Molecular Phylogenetics and Evolution 94 (2016) 146–158

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Effects of missing data on topological inference using a Total Evidence approach ☆

Thomas Guillerme [a,b,*], Natalie Cooper [a,b,c]

[a] School of Natural Sciences, Trinity College Dublin, Dublin 2, Ireland
[b] Trinity Centre for Biodiversity Research, Trinity College Dublin, Dublin 2, Ireland
[c] Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK

ARTICLE INFO

Article history:
Received 8 June 2015
Revised 24 August 2015
Accepted 26 August 2015
Available online 31 August 2015

Keywords:
Morphological characters
Bayesian
Maximum Likelihood
Topology
Fossil
Living

ABSTRACT

To fully understand macroevolutionary patterns and processes, we need to include both extant and extinct species in our models. This requires phylogenetic trees with both living and fossil taxa at the tips. One way to infer such phylogenies is the Total Evidence approach which uses molecular data from living taxa and morphological data from living and fossil taxa.

Although the Total Evidence approach is very promising, it requires a great deal of data that can be hard to collect. Therefore this method is likely to suffer from missing data issues that may affect its ability to infer correct phylogenies.

Here we use simulations to assess the effects of missing data on tree topologies inferred from Total Evidence matrices. We investigate three major factors that directly affect the completeness and the size of the morphological part of the matrix: the proportion of living taxa with no morphological data, the amount of missing data in the fossil record, and the overall number of morphological characters in the matrix. We infer phylogenies from complete matrices and from matrices with various amounts of missing data, and then compare missing data topologies to the "best" tree topology inferred using the complete matrix.

We find that the number of living taxa with morphological characters and the overall number of morphological characters in the matrix, are more important than the amount of missing data in the fossil record for recovering the "best" tree topology. Therefore, we suggest that sampling effort should be focused on morphological data collection for living species to increase the accuracy of topological inference in a Total Evidence framework. Additionally, we find that Bayesian methods consistently outperform other tree inference methods. We therefore recommend using Bayesian consensus trees to fix the tree topology prior to further analyses.

1. Introduction

Although most species that have ever lived are now extinct (Novacek and Wheeler, 1992; Raup, 1981), many large-scale macroevolutionary studies focus solely on living species (e.g. Meredith et al., 2011; Jetz et al., 2012). Ignoring fossil taxa may lead to misinterpretation of macroevolutionary patterns and processes such as the timing of diversification events (e.g. Pyron, 2011), relationships among lineages (e.g. Manos et al., 2007) or niche occupancy (e.g. Pearman et al., 2008). This has led to increasing consensus among evolutionary biologists that fossil taxa should be included in macroevolutionary studies (Jackson and Erwin, 2006; Quental and Marshall, 2010; Dietl and Flessa, 2011; Slater and Harmon, 2013; Fritz et al., 2013). To do this, however, we need to be able to place living and fossil taxa into the same phylogenies; a task that remains difficult despite recent methodological developments (e.g. Pyron, 2011; Ronquist et al., 2012a; Matzke, 2014).

Up to now, three main approaches have been used to place both living and fossil taxa into phylogenies. These approaches differ mainly in how they treat fossil taxa and their data. One can use fossils as tips or as nodes in the phylogeny, and can use only the age of the fossils, only the morphology of the fossils, or age and morphology jointly. Classical cladistic methods use matrices containing morphological data from both living and fossil taxa and treat each taxon as a tip in the phylogeny. Relationships among the taxa are then inferred using optimality criteria such

☆ This paper was edited by the Associate Editor Liliana M. Davalos.

* Corresponding author at: Zoology Building, Trinity College Dublin, Dublin 2, Ireland. Fax: +353 1 6778094.
E-mail address: guillert@tcd.ie (T. Guillerme).

http://dx.doi.org/10.1016/j.ympev.2015.08.023
1055-7903/

as maximum parsimony (Hennig, 1966; Felsenstein, 2004). This approach is commonly used by palaeontologists but it ignores the additional molecular data available from living species and does not allow use of probabilistic methods for dealing with phylogenetic uncertainty. Neontologists, on the other hand, more commonly use probabilistic approaches (e.g. Maximum Likelihood or Bayesian methods) based on matrices containing only molecular data from living species. Because fossil taxa do not usually have available DNA, only fossil occurrence dates are used to time calibrate phylogenies (Zuckerkandl and Pauling, 1965). There have been great improvements in the theory and application of these two approaches (e.g. Bapst, 2013; Stadler and Yang, 2013; Heath et al., 2014) as well as much debate about the "best" approach to use (e.g. Spencer and Wilberg, 2013; Wright and Hillis, 2014). Neither approach, however, uses all the available data.

A final approach, known as the Total Evidence method, uses matrices containing molecular data from living taxa and morphological data from both living and fossil taxa (Eernisse and Kluge, 1993). This approach treats every taxa as a tip in the phylogeny, uses the occurrence age of the fossils to time calibrate the phylogeny (known as tip-dating; Ronquist et al., 2012a), and allows the use of probabilistic methods for estimating phylogenetic uncertainty (Ronquist et al., 2012a). The Total Evidence method is becoming an increasingly popular way of adding fossil taxa to phylogenies (e.g. Pyron, 2011; Ronquist et al., 2012a; Schrago et al., 2013; Slater, 2013; Beck and Lee, 2014; Arcila et al., 2015). Although the Total Evidence approach seems very promising, there is one big drawback in using this approach: it requires both molecular and morphological data, both of which can be difficult (or impossible) to collect for every living and fossil taxon in the tree. Morphological data for living taxa are rarely collected when molecular data are available (e.g. O'Leary et al., 2013 *vs.* Meredith et al., 2011), and for fossil taxa, data can only be collected from features preserved in the fossil record. For example, in vertebrates, the hardest parts of the skeleton are more often preserved than soft parts (Sansom and Wills, 2013); and molecular data are (nearly) always unavailable. Therefore Total Evidence matrices are likely to contain a large proportion of missing data that may affect the method's ability to infer correct topologies, branch lengths and support values (Salamin et al., 2003).

Although missing data do not appear be a major problem in molecular and morphological matrices separately (as long as enough data overlap in each case, and missing data are not phylogenetically biased; Wiens, 2003; Wiens et al., 2005; Wiens, 2006; Wiens and Moen, 2008; Lemmon et al., 2009; Sanderson et al., 2011; Roure and Philippe, 2011; Pattinson et al., 2014), it may become more of an issue in Total Evidence matrices containing both molecular and morphological data for living and fossil taxa. This may be particularly problematic as fossil taxa (generally) do not have molecular data, resulting in a large section of missing data in Total Evidence matrices. Until now, few attempts have been made to study the impact of this missing data issue on phylogenetic inference in a Total Evidence framework (i.e. using both molecular and morphological data; Wiens et al., 2005; Manos et al., 2007; Pattinson et al., 2014). These previous studies assessed the effect of missing data on topology by either (1) comparing a dataset with missing data to subsets without missing data (Wiens et al., 2005); or (2) removing both molecular and some morphological data from living taxa to create artificial fossils (Manos et al., 2007; Pattinson et al., 2014). Both approaches have shown that missing data are not a major problem and should not be an obstacle to combining both living and fossil species in the same phylogenies. The way these studies were conducted, however, means that their conclusions are not generally applicable across all scenarios involving missing data in Total Evidence phylogenies. For example, using an empirical (rather than simulation based) approach limits their conclusions to studies with similar distributions of data across species in the phylogeny. Additionally, one of the three previous studies did not include fossil taxa in their analyses, so their results cannot be used to make conclusions about how missing data may influence the placement of fossils (Wiens, 2003). The other two studies did include fossil taxa, but used the patchiness of the fossil record to determine how to remove data from their matrices (Manos et al., 2007; Pattinson et al., 2014). Data for living species are unlikely to be missing in this patchy way, instead full molecular data with the complete absence of morphological data is a likely pattern (Guillerme and Cooper, 2015). Finally, these previous studies mainly focused on how missing data in fossil taxa affect the placement of fossils, ignoring the effects of missing data in living species (Manos et al., 2007; Pattinson et al., 2014).

In this study, we propose a theoretical assessment of the effect of missing data in the Total Evidence method by removing living taxa with morphological data, fossil data, all data for certain characters and the combination of these three aspects. This is an advance on previous studies because we use large-scale simulations and analyse the effects of three distinct aspects of missing data thus focusing on both neontological and palaeontological parts of the matrix. In addition, we test the effect of missing data by measuring two crucial aspects of topology in both Maximum Likelihood and Bayesian phylogenies: (i) the conservation of clades (based on the Robinson–Foulds distance; Robinson and Foulds, 1981) and (ii) the displacement of wild-card taxa (based on the Triplets distance; Critchlow et al., 1996) rather than just a single measure of clade conservation or clade support (cf. Wiens et al., 2005; Pattinson et al., 2014).

We focus on the effects of missing data on our ability to recover tree topology because it is a crucial aspect of a phylogeny in many macroevolutionary studies, for example when trying to elucidate the evolutionary relationships among species (e.g. Meredith et al., 2011; Jetz et al., 2012), or for studying evolutionary transitions (e.g. Friedman, 2010). Although branch length estimation is also important (namely for timing extinction and/or speciation events; e.g. Ronquist et al., 2012a), we do not consider branch lengths in this study. This is partially due to difficulties with simulating branch lengths and topology simultaneously, but also because previous studies have already empirically assessed the effect of the Total Evidence method on branch length variation but using topological constraints (Ronquist et al., 2012a; Schrago et al., 2013; Slater, 2013; Beck and Lee, 2014). Thus understanding the sensitivity of topology to missing data is important for assessing the accuracy of tree estimation in the Total Evidence framework. To our knowledge, this question has never been formally assessed.

Here we use a simulation approach to assess the effect of missing data on tree topologies inferred from Total Evidence matrices. Since the molecular part of a Total Evidence matrix acts like a "classical" molecular matrix containing only the living taxa (Ronquist et al., 2012a), the effect of missing data on such matrices is well known (Wiens, 2006; Wiens and Moen, 2008; Lemmon et al., 2009; Roure and Philippe, 2011). Therefore, we focus only on missing data in the morphological part of the matrix. We investigate three major parameters that directly affect the completeness and size of the morphological part of the matrix, and reflect empirical biases in data availability: (i) the proportion of living taxa with no morphological data; (ii) the proportion of missing data in the fossil taxa; and (iii) the amount of morphological characters for both living and fossil taxa in the matrix (i.e. the size of the matrix). We remove data from a Total Evidence matrix by changing the values of these three parameters and then assess how this affects the resulting tree topology. We infer the topology from the matrices using both Maximum Likelihood and Bayesian inference methods and measure the differences in topology using two different