Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

How low can you go? The effects of mutation rate on the accuracy of species-tree estimation

Hayley C. Lanier [*,a,b], Huateng Huang [a], L. Lacey Knowles [a]

[a]Department of Ecology and Evolutionary Biology, Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079, USA
[b]Department of Zoology and Physiology, University of Wyoming-Casper, Casper, WY 82601, USA

ARTICLE INFO

Article history:
Received 6 June 2013
Revised 12 August 2013
Accepted 6 September 2013
Available online 21 September 2013

Keywords:
*BEAST
Locus variability
Multispecies coalescent
Phylogenetic accuracy
Sampling effort
STEM

ABSTRACT

Although species-tree methods have been widely adopted for multi-locus data, little consideration has been given to the source and character of the loci used in these approaches. Decisions about which loci to target in empirical studies are typically constrained by availability, technology and funds – characteristics that are not typically considered in simulation studies. As a result, most real-world datasets often combine one or two variable loci (such as mtDNA or chloroplast loci) with multiple lower-variation loci to estimate species trees. These locus selections impact the accuracy and the resolution of a phylogeny. Furthermore, the fact that using a larger sample of loci can result in lower posterior probabilities has been used as an excuse to drop loci from an analysis. Here we address these issues directly through a simulation approach designed to mimic situations arising in empirical datasets by combining loci with differing mutation rates. We show that low-variation loci can be utilized in species-tree analyses that account for gene-tree uncertainty (e.g., a Bayesian framework), whereas maximum likelihood approaches show no improvement in accuracy when low-variation loci are added. We demonstrate that limited phylogenetic signal associated with low-variation loci constrains gains in species-tree estimation accuracy when adding loci. Lastly, we demonstrate that the inclusion of only a handful of loci with higher mutation rates, and hence greater phylogenetic information content, can make a tremendous difference in the accuracy of species-tree estimates, suggesting that empiricists should consider the quality, and not just quantity, of loci in multi-locus phylogenetic analyses.

1. Introduction

Empirical phylogeneticists are increasingly beset by difficult choices regarding marker selection. One might opt for a candidate locus approach, using previously characterized loci that can be broadly amplified across a wide range of taxa or focus on developing anonymous markers for the specific taxa under study (e.g., developing primers from genomic libraries, Carstens and Knowles 2007, or sequence data from next generation sequencing technologies, Faircloth et al., 2012). These alternative approaches generally result in tradeoffs between data quantity and data quality, ease of maker development, cost, amplification and alignment across a relevant set of taxa, and marker variability. Although in general 'more loci are better' (Edwards et al., 2007; Huang et al., 2010; Maddison and Knowles, 2006), empiricists are still left with little guidance as to how to select different types of loci, how many loci are sufficient, and in particular, how the variability of their chosen loci will impact the accuracy of species-tree estimation.

Despite its central importance, the effects of locus variability have been sorely neglected in studies of species-tree methodological efficacy. Most simulation studies examine the accuracy of species-tree methodologies using loci simulated with a set variation rate, and focus on methods and sampling (e.g., Leache and Rannala, 2010; McCormack et al., 2009). These studies have demonstrated that when the number of loci sampled is great, species-tree methods can be much more accurate than concatenation, especially in certain regions of phylogenetic space, such as the anomaly zone (Degnan and Rosenberg, 2009; Huang and Knowles, 2009; Kubatko and Degnan, 2007). However, many empirical studies combine loci with differing levels of variation – often pairing a variable mitochondrial locus with several lower-variation nuclear loci (e.g., Amaral et al., 2012; Nyári and Joseph, 2012; Oneal et al., 2010). It is unknown whether the phylogenetic signal from the variable loci will dominate if the nuclear loci are not very informative (Fig. 1). As many empiricists are publishing datasets which rely primarily on loci amplified using universal primers developed for addressing deep phylogenetic questions, this is not a trivial matter. Often the resulting species tree bears a strong resemblance to the variable locus, with the low-variation loci appearing to add little to the reconstruction. Variability is also an important consideration

[*] Corresponding author.
E-mail addresses: hlanier@uwyo.edu (H.C. Lanier), huatengh@umich.edu (H. Huang), knowlesl@umich.edu (L.L. Knowles).

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.09.006

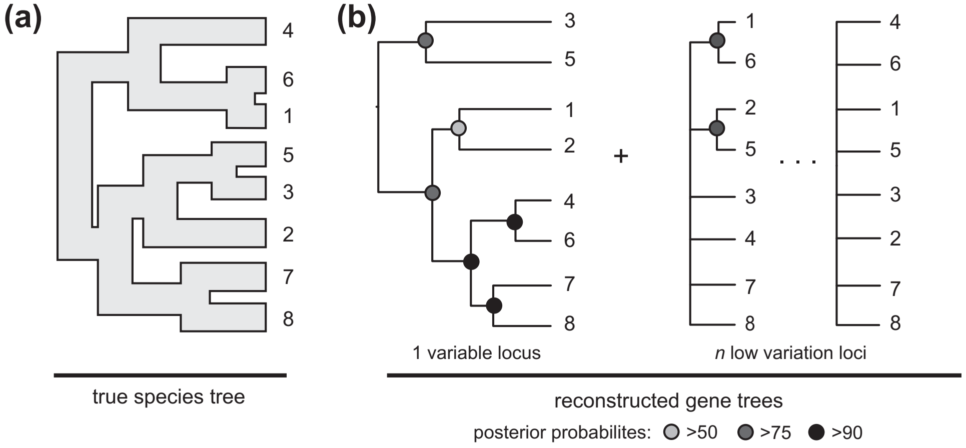

Fig. 1. A. Example topology for a true species tree used to simulate datasets that vary from one another due to coalescent and mutational stochasticity. B. These resulting datasets (shown as reconstructed gene trees) are then combined to estimate the species tree, pairing one variable locus with *n* low-variation loci.

for species-tree reconstruction with loci amplified in next-generation sequencing approaches (e.g., McCormack et al., 2012), where short loci contain only a few variable SNPs and the number of homologous fragments decreases with increasing phylogenetic distance (Althoff et al., 2007).

Using computer simulations, we examine how combining loci with differing levels of information content affects species-tree accuracy. Although the effects of phylogenetic information content have been investigated for traditional (i.e., concatenated) phylogenetic analyses (Moeller and Townsend, 2011; Townsend, 2007; Townsend et al., 2008) this is a new consideration in a species-tree context, where separate resolution of each locus may be paramount. In a traditional phylogenetic analysis, using loci with higher information content results in larger posterior probabilities on key nodes, with the exception of rapid radiations (Townsend, 2007). It is unknown whether these expectations hold for species-tree analyses when information content is mixed where the genealogies of individual loci are treated separately. We expect that the quality of loci (i.e., information content) utilized in species-tree analyses may have as large impact on accuracy as the quantity of loci, especially when comparing among methods that differ in how they account for gene-tree uncertainty (Huang et al., 2010; Knowles et al., 2012). Herein, we examine (1) what affect the addition of low variation loci has in a species-tree framework, (2) whether certain species-tree methodologies are better able to take advantage of the signal from low variation loci, and (3) how low-variation loci affect the support for incorrect relationships if the variable loci in an analysis are inaccurate (due to either coalescent or mutational variance).

2. Materials and methods

2.1. General-simulation approach

The simulation approach utilized consists of the following steps: (1) coalescent gene trees simulated along known species trees, (2) simulation of sequence data along each gene tree to reflect substitution processes corresponding to either variable or low information content loci, (3) estimation of species trees using a combination of variable and low variation loci either (a) directly from simulated nucleotide datasets in a Bayesian framework, or (b) by initially estimating Bayesian gene trees and secondarily calculating the maximum-likelihood species-tree estimate. Accuracy of all tree estimates was measured using the Robinson–Foulds symmetric distances (Robinson and Foulds, 1981), a metric that considers the number of nodes different between two trees without taking into account branch-length differences between estimates.

2.2. Identifying a broad range of species-tree topologies

Because analytical accuracy is intrinsically linked to species-tree topology, we explore the effects of low-variation loci in species trees analyses at two extremes—those especially difficult to estimate accurately and those particularly easy. To achieve this, species trees were selected from a set of 50 trees analyzed in previous simulations studies of species-tree accuracy (Knowles et al., 2012; McCormack et al., 2009). The five trees that showed the greatest or least initial accuracy (i.e., mean accuracy at the 1 individual and 1 locus sampling effort for 1 N total tree depths) were selected from the dataset (see Appendices). In addition, these trees were contrasted with a completely symmetrical and a completely asymmetrical tree, to provide analytical cases where species tree estimation is known to be respectively easy or challenging (Degnan and Salter, 2005; Leache and Rannala, 2010). Relative internode lengths for symmetrical and asymmetrical species-tree cases were fixed (i.e., all internodes within a true species tree were equal).

2.3. Simulation of data

Coalescent gene trees simulated along each species-tree in *ms* (Hudson, 2002) under conditions corresponding to very recent divergences where incomplete lineage sorting is likely to predominate (i.e., total tree depth of 1 N) and at deeper divergence levels where gene tree discord is generally not caused by the retention of ancestral polymorphism (10 N total tree depth) (Maddison and Knowles, 2006). Sequence data were simulated along each gene tree to correspond to situations where loci are known to be variable ($\theta = 0.01$), situations where loci exhibit little variation ($\theta = 0.001$), and situations where θ is drawn from a gamma distribution (i.e., mutation rates are low but highly variable among loci; $\theta \sim \Gamma(0.005, 2)$). All 1000 base-pair DNA sequences were evolved along the coalescent gene tree under an HKY85 model of nucleotide evolution with a gamma shape parameter ($\alpha = 0.8$) in the program *Seq-Gen* (Rambaut and Grassly, 1997). Information content in these loci differed based upon both total tree depth (1 N or 10 N) and the underlying population mutation parameter (Fig. 2), ranging from a mean of 4% divergence in variable loci at greater total tree depths down to no divergences in a small number of the 1 N low variation cases.