Molecular Phylogenetics and Evolution 80 (2014) 308–318

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Should genes with missing data be excluded from phylogenetic analyses?

Wei Jiang [a,b,c], Si-Yun Chen [b], Hong Wang [a,b,c], De-Zhu Li [a,b,c,*], John J. Wiens [d]

[a] Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China
[b] Plant Germplasm and Genomics Center, Germplasm Bank of Wild Species, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China
[c] Kunming College of Life Sciences, University of Chinese Academy of Sciences, Kunming, Yunnan 650201, China
[d] Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721-088, USA

ARTICLE INFO

Article history:
Received 25 January 2014
Revised 15 July 2014
Accepted 3 August 2014
Available online 11 August 2014

Keywords:
Accuracy
Maximum likelihood
Missing data
Phylogeny

ABSTRACT

Phylogeneticists often design their studies to maximize the number of genes included but minimize the overall amount of missing data. However, few studies have addressed the costs and benefits of adding characters with missing data, especially for likelihood analyses of multiple loci. In this paper, we address this topic using two empirical data sets (in yeast and plants) with well-resolved phylogenies. We introduce varying amounts of missing data into varying numbers of genes and test whether the benefits of excluding genes with missing data outweigh the costs of excluding the non-missing data that are associated with them. We also test if there is a proportion of missing data in the incomplete genes at which they cease to be beneficial or harmful, and whether missing data consistently bias branch length estimates. Our results indicate that adding incomplete genes generally increases the accuracy of phylogenetic analyses relative to excluding them, especially when there is a high proportion of incomplete genes in the overall dataset (and thus few complete genes). Detailed analyses suggest that adding incomplete genes is especially helpful for resolving poorly supported nodes. Given that we find that excluding genes with missing data often decreases accuracy relative to including these genes (and that decreases are generally of greater magnitude than increases), there is little basis for assuming that excluding these genes is necessarily the safer or more conservative approach. We also find no evidence that missing data consistently bias branch length estimates.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

The problem of missing data in phylogenetic analysis is an important issue because missing data are common in many data matrices (e.g., Philippe et al., 2004; Fulton and Strobeck, 2006; Burleigh et al., 2009), and are only absent in many others because taxa and genes are deliberately excluded in order to avoid them. For example, the issue of missing data may arise because of gaps in alignments, because data are unavailable for some species for some genes, or because molecular data are lacking entirely (e.g., fossils). There has been extensive debate about whether missing data should be included in phylogenetic analyses or not, and the possible consequences of both approaches (e.g., Huelsenbeck, 1991; Wiens and Reeder, 1995; Wiens, 1998, 2003a,b, 2005; Driskell et al., 2004; Philippe et al., 2004; Wiens et al., 2005, 2010; Wiens and Moen, 2008; Burleigh et al., 2009; Lemmon et al., 2009; Sanderson et al., 2010, 2011; Wiens and Morrill, 2011; Wiens and Tiu, 2012; Roure et al., 2013). In this debate, it is important to remember that missing data cells are only included because excluding missing data also requires excluding some taxa and/or characters from the analysis, which have non-missing data (Wiens, 1998; Cho et al., 2011; Schaefer and Renner, 2011; Zwick et al., 2011). The fundamental question is: when do the benefits of excluding the missing data outweigh the costs of excluding the non-missing data that are associated with them?

Missing data can be added to an analysis by two primary mechanisms: by adding incomplete taxa or by adding incomplete characters (Wiens, 2003a). Many studies have shown that incomplete taxa can often be included with relatively limited negative impacts, especially when the number of characters is large (e.g., Wiens, 2003b; Driskell et al., 2004; Philippe et al., 2004; Wiens and Moen, 2008; Cho et al., 2011; Wiens and Morrill, 2011; Wiens and Tiu, 2012; Roure et al., 2013). Specifically, these studies show

* Corresponding author at: Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China.
E-mail addresses: jiangwei@mail.kib.ac.cn (W. Jiang), chensiyun@mail.kib.ac.cn (S.-Y. Chen), wanghong@mail.kib.ac.cn (H. Wang), dzl@mail.kib.ac.cn (D.-Z. Li), wiensj@email.arizona.edu (J.J. Wiens).

http://dx.doi.org/10.1016/j.ympev.2014.08.006
1055-7903/© 2014 Elsevier Inc. All rights reserved.

that incomplete taxa can be placed correctly in phylogenies (based on simulations with a known true topology or based on concordance with other empirical studies), when sufficient characters have been sampled overall (review in Wiens and Morrill, 2011). Some studies have also shown that adding incomplete taxa can improve the accuracy of estimated relationships among the complete taxa (by breaking up long branches), using both simulated data (Wiens, 2005) and empirical data (Wiens and Tiu, 2012; Roure et al., 2013). In other words, adding incomplete taxa can potentially have similar benefits to adding complete taxa in these cases.

Far fewer studies have addressed the costs and benefits of adding characters with missing data. In a simulation study, Wiens (1998) found that for parsimony analyses adding incomplete characters was often beneficial, but became less beneficial with a greater proportion of missing data. Although this study found little evidence that adding characters with missing data generally decreased accuracy, it also showed that some patterns of missing data could create a problem of long-branch attraction among the species with non-missing data. Lemmon et al. (2009) analyzed simulations of the 4-taxon case and suggested that missing data could cause misleading results in Bayesian analyses with missing data in 2 of 4 taxa, especially when combining data from genes with very low rates of change (approaching invariant data) and very high rates (effectively randomized data). Wiens and Morrill (2011) found that in simulations utilizing rates and numbers of taxa more typical of empirical phylogenetic studies, adding characters with missing data tended to either increase or have little effect on mean accuracy for Bayesian phylogenetics. However, all three of these simulation studies were relatively simplistic. For example, none explored more realistic situations with multiple genes where gene topologies could potentially disagree. Nevertheless, discordance among gene trees is pervasive in empirical multi-locus datasets (Rokas et al., 2003; Cranston et al., 2009), especially when the underlying species topology includes one or more relatively short branches (e.g., Wiens et al., 2008).

Thus, a critical but unresolved question for empirical systematists is whether it is better to include or exclude genes that have some missing data. Specifically, do the benefits of increasing the number of genes outweigh the potential consequences of increasing the overall amount of missing data? This question is particularly relevant for short nodes that are difficult to resolve, nodes which may require the addition of many genes to resolve (e.g., Rokas et al., 2003) but for which gene topologies are especially likely to disagree (e.g., Wiens et al., 2008). Some empirical results on this issue were obtained by Wiens et al. (2005) and Cho et al. (2011), who both found that adding incomplete genes seemed to give more well-supported results that were more consistent with previous taxonomy and phylogenetic estimates (whereas excluding incomplete genes gave weaker support and/or relationships inconsistent with previous taxonomy and phylogenetic hypotheses). However, these authors did not perform detailed experiments examining the impact of missing data in the added genes.

In this study, we use analyses of real datasets to explore the consequences of including versus excluding genes with missing data on the accuracy of concatenated likelihood analyses. We use the similarity of the estimated trees to the phylogeny based on the complete data as a proxy for accuracy (which we define as the similarity of the estimated tree to the true phylogeny). We analyze data from yeast to represent datasets with many genes and extensive genetic divergence among taxa (despite most taxa being congeners in this case) and a dataset from plants representing those with fewer genes and more limited genetic divergence among taxa (despite many species being in different families). We analyze these datasets to address the following questions: (1) is accuracy of concatenated likelihood analyses increased or decreased by adding genes with missing data? (2) If adding genes with missing data is beneficial, is there a proportion of missing data at which adding these incomplete genes ceases to be useful? (3) If adding genes with missing data is detrimental, at what proportion of missing data does this occur? (4) How do the advantages and disadvantages of adding incomplete genes change with the overall number of genes in the analysis?

We also test for potential biases in branch length estimation caused by including versus excluding genes with missing data. Accurate branch-length estimates may be critically important for phylogenetic comparative analyses and for divergence-date estimation. Some authors have suggested that missing data can lead to strongly biased and inaccurate estimates of branch lengths (i.e., Lemmon et al., 2009), whereas other authors have suggested that those results may have been artifacts of the methods used by those authors (e.g., Wiens and Morrill, 2011; Roure et al., 2013). At least two recent studies have tested for biases in branch-length estimation caused by missing data in empirical datasets, and found no evidence for such biases (Pyron et al., 2011; Wiens and Tiu, 2012). Here, we explicitly contrast the impacts of including versus excluding genes with missing data on branch-length estimation, comparing these estimated branch lengths to those from the complete datasets with all sampled genes.

2. Materials and methods

2.1. Yeast data

2.1.1. Basic information on the yeast dataset

We selected an empirical dataset consisting of 8 yeast species (Rokas et al., 2003) and 106 orthologous genes. The dataset includes seven species of *Saccharomyces*, with a more distant relative (*Candida albicans*) included as an outgroup. There are very few missing data in the original data set (only 0.0063%). Separate analyses of each gene revealed considerable discordance among the estimated gene trees (Rokas et al., 2003). However, combining all genes yielded a single tree with 100% likelihood bootstrap values at every branch (Fig. 1; Rokas et al., 2003). The same topology was also found using a coalescent-based species-tree approach (BEST; Edwards et al., 2007). Therefore, we assumed that this tree reflects the true phylogenetic relationships among these eight species.

2.1.2. Design of missing data experiments

The overall design of the yeast experiments was as follows. First, we estimated a phylogeny for the complete data (106 genes). We then created smaller datasets by randomly sampling smaller numbers of genes (5, 10, 20, and 50), creating 100 new data matri-

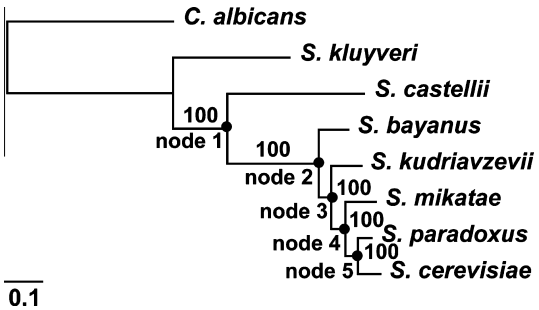

Fig. 1. Maximum likelihood estimate of phylogeny for 8 yeast species (seven species of *Saccharomyces* and an outgroup, *Candida albicans*) based on concatenated analysis of 106 genes (originally from Rokas et al., 2003), showing numbered nodes (for assessing accuracy) and bootstrap support.

Download English Version:

https://daneshyari.com/en/article/2833821

Download Persian Version:

https://daneshyari.com/article/2833821

Daneshyari.com