Molecular Phylogenetics and Evolution 100 (2016) 424–443

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Biases of tree-independent-character-subsampling methods

Mark P. Simmons [a,*], John Gatesy [b]

[a] Department of Biology, Colorado State University, Fort Collins, CO 80523, USA
[b] Department of Biology, University of California, Riverside, CA 92521, USA

ARTICLE INFO

Article history:
Received 9 December 2015
Revised 16 March 2016
Accepted 15 April 2016
Available online 19 April 2016

Keywords:
Character subsampling
Gnetales
Observed Variability
Phylogenetic bias
Substitution rates
Tree Independent Generation of Evolutionary Rates

ABSTRACT

Observed Variability (OV) and Tree Independent Generation of Evolutionary Rates (TIGER) are quick and easy-to-apply tree-independent methods that have been proposed to provide unbiased estimates of each character's rate of evolution and serve as the basis for excluding rapidly evolving characters. Both methods have been applied to multiple phylogenomic datasets, and in many cases the authors considered their trees inferred from the OV- and TIGER-delimited sub-matrices to be better estimates of the phylogeny than their trees based on all characters. In this study we use four sets of simulations and an empirical phylogenomic example to demonstrate that both methods share a systematic bias against characters with more symmetric distributions of character states, against characters with greater observed character-state space, and against large clades in the context of character conflict. As a result these methods can favor convergences and reversals over synapomorphy, exacerbate long-branch attraction, and produce mutually exclusive phylogenetic inferences that are dependent upon differential taxon sampling. We assert that neither OV nor TIGER should be relied upon to increase the ratio of phylogenetic to non-phylogenetic signal in a data matrix. We also assert that skepticism is warranted for empirical phylogenetic results that are based on OV- and/or TIGER-based character deletion wherein a small clade is supported after deletion of characters, yet is contradicted by a larger clade when the entire data matrix was analyzed.

1. Introduction

In phylogenomic datasets that consist of tens or hundreds of thousands of characters, taxon sampling and character quality are generally greater concerns than character number. Character quality is important because of the widespread recognition that biases in these datasets may lead to incorrect, yet highly supported, molecular-evolution and phylogenetic inferences (e.g., Wong et al., 2008; Philippe et al., 2011; Gatesy and Springer, 2014; Simmons and Gatesy, 2015). Hence automated methods that are thought to reduce these biases by excluding characters are widely applied. These methods include excluding ambiguously aligned regions (e.g., Castresana, 2000; Capella-Gutiérrez et al., 2009; but see Tan et al., 2015), excluding third-codon positions (e.g., Chiari et al., 2012; Wickett et al., 2014; but see Källersjö et al., 1999), and analyzing amino-acid rather than nucleotide characters (e.g., Katz and Grant, 2015; Zapata et al., 2015; but see Simmons and Freudenstein, 2002).

Rather than excluding all silent substitutions or all third-codon positions *a priori*, an attractive possibility is to estimate the rate for each individual character along a continuum (rather than assigning it to a pre-defined rate category) in a tree-independent manner such that rate estimates are not dependent upon, and hence not potentially biased by, any particular topology. Two alternative methods have been introduced by authors who have claimed to do just this. Goremykin et al. (2010) described Observed Variability (OV) and Cummins and McInerney (2011) presented Tree Independent Generation of Evolutionary Rates (TIGER).

1.1. Observed Variability

Goremykin et al. (2010:329) introduced OV as a method that "... can arguably improve the net result of phylogeny reconstruction, especially for deeper nodes, where the inference of phylogeny is especially obscured by multiple substitutions and the resulting long-branch attraction." In addition to OV they introduced a gamma-rate (Yang, 1993) sorter. They reported that OV outperformed the gamma-rate sorter, and most empirical phylogenetics papers that have cited Goremykin et al. (2010) only applied OV. Hence we focus on their OV sorter here.

* Corresponding author at: Department of Biology, 200 West Lake Street, Colorado State University, Fort Collins, CO 80523-1878, USA.
E-mail address: psimmons@rams.colostate.edu (M.P. Simmons).

http://dx.doi.org/10.1016/j.ympev.2016.04.022
1055-7903/

OV is calculated for each character individually without reference to any other characters, so it does not take into account character congruence. The OV-score for a given character is simply the number of pairwise character-state matches (scored as 0's) relative to the number of character-state mismatches (scored as 1's) among all terminals for which the character is scored (i.e., excluding terminals with gaps or missing data). Hence, all else equal, lower OV scores are given to characters with fewer observed character states as well as those with more asymmetric distributions of character states. For example, in a matrix with 100 terminals a character with two adenines and 98 thymines has an OV score of just 0.04 because the vast majority of pairwise comparisons are matches of thymines with other thymines. On the other hand a character with 50 cytosines and 50 guanines has an OV score of 0.5 because half of the pairwise comparisons are mismatches between cytosines and guanines.

OV has been broadly applied to delete putatively fast evolving characters from genomic-scale datasets, including those based on mitochondrial (e.g., Lavrov et al., 2013; Liu et al., 2014; Meiklejohn et al., 2014), plastid (e.g., Zhong et al., 2011; Goremykin et al., 2013; Sun et al., 2015), and nuclear (Xi et al., 2013, 2014) genomes. Typically characters are deleted in blocks of 250–1000 positions at a time, and a range of deletions are explored, from <5% up to 50% of the parsimony-informative characters (Xi et al., 2013, 2014) or even 50% of all characters (Sun et al., 2015). The most extreme case that we know of is Xi et al.'s (2014) deletion of all but 5000 of their 142,590 parsimony-informative characters in one of their analyses.

For many authors, the preferred phylogenetic hypothesis is that based on an OV-selected sub-matrix rather than that based on all of the data (e.g., Zhong et al., 2011; Goremykin et al., 2013; Xi et al., 2014). But other authors have expressed concern about OV-based results. Drew et al. (2014:379) noted that OV "... essentially collapses branches by reducing the number of characters," and that this can even apply to "... (virtually) universally accepted clades..." Likewise, Meiklejohn et al. (2014:321) concluded that OV-based character deletion "... eliminated valuable evolutionary signal, resulting in reduced resolution of relationships at a range of taxonomic levels."

More recently, Simmons and Gatesy (2015) demonstrated that Xi et al.'s (2014) application of OV to eliminate half of the parsimony-informative characters from their dataset of 310 nuclear genes resulted in worse performance (as measured by congruence between gene trees with well-established reference clades, the overall success of resolution (number of clades correctly resolved minus number of clades incorrectly resolved), and the averaged overall success of resolution, which incorporates resampling support (Simmons and Webb, 2006)) relative to the complete datasets. Furthermore, the OV-slow parsimony-informative characters had a much lower amount of possible synapomorphy (maximum number of steps possible – minimum number of steps possible in a parsimony context; Farris, 1989) than the OV-fast parsimony-informative characters. The concatenation-based analyses of OV-slow characters actually had lower ensemble retention indices (Farris, 1989) on their most parsimonious trees than did the concatenation-based analyses of OV-fast characters on their most parsimonious trees.

Simmons and Gatesy (2015) also demonstrated that Xi et al.'s (2014) application of OV to analyze the 5000 most conserved characters resulted in a disparate likelihood tree topology (gymnosperms paraphyletic, monocots polyphyletic, eudicots paraphyletic, *Amborella* and *Nuphar* nested within the angiosperms) with all clades consisting of more than five terminals (in an unrooted context) being effectively unsupported. The cause of the effectively unsupported large clades was that none of the 5000 parsimony-informative characters, all of which were binary, included the minority character state in more than two terminals. Finally, Simmons et al. (2016) demonstrated that when Xi et al.'s (2014) 310 gene trees were inferred using the OV-slow characters, they had substantially greater average pairwise topological incongruence than when they were inferred from all characters.

1.2. Tree Independent Generation of Evolutionary Rates

Cummins and McInerney (2011:833) introduced TIGER as a method that "... estimates the relative evolutionary rate of each homologous character," with "... the similarity between characters as a proxy for evolutionary rate." Hence, in contrast to OV, TIGER is based on character congruence and more congruent characters are expected to be slower evolving than less congruent characters. This is a natural application of the idea that character congruence is a measure of phylogenetic signal (Archie, 1989; Faith and Cranston, 1991). The partition-agreement scores produced by TIGER range from zero to one, with higher scores indicating higher congruence.

Partition-agreement scores are calculated such that characters with state distributions that are inclusive of other characters' state distributions receive higher scores. Hence, for a character with two adenines and 98 thymines in a 100-terminal matrix, the thymines (i.e., the character state present in the majority of the terminals) are more likely to be inclusive of other characters' state distributions than would another character with 50 cytosines and 50 guanines. So OV and TIGER would prefer the same character in this example.

Like OV, TIGER has been applied to selectively delete putatively fast evolving characters from genomic datasets (e.g., Morgan et al., 2014; Xi et al., 2014; Katz and Grant, 2015). But unlike OV, TIGER has also been used to partition characters into rate categories for parametric analyses (e.g., Rota and Wahlberg, 2012; Heikkilä et al., 2014; Nakov et al., 2014). Typically character removal is performed by excluding one or more of the ten rate categories that TIGER delimited, with different numbers of characters assigned to each category (e.g., Greene et al., 2014; Morgan et al., 2014). The percentage of characters removed ranged from just 0.6% of all characters (Owen et al., 2015) to 76% of the variable characters (Feuda and Smith, 2015).

Most authors who applied TIGER to partition rate categories or remove characters favored their TIGER-based results. But others have expressed concern. Morgan et al. (2014:642) noted that removal of even one of their 20 TIGER-delimited rate categories "... resulted in an increase in phylogenetic conflict for the remaining 10 [of 13] mtGenes, suggesting that removal of site category 20 could be removing necessary phylogenetic signal." Katz and Grant (2015:411) noted that TIGER-based character deletion "... yields consistent topology with lower support for most clades." Sharma et al. (2015:3) "... observed major loss of phylogenetic signal upon removing sites ranked in one or more of the fastest evolving bins (of 10 equally sized bins), yielding a basal polytomy for two different matrices and the non-monophyly of scorpions."

As with OV, Simmons and Gatesy (2015) demonstrated that Xi et al.'s (2014) application of TIGER to eliminate half of the parsimony-informative characters from their dataset of 310 nuclear genes resulted in worse performance for the same three measures cited above for OV relative to the complete datasets, albeit generally not quite as bad as OV. This pattern also applied to the amount-of-possible-synapomorphy measure. But there was no consistent pattern for the ensemble-retention index (Farris, 1989) for the concatenation-based analyses of TIGER-fast and TIGER-slow characters. Finally, as with OV, Simmons et al. (2016) demonstrated that when Xi et al.'s (2014) 310 gene trees were inferred using the TIGER-slow characters, they had substantially greater average pairwise topological incongruence than when they were inferred from all characters.