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.

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