JOURNAL OF Evolutionary Biology

doi: 10.1111/jeb.13233

SHORT COMMUNICATION

Robustness of the approximate likelihood of the protracted speciation model

C. SIMONET, R. SCHERRER, A. REGO-COSTA & R. S. ETIENNE

Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, The Netherlands

Keywords:

birth-death model;
diversification;
likelihood;
macroevolution;
phylogeny.

Abstract

The protracted speciation model presents a realistic and parsimonious explanation for the observed slowdown in lineage accumulation through time, by accounting for the fact that speciation takes time. A method to compute the likelihood for this model given a phylogeny is available and allows estimation of its parameters (rate of initiation of speciation, rate of completion of speciation and extinction rate) and statistical comparison of this model to other proposed models of diversification. However, this likelihood computation method makes an approximation of the protracted speciation model to be mathematically tractable: it sometimes counts fewer species than one would do from a biological perspective. This approximation may have large consequences for likelihood-based inferences: it may render any conclusions based on this method completely irrelevant. Here, we study to what extent this approximation affects parameter estimations. We simulated phylogenies from which we reconstructed the tree of extant species according to the original, biologically meaningful protracted speciation model and according to the approximation. We then compared the resulting parameter estimates. We found that the differences were larger for high values of extinction rates and small values of speciation-completion rates. Indeed, a long speciation-completion time and a high extinction rate promote the appearance of cases to which the approximation applies. However, surprisingly, the deviation introduced is largely negligible over the parameter space explored, suggesting that this approximate likelihood can be applied reliably in practice to estimate biologically relevant parameters under the original protracted speciation model.

Introduction

A widely observed pattern in empirically reconstructed phylogenies is the slowdown of lineages accumulation through time (McPeek, 2008; Phillimore & Price, 2008). This pattern has been explained by models of diversity-dependent diversification in which the speciation rate declines as species accumulate (Rabosky & Lovette, 2008; Etienne & Rosindell, 2012). There are, however, alternative explanations (Morlon, 2014). One of these is the fact that speciation takes time. Avise *et al.* (1998) already showed that speciation is not an instantaneous process but takes at least 2 million years (My) to complete in various vertebrate clades. Purvis *et al.* (2009) argued that sufficient time must pass for two lineages to 'attract taxonomic attention', that is to be recognized as distinct species.

This protraction in the speciation process is likely to affect species recognition at present and, as a consequence, to modify the resulting reconstructed tree. It has been explicitly implemented in the protracted birth–death model (Etienne & Rosindell, 2012), a generalization of the birth–death process originally introduced by Kendall (1948). Newly arising lineages are regarded as incipient species that will take some time to complete speciation and be regarded as good species. During this time, these incipient species can still

Correspondence: Rampal S. Etienne, Groningen Institute for Evolutionary Life Sciences, University of Groningen, Box 11103, Groningen 9700CC, The Netherlands.
Tel.: +31 50 363 2230; fax: +31 50 363 5205; e-mail: r.s.etienne@rug.nl

become extinct or give rise to new species. Thus, a species identified by a taxonomist comprises a set of related lineages that cannot yet be distinguished, and therefore, the number of species recognized at the present is smaller than the number of actual independent lineages, explaining the slowdown observed in the accumulation of lineages through time (Etienne & Rosindell, 2012). Because protraction seems to be a universal feature of the speciation process, it challenges other more complex explanations for the slowdown in the accumulation of species (Moen & Morlon, 2014), notably diversity-dependent diversification (Etienne *et al.*, 2012).

An approximate likelihood method to estimate the protracted speciation model parameters from the branching times of a phylogenetic tree has been developed by Lambert *et al.* (2015), based on the mathematical theory of coalescent point processes, which provides tools for modelling branching processes (Lambert, 2010; Lambert & Stadler, 2013). From here on, we will refer to this approximation as the LME approximation, referring to the authors of this likelihood (Lambert, Morlon and Etienne). The mathematical derivation of this likelihood requires an approximation of the model which biologically is not entirely satisfactory (Etienne *et al.*, 2014). In short, the approximation often counts fewer species in a tree than what we would conclude biologically (see Methods for more details).

This approximate likelihood has been shown to provide accurate estimations of the model parameters, when using data simulated under the LME approximation, that is simulated trees were modified according to this approximation, and thus often showed fewer species than would be found in an actual reconstructed tree (Etienne *et al.*, 2014). Tree size is known to have a large effect on parameter estimates. This can be understood intuitively as follows. For the pure-birth process (no extinction, just instantaneous speciation), the maximum likelihood estimate of the speciation rate is $(n-2)/s$ where n is the number of tips in the tree and s is the sum of all branch lengths (Nee, 2001). Adding tips with short branch lengths will affect the numerator considerably but hardly affect the denominator, and hence, the estimate will be strongly dependent on tree size. Therefore, given that tree size is not measured correctly in the approximate likelihood, applications of this approximate likelihood on empirical data may lead to systematic biases: for example, it may point to fast completion of speciation in cases where it is actually much slower. Hence, to check for robustness of the method, we need to know whether this approximate likelihood still leads to reliable parameter estimates, when data are generated with the original, biologically relevant model, that is without applying the LME approximation. Here, we assess this robustness.

The most precise way to address this would be to compare the results obtained from an exact likelihood (i.e. making no approximation of the model) with results obtained from this approximate likelihood. However, because an exact likelihood seems unfeasible for this model (Lambert *et al.*, 2015), we adopted a simulation approach. We simulated reconstructed phylogenies under the protracted speciation model, with and without the LME approximation, and compared estimates of the parameters obtained by maximizing the approximate likelihood. The difference between these estimates, and their deviation from the true values used to generate the data, will tell us under which conditions the approximation made by this likelihood introduces a large deviation from the original model, and thus does not provide reliable parameter estimates.

We expected the LME approximation to introduce larger deviations for increasing values of the extinction rate and decreasing values of the speciation-completion rate, where the LME approximation is the most noticeable on the reconstructed trees (see Methods). Our study generally confirms these expectations. However, to our surprise, we found that in most of parameter space, the deviation is so small that it can be safely ignored.

Materials and methods

Outline of the model

The protracted speciation model is a birth–death model in which incipient species are formed at birth events. Both good and incipient species can give rise to new incipient species at rates b_1 and b_2 and can become extinct at rates μ_1 and μ_2, respectively. These incipient species may become good species after a speciation-completion event. Such events happen at rate λ. At the present, all incipient lineages that are not separated by a speciation-completion event are considered to belong to the same species. The speciation-completion events separate different species, but the divergence times are recorded from the speciation-initiation events. For more details and illustrations of the protracted speciation model, we refer to Etienne *et al.* (2014) and Etienne & Rosindell (2012). In line with Etienne *et al.* (2014) who previously investigated the robustness of the approximate likelihood, we used the time-homogeneous Markov version of the model, that is where parameters are independent of time. All topologies are equally likely, and the branching times contain all the information relevant for estimating the model parameters (Lambert & Stadler, 2013). Figure 1a,b shows an example of a tree produced by the protracted speciation model with and without extinct tips.

The LME approximation

Biologically, an incipient species that is alive at present and that has a good but extinct parent species, should

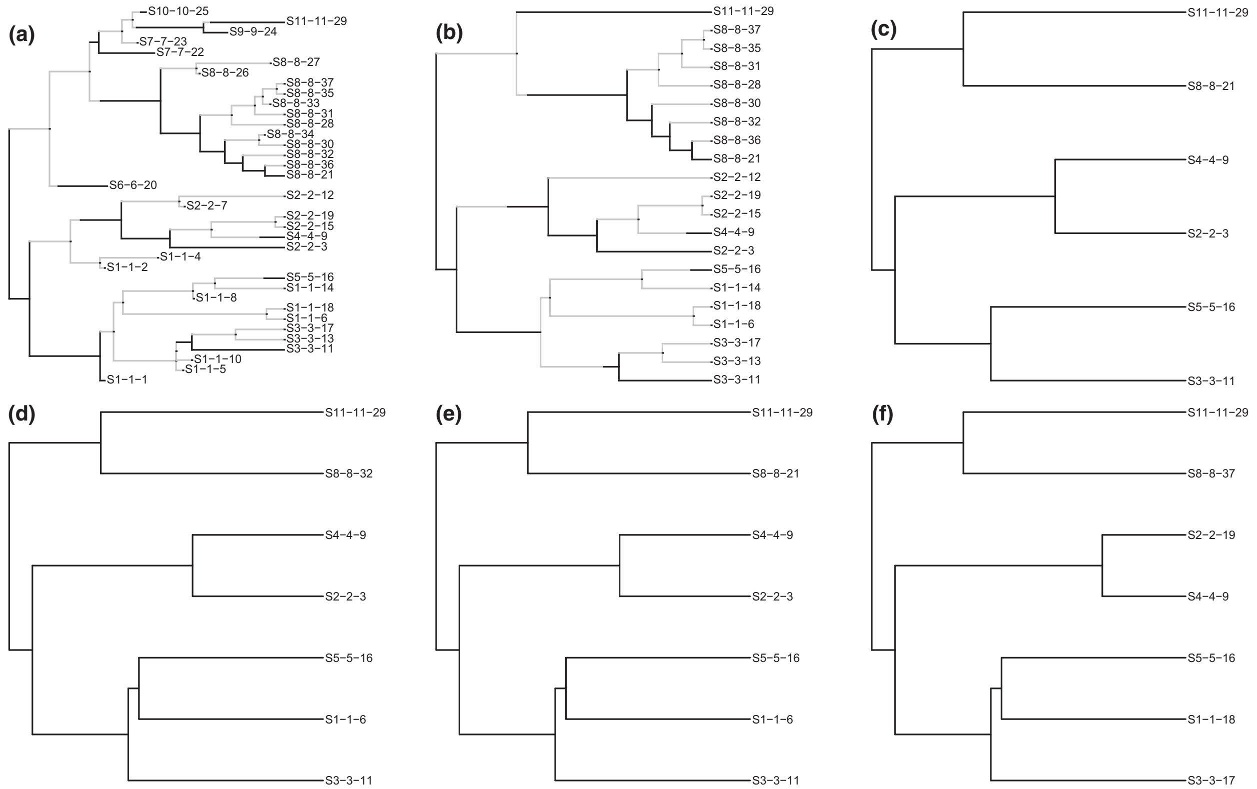

Fig. 1 An example of a phylogenetic tree resulting from the protracted speciation model, simulated by the function pbd_sim of the R package PBD. Note that the trees are oriented such that the mother lineages are at the bottom (see for instance the position of S4-4-9 in panels e and f). (a) Full tree showing extinct and extant species, and incipient (grey) and good (black) species. The tip labels are SX-X-Y, where X-X stands for the species label and Y for the incipient (or sub-)species label. The order in Y denotes the order of appearance; for example, S1-1-18 was formed later than S1-1-6 (in this case it is a daughter of S1-1-6). (b) Same tree as a, but with all extinct species pruned. (c). Species tree resulting from applying the LME approximation to the tree in b. (d). Species tree resulting from sampling one incipient species per species at random in the tree in b. (e) Species tree resulting from choosing the oldest incipient species per species in the tree in b. (f) Species tree resulting from choosing the youngest incipient species per species in the tree in b. Note that the three species trees in the bottom row have one more species than the tree in c. This is an instance of the LME approximation resulting in a biologically unrealistic tree. The ancestral species S1-1-1 has become extinct (see a), but leaves several incipient descendants (S1-1-6, S1-1-14, and S1-1-18) which are representative of species S1-1, and thus, one of them should be included in the total species count. The LME approximation will only count an incipient descendant if it is the oldest extant descendant. However, the oldest extant descendant is S2-2-3 which has already become a different species, and hence, the tree in c has only six species whereas the trees in d-f have seven species. In d, S1-1-6 is randomly sampled to represent species S1-1, and in e and f, the oldest (S1-1-6) and youngest (S1-1-18) are chosen to represent species S1-1. The effect of sampling is not only that different incipient species are being chosen as representatives of the species without changing the shape of the tree; sometimes sampling may also result in a difference in node depths. For example, the node connecting S2-2-3 to S4-4-9 in e is slightly older than the node connecting S2-2-19 to S4-4-9 in f. This can also be seen in b. We note that all branching events are bifurcating. Whenever it seems that three branches appear from a node (e.g. S1-1-5, S1-1-10 and the clade of species S3-3 in a), these are two sequential branching events that rapidly follow each other. Similarly, the completion event leading to the clade of species S3-3 happens very soon after the initiation event and hence almost the entire branch appears black.

be considered as good, because it is ‘representative’ of its ancestor. For mathematical convenience, it would be easier to consider a species as good only if it has been through a speciation-completion event, but this ignores all cases of such representative species. An intermediate solution was developed by Lambert *et al.* (2015) it takes into account representative species, but only if it is the first descendant of the good but extinct parent species. If the incipient species is a younger descendant, then it is not recognized as good, and hence, the LME approximation counts fewer species than there really are. This is illustrated in Figure 1. All species S1-1-x alive at present are representative of their good but extinct parent S1-1-1 (Figure 1a,b). The reconstructed tree of extant species resulting from applying the LME approximation in Figure 1c ignores this case of representative species. By contrast, Figure 1d–f takes this case into account (the only difference between the trees in Figure 1d–f is how incipient species are sampled, see below), and thus, one more tip is present in these trees. As the number of cases of representative species increases, the tree resulting from applying the LME approximation and the 'true' tree will be increasingly different. The difference between these trees will be nonzero if three criteria are met: (i) a good parent must have become extinct, (ii) this parent must have left multiple extant incipient daughter lineages, and (iii) the oldest daughter species must have become good. A necessary condition for the first criterion to be met is a nonzero extinction rate. When extinction is zero, the trees will be identical. The second criterion will be met more often when the speciation-initiation rate is high and the speciation-completion rate is low. The third criterion requires the speciation-completion rate not to be too low, as there would not be any good descendant lineages. Hence, we expected that the LME approximation effect would increase with increasing extinction rate and reach an optimum with respect to the speciation-completion rate.

Apart from studying the effect of the LME approximation, we also studied the effect of sampling. The incipient species tree can have multiple incipient species representing the species. Which one we use to draw our tree makes a difference to the shape of the tree. Figure 1e,f shows an example: S2-2-3 (in e) and S2-2-19 (in f) have different divergence times from S4-4-9, and hence, these trees are different. In summary, the LME approximation and sampling affect the tree, and hence, they potentially affect parameter estimates. Our aim here is to determine how substantial these effects are.

Data simulations and maximum likelihood estimations

We simulated 1000 phylogenetic trees under the protracted speciation model for various parameter sets ($b_1 = b_2 = b = 0.3, 0.4, 0.5, 0.6, 0.7$; $\mu_1 = \mu_2 = \mu = 0, 0.1, 0.2$; $\lambda = 0.1, 0.3, 1$). We kept the speciation-initiation rates for good and incipient species equal, because this is a requirement for the likelihood derivation to hold, and here, we were not interested in deviations from this assumption. We also kept the extinction rates for good and incipient species equal, but this was simply to limit the number of parameter sets. The range of speciation–initiation rates considered covers conditions that generate very small (number of tips ~3) to very large trees (number of tips ~80 000). Speciation-initiation rates outside the range 0.3–0.7 lead to too small or too large trees that are no longer manageable. The range of speciation-completion rates we chose results in trees with severe to almost no lineage accumulation slowdowns (Etienne & Rosindell, 2012). Each phylogeny was simulated in R (R Core Team, 2015), using the package PBD (Etienne, 2016). We used the 'pbd_sim' function with a fixed crown age of 15 My, conditional on the survival of both original crown lineages. An example of the output of this function is shown in Figure 1. This function uses the Gillespie algorithm to generate incipient species phylogenies under the protracted speciation model (Figure 1a,b). The tree of all extant lineages (incipient species tree) is then pruned to sample only one incipient or good extant lineage per species, in order to obtain the final reconstructed species tree (Figure 1d,f). Sampling a single incipient species to represent a species is necessary to obtain a species-level tree, because species are not necessarily monophyletic under the protracted speciation model (e.g. species S2-2 in Figure 1b). The resulting tree is equivalent to the tree that a taxonomist would reconstruct from studying extant species. This sampling can be performed in several ways. One can sample at random (Figure 1d); one can sample the oldest descendant lineage (Figure 1e) or the youngest lineage (Figure 1f). We did not consider the latter option here. We call the first two trees (Figure 1d,e) the 'randomly sampled' and 'oldest sampled' trees. The 'pbd_sim' function also provides the opportunity to apply the LME approximation to the incipient species tree, by sampling the oldest incipient daughter lineage but ignoring certain cases of representative species. We call the resulting species tree the 'approximate tree'. Hence, both the oldest sampled tree and the approximate tree deterministically sample the oldest daughter lineages, but the oldest sampled tree accounts for all cases of representative species whereas the approximate tree ignores certain cases (LME approximation). Thus, comparing the approximate tree with the oldest sampled tree enables us to directly assess the deviation in parameter estimates introduced by the LME approximation. Like the oldest sampled tree, the randomly sampled tree accounts for all cases of representative species, but it samples the incipient lineages tree at random. Furthermore, comparing the randomly sampled tree with the oldest sampled tree enables us to assess the variation introduced by sampling.

One could evaluate the robustness of the LME approximation by assessing how strongly the LME approximation misjudges the tree size as characterized by the difference between the sizes of the approximate tree and the oldest tree. However, it is more important to know how such differences translate to parameter estimates, because this relationship may not be simple and parameter estimates (and quantities derived from them) are what we are ultimately interested in. Hence, the parameters (speciation-initiation rate b, speciation-completion rate λ and extinction rate μ which we assumed to be the same for both incipient and good species) were estimated by maximizing the approximate likelihood developed by Lambert *et al.* (2015), using the 'pbd_ML' function of the PBD package. For each set of parameters we also evaluated and estimated the mean duration of speciation (τ), which is the time it takes for an incipient species to become good or to have an incipient descendent that becomes good. We refer to Etienne & Rosindell (2012) and Etienne *et al.* (2014) for the details of the computation of this quantity (also included in the package PBD). These estimates were compared within each pair of trees (oldest and approximate; oldest and random) by taking the absolute difference between the two estimates. We call the distance between the estimates the LME approximation effect and the sampling effect, respectively. We then investigated how the LME approximation effect and the sampling effect vary across the parameter space.

Results

Robustness of the approximate likelihood

When the LME approximation is applied to the simulated data, we can explore how well the likelihood can estimate the parameters that generated the data when they are produced under the same assumptions as the likelihood computation. We find that the estimates of the speciation-initiation rate (b) and the extinction rate (μ) are biased and highly variable, in particular for low speciation-completion rates and high extinction rates, even for high values of b. The estimates of the net diversification rate ($b - \mu$), the speciation-completion rate (λ) and the mean duration of speciation (τ) are unbiased and quite precise (Figure S1a–e). A larger speciation-initiation rate (b) always leads to a much smaller variance and bias in estimates in all cases. We observed a strong correlation between the tree size and the speciation-initiation rate (Figures S2 and S3); the reduction of variance and bias in parameter estimates for larger speciation-initiation rate is mainly due to the larger size of the trees (Figure S4).

From hereon, we focus on the net diversification rate ($b - \mu$), the mean duration of speciation (τ) and the speciation-completion rate (λ), because these are the only parameters that can be reliably estimated.

LME approximation effect on parameter estimates

For estimates of $b - \mu$ and τ, the deviation in the estimates introduced by the LME approximation (absolute difference between parameter estimates from the oldest sampled tree and the approximate tree) is, as expected, zero when the extinction rate (μ) is equal to zero, and it increases as μ increases and decreases with the speciation-completion rate (λ) (Figures 2 and 3). This observation is true for any value of the speciation-initiation rate (b). By contrast, the size and direction of the effect of the speciation-initiation rate (b) on the deviation introduced by the LME approximation is highly dependent on the other parameters of the model. Depending on the combination of μ and λ, increasing values of simulated b can increase or decrease the LME approximation effect, and the relation is not always monotonic (Figure S5a–c). For estimates of the speciation-completion rate (λ), we observed that the deviation introduced increases with the extinction rate (μ) for any value of the speciation-initiation rate (b). Interestingly, for low values of the speciation-initiation rate (b), it decreases with the simulated speciation-completion rate (λ), whereas for high values, it increases. Finally, we observed a quantitative difference in the deviation introduced in the estimates of these three parameters: the deviation introduced tends to be higher in the estimation of τ and λ than in $b - \mu$ (Figures 2–4, Table S1).

Sampling effect on parameter estimates

In all cases (estimates of $b - \mu$, τ and λ), the sampling effect (absolute difference between parameter estimates from the oldest sampled tree and the randomly sampled tree) introduces deviations for any value of simulated μ, including $\mu = 0$. The deviation introduced by sampling increases for increasing values of μ (Figures 2–4). For estimates of $b - \mu$ and τ, the sampling effect, like the LME approximation effect, becomes more important for small simulated λ (i.e. when speciation takes more time to complete). By contrast, for estimates of the speciation-completion rate (λ), the deviation introduced is more important for high values of simulated λ. The size and direction of the effect of the speciation-initiation rate (b) on the deviation introduced by the sampling depend on the other parameters of the model (Figure S5a–e), as we also observed for the LME approximation effect.

Comparison between LME approximation effect and sampling effect

The sampling effect was generally larger than the LME approximation effect. Only two parameter combinations caused the deviation introduced by the LME approximation to be larger than the deviation introduced by sampling ($b = 0.3$, $\mu = 0.2$ with $\lambda = 0.1$ or 0.3). As soon as b slightly increases, for the same combination of μ and λ, the LME approximation effect becomes smaller than the sampling effect. Their relative prevalence essentially depends on b and μ: for $\mu = 0$, there is no case to which the LME applies, and thus, it introduces no deviation whereas the sampling effect is still present. As μ increases, their relative prevalence becomes more even, but the sampling effect remains almost always higher than the LME effect. The speciation-initiation rate interacts in a complex way with μ and λ to determine the extent of LME and sampling effects (see Discussion) but overall, when μ is different than zero, increasing b tends to increase the relative prevalence of the sampling effect over the LME effect (Figure S5a–c).

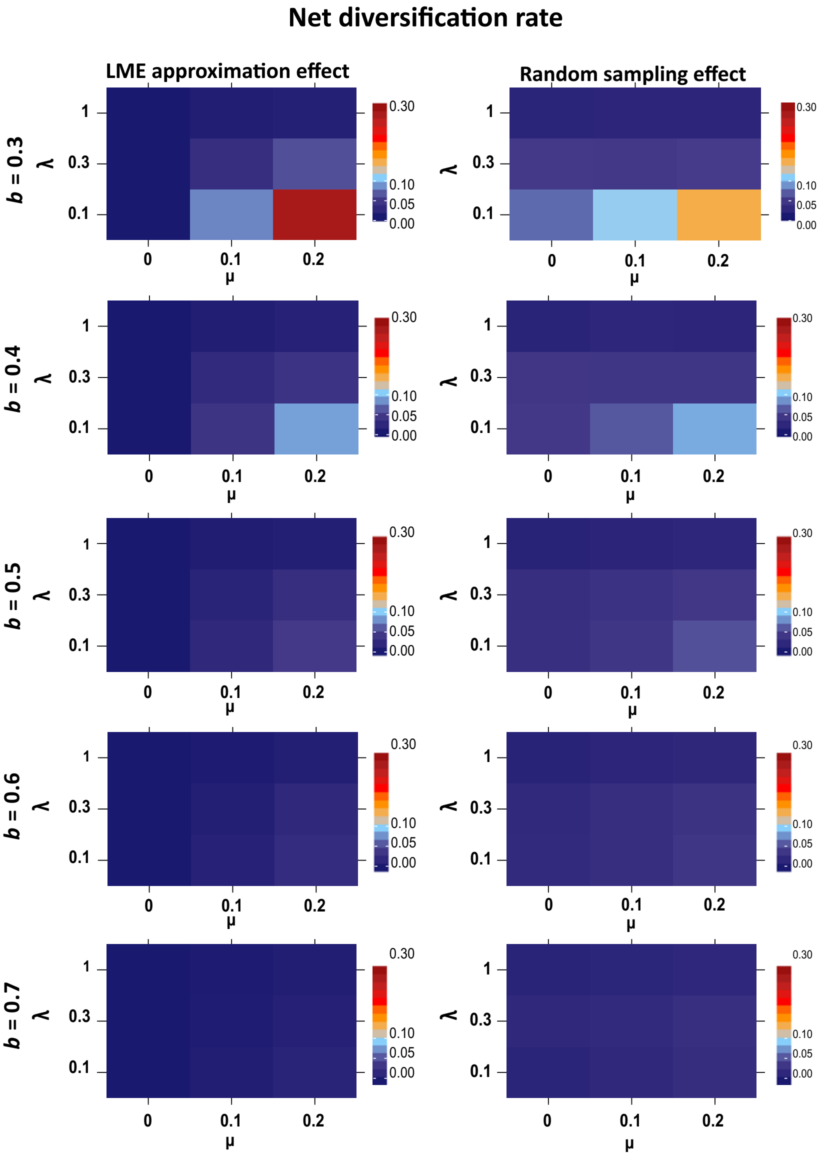

Fig. 2 95th percentile of the deviation introduced by the LME approximation (left) and the sampling effect (right) for the estimation of the net diversification rate ($b - \mu$), for various values of the parameters.

Across all parameter combinations explored, the maximum median deviation (i.e. the maximum of all medians) introduced by either the LME approximation or the sampling effect remained smaller than 0.1 My^{-1} for the estimation of the net diversification rate $b - \mu$ and the speciation-completion rate (λ), and smaller than 0.5 My for the mean duration of speciation (τ). The average of all medians across all parameter combinations was even much lower (Table 1).

Discussion

In this paper, we have studied whether the approximate likelihood of the protracted speciation model derived by Lambert *et al.* (2015) can be reliably used to fit the protracted speciation model to empirical data, despite sometimes counting fewer species in a phylogenetic tree than there are. We first confirmed that when the generating model is the same as the model used to derive the likelihood (i.e. the LME approximation model), the estimates of speciation-initiation rate (b) and extinction rate (μ) are biased and highly variable whereas the estimates of the net diversification rate (b–μ), the mean duration of speciation (τ) and the speciation-completion rate (λ) have little bias. This is in agreement with the results of Etienne *et al.* (2014), as expected because this part of our study (results relative to approximate tree) is a replication of their analysis, but for a wider range of parameters. However, this confirmation was not the main goal of our analysis. Our main concern was whether the likelihood based on the approximate model would deliver biologically meaningful parameters when the generating

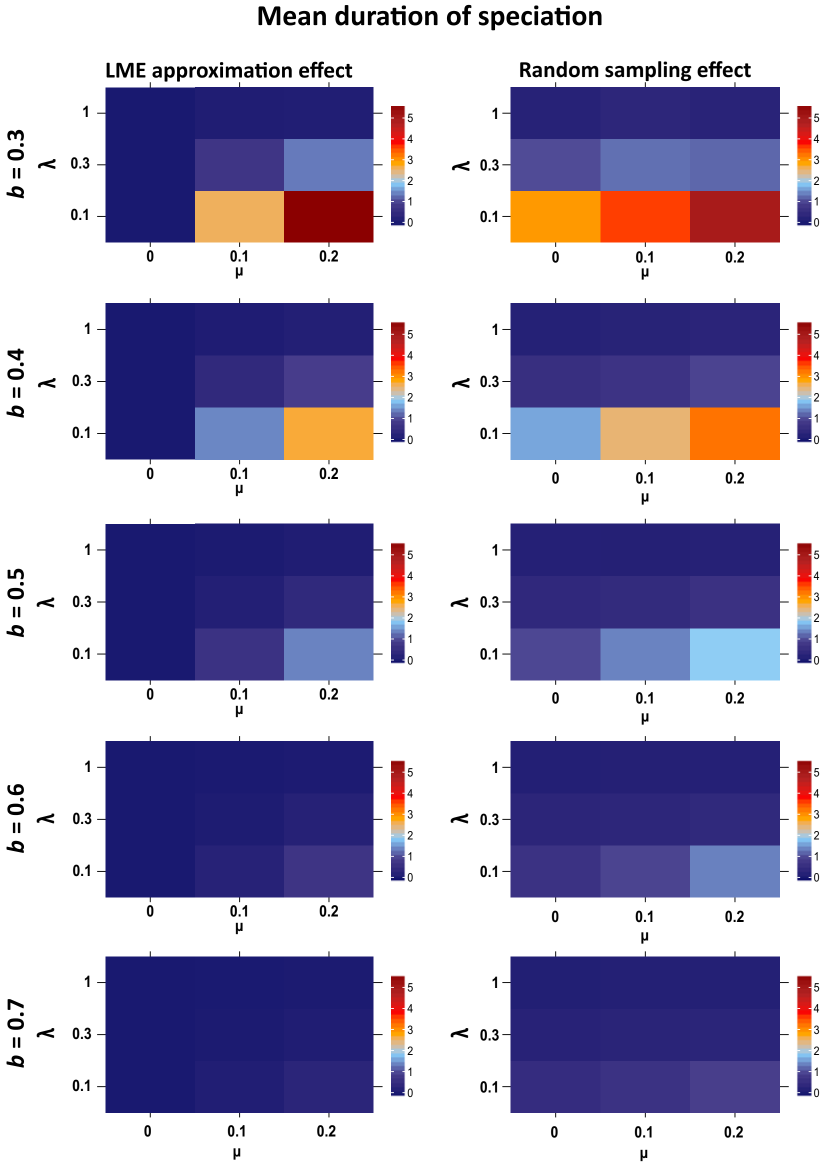

Fig. 3 95th percentile of the deviation introduced by the LME approximation (left) and the sampling effect (right) for the estimation of the mean duration of speciation (τ), for various values of the parameters.

model was, instead of the approximate model, the full protracted speciation model that does not make this approximation and hence sometimes has more tips (Figure 1). To answer this question, it turned out that it was relevant to know how sampling of incipient species to represent the species influences parameter estimates. We will discuss the effects of the approximation and of sampling in order.

LME approximation effect on parameter estimates

Our results show that as simulated μ increases, the deviation introduced by the LME approximation in the parameter estimates becomes more important. This result makes sense, because μ 'controls' how often the cases to which the approximation applies happen. When $\mu = 0$, there is no such case of a good parent species that becomes extinct leaving orphaned incipient daughter species. Thus, the approximate tree and the oldest sampled tree are identical. As μ increases, it becomes more likely that such a case will happen, provided that there is a speciation-initiation rate high enough to ensure that the extinct parent leaves a daughter species before going extinct. We always kept $b > \mu$ in our simulations, which makes this highly likely.

Our results show further that, for the estimates of μ and $b - \mu$, the deviation introduced by the LME approximation is more important for small values of the speciation-completion rate (λ), as expected. Obviously, when the speciation-completion rate is high, and hence, the time to complete speciation is very short, at the present, there are rarely still incipient species left,

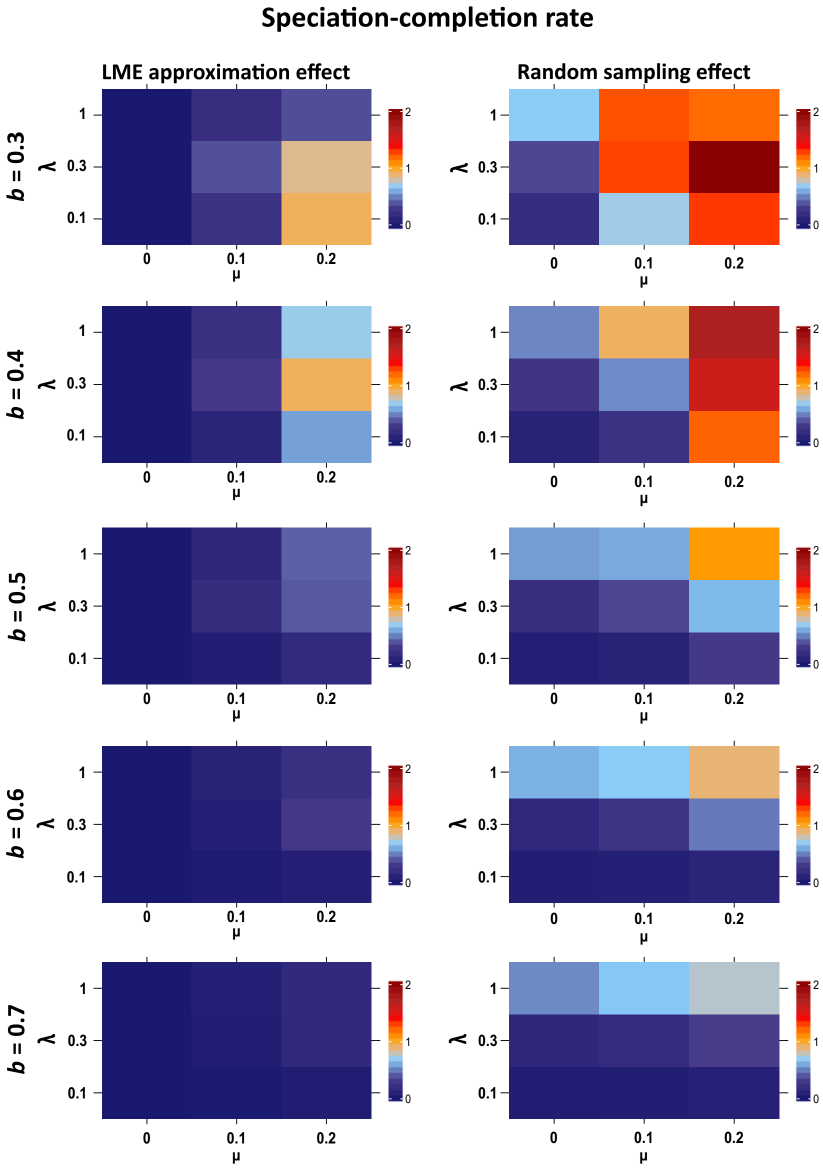

Fig. 4 95th percentile of the deviation introduced by the LME approximation (left) and the sampling effect (right) for the estimation of the speciation-completion rate (λ), for various values of the parameters.

Table 1 Median and 95th percentile deviation introduced by the LME or sampling effect, averaged across all parameters combinations

Parameter estimated	Effect	Median	95th perc.
Net diversification rate (My^{-1})	LME	0.003	0.02
	Sampling	0.009	0.03
Mean duration of speciation (My)	LME	0.06	0.47
	Sampling	0.20	0.90
Speciation-completion rate (My^{-1})	LME	0.01	0.19
	Sampling	0.06	0.58

and thus, no representative species to which the LME approximation can apply. Hence, a combination of high μ and low λ creates many cases of representative species and thus will introduce the largest deviation in the estimates.

The same line of argument leads to the expectation that the deviation introduced by the approximation will decrease for higher speciation- completion rates for the estimates of λ as well: this deviation must vanish when the speciation-completion rate approaches infinity (i.e. as the speciation becomes instantaneous, there are no incipient species). The deviation indeed vanishes for infinite speciation-completion rate, but actually increased with the simulated speciation-completion rate for higher values of b. To explain this, we recall the criteria for the LME approximation effect to be active: (i) a good parent must have become extinct, (ii) this parent must have left multiple extant incipient daughter lineages, and (iii) the oldest daughter species must have become good. A necessary condition for the first criterion to be met is a nonzero extinction rate. The second criterion will be met more often when the speciation-initiation rate is high and the speciation-completion rate is low. The third criterion requires the speciation-completion rate not to be too low, as there would not be any good descendant lineages. Hence, we conclude that for a given b and μ, there is an optimal speciation-completion rate λ where the LME approximation effect is largest. Below this optimum, fewer lineages complete speciation and hence the third criterion is less likely to be satisfied, whereas above it too many lineages complete speciation and hence the second criterion is more often not met. We see this indeed in Figure 3, for b and $\mu > 0$. This optimum shifts upwards for higher values of b, consistent with the second criterion: more incipient species provide more opportunities for the LME approximation effect to occur.

The effect of b on the consequences of the LME approximation is not straightforward for this very reason. Looking across all combinations of λ and μ, increasing b can increase or decrease the amount of deviation introduced by the LME approximation, and the relation can be nonmonotonic for certain combinations of μ and λ. The effect of b is simply to increase the number of new lineages, but the consequence of this depends on what happens to these lineages, which is determined by μ and λ. For example, for low b and high λ, we are likely well to the right of the optimal λ and hence little LME approximation effect; increasing b shifts this optimum upward so that the LME approximation effect becomes more active. By contrast, for low λ, we are likely to the left of optimal λ and increasing b brings us further from this optimum, thus reducing the LME approximation effect.

Sampling effect on parameter estimates

The extinction rate (μ) has little or no impact on the sampling effect whereas the speciation-completion rate (λ) has. This result is intuitive: it is λ that determines the presence of incipient lineages in a clade from which to sample. The effect of b on the sampling effect is, like for the LME approximation effect, very dependent on the combination of the two other parameters. For example, for $\lambda = 1$ (speciation-completion is fast), species are very likely to be good at present, so the sampling becomes deterministic (i.e. all species will be sampled). As b increases, newly initiated lineages very close to the present that are still incipient are more and more likely despite the fast speciation-completion rate. Thus, for high λ, increasing b leads to a more pronounced sampling effect. By contrast, for $\lambda = 0.1$, there is a sampling effect already for small speciation-initiation rate (b) because there are many incipient lineages among which to sample. A higher b leads to many more speciation-initiation events, and so the branching times are closer to each other. This implies that two random samples among these lineages will more likely have similar branching times than when b is small (i.e. two randomly sampled trees will be more similar). Hence, when λ is small, increasing b leads to a less pronounced sampling effect.

Comparison between the LME approximation effect and the sampling effect

We observed that in most cases, the LME effect is smaller than the sampling effect. This means that from the incipient species tree, two sampled trees without LME approximation but randomly sampled will be more different than two trees pruned in the same deterministic way but one with the LME approximation and one without LME approximation. Only for high μ and low λ, the LME effect becomes more substantial than the sampling effect. In these cases, the largest median observed difference between estimates on approximate trees and nonapproximate trees is 0.02 My^{-1} for the net diversification rate ($b-\mu$) which is of the same order of magnitude as empirical estimates of diversification rate in various animals and plants clades (Magallon & Sanderson, 2001; Ricklefs, 2007; Scholl & Wiens, 2016). For the duration of speciation, however, the maximum deviation introduced (0.49 My) remains below empirical estimates. For example, it was estimated to be at least 2 My in various vertebrates clades by Avise *et al.* (1998) and to be between 1 and 5 My in primates (Curnoe *et al.*, 2006). For a more conservative assessment of the deviation introduced by this approximation, one can look at the 95th percentile of the distribution. In this case, in this part of the parameter space, the maximum deviation introduced becomes higher than empirical estimates of these quantities, suggesting that the LME approximation is no longer negligible. However, not only is this restricted to a small part of the parameter space, it also remains quantitatively similar to the amount of uncertainty introduced by sampling of the incipient species trees, which always occurs in the building of empirical phylogenetic trees. Hence, the LME approximation generally introduces a deviation which is negligible compared to the effects of the pruning process inherent to the protracted model.

Overall, our results suggest that the likelihood method developed by Lambert *et al.* (2015) can be reliably applied despite its approximation, at least if the estimates are within the range of values that we tested for. Only a restricted part of the parameter space corresponding to a low speciation-initiation rate, a high extinction rate and a small speciation-completion rate could lead to a substantial amount of deviation introduced by this approximation. We further note that the derivation of this likelihood requires the speciation-initiation of good and incipient species to be equal.

There are several reasons to believe that these rates should be different. For example, interspecific interactions within diversifying clades can reinforce or restrict diversification rate (Drury *et al.*, 2016). Such interactions may affect good and incipient species differently. A simulation study like ours can be used to test this. However, a crucial difference is that in such a study, there are two (or more) initiation rates in the simulation, but only one in the estimation, requiring a proper way to compare simulated and estimated values.

We expected that the biologically unrealistic LME approximation would have a large effect, but we found the opposite: deviations between parameter estimates under the biologically relevant model and the mathematically tractable, but biologically unrealistic model remained small. This is an important lesson for the utility of models. Here, the hypothesis is that speciation takes time. The exact implementation of this idea, that may violate further biological realism, is not so relevant. We can use our approach also for larger differences between original and approximating model. For example, we could construct more mechanistic models of speciation, such as variations on the Bateson–Dobzhansky–Muller model of hybrid incompatibilities (Gavrilets, 2004), or adaptive dynamics models of sexual (and natural) selection (van Doorn *et al.*, 2009), simulate with them and then test whether our simple birth–death model of protracted speciation also captures the essence of these models. This is more difficult than in the current paper, because there is no immediately obvious relationship between parameters of the protracted speciation model and parameters of these more mechanistic models. In fact, such a study would establish such relationships. This is an interesting avenue for future research. Our results justify the use of the approximate likelihood for such analyses.

Acknowledgment

We thank Sander van Doorn for his view on this manuscript. RSE thanks the Netherlands Organization for Scientic Research (NWO) for financial support through a VICI grant.

References

Avise, J.C., Walker, D. & Johns, G.C. 1998. Speciation durations and Pleistocene effects on vertebrate phylogeography. *Proc. R. Soc. Lond. B* **265**: 1707–1712.

Curnoe, D., Thorne, A. & Coate, J.A. 2006. Timing and tempo of primate speciation. *J. Evol. Biol.* **19**: 59–65.

Drury, J., Clavel, J., Manceau, M. & Morlon, H. 2016. Estimating the effect of competition on trait evolution using maximum likelihood inference. *Syst. Biol.* **65**: 700–710.

Etienne, R.S. 2016. *PBD: Protracted Birth-Death Model of Diversification*. R Package Version 1.3 https://cran.r-project.org/web/packages/PBD/index.html

Etienne, R.S. & Rosindell, J. 2012. Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Syst. Biol.* **61**: 204–213.

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

Etienne, R.S., Morlon, H. & Lambert, A. 2014. Estimating the duration of speciation from phylogenies. *Evolution* **68**: 2430–2440.

Gavrilets, S. 2004. *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton, NJ.

Kendall, D.G. 1948. On the generalized birth-and-death process. *Ann. Math. Stat.* **19**: 1–15.

Lambert, A. 2010. The contour of splitting trees is a Levy process. *Ann. Prob.* **38**: 348–395.

Lambert, A. & Stadler, T. 2013. Birth-death models and coalescent point processes : the shape and probability of reconstructed phylogenies. *Theor. Popul. Biol.* **90**: 113–128.

Lambert, A., Morlon, H. & Etienne, R.S. 2015. The reconstructed tree in the lineage-based model of protracted speciation. *J. Math. Biol.* **70**: 367–397.

Magallon, S. & Sanderson, M. 2001. Absolute diversification rates in angiosperm clades. *Evolution* **55**: 1762–1780.

McPeek, M.A. 2008. The ecological dynamics of clade diversification. *Am. Nat.* **172**: E270–E284.

Moen, D. & Morlon, H. 2014. Why does diversification slow down? *Trends Ecol. Evol.* **29**: 190–197.

Morlon, H. 2014. Phylogenetic approaches for studying diversification. *Ecol. Lett.* **17**: 508–525.

Nee, S. 2001. Inferring speciation rates from phylogenies. *Evolution* **55**: 661.

Phillimore, A.B. & Price, T.D. 2008. Density-dependent cladogenesis in birds. *PLoS Biol.* **6**: 483–489.

Purvis, A., Orme, D.L., Toomey, N.H. & Pearson, P.N. 2009. Temporal patterns in diversification rates. In: *Speciation and Patterns of Diversity* (R.K. Butlin, J.R. Bridle & D. Schulter, eds), pp. 278–301. Cambridge University Press, British EC Edition, London.

R Core Team 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rabosky, D.L. & Lovette, I.J. 2008. Density-dependent diversification in North American wood warblers. *Proc. R. Soc. B* **275**: 2363–2371.

Ricklefs, R.E. 2007. Estimating diversification rates from phylogenetic information. *Trends Ecol. Evol.* **22**: 601–610.

Scholl, J.P. & Wiens, J.J. 2016. Diversification rates and species richness across the Tree of Life. *Proc. R. Soc. B*, **283**: 20161334.

van Doorn, G., Edelaar, P. & Weissing, F. 2009. On the origin of species by natural and sexual selection. *Science* **326**: 1704–1707.

Supporting information

Additional Supporting Information may be found online in the supporting information tab for this article: **Figure S1** (a) Error in estimates of the speciation initiation rate (b) for each parameter combination. (b) Error in estimates of the extinction rate (μ) for each parameter combination. (c) Error in estimates of the extinction rate (λ) for each parameter combination. (d) Error in estimates of the extinction rate (b-μ) for each parameter combination. (e) Error in estimates of the extinction rate (τ) for each parameter combination.

Figure S2 Tree size after sampling for the different values of speciation initiation rate simulated (b = 0.3, 0.4, 0.5, 0.6,0.7).

Figure S3 Tree size after sampling (in log scale) for ech parameter combination.

Figure S4 Error in estimates (estimate–true value) of the five parameters as a function of tree size (oldest sampled tree), in log10 scale.

Figure S5 (a) Deviation in estimates of the mean duration of speciation (τ). (b) Deviation in estimates of the net diversification rate ($b-\mu$). (c) Deviation in estimates of the speciation completion rate (λ).

Table S1 95th percentile of the deviation introduced by the LME approximation and the random sampling for the net diversification rate ($b-\mu$), the mean duration of speciation (τ) and the speciation completion rate (λ).

Received 16 July 2017; revised 22 November 2017; accepted 4 December 2017