Journal of Theoretical Biology 364 (2015) 31–39

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

The influence of taxon sampling on Bayesian divergence time inference under scenarios of rate heterogeneity among lineages

André E.R. Soares[*,1], Carlos G. Schrago

Department of Genetics, Federal University of Rio de Janeiro, Rio de Janeiro 21941-590, Brazil

HIGHLIGHTS

- Taxonomic sampling affects Bayesian molecular dating analyses.
- Rates heterogeneity increases the effects of taxonomic sampling.
- Increased taxon sampling improves accuracy and precision of the results.

ARTICLE INFO

Article history:
Received 22 August 2013
Received in revised form
5 August 2014
Accepted 2 September 2014
Available online 11 September 2014

Keywords:
Relaxed molecular clock
Bayesian molecular dating
Molecular dating
Simulation
Molecular clock

ABSTRACT

Although taxon sampling is commonly considered an important issue in phylogenetic inference, it is rarely considered in the Bayesian estimation of divergence times. In fact, the studies conducted to date have presented ambiguous results, and the relevance of taxon sampling for molecular dating remains unclear. In this study, we developed a series of simulations that, after six hundred Bayesian molecular dating analyses, allowed us to evaluate the impact of taxon sampling on chronological estimates under three scenarios of among-lineage rate heterogeneity. The first scenario allowed us to examine the influence of the number of terminals on the age estimates based on a strict molecular clock. The second scenario imposed an extreme example of lineage specific rate variation, and the third scenario permitted extensive rate variation distributed along the branches. We also analyzed empirical data on selected mitochondrial genomes of mammals. Our results showed that in the strict molecular-clock scenario (Case I), taxon sampling had a minor impact on the accuracy of the time estimates, although the precision of the estimates was greater with an increased number of terminals. The effect was similar in the scenario (Case III) based on rate variation distributed among the branches. Only under intensive rate variation among lineages (Case II) taxon sampling did result in biased estimates. The results of an empirical analysis corroborated the simulation findings. We demonstrate that taxonomic sampling affected divergence time inference but that its impact was significant if the rates deviated from those derived for the strict molecular clock. Increased taxon sampling improved the precision and accuracy of the divergence time estimates, but the impact on precision is more relevant. On average, biased estimates were obtained only if lineage rate variation was pronounced.

1. Introduction

Taxon sampling has been considered one of the most important factors in phylogenetic inference (Pollock et al., 2002; Zwickl and Hillis, 2002). The inclusion of taxa is known to reduce problems such as long branch attraction, thereby increasing the overall phylogenetic accuracy of the inference procedure (Hedtke et al., 2006). Nevertheless, taxon sampling has rarely been considered an issue in divergence time estimation problems, where the tree topology is generally fixed or evaluated as a nuisance parameter and only the ages of the nodes are parameters of interest. However, a number of recent studies have investigated the impact of taxon sampling on divergence time inference. These studies have obtained conflicting results. For example, Linder et al. (2005) have shown that the taxonomic composition strongly influences the age estimates of the nodes for the African Restionaceae clade. Poux et al. (2008) also found that the number of terminals affected the chronological timescale obtained for Malagasy tenrecs. Hug and Roger (2007), in contrast, concluded that the effect of

---

[*] Correspondence to: A2-95, Departamento de Genética, Instituto de Biologia, CCS, Universidade Federal do Rio de Janeiro, Ilha do Fundão, Rio de Janeiro 21941-590, Brazil. Tel.: +55 21 2562 6397.

E-mail addresses: aesoares@ucsc.edu (A.E.R. Soares), carlos.schrago@gmail.com (C.G. Schrago).

[1] Present address: Mail Stop EE Biology, Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, United States of America. Tel.: +1 831 459-3009.

http://dx.doi.org/10.1016/j.jtbi.2014.09.004
0022-5193/© 2014 Elsevier Ltd. All rights reserved.

taxon sampling was less pronounced than the position of the calibration points and that taxon sampling played a minor role in molecular dating analysis. Recently, Xiang et al. (2011) also found that the taxonomic density played a minor role in establishing the evolutionary timescale of the Cornales.

Another relevant matter that has not been examined in detail is the relationship between taxon sampling and evolutionary rate variation. However, rate variation among branches is a major problem in phylogenetic inference and several models have been developed to address this problem (Delsuc et al., 2005; Lartillot et al., 2007; Pagel and Meade, 2008). To date, it is not clear whether the addition of taxa is able to circumvent the variation in time estimates if evolutionary rates depart from the strict molecular clock or if only calibration information improves divergence time estimation under heterogeneous rate scenarios.

Both taxon sampling and the evolutionary rate variation are critical to Bayesian divergence time estimates. Choosing which taxa to incorporate into an analysis is the very first step biologists usually take, and the number of available samples of a certain species or genus, or their phylogenetic relationships, usually limits its number. One of the difficulties in studying the effect of taxon sampling on species divergence times is the variety of parameters that may be considered. For instance, the amount, distribution and probability distributions of calibration information along internal nodes is certainly relevant (Heled and Drummond, 2012; Soares and Schrago, 2012; Warnock et al., 2012).

In addition, the degree to which evolutionary rates depart from the strict molecular clock might also influence time estimates in association with taxonomic sampling. Unfortunately, this problem is difficult to tackle analytically because of the complexity of the models used. The use of simulated data can alleviate this by allowing us to have total control over the "evolutionary history" of the samples and also to explore different possible evolutionary scenarios. This approach has been used before to tackle methodological questions, such as the selection of evolutionary models (Posada and Buckley, 2004), adaptive molecular evolution (Yang and Bielawski, 2000; Fay et al., 2001; Wong, 2004) or long branch attraction (Lyons-Weiler and Hoelzer, 1997). It has also been used to improve experimental design on evolutionary studies, like the use of multiple genes in divergence time estimates (Thorne and Kishino, 2002) or synchronicity of evolutionary events (Loss-Oliveira et al., 2012).

This study aims to further explore the robustness of the Bayesian dating methods based on the molecular clock hypothesis. The Bayesian approach allows for the insertion of uncertainty in the rate variation along the phylogenetic tree branches (Yang and Rannala, 2006) and also other sources of information, such as fossil data, or in the case of ancient specimens, radiocarbon dated samples (Shapiro et al., 2004). Modern methods allow for variation of the evolutionary rates among lineages (Thorne et al., 1998; Drummond et al., 2006) or the use of uncorrelated rates that follow a certain statistical distribution (Fay et al., 2001; Drummond et al., 2006). These methods, while widely used, were never tested under a series of simulations aimed to observe their behavior in different scenarios of rate heterogeneity, like an intensive rate variation on few lineages or when these variations are distributed among the tree branches.

In this sense, our main objective is to investigate how taxonomic sampling impacts divergence time estimates, particularly under episodes of rate heterogeneity among lineages. We have examined this issue by performing a series of concise simulation studies and by analyzing an empirical data set. We show that increased taxon sampling enhances the estimation of divergence times by augmenting the precision and accuracy of the estimates. We hypothesize that the effect of taxon sampling is associated with the reduction in the variance of evolutionary rates among lineages; i.e., the addition of terminals reduces the effects of branch rate variation on the estimates of the age of the nodes.

2. Methods

2.1. Simulations

To investigate the impact of taxon sampling on species divergence time inference, we have studied, via simulation, two tree topologies with 12 and 4 taxa, respectively (Fig. 1). The 4-taxon topology was constructed by selecting terminals *sp*1, *sp*6, *sp*11 and *sp*12 from the 12-taxon tree. In both topologies, two nodes are of interest: t_c and $\hat{t}$ (Fig. 1). These nodes represent the calibration node and the node in which the divergence time will be studied from a comparative perspective because it is present in both the 12-taxon and 4-taxon trees.

Three scenarios, termed Cases I, II and III, were adopted. These scenarios were designed to explore situations of evolutionary rate heterogeneity among branches. In all approaches, the bias and standard deviation of the divergence time estimates under different taxonomic compositions (taxon sampling) were measured against a condition of increased rate variation (departure from the strict clock). The simulation was implemented as described in the following paragraphs.

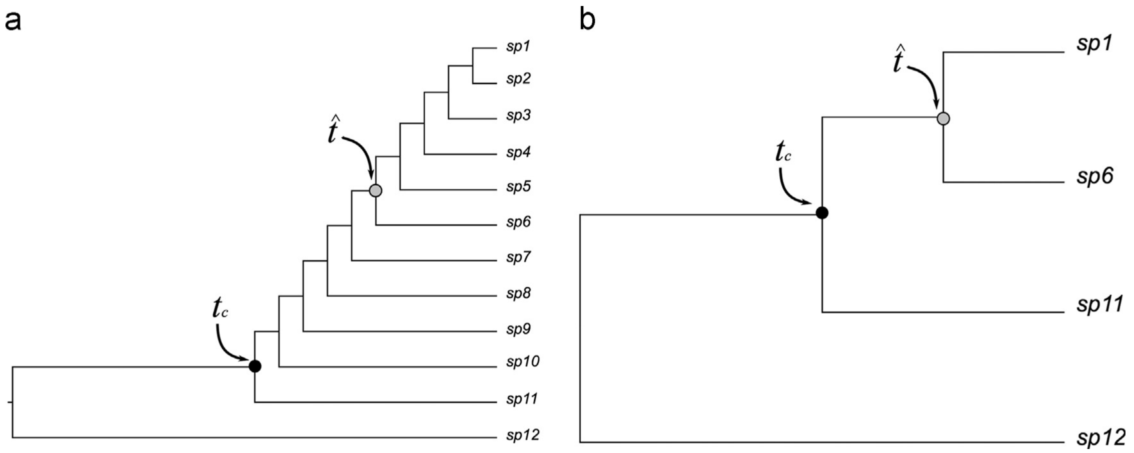

Fig. 1. Tree topologies used in this study. (a) Topology with 12 taxa, in which node t_c (black circle) received calibration information and the age of node $\hat{t}$ (grey circle) was investigated throughout this study. (b) A reduced 4-taxon tree topology was composed by selecting terminals *sp*1, *sp*6, *sp*11 and *sp*12 from the 12-taxon tree. Hence, nodes t_c and $\hat{t}$ were maintained in the reduced topology.