The reconstructed evolutionary process with the fossil record

Gilles Didier, Manuela Royer-Carenzi, Michel Laurin

▶ **To cite this version:**

Gilles Didier, Manuela Royer-Carenzi, Michel Laurin. The reconstructed evolutionary process with the fossil record. Journal of Theoretical Biology, 2012, 315, pp.26-37. 10.1016/j.jtbi.2012.08.046 . hal-01105197

HAL Id: hal-01105197
https://hal.science/hal-01105197

Submitted on 16 Feb 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The reconstructed evolutionary process with the fossil record

Gilles Didier, Manuela Royer-Carenzi and Michel Laurin

February 16, 2022

Abstract

Using the fossil record yields more detailed reconstructions of the evolutionary process than what is obtained from contemporary lineages only. In this work, we present a stochastic process modelling not only speciation and extinction, but also fossil finds. Next, we derive an explicit formula for the likelihood of a reconstructed phylogeny with fossils, which can be used to estimate the speciation and extinction rates. Finally, we provide a comparative simulation-based evaluation of the accuracy of estimations of these rates from complete phylogenies (including extinct lineages), from reconstructions with contemporary lineages only and from reconstructions with contemporary lineages and the fossil record. Results show that taking the fossil record into account yields more accurate estimates of speciation and extinction rates than considering only contemporary lineages.

1 Introduction

The biodiversity and its history were shaped by speciation, extinction, and diversification (the difference between both). The rates of these processes and significant shifts in these, which have occurred in evolutionary radiations and mass extinction events, have fascinated generations of paleontologists and evolutionary biologists (Axelrod and Bailey, 1968; Lewin, 1983; Moore and Donoghue, 2009). For most of the 20$^{\text{th}}$ century, such studies emphasized the fossil record (e.g. Raup and Sepkoski, 1984), to the point that it was believed that the evolutionary history of a taxon could not be know if it lacked a fossil record (Gingerich, 1979, page 454). However, with the rise of molecular phylogenetics in the last two decades, these studies have come to be based mostly on phylogenies of extant taxa (Moore and Donoghue, 2009). Most sophisticated analytical methods developed in the last two decades were developed for molecular phylogenies of extant taxa (Nee et al., 1994; Paradis, 2004). Few recent studies have used the fossil record, and fewer still, phylogenies incorporating fossils, to track changes in biodiversity (Ruta et al., 2007; Marjanović and Laurin, 2008). This situation is suboptimal because the fossil record provides the most direct data on the evolution of biodiversity. Several reasons may explain the relative neglect of fossil data in recent studies on the evolution of biodiversity. First, the incompleteness of the fossil record requires more complex methods to deal with the fact that we can never hope to have a complete sample of all extinct lineages of a clade (Foote and Raup, 1996), whereas it is possible to sample

all extant lineages, at least for well-known organisms such as vertebrates (e.g. Bininda-Emonds et al., 2007). This problem is surely not unsurmountable because most statistical methods can deal with random samples (very few methods require having exhaustive data on populations). Second, placing fossils in cladograms may be a little more difficult than for extant taxa, although this problem has certainly been exaggerated (Donoghue et al., 1989), and its magnitude probably depends on the taxon morphological complexity of the taxon (vertebrates are much easier to deal with than bacteria, for instance). Indeed, fossils have even been shown to play a critical role in phylogenetic inference, despite the fact that incomplete information is available about their morphology (e.g. Huelsenbeck, 1991; Lee, 2009). Third, branch lengths are more difficult to estimate with fossil data than for extant taxa because the absence of a universal morphological centralized databank equivalent to Genbank means that most large-scale studies need to rely on supertrees (e.g. Hone et al., 2005), rather than on original data matrices. It is also more difficult to use morphological than molecular datasets to infer branch lengths because morphological matrices are more time-consuming to compile than molecular ones, and because such matrices have generally been compiled to solve phylogenetic problems, which means that autapomorphies have tended to be disregarded and the character sample can hardly be considered to be random. Furthermore, morphological characters are more complex than molecular ones, and cannot be easily partitioned (there is no equivalent to nuclear vs. mitochondrial, coding vs. non-coding, or codon position). Consequently, our models of morphological character evolution are not as reliable as molecular ones. However, attempts at using morphological data to infer branch length have recently been made (Pyron, 2011), and various other solutions have been proposed to obtain very approximate branch lengths (e.g. Laurin, 2004; Marjanović and Laurin, 2007).

Thus, these problems inherent in the use of fossil data, which may have discouraged many evolutionary biologists from using them recently, do not appear to be unsurmountable. The benefits of incorporating fossil data are great, simply because these data, when sufficiently abundant, provide direct evidence about the timing of major events in evolution, as shown by several studies on biological crises (e.g. Axelrod and Bailey, 1968; Lewin, 1983; Ward et al., 2005).

Taxonomic diversification is classically modeled as a birth-death model (Kendall, 1948; Nee et al., 1994). The simplest model assumes that birth (speciation) and death (extinction) rates are constant over time and lineages. Throughout this paper, "speciation" designates a cladogenesis, and species are considered to be evolutionary lineages that exist between two nodes on an evolutionary tree (de Queiroz, 1998). In the case where such a birth-death process is continuously observed, the properties of the maximum likelihood estimators of its parameters have been well studied (Keiding, 1975). Unfortunately, this is clearly not the case here since the times involved in the evolution of complex species are far beyond our own timescale. It follows that to be studied, the speciation-extinction process has to be reconstructed from what can be observed from the present day. To do so, essentially two kinds of data are available: the extant taxa and the fossil record. Nee et al. (1994) show how to use the contemporary taxa to estimate the speciation and extinction rates, under the assumption that their phylogenetic tree can be reconstructed in an accurate way by molecular methods. This reconstructed tree does not include any extinct lineage but it is shown that it can be seen as the realization of a new stochastic process, called

the reconstructed process, which is a nonhomogeneous pure birth process with a time-dependent birth rate explicitly expressed from birth and death rates of the initial model of speciation-extinction (Nee et al., 1994). The likelihood of a phylogenetic tree of contemporary lineages can be computed as a function of speciation and extinction rates, allowing estimation of these rates by maximum likelihood. Paradis (2004) evaluated the performances of these estimations with a simulation-based protocol. The so-called reconstructed process has been studied recently by Gernhard (2008) and Hallinan (2012).

In this work, we propose the same general approach by considering reconstructions of the speciation-extinction process taking into account not only the contemporary lineages, but also the fossil record. As in the case of reconstruction from contemporary taxa, we assume that the reconstruction is perfectly accurate for both contemporary and fossilized lineages (see Figure 1 for an example of realization of the whole process and its reconstructions with and without the fossil record). A first step is to incorporate the fossil record into the general model of evolution of species. This is done by assuming that we find a fossil dated at a time t of a lineage alive at t with a given rate, assumed constant over time and lineages. In other words, we model the fossil finds as a Poisson process running on the whole phylogenetic tree as in (Wilkinson and Tavaré, 2009). We show that the reconstruction with fossils of a realization of the whole process of evolution can be seen as a the realization of another stochastic process, roughly speaking a nonhomogeneous time-dependent birth-death process with rates expressed from the initial parameters: speciation, extinction and fossil recovery rates. We derive a formula for the likelihood of such a reconstruction, which allows us to estimate the speciation and extinction rates. Remark that taking into account uncertainty about topologies of the phylogenies, speciation times estimated from molecular data and/or fossil ages can be done by using our formula in the standard Bayesian framework with suitable a priori distributions and MCMC methods.

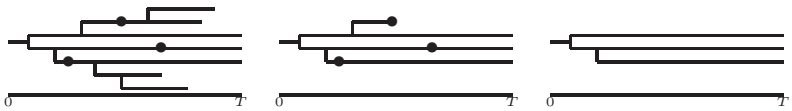
Finally, we evaluate the respective accuracy of estimations from complete realizations and their reconstructions with and without fossils in terms of absolute errors over simulated evolutions. We run simulations from several sets of speciation, extinction and fossil find rates. As expected, better estimations are obtained from complete realizations, but taking into account fossil finds into the reconstruction improves the accuracy of the estimations, even if the fossil find rate is quite small.

The software used to simulate the evolutionary trees and estimate speciation and extinction rates was developed in the C language and uses the GNU Scientific Library (Galassi et al., 2003) for random generators and maximum likelihood estimations. Source code and Linux executable are available upon request.

2 Evolutionary process and reconstructions

2.1 The complete process

Starting at time 0 with a single lineage, the stochastic process models three types of events which can occur over any lineage: birth, which gives rise to a new lineage, death, which extinguishes the lineage, and discovery of a fossil that



(a) (b) (c)

Figure 1: (a) a realization of the process $(Z(t), Y(t))$, (b) the reconstruction of this realization taking into account contemporary lineages and the fossil record, (c) the reconstruction of this realization from contemporary lineages only. Dots ($\bullet$) represent fossils finds.

is correctly dated and placed on a phylogenetic tree (actually the time when a fossil find event occurs in the process is the fossil date, i.e. its time of burial in the sediments), with respective per lineage rates λ, μ and γ. More formally, let $Z(t)$ denote the number of lineages alive at time t and $Y(t)$ the cumulated number of fossil finds from the beginning of the evolutionary process to t. The joint process $(Z(t), Y(t))$ is described by

$$(Z(t+h), Y(t+h)) = \begin{cases} (Z(t)+1, Y(t)) & \text{w.p.} \quad Z(t)\lambda h + o(h) \\ (Z(t)-1, Y(t)) & \text{w.p.} \quad Z(t)\mu h + o(h) \\ (Z(t), Y(t)+1) & \text{w.p.} \quad Z(t)\gamma h + o(h) \\ (Z(t), Y(t)) & \text{w.p.} \quad 1 - Z(t)(\lambda+\mu+\gamma)h + o(h) \end{cases}$$

where h is an interval of time. Abbreviation "w.p." stands for "with probability" and notation "$o(h)$" stands for "a function becoming insignificant relative to h as h tends to O". Basically, having $Z(t+h) = Z(t)+1$ means that a speciation has occurred between t and $t+h$ (first line of the evolution equations). In the same way, $Z(t+h) = Z(t)-1$ means that a lineage became extinct between t and $t+h$ (second line). We have $Y(t+h) = Y(t)+1$ if a fossil find is dated between t and $t+h$ (third line). If the numbers of lineages and fossils finds remain the same between t and $t+h$, then no event occurs during this interval of time (last line). The probability that more than a single event occur between t and $t+h$ is $o(h)$ thus negligible for an infinitesimal h.

Below, we assume that the evolutionary process starts with a single lineage and no fossil, more formally, that the initial condition $(Z(0), Y(0)) = (1, 0)$ is granted. We also make the technical (and quite usual) assumption that $\lambda > \mu$.

We first derive an expression for $\mathrm{P}(n,t) = \mathbb{P}((Z(t), Y(t)) = (n, 0))$, that is the probability to have n lineages alive at time t and no fossil discovery for the period between 0 and t. The evolution equations for the joint process without any fossil discovery between 0 and $t+h$ are

$$(Z(t+h), 0) = \begin{cases} (Z(t)+1, 0) & \text{w.p.} \quad Z(t)\lambda h + o(h) \\ (Z(t)-1, 0) & \text{w.p.} \quad Z(t)\mu h + o(h) \\ (Z(t), 0) & \text{w.p.} \quad 1 - Z(t)(\lambda+\mu+\gamma)h + o(h) \end{cases}$$

It follows that $\mathrm{P}(n,t)$ satisfies the differential equations

$$\frac{\mathrm{dP}(n,t)}{\mathrm{d}t} = (n-1)\lambda \mathrm{P}(n-1,t) + (n+1)\mu \mathrm{P}(n+1,t) - n(\lambda+\mu+\gamma)\mathrm{P}(n,t)$$

if $n > 0$, and

$$\frac{\mathrm{dP}(0,t)}{\mathrm{d}t} = \mu \mathrm{P}(1,t)$$

The corresponding generating function

$$\phi(x,t) = \sum_n x^n \mathrm{P}(n,t)$$

satisfies the linear partial differential equation

$$\frac{\partial \phi}{\partial t} + (-\lambda x^2 + (\lambda + \mu + \gamma)x - \mu)\frac{\partial \phi}{\partial x} = 0$$

which can be solved by using the method of characteristics to obtain

$$\phi(x,t) = \frac{\alpha\beta(1 - \exp(-\lambda(\beta-\alpha)t) - x(\alpha - \beta\exp(-\lambda(\beta-\alpha)t))}{\beta - \alpha\exp(-\lambda(\beta-\alpha)t)}$$
$$\times \left(1 - x\frac{1-\exp(-\lambda(\beta-\alpha)t)}{\beta - \alpha\exp(-\lambda(\beta-\alpha)t)}\right)^{-1}$$

where $\alpha < \beta$ are the roots of $-\lambda x^2 + (\lambda + \mu + \gamma)x - \mu = 0$, which are

$$\frac{\lambda + \mu + \gamma \pm \sqrt{(\lambda+\mu+\gamma)^2 - 4\lambda\mu}}{2\lambda}$$

2.2 Reconstructing evolution with the fossil record

Let us start by formalizing which parts of the evolutionary process can be actually reconstructed at the present day, namely at time T, from data about contemporary lineages and the fossil record. We assume from now on that we know the time origin of the evolution process. Namely it starts at time 0 with a single lineage.

A lineage ℓ, alive at time t, is said to be *observable at time t* if it does not go extinct before T or if a fossil of ℓ or of one of its descendants is found at a time greater than t. We say that ℓ is *observable* without further precisions, if it is observable at (at least) some time $t \leq T$.

The probability for lineage alive at time t to be observable at t is exactly its probability of not going extinct without leaving any fossil from time t, which is

$$\begin{aligned}\mathrm{P_o}(t) &= 1 - \mathrm{P}(0, T-t) \\ &= 1 - \phi(0, T-t) \\ &= 1 - \frac{\alpha\beta(1-\exp(-\lambda(\beta-\alpha)(T-t))}{\beta - \alpha\exp(-\lambda(\beta-\alpha)(T-t))}\end{aligned}$$

The part of a realization of the evolution which can be reconstructed from extant taxa and the fossil record is that which is observable (Figure 1-b).

It follows from the definition that a lineage observable at time t is observable at any time of its existence prior to t. Moreover all its ancestors are observable. As a corollary, a lineage giving birth to a lineage observable is itself observable at least until the time of this birth. In the same way, a fossil find dated at t can

occur only on a lineage observable at any time prior to t. Conversely, a lineage dying at time t cannot be observable at t (the model doesn't take into account that, though human scale is small with regard of that of evolution and some species did become extinct during the human scientific era, in which cases the exact time of extinction can be determined).

Let $Z_o(t)$ be the number of observable lineages living at time t. If we distinguish between observable and unobservable lineages in the stochastic process $(Z(t), Y(t))$, only five types of events can occur during an infinitesimal interval h starting at time t:

1. a speciation giving birth to an observable lineage w.p. $P_o(t)Z_o(t)\lambda h + o(h)$
2. a speciation giving birth to an unobservable lineage w.p. $((1-P_o(t))Z_o(t)+ Z(t) - Z_o(t))\lambda h + o(h)$
3. a fossil find of a lineage observable afterwards w.p. $P_o(t)Z_o(t)\gamma h + o(h)$
4. a fossil find of a lineage unobservable afterwards w.p. $(1-P_o(t))Z_o(t)\gamma h + o(h)$
5. the extinction of a lineage, which by definition was not observable at this time, w.p. $(Z(t) - Z_o(t))\mu h + o(h)$

Since we cannot observe the extinction of a lineage, the "death" event for an observable lineage corresponds with becoming unobservable, i.e. the end of its known stratigraphic range (a Type 4 event).

By considering exclusively the events modifying the numbers of observable lineages and/or of fossil finds, we get that the observable process $(Z_o(t), Y(t))$ is described by

$$(Z_o(t+h), Y(t+h)) = \begin{cases} (Z_o(t)+1, Y(t)) & \text{w.p.} \quad P_o(t)Z_o(t)\lambda h + o(h) \\ (Z_o(t), Y(t)+1) & \text{w.p.} \quad P_o(t)Z_o(t)\gamma h + o(h) \\ (Z_o(t)-1, Y(t)+1) & \text{w.p.} \quad (1-P_o(t))\, Z_o(t)\gamma h + o(h) \\ (Z_o(t), Y(t)) & \text{w.p.} \quad 1 - Z_o(t)\,(P_o(t)\lambda + \gamma)\, h + o(h) \end{cases}$$

and, under the assumption that $(Z(0), Y(0)) = (1, 0)$ (*i.e.* the underlying evolutionary process starts with a single lineage and no fossil),

$$(Z_o(0), Y(0)) = \begin{cases} (1,0) & \text{w.p.} \quad P_o(0) \\ (0,0) & \text{w.p.} \quad 1 - P_o(0) \end{cases}$$

This initial condition comes from the fact that, even though the underlying evolutionary process starts with a single lineage, this one is observable only with probability $P_o(0)$. In the complementary case, the process remains in the absorbing state $(0, 0)$, in other words, empty. We insist here on the fact that $(Z_o(t), Y(t))$ is the reconstructed process of $(Z(t), Y(t))$ and is not conditioned to any event (for instance to the fact that at least one lineage is observed).

In particular, the stochastic process $(Z_o(t))$ is a nonhomogeneous time-dependent birth-death process.

2.3 Waiting times between two observable events

In order to derive the likelihood of a realization of the process $(Z_o(t), Y(t))$, or equivalently, the likelihood of the reconstruction from the fossil record of a realization of the process $(Z(t), Y(t))$, we compute the distribution of waiting times between two successive events of $(Z_o(t), Y(t))$. We proceed as in (Nee et al., 1994) and define $W^o_{n,t}(s)$ as the probability of waiting more than s from time t until an event occurs, if there are n observable lineages alive at time t. We have

$$W^o_{n,t}(s + ds) = (1 - n\,(\lambda P_o(t + s) + \gamma)\,ds)\,W^o_{n,t}(s)$$

Solving the corresponding differential equation gives us

$$W^o_{n,t}(s) = \exp(-n(\gamma + \lambda - \lambda\alpha)s)\left(\frac{\beta - \alpha\exp(-\lambda(\beta-\alpha)(T-t-s))}{\beta - \alpha\exp(-\lambda(\beta-\alpha)(T-t))}\right)^n$$

The probability density function of waiting times is then

$$\begin{aligned} w^o_{n,t}(s) = n\exp(-n(\gamma + \lambda(1-\alpha))s)&\frac{(\beta - \alpha\exp(-\lambda(\beta-\alpha)(T-t-s)))^{n-1}}{(\beta - \alpha\exp(-\lambda(\beta-\alpha)(T-t)))^n} \\ &\times [\beta(\gamma + \lambda(1-\alpha)) - \alpha(\gamma + \lambda(1-\beta))\exp(-\lambda(\beta-\alpha)(T-t-s))] \end{aligned}$$

2.4 Probabilities of observable events

Given that an event occurs at time t, the relative probabilities of a birth (a speciation giving birth to an observable lineage), a death (actually a fossil find of a lineage becoming unobservable after t) and a fossil find (of a lineage still observable after t) do not depend on the number of lineages alive at t and are respectively

$$\begin{aligned} p_b(t) &= \frac{\lambda P_o(t)}{\lambda P_o(t) + \gamma} \\ p_d(t) &= \frac{\gamma(1 - P_o(t))}{\lambda P_o(t) + \gamma} \\ p_f(t) &= \frac{\gamma P_o(t)}{\lambda P_o(t) + \gamma} \end{aligned}$$

2.5 Likelihood of a reconstructed realization with fossils

From its initial condition, the likelihood of an empty realization of the stochastic process $(Z_o(t), Y(t))$ (*i.e.* a realization of $(Z(t), Y(t))$ without any observable lineage) is $1 - P_o(0)$. In order to compute the likelihood of a non-empty realization of $(Z_o(t), Y(t))$, we sort the times when an event occurs in increasing order $t_1 < t_2 < \ldots < t_k$ (Figure 2) and we set $t_0 = 0$. We note e_1, e_2, …, e_k the sequence storing the nature of the corresponding events. We have $e_i = \mathrm{b}$ (*resp.*

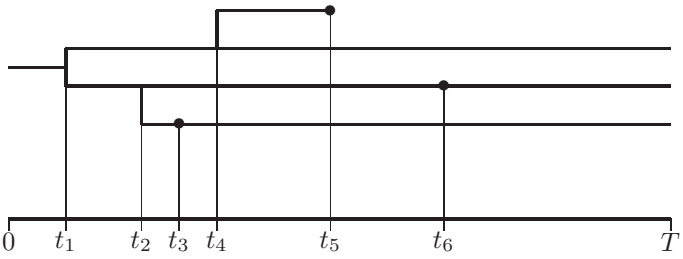


Figure 2: A realization of the reconstructed process with fossils where times of events are reported on the time axis. The sequence of kinds of corresponding events is $(e_i)_{1\leq i\leq 6} = (\mathrm{b}, \mathrm{b}, \mathrm{f}, \mathrm{b}, \mathrm{d}, \mathrm{f})$ where b stands for a birth of an observable lineage and f (resp. d) stands for a fossil find on a lineage observable (resp. unobservable) afterwards.

$e_i = \mathrm{d}$, $e_i = \mathrm{f}$), if the i^{th} event is a birth (*resp.* a death, a fossil find – see Figure 2). The likelihood of a non-empty realization with k observable events is

$$\mathrm{P}_{\mathrm{o}}(0) \left(\prod_{i=1}^{k} \mathrm{p}_{e_i}(t_i) \right) \left(\prod_{i=1}^{k} \mathrm{w}^{\mathrm{o}}_{Z_{\mathrm{o}}(t_i),t_{i-1}}(t_i - t_{i-1}) \right) \mathrm{W}^{\mathrm{o}}_{Z_{\mathrm{o}}(T),t_k}(T - t_k) \qquad (1)$$

This likelihood is written as a product of four factors. The first one comes from the fact that, if the realization is not empty, then we have necessarily $(Z_{\mathrm{o}}(0), Y(0)) = (1, 0)$, which occurs with probability $\mathrm{P}_{\mathrm{o}}(0)$. The second one is the product of all the relative probabilities of events occurring during the realization. The third factor is the product of the probability densities of the waiting time from the origin of time to the first event and all those between two successive events. The fourth one is for the waiting time between the last event and the present time. This last factor takes into account the fact that the evolutionary process is right-censored (unknown after time T).

The likelihood of a realization of the complete process $(Z(t), Y(t))$ or of one of the reconstructed process from contemporary lineages only can be written in the very same way.

2.6 Reconstructing without fossils

We apply here the same approach as above for reconstructing the evolutionary process from extant taxa only (i.e. without the fossil record), notably in order to point out and explain some differences between the likelihood we use and that of (Nee et al., 1994).

Since fossil finds are not taken into account here, we go back to the (single) process $(Z(t))$, described by

$$Z(t+h) = \begin{cases} Z(t)+1 & \text{w.p.} \quad Z(t)\lambda h + o(h) \\ Z(t)-1 & \text{w.p.} \quad Z(t)\mu h + o(h) \\ Z(t) & \text{w.p.} \quad 1 - Z(t)(\lambda+\mu)h + o(h) \end{cases}$$

where h is an interval of time. As for the joint process, we have the initial condition $Z(0) = 1$ (the process starts with a single lineage).

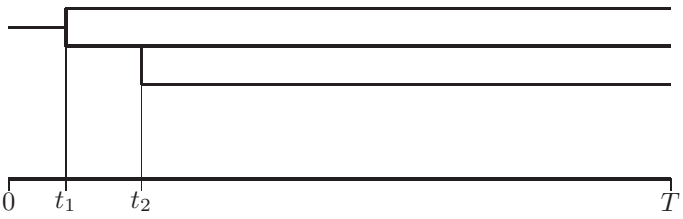


Figure 3: A realization of the reconstructed process without fossils where times of events are reported on the time axis.

Under this model, the probability for a lineage alive at time t not to go extinct before time T, is (Kendall, 1948)

$$\mathrm{P}_{\mathrm{a}}(t) = \frac{\lambda - \mu}{\lambda - \mu \exp(-(\lambda - \mu)(T - t))}$$

Let $Z_{\mathrm{a}}(t)$ be the number of lineages both alive at time t and not extincted before T. The process $(Z_{\mathrm{a}}(t))$ is the nonhomogeneous pure birth process described by

$$Z_{\mathrm{a}}(t+h) = \begin{cases} Z_{\mathrm{a}}(t) + 1 & \text{w.p.} \quad \mathrm{P}_{\mathrm{a}}(t) Z_{\mathrm{a}}(t) \lambda h + o(h) \\ Z_{\mathrm{a}}(t) & \text{w.p.} \quad 1 - \mathrm{P}_{\mathrm{a}}(t) Z_{\mathrm{a}}(t) \lambda h + o(h) \end{cases}$$

where h is an interval of time (Nee et al., 1994).

Under the assumption that the process $(Z(t))$ starts with a single lineage (*i.e.* $Z(0) = 1$), the initial condition of the process $(Z_{\mathrm{a}}(t))$ is

$$Z_{\mathrm{a}}(0) = \begin{cases} 1 & \text{w.p.} \quad \mathrm{P}_{\mathrm{a}}(0) \\ 0 & \text{w.p.} \quad 1 - \mathrm{P}_{\mathrm{a}}(0) \end{cases}$$

Remark that this initial condition is claimed to be $Z_{\mathrm{a}}(0) = 1$ in (Nee et al., 1994), which is in contradiction with the fact that a realization of the complete process $(Z(t))$ may become extinct before the present day.

However, it doesn't change the expression of the probability of waiting more than s between two birth events of the reconstructed process $(Z_{\mathrm{a}}(t))$ which is

$$\mathrm{W}^{\mathrm{a}}_{n,t}(s) = \exp(-n(\lambda - \mu)s) \left(\frac{\lambda - \mu \exp(-(\lambda - \mu)(T - t - s))}{\lambda - \mu \exp(-(\lambda - \mu)(T - t))} \right)^n$$

as well as that of the probability density function of waiting times

$$\mathrm{w}^{\mathrm{a}}_{n,t}(s) = n\lambda(\lambda - \mu) \exp(-n(\lambda - \mu)s) \frac{(\lambda - \mu \exp(-(\lambda - \mu)(T - t - s)))^{n-1}}{(\lambda - \mu \exp(-(\lambda - \mu)(T - t)))^n}$$

both given in (Nee et al., 1994).

From its initial condition, the likelihood of an empty realization of the stochastic process $(Z_{\mathrm{a}}(t))$ (*i.e.* a realization of $(Z(t))$ without lineage alive at T) is $1 - \mathrm{P}_{\mathrm{a}}(0)$. As above, in order to compute the likelihood of a non-empty

realization of $(Z_{\mathrm{a}}(t))$, we sort the times when an event, here only births, occurs in increasing order $t_1 < t_2 < \ldots < t_k$ (Figure 3), and set $t_0 = 0$. Remark that the number of lineages alive at time $t \in (t_i, t_{i+1}]$ and not going extinct before T is equal to $i + 1$. The likelihood of a non-empty realization of $(Z_{\mathrm{a}}(t))$ with k births is

$$\mathrm{P}_{\mathrm{a}}(0) \left(\prod_{i=1}^{k} \mathrm{w}^{\mathrm{a}}_{i,t_{i-1}}(t_i - t_{i-1}) \right) \mathrm{W}^{\mathrm{a}}_{k+1,t_k}(T - t_k) \tag{2}$$

which is not equal to the formula derived in (Nee et al., 1994)

$$k!\lambda^{k-1} \left(\prod_{i=2}^{k} \mathrm{P}_{\mathrm{a}}(t_i) \right) (1 - u_{(T-t_1)})^2 \prod_{i=2}^{k} (1 - u_{(T-t_i)})$$

where

$$u_s = \frac{\lambda(1 - \exp(-(\lambda - \mu)s))}{\lambda - \mu \exp(-(\lambda - \mu)s)}$$

Some calculations show that this last formula can be equivalently written

$$\left(\prod_{i=2}^{k} \mathrm{w}^{\mathrm{a}}_{i,t_{i-1}}(t_i - t_{i-1}) \right) \mathrm{W}^{\mathrm{a}}_{k+1,t_k}(T - t_k) \tag{3}$$

which indicates that it can be interpreted as the probability density of the realization of the reconstructed process, conditioned to the fact that it is not empty and that its first event does occur at time t_1.

This conditioning comes from the fact that Nee et al. (1994) assumed that the time origin of the evolution process was unknown. An alternative way to deal with the time origin uncertainty could be to sum over all the possible times of origin t_0, *i.e.* from $-\infty$ to t_1, accordingly to a suitable prior probability density. We will not go further in this direction, since we make here the assumption that the time origin is known. We compare the performances obtained from likelihoods 2 and 3 respectively (see below).

3 Evaluation protocol

In the simulations designed to test the performance of our method as well as the improvement in accuracy that can be obtained by using the fossil record, we consider four pairs of speciation and extinction rates: $(0.55, 0.45)$, $(0.60, 0.40)$, $(0.67, 0.33)$ and $(0.75, 0.25)$, corresponding to ratios speciation/extinction of about 1.2, 1.5, 2 and 3, respectively, combined with three fossil finds rates: 0.1, 0.5 and 1. Remark that, with these rates, we observe, on average, a fossil discovery each 10, 2 or 1 speciation-extinction event(s), respectively. With these simulation settings, some of the speciation and extinction events are unobservable, even when fossils are considered.

In practice, we first simulate a phylogenetic tree with speciation and extinction rates, and then, we run a Poisson process on it (three times, one for each fossil recovery rate) to simulate the fossil finds. The speciation-extinction process is simulated by iteratively drawing the waiting times between two successive events following exponential distributions (parametrized by the speciation and

extinction rates and the number of lineages alive at that time), then by drawing the event type following the relative probabilities of speciation and extinction (and the lineage on which the event occurs uniformly) and by repeating these steps with the updated number of lineages until the sum of waiting times becomes greater than the evolution time. The fossil finds are next simulated by drawing waiting times, again following suitable exponential distributions, on the tree thus obtained. Figure 5 displays an evolutionary tree simulated with parameters $(\lambda, \mu, \gamma) = (0.67, 0.33, 0.5)$ and $T = 5$ and its reconstructions.

For each speciation and extinction rates and each evolution time (tree depth) from 5 to 10 with a 0.25 step, we simulated 1000 realizations of the stochastic process $(Z(t), Y(t))$, i.e. a total of 252.000 simulations. As in (Paradis, 2004), a simulation leading to a number of contemporary lineages strictly smaller than 3 is discarded (actually, we run the simulation process until getting 1000 artificial phylogenies with at least 3 contemporary lineages). In Figure 4, we displayed the mean numbers of speciations, extinctions and contemporary lineages observed on accepted simulations as well as the corresponding expectations not conditioned to have at least 3 living taxa at the present day. Remark that these plots contain some redundant data since the number of contemporary taxa is nothing but that of speciations plus 1 minus the number of extinctions. These additional data are presented to facilitate interpretation of these graphs.

We compute the maximum likelihood estimations of the speciation and extinction rates in the three following cases:

1. from the complete realization (Figures 1-a and 5-a), which gives us the ideal situation of complete information, unfortunately not available in practice;

2. from its reconstruction from contemporary lineages only (Figures 1-c and 5-c), as in (Paradis, 2004);

3. from its reconstruction from contemporary lineages and fossil finds (Figures 1-b and 5-b), to see how incorporating the fossil record improves our estimates.

In Case 1, the rates are estimated by dividing the observed numbers of speciations and extinctions by the total length of the evolutionary tree (Keiding, 1975). In Case 2, the pair of parameters maximizing the likelihood 2 are numerically computed with a conjugate gradient method. In Case 3, we first determine the fossil find rate by dividing the observed numbers of (intern) fossil finds by the total length of the evolutionary tree. This estimated fossil find rate is next used as a constant in the likelihood (1) to compute numerically the speciation and extinction rates maximizing this likelihood, again with a conjugate gradient method.

We plot the mean absolute errors of these estimations and the corresponding standard deviations (Figure 6 and 7). Estimations from complete realizations or their reconstructions from contemporary lineages only, do not take into account the fossil record thus do not depend on the fossil find rate. This is why each graphic shows a single curve for the absolute error for inferences based on the complete realizations and on extant lineages only, for each pair of speciation and extinction rates, whereas three curves represent various fossil record densities.

Finally, we focus on the reconstruction from contemporary lineages only by plotting means and standard deviations of the absolute error of estimations

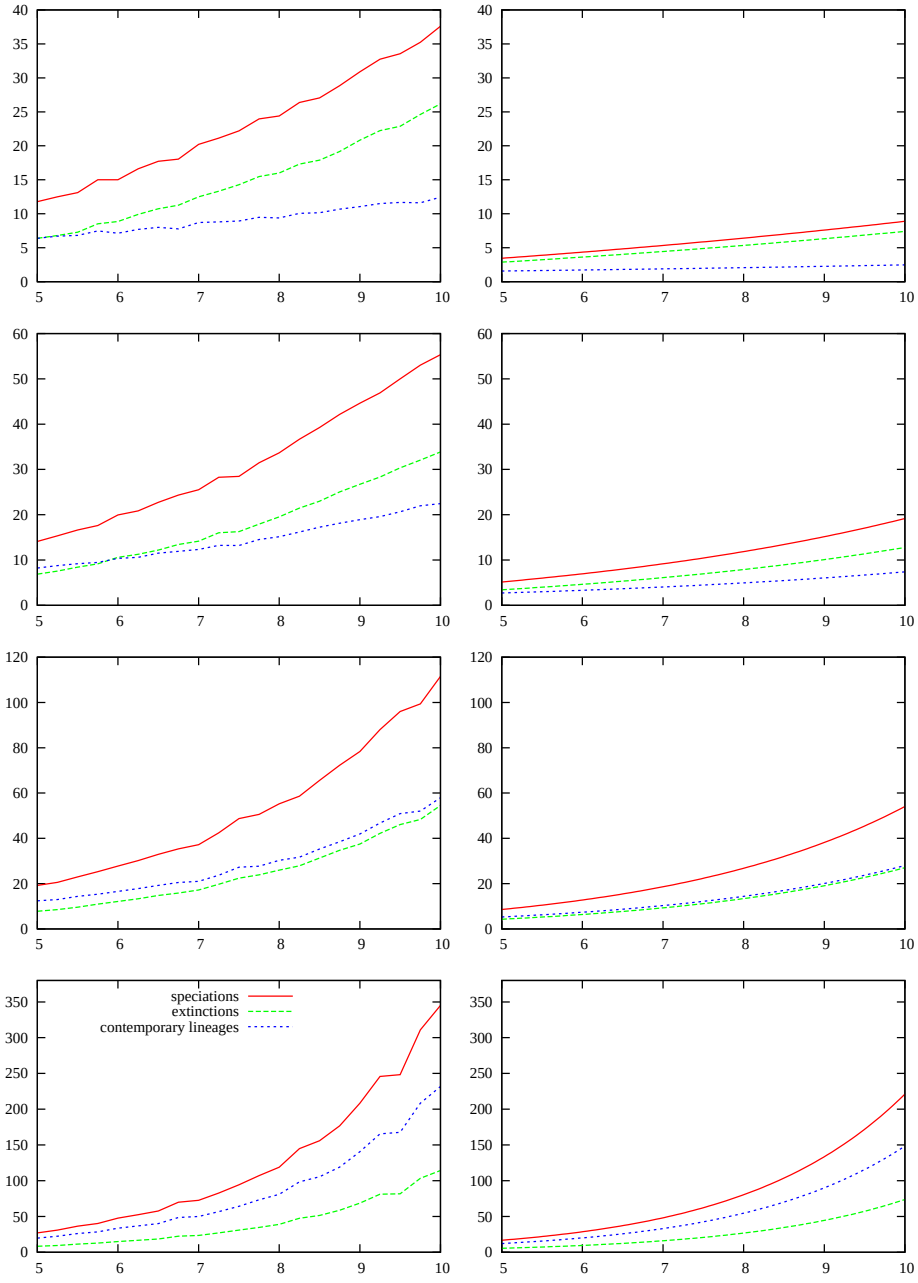
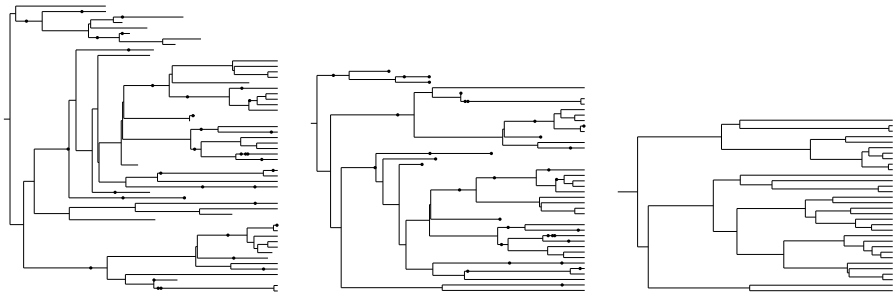


Figure 4: Mean numbers of speciations, extinctions and contemporary lineages observed over accepted simulations, i.e. with at least 3 contemporary taxa (left column) and expectations of the same quantities by taking into account realizations with less than 3 contemporary lineages (right column) *vs* evolution time T. Evolutionary trees are simulated with $(\lambda, \mu) = (0.55, 0.45)$, $(0.60, 0.40)$, $(0.67, 0.33)$ and $(0.75, 0.25)$, from top to bottom.



(a) (b) (c)

Figure 5: (a) an evolutionary tree simulated with parameters $(\lambda, \mu, \gamma) = (0.67, 0.33, 0.5)$ and $T = 5$, (b) its reconstruction from contemporary lineages and fossil finds, (c) its reconstruction from contemporary lineages only. Dots ($\bullet$) represent fossil finds.

obtained from likelihood (2) (referred as "full likelihood") and likelihood (3) (referred as "conditioned likelihood") in Figure 8 and 9.

4 Discussion

Taking into account the fossil record improves the speciation and extinction rate estimations (Figures 6 and 7). The greater the fossil find rate, the better this improvement. Whatever the evolution time T and the parameters (λ, μ) used to simulate the evolutionary trees, estimating the rates by using the fossil record provides significantly more accurate results than by disregarding the fossil record (Student's t-test for paired samples, all p-values $\leq 2.66 \times 10^{-4}$ for speciation rate and all p-values $\leq 9.81 \times 10^{-13}$ for extinction rate). Even the relatively low 0.1 fossil recovery rate yields noticeably lower error rate on speciation and extinction rates than disregarding the fossil record (roughly, half of the error reduction yielded by a fossil recovery rate of 0.5). A Student's t-test for paired sample shows that these differences are significant (p = 0.0003 and 8E-11, respectively). The difference in mean error rate estimates on speciation and extinction rates between fossil recovery rates of 1 and 0.5 is slight, suggesting that the fossil record needs not be nearly complete to obtain near-optimal results.

As previously reported (Paradis, 2004), the mean error on extinction rate is much greater than for speciation rate. However, this patterns is most noticeable for estimates based on extant lineages only. When fossils are incorporated, this difference decreases. Error on extinction rate estimates is approximately proportional to the ratio between actual extinction and speciation rates (Fig. 7). Paradis (2004) similarly reported that estimates of extinction rates were highly biased in a wide range of situations when fossils were not incorporated, and for speciation rates, estimates were biased in the absence of fossils, when the real extinction rate was high, in comparison with the origination rate.

Our new method uses (and requires) more detailed data about the fossil record than previously proposed methods. Thus, it may be relevant to discuss to what extent these requirements may be limiting. We showed that with a fossil find rate of 0.1, mean absolute error on speciation and extinction rates

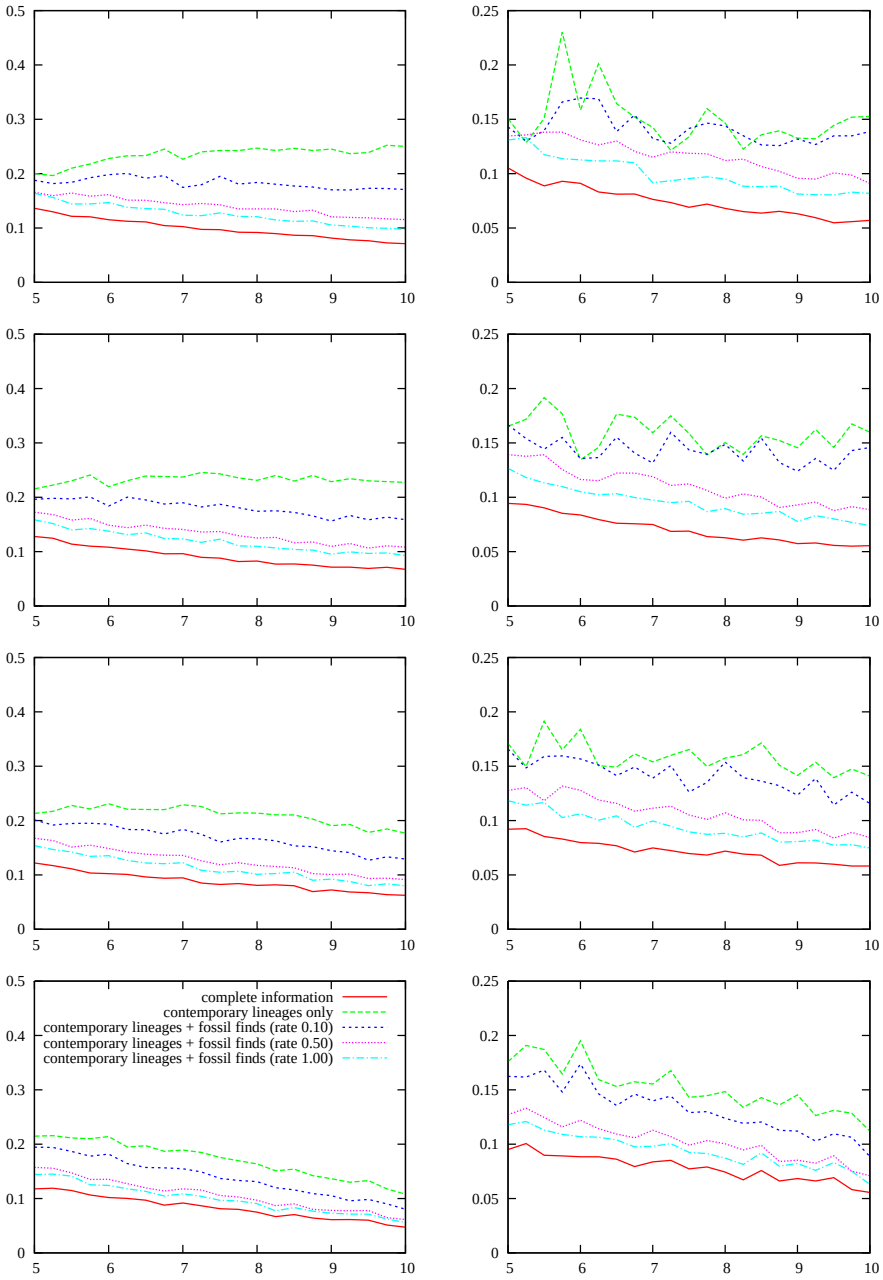


Figure 6: Mean (left column) and standard deviation (right column) of absolute errors of speciation rates estimations *vs* evolution time T. Evolutionary trees are simulated with $(\lambda, \mu) = (0.55, 0.45)$, $(0.60, 0.40)$, $(0.67, 0.33)$ and $(0.75, 0.25)$, from top to bottom.

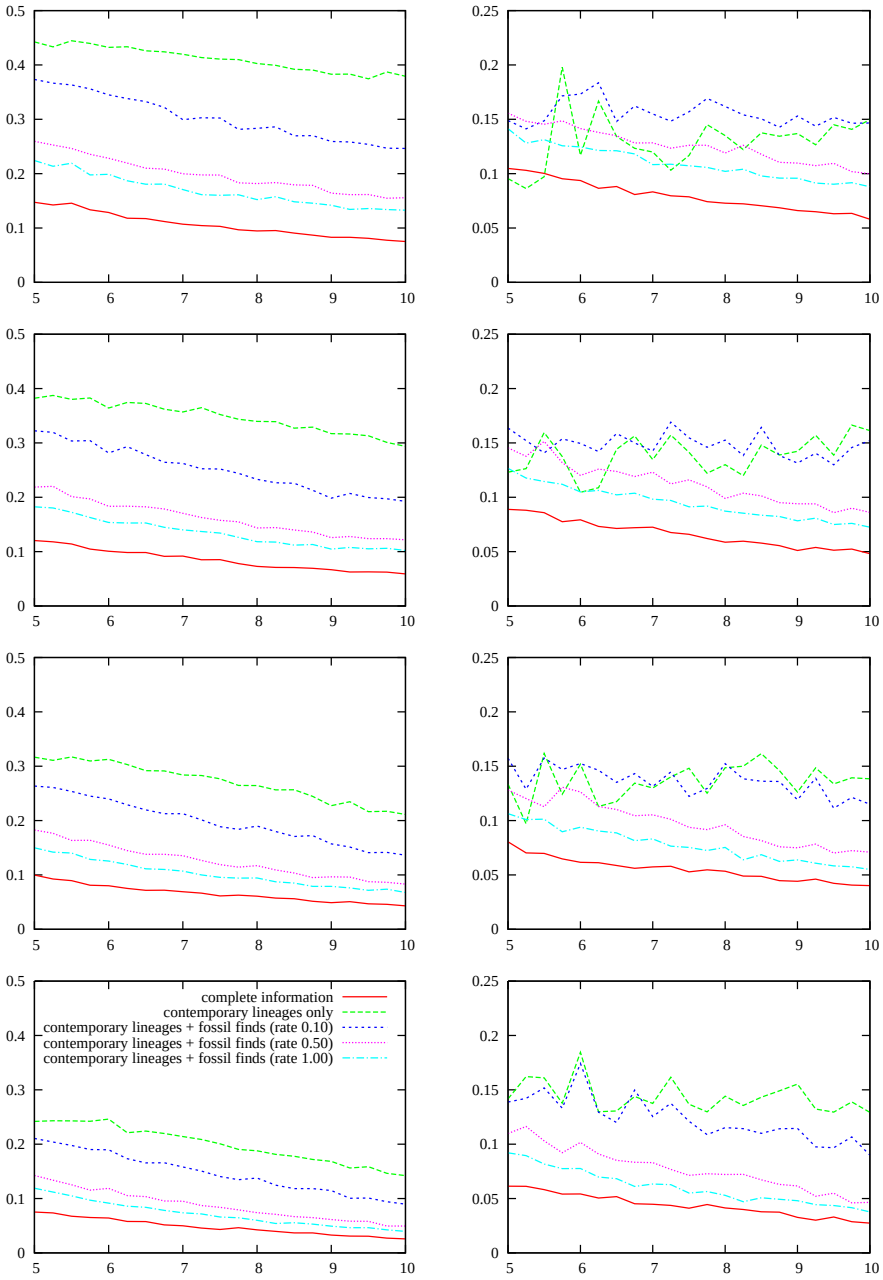


Figure 7: Mean (left column) and standard deviation (right column) of absolute errors of extinction rates estimations *vs* evolution time T. Evolutionary trees are simulated with $(\lambda, \mu) = (0.55, 0.45)$, $(0.60, 0.40)$, $(0.67, 0.33)$ and $(0.75, 0.25)$, from top to bottom.

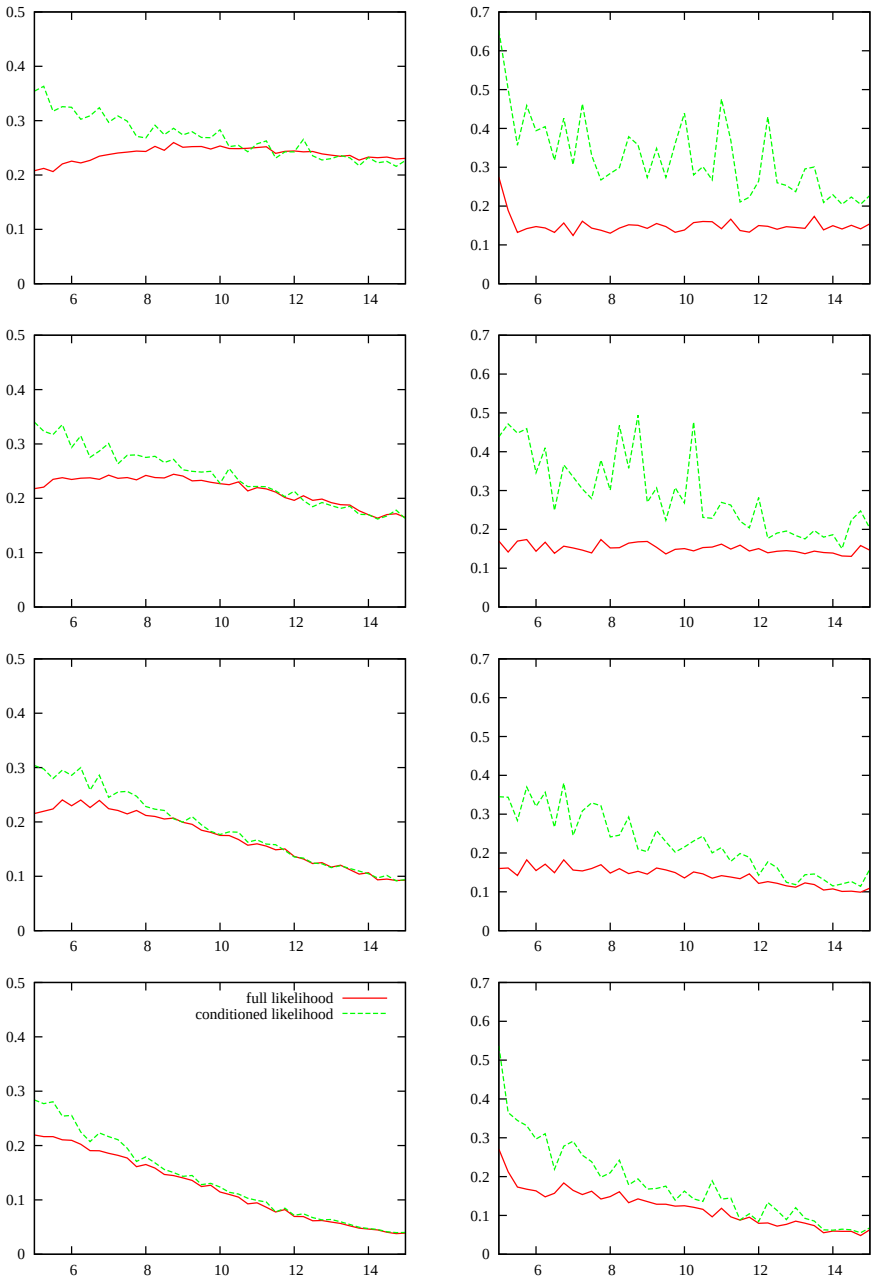


Figure 8: Mean (left column) and standard deviation (right column) of absolute errors of speciation rates estimations from full and conditioned likelihoods *vs* evolution time T. Evolutionary trees are simulated with $(\lambda, \mu) = (0.55, 0.45)$, $(0.60, 0.40)$, $(0.67, 0.33)$ and $(0.75, 0.25)$, from top to bottom.

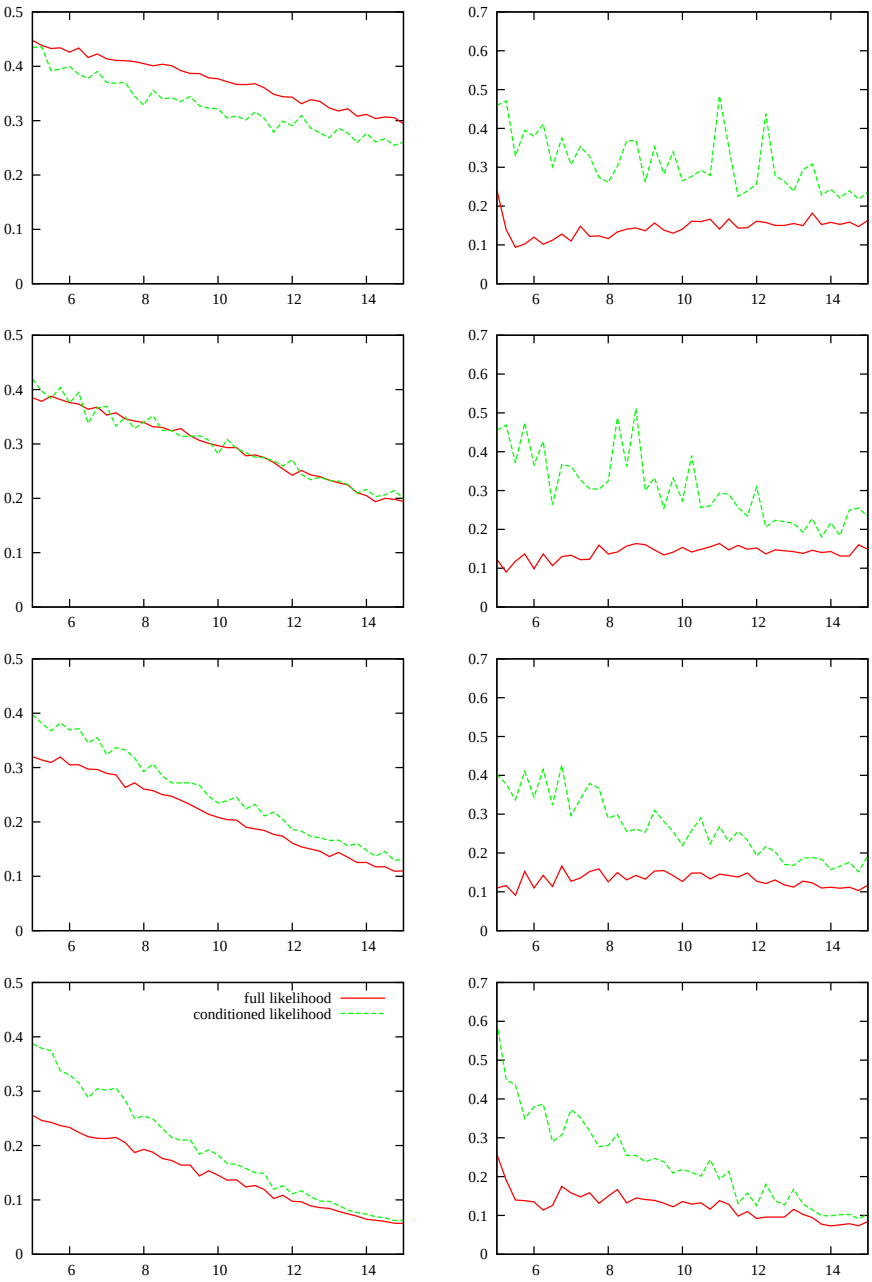


Figure 9: Mean (left column) and standard deviation (right column) of absolute errors of extinction rates estimations from full and conditioned likelihoods *vs* evolution time T. Evolutionary trees are simulated with $(\lambda, \mu) = (0.55, 0.45)$, $(0.60, 0.40)$, $(0.67, 0.33)$ and $(0.75, 0.25)$, from top to bottom.

is reduced by about 20-30% for a wide range of values of the parameters, and with a fossil find rate of 0.5, error is often reduced by about half. These fossil find rates correspond roughly with a species-level completeness of about 20-90% if species are conceptualized as evolutionary lineages between two cladogeneses (de Queiroz, 1998). These values of species-level completeness may seem high compared to the overall proportion of species that is thought to have a fossil record for all life forms and all geological times. That completeness is estimated at about 0.25 to 38% for marine animals, for the Phanerozoic, by Foote (1996, table 1), depending on the assumed diversification model, and at 1-5% by Forey et al. (2004, page 642). However, these figures reflect the fact that large clades (like annelids, nematodes, and priapulids) have left little or no fossil record, usually because they lack mineralized parts. The fossil record of clades of organisms with a well-mineralized skeleton is much better and has been estimated (perhaps a little optimistically) to be 60-90% complete (Foote, 1996, table 2). Thus, even the highest fossil find rate that we used (1, amounting to a nearly complete specific-level fossil record with an average of 1.5-2 fossil per species) is not unrealistic for such clades. With such densely-sampled fossil records, our method should yield much better results than using only extant taxa.

In some paleontological trees, many fossils represent the ancestors of other fossils or of extant organisms. There is currently a controversy about whether or not ancestors can be positively identified in the fossil record, although simulations show that the probability of discovering an actual ancestor in the fossil record is not negligible (Foote, 1996). Until at least the 1960s, most paleontologists actively looked for ancestors in the fossil record and often claimed to have found them (Romer, 1966). The frequent lack of rigor with which such claims were made led (Hennig, 1965) to emphasize that it was very difficult to be sure that an ancestor had been identified because this identification rests partly on negative evidence (the lack of autapomorphies not also found in the presumed descendants). However, other paleontologists have argued that when a fossil is older than its presumed descendants, when it shares some apomorphies that demonstrate close affinities with them, and when no autapomorphies of the prospective ancestor can be found, it is reasonable (more parsimonious in the sense of not requiring the hypothesis that a distinct lineage existed) to consider that it does represent the ancestor (Bonde, 2001). Of course, uncertainty remains, but that is always true of phylogenetic inference, as shown by the proliferation of methods to assess such uncertainties, such as bootstrapping (Felsenstein, 1985), the Bremer index (Bremer, 1988), and more recently, posterior probability obtained from Bayesian methods (Huelsenbeck et al., 2001). In any case, several paleontologists still think that they can recognize ancestors (Clauset and Erwin, 2008; Forey et al., 2004, figs 5, 6), especially when the fossil record is abundant, as is often the case in micropaleontology, and methods to better do this have been developed in the last decades (e.g. Alroy, 1995; Dzik, 2005). Even the most convinced cladists who may be reluctant to consider an extinct species ancestral to another may accept to consider that a given species is represented by series of diachronous populations that are probably ancestral to each other, as often occurs in various levels of a given fossiliferous locality or in several localities of various ages (e.g. Marshall, 1990). For instance, many fossil *Homo sapiens* presumably represent ancestors of present-day humans (e.g. White et al., 2003). Thus, the fact that our method allows (but does not require) fossils to represent ancestors of observed lineages is coherent with current

paleontological practice. Note that our method assumes that the phylogeny is known without error. Incorporating phylogenetic uncertainty would be possible using Bayesian methods, non-parametric bootstrap and other such methods, but this is beyond the scope of this paper.

In the mathematical developments above, we have considered infinitesimally small time units. However, the age of fossils, especially when extracted from compilations (e.g. Benton, 1993; Marjanović and Laurin, 2007), is usually available at the geological period, stage, or other conceptually similar temporal scales (Gradstein and Ogg, 2004), such as the Land Mammal Ages (Wood and Clark, 1941; Evander, 1986). This does not create a serious problem as long as the time subdivisions used area sufficiently fine compared to the envisioned timespan. Thus, if the evolutionary radiation of a 300 Ma-old clade is studied, stage-level resolution (roughly of 5 Ma) is probably sufficient. For clades spanning most of the Cenozoic, the much finer North American Land Mammal Ages scale, with substages of less than 1 Ma (Evander, 1986), is fine enough, and this scale is routinely used in studies that assess the evolution of biodiversity (e.g. Alroy, 2000). Anyway, the precision in which one can estimate the speciation an extinction rates depends on the precision of the age of the relevant fossils, even in the context of a molecular tree of extant taxa, given that such trees a customarily calibrated using the fossil record. Molecular ages of the nodes have their own associated uncertainty, but in the current implementation, only best estimates can be used. Again, further developments could allow incorporating the associated uncertainty.

Our method does not require an exhaustive use of the fossil record. It is sufficient that the fossil data incorporated into the analysis be a representative random sample of the existing data. Thus, not all taxa known from fossils need to be incorporated, which might create problems because the systematic position of fragmentary fossils is often poorly constrained. Thus, if a study incorporated only fossils that can be placed accurately in a phylogeny, this should not be problematic. Similarly, the data entered could represent the number of individual fossil finds on a given branch in a give time interval (a strategy feasible only for small trees or those with a very scanty fossil record), or it could simply be presence/absence data of a fossil record of a given branch on a tree in a given time unit. Provided that the adopted strategy is consistently used throughout the tree, our method should remain valid. Thus, our method could be used for a potentially wide range of taxa with a fossil record.

Our estimation process assumes that we do know the time of origin of the evolution process, which can be assimilated with the age of a branch-based taxon (Cantino and de Queiroz, 2010). In other words, it assumes that we have a date at which a common ancestor of the considered clade originated. This date can be obtained from a fossil discovery or from molecular data, either of which can date the appearance of the stem of the clade. This assumption allows us to use the full likelihood (2) rather than the likelihood (3) conditioned to the the first observed speciation event, to estimate the speciation and extinction rates. Nee et al. (1994) (likelihood 3) thus basically worked on node-based taxa (delimited by the basalmost node linking two extant taxa), whereas our method works on branch-based taxa (Cantino and de Queiroz, 2010). Our results (Figure 8 and 9) show that our procedure generally improves the estimation both in terms of means and standard deviations of the absolute error, except for the mean absolute error of the extinction rate estimation for the pair of parameters $(\lambda, \mu) = (0.55, 0.45)$,

probably for the small number of contemporary taxa expected in this case. As evolution time increases, performance of both estimation methods tends to converge, very fast for the speciation rate, and more slowly for extinction rate.

References

Alroy, J. (1995). Continuous track analysis: a new phylogenetic and biogeographic method. *Systematic Biology*, 44(2):152.

Alroy, J. (2000). New methods for quantifying macroevolutionary patterns and processes. *Paleobiology*, 26(4):707.

Axelrod, D. and Bailey, H. (1968). Cretaceous dinosaur extinction. *Evolution*, pages 595–611.

Benton, M. (1993). *The fossil record 2*. Chapman & Hall. London. GB.

Bininda-Emonds, O., Cardillo, M., Jones, K., MacPhee, R., Beck, R., Grenyer, R., Price, S., Vos, R., Gittleman, J., and Purvis, A. (2007). The delayed rise of present-day mammals. *Nature*, 446(7135):507–512.

Bonde, N. (2001). L'espèce et la dimension du temps. *Biosystema*, (19):29–62.

Bremer, K. (1988). The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. *Evolution*, pages 795–803.

Cantino, P. and de Queiroz, K. (2010). International code of phylogenetic nomenclature. *Version 4c*.

Clauset, A. and Erwin, D. (2008). The evolution and distribution of species body size. *Science*, 321(5887):399.

de Queiroz, K. (1998). The general lineage concept of species, species criteria, and the process of speciation. In Howard, D. and Berlocher, S., editors, *Endless Forms: Species and Speciation*, pages 57–75. Oxford University Press.

Donoghue, M., Doyle, J., Gauthier, J., Kluge, A., and Rowe, T. (1989). The importance of fossils in phylogeny reconstruction. *Annual Review of Ecology and Systematics*, 20:431–460.

Dzik, J. (2005). The chronophyletic approach: stratophenetics facing an incomplete fossil record. *Spec Pap Palaeont*, 73:159–183.

Evander, R. (1986). Formal redefinition of the Hemingfordian–Barstovian land mammal age boundary. *Journal of Vertebrate Paleontology*, 6(4):374–381.

Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, pages 783–791.

Foote, M. (1996). On the probability of ancestors in the fossil record. *Paleobiology*, pages 141–151.

Foote, M. and Raup, D. (1996). Fossil preservation and the stratigraphic ranges of taxa. *Paleobiology*, pages 121–140.

Forey, P., Fortey, R., Kenrick, P., and Smith, A. (2004). Taxonomy and fossils: a critical appraisal. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 359(1444):639–653.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Booth, M., and Rossi, F. (2003). *Gnu Scientific Library: Reference Manual*. Network Theory Ltd.

Gernhard, T. (2008). The conditioned reconstructed process. *Journal of Theoretical Biology*, 253(4):769 – 778.

Gingerich, P. (1979). Paleontology, phylogeny, and classification: an example from the mammalian fossil record. *Systematic Biology*, 28(4):451.

Gradstein, F. and Ogg, J. (2004). Geologic time scale 2004–why, how, and where next! *Lethaia*, 37(2):175–181.

Hallinan, N. (2012). The generalized time variable reconstructed birth-death process. *Journal of Theoretical Biology*, 300(0):265 – 276.

Hennig, W. (1965). Phylogenetic systematics. *Annual Review of Entomology*, 10(1):97–116.

Hone, D., Keesey, T., Pisani, D., and Purvis, A. (2005). Macroevolutionary trends in the Dinosauria: Cope's rule. *Journal of evolutionary biology*, 18(3):587–595.

Huelsenbeck, J. (1991). When are fossils better than extant taxa in phylogenetic analysis? *Systematic Biology*, 40(4):458.

Huelsenbeck, J., Ronquist, F., et al. (2001). Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755.

Keiding, N. (1975). Maximum likelihood estimation in the birth-and-death process. *The Annals of Statistics*, 3(2):363–372.

Kendall, D. (1948). On the generalized "birth-and-death" process. *The annals of mathematical statistics*, 19(1):1–15.

Laurin, M. (2004). The evolution of body size, Cope's rule and the origin of amniotes. *Systematic Biology*, 53(4):594.

Lee, M. (2009). Hidden support from unpromising data sets strongly unites snakes with anguimorph 'lizards'. *Journal of Evolutionary Biology*, 22(6):1308–1316.

Lewin, R. (1983). Extinctions and the History of Life: Now that, for many at least, asteroid impact has been accepted as a causative agent in mass extinction, attention turns to the wider view. *Science (New York, NY)*, 221(4614):935.

Marjanović, D. and Laurin, M. (2007). Fossils, molecules, divergence times, and the origin of lissamphibians. *Systematic biology*, 56(3):369–388.

Marjanović, D. and Laurin, M. (2008). Assessing confidence intervals for stratigraphic ranges of higher taxa: the case of Lissamphibia. *Acta Palaeontologica Polonica*, 53(3):413–432.

Marshall, C. (1990). Confidence intervals on stratigraphic ranges. *Paleobiology*, pages 1–10.

Moore, B. and Donoghue, M. (2009). A Bayesian approach for evaluating the impact of historical events on rates of diversification. *Proceedings of the National Academy of Sciences*, 106(11):4307.

Nee, S., Holmes, E., May, R., and Harvey, P. (1994). Extinction rates can be estimated from molecular phylogenies. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1307):77–82.

Paradis, E. (2004). Can extinction rates be estimated without fossils? *Journal of theoretical biology*, 229(1):19–30.

Pyron, R. (2011). Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. *Systematic biology*, 60(4):466.

Raup, D. and Sepkoski, J. (1984). Periodicity of extinctions in the geologic past. *Proceedings of the National Academy of Sciences*, 81(3):801.

Romer, A. (1966). *Vertebrate paleontology*. University of Chicago press Chicago.

Ruta, M., Pisani, D., Lloyd, G., and Benton, M. (2007). A supertree of temnospondyli: cladogenetic patterns in the most species-rich group of early tetrapods. *Proceedings of the Royal Society B: Biological Sciences*, 274(1629):3087.

Ward, P., Botha, J., Buick, R., De Kock, M., Erwin, D., Garrison, G., Kirschvink, J., and Smith, R. (2005). Abrupt and gradual extinction among late permian land vertebrates in the Karoo Basin, South Africa. *Science*, 307(5710):709.

White, T., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G., Suwa, G., and Clark Howell, F. (2003). Pleistocene homo sapiens from Middle Awash, Ethiopia. *Nature*, 423(6941):742–747.

Wilkinson, R. and Tavaré, S. (2009). Estimating primate divergence times by using conditioned birth-and-death processes. *Theoretical population biology*, 75(4):278–285.

Wood, H. and Clark, J. (1941). Nomenclature and correlation of the North American continental Tertiary. *Geological Society of America Bulletin*, 52(1):1–48.