Molecular Phylogenetics and Evolution 69 (2013) 1057–1062

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

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

Effects of missing data on species tree estimation under the coalescent

Rasmus Hovmöller [a], L. Lacey Knowles [b], Laura S. Kubatko [a,c,*]

[a] Department of Statistics, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, United States
[b] Department of Ecology and Evolutionary Biology, Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079, United States
[c] Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210, United States

ARTICLE INFO

Article history:
Received 2 March 2013
Accepted 5 June 2013
Available online 13 June 2013

Keywords:
Species tree inference
Multilocus data
Coalescent model

ABSTRACT

With recent advances in genomic sequencing, the importance of taking the effects of the processes that can cause discord between the speciation history and the individual gene histories into account has become evident. For multilocus datasets, it is difficult to achieve complete coverage of all sampled loci across all sample specimens, a problem that also arises when combining incompletely overlapping datasets. Here we examine how missing data affects the accuracy of species tree reconstruction. In our study, 10- and 100-locus sequence datasets were simulated under the coalescent model from shallow and deep speciation histories, and species trees were estimated using the maximum likelihood and Bayesian frameworks (with STEM and *BEAST, respectively). The accuracy of the estimated species trees was evaluated using the symmetric difference and the SPR distance. We examine the effects of sampling more than one individual per species, as well as the effects of different patterns of missing data (i.e., different amounts of missing data, which is represented among random taxa as opposed to being concentrated in specific taxa, as is often the case for empirical studies). Our general conclusion is that the species tree estimates are remarkably resilient to the effects of missing data. We find that for datasets with more limited numbers of loci, sampling more than one individual per species has the strongest effect on improving species tree accuracy when there is missing data, especially at higher degrees of missing data. For larger multilocus datasets (e.g., 25–100 loci), the amount of missing data has a negligible effect on species tree reconstruction, even at 50% missing data and a single sampled individual per species.

1. Introduction

Traditional single-locus phylogenetic tree inference typically involves consideration of only the mutational process for explaining the variation seen in the sampled DNA sequences, with the result that the history of the gene analyzed is assumed to reflect the history of the species. However, it has long been recognized (Maddison, 1997, and references therein) that other processes influence sequence variation, and that their effects should be taken into account when performing phylogenetic analysis of molecular data. Such processes can create discord between the speciation history and the trees inferred from the observed gene sequences by moving genes across species boundaries (horizontal transfer), obscuring the difference between paralogous and orthologous genes (gene duplication and extinction), and by the retention of ancestral polymorphisms resulting in a gene genealogy that does not reflect the speciation history (deep coalescence).

While standard sequencing practices target specific areas of the genome, next generation sequencing technologies generate randomly sampled sequences from across the genome. Consequently, datasets collected by such sequencing (e.g., data from the 454-sequencing and Illumina platforms) will have characteristics that differ from traditional sequencing datasets (i.e., generated by Sanger sequencing) from a phylogenetic perspective: they will have many loci where the gene history is more strongly influenced by the coalescent than the mutational process (i.e., they have not been vetted for specific levels of nucleotide variation), and since the sampling of loci is more or less random for each sampled individual, there will be a large portion of missing data in the full dataset when loci are counted as missing/present. This suggests the need for methods for analyzing multilocus data that take the coalescent process into account while remaining robust with respect to a large proportion of missing data.

The coalescent process (Kingman, 1982; Hudson, 1983; Tajima, 1983) is commonly used to model the process of deep coalescence in phylogenetics to estimate species trees (Knowles and Kubatko, 2010). The coalescent process describes how gene histories are influenced by historical population size and time intervals between speciation events: allelic variation is more likely to be retained in a large population, or across speciation events within short time

* Corresponding author at: Department of Statistics, The Ohio State University, 404 Cockins Hall, 1958 Neil Avenue, Columbus, OH 43210, United States. Tel. +1 614 247 8846.
E-mail address: Kubatko.2@osu.edu (L.S. Kubatko).

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.06.004

intervals as one allele is less likely to become fixed under those conditions, increasing the probability of each gene history being affected by deep coalescence. Because the coalescent can be used to compute the likelihood of observing each gene history given a species tree, algorithms can be devised that take the effect of deep coalescence into account when estimating a phylogeny from a multilocus dataset.

Methods that explicitly include the coalescent process as part of the estimation method in a probabilistic framework can be classified into two groups: those that use sequence data as input, and those that use estimated gene trees as input. The two popular methods that take sequence data as input are BEST (Liu and Pearl, 2007; Liu et al., 2008) and *BEAST (Heled and Drummond, 2010). Both of these methods use a Bayesian framework to estimate the species trees given a multilocus dataset under the coalescent model. The most widely used probabilistic method based on gene tree data as input is STEM (Kubatko et al., 2009), which finds the maximum likelihood species tree for the input collection of gene trees under the coalescent model. The method is based on the Maximum Tree algorithm (Liu, 2006; Mossel and Roch, 2010; Liu et al., 2010a,b).

A primary disadvantage of methods like STEM (see also methods such as MDC (Maddison and Knowles, 2006; Than and Nakhleh, 2009, 2010), ST-ABC (Fan and Kubatko, 2011), STAR and STEAC (Liu et al., 2009) and NJ_{st} (Liu and Yu, 2011)) that use gene trees as input is that variability in the gene trees is not taken into account (but see Knowles et al., submitted for publication). When gene trees are assumed to be known, variation in the sequence data leading to the estimated gene trees is ignored. Several authors (Huang et al., 2010; Knowles, 2009; Liu et al., 2009; McCormack et al., 2009; Than and Rosenberg, 2011) have studied the effect of ignoring this source of variability, and have found that in general these methods perform well as long as their particular basic assumptions are not violated. Although methods based on sequence data, such as BEST and *BEAST, do not suffer from this shortcoming, this comes at the expense of increased time required to carry out the computations. In some cases, convergence is difficult to achieve and/or the amount of run time required to achieve convergence in all parameters is prohibitive (e.g., Cranston et al., 2009; Kubatko et al., 2011). Moreover, the amount of computation time required for sequence-based methods increases with the number of sequences included in the analysis, whereas for methods like STEM that use gene trees as input, the estimate of the species tree is returned rapidly, regardless of the size of the sample.

However, of the current software implementations, STEM (Kubatko et al., 2009) is the only application that explicitly allows for missing data. This is made possible in STEM because the Maximum Tree algorithm (Liu, 2006; Liu et al., 2010a,b) provides an analytical solution to the problem of finding the ML species tree that is based on the set of observed coalescent times across gene trees. When a coalescent time is not able to be estimated for a particular gene, as would be the case for any coalescent event involving a taxon that does not have a sequence available for the gene under consideration, that coalescent time simply does not contribute to forming the ML estimate of the divergence time for that node in the species tree. Thus missing data can be handled naturally in STEM without invoking any special assumptions, except that missing data are missing at random (i.e., the fact that the time is missing is independent of the coalescent time itself, and thus, missing data does not introduce a bias for the method). Methods based on sequence data, on the other hand, are expected to be adversely affected by a large proportion of missing data. At present, neither BEST nor *BEAST explicitly allow missing data. However, a user could fill an entire locus with the '?' character to reflect missing data for a taxon at that locus, but the likely result of such a procedure is increased difficulty in achieving convergence. This would be expected to occur because there will be no information in the data concerning the placement of such a taxon in the genealogy.

Here we present a simulation study to examine the effect of missing data on the performance of species tree inference using STEM and *BEAST. The aim is to determine how adding more loci, more coverage or more individuals per species improves the accuracy of reconstructing the correct species tree, and how efficient the method is under a range of parameters such as tree depth, number of individuals per species, and total number of sampled loci and individuals. We make general conclusions with regard to all of these factors for STEM, while examining a subset of our simulated data using *BEAST.

2. Methods

The workflow for the simulation study can be summarized by the following steps: (1) generating species trees, (2) generating gene trees from the species trees under the coalescent, (3) generating sequence data from the gene trees, (4) estimating maximum likelihood gene trees from the sequence data, (5) reconstructing species trees from the estimated gene trees, and (6) comparing accuracy of the estimated species trees to the trees originally generated in step 1. Generation of species trees, gene trees and sequence data follows the procedure outlined by Huang et al. (2010), and will be briefly summarized below.

Species trees with eight terminal taxa were generated by Mesquite (Maddison and Maddison, 2010) under the Yule process. These species trees were used as input for generating gene tree topologies using ms (Hudson, 2002) under a basic coalescent model that assumes a constant population size, no migration or horizontal transfer, and free recombination between loci. We considered total sample sizes of 10 and 100 loci. For the 10 locus datasets, gene trees were generated with 1, 3 or 9 individuals per species, while for the 100 locus datasets, 1 or 3 individuals per species were used. Gene trees were generated under assumed tree depths of 1N and 10N generations, where N is the effective population size.

Sequence data were generated with Seq-Gen (Rambaut and Grassly, 1997) under the HKY85 model (Hasegawa et al., 1985). The fixed parameters for the sequence generation are the same as those used in Huang et al. (2010): a transition/transversion ratio of 3.0, gamma mutation rate shape = 0.8 and a nucleotide frequency distribution of ($\pi_A = 0.3$, $\pi_C = 0.2$, $\pi_G = 0.3$, $\pi_T = 0.2$). For each gene tree, 1000 base pairs were generated for each terminal taxon.

For each dataset, 10%, 25% or 50% of the sequences were removed. The missing data were generated by deleting entire sequences at a locus using one of two patterns: 'bad terminals' or random allocation. The 'bad terminal' process is intended to simulate degraded input material (i.e. specimens with degraded DNA) as would be expected from historical material, where sequencing was unsuccessful for several taxa at one or more loci. In this case, a number of terminal taxa are designated as 'bad' following a Poisson distribution with λ set to half the number of terminal taxa, and the missing loci are distributed over the 'bad' taxa according to a multinomial distribution with equal probabilities for all categories. The random allocation of bad loci is equally likely to pick any locus from any terminal taxon. For the 100 locus datasets, missing data was only distributed by the random pattern, so as to emulate the pattern of missing data from next generation sequencing. This decision also reflects the fact that we observed very little difference in the two missing-data patterns in the 10-locus datasets (see below for details). We also included a series of simulations for the random pattern in which 25, 50, 75, or 100 loci were used at both 1N and 10N total tree depths over the range of missing data percentages.