Accounting for Calibration Uncertainty: Bayesian Molecular Dating as a "Doubly Intractable" Problem

Stéphane Guindon

Syst. Biol. 67(4):651–661, 2018

DOI:10.1093/sysbio/syy003
Advance Access publication January 27, 2018

Accounting for Calibration Uncertainty: Bayesian Molecular Dating as a "Doubly Intractable" Problem

STÉPHANE GUINDON*

Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS, Université de Montpellier, Montpellier, France
**Correspondence to be sent to: Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, UMR 5506, CNRS, Université de Montpellier, Montpellier, France; E-mail: guindon@lirmm.fr.*

Received 6 February 2017; reviews returned 19 January 2018; accepted 23 January 2018
Associate Editor: David Bryant

Abstract.—This study introduces a new Bayesian technique for molecular dating that explicitly accommodates for uncertainty in the phylogenetic position of calibrated nodes derived from the analysis of fossil data. The proposed approach thus defines an adequate framework for incorporating expert knowledge and/or prior information about the way fossils were collected in the inference of node ages. Although it belongs to the class of "node-dating" approaches, this method shares interesting properties with "tip-dating" techniques. Yet, it alleviates some of the computational and modeling difficulties that hamper tip-dating approaches. The influence of fossil data on the probabilistic distribution of trees is the crux of the matter considered here. More specifically, among all the phylogenies that a tree model (e.g., the birth–death process) generates, only a fraction of them "agree" with the fossil data. Bayesian inference under the new model requires taking this fraction into account. However, evaluating this quantity is difficult in practice. A generic solution to this issue is presented here. The proposed approach relies on a recent statistical technique, the so-called exchange algorithm, dedicated to drawing samples from "doubly intractable" distributions. A small example illustrates the problem of interest and the impact of uncertainty in the placement of calibration constraints in the phylogeny given fossil data. An analysis of land plant sequences and multiple fossils further highlights the pertinence of the proposed approach. [Bayesian inference, land plants, MCMC, molecular dating.]

INTRODUCTION

Inferring times of divergence between species from the analysis of genetic and fossil data has led to spectacular advances in our understanding of evolution. One of the most striking illustrations is given by the work of Sarich and Wilson (1967) which led to a reappraisal of the timing of divergence between African apes and humans. Still, "molecular estimates" are generally older than that suggested by the fossil record (Benton and Ayala 2003). Statistical modeling of the age of the most recent common ancestor (MRCA) of extant primates based on the fossil record points to estimates close to those obtained from the analysis of molecular data nonetheless (Tavaré et al. 2002). Still, discrepancies between molecular and paleontological dates are frequent and are generally attributed to deficiencies in the models used to infer divergence times from molecular data (Yang 2006; Puttick et al. 2016).

Overly simplistic models of substitution rate variation during the course of evolution are a cause of concern amongst others. In fact, Bromham et al. (2000) provide a clear example whereby enforcing a strict molecular clock leads to inaccurate estimates of divergence times between rodents and primates. Sanderson (1997) was the first to propose a statistical framework and a corresponding inference technique (Sanderson 2002) to accommodate for the variation of substitution rates across lineages. Thorne et al. (1998) devised a similar yet more explicit statistical model of a "relaxed clock" and based the inference on the posterior distribution of model parameters.

Thorne et al. (1998) applied Markov Chain Monte Carlo (MCMC) techniques to Bayesian inference of hierarchical model parameters in phylogenetics. The Bayesian approach enjoyed a considerable popularity in the decades that followed (see dos Reis et al. 2016 for a recent review). Part of this success comes from the ease with which new models can be integrated without affecting the inference techniques (see for instance the "plug-in" architecture implemented in BEAST2, Bouckaert et al. 2014).

The first technique that combined fossil and genetic data in Bayesian molecular dating relied on the so-called "node-dating" approach. The age of a given fossil's stratigraphic range along with its morphological features are examined in a first step. This analysis is usually conducted by experts, following a well-established protocol (Benton and Donoghue 2007; Parham et al. 2011). As a result, subsets of extant (and sampled) species are identified that share a MRCA with one or more fossils. If a particular subset of taxa is deemed to share its MRCA with more than one fossil, then only the oldest fossil conveys information about the minimum, that is younger age of the corresponding clade. Maximum clade ages are frequently determined from the age of older fossil deposits that do not contain the fossil of the clade under scrutiny (Benton and Donoghue 2007). Despite these clear guidelines, spectacular failures in properly accommodating for calibration constraints are still not uncommon (see dos Reis et al. 2014).

There has been a substantial number of developments around node-dating techniques and software implementing these in the last decade or so

(dos Reis et al. 2016; Kumar and Hedges 2016). Yet, current approaches have serious limitations. One of these originates in the mathematics underlying the tree model, that is the distribution of topology and node ages, given calibration data. Although it is commonplace to define a marginal distribution for each calibrated node, Rannala's (2016) theoretical investigations indicate that it is generally not possible to specify a tree model that "agrees" with these distributions. In other words, when ignoring genetic data, the marginal distributions of calibrated node ages as defined by the user differ from that derived from the joint distribution of ages given calibration data returned by the MCMC analysis. A corollary is that the models implemented in several popular statistical software are in fact distinct from those intended (Warnock et al. 2015). Beside these mathematical difficulties, node-dating techniques lack flexibility in accommodating for uncertainty associated to fossil data. In particular, it is not always straightforward to delineate the subset of species a given fossil shares its MRCA with (see e.g., Sauquet et al. 2011; Saladin et al. 2017). Yet, in practice, only one subset of taxa is associated to a given calibration constraint, thereby overlooking the uncertainty inherent to the process that translates fossil data into calibration constraints.

"Tip-dating" approaches avoid some of the difficulties that hamper node-dating methods by treating fossils as *bona fide* taxa. The fossilized birth–death (FBD) process (Stadler 2010; Didier et al. 2012; Heath et al. 2014), for instance, belongs to this class of methods as it considers the phylogenetic position and age of fossils as latent variables (but see Arcila et al. 2015 for a different viewpoint). The FBD model accommodates for birth and death of lineages along with fossilization events in a unified mathematical framework. The direct ancestor of each fossil, that is the internal node that connects the external edge leading to a fossil tip to the rest of the tree, has to be older than the fossil itself. This constraint imposes a "hard" younger bound on the age of this internal node. The corresponding older bound is determined by the time elapsed between the speciation event that gave rise to the fossil's direct ancestor and the fossilization event itself. The distribution of this time is determined by the parameters inferred under the FBD process. The corresponding older bound is thus less constrained than the younger bound and can be as old as the (unknown) age of the crown node of the clade that the fossil calibrates. As a likely consequence, node age estimates obtained under the FBD model are generally older than that derived from node-dating techniques (Grimm et al. 2014; Arcila et al. 2015; Zhang et al. 2016; Saladin et al. 2017).

The "total-evidence" approach also belongs to the class of tip-dating methods. It was proposed in an attempt to analyze genetic sequences along with morphological characters in a unified statistical framework (Pyron 2011; Ronquist et al. 2012). The inferred position of fossil taxa in the phylogeny is here determined by the similarities of morphological features displayed by extant and extinct (i.e., fossil) taxa. While node-dating relies on expert knowledge to translate fossil data into calibration information, total-evidence instead tackles this issue using a statistical modeling approach. Uncertainty around the placement of calibration constraints is thus dealt with in a proper probabilistic framework, thereby giving this tip-dating technique a clear edge on node-dating approaches. The implementation of the total-evidence technique proposed by Pyron (2011) and Ronquist et al. (2012) involved tree generating processes that did not accommodate for the specificities of fossil taxa. Gavryushkina et al. (2017) and Zhang et al. (2016) recently combined the total-evidence approach with the FBD model in order to draw inference from both morphological and genetic data under a probabilistic model that is well suited to the hybrid type of data at hand.

Current tip-dating techniques have serious limitations however. First, treating fossils as actual taxa requires exploring the space of their potential positions in the phylogeny, making this approach more computationally intensive compared to node-dating. But most importantly, modeling morphological character evolution, as is done in the total evidence approach, is challenging. Indeed, as opposed to DNA or protein sequences where the state space is well defined (i.e., the four nucleotides and the 20 amino-acids for DNA and protein sequences, respectively), each morphological feature has its own state space. Correlation between characters and ascertainment biases (only parsimony-informative characters are usually collected) are also a source of concern (dos Reis et al. 2016). Finally, current tip-dating techniques ignore relevant information about fossil deposits that ought to contain fossils of the clade of interest, but do not. As already mentioned, the lack of a given fossil in a particular fossiliferous horizon is a clear indicator of the maximum age of a clade (Benton and Donoghue 2007). Yet, current tree models, including the FBD, do not accommodate for this type of evidence while node-dating techniques rely on it, at least in theory.

This study introduces a new Bayesian dating technique that alleviates some of the issues currently hindering both node- and tip-dating approaches. The new method belongs to the family of node-dating techniques. Yet, it brings flexibility in handling uncertainty around fossil data comparable to that achieved by tip-dating approaches. More specifically, a given calibration constraint can apply to various clades with different probabilities. The proposed model thereby provides a relevant framework to deal with ambiguities related to the interpretation of fossil data. I show that the joint distribution of node ages under the new model belongs to the class of "doubly intractable" distributions (Murray et al. 2012). Efficient computational solutions exist to tackle this class of problems. I present one of them in the context of molecular dating. An illustration of the proposed

TABLE 1. Definitions of the main symbols used in this article

Symbol	Definition
n	Number of taxa
d_n	Aligned genetic sequences
$\boldsymbol{\alpha}$	Vector of calibration constraints probabilities
τ	Ranked tree topology
$\mathbf{t}$	Vector of internal node ages
$\boldsymbol{\theta}$	Vector of tree model parameters
$\mathbf{e}$	Vector of calibrated subsets of taxa
$\mathbf{i}$	Vector of calibration time intervals

technique on the timing of speciation events in land plants is also provided.

THEORY

A labeled history or ranked tree is defined as a labeled tree with temporally ordered internal nodes (Edwards 1970). Let n be the number of taxa and $t_1 \geq t_2 \geq \cdots t_{n-1} \geq 0$ denote the ages of internal nodes from the oldest to the youngest. $\mathbf{t}=(t_1,\ldots,t_{n-1})$ is the vector of these ranked times. The parameter $\boldsymbol{\theta}$ denotes one or more numerical parameters involved in the definition of the processes generating the tree and calibration data (e.g., $\boldsymbol{\theta}:=\{\lambda,\mu\}$, where λ and μ are the birth and death rates in the birth–death model with full sampling). τ is the ranked tree topology, that is the ranked tree with information about the age of internal nodes removed. Let d_n denote a set of n homologous sequences collected for the inference of τ and $\mathbf{t}$.

Fossil data convey information about the age of the MRCA of subsets of extant species that define a clade. In what follows, these clades may or may not be monophyletic. Also, in case the tree topology is not fixed throughout the inference, the internal node corresponding to the MRCA of interest may move about. The age of the oldest fossil sharing one or multiple apomorphies with this subset of taxa defines the minimum age of the MRCA in question. The maximum age of the same MRCA is often derived from the age of the youngest stratigraphic range that does not contain any fossil of the clade of interest (Benton and Donoghue 2007).

Let $I^{(k)}$ be the random interval corresponding to the age range defining the kth calibration constraint. $i^{(k)}$ denote a particular time interval, that is a value taken by the random variable $I^{(k)}$. $i_-^{(k)}$, and $i_+^{(k)}$ give the younger and corresponding older bounds for that interval. Note that $I^{(k)}$ characterizes the age of the MRCA of a subset of sampled species, which is distinct from the age of the fossil itself. $E^{(k)}$ is the random subset of species associated to the kth calibration constraint. $e^{(k)}$ corresponds to one such subset, that is a realization of the random variable $E^{(k)}$. In the following, $\mathbf{i}=(i^{(1)},\ldots,i^{(m)})$ and $\mathbf{e}=(e^{(1)},\ldots,e^{(m)})$ will denote the vectors of time intervals and subsets of species corresponding to m calibration constraints. For the sake of brevity, I will omit the superscript (k) when referring to a given subset of species and a time interval in some cases.

The random variables $I^{(k)}$ and $E^{(k)}$ belong to the set of parameters of the model rather than data. In the context of interest, fossil data arise as a random variable, noted as $\mathbf{F}=\{F^{(1)},\ldots,F^{(m)}\}$, whereby $F^{(k)}$ is the set of fossils that are used to define the kth calibration constraint. Typically, the value taken by $F^{(k)}$ will correspond to a series of quantitative or qualitative measurements made from the analysis of morphological characters found in one or more fossils and extant taxa. Values taken by these random variables define the probability of the different outcomes of $I^{(k)}$ and $E^{(k)}$. In the following, I will use $\Pr(e,i|f,\alpha)$ instead of $\Pr(E^{(k)}=e,I^{(k)}=i|F^{(k)}=f,\alpha)$ for the sake of brevity. One may then have $\Pr(e,i|f,\alpha)=\alpha$, $\Pr(e',i'|f,\alpha)=1-\alpha$, and $\Pr(e',i|f,\alpha)=\Pr(e,i'|f,\alpha)=0$, with $0<\alpha<1$. In other words, fossil data suggest here that a calibration constraint (the kth) applies to the subset of taxa e (with a MRCA that lived in the time interval i) with probability α, or to the subset e' (with time interval i') with probability $1-\alpha$. In what follows, $\boldsymbol{\alpha}$ will denote the vector of values of α as introduced above. The length of this vector corresponds to the number of calibration constraints, that is m. Since $\boldsymbol{\alpha}$ conveys all the information from fossil data about calibration constraints as determined by expert knowledge, it will replace $\mathbf{F}$ in the following. Table 1 provides a summary of the main mathematical symbols used in this article and their definitions.

The posterior distribution of model parameters ($\mathbf{t}$, τ, $\mathbf{e}$, $\mathbf{i}$, and $\boldsymbol{\theta}$) given genetic sequences (d_n) and fossil data ($\boldsymbol{\alpha}$) can be written as follows:

$$p(\mathbf{t},\tau,\mathbf{e},\mathbf{i},\boldsymbol{\theta}|d_n,\boldsymbol{\alpha}) \propto \Pr(d_n|\mathbf{t},\tau)p(\mathbf{t},\tau,\mathbf{e},\mathbf{i},\boldsymbol{\theta}|\boldsymbol{\alpha}).$$

The term $\Pr(d_n|\mathbf{t},\tau)$ corresponds to the likelihood of the phylogenetic model given the alignment of genetic sequences (the average rate at which substitutions take place along the tree is omitted here for the sake of clarity of notations). It is evaluated using Felsenstein's pruning algorithm (Felsenstein 1981). The term $p(\mathbf{t},\tau,\mathbf{e},\mathbf{i},\boldsymbol{\theta}|\boldsymbol{\alpha})$ is the focus of this study. There are multiple ways to define this posterior density, corresponding to various modeling assumptions. I propose the following definition:

$$p(\mathbf{t},\tau,\mathbf{e},\mathbf{i},\boldsymbol{\theta}|\boldsymbol{\alpha}) := p_{\mathrm{T}}(\mathbf{t},\tau|\boldsymbol{\theta},\mathbf{e},\mathbf{i})g(\mathbf{e},\mathbf{i}|\boldsymbol{\theta},\boldsymbol{\alpha})f(\boldsymbol{\theta}). \quad (1)$$

$f(\boldsymbol{\theta})$ is the prior density for the tree and calibration model parameters. Exponential distributions for the birth and death parameters of the tree model are fairly standard. The definition of the probability density $p_{\mathrm{T}}(\mathbf{t},\tau|\boldsymbol{\theta},\mathbf{e},\mathbf{i})$ depends on the tree process and calibration data. More specifically, this term can be understood as a truncated density whereby the truncation comes from calibration constraints and applies to the tree process. Its precise

definition is the crux of the present study and is discussed in detail below. Note that the definition above implies that $p(\boldsymbol{\alpha}|\boldsymbol{\theta}) := p(\boldsymbol{\alpha})$, that is the likelihood of a given value of $\boldsymbol{\theta}$ does not depend on $\boldsymbol{\alpha}$. In words, the position of calibration constraints in the tree and the associated uncertainty (represented by $\boldsymbol{\alpha}$) is not influenced by the tree process parameters (represented by $\boldsymbol{\theta}$). This assumption is built into the model and amounts to considering that all the signal in the data about the tree process parameters is conveyed by the genetic sequences through the inferred phylogenetic tree, which does not seem unreasonable.

The density $g(\mathbf{e}, \mathbf{i}|\boldsymbol{\alpha}, \boldsymbol{\theta})$ is defined as a product over all calibrations. This term is the part of the model that incorporates expert knowledge about the uncertainty of calibration constraints given fossil data. Unlike for the "product-of-marginals" approach, the assumption of independence of calibration constraints is sensible here as the phylogeny is not involved on that level of the hierarchical model:

$$g(\mathbf{e}, \mathbf{i}|\boldsymbol{\alpha}, \boldsymbol{\theta}) = \prod_{k=1}^{m} \Pr(E^{(k)} = e^{(k)}, I_+^{(k)} = i_+^{(k)}, I_-^{(k)} = i_-^{(k)}|\alpha^{(k)}), \quad (2)$$

where $\alpha^{(k)}$ is the kth element of $\boldsymbol{\alpha}$ while $I_+^{(k)}$ and $I_-^{(k)}$ correspond to the younger and older age bounds defined by $I^{(k)}$. I will often write $g(\mathbf{e}, \mathbf{i}|\boldsymbol{\alpha})$ instead of $g(\mathbf{e}, \mathbf{i}|\boldsymbol{\alpha}, \boldsymbol{\theta})$ as the density defined above is not a function of $\boldsymbol{\theta}$ by definition. As mentioned before, fossils typically define precise younger bounds for clade ages while older bounds are less well known (Benton and Donoghue 2007). This asymmetry can be dealt with by defining the conditional distribution of the older clade age $I_+^{(k)}|I_-^{(k)}$ and the marginal distribution of the younger age $I_-^{(k)}$ accordingly. The corresponding densities then serve as a basis for the definition of $\Pr(E^{(k)} = e, I_+^{(k)} = i_+, I_-^{(k)} = i_-|\alpha)$ (the subscripts (k) on e, i_+, i_-, and α are omitted in this paragraph in a slight abuse of notation). In the present study, I will mostly focus on the simple case where $\Pr(E^{(k)} = e, I_+^{(k)} = i_+, I_-^{(k)} = i_-|\alpha) = 1$, that is there is no uncertainty in the subset of species and the corresponding time interval calibrated by the kth fossil datum. The case where $\Pr(E^{(k)} = e, I_+^{(k)} = i_+, I_-^{(k)} = i_-|\alpha) = \alpha$ and $\Pr(E^{(k)} = e', I_+^{(k)} = i'_+, I_-^{(k)} = i'_-|\alpha) = 1-\alpha$ with $0 < \alpha < 1$ will also be examined in the setting of a three-taxon data set. In this situation, the calibration constraint defined by the kth fossil datum calibrates the subset of species e in the time interval i with probability α and the subset e' with time interval i' with probability $1-\alpha$.

The modeling approach proposed above provides a simple mean to incorporate expert knowledge derived from the analysis of fossils into the molecular dating experiment. More sophisticated mechanistic models such as that described in Tavaré et al. (2002) and Wilkinson et al. (2010) could be used here instead. In fact, accounting for relevant assumptions about the fossilization process and the sampling intensities in different stratigraphic layers is a useful and potentially important feature of the technique proposed in this study.

As stated before, the main focus is on the probabilistic distribution of the phylogeny given calibration constraints. The density of interest is given below:

$$p_{\mathrm{T}}(\mathbf{t}, \tau|\boldsymbol{\theta}, \mathbf{e}, \mathbf{i}) = \begin{cases} p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta})\Pr_{\mathrm{T}}(\tau)/Z_{\boldsymbol{\theta}} & \text{if } \chi(\mathbf{t}, \tau, \mathbf{e}, \mathbf{i}) = 1 \\ 0 & \text{if } \chi(\mathbf{t}, \tau, \mathbf{e}, \mathbf{i}) = 0, \end{cases} \quad (3)$$

or simply $p_{\mathrm{T}}(\mathbf{t}, \tau|\boldsymbol{\theta}, \mathbf{e}, \mathbf{i}) = p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta})\Pr_{\mathrm{T}}(\tau)\chi(\mathbf{t}, \tau, \mathbf{e}, \mathbf{i})/Z_{\boldsymbol{\theta}}$, where $\chi(\mathbf{t}, \tau, \mathbf{e}, \mathbf{i}) = 1$ whenever all calibration constraints are "satisfied" and 0 otherwise. The kth calibration constraint is said to be satisfied when the internal node corresponding to the MRCA of the set of taxa making up $e^{(k)}$ has an age that falls within the time interval defined by $i^{(k)}$. If multiple calibration constraints are associated to a single node, a conservative criterion applies. The older bound for the age of that node is then set to the youngest of the older bounds of all calibration intervals pointing to this node. In a symmetric fashion, the younger bound is set to the oldest of the younger bounds of all corresponding calibration intervals. Also, the distribution on ranked tree topologies with n tips is uniform under classical tree models (Stadler 2008). The equality $\Pr_{\mathrm{T}}(\tau) = 2^{n-1}/(n!(n-1)!)$ therefore holds in the present study, and this probability can be safely ignored throughout the inference.

$Z_{\boldsymbol{\theta}}$ is the normalization factor for the density of interest. We have:

$$Z_{\boldsymbol{\theta}} = \sum_{\psi} \int p_{\mathrm{T}}(\mathbf{u}|\boldsymbol{\theta})\Pr_{\mathrm{T}}(\psi)\chi(\mathbf{u}, \psi, \mathbf{e}, \mathbf{i})\mathrm{d}\mathbf{u}, \quad (4)$$

where the sum is over all ranked trees ψ, the integral is over all vectors of internal node ages $\mathbf{u}$ with no reference to calibration constraints, and $\chi(\mathbf{u}, \psi, \mathbf{e}, \mathbf{i})$ is the probability defined above. The value of $Z_{\boldsymbol{\theta}}$ is thus a function of $\boldsymbol{\theta}$. Therefore, it cannot be considered as a constant that would cancel out in the Metropolis ratio of a MCMC operator updating $\boldsymbol{\theta}$.

Altogether, the probability density of interest is thus:

$$p(\mathbf{t}, \tau, \mathbf{e}, \mathbf{i}, \boldsymbol{\theta}|\boldsymbol{\alpha}) := \frac{p_{\mathrm{T}}(\mathbf{t}, \tau|\boldsymbol{\theta})\chi(\mathbf{t}, \tau, \mathbf{e}, \mathbf{i})}{Z_{\boldsymbol{\theta}}} g(\mathbf{e}, \mathbf{i}|\boldsymbol{\alpha}, \boldsymbol{\theta}) f(\boldsymbol{\theta}).$$

If one uses $g(\mathbf{e}, \mathbf{i}|\boldsymbol{\alpha}, \boldsymbol{\theta}) := Z_{\boldsymbol{\theta}}$ instead of the definition given in Equation 2, the above density simplifies to give the following:

$$p(\mathbf{t}, \tau, \mathbf{e}, \mathbf{i}, \boldsymbol{\theta}|\boldsymbol{\alpha}) := p_{\mathrm{T}}(\mathbf{t}, \tau|\boldsymbol{\theta})\chi(\mathbf{t}, \tau, \mathbf{e}, \mathbf{i}) f(\boldsymbol{\theta}). \quad (5)$$

This expression is in fact similar to that used by Gavryushkina et al. (2017) in the context of a total-evidence approach combined to the FBD tree process. It also corresponds to using uniform priors on calibrated node ages in the "product-of-marginals" technique implemented in the BEAST software. This particular approach thus relies implicitly on substituting $g(\mathbf{e}, \mathbf{i}|\boldsymbol{\alpha}, \boldsymbol{\theta})$ with $Z_{\boldsymbol{\theta}}$, which indeed corresponds to the probability of the calibration constraint given the fossil data under

the tree process considered. This choice makes perfect sense from a mathematical and a modeling perspective. Yet, as shown above, using the tree process to define the distribution of $\mathbf{e},\mathbf{i}|\boldsymbol{\alpha},\boldsymbol{\theta}$ is not a requirement. In fact, choosing one model instead of the other should be governed by the way fossil and genetics data are sampled (see Discussion section). Notwithstanding the design of the experiment, practical considerations are also important. Defining $g(\mathbf{e},\mathbf{i}|\boldsymbol{\alpha},\boldsymbol{\theta})$ in a manner that reflects uncertainty in the translation of fossil data into calibration constraints is straightforward. The ease with which expert knowledge can be incorporated in the model therefore makes this approach appealing from a practical point of view. In the following, the modeling approaches defined by Equations 1 and 5 will be referred to as the "tree-independent" and "tree-dependent" calibration densities respectively.

A "Doubly Intractable" Problem

If the tree topology is to be estimated, the calculation of $Z_{\boldsymbol{\theta}}$ requires summing over all ranked tree topologies and, for each of these, integrating over node heights (see Equation 4). It might not be feasible to enumerate such topologies for a large number of constraints (see Gavryushkina et al. 2014) The present study thus uses a new route to tackle this problem. The proposed approach relies on efficient numerical techniques that are relevant to Bayesian inference using MCMC. Below is a description of one of these techniques, namely the "exchange algorithm."

The posterior distribution of model parameters ($\mathbf{t}$, τ, $\boldsymbol{\theta}$, $\mathbf{e}$ and $\mathbf{i}$) given genetic sequences (d_n) and fossil data ($\boldsymbol{\alpha}$) is given below:

$$p(\tau,\mathbf{t},\mathbf{e},\mathbf{i},\boldsymbol{\theta}|d_n,\boldsymbol{\alpha}) = \frac{\Pr(d_n|t,\tau)p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta})\Pr_{\mathrm{T}}(\tau)\chi(\mathbf{t},\tau,\mathbf{e},\mathbf{i})}{Z_{\boldsymbol{\theta}}\Pr(d_n)}\frac{g(\mathbf{e},\mathbf{i}|\boldsymbol{\alpha})f(\boldsymbol{\theta})}{},$$

which is rewritten as follows:

$$p(\tau,\mathbf{t},\mathbf{e},\mathbf{i},\boldsymbol{\theta}|d_n,\boldsymbol{\alpha}) = \frac{\Pr(d_n|\mathbf{t},\tau)}{\Pr(d_n)}h(\mathbf{e},\mathbf{i},\boldsymbol{\alpha},\boldsymbol{\theta})\frac{f_{\mathrm{T}}(\mathbf{t},\boldsymbol{\theta},\tau,\mathbf{e},\mathbf{i})}{Z_{\boldsymbol{\theta}}},$$

whereby $f_{\mathrm{T}}(\mathbf{t},\boldsymbol{\theta},\tau,\mathbf{e},\mathbf{i}) := p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta})\Pr_{\mathrm{T}}(\tau)\chi(\mathbf{t},\tau,\mathbf{e},\mathbf{i})$ and $h(\mathbf{e},\mathbf{i},\boldsymbol{\alpha},\boldsymbol{\theta}) := g(\mathbf{e},\mathbf{i}|\boldsymbol{\alpha})f(\boldsymbol{\theta})$.

I assume that neither $Z_{\boldsymbol{\theta}}$ nor $\Pr(d_n)$ can be computed. For that reason, the posterior density of interest can be considered as a *doubly intractable* distribution (Murray et al. 2012). Updating the value of $\boldsymbol{\theta}$ using a traditional Metropolis–Hastings (MH) algorithm is not feasible as the calculation of the MH acceptance ratio α requires knowing the values of both $Z_{\boldsymbol{\theta}}$ and $Z_{\boldsymbol{\theta}'}$ in the expression below:

$$\alpha = \min\left\{1, \frac{h(\mathbf{e},\mathbf{i},\boldsymbol{\alpha},\boldsymbol{\theta}')}{h(\mathbf{e},\mathbf{i},\boldsymbol{\alpha},\boldsymbol{\theta})}\cdot\frac{p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta}')}{p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta})}\cdot\frac{Z_{\boldsymbol{\theta}}}{Z_{\boldsymbol{\theta}'}}\cdot\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}, \quad (6)$$

where $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are the current and proposed values of the parameter respectively and $q(\cdot|\cdot)$ is the proposal density. One way to circumvent this issue is to introduce an auxiliary variable, $\zeta = \{\mathbf{u},\psi\}$, which is a composite parameter made of $\mathbf{u}$, a vector of non-negative real numbers that has length $n-1$, that is the same as that of $\mathbf{t}$, corresponding to the number of internal nodes in the tree, and ψ, the corresponding ranked tree topology.

The proposed algorithm then relies on the following joint posterior probability density:

$$p(\tau,\mathbf{t},\mathbf{e},\mathbf{i},\boldsymbol{\theta},\psi,\mathbf{u},\boldsymbol{\theta}'|d_n,\boldsymbol{\alpha}) := \frac{\Pr(d_n|\tau,\mathbf{t})}{\Pr(d_n)}h(\mathbf{e},\mathbf{i},\boldsymbol{\alpha},\boldsymbol{\theta})\frac{f_{\mathrm{T}}(\mathbf{t},\boldsymbol{\theta},\tau,\mathbf{e},\mathbf{i})}{Z_{\boldsymbol{\theta}}}q(\boldsymbol{\theta}'|\boldsymbol{\theta})\frac{f_{\mathrm{T}}(\mathbf{u},\boldsymbol{\theta}',\psi,\mathbf{e},\mathbf{i})}{Z_{\boldsymbol{\theta}'}}, \quad (7)$$

which, when marginalizing over ψ, $\mathbf{u}$ and $\boldsymbol{\theta}'$ gives the posterior density of interest. Consider that the current instance of the (augmented) model is $\{\tau,\mathbf{t},\boldsymbol{\theta},\psi,\mathbf{u},\boldsymbol{\theta}'\}$, where $\mathbf{u}$ and ψ were obtained by sampling from $f_{\mathrm{T}}(\cdot,\boldsymbol{\theta}',\cdot,\mathbf{e},\mathbf{i})/Z_{\boldsymbol{\theta}'}$ and $\boldsymbol{\theta}'$ was sampled from $q(\cdot|\boldsymbol{\theta})$, in accordance with the density above. A new instance of the model is then proposed by swapping $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. The proposed state is thus $\{\tau,\mathbf{t},\boldsymbol{\theta}',\psi,\mathbf{u},\boldsymbol{\theta}\}$ and the Hastings ratio for that move is equal to one because the exchange $\boldsymbol{\theta}\leftrightarrow\boldsymbol{\theta}'$ is deterministic. The acceptance ratio is therefore given by the ratio of the relevant posterior densities:

$$\alpha = \min\left\{1, \frac{p(\tau,\mathbf{t},\mathbf{e},\mathbf{i},\boldsymbol{\theta}',\psi,\mathbf{u},\boldsymbol{\theta}|d_n,\boldsymbol{\alpha})}{p(\tau,\mathbf{t},\mathbf{e},\mathbf{i},\boldsymbol{\theta},\psi,\mathbf{u},\boldsymbol{\theta}'|d_n,\boldsymbol{\alpha})}\right\} \quad (8)$$

$$= \min\left\{1, \frac{h(\mathbf{e},\mathbf{i},\boldsymbol{\alpha},\boldsymbol{\theta}')}{h(\mathbf{e},\mathbf{i},\boldsymbol{\alpha},\boldsymbol{\theta})}\cdot\frac{p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta}')}{p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta})}\cdot\frac{p_{\mathrm{T}}(\mathbf{u}|\boldsymbol{\theta})}{p_{\mathrm{T}}(\mathbf{u}|\boldsymbol{\theta}')}\cdot\frac{q(\boldsymbol{\theta}|\boldsymbol{\theta}')}{q(\boldsymbol{\theta}'|\boldsymbol{\theta})}\right\}, \quad (9)$$

which does not require evaluating Z_{θ} nor $Z_{\theta'}$. This approach corresponds to the "exchange algorithm" first described in Murray et al. (2012).

The distribution of the auxiliary variables as defined by the joint posterior density (Equation 7) is determined by $f_{\mathrm{T}}(\mathbf{u},\boldsymbol{\theta}',\psi,\mathbf{e},\mathbf{i})/Z_{\boldsymbol{\theta}'}$. In other words, random draws from the target distribution for $\mathbf{u}$ and ψ could be obtained through exact simulation under the tree model with calibration constraints. I was unable to design a suitable technique for that step unfortunately. It is nonetheless possible to obtain valid draws from the relevant distribution using traditional MH. Indeed, for a given value of $\boldsymbol{\theta}$, the term $Z_{\boldsymbol{\theta}}$ cancels out in the Metropolis ratio and the acceptance ratio for updating the value of $\mathbf{u}$ and ψ in a MH step is as follows:

$$\beta = \min\left\{1, \frac{p_{\mathrm{T}}(\mathbf{u}^*|\boldsymbol{\theta})}{p_{\mathrm{T}}(\mathbf{u}|\boldsymbol{\theta})}\cdot\frac{q(\mathbf{u},\psi|\mathbf{u}^*,\psi^*)}{q(\mathbf{u}^*,\psi^*|\mathbf{u},\psi)}\right\},$$

where $\mathbf{u}$ and ψ refer to the state currently occupied by the chain built in this MCMC-within-MCMC step of the analysis, while $\mathbf{u}^*$ and ψ^* are the proposed states. New values of (auxiliary) node ages and ranked tree topologies are proposed using standard operators in statistical phylogenetics. Therefore, updating values of ζ does not present any particular difficulty. In practice, 100 MH steps were taken in order to obtain what was considered as a valid draw from the target distribution.

RESULTS

An Example with Three Taxa

It is possible to derive an analytical expression for the posterior density of model parameters in the special case where only three taxa are analyzed and sequences of infinite length are considered. I will use this simple setting to illustrate the technique proposed in this study. In particular, differences between the posterior distributions of tree model parameters derived under the tree-dependent (Equation 5) and the new tree-independent (Equation 1) modeling strategies will be examined. The impact of uncertainty around calibration constraints will be illustrated afterwards.

Let a, b, and c denote the three taxa. Also, e and e' are the subsets of species $\{a,b\}$ and $\{a,b,c\}$, respectively. i and i' are two nonoverlapping time intervals $[u,v]$ and $[x,y]$ such that $x>v$ (i.e., the time point x is older than v). The MRCAs of $\{a,b\}$ and $\{a,b,c\}$ lived in time intervals $[u,v]$ and $[x,y]$ respectively, without ambiguity, that is $g(\mathbf{e}=(e,e'),\mathbf{i}=(i,i')|\boldsymbol{\alpha})=1$.

Because the sequences are of infinite length and a strict molecular clock with known substitution rate applies, we have $\Pr(d_n|\mathbf{t},\tau)=\Xi(\mathbf{t},\mathbf{t}^*,\tau,\tau^*)$, where $\mathbf{t}^*$ and τ^* are the maximum likelihood estimates of node ages and ranked tree topology. $\Xi(\mathbf{t},\mathbf{t}^*,\tau,\tau^*)$ is equal to one for $\mathbf{t}=\mathbf{t}^*$ and $\tau=\tau^*$, and zero otherwise. The posterior density of interest takes the following expression:

$$p(\tau,\mathbf{t},\mathbf{e},\mathbf{i},\boldsymbol{\theta}|d_n,\boldsymbol{\alpha})=\begin{cases} p_{\mathrm{T}}(\tau,\boldsymbol{t}|\mathbf{e},\mathbf{i},\boldsymbol{\theta})g(\mathbf{e},\mathbf{i}|\boldsymbol{\alpha})f(\boldsymbol{\theta})/K \\ \qquad\qquad \text{if } \mathbf{t}=\mathbf{t}^* \text{ and } \tau=\tau^*, \\ 0 \qquad\qquad \text{otherwise.} \end{cases}$$

K is a normalization factor (distinct from $Z_{\boldsymbol{\theta}}$) that ensures that $p(\tau,\mathbf{t},\mathbf{e},\mathbf{i},\boldsymbol{\theta}|d_n,\boldsymbol{\alpha})$ as defined above is proper. Its expression is given below:

$$K=\int_0^\infty p_{\mathrm{T}}(\tau^*,\mathbf{t}^*|\boldsymbol{\theta},\mathbf{e},\mathbf{i})g(\mathbf{e},\mathbf{i}|\boldsymbol{\alpha})f(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}. \tag{10}$$

The expression for $p_{\mathrm{T}}(\tau^*,\mathbf{t}^*|\boldsymbol{\theta},\mathbf{e},\mathbf{i})$ is given by Equation 3. I assume that the tree process is a critical birth–death model (i.e., birth and death rates are equal) with parameter $\boldsymbol{\theta}$. The joint density of node ages under this model is as follows (see Equation 3.19 in Stadler 2008, with $\mu\to\lambda$):

$$p_{\mathrm{T}}(t_1,t_2|\boldsymbol{\theta},n=3)=3!\frac{\boldsymbol{\theta}}{(1+\boldsymbol{\theta}t_1)^3}\frac{\boldsymbol{\theta}}{(1+\boldsymbol{\theta}t_2)^2}.$$

Only one ranked tree topology (τ^*) has nonzero probability (see Fig. 1). More precisely $\Pr_{\mathrm{T}}(\tau)\chi(\mathbf{t}^*,\tau,\mathbf{e},\mathbf{i})=1$ when $\tau=\tau^*$ and $\Pr_{\mathrm{T}}(\tau)\chi(\mathbf{t}^*,\tau,\mathbf{e},\mathbf{i})=0$ otherwise. In fact, if $\tau\neq\tau^*$, then $\Pr_{\mathrm{T}}(\tau)\chi(\mathbf{t},\tau,\mathbf{e},\mathbf{i})=0$ for all $\mathbf{t}$. Considering the special case where $v<x$ (i.e., the two calibration time intervals do not overlap), $p_{\mathrm{T}}(\tau,\mathbf{t}|\mathbf{e},\mathbf{i},\boldsymbol{\theta})$ is a function of $Z_{\boldsymbol{\theta}}$ which is expressed as follows:

$$Z_{\boldsymbol{\theta}}=\sum_\tau\int p_{\mathrm{T}}(\mathbf{t}|\boldsymbol{\theta})\Pr_{\mathrm{T}}(\tau)\chi(\mathbf{t},\tau,\mathbf{e},\mathbf{i})\mathrm{d}\mathbf{t}$$

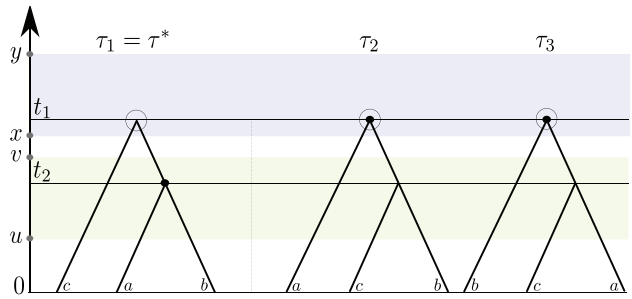

FIGURE 1. An example with three taxa. τ_1, τ_2, and τ_3 are the three ranked tree topologies. τ_1 corresponds to the maximum likelihood ranked tree topology (τ^*). t_1 and t_2 are node ages. They also correspond to the maximum likelihood estimates of these parameters (i.e., if $\tau=\tau^*$, then $t_1^*=t_1$ and $t_2^*=t_2$). u and v are the younger and older bounds for the calibration time interval i with $e=\{a,b\}$ the corresponding subset of taxa being calibrated. x and y are the younger and older bounds for the calibration interval i' and $e'=\{a,b,c\}$ is the subset of taxa that defines this calibration. The black disks and open circles indicate the MRCA of e and e', respectively. For τ_2 and τ_3, the age of the MRCA for a and b (respectively a, b, and c) cannot fall within its calibration interval, provided the age of the MRCA of a, b and c (respectively a and b) is inside its calibration interval. Therefore, $\Pr_{\mathrm{T}}(\tau_2)\chi(t_1,t_2,\tau_2,\mathbf{e},\mathbf{i})=0$ and $\Pr_{\mathrm{T}}(\tau_3)\chi(t_1,t_2,\tau_3,\mathbf{e},\mathbf{i})=0$ for all t_1 and t_2.

$$\begin{aligned}&=\int_x^y\int_u^v 2\frac{\boldsymbol{\theta}}{(1+\boldsymbol{\theta}t_1)^3}\frac{\boldsymbol{\theta}}{(1+\boldsymbol{\theta}t_2)^2}\mathrm{d}t_2\mathrm{d}t_1\\&=\frac{\boldsymbol{\theta}^2(u-v)(\boldsymbol{\theta}x^2-\boldsymbol{\theta}y^2+2x-2y)}{(\boldsymbol{\theta}u+1)(\boldsymbol{\theta}v+1)(\boldsymbol{\theta}x+1)^2(\boldsymbol{\theta}y+1)^2}.\end{aligned}$$

Taking $f(\boldsymbol{\theta})\propto 1$, the posterior density for $\boldsymbol{\theta}$ is then:

$$p(\boldsymbol{\theta}|d_n,\boldsymbol{\alpha})=\frac{p_{\mathrm{T}}(\mathbf{t}^*|\boldsymbol{\theta},n=3)}{Z_{\boldsymbol{\theta}}K}, \tag{11}$$

where Equation 10 gives:

$$K\propto\int_0^\infty\frac{p_{\mathrm{T}}(\mathbf{t}^*|\boldsymbol{\theta},n=3)}{Z_{\boldsymbol{\theta}}}\mathrm{d}\boldsymbol{\theta}.$$

When ignoring $Z_{\boldsymbol{\theta}}$, that is using the tree-dependent approach, the posterior density of $\boldsymbol{\theta}$ is instead:

$$p^*(\boldsymbol{\theta}|d_n,\boldsymbol{\alpha})=\frac{p_{\mathrm{T}}(\mathbf{t}^*|\boldsymbol{\theta},n=3)}{K^*} \tag{12}$$

where

$$K^*\propto\int_0^\infty p_{\mathrm{T}}(\mathbf{t}^*|\boldsymbol{\theta},n=3)\mathrm{d}\boldsymbol{\theta}.$$

Values of K and K^* were computed for different $\mathbf{t}^*$, u, v, x, and y using numerical integration routines available in Maple 17 (http://www.maplesoft.com/).

As mentioned previously, I first consider the case where there is no uncertainty in calibration constraints such that $g(\mathbf{e}=(e,e'),\mathbf{i}=(i,i')|\boldsymbol{\alpha})=1$. Figure 2 shows the impact of the width of the calibration time interval for the MRCA of $\{a,b\}$ on the marginal posterior of $\boldsymbol{\theta}$. While that width does not affect the tree-dependent densities (in green), the tree-independent ones (in red) behave differently. When the calibration constraint is tight (e.g., $u=0.45$ and $v=0.55$), the posterior distribution of the birth–death parameter is virtually uniform under

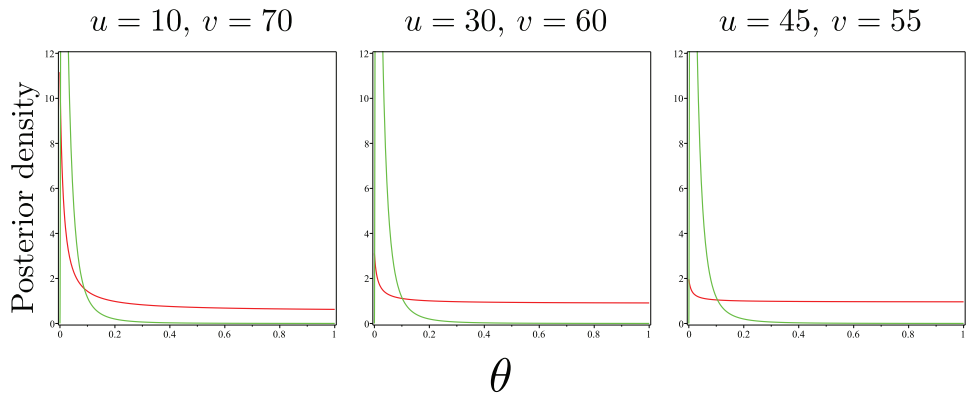

FIGURE 2. Impact of tightening calibration bounds on the posterior distributions of $\boldsymbol{\theta}$ using the tree-independent (in red) and the tree-dependent (in green) models. The values of u and v, defining the calibration time interval for the age of clade $\{a,b\}$ are given above each plot while that of t_1^* and t_2^* are set to 100 and 50 time units throughout, respectively. The younger and older bounds for the age of the MRCA of $\{a,b,c\}$ are $x=90$ and $y=110$, respectively.

the tree-independent scheme. This flattening of the posterior distribution is expected under that model. Indeed, among all the possible birth–death trees, one only considers those where t_2 falls within $[u,v]$ (and t_1 within $[x,y]$). Hence, the data-generating process is heavily censored here and only a small fraction of all possible birth–death trees are observable when the time interval $[u,v]$ is narrow, thereby decreasing the signal conveyed by the data about $\boldsymbol{\theta}$.

The posterior distribution of $\boldsymbol{\theta}$ is not a function of $\mathbf{e}$ and $\mathbf{i}$ with the tree-dependent approach (provided the calibration constraints are fulfilled). The green curves in the three settings tested in Figure 2 are thus all identical indeed. The difference of behavior between the two techniques illustrates the fact that all the information about $\boldsymbol{\theta}$ is conveyed by the tree (i.e., τ and $\mathbf{t}$) under the tree-dependent approach. As already mentioned above, the tree-independent technique acknowledges instead that, while τ and $\mathbf{t}$ conveys information about $\boldsymbol{\theta}$, $\mathbf{e}$, and $\mathbf{i}$ only allow a fraction of all realizations of the tree variable to be observed. It is not obvious whether one approach is more relevant than the other since they both tackle the same problem from two distinct modeling angles (but see Discussion section).

I next considered the impact of uncertainty around calibration constraints. The calibrations of the MRCAs of $\{a,b\}$ and $\{a,b,c\}$ correspond here to the time intervals $[u,v]$ and $[x,y]$ with probability α, whereby $u=1$, $v=90$, $x=90$, and $y=110$. Also, with probability $1-\alpha$, the two subsets of taxa are calibrated with the same time interval $[u,y]$. Let i'' denote this last time interval. We thus have $g(\boldsymbol{e}=(e,e'),\boldsymbol{i}=(i,i')|\alpha)=\alpha$ and $g(\boldsymbol{e}=(e,e'),\boldsymbol{i}=(i'',i'')|\alpha)=1-\alpha$. The proposed calibration model therefore accommodates for the uncertainty around the maximum (older) age of the MRCA of $\{a,b\}$ and the minimum (younger) age of that of $\{a,b,c\}$. In the next paragraph, I will refer to these two alternatives as the two-interval and the one-interval scenarios. I will focus on the interplay between the strength of fossil evidence for the two-interval scenario and the signal conveyed by molecular data. If molecular data tend to push the values of t_2^* upwards (i.e., close to t_1^*), then it favors the one-interval scenario, which may go against a strong support for the two-interval scenario conveyed by fossil data (if the value of α is close to one).

Figure 3 gives the posterior probability of model parameters as a function of t_2^* and α. The contour plot indicates that α mitigates the impact of molecular data, as expected, with larger values of this scalar increasing the support for the two-interval scenario, no matter how close t_2^* is from t_1^*. In other words, if fossil data strongly suggest that the two-interval scenario is the most plausible ($\alpha>0.8$), then the signal conveyed by molecular data only weakly impacts that belief. For values of α smaller than $\sim$0.2, the posterior probability of the two-interval scenario quickly drops as t_2^* gets closer to t_1^*, conveying the idea that the one-interval scenario is indeed preferable when both fossil and molecular data weakly support the two-interval scenario (α small, t_2^* close to t_1^*).

THE ORIGINS OF FLOWERING PLANTS

Smith et al. (2010) conducted a thorough analysis of the timing of speciation in land plants. They used a nucleotide sequence data set with 154 taxa and three genes (18S, *atp*B, and *rbc*L) totaling 4533 bp. The fossil data available provide calibration time intervals for 33 sets of taxa. The authors performed two analyses: one with a maximum age for the origin of eudicots set to 125 Ma and another without this particular constraint. Because geographical and morphological evidence suggest an earlier origin for that clade, this datum was discarded and the analysis conducted here focuses on the remaining 32 calibration intervals.

Smith et al. (2010) used the "product-of-marginals" approach implemented in BEAST 1.4.7. A log-normal probability density was used to model the marginal distribution corresponding to each fossil.

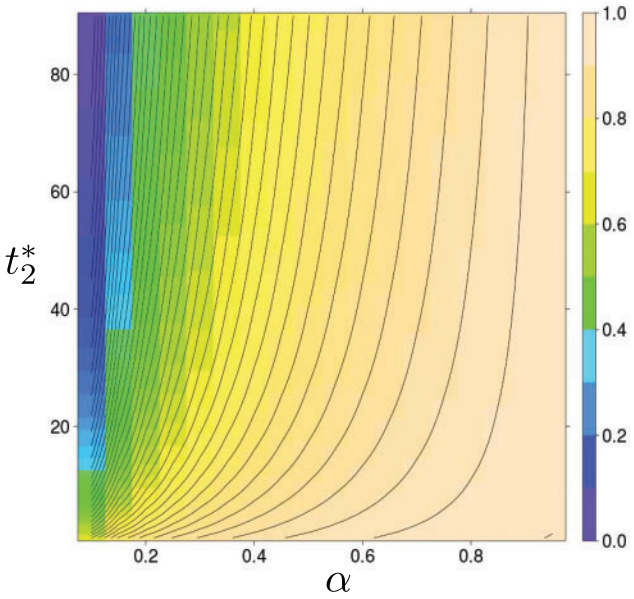

FIGURE 3. Posterior probability of the model parameters for variable calibration constraints. The contour plot gives the values of $\Pr(\tau = ((a,b),c), t_1^* = 100, t_2^* = \cdot, \boldsymbol{e} = (\{a,b\},\{a,b,c\}), \boldsymbol{i} = ([u,v],[x,y]) | d_n, \alpha = \cdot)$ as a function of t_2^* and α, with $u = 1$, $v = 90$, $x = 90$ and $y = 110$.

Each distribution was offset by a value corresponding to the minimum age of each clade (see Table S2 in their article). These values were used in my own analysis to define the younger bounds for the ages of the same clades. The corresponding older bounds are less straightforward to define as fossil data do not provide precise information about them. A preliminary analysis using the 95% quantiles of every lognormal distribution with mean and standard deviation as determined by the authors (given in their Table S2) revealed that the timing of some events (e.g., the origins of Eudicots) was largely defined by this soft older bound (i.e., increasing the standard deviation of the lognormal prior distributions also increased the median posterior ages). I thus elected to use a less stringent strategy whereby all calibrated nodes were constrained to be younger than the older bound of the oldest calibration (corresponding to the stem age of the clade *Tracheophyta*). As in the preliminary analysis, this older bound was given by the 95% quantile of the corresponding lognormal, giving an age of 452 Ma.

The sequence alignment resulting from the concatenation of the three genes was analyzed under the HKY nucleotide substitution model (Hasegawa et al. 1985) and the FreeRate model (Soubrier et al. 2012), which is a nonparametric mixture model (with three classes here) that accommodates for the heterogeneity of rates across sites. Truncated normals were used to model the distributions of substitution rates on the edges of the phylogeny. Let $w_i := r_i c$ be the average substitution rate on edge i. The parameter c corresponds to the "clock rate" of substitution which is common to all edges, while r_i corresponds to a multiplicative factor that is specific to edge i. The value of w_i was assumed to be a random draw from a normal distribution truncated to positive values, with mode set to c and standard deviation cv. Therefore, rates are not autocorrelated *a priori* under this model, following (Smith et al., 2010) analysis. The parameter v measures here the deviation from the strict clock assumption. Its posterior distribution was estimated from the data. Lastly, the tree process was considered to be a birth-death model with birth and death parameters λ and μ. Complete sampling of lineages was assumed here since the fraction of sampled lineages cannot be estimated whenever the birth and death of lineages are considered as two separate parameters (Stadler 2009).

Two series of experiments were performed. In the first, five MCMC analyses were run separately using different random seeds to initiate the analysis. The values of λ and μ were updated using the exchange algorithm, under the tree-independent model. The second series consisted in five separate analyses where the same two parameters were updated using the tree-dependent approach, that is ignoring the normalization factor $Z_{\boldsymbol{\theta}}$ in Equation 6. The analysis of the trace files produced showed that the effective sample size for each parameter was generally well beyond 200. Comparison of the five replicates for each of the two methods also indicated that the sampling had systematically converged to the same ranges of parameter values.

The analysis involving 32 fossils did not reveal any substantial difference between node ages estimated under the tree-dependent and tree-independent techniques. The 95% posterior credibility intervals for the timing of diversification of angiosperms were [244; 307] and [247; 304] Ma, respectively. Similarly, the origin of eudicots was estimated to have taken place in [184; 240] and [189; 243] Ma with these two approaches. These estimates are older than those reported in Smith et al. (2010), although the credibility intervals for the origins of angiosperms reported here overlap with that reported in their study. Also, when using BEAST 1.7.4 in the same conditions as in Smith et al. (2010), increasing the standard deviation for every calibration distribution from 0.5 to 10 except for that of the oldest fossil leads to node age estimates similar to those obtained here. This result suggests a high sensitivity of posterior age estimates to the prior distributions used for calibration in the Smith et al. (2010) study. Similarly, Bell et al. (2010) reported younger age estimates for the origins of angiosperms but only considered tight calibration intervals in their Bayesian analysis without clear justification for that choice.

In order to further investigate the impact of the amount of fossil data available, I randomly picked 16 out of the 32 fossil data points available and ran five independent repeats of the analyses under the tree-dependent and tree-independent models in conditions identical to those used before. The time estimates obtained with the tree-independent model are noticeably younger than those returned by the tree-dependent one. Figure 4 shows the posterior distributions and node ages corresponding to the origins of eudicots, angiosperms

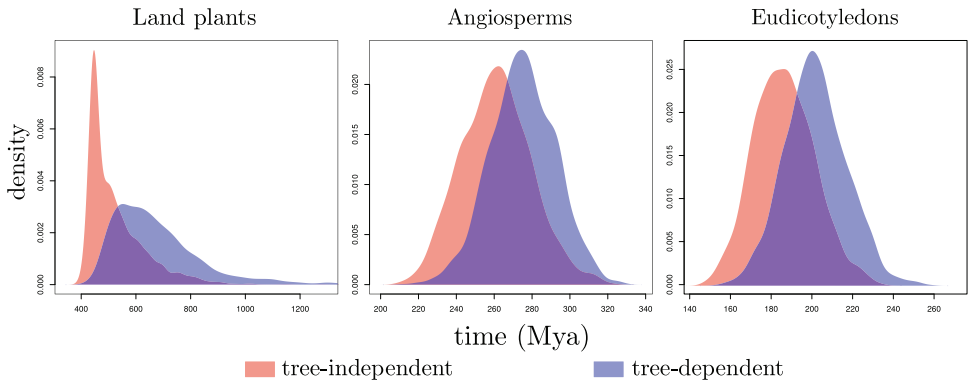

FIGURE 4. Impact of ignoring Z_θ on the inferred timing of speciation in lands plants. Smith et al. (2010) data set was analyzed with a subsample of 16 fossils (randomly sampled in the full set of 32 fossils). The timing of diversification of land plants, angiosperms and eudicots were estimated using the "traditional" tree-dependent approach that ignores Z_θ (in blue) and the exchange algorithm that implements the tree-independent technique (in pink).

and land plants. The 95% credibility intervals for these three events are [159;225], [221;303] and [428;990] Ma, respectively under the tree-independent model. Using the tree-dependent approach, the equivalent intervals are [177;337], [240;448, and [475;1367] Ma. Substantial differences in the posterior distribution of the birth parameter are also observed: the 95% credibility intervals with the tree-independent and the tree-dependent techniques are [0.006; 0.014] and [0.005; 0.008] respectively. Conversely, the posterior distributions of the death parameter do not show any noticeable difference between the two approaches.

In conclusion, the analysis using the full set of 32 fossils does not reveal any obvious difference between node age estimates using the technique introduced in this study compared to the traditional tree-dependent approach. Yet, the analysis based on a reduced number of fossils gives substantial differences in node age and tree process parameter estimates. In particular, the tree-independent approach produces younger node ages compared to that obtained with the more traditional model. The "new" estimates thus provide a closer match to the stratigraphic record, even though the gap between the two is still considerable. Inference of the birth parameter is also impacted with a wider credibility interval obtained under the tree-independent approach, as expected.

DISCUSSION

Hierarchical Bayesian modeling provides a suitable framework for inferring the timing of evolutionary events from the joint analysis of molecular and fossil data. On the first level of the hierarchy, molecular data convey evidence about the evolutionary history of sampled species. This history forms the basis of the second level of the hierarchy whereby fossil data help disentangling times and rates of evolution. Although this construction is fairly standard in statistics, meaningful Bayesian inference requires proper mathematical modeling of all aspects of the hierarchy.

The top level of this hierarchy, corresponding to the probability of the sequence alignment given a phylogenetic tree, suffers no ambiguity. The lower level, however, is more difficult to comprehend. Although the product-of-marginals approach is very popular and fairly straightforward at first sight, it has conceptual issues. As already pointed out in Rannala (2016) and elsewhere (see e.g., Warnock et al. 2011), the distributions of node ages defined by the tree process with calibration constraints generally conflict with the user-defined distributions of ages for specific groups of species, thereby limiting the relevance of the latter.

The present study relies on a new approach to accommodate for the effect of fossil data on the probabilistic distribution of trees generated by a birth and death process. While the traditional approach consists in modeling the distribution of calibration constraints conditional on the underlying phylogeny, the tree-independent model considers instead that the distribution of ranked trees is conditioned on the calibration constraints. This last approach gives leeway to accommodate for expert knowledge about the placement of these constraints in the tree and the associated uncertainty. A given constraint can apply to several groups of taxa, with an associated probability distribution. Inference under this new modeling approach relies on an original statistical technique that was designed to deal with doubly intractable distributions. Murray et al. (2012) recently described a sampling approach—the so-called "exchange algorithm"—that generates valid random draws from this type of distribution. This algorithm involves a modest computational overhead compared to the standard approach and is relatively straightforward to implement.

The analysis of a three-taxon data set with sequences of infinite length shows that the precision with which the tree process parameters are estimated depends on the width of the time intervals defining the calibration constraints. The narrower the time range, the broader the posterior distribution. This result appears to be

counter-intuitive at first sight. It is in fact well explained when examining the underlying experimental design assumed here. The tree-independent model considers that the morphological features of the fossil(s) influence which species are selected for the molecular dating analysis. The timing of divergence is thus informed by the calibration time interval, not the other way around. This time interval does not convey any information about the rate at which lineages arise or vanish. Instead, it may reflect the difficulty or ease with which particular geological layers can be accessed for collecting fossils. In this context, a short calibration time interval constrains the species divergence times to occur in a narrow "window of observation," thereby explaining the lack of signal in the data for the tree-process parameters. In the tree-dependent model, species are selected first and the sampling of fossil data is conditioned on the morphological features of the sampled lineages. In the given example, all the information about the tree-process parameters is conveyed by the phylogeny in as sequences are of infinite lengths and the substitution rate is considered as known. Therefore, fossil data do not play any role in the estimation of these parameters and the three posterior distributions obtained under this model are thus all identical.

The reanalysis of a land plant data set indicates that the standard product-of-marginals, the tree-dependent and the tree-independent methods return similar node age estimates when using comparable priors and a substantial number of calibration constraints, which is reassuring. However, noticeable differences between node age estimates are observed with a smaller number of calibration constraints. Younger ages are derived under the tree-independent model compared to the tree-dependent approach. Although it is not clear whether this particular observation corresponds to a real trend, the gap between node ages obtained from fossils and molecules appears to be smaller here with the tree-independent technique.

The exchange algorithm relies on simulating the tree process conditional on time constraints coming from fossil data. In the present study, this task involved a series of Metropolis–Hastings steps updating different components of the model parameters. This approach is efficient from a computational perspective. Nonetheless, direct simulation from the generating process would be preferable. Although generating birth–death or coalescent trees is straightforward, incorporating time constraints for some clades in these simulations is challenging. Efficiently generating random trees conditional on calibration constraints would also help testing the correctness of the implementation of Bayesian samplers (through the comparison of sampled and simulated tree distributions, ignoring sequence data). Furthermore, such a generator would also help assessing the impact of calibration data on divergence time estimates through simulations.

Finally, the method proposed in this study belongs to the family of "node-dating" techniques. It does not rest on stochastic models describing the evolution of morphological characters for calibrating the molecular clock. This lesser degree of sophistication compared to the total-evidence technique can be perceived as a weakness. Yet, both approaches achieve the same goal in accounting for the uncertainty in the placement of calibration constraints in the tree. Moreover, the technique described here provides a suitable statistical framework for modeling the occurrence of fossils across stratigraphic ranges that can accommodate for the complexities of the underlying fossilization and sampling processes. It should therefore contribute to improving the techniques for dating evolutionary events from the analysis of genetic and fossil data.

Software

The model and sampling techniques described in this article are available in the PhyTime software (Guindon 2013), which is part of the PhyML package available at https://github.com/stephaneguindon/phyml.

Acknowledgments

The work presented in this article greatly benefited from in depth discussions with Alexandra Gavryushkina. I would also like to thank Pierre Pudlo for pointing me to Murray et al. (2012) article about sampling from doubly intractable distributions; Jeremy Beaulieu along with Michael Donoghue for sharing with me the plant data set; Tanja Stadler, David Welch, Emmanuel Douzery and Sophie Mignon for discussions; David Bryant, Mario dos Reis and an anonymous reviewer for excellent remarks and suggestions that improved this work.

References

Arcila D., Pyron R.A., Tyler J.C., Ortí G., Betancur-R R. 2015. An evaluation of fossil tip-dating versus node-age calibrations in tetraodontiform fishes (teleostei: Percomorphaceae). Mol. Phylogenet. Evol. 82:131–145.

Bell C.D., Soltis D.E., Soltis P.S. 2010. The age and diversification of the angiosperms re-revisited. Am. J. Bot. 97:1296–1303.

Benton M.J., Ayala F.J. 2003. Dating the tree of life. Science 300:1698–1700.

Benton M.J., Donoghue P.C. 2007. Paleontological evidence to date the tree of life. Mol. Biol. and Evol. 24:26–53.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.-H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Comput. Biol. 10:e1003537.

Bromham L., Penny D., Rambaut A., Hendy M.D. 2000. The power of relative rates tests depends on the data. J. Mol. Evol. 50:296–301.

Didier G., Royer-Carenzi M., Laurin M. 2012. The reconstructed evolutionary process with the fossil record. J. Theoret. Biol. 315:26–37.

dos Reis M., Donoghue P.C., Yang Z. 2014. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biol. Lett. 10:20131003.

dos Reis M., Donoghue P.C., Yang Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nat. Rev. Genet. 17:71–80.

Edwards A.W. 1970. Estimation of the branch points of a branching diffusion process. J. R. Stat. Soc. B 32:155–174.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.
Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Syst. Biol. 66:57–73.
Gavryushkina A., Welch D., Stadler T., Drummond A.J. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS Comput. Biol. 10:e1003919.
Grimm G.W., Kapli P., Bomfleur B., McLoughlin S., Renner S.S. 2014. Using more than the oldest fossils: dating Osmundaceae with three Bayesian clock approaches. Syst. Biol. 64:396–405.
Guindon S. 2013. From trajectories to averages: an improved description of the heterogeneity of substitution rates along lineages. Syst. Biol. 62:22–34.
Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial-DNA. J. Mol. Evol. 22:160–174.
Heath T.A., Huelsenbeck J.P., Stadler T. 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proc. Natl. Acad. Sci. USA 111:E2957–E2966.
Kumar S., Hedges S.B. 2016. Advances in time estimation methods for molecular data. Mol. Biol. Evol. 33:863–869.
Murray I., Ghahramani Z., MacKay D. 2012. MCMC for doubly-intractable distributions. arXiv preprint arXiv:1206.6848.
Parham J.F., Donoghue P.C., Bell C.J., Calway T.D., Head J.J., Holroyd P.A., Inoue J.G., Irmis R.B., Joyce W.G., Ksepka D.T., Patané J.S.L., Smith N.D., Tarver J.E., van Tuinen M., Yang Z., Angielczyk K.D., Greenwood J.M., Hipsley C.A., Jacobs L., Makovicky P.J., Müller J., Smith K.T., Theodor J.M., Warnock R.C.M., Benton M.J. 2011. Best practices for justifying fossil calibrations. Syst. Biol. 61:346–359.
Puttick M.N., Thomas G.H., Benton M.J. 2016. Dating placentalia: Morphological clocks fail to close the molecular fossil gap. Evolution 70:873–886.
Pyron R.A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Syst. Biol. 60:466–481.
Rannala B. 2016. Conceptual issues in Bayesian divergence time estimation. Philos. Trans. R. Soc. B 371:20150134.
Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Syst. Biol. 61:973–999.
Saladin B., Leslie A.B., Wüest R.O., Litsios G., Conti E., Salamin N., Zimmermann N.E. 2017. Fossils matter: improved estimates of divergence times in pinus reveal older diversification. BMC Evol. Biol. 17:95.
Sanderson M. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14:1218–1231.
Sanderson M. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19:101–109.
Sarich V., Wilson A. 1967. Immunological time scale for hominid evolution. Science 158:1200–1203.
Sauquet H., Ho S.Y., Gandolfo M.A., Jordan G.J., Wilf P., Cantrill D.J., Bayly M.J., Bromham L., Brown G.K., Carpenter R.J., Lee D.M., Murphy D.J., Sniderman J.M.K., Udovicic F. 2011. Testing the impact of calibration on molecular divergence times using a fossil-rich group: the case of Nothofagus (Fagales). Syst. Biol. 61:289–313.
Smith S.A., Beaulieu J.M., Donoghue M.J. 2010. An uncorrelated relaxed-clock analysis suggests an earlier origin for flowering plants. Proc. Natl. Acad. Sci. USA 107:5897–5902.
Soubrier J., Steel M., Lee M.S., Der Sarkissian C., Guindon S., Ho S.Y., Cooper A. 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol. Biol. Evol. 29:3345–3358.
Stadler T. 2008. Evolving trees: models for speciation and extinction in phylogenetics [Ph.D. thesis]. Technische Universität München, Zentrum Mathematik. Munchen, Germany.
Stadler T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. 261:58–66.
Stadler T. 2010. Sampling-through-time in birth–death trees. J. Theor. Biol. 267:396–404.
Tavaré S., Marshall C.R., Will O., Soligo C., Martin R.D. 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416:726–729.
Thorne J., Kishino H., Painter I. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.
Warnock R.C., Parham J.F., Joyce W.G., Lyson T.R., Donoghue P.C. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proc. R. Soc. B 282:20141013.
Warnock R.C., Yang Z., Donoghue P.C. 2011. Exploring uncertainty in the calibration of the molecular clock. Biol. Lett. rsbl20110710.
Wilkinson R., Steiper M., Soligo C., Martin R., Yang Z., Tavaré S. 2010. Dating primate divergences through an integrated analysis of palaeontological and molecular data. Syst. Biol. 60:16–31.
Yang Z. 2006. Computational molecular evolution. Oxford, United Kingdom: Oxford University Press. Oxford, United Kingdom.
Zhang C., Stadler T., Klopfstein S., Heath T.A., Ronquist F. 2016. Total-evidence dating under the fossilized birth–death process. Syst. Biol. 65:228–249.