Molecular Phylogenetics and Evolution 69 (2013) 265–275

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

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

An artifact caused by undersampling optimal trees in supermatrix analyses of locally sampled characters

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

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

ARTICLE INFO

Article history:
Received 20 April 2013
Revised 28 May 2013
Accepted 1 June 2013
Available online 10 June 2013

Keywords:
Branch swapping
Maximum likelihood
Parsimony
Random-addition sequence
Tree-bisection-reconnection collapsing
Wagner tree

ABSTRACT

Empirical and simulated examples are used to demonstrate an artifact caused by undersampling optimal trees in data matrices that consist mostly or entirely of locally sampled (as opposed to globally, for most or all terminals) characters. The artifact is that unsupported clades consisting entirely of terminals scored for the same locally sampled partition may be resolved and assigned high resampling support—despite their being properly unsupported (i.e., not resolved in the strict consensus of all optimal trees). This artifact occurs despite application of random-addition sequences for stepwise terminal addition. The artifact is not necessarily obviated with thorough conventional branch swapping methods (even tree-bisection-reconnection) when just a single tree is held, as is sometimes implemented in parsimony bootstrap pseudoreplicates, and in every GARLI, PhyML, and RAxML pseudoreplicate and search for the most likely tree for the matrix as a whole. Hence GARLI, RAxML, and PhyML-based likelihood results require extra scrutiny, particularly when they provide high resolution and support for clades that are entirely unsupported by methods that perform more thorough searches, as in most parsimony analyses.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Many contemporary supermatrices (Sanderson et al., 1998) include hundreds or even thousands of terminals that are only scored for a minority of the characters sampled because they were primarily or entirely assembled by using publicly available sequences that were originally generated for more narrowly focused phylogenetic studies. Recently published supermatrix analyses have included 226–73,060 terminals with 70% to 97.5% missing data (e.g., McMahon and Sanderson, 2006; Fabre et al., 2009; Goloboff et al., 2009; Couvreur et al., 2010; Peters et al., 2011).

Goloboff et al. (2009) implemented tree fusing and sectorial searches (Goloboff, 1999) with tree-bisection-reconnection (TBR) to search for the most parsimonious trees. Similarly, McMahon and Sanderson (2006) and Couvreur et al. (2010) both implemented parsimony-based ratchet searches (Nixon, 1999) with TBR to search for the most parsimonious trees. In contrast, Fabre et al. (2009) and Peters et al. (2011) restricted their phylogenetic analyses to RAxML (Stamatakis, 2006), which is limited to "lazy" and local subtree-pruning-and-regrafting (SPR) branch swapping and only saves a single fully resolved most likely tree for the matrix as a whole as well as for each bootstrap (BS; Felsenstein, 1985) pseudoreplicate. Fabre et al. (2009) performed just 100 optimal-tree searches and 100 BS pseudoreplicates while Peters et al. (2011) relied upon rapid bootstrapping (Stamatakis et al., 2008) with 560 pseudoreplicates and presumably just 112 optimal-tree searches. Phylogenetic analyses that are restricted to such a limited number of low quality heuristic searches may be particularly vulnerable to undersampling artifacts that favor clades resolved in a subset of optimal topologies over equally optimal alternative resolutions of those terminals in a manner that is determinate to phylogenetic inference.

In describing the Wagner Method of tree construction, Farris (1970) noted that the algorithm could be modified by changing the order in which terminals are added to the tree in three different ways, though none of these were the random addition sequence (RAS) that is now widely employed as the basis for initial parsimony- and likelihood-tree construction prior to branch swapping. In describing the importance of conducting multiple independent hill-climbing tree searches to identify multiple islands of optimal trees, Maddison (1991, p. 319) asserted that, "PAUP's facilities for generating an unlimited number of [RAS] starting trees make it ideal for discovery of multiple islands." Indeed, Maddison's (1991) assertion has been widely supported, but there is an implicit expectation that with enough RAS searches, all islands of optimal trees can be found.

Källersjö et al. (1998, p. 261) stated that, "To ensure that the addition order of taxa [in each jackknife pseudoreplicate] did not

* 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.2013.06.001

influence the results, five random-addition sequences were performed for each replicate." Tehler et al. (2003, p. 903) asserted that, "The [Xac jackknifing] program automatically discards groups found in less than 50% of the trees for pseudoreplicates, thus eliminating unjustified (poorly supported) resolution caused by ambiguous data sets." The expectation from those statements is that resampling and randomized addition sequences necessarily lead to ambiguously supported clades being collapsed. Källersjö et al.'s (1998) and Tehler et al.'s (2003) expectations were probably met for the datasets that they analyzed (one gene region scored for all terminals sampled), but they do not necessarily generalize to many contemporary supermatrix analyses in which there is a high percentage of missing entries that are non-randomly distributed.

It is well understood that the topology of the initial tree constructed can be determinate to the optimal tree found within a given heuristic hill-climbing search (Maddison, 1991; Davis et al., 2005). Hence ⩾1000 independent searches are typically applied in rigorous parsimony and likelihood phylogenetic analyses and their results are combined to create a strict consensus. If the initial trees (one for each heuristic search) consistently favor clades resolved in only a subset of the optimal topologies then not only may the strict consensus include unsupported clades, but those unsupported clades may also receive high resampling (bootstrap and jackknife; Farris et al., 1996) and Bremer support (Goodman et al., 1985; Bremer, 1988). The reason is that the same preference for groups entailed in the search for the optimal trees for the entire matrix may also be expressed in each resampling pseudoreplicate. Likewise when suboptimal trees found during the searches are used to calculate Bremer support.

To avoid artifacts of unsupported clades in the strict consensus and inflated branch support for those clades, any consistent group preference among the initial trees for a subset of optimal topologies should be minimized. The initial-tree-construction method with the most (potentially) consistent preference for a subset of optimal trees is simple addition sequence. The preference is lower for RAS, and least for entirely random trees. Ideally, branch swapping will overcome any consistent preference in construction of the initial trees. Specifically, the most thorough conventional branch-swapping method is TBR, followed by SPR, then nearest-neighbor interchange (NNI), and finally no swapping at all (Swofford et al., 1996). Yet it is doubtful whether even the most thorough branch-swapping method can overcome a consistent preference in construction of the initial trees when only a single optimal tree can be held despite there being multiple equally optimal trees (Goloboff and Farris, 2001).

A simple test of a potentially consistent group preference in construction of the initial trees and the extent to which branch swapping can overcome any such preference is to conduct a very thorough tree search (Goloboff, 1999; Nixon, 1999; Davis et al., 2005) to rigorously identify the (hopefully correct) strict consensus of all most optimal trees and then compare the majority-rule consensus of the other searches to this. The more consistent the group preference, the higher the number of properly unsupported clades (i.e., any clades that are unresolved in the rigorously constructed strict consensus of all optimal trees; Goloboff et al., 2003) that will be resolved. A complementary test is to quantify the inferred resampling support for those unsupported clades – the more consistent the group preference, the stronger and more misleading the inferred resampling support.

Ideally, those methods with the least consistent preference in initial-tree construction and those trees subsequently found by branch swapping should not just reduce branch support for all clades but rather preferentially reduce inferred branch support for the properly unsupported clades while maintaining support for the properly supported clades. Hence the ratio of support assigned to the properly supported clades should increase as progressively more effective methods for initial-tree construction and branch swapping are applied.

2. Materials and methods

2.1. Empirical examples

The empirical examples consist of 347 terminals sampled for the internal-transcribed-spacer (ITS) region of nuclear rDNA (including the 3′ terminus of the 18S subunit, ITS 1, the entire 5.8S subunit, ITS 2, and the 5′ start of the 26S subunit for most sequences) from the plant order Celastrales. The sequence data were taken from Coughenour et al. (2010, 2011) and Simmons et al. (2012a, 2012b), to which 51 Madagascan terminals were added by Bacon et al. (unpublished data).

Because of alignment ambiguity in the ITS 1 and ITS 2 regions when attempting to align those regions across the entire order, an unconventional alignment approach was implemented whereby the 18S, 5.8S, and 26S regions (together with three adjacent positions from ITS 1 and nine adjacent positions from ITS 2) were globally aligned across the Celastrales whereas the remaining positions of ITS 1 and ITS 2 were only locally aligned within each of seven monophyletic or paraphyletic groups consisting of 26–88 terminals that were well supported in previous analyses and/or trees generated by preliminary analyses of four plastid loci (*atpB*, *matK*, *rbcL*, and *trnL-F*). The two paraphyletic groups are well supported in the sense that they are bracketed by well supported branches. This alignment approach was derived from a presentation by K.S. MacDonald and M.E. Siddall at Hennig XXVI in 2007, which was based on Barta's (1997) proposal on how to integrate hypervariable regions into molecular phylogenetic analyses.

Preliminary nucleotide alignments were obtained using MAFFT ver. 6.5 (Katoh and Toh, 2008a). Q-INS-i, which considers inferred secondary structure of rDNA (Katoh and Toh, 2008b), was used for the local alignments, whereas the less computationally intensive G-INS-i was used for the global alignment. The 20PAM nucleotide scoring matrix was used for all alignments. The default gap opening penalty was applied (1.53) and the gap offset value was set to 0.1. Manual adjustments to the alignments were then performed using the similarity criterion (Zurawski and Clegg, 1993; Simmons, 2004). Ambiguously aligned regions (as identified using the similarity criterion; ranging from 0 to 110 positions in the local alignments) across all terminals were excluded and ambiguously aligned regions from individual terminals were re-scored as ambiguously aligned ("?") for those terminals. Although gap characters should normally be included in sequence-based phylogenetic analyses (Simmons and Ochoterena, 2000; Simmons et al., 2001), they were excluded here so that the parsimony and likelihood analyses (see below) both sampled the same characters (i.e., nucleotides only).

The seven blocks of locally aligned characters (from 477 to 544 characters per block after exclusion of ambiguously aligned regions) were concatenated, one after the other, to the block of 260 globally aligned characters to create the "ITS_all" matrix, which consists of 3814 characters, including 2252 variable and 1796 parsimony-informative characters with 111 – 700 characters scored per terminal (mean = 623 characters). A second matrix ("ITS_conserved"), consisting of only the 260 globally aligned characters, was also analyzed. This matrix includes 90 variable and 58 parsimony-informative characters with 8 – 249 characters scored per terminal (mean = 219). A third matrix ("ITS_no_overlap") was also analyzed wherein the 260 globally aligned characters were staggered in the same manner as the 3554 locally aligned characters such that no characters were scored between any terminals among the seven monophyletic or paraphyletic groups. That is, the 260

globally aligned characters were changed into seven blocks each consisting of 260 locally aligned characters. This third matrix included 2371 variable and 1865 parsimony-informative characters. All three matrices are posted as supplemental online data at: http://rydberg.biology.colostate.edu/Research/.

2.2. Simulated examples

The simulated examples consist of 64 terminals simulated for 10,000 characters. Nucleotide characters were simulated by using the Evolver program with MCbase.dat from the PAML ver. 4.1 package (Yang, 2007). The characters were simulated on a completely pectinate tree for which all branch lengths are 0.004. The relationships among the terminals reflect the terminal numbering (e.g., terminal 10 is more closely related to terminals 11 – 64 than is terminal 9, terminal 9 is more closely related to terminals 10 – 64 than is terminal 8). With 10,000 simulated characters, a 0.004 branch length equates to an average of 40 substitutions per branch. The Jukes and Cantor (1969) model was applied with equal nucleotide frequencies and no rate heterogeneity among characters. These simulation parameters were selected to minimize homoplasy while providing a large number of potential synapomorphies for phylogenetic methods to resolve each branch without confounding effects that may be caused by long-branch attraction (Felsenstein, 1978b) or model-mis-specification (Gaut and Lewis, 1995). Based on these simulation conditions we expected parsimony and likelihood to produce similar results. Ten replicate matrices were created.

The original 10,000 simulated characters were separated into two partitions of 10,000 characters each, with 50% missing data in each partition. The odd-numbered terminals were scored for only the first partition and the even-numbered terminals were scored for only the second partition. Although relationships are clear among the odd terminals relative to each other and among the even terminals relative to each other, there is no comparable information shared between those two groups of terminals. Because the lack of comparable information precludes inferring relationships between the odd and even terminals, the strict consensus should be completely unresolved.

2.3. Parsimony analyses

To test the effect of different terminal-addition methods for generation of the initial tree for subsequent branch swapping, two different methods were implemented in TNT ("rseed [" and "rseed]; Goloboff et al., 2008) and" three different methods were implemented in PAUP* 4.0b10 ("addseq = simple", "addseq = random", and "randomize = trees"; Swofford, 2001). In TNT, both rseed [and rseed] implement a RAS, but they differ with respect to where in the tree each new terminal is attempted to be inserted. Whereas rseed [selects branches on the existing tree in a random sequence to try and insert each newly added terminal, rseed] selects branches for insertions following either a postorder (i.e., downward pass) or a preorder (i.e., upward pass) traversal (each chosen equiprobably for a given terminal to be added to the tree; http://tnt.insectmuseum.org/index.php/Commands/rseed). For completely uninformative data sets (i.e., those that lack parsimony informative characters), rseed [creates approximately random trees whereas rseed] creates pectinate trees.

"Addseq = simple" is the default method in PAUP* and implements the original Wagner Method as described by Farris (1970) without any variation in the order of terminal addition (i.e., only a single search is conducted irrespective of the number of replications specified). "Addseq = random" implements RAS. Unlike rseed [in TNT, PAUP* does not randomly select branches on the existing tree to try and insert each newly added terminal. Rather, PAUP* appears to implement an approach similar to rseed] in TNT because it produces entirely pectinate topologies when applied to an uninformative dataset without branch swapping or collapsing branches. "Randomize = trees" does not implement the Wagner Method at all, but rather uses random topologies as the starting points for branch swapping.

Following the initial-tree construction, three alternative branch-swapping methods (or lack thereof) were tested in TNT (no swapping, SPR, and TBR) and four alternative methods were tested in PAUP* (no swapping, NNI, SPR, and TBR). To make the TNT and PAUP* results more comparable, the options "collapse 3" and "collapse]" were implemented in TNT and the option "collapse = minbrlen" was implemented in PAUP*. By doing so, only branches for which the minimum optimized length is >0 are retained in the saved trees. To emulate the tree-search strategies of GARLI (Zwickl, 2006), RAxML, and PhyML (Guindon et al., 2010), only a single tree was held per search and only a single optimal tree from all searches was held.

For each combination of terminal-addition and branch-swapping methods, 1000 searches were conducted for the optimal tree (albeit only a single search is applicable for "addseq = simple"). One thousand BS pseudoreplicates using the same methods and a single search per pseudoreplicate were also conducted. The BS values were mapped onto the first optimal tree saved by using SumTrees ver. 3.11 (Sukumaran and Holder, 2010).

In addition to the TNT bootstrap analyses that applied the "collapse 3" and "collapse]" options, an additional set of analyses were performed using the "collapse 6" option to implement TBR-collapsing (Goloboff and Farris, 2001). As described by Goloboff and Farris (2001), TBR-collapsing is a way to take into account the ambiguity in results but without saving large numbers of equally parsimonious trees. With this method, the tree to be collapsed is subject to TBR, and for every move that would produce a tree of the same score as the tree being swapped, all the nodes in the path between the two nodes are marked to be eliminated at the end. This is equivalent to producing the strict consensus of all the equally parsimonious trees found by TBR, but without actually saving them to RAM. By using the same time-saving shortcuts used during true searches (e.g., Goloboff, 1993, 1996, 1999), trees can be collapsed with almost no extra time.

A second round of analyses was conducted for the RAS, random-tree, and rseed] methods of initial-tree construction in PAUP* and TNT to demonstrate that the consistent group preference found in the initial-tree construction is not restricted to resampling support. Rather, the same group preference also applies to searches for the optimal trees for the matrix as a whole. For each of the four methods of branch swapping in PAUP*, 10,000 searches for the optimal tree were performed and the single optimal tree found in every search was saved (savereps = yes). The majority-rule consensus (Margush and McMorris, 1981) of all (non-identical, when applicable) 10,000 trees was then saved and the resolution on the tree quantified at both 50% and 90% cut-offs. This approach was implemented rather than reporting the consensus of only the optimal trees found across all 10,000 searches because in many cases only a single optimal tree was found for the ITS matrices. For TNT, the $\geqslant$10,000 tree searches were performed by applying the macro command "loop." The majority-rule consensus for the GARLI results is based on the original 1000 optimal-tree searches.

A relatively thorough set of TNT analyses of the ITS matrices was conducted by using a two-part tree search. Branches with a minimum possible optimized length of zero were collapsed to improve efficiency of tree searches and help minimize artifacts caused by missing data (Kitching et al., 1998; Kearney and Clark, 2003; Davis et al., 2005). In the first part up to 50 trees were held (Davis et al., 2005) within each of 5000 RAS TBR searches that also implemented 100 ratchet iterations, which alternated between

equal character weighting and each character having a 10% chance of being upweighted and a 5% chance of being downweighted. The second part of the search consisted of TBR swapping on all trees obtained from the first part of the search with up to 500,000 trees retained, after which the strict consensus was calculated.

In the case of the ITS_all matrix, we also confirmed the correct strict consensus of optimal trees by running this matrix with a driven search (xmult command) on 44 processors. Each of the replicates incorporated 6 rounds of tree drifting (with 30 substitutions in the perturbation phase, accepting rearrangements within a relative fit difference of 0.1), constrained, random, and exclusive sectorial searches (the random and auto-constrained searches with default settings; the exclusive searches using 20 selections dividing the tree in 3 or 2 parts, using a combined search strategy for each sector, with 6 starts and 6 rounds of tree-drifting per sector).

One thousand BS pseudoreplicates were implemented. Each pseudoreplicate consisted of 100 TBR searches that each held up to 50 most parsimonious trees. Although bootstrapping has known problems (Harshman, 1994; Carpenter, 1996; Goloboff et al., 2003; Freudenstein and Davis, 2010), we used this resampling method because it is the only one implemented in GARLI and RAxML and consistent usage of the BS enables a more direct comparison with the resampling values generated by PAUP* and TNT. For the problems illustrated in this paper, it is irrelevant whether modified resampling procedures (such as in Farris et al., 1996; Goloboff et al., 2003) are used to eliminate the influence of uninformative characters, the influence of differential weights or costs, or the problem that summarizing results with group frequencies may be less appropriate than frequency differences because all of those modified methods are affected in the same way by the quality of the tree-search for each pseudoreplicate.

A relatively thorough TNT analysis of the simulated examples consisted of 1000 RAS TBR searches that each held up to 50 most parsimonious trees. Branches with a minimum possible optimized length of zero were collapsed. The strict consensus of all most parsimonious trees found was then saved. All 1000 BS pseudoreplicates implemented the same tree-search strategy, albeit with 100 RAS TBR searches per pseudoreplicate.

2.4. Likelihood analyses

Likelihood analyses were performed with GARLI ver. 2.0.1019. Following the recommended setting in GARLI, branches with a length of 1×10^{-8} (i.e., effectively zero; GARLI will never present branches with a length of zero; Zwickl, 2012) were collapsed in the same manner that PAUP* collapses these branches in likelihood analyses (Swofford, 2001). The GARLI analyses were performed by using the least rigorous settings for an intensive search recommended by Zwickl (2009; streefname = stepwise; attachmentspertaxon = 50, genthreshfortopoterm = 20,000, numberofprecreductions = 20, treerejectionthreshold = 100) for both optimal-tree searches (1000 search replicates) and the BS (1000 pseudoreplicates, each with 1 search for the empirical examples and 10 searches for the simulated examples).

The best-fit model identified by the Akaike Information Criterion (Akaike, 1974) in jModelTest ver. 0.1.1 (Posada, 2008) for the ITS_conserved and ITS_all matrices was TPM1uf + Γ, which entails three substitution rates. Because this Q-matrix is not implemented in GARLI, the GTR + Γ model with four rate categories was applied to the three empirical matrices instead. Both unpartitioned and partitioned GARLI analyses of these empirical matrices were performed for the ITS_all and ITS_no_overlap matrices. For the ITS_all matrix, the 260 globally aligned characters were treated as one partition and each of the seven groups of locally aligned characters was treated as a separate partition. For the ITS_no_overlap matrix, each of the seven groups of locally aligned characters (including the formerly globally aligned characters) was treated as a separate partition. Different model parameter values and a different rate multiplier were allowed for each partition (linkmodels = 0; subsetspecificrates = 1).

Because the simulated characters were generated using a Jukes-Cantor model without any invariant sites or rate heterogeneity among characters, the same model was applied for the GARLI analyses of the simulated examples. Each character partition was also delimited in the GARLI analyses by using charset commands. The same model was applied to both partitions but each partition was allowed to have its own rate multiplier (linkmodels = 1; subsetspecificrates = 1).

RAxML ver. 7.2.6 likelihood analyses were run on the simulated examples by using the GTRGAMMA model (no simplifications of the GTR Q-matrix are implemented in RAxML and all nucleotide models incorporate gamma-based rate heterogeneity) and 1000 independent searches starting from randomized parsimony trees. Bootstrap support for the clades resolved on the single optimal tree was calculated in RAxML ver. 7.0.3 in order to apply the most thorough bootstrapping algorithm that had been implemented in RAxML and perform multiple separate tree searches per pseudoreplicate. These options are not implemented in RAxML ver. 7.2.6 or 7.5.3. Each of the 1000 BS pseudoreplicates consisted of ten searches using the "–f i" option, which "refine[s] the final BS tree under GAMMA and a more exhaustive algorithm" (Stamatakis, 2008, p. 9). Each of the character partitions was allowed to have different model-parameter values, but no rate multiplier for partitions is implemented in RAxML.

2.5. Quantification of branch support

Branch support was quantified by linearly scaling BS support to that provided by one to four uncontradicted synapomorphies, while also incorporating those clades with 50–62% support (Simmons and Webb, 2006; Simmons et al., 2010). Clades with 50–62% support (less than that provided by one uncontradicted synapomorphy) were set to 0.2, 63–85% support to 0.4, 86–94% support to 0.6, 95–97% support to 0.8, and 98–100% support (equivalent to at least four uncontradicted synapomorphies) to 1. Support was then added across all resolved clades. Averages across all 10 replicates are presented for the simulated examples.

3. Results

3.1. Resolution and scaled support

The resolution (on the first optimal tree held, when applicable) and scaled-support results for the empirical 347-terminal ITS analyses are presented in Table 1. The strict consensus of the TNT ratchet searches was entirely unresolved for the ITS_conserved and ITS_no_overlap matrices, while a mere 17 clades with scaled support of 8.2 were resolved in the ITS_all strict consensus. For the ITS_all matrix, trees of 6492 steps (the most parsimonious trees found) were obtained in 368 of the 5000 searches, and in 846 independent hits from the driven searches. In contrast to the TNT ratchet searches, numerous clades were resolved and assigned $\geqslant$50% BS support for all three matrices by most other methods.

Among the three methods of initial-tree construction applied in PAUP*, using random trees generally provided the lowest scaled support, followed by RAS, and then simple addition sequence for the ITS_all and ITS_no_overlap matrices, with the more dramatic differences observed for the ITS_no_overlap matrix (Table 1). The differences are less pronounced for the ITS_all matrix and in some cases the ordering among methods differs for the ITS_conserved

Table 1
Resolution (on the first optimal tree held, when applicable) and scaled-support results for the empirical 347-terminal ITS analyses.

Method	Conserved		All		17 ratchet clades[a]		No overlap	
	Resolution	Support	Resolution	Support	Support	%	Resolution	Support
Parsimony:								
TNT:								
Ratchet	0	N/A	17	8.2	8.2	100	0	N/A
TBR RAS rseed [	122	9.8	314	107.2	10.2	9.5	247	32
TBR RAS rseed]	124	9.2	317	109.4	10	9.1	280	50.8
TBR RAS [collapse	N/A	0	N/A	44.2	4.6	10.4	N/A	0
TBR RAS] collapse	N/A	0	N/A	44	4.4	10	N/A	0
SPR RAS rseed [	122	10	312	104.2	9.8	9.4	257	34.8
SPR RAS rseed]	125	9.8	320	109.6	10.4	9.5	282	51
SPR RAS [collapse	N/A	0	N/A	37.4	4	10.7	N/A	0
SPR RAS] collapse	N/A	0	N/A	38	4	10.5	N/A	0
No-swap RAS rseed [	122	8.6	305	95	8.6	9.1	284	38.2
No-swap RAS rseed]	123	9	308	96.8	8.8	9.1	272	46.4
No-swap RAS [coll.	N/A	0	N/A	3.8	0.4	10.5	N/A	0
No-swap RAS] coll.	N/A	0	N/A	4	0.4	10	N/A	0
PAUP*:								
TBR RAS	121	10.2	311	110.4	9.4	8.5	277	69
TBR simple	122	8.8	311	113.2	10.2	9	293	154.2
TBR random tree	123	9.4	311	86	6.8	7.9	179	0
SPR RAS	125	9	305	98.6	9.4	9.5	285	67
SPR simple	121	9.2	310	113.8	9.4	8.3	295	154.4
SPR random tree	123	9.6	308	84.4	7.2	8.5	156	0
NNI RAS	124	9	306	104	9	8.7	282	71.8
NNI simple	121	9	303	108.2	9	8.3	295	147.2
NNI random tree	104	0	182	0	0	N/A	91	0
No-swap RAS	120	9.2	308	104.2	9.6	9.2	291	72.2
No-swap simple	123	8.2	297	107.6	8.6	8.0	302	152.6
No-swap random tree	48	0	114	0	0	N/A	69	0
Likelihood:								
GARLI unpartitioned	134	8.6	321	103.6	8.6	8.3	325	88.4
GARLI partitioned	N/A	N/A	314	107.6	9.4	8.7	320	86.8

[a] 17 clades resolved on the ratchet-based strict consensus for the matrix of all ITS characters.

matrix, though only modest differences in scaled support were apparent in the latter for the SPR and TBR results. Between the two methods of initial-tree construction applied in TNT, rseed [provided lower scaled support than did rseed] for both the ITS_all and ITS_no_overlap matrices.

Among the four methods of branch swapping (or lack thereof) applied in PAUP*, there was no consistent pattern across all of the initial-tree construction methods. None of the branch-swapping methods provided any scaled support in the ITS_no_overlap tree when randomized trees were applied. Alternatively, increased thoroughness of branch swapping generally decreased scaled support for the ITS_all and ITS_no_overlap trees when RAS was applied, at least for no swapping vs. NNI vs. SPR. For TNT the same RAS pattern was observed for no swapping vs. SPR vs. TBR for the ITS_no_overlap trees, whereas the opposite pattern was observed for the ITS_all trees.

Other than when random initial trees and either no swapping or NNI swapping were implemented, the percentage of scaled support assigned to the 17 clades present in the strict consensus inferred from the TNT ratchet analysis varied from just 7.9% to 10.7% among the different methods. The percentage of scaled support assigned to those 17 clades generally slightly increased as more thorough branch-swapping methods were applied, but not universally so.

The TBR-collapsing bootstrap results for both rseed [and rseed], irrespective of whether no-swapping, SPR swapping, or TBR swapping were applied, are outliers relative to the other RAS searches (Table 1). The scaled support is zero for every bootstrap analysis of the ITS_conserved and ITS_no_overlap matrices. The TBR-collapsing bootstrap analyses did provide scaled support for the ITS_all matrix, but the scaled support was just 4–41% that provided when "collapse 3" was implemented. The TBR-collapsing bootstrap analyses also provided lower scaled support than when "collapse 3" was implemented for the 17 clades present in the strict consensus inferred from the TNT ratchet analysis. But these were the only analyses, aside from the ratchet, in which ⩾10% of that scaled support was assigned to the 17 clades.

The scaled support provided by GARLI was similar to that provided by the parsimony-based analyses for the ITS_conserved and ITS_all matrices. But GARLI provided ⩾20% additional scaled support for the ITS_no_overlap matrix than did any of the parsimony analyses with the exception of when simple-addition sequence was implemented.

The consistent group preference in initial-tree construction that was observed is that in data matrices consisting mostly or entirely of characters that are sampled locally (as opposed to globally, for most or all terminals), properly unsupported clades that consist entirely of terminals scored for the same locally sampled partition were resolved and in many cases highly supported. To quantify this consistent preference, the number and support for such clades were recorded and contrasted to the number and support for clades that do *not* consist entirely of terminals scored for the same locally sampled partition. Specifically, this quantification was conducted for the ITS_no_overlap matrix, for which there are no properly supported clades. Among the 2675 clades assigned ⩾50% support by the various methods for the ITS_no_overlap trees, only 88 of them did not exclusively contain terminals scored for the same local partition (Table S1). All of these 88 clades were restricted to the four simple-addition-sequence analyses and constituted an average of just 10.1% of the clades assigned ⩾50% support by those methods and even less (7.6%) of the scaled support.

The resolution (on the first optimal tree held, when applicable) and scaled-support results for the simulated 64-terminal analyses

are presented in Table 2. As expected, the strict consensus of the TNT ratchet searches was entirely unresolved. In contrast, numerous clades were resolved and assigned ⩾50% BS support by most other methods.

Among the three methods of initial-tree construction applied in PAUP*, the same pattern was observed as for the ITS matrices: using random trees consistently provided the lowest scaled support, followed by RAS, and finally simple-addition sequence (Table 2). But in contrast to the ITS matrices, rseed] in TNT consistently provided lower scaled support than rseed [. When TBR-collapsing was implemented with either rseed [or rseed], the scaled support was zero for every case.

Among the four methods of branch swapping (or lack thereof) applied in PAUP*, there were no large shifts in scaled support within any of the three initial-tree-construction methods. Contrary to the ITS_all and ITS_no_overlap matrices, increased thoroughness of branch swapping generally decreased scaled support when RAS was applied, at least for no swapping vs. NNI vs. SPR. But the opposite pattern was observed for RAS among the three methods of branch swapping applied in TNT. In both cases the differences are relatively minor.

The average scaled support provided by both GARLI and RAxML is ⩾79% greater than that provided by any of the parsimony methods with the exception of cases in which simple-addition sequence was implemented.

Among the 5408 clades (including all ten replicate matrices) assigned ⩾50% support by the various methods for the simulated 64-terminal trees, every one of them exclusively contain terminals scored for the same local partition, including those 2320 clades resolved on trees that were initially constructed using simple addition sequence (Table S2).

3.2. Correlations among optimal-tree searches

Majority-rule-consensus results for the RAS and random-tree methods of initial-tree construction in PAUP*, TNT, and GARLI when applied to the original data matrix (rather than BS pseudoreplicates) are presented in Table 3 for the empirical 347-terminal ITS analyses. Scaled support is not applicable to these results. Among the four methods of branch swapping (or lack thereof) applied in PAUP*, there was a consistent increase in resolution on the 50% majority-rule consensus as thoroughness of branch swapping increased, but this was not necessarily reflected when the 90% cut-off was applied because the amount of resolution actually decreased for RAS when applied to the ITS_no_overlap matrix. A consistent increase in resolution was also observed among the four methods of branch swapping applied in TNT, in this case for both the 50% and 90% cut-offs.

Disparate results were sometimes obtained for the method of initial-tree construction applied, in particular for the ITS_no_overlap matrix and when NNI or no-swapping was applied in PAUP* (Table 3). As may be expected, there was little congruence among the random-tree search replicates when minimal or no branch swapping was applied. But, more interestingly, use of random trees provided similar results to RAS for the ITS_conserved and ITS_all matrices but not for the ITS_no_overlap matrix. There was little congruence (just two clades in the 50% majority-rule consensus) among the 10,000 optimal-tree searches even when the random trees were swapped upon using TBR.

Congruence among the 1000 optimal-tree searches was consistently greater for GARLI than among any of the parsimony-based 10,000 optimal-tree searches (Table 3). The 17 clades present in the ITS_all ratchet-based strict consensus were consistently resolved by all methods of initial-tree construction when SPR or TBR branch-swapping was applied. Among the 2573 clades resolved in all of the 50% majority-rule consensus trees for the ITS_no_overlap analyses, not a single one included terminals sampled for different partitions.

Majority-rule-consensus results for the RAS and random-tree methods of initial-tree construction in PAUP*, TNT, and GARLI when applied to the original data matrix (rather than BS pseudoreplicates) are presented in Table 4 for the 64-terminal analyses. For both PAUP* and TNT the results are generally stable across the different methods of branch swapping (or lack thereof). Between the different methods of initial-tree construction, there was far greater congruence among the 10,000 optimal-tree searches for RAS than random trees in PAUP*, and a consistent but less pronounced difference between rseed [(greater congruence) and rseed] (less congruence) in TNT.

Congruence among the 1000 optimal-tree searches was far greater for GARLI than among any of the parsimony-based 10,000 optimal-tree searches (Table 4). Among the 2995 clades resolved in the 50% majority-rule consensus trees from each of the 10 replicates, not a single one included terminals sampled for different partitions.

Table 2
Resolution (on the first optimal tree held, when applicable), scaled support, and bootstrap-percentages results for the 64-terminal analyses.

Method	Local partitions		Bootstrap ⩾ 50%	
	Resol.	Support	Minimum	Maximum
Parsimony:				
TNT:				
thorough	0	0	N/A	N/A
TBR RAS rseed [	55.9	6.6	50	92
TBR RAS rseed]	56.9	4.1	50	82
SPR RAS rseed [	56.8	6.8	50	92
SPR RAS rseed]	55.9	2.8	50	83
No-swap RAS rseed [	55.6	7.4	50	92
No-swap RAS rseed]	56.9	4	50	83
PAUP*:				
TBR RAS	57.3	11.3	93	50
TBR simple	58	58	100	100
TBR random tree	52.3	0.6	50	65
SPR RAS	57	11.9	93	50
SPR simple	58	58	100	100
SPR random tree	52.5	0.5	51	63
NNI RAS	56.8	11.3	92	50
NNI simple	58	58	100	100
NNI random tree	39.5	0	N/A	N/A
No-swap RAS	57.6	10	94	51
No-swap simple	58	58	100	100
No-swap random tree	27.3	0	N/A	N/A
Likelihood:				
GARLI partitioned	61	22.1	50	98
RAxML partitioned	61	21.3	50	99

4. Discussion

4.1. Underlying cause of the artifact

We identified a consistent group preference in initial-tree construction in data matrices consisting mostly or entirely of locally sampled characters such that properly unsupported clades consisting entirely of terminals scored for the same locally sampled partition were resolved and in many cases highly supported. The underlying cause of this artifact in supermatrix analyses is identical to that described by Sharkey and Leathers (2001) and Sumrall et al. (2001) in their criticisms of using majority-rule consensus trees to resolve ambiguity, as well as Goloboff's (1999) and Goloboff and Farris' (2001) description of how to efficiently calculate the strict consensus for matrices with hundreds or thousands of terminals. As stated by Goloboff and Farris (2001, pp. S27–S28):

Table 3
Majority-rule-consensus results (at 50% and 90% cutoffs) of 10,000 optimal-tree searches for the empirical 347-terminal ITS analyses.

Method	Conserved		All		17 ratchet clades[a]		No overlap	
	50%	90%	50%	90%	50%	90%	50%	90%
Parsimony:								
TNT:								
TBR RAS rseed [	76	35	255	175	17	17	194	0
TBR RAS rseed]	74	32	254	176	17	17	211	12
SPR RAS rseed [	75	33	250	163	17	15	192	0
SPR RAS rseed]	71	31	250	154	17	15	209	12
No-swap RAS rseed [	65	29	224	137	17	12	175	0
No-swap RAS rseed]	65	28	227	139	17	12	190	11
PAUP*:								
TBR RAS	75	30	252	176	17	17	235	71
TBR random tree	77	36	249	163	17	13	2	0
SPR RAS	73	30	249	171	17	17	235	71
SPR random tree	74	37	245	159	17	13	2	0
NNI RAS	68	30	236	148	17	13	224	84
NNI random tree	0	0	0	0	0	0	0	0
No-swap RAS	70	30	230	143	17	12	217	96
No-swap random tree	0	0	0	0	0	0	0	0
Likelihood:								
GARLI unpartitioned	93	40	267	207	17	16	250	118
GARLI partitioned	N/A	N/A	273	205	17	16	237	88

[a] 17 Clades resolved on the ratchet-based strict consensus for the matrix of all ITS characters.

Table 4
Majority-rule-consensus results (at 50% and 90% cutoffs) of 10,000 optimal-tree searches for the 64-terminal analyses.

Method	Local partitions	
	50%	90%
Parsimony:		
TNT:		
TBR RAS rseed [	28	1
TBR RAS rseed]	19	0
SPR RAS rseed [	28.2	1.2
SPR RAS rseed]	19.6	0
No-swap RAS rseed [	28.1	1.1
No-swap RAS rseed]	19.2	0
PAUP*:		
TBR RAS	37.9	4
TBR random tree	4	0
SPR RAS	37.2	4
SPR random tree	4	0
NNI RAS	36.9	4
NNI random tree	0	0
No-swap RAS	37.4	4
No-swap random tree	0	0
Likelihood:		
GARLI partitioned	55.5	15.2

> The reason for consistently recovering unsupported groups [in the "double-consensus" method] is that some groups [i.e., clades] may be more frequent in optimal or nearly optimal trees, even if unsupported. It is then much more likely that a given individual search will end up in one tree having the frequent group, even if the group is entirely unsupported. Thus, unsupported but frequent groups are consistently recovered. Since this effect is systematic, not random, repeating the estimation will often retrieve the same unsupported groups. This problem affects more than just the estimation methods described here; unless very large numbers of trees are saved (or minimum length is hit many times independently), it also affects more traditional search strategies, such as multiple random addition sequences.

Goloboff and Pol (2005) provided a contrived example of 24 terminals that are fully resolved on a pectinate tree by a matrix of Hennigian (i.e., completely congruent) parsimony-informative characters. To this they added one wildcard terminal (Nixon and Wheeler, 1991) that was not scored for any characters. The wildcard terminal can be equally parsimoniously resolved on any of the 45 branches connecting the other 24 terminals and is not included in the two-terminal clades in 43 of those 45 trees. Consequently, the frequency-within-replicates BS value assigned to the two-terminal clades is (43/45 branches) $\times$ 100 = 96%. Likewise, in initial (or "starting") trees prior to branch swapping the two-terminal clade would be resolved an average of 96% of the time.

In the context of supermatrices that consist mostly or entirely of locally sampled characters, this artifact is manifested by consistently favoring smaller clades that contain terminals exclusively scored for the same locally sampled partition. There are more topologies that resolve smaller clades which contain terminals that are all scored for the same locally sampled partition than there are topologies containing equivalently sized clades that consist of terminals scored for different locally sampled partitions when both of these alternatives are equally optimal. The reason for this is as follows. For two groups of terminals that are not scored for any common characters but for which the relationships among terminals within each group are supported, the two sub-topologies (i.e., one topology for each group of terminals) may be connected to each other at any of their branches in the optimal trees. The two sub-topologies may be connected by one branch or they may partially or even entirely overlap with each other such that their terminals are intermingled on the global topology. Of those many alternative connections and regions of overlap, only a minority connects to or overlaps with each of the smaller clades within the two sub-topologies.

A contrived example to demonstrate the cause of the artifact is presented in Fig. 1. The model tree (Fig. 1A, with terminal branches of negligible length, all other branches of similar length) is used to generate data for a gray and a black gene. As in the simulated matrices (Tables 2 and 4), only the odd-numbered terminals are sequenced for the gray gene, and only the even terminals are sequenced for the black gene. Under maximum likelihood, if the branch lengths for each of the genes are correctly estimated, fewer optimal trees than under parsimony will be obtained. But even under maximum likelihood, the two subtrees can be intertwined in different ways (three examples are shown in Fig. 1B–D) while

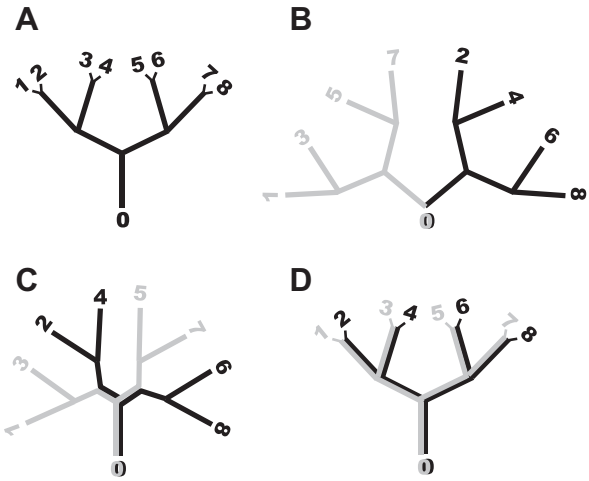

Fig. 1. A contrived example to demonstrate the cause of the artifact. The model tree (A) is used to generate data for a gray and a black gene. Only the odd-numbered terminals are sequenced for the gray gene, and only the even terminals are sequenced for the black gene. Under both parsimony and likelihood there will be more alternative ways to intertwine the two subtrees such that smaller groups (e.g., group 1 + 3, which is present in 2/3 of the tree shown; A, B, C) are monophyletic in a greater proportion of the subtree combinations than are larger groups (e.g., group 2 + 4 + 6 + 8, which is present in 1/3 of the trees shown).

fulfilling the requirement that all branch lengths are correctly identified for both genes. Under both parsimony and likelihood there will be more alternative ways to intertwine the subtrees such that smaller groups (e.g., group 1 + 3, which is present in 2/3 of the trees shown) are monophyletic in a greater proportion of the subtree combinations than are larger groups (e.g., group 2 + 4 + 6 + 8, which is present in 1/3 of the trees shown).

Many supermatrix analyses have been created entirely from publicly available sequences that were originally generated for more narrowly focused phylogenetic studies (e.g., Fabre et al., 2009; Lanfear and Bromham, 2011; Pyron and Wiens, 2011). When the more narrowly focused phylogenetic studies each sampled different loci that are then included in the supermatrix, the artifact described here will tend to favor congruence between the supermatrix tree and the smaller clades resolved in the earlier studies. Consequently, in order to maximize the severity of the test of earlier studies' inferred phylogenies, the supermatrix analysis must rigorously construct a strict consensus for the matrix as a whole as well as within each resampling pseudoreplicate when quantifying branch support.

4.2. Parsimony-based results

Our hypothesis regarding the methods of tree construction was that simple-addition sequence would most consistently favor clades resolved in a subset of optimal topologies over equally optimal alternative resolutions of those terminals, followed by RAS, and then entirely random trees. This hypothesis was supported irrespective of any subsequent branch swapping based on the scaled support for the ITS_all, ITS_no_overlap, and 64-terminal matrices (Tables 1 and 2). The more consistent group preference by RAS relative to entirely random trees was also observed for the same matrices when comparing the majority-rule consensus (at both 50% and 90% cutoffs) of the 10,000 optimal-tree searches (Tables 3 and 4).

More thorough branch-swapping methods generally were not more effective in overcoming any consistent group preference when only a single tree was held, as observed and predicted by Goloboff and Farris (2001). Rather, in most cases for the ITS_all, ITS_no_overlap, and 64-terminal matrices, the scaled support actually increased as progressively more thorough branch-swapping methods were applied (in both PAUP* and TNT; Tables 1 and 2). Likewise for the majority-rule-consensus results (at both 50% and 90% cutoffs) of 10,000 optimal-tree searches for those same matrices.

Use of the simple-addition sequence consistently provided the highest scaled support for the ITS_all, ITS_no_overlap, and 64-terminal matrices (Tables 1 and 2). For the ITS_no_overlap and 64-terminal matrices, where there is no properly supported resolution, the scaled support provided by simple-addition sequence was at least twice that of RAS, irrespective of any subsequent branch swapping. For the matrix as a whole, there is only a single simple-addition sequence that is applied. But different simple-addition sequences may be applied among the BS pseudoreplicates given that the most similar terminals in one pseudoreplicate are not necessarily the most similar terminals in another pseudoreplicate because of variations caused by resampling the original matrix. Given the high scaled support provided by simple-addition sequence for otherwise unsupported clades in the ITS_all, ITS_no_overlap, and 64-terminal matrices without there being any clear advantage over RAS for the ITS_conserved matrix or the 17 ratchet clades for the ITS_all matrix (Table 1), we follow Maddison (1991) in recommending against reliance on simple-addition sequence—even when relying upon just a single tree search for each resampling pseudoreplicate, as is the default in PAUP*. Likewise, one of the main points of Farris et al. (1996) was showing that, for proper evaluations of support, addition sequences must vary between pseudoreplicates.

The use of random starting trees in PAUP* was successful in eliminating or substantially decreasing (relative to RAS) BS support $\geqslant$50% for otherwise unsupported clades in the ITS_all, ITS_no_overlap, and the 64-terminal matrices (Tables 1 and 2). Yet this method does not appear to be generally effective for phylogenetic inference because it did not preferentially decrease BS support for unsupported clades over the 17 supported clades in the ITS_all matrix when SPR and TBR swapping were applied (Table 1). Given the vast number of possible bifurcating trees (Felsenstein, 1978a), the difficulty of effectively traversing tree space when applying conventional hill-climbing branch-swapping methods (Goloboff, 1999; Nixon, 1999; Davis et al., 2005), and the frequent existence of multiple islands of most parsimonious trees (Maddison, 1991), incorporation of some searches using random starting trees may be helpful for eliminating properly unsupported clades from the strict consensus when multiple optimal trees are held, but this method should not be relied upon exclusively for constructing initial trees for subsequent branch swapping.

The PAUP* RAS and TNT RAS (both rseed [and rseed]) results are roughly similar for the ITS_conserved and ITS_all matrices, yet the results are disparate for the ITS_no_overlap and 64-terminal matrices. Irrespective of whether no swapping, SPR, or TBR branch swapping was subsequently applied, PAUP* RAS provided 31–116% higher scaled support relative to that provided by TNT for the ITS_no_overlap matrix (Table 1) and an average of 35–325% higher scaled support for the 64-terminal matrix (Table 2). Furthermore, PAUP* RAS consistently provided more resolution than TNT RAS in both majority-rule consensus trees (at both 50% and 90% cutoffs) for the same two matrices. The disparities were particularly dramatic when the 90% cutoff was applied, with 71–96 clades resolved by PAUP* and 0–12 clades resolved by TNT for the ITS_no_overlap matrix (Table 3), and an average of 4 clades resolved by PAUP* and an average of 0–1.2 clades resolved by TNT for the 64-terminal matrix (Table 4). Taken together, these results appear to indicate that details of the RAS implementations differ between PAUP* and TNT. At least in the specific case of these matrices, the TNT implementations seem to provide a better

exploration of tree space, and hence more accurate results, than PAUP*. This is somewhat unexpected, because (speed differences aside) the implementation of search algorithms in PAUP* is quite thorough, and there is only a limited number of details that can vary in the implementation of RAS and TBR (none of which would obviously produce the observed disparity in results).

TBR-collapsing, as implemented in TNT using the "collapse 6" command, clearly outperformed collapsing branches with a minimum length of zero, as done using the "collapse = minbrlen" command in PAUP* and the "collapse 3" command in TNT. Based on these results, we suggest that TBR-collapsing be widely applied to any data matrices in which ambiguous phylogenetic resolution may be expected, such as empirical supermatrices that contain large amounts of non-randomly distributed missing data. Implementation of TBR-collapsing is particularly important when few trees are to be saved (during searches for optimal trees, as in Goloboff, [1999, p. 425], or within each resampling pseudoreplicate). Although TBR-collapsing is more computationally demanding than just collapsing zero-length branches (which is only based on optimization of the final tree[s]), the cost of this method under parsimony is not excessive given the high efficiency of the TBR implementation in TNT. But the cost of using a similar method, however desirable, may be prohibitive in the case of maximum likelihood programs that only do SPR (or a simplified version of SPR, as in RAxML).

4.3. Implications for likelihood analyses

Aside from the ratchet analysis, all of the results were based on holding only a single optimal tree for the matrix as a whole as well as within each BS pseudoreplicate. Holding only a single optimal tree is almost never implemented for the matrix as a whole in parsimony analyses, though it is sometimes implemented for parsimony-based BS and jackknife pseudoreplicates (e.g., DeBry and Olmstead, 2000; McMahon and Sanderson, 2006; Richardson et al., 2006). By contrast, retaining a single optimal tree is the *only option* in every likelihood-based GARLI, PhyML, and RAxML analysis for both the matrix as a whole as well as within every BS pseudoreplicate. Hence the results presented here are most directly relevant to likelihood-based studies that are based on those programs. Arguably those three programs are used for the vast majority of contemporary likelihood-based phylogenetic analyses. For example, among the 35 regular articles in the January 2013 issue of *Molecular Phylogenetics and Evolution*, 22 included maximum likelihood-based phylogenetic analyses and of those, 13 used RAxML, five used PhyML, two used PAUP*, one used GARLI, and one used TreeFinder (Jobb et al., 2004).

Of the 13 regular articles that used RAxML, seven explicitly applied the rapid bootstrapping algorithm. This implementation of likelihood bootstrapping is particularly problematic because the pseudoreplicates are not independent. Rather, the starting parsimony tree for every tenth pseudoreplicate is based on the optimal tree inferred for the original matrix as a whole and the starting tree for each of the nine subsequent pseudoreplicates is based on the final tree from the preceding pseudoreplicate (Stamatakis et al., 2008). This approach adds a bias that is liable to further inflate support values for properly unsupported clades (Siddall, 2010; Simmons and Norton, 2013). As is clear from our results, a consistent preference for clades resolved in a subset of optimal topologies over equally optimal alternative resolutions of those terminals is not necessarily obviated even if relatively thorough branch-swapping methods are applied while holding just a single tree.

Currently, GARLI, PhyML, and RAxML all rely upon "lazy" SPR, wherein only local branch lengths are re-optimized after branch swapping, as the means of exploring tree space (Stamatakis et al., 2005, 2008; Guindon et al., 2010). Furthermore, RAxML only performs local, rather than global, SPR swaps (Stamatakis et al., 2005). But even if future versions of these programs implement TBR-quality branch swapping, this will not necessarily overcome a consistent group preference for clades that are more frequent in optimal topologies; only saving multiple trees (explicitly, or implicitly, as in TBR-collapsing) will accomplish this, but the time required for an analysis will increase enormously. GARLI ver. 2 has an option wherein branches that have a length of 1×10^{-8} (i.e., effectively zero) can be collapsed. That is a step forward relative to the fully resolved trees that are always output by PhyML and RAxML, but it is clearly insufficient. The programs also should be able to retain multiple equally optimal trees and not ignore alternative topologies that differ by as little as one-hundred thousandth (PhyML), one-millionth (RAxML), or one-hundred millionth (GARLI) of a log-likelihood (see also Morrison, 2007).

The results presented here are primarily from parsimony-based analyses rather than likelihood-based analyses. The likelihood results are from unpartitioned and partitioned GARLI analyses that performed relatively thorough tree searches (following Zwickl, 2009). The GARLI results should be a strong test of whether a consistent group preference is determinate to GARLI, PhyML, and RAxML analyses because GARLI uses both global and local swapping (unlike RAxML; Stamatakis et al., 2005), and does not limit most SPR swaps to those that pass a parsimony-based filter (as in PhyML; Guindon et al., 2010). Also note that the default setting in RAxML is to use a parsimony-based starting tree (Stamatakis, 2008).

Relative to the RAS-based parsimony analyses, GARLI provided higher scaled support for the ITS_no_overlap and 64-terminal matrices. Furthermore, congruence among the 1000 GARLI optimal-tree searches was consistently greater than among any of the parsimony-based 10,000 optimal-tree searches for all three ITS matrices as well as the 64-terminal matrices. Based on these results it is clear that a consistent group preference, favoring clades resolved in a subset of optimal topologies over equally optimal alternative resolutions of those terminals, also applies to likelihood analyses.

4.4. Connection to empirical supermatrices

The consistent results between the empirical 347-terminal ITS and contrived 64-terminal analyses demonstrate the following four points. First, problems caused by a consistent group preference in tree-searches are not only a concern for supermatrices with hundreds of terminals, but can also apply to matrices with fewer terminals, as in more traditional phylogenetic analyses. Second, these problems are not limited to matrices in which some terminals are represented by just a few hundred characters but also can occur in matrices in which every terminal is sampled for thousands of parsimony-informative characters. The 64-terminal matrices have an average of 2099 parsimony-informative characters scored per terminal. Third, these problems are not limited to matrices with moderate amounts of character conflict. The ensemble consistency and retention indices (excluding parsimony-uninformative characters; Kluge and Farris, 1969; Farris, 1989) are 0.48 and 0.79, respectively, for the ITS_all PAUP* RAS TBR trees and 0.49 and 0.79, respectively, for the ITS_no_overlap trees. In contrast, those indices averaged 0.92 and 0.99 for the 64-terminal PAUP* RAS TBR trees. Fourth, these problems are not limited to matrices that consist of just two locally sampled partitions (as in the 64-terminal matrices). Rather, they can also occur in matrices with multiple locally sampled partitions as well as a globally sampled partition (as for the ITS_all matrix).

Neither the empirical nor the simulated examples examined here appear similar to most published empirical supermatrices, because of the few characters scored for the globally sampled

partition (ITS empirical examples) and the entire lack of overlap between the locally sampled partitions (both the ITS empirical examples and the 64-terminal simulated examples). These differences may apply for the matrices as a whole, but they often do not apply to individual clades resolved by published empirical supermatrices. For example, Simmons (2012a) identified numerous clades resolved in McMahon and Sanderson's (2006) supermatrix strict consensus that could not be unequivocally justified because they included terminals with no comparable information (i.e., no scored characters in common). Likewise, Simmons (2012b) pointed out that Wolsan and Sato's (2010) basis for asserting superior performance of Bayesian MCMC over parsimony in their supermatrix analysis was the resolution provided by Bayesian MCMC, but not parsimony, for terminals with no comparable information.

Such cases are apparent when authors explicitly indicate their character sampling for each terminal, as did Fabre et al. (2009) for their 27 sampled gene regions that were analyzed using RAxML. For example, their figure 3 shows *Propithecus edwardsi* as sister to four other species from the same genus, while *P. diadema* is sister to all five of those species. Yet *P. diadema* and *P. edwardsi* are not scored for any common characters so there can be no observational basis other than extrapolating branch lengths among character partitions for asserting that *P. edwardsi* is more closely related to those other four species than is *P. diadema*, let alone assign that clade 70–95% BS support. Other clear examples from the same paper involve *Ateles belzebuth* vs. *A. fusciceps* (70–95% BS), *Cebus libidinosus* + *C. nigritus* vs. *C. xanthosternos* ($\geqslant$95% BS), *Callicebus brunneus* vs. *C. hoffmannsi* (70–95% BS), and *Cercopithecus lowei* vs. *C. pogonias* (70–95% BS). As Sanderson et al. (2010) described, just because a data matrix includes a large number of characters and overlap in character sampling among most closely related terminals does not mean that it is decisive for all trees.

For phylogenomic supermatrices that include hundreds of loci and for which the missing data are ± randomly distributed (e.g., as may occur with anchored hybrid enrichment; Lemmon et al., 2012), we do not expect the artifact described here to be particularly severe given that a consistent group preference from any particular locus would be countered by many other loci that have different distributions of missing data. For phylogenomic supermatrices one may apply the partition bootstrap wherein entire character partitions (or loci) are resampled in or excluded from a given pseudoreplicate (Siddall, 2010). We do not anticipate the partition bootstrap to be any more or less susceptible than regular bootstrapping to the artifact described here as long as numerous of loci are sampled in the phylogenomic supermatrix.

5. Conclusions

Irrespective of whether RAS rseed [or RAS rseed] is applied, there is a consistent group preference during tree-searches on data matrices that consist mostly or entirely of characters that are sampled locally (as opposed to globally, for most or all terminals). This preference is for smaller clades consisting entirely of terminals that are scored for the same locally sampled partition such that these clades may be resolved and assigned high resampling support despite their being properly unsupported. These smaller clades are properly unsupported because of the lack of comparable information among terminals for those characters that provide synapomorphies for their apparent resolution.

Neither simple-addition sequence nor random tree topologies should be relied exclusively upon for initial-tree construction because the former can consistently resolve properly unsupported clades and the latter provides an inefficient starting point from which to identify properly supported clades.

The consistent group preference in tree-searches that favors clades resolved in a subset of optimal topologies over equally optimal alternative resolutions of those terminals is not necessarily obviated when thorough conventional branch swapping methods (even TBR) are subsequently applied as long as just a single tree is held.

Application of TBR-collapsing within each resampling pseudoreplicate is an effective approach to help address ambiguous phylogenetic resolution in supermatrices that contain large amounts of non-randomly distributed missing data and when low quality tree searches are applied within each resampling pseudoreplicate.

The consistent group preference in tree searches affects both parsimony and likelihood optimality criteria.

The consistent group preference in tree-searches is liable to be particularly problematic when only a single tree is saved for the matrix as a whole and/or in each resampling pseudoreplicate. Hence GARLI, RAxML, and PhyML-based results require extra scrutiny, particularly when they provide high resolution and strong support for clades that are entirely unsupported by methods that perform more thorough searches, as in most parsimony analyses.

Acknowledgments

We thank John Gatesy, Mark Siddall, and an anonymous reviewer for suggestions with which to improve our manuscript; Christine Bacon for permission to use unpublished ITS sequences; Jerry Davis for suggestions on an earlier draft of this manuscript; John Freudenstein, John Gatesy, and Todd Gilligan for helpful discussions; and Todd Gilligan for help running the phylogenetic analyses and SumTrees.

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.2013.06.001.

References

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19, 716–723.
Barta, J.R., 1997. Investigating phylogenetic relationships within the Apicomplexa using sequence data: the search for homology. Methods 13, 81–88.
Bremer, K., 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42, 795–803.
Carpenter, J.M., 1996. Uninformative bootstrapping. Cladistics 12, 177–181.
Coughenour, J.M., Simmons, M.P., Lombardi, J.A., Cappa, J.J., 2010. Phylogeny of Celastraceae subfamily Salacioideae and tribe Lophopetaleae inferred from morphological characters and nuclear and plastid genes. Syst. Bot. 35, 358–366.
Coughenour, J.M., Simmons, M.P., Lombardi, J.A., Yakobson, K., Archer, R.H., 2011. Phylogeny of Celastraceae subfamily Hippocrateoideae inferred from morphological characters and nuclear and plastid genes. Mol. Phylogenet. Evol. 59, 320–330.
Couvreur, T.L.P., Franzke, A., Al-Shehbaz, I.A., Bakker, F.T., Koch, M.A., Mummenhoff, K., 2010. Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol. Biol. Evol. 27, 55–71.
Davis, J.I., Nixon, K.C., Little, D.P., 2005. The limits of conventional cladistic analysis. In: Albert, V.A. (Ed.), Parsimony, Phylogeny, and Genomics. Oxford University Press, Oxford, pp. 119–147.
DeBry, R.W., Olmstead, R.G., 2000. A simulation study of reduced tree-search effort in bootstrap resampling analysis. Syst. Biol. 49, 171–179.
Fabre, P.-H., Rodrigues, A., Douzery, E.J.P., 2009. Patterns of macroevolution among primates inferred from a supermatrix of mitochondrial and nuclear DNA. Mol. Phylogenet. Evol. 53, 808–825.
Farris, J.S., 1970. Methods for computing Wagner trees. Syst. Zool. 19, 83–92.
Farris, J.S., 1989. The retention index and the rescaled consistency index. Cladistics 5, 417–419.
Farris, J.S., Albert, V.A., Källersjö, M., Lipscomb, D., Kluge, A.G., 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12, 99–124.
Felsenstein, J., 1978a. The number of evolutionary trees. Syst. Zool. 27, 27–33.
Felsenstein, J., 1978b. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27, 401–410.
Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.

Freudenstein, J.V., Davis, J.I., 2010. Branch support via resampling: an empirical study. Cladistics 26, 643–656.

Gaut, B.S., Lewis, P.O., 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12, 152–162.

Goloboff, P.A., 1993. Character optimization and calculation of tree lengths. Cladistics 9, 433–436.

Goloboff, P.A., 1996. Methods for faster parsimony analysis. Cladistics 12, 199–220.

Goloboff, P.A., 1999. Analyzing large data sets in reasonable times: solutions for composite optima. Cladistics 15, 415–428.

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.

Goloboff, P.A., Farris, J.S., Källersjö, M., Oxelman, B., Ramírez, M.J., Szumik, C.A., 2003. Improvements to resampling measures of group support. Cladistics 19, 324–332.

Goloboff, P.A., Farris, J.S., Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774–786.

Goloboff, P.A., Catalano, S.A., Mirande, J.M., Szumik, C.A., Arias, J.S., Källersjö, M., Farris, J.S., 2009. Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. Cladistics 25, 211–320.

Goodman, M., Olson, C.B., Beeber, J.E., Czelusniak, J., 1985. New perspectives in the molecular biological analysis of mammalian phylogeny. Acta Zoologica Fennica 169, 19–35.

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.

Harshman, J., 1994. The effect of irrelevant characters on bootstrap values. Syst. Biol. 43, 419–424.

Jobb, G., von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 18.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Murto, H.N. (Ed.), Mammalian Protein Metabolism, vol. 3. Academic Press, New York, pp. 21–132.

Källersjö, M., Farris, J.S., Chase, M.W., Bremer, B., Fay, M.F., Humphries, C.J., Petersen, G., Seberg, O., Bremer, K., 1998. Simultaneous parsimony jackknife analysis of 2538 *rbc*L DNA sequences reveals support for major clades of green plants, land plants, seed plants, and flowering plants. Plant Syst. Evol. 213, 259–287.

Katoh, K., Toh, H., 2008a. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9, 286–298.

Katoh, K., Toh, H., 2008b. Improved accuracy of multiple ncRNA alignment by incorporating structural information into a MAFFT-based framework. BMC Bioinformatics 9, 212.

Kearney, M., Clark, J.M., 2003. Problems due to missing data in phylogenetic analyses including fossils: a critical review. J. Vertebr. Paleontol. 23, 263–274.

Kitching, I.J., Forey, P.L., Humphries, C.J., Williams, D.M., 1998. Cladistics: The Theory and Practice of Parsimony Analysis. Oxford University Press, Oxford, Oxford.

Kluge, A.G., Farris, J.S., 1969. Quantitative phyletics and the evolution of Anurans. Syst. Zool. 18, 1–32.

Lanfear, R., Bromham, L., 2011. Estimating phylogenies for species assemblages: a complete phylogeny for the past and present native birds of New Zealand. Mol. Phylogenet. Evol. 61, 958–963.

Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.

Maddison, D.R., 1991. The discovery and importance of multiple islands of most-parsimonious trees. Syst. Zool. 40, 315–328.

Margush, T., McMorris, F.R., 1981. Consensus *n*-trees. Bull. Math. Biol. 43, 239–244.

McMahon, M.M., Sanderson, M.J., 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Syst. Biol. 55, 818–836.

Morrison, D.A., 2007. Increasing the efficiency of searches for the maximum likelihood tree in a phylogenetic analysis of up to 150 nucleotide sequences. Syst. Biol. 56, 988–1010.

Nixon, K.C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15, 407–414.

Nixon, K.C., Wheeler, Q.D., 1991. Extinction and the origin of species. In: Wheeler, Q.D., Novacek, M. (Eds.), Extinction and Phylogeny. Columbia University Press, New York, pp. 119–143.

Peters, R.S., Meyer, B., Krogmann, L., Borner, J., Meusemann, K., Schutte, K., Niehuis, O., Misof, B., 2011. The taming of an impossible child: a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. BMC Biol. 9, 55.

Posada, D., 2008. JModelTest: phylogenetic model averaging. Mol. Biol. Evol. 25, 1253–1256.

Pyron, R.A., Wiens, J.J., 2011. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. Mol. Phylogenet. Evol. 61, 543–583.

Richardson, D., Simmons, M.P., Reddy, A.S.N., 2006. Comprehensive comparative analysis of kinesins in photosynthetic eukaryotes. BMC Genomics 7, 18.

Sanderson, M.J., Purvis, A., Henze, C., 1998. Phylogenetic supertrees: assembling the trees of life. Trends Ecol. Evol. 13, 105–109.

Sanderson, M.J., McMahon, M.J., Steel, M., 2010. Phylogenomics with incomplete taxon coverage: the limits of inference. BMC Evol. Biol. 10, 155.

Sharkey, M.J., Leathers, J.W., 2001. Majority does not rule: the trouble with majority-rule consensus trees. Cladistics 17, 282–284.

Siddall, M.E., 2010. Unringing a bell: metazoan phylogenomics and the partition bootstrap. Cladistics 26, 444–452.

Simmons, M.P., 2004. Independence of alignment and tree search. Mol. Phylogenet. Evol. 31, 874–879.

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., 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.

Simmons, M.P., Ochoterena, H., 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49, 369–381.

Simmons, M.P., Webb, C.T., 2006. Quantification of the success of phylogenetic inference in simulations. Cladistics 22, 249–255.

Simmons, M.P., Ochoterena, H., Carr, T.G., 2001. Incorporation, relative homoplasy, and effect of gap characters in sequence-based phylogenetic analyses. Syst. Biol. 50, 454–462.

Simmons, M.P., Müller, K.F., Norton, A.P., 2010. Alignment of, and phylogenetic inference from, random sequences: the susceptibility of alternative alignment methods to creating artifactual resolution and support. Mol. Phylogenet. Evol. 57, 1004–1016.

Simmons, M.P., Bacon, C.D., Cappa, J.J., McKenna, M.J., 2012a. Phylogeny of Celastraceae subfamilies Cassinoideae and Tripterygioideae inferred from morphological characters and nuclear and plastid loci. Syst. Bot. 37, 456–467.

Simmons, M.P., McKenna, M.J., Bacon, C.D., Yakobson, K., Cappa, J.J., Archer, R.H., Ford, A.J., 2012b. Phylogeny of Celastraceae tribe Euonymeae inferred from morphological characters and nuclear and plastid genes. Mol. Phylogenet. Evol. 62, 9–20.

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

Stamatakis, A., 2008. The RAxML 7.0.4 manual. <http://icwww.epfl.ch/_stamatak/index-Dateien/Page443.htm> (downloaded 16.09.08).

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.

Sukumaran, J., Holder, M.T., 2010. DendroPy: a python library for phylogenetic computing. Bioinformatics 26, 1569–1571.

Sumrall, C.D., Brochu, C.A., Merck, J.W., 2001. Global lability, regional resolution, and majority-rule consensus bias. Paleobiology 27, 254–261.

Swofford, D.L., 2001. PAUP*: Phylogenetic Analysis using Parsimony (*and other methods). Sinauer Associates, Sunderland.

Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M., 1996. Phylogenetic inference. In: Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), Molecular Systematics. Sinauer Associates, Inc., Sunderland, pp. 407–514.

Tehler, A., Little, D.P., Farris, J.S., 2003. The full-length phylogenetic tree from 1551 ribosomal sequences of chitinous fungi. Mycol. Res. 107, 901–916.

Wolsan, M., Sato, J.J., 2010. Effects of data incompleteness on the relative performance of parsimony and Bayesian approaches in a supermatrix phylogenetic reconstruction of Mustelidae and Procyonidae (Carnivora). Cladistics 26, 168–194.

Yang, Z., 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24, 1586–1591.

Zurawski, G., Clegg, M.T., 1993. *Rbc*L sequence data and phylogenetic reconstruction in seed plants: forward. Ann. Mo. Bot. Gard. 80, 523–525.

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. <https://www.nescent.org/wg_garli/GARLI_Configuration_Settings> (downloaded 30.05.12).