Molecular Phylogenetics and Evolution 74 (2014) 1–14

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Limitations of locally sampled characters in phylogenetic analyses of sparse supermatrices

Mark P. Simmons *

Department of Biology, Colorado State University, Fort Collins, CO 80523, USA

ARTICLE INFO

Article history:
Received 10 June 2013
Revised 30 January 2014
Accepted 30 January 2014
Available online 14 February 2014

Keywords:
Bayesian unlinked
Character partition
Maximum likelihood
Missing data
Parsimony
Strict consensus

ABSTRACT

Empirical and simulated examples were used to demonstrate the following four points in the context of sparse supermatrices. First, locally sampled characters, when analyzed with low quality heuristic parsimony, likelihood, or Bayesian searches, can create high resolution and resampling values for clades that are properly unsupported because there is no comparable information among sets of terminals. Second, arbitrary factors that should have no effect on phylogenetic inference can create large fluctuations in congruence of trees inferred by parsimony, likelihood, and Bayesian methods with the simulated topology. Third, phylogenetic signal present in locally sampled characters may be interpreted in radically different ways depending upon the phylogenetic signal present in globally sampled characters. Fourth, application of Bayesian MCMC analyses with unlinked branch lengths among character partitions cannot be expected to universally obviate missing-data artifacts, even when numerous characters are sampled from each partition. The first three points may be addressed by conducting thorough tree searches while allowing numerous equally optimal trees to be saved from each replicate rather than relying entirely upon subtree pruning and regrafting (SPR) while saving a single optimal tree, as is the case in many contemporary empirical sparse-supermatrix analyses.

1. Introduction

It is widely recognized that locally sampled characters contain phylogenetic signal and generally improve phylogenetic inference (Wiens, 1998, 2003). Furthermore, adding terminals with extensive missing data may improve phylogenetic inference by breaking long branches (Wiens, 2005). Therefore, the serendipitous scaffolding approach (Johnson et al., 2012) of integrating all available characters largely irrespective of their missing-data content into a single supermatrix has been widely embraced. But locally sampled characters have their own limitations, including the potential to mislead parametric methods (Gatesy et al., 2002; Lemmon et al., 2009; Simmons, 2012a,b) and cause tree-search artifacts (Simmons and Goloboff, 2013).

Phylogenetic analyses that are based on sparse supermatrices are increasingly common, ranging from broad-scale studies that are based largely or entirely on publicly available sequence data (e.g., Peters et al., 2011; Springer et al., 2012), to studies that are based on partial genomic resources (e.g., restriction-site-associated DNA, transcriptomes; Meusemann et al., 2010; Wagner et al., 2013), studies that integrate novel organellar-genome data to resolve ancient lineages while also sampling a large number of taxa with comparatively few genes sampled to resolve more recently derived clades (Davis et al., 2013), as well as more conventional studies that are based on individually amplifying several genes while also integrating publicly available data (e.g., Chatrou et al., 2012; Simmons, 2012a,b; Guo et al., 2013). There are often dramatic disparities in the amount of missing (and inapplicable) data between character partitions, such that some partitions are sampled for nearly all terminals (i.e., "globally sampled"), whereas many others are sampled for just a minority of the terminals (i.e., "locally sampled"; e.g., Simmons and Goloboff, 2013). Note that "locally sampled" is based simply on the fraction of terminals scored and has no necessary correlation with being sampled for a single lineage.

Depending on the inferred tree topology and the distribution of missing data within each locally sampled character or character partition, that character may or may not be able to contribute an unambiguously optimized synapomorphy for a given clade (Malia et al., 2003; Sanderson et al., 2010; Simmons, 2012b). For example, consider a phylogenetic analysis that encompasses the Vertebrata, including 100 species of Amniota. If humans and chimps are the only members of Amniota scored for presence or absence of an amniotic egg, then presence of an amniotic egg would be

* Fax: +1 970 491 0649.
E-mail address: psimmons@lamar.colostate.edu

http://dx.doi.org/10.1016/j.ympev.2014.01.030
1055-7903/

ambiguously optimized on every branch between the clade of Amniota as a whole and the two-terminal clade of humans + chimps because the two-terminal clade is highly nested within Amniota. Alternatively, if humans and swans are the only members of Amniota scored for presence or absence of an amniotic egg, then presence of an amniotic egg would be unambiguously resolved as a synapomorphy for the Amniota as a whole because the most recent common ancestor of humans and swans is the most recent common ancestor of (extant) Amniota. In this second case the optimization obtained from the incompletely scored amniotic-egg character is identical to that which would be obtained if all members of Amniota were scored for presence or absence of an amniotic egg.

The effect of including a locally sampled character that provides phylogenetic signal albeit with ambiguous optimization, as with the human + chimp example described above, may vary depending upon how the character interacts with globally sampled characters and how the data are analyzed. Hence the locally sampled character may be clearly beneficial in one context yet misleading in another. An exploration of these different contexts is needed in order to provide general guidelines for assembly and phylogenetic analysis of sparse supermatrices, and that is the focus of this study.

In this study empirical and simulated examples were used to examine interactions between partitions of locally sampled characters relative to each other as well as to globally sampled characters in the context of different optimality criteria (parsimony, maximum likelihood, and Bayesian posterior probabilities; Kluge and Farris, 1969; Fitch, 1971; Felsenstein, 1973; Yang and Rannala, 1997), tree searches, and model partitioning. The results were used to demonstrate misleading artifacts that may occur based on interactions between globally and locally sampled partitions and how different distributions of missing data among locally sampled partitions can contribute to unstable results. Furthermore, the results were used to demonstrate the importance of thorough tree searches that allow multiple equally optimal trees to be held irrespective of which optimality criterion is employed, the extent of phylogenetic-inference artifacts that may occur when low quality tree searches are applied, and the limits of additional character sampling and model partitioning to address these artifacts.

The potential phylogenetic-inference artifacts described in this manuscript need to be considered at the following three stages of supermatrix analyses: (1) when determining which characters and terminals to include in a supermatrix, (2) which phylogenetic-inference methods to apply to the supermatrix, and (3) when assessing the branch support conferred by synapomorphies from locally sampled characters.

1.1. Three general principles

Three principles described in the following paragraphs may generally (but not always) apply when comparing resolution in the strict consensus of all optimal trees between a matrix that consists entirely of characters that are globally sampled for all terminals in the study, and the same matrix that also includes characters that are only locally sampled for some terminals in the study. These principles are relevant whenever one quantifies the phylogenetic signal contributed by undersampled character partitions to an existing data matrix that is largely complete. For the vast majority of macroscopic eukaryote phylogenetic studies being conducted, there are some publicly available data that can be scored for at least a minority of the terminals included in novel phylogenetic analyses. The added resolution and branch support may initially appear impressive, but cannot always be unequivocally justified, particularly in the context of low-quality heuristic tree searches.

The first principle is that any clade resolved in the strict consensus for the global + local matrix can generally only be unequivocally supported if it is a convex group (i.e., monophyletic or paraphyletic) that is bounded by branches that are supported by synapomorphies from the globally sampled partition. The reason for this first principle is that the locally sampled characters are only informative for inferring relationships among terminals that are sampled for those characters relative to each other. By definition those locally sampled characters cannot exclude any terminals that they are not sampled for from any of the clades that they delimit. That is, the unsampled terminals should behave as wildcards (Nixon and Wheeler, 1991) and collapse any such clades in the strict consensus unless those terminals are excluded from the clade in question by the globally sampled characters.

The first principle is fulfilled whenever one of the following three cases apply: (1) the clade that is present in the strict consensus for the global + local matrix is also present in the global-only matrix, (2) the clade in question is nested within another larger clade (in an unrooted sense) that is also present in the strict consensus for the global-only matrix and that larger clade only includes terminals scored for the same locally sampled characters, or (3) the clade in question encompasses another smaller clade (in an unrooted sense) that is also present in the strict consensus for the global-only matrix; this smaller clade includes any terminals that are not scored for the locally sampled character(s) that support the clade in question. The difference between the second and third cases is that the second case focuses on larger clades that encompass the clade in question whereas the third case focuses on a sub-clade within the clade in question.

The second principle is that any clades present in the strict consensus for the global + local matrix but not in the strict consensus for the global-only matrix can generally only be unequivocally supported if either the second or third conditions described for the first principle apply. Otherwise, any unsampled terminal should act as a wildcard within the larger clade and collapse any such nested clades in the strict consensus for the same reason cited above for the first principle.

The third principle is that the resampling-based branch-support assigned to a given clade resolved by the global + local matrix generally should not be higher than that provided for the same clade (or a larger clade that contains it when the second condition applies, or a smaller clade that includes the terminal that is not scored for the locally sampled characters when the third condition applies) in the strict consensus for the global-only matrix. The rationale for this principle is that if the globally sampled character(s) that provide support for the given clade are not sampled in a bootstrap (BS; Felsenstein, 1985) or jackknife (Farris et al., 1996) pseudoreplicate then neither that clade nor any nested clades within it can be unequivocally supported by the locally sampled characters alone.

An example wherein the first and second principles do not hold is presented in Fig. 1. An example that demonstrates the second principle and also shows a case where the third principles does not hold is presented in Fig. 2.

2. 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,b)), to which 51 Madagascan terminals were added by C.D. Bacon et al. (unpublished data). This is the same matrix