accuracy

In no case did the 95% confidence interval for the 70% bootstrap cutoff overlap with the 5% error rate; rather, their error rates were higher – often much higher (Tables 1 and 2). In contrast, the 95% confidence interval for the ⩾95% bootstrap cutoff overlapped with the 5% error rate for most bootstrapping methods in both the conflict and star matrices (the exceptions being RAxML rapid bootstrapping with the GTRCAT model for both sets of matrices, and IQ-TREE UFBoot and PhyML GTR + Γ for the star matrices only). These results corroborate Zharkikh and Li's (1992) and Felsenstein and Kishino's (1993) investigations into polytomies, wherein they found that the 95% bootstrap cutoff should be more appropriate than the 70% cutoff, and enables their findings to be generalized to most current implementations of bootstrapping that apply the likelihood optimality criterion.

Note that several factors in our study favor obtaining accurate bootstrap values: numerous characters with a slow rate of evolution, no missing data, no nucleotide-model violation, and simple tree searches. As such, the bootstrap values obtained here (other than the extremes found for RAxML rapid bootstrapping with the GTRCAT model) are likely conservative relative to empirical analyses that have underparameterized models, numerous terminals, and missing and inapplicable data – all of which can contribute to inflated bootstrap support (Erixon et al., 2003; Simmons and Freudenstein, 2011; Simmons, 2012a, 2012b; Simmons and Goloboff, 2013). Based on that information, together with results from Huelsenbeck and Rannala (2004) and the criticisms of

interpreting bootstrap values as a measure of accuracy (Brown, 1994; Holmes, 2003; Anisimova et al., 2011), we suggest that the idea that ⩾70% bootstrap generally equates to 95% probability of accuracy in empirical analyses finally be abandoned.

4.5. Precision of bootstrap values

The results from our sixth set of analyses on the 100-seed results are almost entirely consistent with Hedges' (1992) assertion that the precision of bootstrap values follows the binomial distribution (Tables 3–5). Our a priori hypothesis that RAxML rapid bootstrapping with GTRCAT would have significantly lower precision than the other methods sampled was corroborated; the observed standard error for the second matrix replicate from the star simulations was about three times that expected from the binomial error distribution (Table 5). This result reinforces our earlier recommendation that investigators avoid implementing rapid bootstrapping with the GTRCAT model in RAxML and also further calls into question bootstrap values reported by published empirical studies that have implemented this method. These dubious results generated by RAxML rapid bootstrapping with the GTRCAT model are of particular concern for those empirical studies that implemented << 1,000 bootstrap pseudoreplicates and relied entirely upon RAxML for phylogenetic inference (e.g., Hedtke et al., 2013; Hinchliff and Roalson, 2013; Soltis et al., 2013).

Acknowledgments

We thank an anonymous reviewer for suggestions with which to improve the manuscript; Jerry Davis and Pablo Goloboff for helpful discussions; and Andre Aberer and Alexandros Stamatakis for clarifications regarding RAxML.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ympev.2014.01.018.

References

Anisimova, M., Gascuel, O., 2006. Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.
Anisimova, M., Gil, M., Dufayard, J.-F., Dessimoz, C., Gascuel, O., 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Syst. Biol. 60, 685–699.
Antiabong, J.F., Boardman, W., Smith, I., Brown, M.H., Ball, A.S., Goodman, A.E., 2013. A molecular survey of a captive wallaby population for periodontopathogens and the co-incidence of *Fusobacterium necrophorum* subspecies *necrophorum* with periodontal diseases. Vet. Microbiol. 163, 335–343.
Bacon, C.D., Michonneau, F., Henderson, A.J., McKenna, M.J., Milroy, A.M., Simmons, M.P., 2013. Geographic and taxonomic disparities in species diversity: dispersal and diversification across Wallace's Line. Evolution 67, 2058–2071.
Brown, J.K.M., 1994. Bootstrap hypothesis tests for evolutionary trees and other dendrograms. P. Natl. Acad. Sci. USA 91, 12293–12297.
Buckley, T.R., Cunningham, C.W., 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. Mol. Biol. Evol. 19, 394–405.
Ceccarelli, F.S., Zaldivar-Riveron, A., 2013. Broad polyphyly and historical biogeography of the neotropical wasp genus *Notiospathius* (Braconidae: Doryctinae). Mol. Phylogenet. Evol. 69, 142–152.
Colston, T.J., Grazziotin, F.G., Shepard, D.B., Vitt, L.J., Colli, G.R., Henderson, R.W., Hedges, S.B., Bonatto, S., Zaher, H., Noonan, B.P., Burbrink, F.T., 2013. Molecular systematics and historical biogeography of tree boas (*Corallus* spp.). Mol. Phylogenet. Evol. 66, 953–959.
Davis, J.I., Simmons, M.P., Stevenson, D.W., Wendel, J.F., 1998. Data decisiveness, data quality, and incongruence in phylogenetic analysis: an example from the monocotyledons using mitochondrial *atp*A sequences. Syst. Biol. 47, 282–310.
De Laet, J., Farris, S., Goloboff, P., 2004. Treatment of multiple trees in resampling analyses. Cladistics 20, 590.
Efron, B., Halloran, E., Holmes, S., 1996. Bootstrap confidence levels for phylogenetic trees. P. Natl. Acad. Sci. USA 93, 13429–13434.
Engelbrecht, H.M., van Niekerk, A., Heideman, N.J.L., Daniels, S.R., 2013. Tracking the impact of Pliocene/Pleistocence sea level and climatic oscillations on the cladogenesis of the Cape legless skink, *Acontias meleagris* species complex, in South Africa. J. Biogeogr. 40, 492–506.
Erixon, P., Svennblad, B., Britton, T., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52, 665–673.
Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.
Felsenstein, J., Kishino, H., 1993. Is there something wrong with the bootstrap on phylogenies? A reply to Hillis and Bull. Syst. Biol. 42, 193–200.
Freudenstein, J.V., Davis, J.I., 2010. Branch support via resampling: an empirical study. Cladistics 26, 643–656.
Freudenstein, J.V., van den Berg, C., Goldman, D.H., Kores, P.J., Molvray, M., Chase, M.W., 2004. An expanded plastid DNA phylogeny of Orchidaceae and analysis of jackknife branch support strategy. Am. J. Bot. 91, 149–157.
Goloboff, P.A., Farris, J.S., 2001. Methods for quick consensus estimation. Cladistics 17, S26–S34.
Goloboff, P.A., Pol, D., 2005. Parsimony and Bayesian phylogenetics. In: Albert, V.A. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 148–159.
Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.
Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
Guo, X., Wang, R.-J., Simmons, M.P., But, P.P.-H., Yu, J., 2013. Phylogeny of the Asian *Hedyotis-Oldenlandia* complex (Spermacoceae, Rubiaceae): Evidence for high levels of polyphyly and parallel evolution of diplophragmous capsules. Mol. Phylogenet. Evol. 67, 110–122.
Guz, N., Kocak, E., Kilincer, N., 2013. Molecular phylogeny of *Trissolcus* species (Hymenoptera: Scelionidae). Biochem. Syst. Ecol. 48, 85–91.
Hasegawa, M., Kishino, H., 1994. Accuracies of the simple methods for estimating the bootstrap probability of a maximum-likelihood tree. Mol. Biol. Evol. 11, 142–154.
Hedges, S.B., 1992. The number of replications needed for accurate estimation of the bootstrap *p* value in phylogenetic studies. Mol. Biol. Evol. 9, 366–369.
Hedtke, S.M., Patiny, S., Danforth, B.N., 2013. The bee tree of life: a supermatrix approach to apoid phylogeny and biogeography. BMC Evol. Biol. 13, 138.
Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analyses. Syst. Biol. 42, 182–192.
Hinchliff, C.E., Roalson, E.H., 2013. Using supermatrices for phylogenetic inquiry: an example using the sedges. Syst. Biol. 62, 205–219.
Holmes, S., 2003. Bootstrapping phylogenetic trees: theory and methods. Stat. Sci. 18, 241–255.
Huelsenbeck, J.P., Rannala, B., 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53, 904–913.
Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N., (Ed.), Mammalian Protein Metabolism. Academic Press, New York, vol. 3, pp. 21–132.
Keskin, E., Agdamar, S., Tarkan, A.S., 2013. DNA barcoding common non-native freshwater fish species in Turkey: low genetic diversity but high population structuring. Mitochondr. DNA 24, 276–287.
Kumar, R., Nongkhlaw, M., Acharya, C., Joshi, S.R., 2013. Uranium (U)-tolerant bacterial diversity from U ore deposit of Domiasiat in north-east India and its prospective utilisation in bioremediation. Microbes Environ. 28, 33–41.
Kwek, J.H.L., Wynne, A., Lefevre, C., Familari, M., Nicholas, K.R., Sharp, J.A., 2013. Molecular evolution of a novel marsupial S100 protein (S100A19) which is expressed at specific stages of mammary gland and gut development. Mol. Phylogenet. Evol. 69, 4–16.
Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference. Syst. Biol. 58, 130–145.
Lewis, P.O., Holder, M.T., Holsinger, K.E., 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54, 241–253.
Miner, B.E., Knapp, R.A., Colbourne, J.K., Pfrender, M.E., 2013. Evolutionary history of alpine and subalpine *Daphia* in western North America. Freshwater Biol. 58, 1512–1522.
Minh, B.Q., Nguyen, M.A.T., von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. 30, 1188–1195.
Olsson, U., Irestedt, M., Sangster, G., Ericson, P.G.P., Alstrom, P., 2013. Systematic revision of the avian family Cisticolidae based on a multi-locus phylogeny of all genera. Mol. Phylogenet. Evol. 66, 790–799.
Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., Moret, B.M.E., Stamatakis, A., 2009. How many bootstrap replicates are necessary? Lect. Notes Comput. Sc. 5541, 184–200.
Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R.P., Moret, B.M.E., Stamatakis, A., 2010. How many bootstrap replicates are necessary? J. Comput. Biol. 17, 337–354.
Pyron, R.A., Burbrink, F.T., Wiens, J.J., 2013. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. BMC Evol. Biol. 13, 93.
Ribeiro, P.L., Rapini, A., Soares e Silva, U.C., van den Berg, C., 2012. Using multiple analytical methods to improve phylogenetic hypotheses in *Minaria* (Apocynaceae). Mol. Phylogenet. Evol. 65, 915–925.
Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.

SAS Institute, 2007. JMP®, version 9.0.2. SAS Institute Inc., Cary.
Schuh, R.T., Polhemus, J.T., 1980. Analysis of taxonomic congruence among morphological, ecological, and biogeographic data sets for the Leptopodomorpha (Hemiptera). Syst. Zool. 29, 1–26.
Shimodaira, H., Hasegawa, M., 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16, 1114–1116.
Siddall, M.E., 2010. Unringing a bell: metazoan phylogenomics and the partition bootstrap. Cladistics 26, 444–452.
Simmons, M.P., 2012a. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Mol. Phylogenet. Evol. 62, 472–484.
Simmons, M.P., 2012b. Misleading results of likelihood-based phylogenetic analyses in the presence of missing data. Cladistics 28, 208–222.
Simmons, M.P., Freudenstein, J.V., 2011. Spurious 99% bootstrap and jackknife support for unsupported clades. Mol. Phylogenet. Evol. 61, 177–191.
Simmons, M.P., Goloboff, P.A., 2013. An artifact caused by undersampling optimal trees in supermatrix analyses of locally sampled characters. Mol. Phylogenet. Evol. 69, 265–275.
Simmons, M.P., Norton, A.P., 2013. Quantification and relative severity of inflated branch-support values generated by alternative methods: an empirical example. Mol. Phylogenet. Evol. 67, 277–296.
Soltis, P.S., Soltis, D.E., 2003. Applying the bootstrap in phylogeny reconstruction. Stat. Sci. 18, 256–267.
Soltis, D.E., Mort, M.E., Latvis, M., Mavrodiev, E.V., O'Meara, B.C., Soltis, P.S., Burleigh, J.G., Rubio de Casas, R., 2013. Phylogenetic relationships and character evolution analysis of Saxifragales using a supermatrix approach. Am. J. Bot. 100, 916–929.
Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463.
Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML web servers. Syst. Biol. 57, 758–771.
Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. P. Natl. Acad. Sci. USA 99, 16138–16143.
Swofford, D.L., 2001. PAUP*: Phylogenetic Analysis using Parsimony (*and other Methods). Sinauer Associates, Sunderland.
Tavaré, S., 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lect. Math. Life Sci. 17, 57–86.
Vinh, L.S., von Haeseler, A., 2004. IQPNNI: moving fast through tree space and stopping in time. Mol. Biol. Evol. 21, 1565–1571.
Whelan, S., Money, D., 2010. The prevalence of multifurcations in tree-space and their implications for tree-search. Mol. Biol. Evol. 27, 2674–2677.
Wilson, E.B., 1927. Probable inference, the law of succession and statistical inference. J. Am. Stat. Assoc. 22, 209–212.
Yang, Z., 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10, 1396–1401.
Yang, Z., 2007a. Fair-balance paradox, star-tree paradox, and Bayesian phylogenetics. Mol. Biol. Evol. 24, 1639–1655.
Yang, Z., 2007b. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.
Yang, Z., Rannala, B., 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Mol. Biol. Evol. 14, 717–724.
Zharkikh, A., Li., W.-H., 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. I. Four taxa with a molecular clock. Mol. Biol. Evol. 9, 1119–1147.
Zwickl, D.J., 2006. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets under the Maximum Likelihood Criterion. Ph.D. Dissertation, The University of Texas at Austin.
Zwickl, D.J., 2009. GARLI 0.96 Settings Cheat Sheet (Smithsonian, June 09). Distributed by the Author.
Zwickl, D.J., 2012. GARLI Configuration Settings. Downloaded on 30 May 2012. <https://www.nescent.org/wg_garli/GARLI_Configuration_Settings>.