Advance Access publication September 8, 2012

How Can We Improve Accuracy of Macroevolutionary Rate Estimates?

TANJA STADLER

Institute of Integrative Biology, ETH Zürich, Switzerland;
**Correspondence to be sent to: Institute of Integrative Biology, ETH Zürich, Universitätsstr. 16, 8092 Zürich, Switzerland;*
E-mail: tanja.stadler@env.ethz.ch

Received 19 February 2012; reviews returned 17 April 2012; accepted 10 August 2012
Associate Editor: Luke Harmon

Nee et al. (1994) presented likelihood equations for estimating speciation and extinction rates based on phylogenies of only extant species; in particular their method can infer extinction patterns without extinct species data. Meanwhile, even for the simplest model of speciation and extinction, namely, the constant rate birth–death process, a number of studies have been published using different likelihood equations (Thompson 1975; Rannala and Yang 1996; Yang and Rannala 1997; Gernhard 2008; Stadler 2009). The likelihood functions differ due to conditioning the likelihood on different quantities, like the age of the tree, survival of the tree, or the number of species in the tree.

Which conditionings yield the most accurate speciation and extinction rate estimates? In order to answer this question, I present an overview of 7 likelihood functions (which have been published in previous articles), conditioning on different aspects of the tree. I investigate and discuss the impact of the different conditionings toward accuracy of the maximum-likelihood rate estimates by inferring rates based on simulated phylogenies.

The second part of this article discusses a possible bias in speciation and extinction rate estimates when analyzing incomplete phylogenies, that is, phylogenies in which not all extant species are included. The analytic considerations reveal that we cannot estimate the fraction of nonsampled species, but have to know it, when estimating speciation and extinction rates.

The conclusions reached in this article, assuming the simple constant rate birth–death model, will also apply when assuming the more realistic macroevolutionary models allowing for nonconstant rates (Rabosky 2007; Alfaro et al. 2009; FitzJohn et al. 2009; Morlon et al. 2011; Stadler 2011a; Silvestro et al. 2011; Etienne et al. 2012), as these general models all contain the constant rate birth–death model as a special case. This article ends with contrasting these different method implementations (Table 1) and providing some recommendations for end users in order to facilitate model comparison across packages.

SEVEN TREE LIKELIHOOD FUNCTIONS

I will first define the considered macroevolutionary model, the constant rate birth–death process, and then present seven functions describing the likelihood of a phylogeny on only extant species under this model.

The constant rate birth–death process starts at a time t_0 in the past with one species. At all times, each species gives rise to new species with a constant rate λ and goes extinct with a constant rate μ. After a speciation event, we distinguish between the two descending species (e.g., call them "left" and "right"). The process is stopped after time t_0, the present. Each extant species is sampled with a probability ρ, and we denote the number of sampled extant species by n. The resulting tree where all nonsampled and extinct species are pruned is called the reconstructed tree. The $n-1$ speciation times in a reconstructed tree with n species are $t_1 > t_2 > \cdots > t_{n-1}$, with the present being time 0, meaning a speciation time measures the amount of time before present; Figure 1 shows an example of a reconstructed tree on $n=5$ species. Note that in the reconstructed tree, we distinguish between the "left" and "right" descendant of each branching event, such trees are also called *oriented trees* (Ford et al. 2009). I emphasize that the orientation "left" and "right" is introduced to distinguish all species at all points in time in a convenient way when deriving the likelihood functions (namely each species is characterized by a sequence of "left" and "right" when following edges from the origin to the species). A tree with the extant species being labeled can readily be obtained from the oriented tree by labeling the extant tips and possibly dropping the "left" and "right" labels. As pointed out below, parameter inference will not be influenced by a particular orientation or labeling.

We emphasize that the birth–death process has four parameters t_0, λ, μ, and ρ, and typically λ and μ are the parameters that are being inferred based on a reconstructed phylogeny using the likelihood function, whereas t_0 and ρ are fixed prior to the inference.

TABLE 1. Comparison of different implementations estimating speciation and extinction rates

Software	Function	Parameter setting	Likelihood
R ape	birthdeath		– 2 log(Equation (5)) - 2 log $((n-1)!)$
R laser	CalcLHbd		log(Equation (5)) + log $((n-1)!)$
R geiger	Medusa		log(Equation (5))
R diversitree	make.bisse		log(Equation (5))
R diversitree	make.bd		log(Equation (5)) + log $((n-1)!)$
R DDD	ddd_loglik	Cond=TRUE	log(Equation (5)) + log $((n-1)!)$
R DDD	ddd_loglik	Cond=FALSE	log(Equation (5)) + log $((n-1)!)$
R TreePar	bd.shifts.optim	Survival=1	−log(Equation (5))-$(n-2)$ log 2
R TreePar	bd.shifts.optim	Survival=0	−log(Equation (4))-$(n-2)$ log 2
R code Morlon			log(Equation (2))
Python BayesRates			log(Equation (5)) + log $((n-1)!)$

The expressions in the column "likelihood" specifies the value being returned by the implementation. If using different packages for the same data set, it is essential to make sure to add the appropriate constants such that likelihoods are comparable across packages.

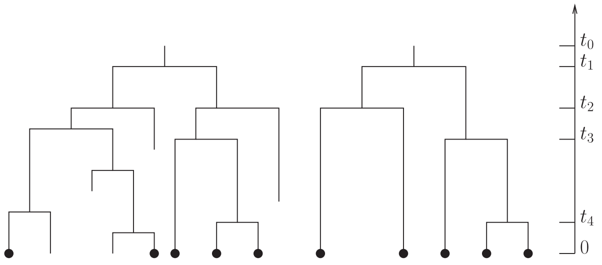

FIGURE 1. Complete phylogeny of age t_0 (left) and corresponding reconstructed phylogeny (right). In the reconstructed phylogeny, all extinct and nonsampled extant species are pruned, only the sampled species (denoted with a black circle) are included. The time t_0 is also called stem age, and the time t_1 is also called crown age.

In order to obtain the likelihood function of a reconstructed phylogeny, we need the following definitions. We write $L(t)=k$ for k species existing at time t, and $L_s(t)=k$ for observing k species at time t in the sampled tree. Furthermore, let $p_n(t)$ be the probability of observing n sampled lineages after sampling in a birth–death tree of age t. Furthermore, let $q(t):=\rho\lambda(1-e^{-(\lambda-\mu)t})/(\lambda\rho+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t})$. We have, from Kendall (1949) for $\rho=1$, Yang and Rannala (1997) for $n=0,1$, and Stadler (2010) for the general case,

$$p_0(t)=1-\frac{\rho(\lambda-\mu)}{\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t}},$$

$$p_1(t)=\frac{\rho(\lambda-\mu)^2e^{-(\lambda-\mu)t}}{(\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t})^2},$$

$$p_n(t)=p_1(t)q(t)^{n-1}.$$

Note that for $\rho=1$ and $\mu>0$, we have $q(t)=\frac{\lambda}{\mu}p_0(t)$.

In the literature, the probability density (which is the normalized likelihood function) of a reconstructed phylogeny is calculated frequently; the probability density is essentially a function of $p_0(t), p_1(t)$, and $q(t)$. However, due to different assumptions, slightly different formulae were obtained for the probability density. In the following, I will present, discuss, and compare the different probability densities and point out where they were originally derived.

We recall that the constant rate birth–death process has 4 parameters t_0, λ, μ, and ρ, and we typically want to estimate λ and μ based on a reconstructed phylogeny. When calculating the probability density of a reconstructed phylogeny, we assume ρ to be a fixed parameter. Seven different possibilities for t_0 are considered:

We can (1) simply condition on the process starting at a time t_0 (i.e., fix the parameter t_0, meaning $L(t_0)=1$). Additionally, we can (2) condition on the process between time t_0 and 0 to survive, that is, to produce at least one sampled species, yielding a stem age t_0 (for this condition, we use the shorthand $t_0=t_{\text{stem}}$, instead of $L(t_0)=1$ and survival). The rationale for this condition is that we typically do not have information on how many clades are unobserved due to extinction of the clade, or nonsampling of the clade. Thus our data analysis being based on an observed reconstructed phylogeny is implicitly conditioned on observing a tree, that is, on survival. Additionally, we can (3) also condition on observing precisely n sampled species ($L_s(0)=n$), the rationale being that we know the number of sampled species n with certainty and thus may want to condition on it.

Often, we do not have knowledge about t_0, the stem age of the tree. However, at time t_1, we observe two lineages, thus analog to above, we can (4) condition on two birth–death processes starting at time t_1 ($L(t_1)=2$). We can (5) also condition on both of these processes surviving to the present, meaning t_1 is the age of the most recent common ancestor of the extant species (also called crown age t_{crown}, we write for this condition $t_1=t_{\text{crown}}$, being short for $L(t_1)=2$ and survival), or we can (6) condition on the number of species being n with $t_{\text{crown}}=t_1$.

Finally, when no knowledge about t_0 is available, the process has alternatively been conditioned (7) on n extant species ($L_s(0)=n$), integrated over all possible t_0 (assuming a uniform prior on $(0,\infty)$ for t_0). The rationale here is that we are certain about the number of sampled species n while we are uncertain about the stem age t_0 (or the crown age t_1). Note that a uniform prior on t_0 is assumed without knowing the number of extant species in the clade. The probability density of t_0 given n sampled extant species in the clade is nonuniform and given in Gernhard (2008, Theorem 3.2). Interestingly, assuming that the clade of consideration and its sister clade have the same speciation and extinction rate yields the same probability for t_0 given n extant species as the uniform prior assumption (Hallinan 2012), and thus may be a biological justification for the uniform prior assumption.

We now state the probability density of a reconstructed phylogeny (with branching times $t_1,\ldots,t_{n-1}$) under the seven conditions:

$$f(\mathcal{T}|L(t_0)=1) = p_1(t_0)\prod_{i=1}^{n-1}\lambda p_1(t_i) \tag{1}$$

$$f(\mathcal{T}|t_0=t_{\text{stem}}) = \frac{p_1(t_0)}{1-p_0(t_0)}\prod_{i=1}^{n-1}\lambda p_1(t_i) \tag{2}$$

$$f(\mathcal{T}|t_0=t_{\text{stem}},L_s(0)=n) = \prod_{i=1}^{n-1}\lambda\frac{p_1(t_i)}{q(t_0)} \tag{3}$$

$$f(\mathcal{T}|L(t_1)=2) = p_1(t_1)^2\prod_{i=2}^{n-1}\lambda p_1(t_i) \tag{4}$$

$$f(\mathcal{T}|t_1=t_{\text{crown}}) = \left(\frac{p_1(t_1)}{1-p_0(t_1)}\right)^2\prod_{i=2}^{n-1}\lambda p_1(t_i) \tag{5}$$

$$f(\mathcal{T}|t_1=t_{\text{crown}},L_s(0)=n) = \frac{1}{(n-1)}\prod_{i=2}^{n-1}\lambda\frac{p_1(t_i)}{q(t_0)} \tag{6}$$

$$f(\mathcal{T}|L_s(0)=n) = n\frac{p_1(t_1)}{1-p_0(t_1)}\prod_{i=1}^{n-1}\lambda p_1(t_i) \tag{7}$$

Equation (1) has been derived in Stadler (2010, Theorem 3.5). Dividing Equation (1) by the probability of the process having sampled species at the present $(1-p_0(t_0))$ leads to Equation (2). Dividing Equation (1) by the probability of observing n extant species at the present $(p_n(t_0))$ yields Equation (3). Equation (4) follows directly from Equation (1) by acknowledging that the two trees descending the two lineages at time t_1 can be calculated using Equation (1). Equation (5) is established equivalent to Equation (2), that is, by dividing with the probability of survival of the two lineages at time t_1. Equation (7) has been established in Stadler (2009) by integrating Equation (3) over t_0. It remains to derive Equation (6).

Derivation of Equation (6): We have,

$$\begin{aligned}
&f(\mathcal{T}|L_s(0)=n,t_1=t_{\text{crown}})\\
&\quad=\frac{f(\mathcal{T},L_s(0)=n,t_1=t_{\text{crown}}|L(t_1)=2)}{p(L_s(0)=n,t_1=t_{\text{crown}}|L(t_1)=2)}\\
&\quad=\frac{f(\mathcal{T}|L(t_1)=2)}{p(L_s(0)=n,t_1=t_{\text{crown}}|L(t_1)=2)}.
\end{aligned}$$

We derive $p(L_s(0)=n,t_1=t_{\text{crown}}|L(t_1)=2)$, the probability of sampling n sampled species with crown age t_1, by calculating the probability for one subtree descending the root at time t_1 yielding i sampled species, and the other subtree yielding $n-1$ sampled species (for all possible i):

$$\begin{aligned}
p(L_s(0)=n,t_1=t_{\text{crown}}|L(t_1)=2) &= \sum_{i=1}^{n-1}p_i(t_1)p_{n-i}(t_1)\\
&=(n-1)p_1(t)^2q(t)^{n-2}.
\end{aligned}$$

This establishes Equation (6).

Remarks:

1. Equation (4) appears in Thompson (1975, p. 58), Equation (3.4.6) for $\rho=1$. The author states that this expression $p_1(t_1)^2\prod_{i=2}^{n-1}\lambda p_1(t_i)$ is the density of a tree conditioned on the most recent common ancestor of the extant species being at time t_1, that is, $f(\mathcal{T}|t_1=t_{\text{crown}})$ (see also Thompson's Figure 3.4). However, Thompson actually calculated $f(\mathcal{T}|L(t_1)=2)$ (Equation (4)).

2. Equation (5) has been derived in Nee et al. (1994, p. 308).

3. Equation (6) is provided in Yang and Rannala (1997). For $\rho=1$, Equation (6) was first derived in Rannala and Yang (1996, Equation (8)). Note that the authors derive Equation (6) by using $f(\mathcal{T}|t_1=t_{\text{crown}})$ from Thompson (1975). As stated in Remark 1, Thompson actually derived $f(\mathcal{T}|L(t_1)=2)$ (and not $f(\mathcal{T}|t_1=t_{\text{crown}})$). Because Rannala and Yang (1996) divide this density by $p(L_s(0)=n,t_1=t_{\text{crown}}|L(t_1)=2)$, the correct density for $f(\mathcal{T}|L_s(0)=n,t_1=t_{\text{crown}})$ is obtained.

4. For a tree with unknown origin t_0, we could estimate t_0,λ,μ (instead of fixing t_0 and estimating only λ,μ) based on a tree with speciation times $t_1,\ldots,t_{n-1}$.

5. For $t_0=t_1$, Equation (2) multiplied by n equals Equation (7). In particular, the maximum-likelihood parameter estimates are the same if t_0 approaches t_1.

6. Equations (1–7) hold for reconstructed trees where the descendants of each branching event are distinguished by "left" and "right" (oriented trees). For obtaining the probability density of

the more commonly used *labeled tree*, that is, the tree where each tip has a unique label but the "left" and "right" is ignored, we need to divide Equations (1–7) by the number of labelings ($n!$) and multiply by the number of "left" and "right" assignments (2^{n-1}). Note, however, that the factor $2^{n-1}/n!$ does not change the likelihood of a tree. Furthermore, under the constant rate birth–death process, it has been established that branching times are independent of the tree shape (Aldous 2001). In fact, for a given vector of branching times, there are $(n-1)!$ different oriented trees, and each of them is equally likely. Thus, for obtaining the probability density of the branching times, we need to multiply Equations (1–7) by $(n-1)!$.

Implementations

The R packages ape, laser, geiger, diversitree, DDD, and TreePar (Paradis et al. 2004; Harmon et al. 2008; FitzJohn et al. 2009; Rabosky 2009; Stadler 2011a; Etienne and Haegeman 2012), as well as R code provided in Morlon et al. (2011) and the Python implementation BayesRates (Silvestro et al. 2011), all estimating macroevolutionary rates under a range of models, allow to condition on survival for most of the implemented models, in particular for all methods assuming the constant rate birth–death process (Table 1). Note in particular that if a method assumes no extinction and complete sampling ($\mu=0, \rho=1$), then conditioning on survival or not yields the same result (as $p_0(t)=0$ for all t).

The R packages diversitree, TreePar, BayesRates, and Morlon's implementation allow for uniform taxon sampling where each extant species has probability ρ of being sampled; DDD allows for uniform taxon sampling where n out of m species are being sampled (see Supplementary Material of Etienne et al. 2012). BayesRates allows for clade-specific sampling probabilities, and diversitree allows for trait-specific sampling probabilities. Medusa in geiger and TreePar can further analyze incomplete phylogenies in which some clades are collapsed to a single tip, and only the number of species represented by that tip is known.

Note that TreePar multiplies each branching event by two in order to allow the likelihood to be compared with general models where branching events are asymmetric such that we need to label the two descendants; for example, the two descendants may be ancestor and offspring, resulting from peripatric speciation (or transmission between hosts if considering virus phylogenies). For the simple constant rate birth–death process, assigning the left descendant "ancestor" and the right descendant "offspring" or vice versa result in the same likelihood of the tree, as the constant rate birth–death process is memoryless, thus the factor 2.

Which Likelihood Function Is The Best?

I simulated 1000 trees on 100 and 1000 extant species using the function sim.bd.taxa in the R package TreeSim (Stadler 2011b), and then reestimated the speciation and extinction rates. I chose $\lambda=1, \rho=1$, and varied the turnover $\mu/\lambda=0.00, 0.05, 0.25, 0.50, 0.75, 0.95$ (note that for different values of λ, the time scale is changed, but the results are the same). When estimating parameters, I fixed $\rho=1$ (because we cannot estimate the sampling fraction, see the section below). The parameter estimates obtained using Equations (1–7) are shown in Figures 2 and 3.

First, I note that simulating trees on a fixed number of species (here 100 and 1000) is performed by choosing any tree with n species uniformly at random. This procedure implicitly assumes a uniform prior on $(0,\infty)$ for the age of the tree (Hartmann et al. 2010), and thus Equation (7) is the appropriate equation for this simulated data.

The seven methods perform similarly well, and in particular for the trees of size 1000, the seven methods produce on average correct estimates with a very small variance across the simulated trees. Empirical trees are typically smaller than size 1000. Based on the simulations for trees with moderate size (100 tips), I observe important differences in performance of the seven likelihood functions, in summary:

1. Is the stem age t_0 important? Equations (1–3) use the information of the stem age t_0, whereas Equations (4–6) neglect it. We observe that the corresponding Equations (1) vs. (4), (2) vs. (5), and (3) vs. (6) produce equivalent results, that is, the additional *stem age does not provide much additional information for rather large trees*. This agrees with our expectation: the stem age is the 100th datapoint, while both methods use 99 datapoints (the 99 branching times), that is, the stem age only adds about 1% information when analyzing trees with 100 species. However, when analyzing empirical phylogenies, some branching times might be estimated less accurate than others, and an accurate stem age estimate might therefore add more than just 1% of information. Thus, whenever an accurate stem age estimate is available, it should be included. This inclusion is in particular relevant for small trees.

2. Shall we condition on survival? Conditioning on survival increases the estimated extinction rates compared with only conditioning on tree age (Equation (1) vs. (2); (4) vs. (5)). This is because we only analyze surviving trees: if we use Equation (1) (respectively, 4), that us, not condition on survival, the method estimates low extinction rates as the method only sees surviving trees and therefore is implicitly informed that no trees went extinct; thus that extinction was low. I conclude that *for high turnover μ/λ, it is very important to condition on survival.* However, for very low extinction rates, the parameter estimates are tighter

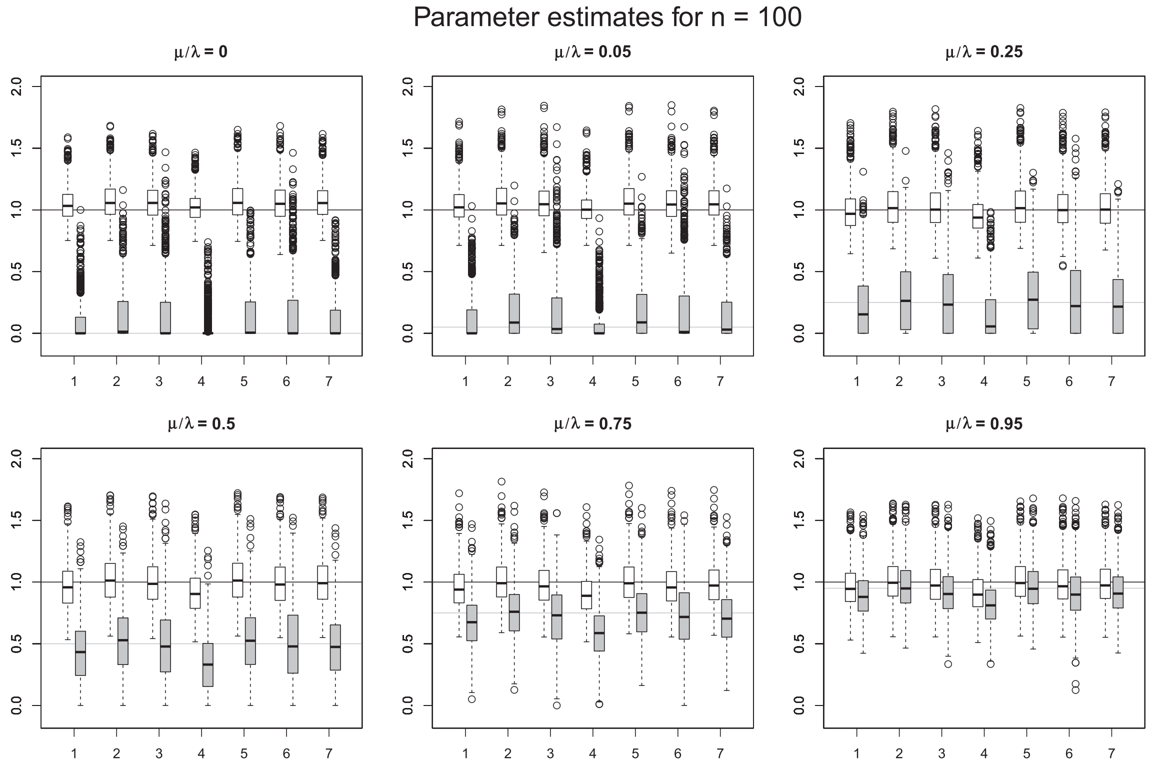

FIGURE 2. Maximum-likelihood speciation rate estimates (white boxplot) and extinction rate estimates (gray boxplot) based on 1000 simulated trees on 100 species (for each parameter combination). The true speciation and extinction rates are indicated by the horizontal lines: speciation rate is set to 1; each panel corresponds to a different extinction rate. In each panel, the estimates from left to right correspond to using Equations (1)–(7).

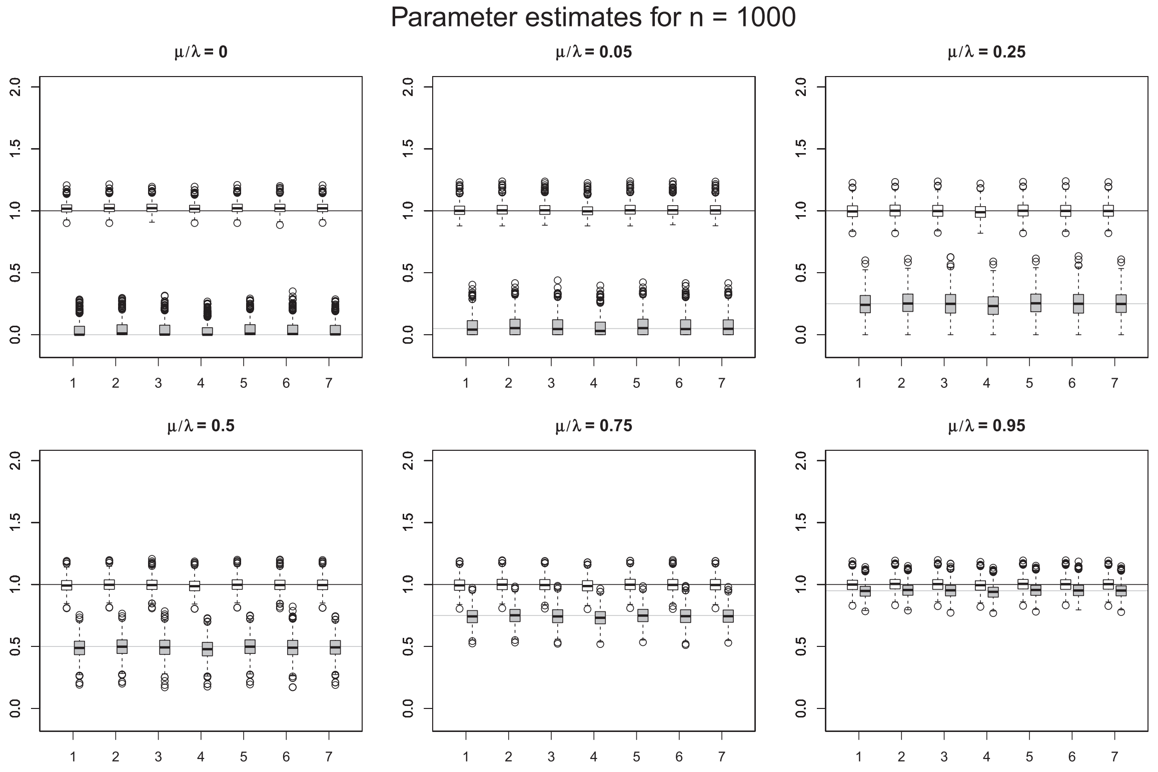

FIGURE 3. Maximum-likelihood speciation rate estimates (white boxplot) and extinction rate estimates (gray boxplot) based on 1000 simulated trees on 1000 species (for each parameter combination). The true speciation and extinction rates are indicated by the horizontal lines: speciation rate is set to 1; each panel corresponds to a different extinction rate. In each panel, the estimates from left to right correspond to using Equations (1)–(7).

using Equation (1) (respectively, 4) rather than Equation (2) (respectively, 5). This is because μ is bounded from below by 0. Now if we condition on survival, also some nonzero μ are likely, and thus the variance in estimated extinction rates becomes wider. As for most clades we expect a high extinction rate based on the fossil record, I suggest to condition on survival in the analyses.

3. Shall we condition additionally on the number of sampled species? If we condition on age and additionally on the number of species n, the variance in parameter estimates becomes larger compared with not conditioning on n. This is expected since by conditioning on n, the parameter estimates are not informed any more by the number of species after time t_0 (respectively, t_1). We expect after time t_0 (respectively, t_1) to observe $n = \mathrm{e}^{(\lambda-\mu)t_0}$ (respectively, $n = 2\mathrm{e}^{(\lambda-\mu)t_1}$) species (for $\rho = 1$), and the estimation method makes use of this information for determining $\lambda - \mu$ as long as we do not condition on n. Thus, I recommend to *not condition on both age and number of sampled species* (i.e., to not use Equation (3) or (6)).

4. Too much conditioning is bad! In general, if we condition on more quantities (e.g., n in addition to age as in the last point), we take away information in the data (by conditioning on it), and thus the estimates become less precise. In the extreme case of conditioning on all speciation times $t_1, \ldots, t_{n-1}$ of the phylogeny, each parameter combination has the same likelihood ($f(\mathcal{T}|t_0, t_1, \ldots, t_{n-1}, n, \lambda, \mu) = 1$), thus no information is left in the data.

5. What is the disadvantage of Equation (7)? I simulated trees on a fixed number n of species, thus Equation (7) is the appropriate equation for our simulation study. We expect Equations (1)–(6), which are slightly violating the simulation study design, to not perform better. *It turns out that Equations (2) and (5) perform as well as Equation (7), meaning these equations are robust toward slight violations of the simulation study assumption. Due to this robustness, I suggest to use Equations (2) and (5).* Furthermore, from a conceptual point of view, people might hesitate using the assumption of a uniform prior for stem age made in Equation (7), and in fact for more complicated models, it might not be possible to integrate over all possible stem ages which is required to obtain Equation (7).

6. We can estimate high turnover better than low extinction! This final conclusion is made based on the observation that the variance in extinction rate estimates decreases for increasing turnover μ/λ. This indicates that for large μ/λ, small changes in μ/λ change the tree distribution significantly, whereas for small μ/λ, small changes in μ/λ barely change the tree distribution.

Can we Estimate Parameters Based on Incomplete Phylogenies?

I performed all simulations using $\rho = 1$. Due to parameter correlations, this actually give us information also for $\rho < 1$:

Let $p_i(t|\lambda, \mu, \rho)$ (respectively, $q(t|\lambda, \mu, \rho)$ be the probability $p_i(t)$ (respectively, $q(t)$) with parameters λ, μ, ρ. Let λ, μ, ρ and λ', μ', ρ' be birth–death parameters. If and only if these parameters fulfil,

$$\lambda - \mu = \lambda' - \mu' \qquad \text{and} \qquad \lambda\rho = \lambda'\rho'.$$

we have for all t,

$$\begin{aligned}(1 - p_0(t|\lambda, \mu, \rho))/\rho &= (1 - p_0(t|\lambda', \mu', \rho'))/\rho',\\ p_1(t|\lambda, \mu, \rho))/\rho &= p_1(t|\lambda', \mu', \rho')/\rho',\\ p_n(t|\lambda, \mu, \rho))/\rho &= p_n(t|\lambda', \mu', \rho')/\rho',\\ q(t|\lambda, \mu, \rho)) &= q(t|\lambda', \mu', \rho').\end{aligned}$$

The "if" can be verified directly by plugging the parameters into the above equations. The "and only if" follows, because: (i) we need $\lambda - \mu$ constant in order to have the exponentials of the functions p_i and q to be invariant for all t and (ii) we need $\lambda\rho$ invariant in order to have invariant factors in front of the exponentials.

Incomplete Sampling in Equations (2),(3),(5)–(7)

Based on this last derivation, all tree densities except of Equations (1) and (4) are functions of only $\lambda - \mu$ and $\lambda\rho$. This means that *we cannot estimate the 3 parameters λ, μ, ρ simultaneously, one parameter has to be fixed.*

Furthermore, instead of simulating or calculating the likelihood for the original parameters λ', μ', ρ' with $\rho' < 1$, we can set $\rho = 1$ and $\lambda = \lambda'\rho', \mu = \mu' - \lambda'(1-\rho')$. Because $\mu \geq 0$, this transformation implicitly assumes $\mu'/\lambda' \geq 1 - \rho'$, see also Stadler and Steel (2012). Thus, we can use previously published rate estimation methods which assume $\rho = 1$ even though for our data we have $\rho' < 1$: we simply transform the estimates λ, μ obtained by assuming $\rho = 1$ into

$$\lambda' = \lambda/\rho', \qquad \mu' = \mu - \lambda(1 - 1/\rho').$$

If we want to allow $\mu'/\lambda' < 1 - \rho'$, we need to use the equations with $\rho' < 1$, as no transformation is available in that case (Stadler and Steel 2012). As an example, I transformed the parameters in Figure 2 using $\rho = 0.5$ (Figure 4). The qualitative pattern of Figure 4 follows the pattern of Figure 2, thus the conclusions drawn in the last section still hold under incomplete sampling.

Incomplete Sampling in Equations (1) and (4)

Because Equations (1) and (4) depend on $\lambda - \mu$, $\lambda\rho$ and λ, we might hope to estimate the three parameters λ, μ, ρ based on Equations (1) and (4). However, this is not possible either: Both Equations (1) and (4) multiplied by λ only depend on $\lambda - \mu$ and $\lambda\rho$. This means that for

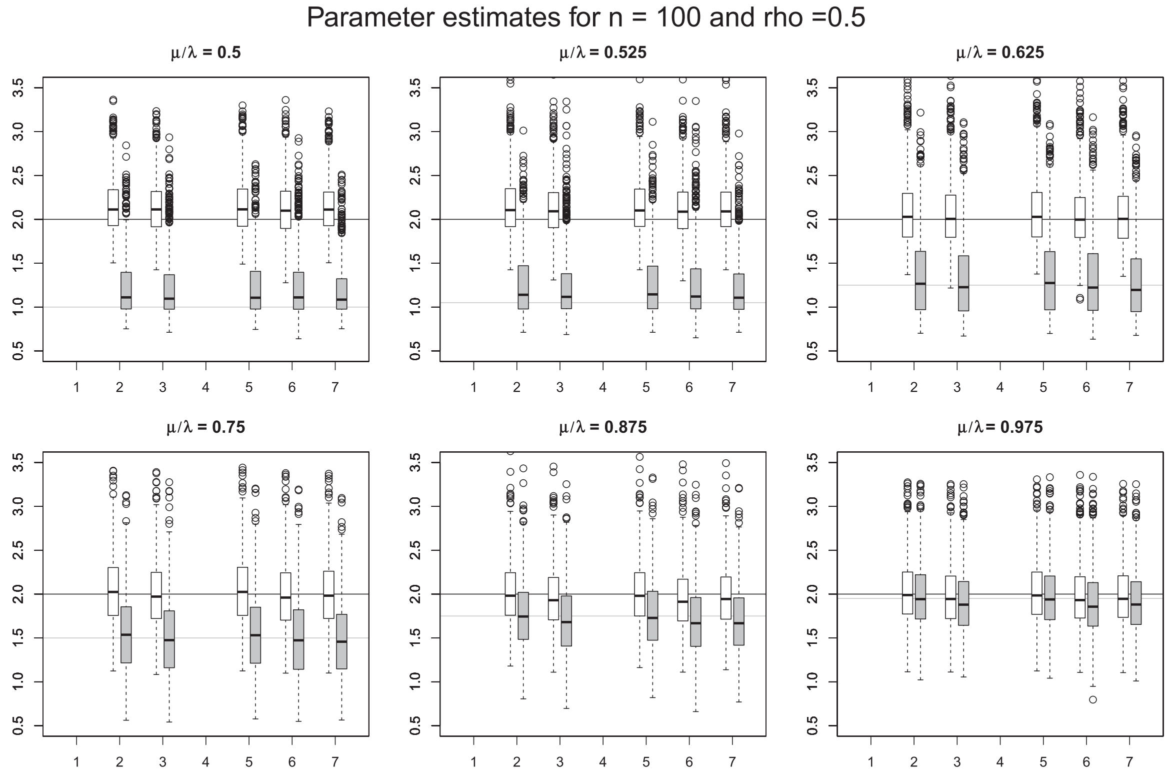

FIGURE 4. Maximum-likelihood speciation rate estimates (white boxplot) and extinction rate estimates (gray boxplot) based on 1000 simulated trees on 100 species (for each parameter combination) with incomplete sampling ($\rho=0.5$). The true speciation and extinction rates are indicated by the horizontal lines; each panel corresponds to a different extinction rate. In each panel, the estimates from left to right correspond to using Equations (2), (3), (5)–(7). Note that we could not use Equations (1) and (4) as we used theory for transforming the estimates in Figure 2 such that we account for $\rho=0.5$; this theory only holds for Equations (2), (3), (5)–(7).

any given λ, μ, ρ, we can increase the tree likelihood by decreasing λ and keeping $\lambda-\mu$ and $\lambda\rho$ constant; thus the parameters $\lambda'=\lambda\rho/\rho'$ and $\mu'=\mu-(1-\rho/\rho')\lambda$ with $\rho'>\rho$ yield a higher tree likelihood than λ, μ, ρ; the speciation rate λ' decreases (and thus the tree likelihood increases) for ρ' increasing. We require $\rho'\leq 1$ as well as $\mu'\geq 0$, which means that our tree likelihood is maximized for $\rho'=\min(1, \frac{\rho\lambda}{\lambda-\mu})$. If we simulate trees under parameters λ, μ, ρ, and then reestimate the speciation, extinction and sampling parameter, we obtain on average biased parameter estimates λ', μ', ρ'; in particular extinction is underestimated:

$$\lambda'=\max(\lambda-\mu, \rho\lambda), \qquad \mu'=\max(0, \mu-(1-\rho)\lambda),$$
$$\rho'=\min\left(1, \frac{\rho\lambda}{\lambda-\mu}\right).$$

This means that only two out of three parameters can be estimated (ρ' will always be chosen as big as possible). Always obtaining biased estimates might seem paradoxical. However, the reason for overestimating sampling and underestimating speciation as well as extinction is simple: The tree densities in Equation (1) (respectively, (4)) describe the probability density of the process after time t_0 (respectively, t_1). If we were to simulate many trees of age t_0 (respectively, t_1) and estimate the parameters based on the collection of trees, including the extinct trees, we would obtain nonbiased estimates. However, by only estimating parameters based on nonextinct trees, we bias our estimates toward low extinction rates.

CONCLUSIONS

I conclude by summarizing the main results of this article. The macroevolutionary model considered in this article, the constant rate birth–death process, has 4 parameters, the speciation rate λ, the extinction rate μ, the sampling probability ρ, and the stem age of the tree (t_0). Instead of stem age, we can also consider crown age as a parameter; crown age is the stem age of the two clades subtending the first branching event in the reconstructed tree.

First, the sampling probability ρ cannot be estimated together with λ and μ based on a reconstructed phylogeny. The likelihood function can only inform us about two parameters while we have to fix the third. Thus, one answer to the question posed in the title of the article, "How can we improve accuracy of macroevolutionary rate estimates?", is by improving a priori knowledge about the sampling probability ρ.

Second, the stem age/crown age of the tree is typically fixed to the observed value when estimating speciation and extinction rates. When we estimate rates based on reconstructed phylogenies of moderate size (here 100 tips), we should condition the likelihood function on survival (Equation (2) or (5)), in order to improve accuracy of macroevolutionary rate estimates. If not conditioning on survival, we underestimate extinction. The bias of underestimating extinction would disappear if we were to include knowledge about the fraction of clades having gone extinct; however, this information is typically not available.

As an alternative, conditioning the likelihood function on the number of species but averaging over all possible stem ages (Equation (7)) is appropriate, provided that a uniform prior for the time of origin seems plausible.

We should not combine the two scenarios though: if we condition on the number of species in the tree in addition to fixing the tree age (Equation (3) or (6), the estimates become less accurate, as we take away information from the data (by conditioning on it)!

For large phylogenies (here 1000 tips), the different conditionings performed equally well.

In general, the simulations revealed that given the analyzed trees evolved under a constant rate birth–death process, the extinction rate can be estimated accurately, in particular if the turnover (extinction/speciation) is large. I want to emphasize though that under more complex models with varying extinction rates, turnover becomes increasingly difficult to estimate while diversification rate (speciation–extinction) can be obtained reliably, see for example, the simulation results of Stadler (2011a). An additional difficulty in estimating turnover arises as turnover estimates are very sensitive toward model misspecification: Rabosky (2010) simulated phylogenies (with 50 tips) under a model with constant turnover while speciation rates varied in a heritable fashion across lineages; fitting a constant rate birth–death model to these trees yielded bimodal turnover estimates, both a turnover of zero and one was supported most. Furthermore, many empirical data sets reveal an extinction rate estimate of zero while the fossil record clearly shows patterns of extinction (Purvis 2008). Thus, a future challenge will be to develop methodology which uses phylogenetic trees as well as fossil data in order to obtain more reliable turnover estimates.

In this article, accuracy and precision of parameter estimates were investigated by simulating a set of trees, and then inspecting the distribution of maximum likelihood parameter estimates; the obtained confidence intervals are parametric bootstrap confidence intervals. For empirical trees, a set of posterior trees or bootstrap trees can be analyzed analogously. It is important to note that the obtained parameter intervals reflect the sensitivity of the maximum likelihood estimates toward changes in the phylogeny.

In order to get confidence regions reflecting the shape of the likelihood function, the contour plot of the likelihood has to be considered. The contour plot displays the region in parameter space where the likelihood is at most x units away from the maximum likelihood. The value x depends on the number of parameters and the chosen confidence region (95%, 99%, etc.). Alternatively, the likelihood functions can be used in an MCMC analysis (instead of a maximum likelihood analysis as done in this article) yielding credible intervals for each parameter. I strongly recommend to analyze the posterior trees or the bootstrap trees directly, rather than the summarized maximum clade credible tree or the consensus tree, as these summary trees do not reflect branch lengths very well (e.g., branch lengths may be negative).

When testing the simple constant rate birth–death process against a more complex model, it is important that the two likelihoods are comparable, meaning that they were obtained using the same conditionings and normalizing constants. For example, if the probability of extinction of the clade ($p_0(t_1)$) is 0.5, then the likelihood obtained from Equation (5) is larger than that obtained from Equation (4) by a factor of 4. If $p_0(t_1)$ is 0.9, then the factor increases even to 100. As a consequence, if using Equation (5) for the constant rate birth–death model while using the analog of Equation (4) for the complex model, the likelihood ratio test is too conservative in rejecting the constant rate birth–death process. Also, likelihoods do not have to be normalized; the different available packages use different normalization constants. Table 1 summarizes which conditionings and normalizing constants are used in the available packages, which will hopefully facilitate comparison of models implemented in different packages. Any novel packages should explicitly state which assumptions are made and normalizing constants are used to facilitate comparison with the existing packages.

I want to conclude with some recommendations to end users. Whenever comparing models across packages, I suggest that end users verify that, under the simple constant rate birth–death model, the same maximum-likelihood value and parameter estimates are obtained with any of the packages used. For the packages discussed here, normalizing constants have to be added to the likelihoods as listed in Table 1. In order to obtain unbiased extinction rate estimates, I suggest using the option "condition on survival" if it is available in all of the considered packages.

Funding

This work was supported by the Swiss National Science Foundation.

Acknowledgements

I thank the editor, the associate editor, Matt Pennell, Rampal Etienne, and an anonymous reviewer for very valuable comments on the article.

References

Aldous D.J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Statist. Sci. 16:23–34.

Alfaro M., Santini F., Brock C., Alamillo H., Dornburg A., Rabosky D., Carnevale G., and Harmon L. 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. Proc. Natl. Acad. Sci. USA 106:13410.

Etienne R.S., Haegeman B. 2012. DDD: Diversity-dependent diversification. Available from: URL http://cran.r-project.org/web/packages/DDD/index.html.

Etienne R.S., Haegeman B., Stadler T., Aze T., Pearson P., Purvis A., Phillimore A. 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proc. R. Soc. Ser. B 279:1300–1309.

FitzJohn R., Maddison W., Otto S. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595.

Ford D., Matsen E., Stadler T. 2009. A method for investigating relative timing information on phylogenetic trees. Syst. Biol. 58: 167–183.

Gernhard T. 2008. The conditioned reconstructed process. J. Theo. Biol. 253:769–778.

Hallinan N. 2012. The generalized time variable reconstructed birth–death process. J. Theor. Biol. 300:265–276.

Harmon L., Weir J., Brock C., Glor R., Challenger W. 2008. Geiger: investigating evolutionary radiations, Bioinformatics 24: 129–131.

Hartmann K., Wong D., Stadler T. 2010. Sampling trees from evolutionary models. Syst. Biol. 59:465–476.

Kendall D.G. 1949. Stochastic processes and population growth. J. R. Statist. Soc. Ser. B. 11:230–264.

Morlon H., Parsons T., Plotkin J. 2011. Reconciling molecular phylogenies with the fossil record. Proc. Natl. Acad. Sci. USA 108:16327–16332.

Nee S.C., May R.M., Harvey P. 1994. The reconstructed evolutionary process. Philos. Trans. R. Soc. Lond. Ser. B 344: 305–311.

Paradis E., Claude J., Strimmer K. 2004. Ape: Analyses of phylogenetics and evolution in R language. Bioinformatics 20: 289–290.

Purvis A. 2008. Phylogenetic approaches to the study of extinction. Annu. Rev. Ecol., Evol. Syst. 39:301–319.

Rabosky D. 2007. Likelihood methods for detecting temporal shifts in diversification rates. Evolution 60:1152–1164.

Rabosky D. 2009. laser: Likelihood analysis of speciation/extinction rates from phylogenies. Available from: URL http://cran.r-project.org/web/packages/laser/index.html.

Rabosky D. 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution 64:1816–1824.

Rannala B., Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. J. Mol. Evol. 43:304–311.

Silvestro D., Schnitzler J., Zizka G. 2011. A bayesian framework to estimate diversification rates and their variation through time and space. BMC Evol. Biol. 11:311.

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.

Stadler T. 2011a. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. USA 108:6187–6192.

Stadler T. 2011b. Simulating trees with a fixed number of extant species. Syst. Biol. 60:676–684.

Stadler T., Steel M. 2012. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. J. Theor. Biol. 297:33–40.

Thompson E.A. 1975. Human evolutionary trees. Cambridge University Press.

Yang Z., Rannala B. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov Chain Monte Carlo Method. Mol. Biol. Evol. 17:717–724.