bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

DETECTING PHYLODIVERSITY-DEPENDENT DIVERSIFICATION WITH A GENERAL PHYLOGENETIC INFERENCE FRAMEWORK

A PREPRINT

Francisco Richter
Institute of Computing
Università della Svizzera italiana (USI)
richtf@usi.ch

Ernst C. Wit
Institute of Computing
Università della Svizzera italiana (USI)
ernst.jan.camiel.wit@usi.ch

Thijs Janzen
Groningen Institute for Evolutionary Life Sciences,
University of Groningen,
t.janzen@rug.nl

Hanno Hildenbrandt
Groningen Institute for Evolutionary Life Sciences,
University of Groningen,
h.hildenbrandt@rug.nl

Rampal S. Etienne
Groningen Institute for Evolutionary Life Sciences,
University of Groningen,
r.s.etienne@rug.nl

July 1, 2021

ABSTRACT

Diversity-dependent diversification models have been extensively used to study the effect of ecological limits and feedback of community structure on species diversification processes, such as speciation and extinction. Current diversity-dependent diversification models characterise ecological limits by carrying capacities for species richness. Such ecological limits have been justified by niche filling arguments: as species diversity increases, the number of available niches for diversification decreases.

However, as species diversify they may diverge from one another phenotypically, which may open new niches for new species. Alternatively, this phenotypic divergence may not affect the species diversification process or even inhibit further diversification. Hence, it seems natural to explore the consequences of phylogenetic diversity-dependent (or phylodiversity-dependent) diversification. Current likelihood methods for estimating diversity-dependent diversification parameters cannot be used for this, as phylodiversity is continuously changing as time progresses and species form and become extinct.

Here, we present a new method based on Monte Carlo Expectation-Maximization (MCEM), designed to perform statistical inference on a general class of species diversification models and implemented in the R package `emphasis`. We use the method to fit phylodiversity-dependent diversification models to 14 phylogenies, and compare the results to the fit of a richness-dependent diversification model. We find that in a number of phylogenies, phylogenetic divergence indeed spurs speciation even though species richness reduces it. Not only do we thus shine a new light on diversity-dependent diversification, we also argue that our inference framework can handle a large class of diversification models for which currently no inference method exists.

Keywords Phylogenetic trees · Network sciences · Point processes

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

1 Introduction

The hypothesis of diversity-dependent diversification posits that diversification processes at macro-evolutionary scales are affected by community structure, and particularly by diversity [Walker and Valentine, 1984, Gould et al., 1977]. One of the underlying ideas is that there are ecological limits to diversity (there is a limited number of niches that can be filled with species) and hence to diversification [Rabosky, 2009]. The hypothesis has been extensively studied both empirically and theoretically [Condamine, 2018, Gibb et al., 2016, Cunha et al., 2017, Pouchon et al., 2018, Chen et al., 2017, Pinto-Ledezma et al., 2017, McGuire et al., 2014, Pyron and Wiens, 2013, Xu and Etienne, 2018, Etienne et al., 2016, Liow et al., 2010, Herrera-Alsina et al., 2018, Morlon, 2014, Rabosky and Hurlbert, 2015, Jønsson et al., 2012]. However, currently developed inference models for detecting diversity-dependent diversification from molecular phylogenies consider only species richness as a proxy for diversity [Etienne et al., 2012a].

Phylogenetic diversity, quantifying the genetic differences among a group of species, has been identified as a key feature of diversity [Kling et al., 2018, Scheiner et al., 2017] to be taken into account in conservation biology [Laity et al., 2015, Faith and Baker, 2006] (but see Cantalapiedra et al. [2019], Mazel et al. [2018]), community ecology [Stadler et al., 2017, Tucker et al., 2016, Webb et al., 2006, Violle et al., 2011], evolutionary biology [Kling et al., 2018] and the intersection of these fields. Phylogenetic diversity, or phylodiversity, provides a different perspective on diversity and ecological limits. On the one hand, species richness models suggest that as species diverge there may be less opportunity to speciate further because the growing phenotypic space between species leaves less room to be occupied. However, one could argue that on the other hand, the divergence provides access to increased phenotypic space to speciate into. Hence, extending diversity-dependence to phylodiversity and developing methods to infer such phylodiversity-dependent diversification from molecular phylogenies seems worthwhile. It will allow us to consider the dynamic nature of ecological limits [Costa et al., 2008, Lister, 1976, Soininen et al., 2011] and thus relax the assumption of fixed limits [Marshall and Quental, 2016, Etienne et al., 2012a].

The incorporation of phylogenetic diversity is not possible with the current simulation-free methods for inferring diversity-dependent diversification using the Q-approach introduced by Etienne et al. [2012a] and Laudanno et al. [2020] and implemented in the R package `DDD`. The Q-approach is based on a hidden Markov model approach where the probability of an extant-species tree is integrated over the infinite set of complete trees compatible with it, i.e., the trees that also contain now-extinct species. By only taking into account the number of species at any point in time, the Q-approach does not depend on tree topology. Contrastingly, phylogenetic diversity, defined as the sum of the lengths of all branches in a phylogenetic tree [Faith, 1992], highly depends on the topology of the tree as well as the branching times. Hence, a new methodology is needed to incorporate topological characteristics of the diversification processes.

To do so we generalize a recently developed statistical framework [Richter et al., 2020] based on Monte Carlo Expectation-Maximization (MCEM) that allows inference on a general class of diversification models, including models with phylodiversity-dependent diversification. In this EMPHASIS (Expectation-Maximization in PHylogenetic Analysis with Simulations and Importance Sampling) framework, maximum likelihood estimation is performed on an augmented data set generated by Monte Carlo simulations.

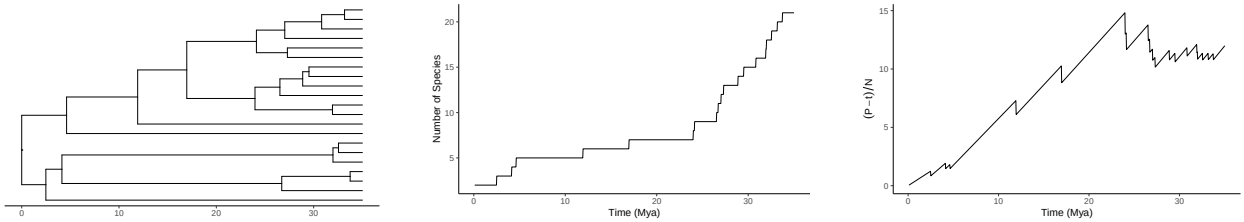
The general class of Species Diversification Models (SDM) introduced in Richter et al. [2020] contains a broad spectrum of scenarios considered in the literature where rates can be constant [Nee et al., 1994], related to the age of the species [Hagen et al., 2015], to the (changing) paleo-enviroment [Descombes et al., 2018], to geographic patterns [Goldberg et al., 2011] or to temperature and diversity [Condamine et al., 2019], just to name a few. For each of these models specific likelihood formulas have been derived (implemented in different packages), but our new method can handle all of these within a single framework, and also applies to combinations of these models for which no such likelihood formula is available and is often impossible to derive or compute numerically. It also applies to new models such as the phylodiversity-dependent models discussed in detail here, and other models with possibly complex interactions between ecological factors and macroevolution, thereby opening endless opportunities for macroevolutionary diversification analysis. The main challenge of our framework is computationally: the Monte Carlo integration is very demanding. In this manuscript we therefore provide a method to perform this integration efficiently.

We illustrate our inference method by applying it to 14 phylogenies, comparing a phylodiversity-dependent diversification model to a diversity-dependent diversification model. We generally find little difference between these two models, although the phylodiversity-dependent diversification model provides an additional narrative for the evolution of global speciation through time in several cases.

2 Diversity-Dependent Diversification Models

Diversity-dependent species diversification models are typically used to quantify the effect that diversity has on diversification [Etienne et al., 2012b, Cunha et al., 2017, Etienne and Haegeman, 2012, Foote et al., 2018, Condamine

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) Phylogeny of Megapodidae (b) Number of species. (c) Phylogenetic diversity per species

Figure 1: Phylogeny, number of species and phylogenetic diversity per species.

et al., 2019]. Under the classical linear diversity-dependent diversification (LDD) model, it is assumed that speciation rate is a linear function of species richness:

$$\lambda_t = \lambda_0 + \beta_N N_t; \qquad \mu_t = \mu_0, \tag{1}$$

where λ_t is the per species speciation rate, N_t the number of species and μ_t represents the per species extinction rate, at time t. Assuming that λ_0 is positive, if β_N is negative the quantity $K' = -\lambda_0\beta_N$ is called the carrying capacity, which denotes the value for which a clade approaches a niche limit and consequently experiences a slow-down in speciation. If $\beta_N = 0$, then the model reduces to the diversity-independent diversification model, i.e. the constant-rate model.

Phylogenetic diversity is recognised as a critical feature to take into consideration in several fields such as conservation ecology, macroecology and macroevolution. So far, it has been studied mostly in a qualitative way and only as a single number at the present instead of considering it as a dynamical quantity that changes through macroevolutionary time. Current diversity-dependent diversification models do not consider phylodiversity, and assume that diversity slows down diversification (e.g. due to niche filling), while qualitative studies suggest that diversity can spur diversification [Jarne et al., 2017, Hamilton et al., 2020]. Likelihood-based inference approaches ignore phylodiversity; they fully describe the processes by considering the probability that the clade has N_t lineages at time t, but ignore the topology of the trees. We here introduce a generalised diversity-dependent diversification model, i.e., a phylodiversity-dependent diversification (LPD) model, where we assume that the speciation rate also depends on the phylogenetic diversity per species:

$$\lambda_{t;\beta} = \lambda_0 + \beta_N N_t + \beta_P \frac{P_t - t}{N_t}; \qquad \mu_{t;\beta} = \beta_0 \tag{2}$$

where P_t is the phylogenetic diversity at time t defined as the total branch length of the reconstructed tree up until t. The quantity $\frac{P_t - t}{N_t}$ corresponds to the phylogenetic diversity per species at time t. Note that by subtracting t from the phylogenetic diversity P_t, the phylogenetic diversity per species remains 0 for a single species. In Figure 1, an example tree is plotted, with the species richness though time and the phylogenetic diversity per species through time.

Here, we make use of the statistical methods described in Richter et al. [2020], combined with an efficient importance sampler, in order to perform statistical inference assuming diversification dynamics given by the LPD model and compare it with the diversification dynamics given by the simple LDD model. In this way, we quantify the signal that phylodiversity leaves in species diversification and evaluate if its incorporation in diversity-dependent diversification models is promising for further studies.

3 Materials and Methods

Phylogenetic trees are branching diagrams, reconstructed from DNA sequences, representing the evolutionary history of species diversification [Kapli et al., 2020]. Mathematically, they are represented by a discrete part given by the topology of the tree and a continuous part given by its branching times. We define a tree $x = \{\mathbf{t}, \tau\}$ as a combination of branching times and topology. More precisely, the branching times are defined by a chronological sequence vector of times $\mathbf{t} = \{t_0,\ t_1,\ t_2, ...,\ t_p\}$, with $t_0 = 0$ and t_p being the present time. The topology is defined by a succession of allocation values which can be characterized in multiple ways such as in network or matrix notation. Here, we consider the succession of species names $s_1^*, s_2^*, ..., s_{p-1}^*$ to be the species that diversified (or became extinct) at branching time t_i. Moreover, we define the subsets of subindex $\mathcal{C}_x \subset \{1, ..., p-1\}$ and $\mathcal{E}_x \subset \{1, ..., p-1\}$ to be the indices corresponding to speciation and extinction events, respectively. This means that if $i \in \mathcal{C}_x$ then t_i is a branching time corresponding to a speciation event while if $i \in \mathcal{E}_x$ then t_i is a branching time corresponding to an extinction event.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

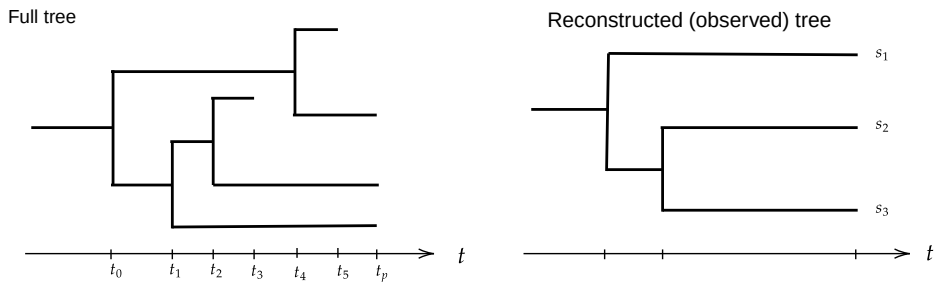


Figure 2: Full phylogenetic trees (left) and the corresponding reconstructed tree (right). Each branch represents a species.

Figure 2 shows a representation of a tree describing a full evolutionary process (speciation and extinction events), and the corresponding reconstructed tree, considering the evolutionary history of currently extant species. In ths case $\mathcal{C}_x = \{0, 1, 2, 4\}$ and $\mathcal{E}_x = \{3, 5\}$.

Throughout, we consider the extant species trees to be accurate (i.e., no uncertainty in branching times or topology). Statistically, extant species trees are our observed data and extinct species are usually latent or unobserved variables, which in the case of diversity-dependent diversification also affect diversification rates.

3.1 Diversification of species as a point process

We consider the species diversification process as a general Point Process where each species has a waiting time to speciate into two daughter species that follows an exponential probability distribution with rate $\lambda_{t,s|\theta}$, for any time t, species s and parameters θ. Species can also become extinct with an exponential distribution with rate $\mu_{t,s|\theta}$ for the waiting time to extinction. We will denote the set of extant species at time t by $\mathcal{S}_t = \{s_1, ..., s_{N_t}\}$, and the number of extant species at time t by N_t. These quantities are described by a Non-Homogenous Poisson Process (NHPP) [Daley and Vere-Jones, 2007]. Typically, we consider

$$\lambda_{t,s|\theta} = g\left(\sum_i \theta_i v_{i,t,s}\right)$$

for a set of covariates $v_{i,t,s}$ and a link function $g : \mathbb{R} \to \mathbb{R}$ [Dobson and Barnett, 2008]. The loglikelihood function of the full process represented by a complete tree (Figure 2, left) is given by

$$\ell_x(\theta) = \sum_{\mathcal{C}_x} \log\left(\lambda_{t_i,s_i^*|\theta}\right) + \sum_{\mathcal{E}_x} \log\left(\mu_{t_i,s_i^*|\theta}\right) - \sum_{i=1}^{p}\left[\int_{t_{i-1}}^{t_i} \sum_{s\in\mathcal{S}_{t_i}} (\lambda_{t,s|\theta} + \mu_{t,s|\theta})dt\right] \tag{3}$$

Phylogenies that are derived from molecular data (e.g. DNA sequences) are, however, not full trees, as they do not contain the extinct species (Figure 2 right). The likelihood for an observed tree can be written in terms of the likelihood of compatible full trees. In principle, this is simply the integration over all possible full trees that are in agreement with the observed tree x_{obs}:

$$f(x_{obs}|\theta) = \int_{x\in\mathcal{X}(x_{obs})} \exp(\ell_x(\theta|x_{obs}))dx \tag{4}$$

This integration is usually impossible to compute in practice for most diversification models. Here, we present a method where maximum likelihood estimation is possible without calculating directly the likelihood function (4), by implementing a combination of statistical inference and a data augmentation algorithm.

3.2 The EMPHASIS Statistical Framework

Our statistical framework is a generalisation of that of Richter et al. [2020], which makes use of an Expectation-Maximization algorithm for maximising the likelihood [Dempster et al., 1977]. The EM algorithm is an iterative

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

procedure consisting of two steps: the E-step and the M-step. Starting from an initial value for the parameters, the E-step involves computing the expected loglikelihood of the observed tree for the given parameters and the M-step involves computing the parameters that maximise that expectation of the loglikelihood. Each iteration the parameters are updated with the values obtained in the M-step of the previous iteration. The E- and M-steps are run iteratively until convergence is reached. The parameters thus obtained have been shown to be the maximum likelihood estimators [Dempster et al., 1977].

Because the expectation in the E-step cannot be computed exactly (or numerically) due to the high dimensionality of the space of complete trees, Richter et al. [2020] proposed to use a stochastic approximation and data augmentation [Tanner and Wong, 1987], specifically a Monte-Carlo method [Chan and Ledolter, 1995] in combination with importance sampling [Glynn and Iglehart, 1989] in the E-step of their EM-algorithm [McLachlan and Krishnan, 2007], and calculated

$$\begin{aligned} Q_\theta = \mathbb{E}_{\theta^*}[\ell_x(\theta) \mid x_{obs}] &\approx \frac{1}{N} \sum_{x_i \sim f(x_i|\theta, x_{obs})} \ell_{x_i}(\theta) \\ &= \frac{1}{N} \sum_{x_i \sim f_\alpha(x_i|\theta, x_{obs})} \ell_{x_i}(\theta) w_i \end{aligned} \tag{5}$$

where

$$w_i = \frac{f\left(x_i|\theta, x_{obs}\right)}{f_\alpha\left(x_i|\theta, x_{obs}\right)} \tag{6}$$

are called the *importance weights* and are highly dependent of the importance sampler f_α. The importance weights reflects how accurate the data augmentation is in comparison with the desired distribution, importance weights equal to 1 shows that the importance sampler is the same distribution as the likelihood f distribution that generates the process of interest. The data augmentation scheme used by Richter et al. [2020] was mathematically correct but computationally inefficient, as the paper was aimed at the conceptual framework rather than performance. Here we present an improved version of the framework, hereafter called *emphasis*, with a very efficient data augmentation scheme, because the choice of data augmentation scheme is crucial for computational performance [Van Dyk and Meng, 2001].

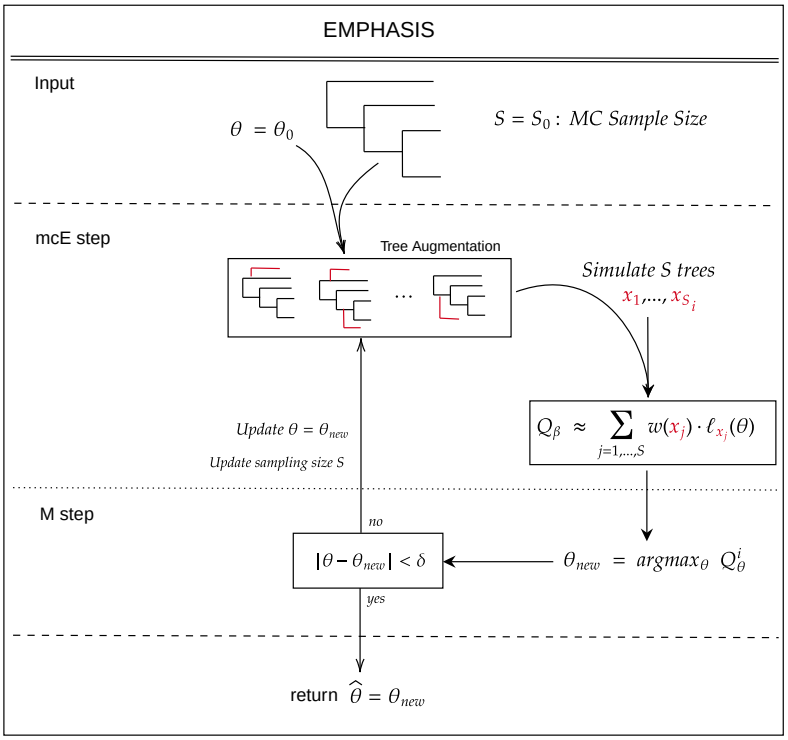


Figure 3: Monte-Carlo EM algorithm diagram in the context of phylogenetic trees.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3.3 Augmentation of observed trees, a novel importance sampler for phylogenetic inference

Richter et al. [2020] presents an MCEM algorithm where trees are augmented by drawing uniformly the number of branching events and its corresponding branching times. The method worked well for small trees, but the variance of the estimates grows fast as the tree gets larger for constant number of samples, making the method computationally intractable for medium-sized to large clades. This is due to the curse of dimensionality, i.e., the problem of exploring high-dimensional spaces efficiently [Friedman, 1997]. Our proposed alternative for the data augmentation algorithm augments trees according to the underlying diversification model, encouraging samples in the regions of parameter space that are likely under the proposed SDM.

To sample the extinct species in the tree, we approximate the diversification process of extinct lineages conditional on the extant species in the data as a birth-death process with rates

$$\lambda^m_{t,s|\theta} = \lambda_{t,s|\theta} P_\alpha(t, t_p), \quad \mu^m_{t,s|\theta} = \frac{\mu_{t,s|\theta}}{P_\alpha(t, t_p)} \tag{7}$$

where $P_\alpha(t, t_p)$ is an approximation of the probability that a species observed at time t will not have any descendants at time t_p. Kendall [1948] showed that, for lineage-independent models, the exact probability is given by

$$P_0(t_c, t_p) = \frac{\int_{t_c}^{t_p} \mu_{\tau,s|\theta} e^{-\int_{t_c}^{\tau} (\lambda_{r,s|\theta} - \mu_{r,s|\theta}) dr} d\tau}{1 + \int_{t_c}^{t_p} \mu_{\tau,s|\theta} e^{-\int_{t_c}^{\tau} (\lambda_{r,s|\theta} - \mu_{r,s|\theta}) ds} d\tau} \tag{8}$$

Note that the probability depends on information on $\lambda_{ts|\theta}$ and $\mu_{ts|\theta}$ for $t_c < t < t_p$. For constant rates this information is available and calculation of Eq. 8 is easy. However, for most SDM information on the full process is not available. For instance, in the case of diversity-dependent diversification models the quantity N_t is unknown.

We augment the observed tree with hidden speciation events. These events can be allocated to all lineages, but not with equal probability. Speciation events occurring on an observed lineage have twice the weight of speciation events occurring on an unobserved lineage [Etienne et al., 2012a]. Figure 4 shows an example when a tree is augmented with a new speciation event at a time that there are two extant lineages and one extinct lineage. In that case, there are five possible allocations. More generally, there are $N^e_{t^-} + 2N^o_{t^-}$ possible allocations, where $N^e_{t^-}$ is the number of currently extinct lineages alive just before time t and $N^o_{t^-}$ is the number of currently extant lineages just before time t. Therefore, we can compute the probability distribution for the waiting times for the augmented speciation events (which we will call missing speciation events) considering the $N^e_{t^-} + 2N^o_{t^-}$ non-homogenous Poisson processes together. Because the minimum waiting time for exponential distributed processes is also an exponential process, given a time t_0, the waiting time for the first missing speciation event to occur is given by an exponential distribution with rate

$$\sigma_{t|\theta} = \sum_{s \in \mathcal{S}^m_t} \lambda^m_{t,s|\theta} + 2 \sum_{s \in \mathcal{S}^0_t} \lambda^m_{t,s|\theta}$$

where $\mathcal{S}^m_t$ and $\mathcal{S}^o_t$ are the sets of observed and missing species at time t respectively. Hence, the probability density of the waiting time for any speciation to occur at time t, starting the process at initial time t_i, is a non-homogeneous exponential distribution with rate $\sigma_{t|\theta}$, that is

$$f_B(t_c | \, t_i \, , \theta) = \sigma_{t|\theta} e^{-\int_{t_i}^{t_c} \sigma_{t|\theta} dt}.$$

Once a missing speciation event has occurred, the new lineage needs to get an allocation and a extinction time assigned to be included in the tree. In a model where speciation rates are the same for all lineages, all allocations have the same probability

$$\mathbb{P}_A(\tau | t_c, \theta) = \frac{1}{N^e_{t^-} + 2N^o_{t^-}}.$$

The lineage produced at the missing speciation event must become extinct before the present. The extinction time of the species s born at time t_c is a random variable with a density distribution that is conditioned on extinction occurring before time t_p,

$$f(t_e | s, t_c) = \mu_{t_e,s|\theta} \frac{e^{-\int_{t_c}^{t_e} \mu_{q,s|\theta} dq}}{1 - e^{-\int_{t_c}^{t_p} \mu_{q,s|\theta} dq}}.$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

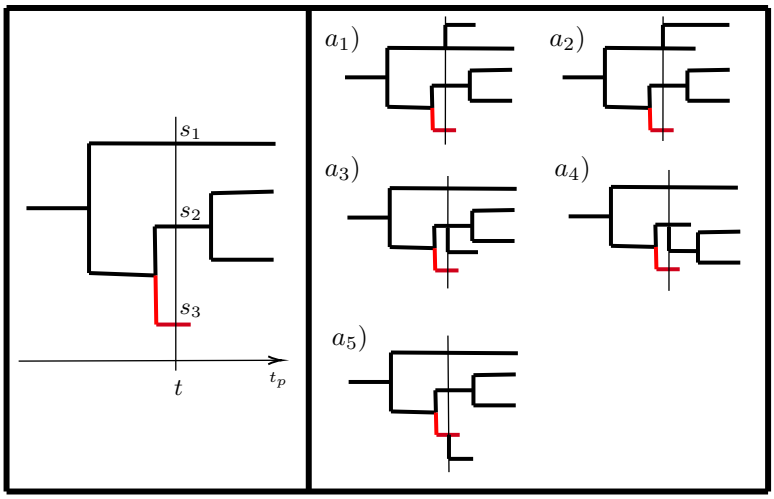


Figure 4: Phylogenetic tree with 2 observed species and 1 missing species at time t. When a new species is created there are $2N_0 + N_m$ possible allocations, in this case $2 * 2 + 1$.

Note that this probability also depends on the extinction rate of the full process (i.e., at times later than t_c), which is not always available, as it may depend, for example, on diversity at those later times. Hence, we propose to sample the extinction time from the truncated distribution

$$f_D(t_e|s,t,\theta) = \mu_{t,s|\theta}\frac{e^{-\mu_{t,s|\theta}(t_e-t)}}{1-e^{-\mu_{t,s|\theta}(t_p-t)}}.$$

The full sampling probability of the missing part of a tree under this scheme is then given as

$$f_m(x|\theta) = \prod_{i\,\in\,\mathcal{M}_\tau} f_B(t_i|\,t_{i-1})\mathbb{P}_A(a_i|t_i)f_D(t_i^e|a_i,t_i). \tag{9}$$

The data augmentation algorithm (DAA)

The main idea for our proposed data augmentation algorithm is to replace $P_\alpha(t)$ by the probability that the newly created species (and not the entire clade that will descend from it) will become extinct before the present time

$$P_1(t_c,t_p) \;=\; 1-e^{-\mu_{t_c}(t_p-t_c)}$$

Thus, we consider the evolutionary process with diversification rates

$$\lambda^m_{t,s|\theta} \;= \lambda_{t,s|\theta}(1-e^{-\mu_{t_c}(t_p-t_c)}), \quad \mu^m_{t,s|\theta} \;= \frac{\mu_{t,s|\theta}}{(1-e^{-\mu_{t_c}(t_p-t_c)})} \tag{10}$$

The algorithm is based on a Gillespie-type simulation algorithm which is a computationally simple and relatively simple digital computer algorithm [Gillespie, 1976, Kieu, 2018].

The algorithm proceeds as follows:

1. Input: Set $t_0 = 0, i \;=\; 1$.
2. Draw a **missing speciation time** t from distribution

$$f_B(t|\,t_i\,,\theta) \;= \sigma_{t|\theta}e^{-\int_{t_i}^{t}\sigma_{t|\theta}dt}.$$

   where

$$\sigma_{t|\theta} = \sum_{s\in\mathcal{S}_t^m}\lambda_{t,s|\theta}(1-e^{-\mu_t(t_p-t)}) + 2\sum_{s\in\mathcal{S}_t^0}\lambda_{t,s|\theta}(1-e^{-\mu_t(t_p-t)})$$

3. Draw an **allocation** for the species from distribution

$$P_A(\tau|t,\theta) \;= \frac{1}{N^e_{t^-}+2N^o_{t^-}}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4. Draw the corresponding **extinction time** from distribution

$$f_D(t_e|s(\tau),t,\theta) = \mu_{t,s|\theta}\frac{e^{-\mu_{t,s|\theta}(t_e-t)}}{1-e^{-\mu_{t,s|\theta}(t_p-t)}}$$

5. Set $t_i = t$, if $t_i < t_p$ update the tree with the new species (speciation time, extinction time and allocation) and go to step 2; if $t_i > t_p$ stop the algorithm and return the augmented tree.

An interpretation of this process is as follows:

- We observe a process with varying extinction rates, but as soon as a new missing species arises the extinction rate of that species is fixed throughout the rest of the process.
- When allocating a new species we assume that all possible allocations have a uniform probability distribution.
- We consider alternative probabilities $P_\alpha(t,T)$, thus indexed by α, instead of the probability of not having any descendants $P_0(t,T)$. In our proposed importance sampler we consider the probability of extinction P_1 of the just created lineage.

Figure 5 shows a diagram with the steps of the proposed data augmentation algorithm.

Using equation (9) with the data augmentation scheme described above we have the following sampling probability of the full augmentation process:

$$f_m((t,\tau)|x_{obs},\theta) = \prod_{i\,\in\,\mathcal{M}_\tau} f_B(t_i|\,t_{i-1})P_A(\tau_i|t_i)f_D\left(t_i^e|\tau_i,t_i\right) = \tag{11}$$

$$\left[\prod_{i\,\in\,\mathcal{M}_\tau}\frac{\sum_{s\in\mathcal{S}_{t_i}}\lambda_{t_i,s|\theta}}{N^e_{t_i^-}+2N^o_{t_i^-}}\mu_{t_i,s_i^*|\theta}e^{-\mu_{t_i,s_i^*|\theta}(t_i^e-t_i)}\right]\left[\prod_{i\,\in\,\{1,\ldots,p\}} e^{-\int_{t_{i-1}}^{t_i}\left[\sum_{b\in\mathcal{S}_t}\lambda_{r,b|\theta,v}\left(1-e^{-\mu_{r,b|\theta,v}(t_p-r)}\right)\right]dt}\right] \tag{12}$$

Taking the logarithm we have

$$\begin{aligned}\ell_m(\theta) \;=\; & -\int_{t_0}^{t_p}\left[\sum_{s\in\mathcal{S}_t}\lambda_{t,s|\theta}\left(1-e^{-\mu_{t,s|\theta}(t_p-t)}\right)\right]dt + \sum_{i\in\mathcal{M}_\tau}\log\left(\sum_{s\in\mathcal{S}_{t_i}}\lambda_{t_i,s|\theta}\left(1-e^{-\mu_{t_i,s|\theta}(t_p-t_i)}\right)\right)\\ & -\log\left(N^e_{t_i^-}+2N^o_{t_i^-}\right)+\log\left(\mu_{t_i,s_i^*|\theta}\right)\;-\mu_{t_i,s_i^*|\theta}\left(t_i^e-t_i\right)\end{aligned} \tag{13}$$

Example. Consider a model with a speciation rate that is the same for all lineages and with a constant extinction rate, i.e.,

$$\lambda_{t,s;|\theta} = \lambda_{t|\theta}, \forall s\in\mathcal{S}_t, \;\text{ and }\; \mu_{t,s;|\theta} = \mu_o, \forall t,s;|\theta,$$

then, the sampling probability of the DAA is

$$f_m((t,\tau)|x_{obs},\theta) = \mu_0^{\#\mathcal{M}_\tau}\prod_{i\,\in\,\mathcal{M}_\tau} e^{-\mu_0(t_i^e-t_i)}\frac{N_{t_i^-}\lambda_{t_i|\theta}}{N^e_{t_i^-}+2N^o_{t_i^-}}\prod_{i\,\in\,\{1,\ldots,p\}} e^{-N_{t_i}\int_{t_{i-1}}^{t_i}\left[\lambda_{t|\theta}\left(1-e^{-\mu_0(t_p-t)}\right)\right]dt}.$$

3.3.1 Sample size

In Monte-Carlo methods, the variance of the estimates and the convergence time are determined by the sample size, the explored region of the parameter space and the type of data. From these three factors, we have control only over the sample size. MC methods require a sensible choice of the sample size, and it much depends on the type of problem. In iterative algorithms such as the MCEM algorithm, it is usually efficient to start with small sample size and increase it while parameters are approaching the MLE [Delyon et al., 1999], but there is no general rule for the choice of sampling sizes [Atanassov and Dimov, 2008].

To determine the required sample size in the emphasis method, we consider the estimator the distribution f_m.

$$f\left(x_{obs}|\theta\right) = \int_{x\in\mathcal{X}(x_{obs})} f(x,x_{obs}|\theta)\,dx \approx \frac{1}{M}\sum_{x_i\sim f_m}\frac{f(x,x_{obs}|\theta)}{f_m(x_i|x_{obs},\theta)} = f\,\widehat{(x_{obs}|\theta)}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

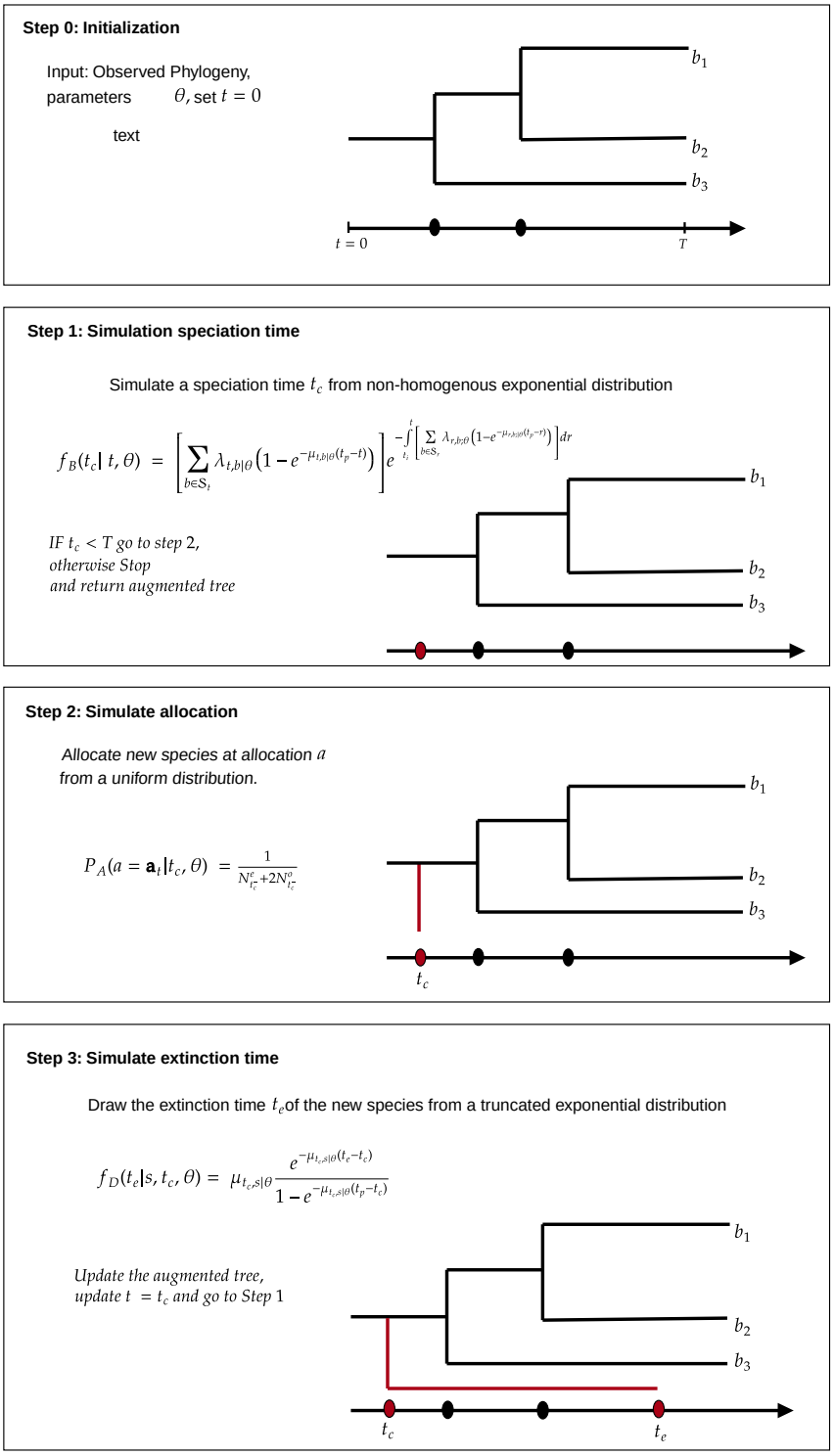


Figure 5: Tree augmentation algorithm based on the underlying non-homogeneous Poisson process.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where $\{x_1, ..., x_M\}$ are full trees sampled from $f_m((t,\tau)|x_{obs},\theta)$. We will assume that if $SE\left(\widehat{\ell(\theta)}\right) < C$ then $\widehat{\ell(\theta)}$ is good enough, for a small constant C. Note that, by taking a Taylor expansion of the logarithm of the estimated likelihood around the observed tree x_{obs}, we can write

$$\begin{aligned}\mathbb{E}\left[\log \ f\left(\widehat{x_{obs}|\theta}\right)\right] &\approx \mathbb{E}[\log f\left(x_{obs}|\theta\right) + \left(f\left(\widehat{x_{obs}|\theta}\right) - f\left(x_{obs}|\theta\right)\right)\frac{1}{f\left(x_{obs}|\theta\right)} \\ &\quad + \frac{1}{2}\left(f\left(\widehat{x_{obs}|\theta}\right) - f\left(x_{obs}|\theta\right)\right)^2 \frac{-1}{f^2\left(x_{obs}|\theta\right)}] \\ &= \ell(\theta) - \frac{1}{2}\frac{V\left(f(\widehat{x_{obs}|\theta})\right)}{f^2(x_{obs}|\theta)}\end{aligned}$$

where the last term represents the first-order bias. So, typically our method will underestimate the loglikelihood. Furthermore, the estimation tends to be variable. The variability can be assessed by a first order Taylor expansion, i.e.,

$$\begin{aligned}\mathbb{V}\left[\log \ f\left(\widehat{x_{obs}|\theta}\right)\right] &\approx \mathbb{V}\left[\log f\left(x_{obs}|\theta\right) + \left(f\left(\widehat{x_{obs}|\theta}\right) - f\left(x_{obs}|\theta\right)\right)\frac{1}{f\left(x_{obs}|\theta\right)}\right] \\ &= \frac{V\left(f\left(\widehat{x_{obs}|\theta}\right)\right)}{f^2\left(x_{obs}|\theta\right)}\end{aligned}$$

The variance of $f\left(\widehat{x_{obs}|\theta}\right)$ can be easily estimated by the sample variance of the importance weights divided by the sample size, i.e., $V\left(f\left(\widehat{x_{obs}|\theta}\right)\right) \approx V(w_1, ..., w_M)/M$. To assess the total possible deviation of the MC estimation we consider the bias and the standard error combined:

$$\widehat{\text{Deviation}} = \frac{1}{2}\frac{V(w_1, ..., w_M)}{M\left(f\left(\widehat{x_{obs}|\theta}\right)\right)^2} + \frac{\sqrt{V(w_1, ..., w_M)}}{\sqrt{M} f\left(\widehat{x_{obs}|\theta}\right)}$$

where the weights w_i are given by equation 6. The first term of the equation is an estimate of the bias and the second term an estimate of the standard error.

If it is feasible to perform a large number of trial simulations, then the standard error becomes of a lower order than the bias and, thus, we have that approximately,

$$\overline{\widehat{\ell(\theta)}} > \ell(\theta) - \widehat{\text{Bias}} \approx \ell(\theta) - \frac{K_1}{M}$$

for a constant value $K_1 = V(w_1, ..., w_M)/\left(2f^2\left(\widehat{x_{obs}|\theta}\right)\right)$, which can be further used to do a bias-correction of our likelihood. With this, we can calculate an approximation of the required sample size M to reach a desired level of accuracy. Figure 6 shows an illustration on the method we use to assess the required sample size. We sample trees with different sample sizes in order to obtain different estimates of the loglikelihood, and we can fit a curve of the form $c_1 + \frac{c_2}{M}$. With the fitted model we can calculate the asymptotic value of the loglikelihood c_1. We set the sample size M such that for a given tolerance level ϵ, $\frac{c_1}{M} < \epsilon$.

The MCEM algorithm can be replaced by the SAEM, MCMC or variations and combinations of them [Delyon et al., 1999, Celeux et al., 1995, Rydén et al., 2008, Wang, 2007, Kuhn and Lavielle, 2004]. All these algorithms rely on a sampling scheme and importance samplers. Our sampling scheme and sample size determination strategy can be used in any of these methods.

3.4 Model Selection

It is possible to apply standard model selection tools such as AIC or BIC [Wit et al., 2012] to the obtained loglikelihood. Furthermore, in the context of phylogenetic trees, specific statistics have been developed to test how well a model describes an observed tree. An informative summary statistic is the lineage-through-time (LTT) statistic [Janzen et al., 2015], defined as

$$LTT(1,2) = \int_{t_0}^{t_p} |N_t^{(1)} - N_t^{(2)}| dt$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

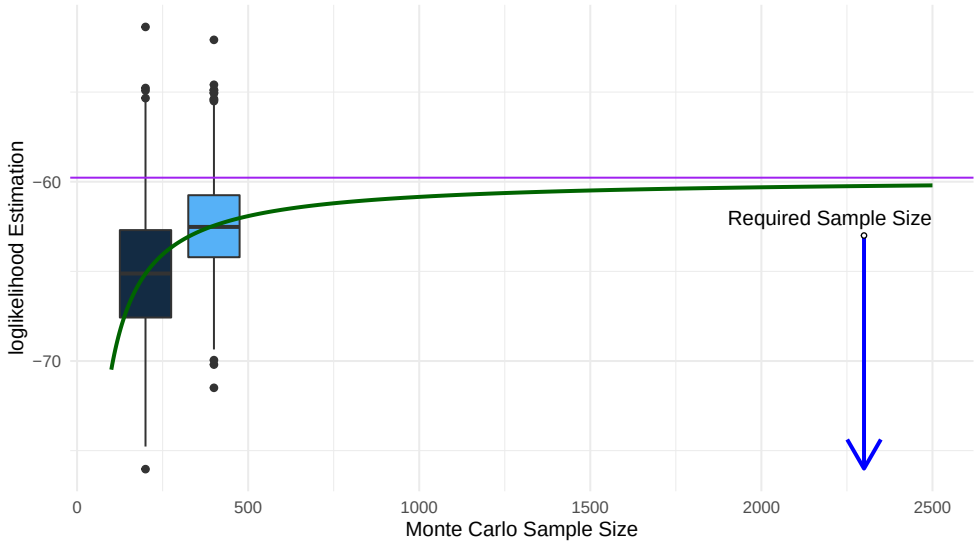


Figure 6: To calculate the required sample size, we simulate trees and estimate the loglikelihood via Monte-Carlo with at least two different sample sizes. We then calculate the curve that fits the relationship between the sample size and the estimated MC loglikelihood. This curve indicates the sample size required under a given tolerance level.

where $N_t^{(i)}$ is the number of species of a tree i at time t. This statistic can also be used to assess how well a model describes an observed tree, simulating trees from the desired model and then calculating the LTT statistic between each simulated tree and the observed tree. It is also possible to calculate the mean number of species through time into a single "average" tree and calculate the LTT statistic of that tree compared to the observed tree.

The LTT statistic and model 1 are mathematical expressions that take into account the branching times of the tree, but ignore the topology. That is, the parent-child relationship among species is not relevant; only the branching times are considered. In this manuscript, we introduce an alternative to the LTT statistic, considering phylogenetic diversity instead of species richness. We define the *phylodiversity-through-time* (PTT) statistic as

$$PTT(1,2) = \int_{t_0}^{t_p} |P_t^{(1)} - P_t^{(2)}| dt$$

where $P_t^{(i)}$ is the phylogenetic diversity for tree i at time t. In Figure 7 we present two example trees and the species richness for both trees as well as the phylogenetic diversity. The blue area represents the LTT statistic, while the green area represents the PTT statistic.

Here, we will consider the LTT statistic, the PTT statistic and the AIC weights for model comparison and general goodness-of-fit considerations.

4 Application

To illustrate our method, we quantitatively compare model (2) with model (1) for 14 phylogenies obtained from [Condamine et al., 2019], with sizes ranging between 16 and 141 species and crown ages between 5 My and 65 My. Figure 8 represents the distribution of the number of species and crown age of the clades.

In this application, the ultimate use of the emphasis framework is to quantify the impact of phylodiversity-dependent diversification by finding the maximum likelihood estimates for model (2) and compare them with model (1). But first, we perform some initial steps to evaluate the required sampling size for different phylogenetic trees. This will give insight about which phylogenies we can apply emphasis to and at what computational cost.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

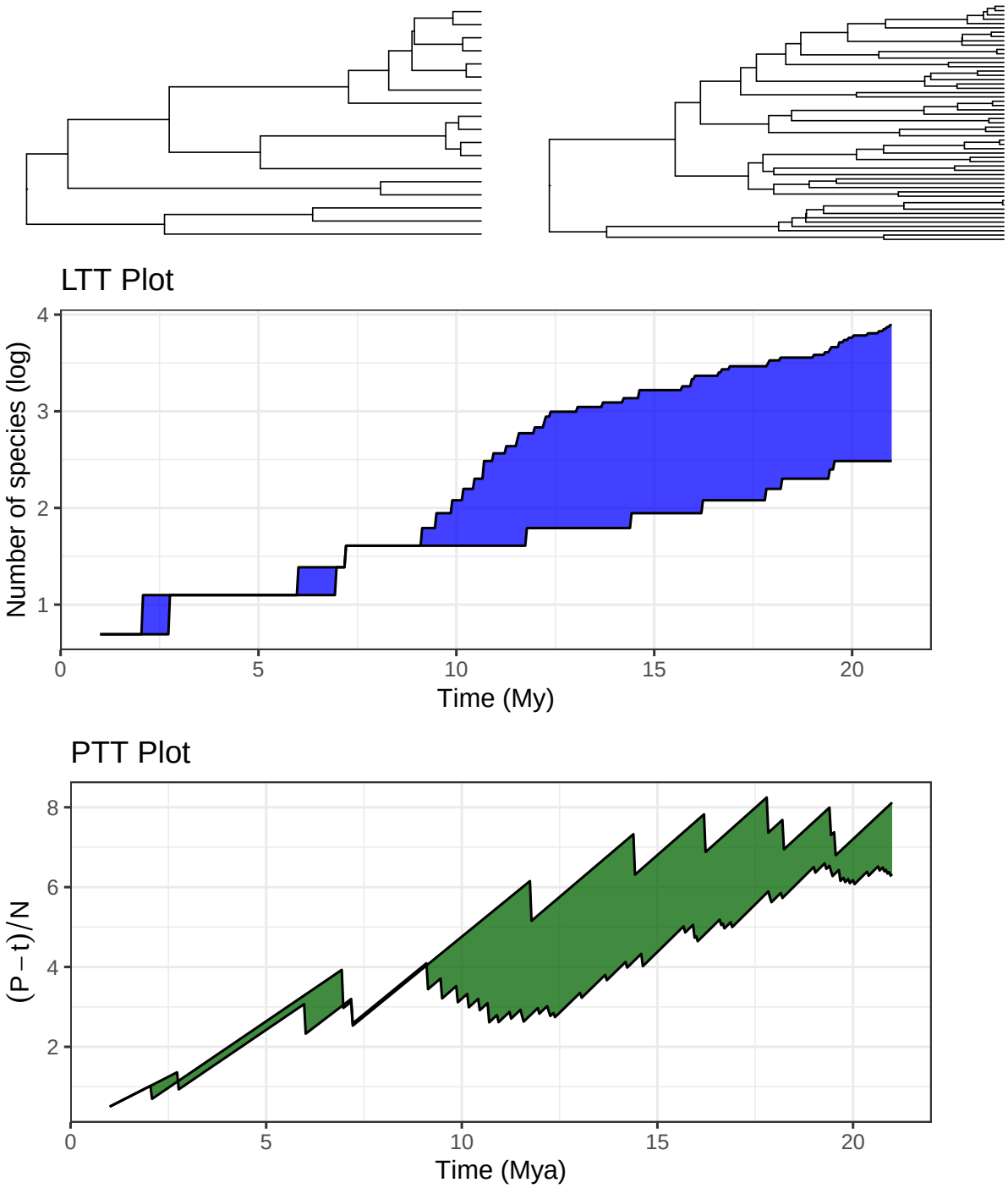
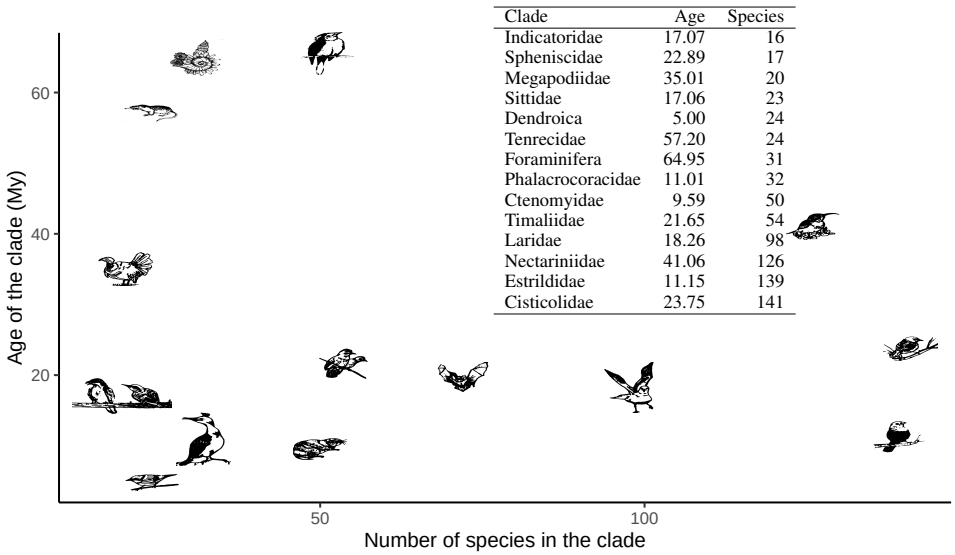


Figure 7: Comparison between two phylogenetic trees using the LTT (Number of lineages) and PTT (Phylogenetic diversity) through time. The area represents the distance between the trees.

4.1 Monte-Carlo approximation with the proposed importance sampler

Before performing an analysis with the model (2), we want to test the efficiency of the Monte-Carlo method with the importance sampler introduced in Section 3.3. Monte-Carlo methods require a sensible choice of the sample size, and this largely depends on the type of problem. For sampling full trees, the relationship between accuracy and sample size is complex. For the uniform importance sampler presented in Richter et al. [2020], the required MC sample size

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



Clade	Age	Species
Indicatoridae	17.07	16
Spheniscidae	22.89	17
Megapodiidae	35.01	20
Sittidae	17.06	23
Dendroica	5.00	24
Tenrecidae	57.20	24
Foraminifera	64.95	31
Phalacrocoracidae	11.01	32
Ctenomyidae	9.59	50
Timaliidae	21.65	54
Laridae	18.26	98
Nectariniidae	41.06	126
Estrildidae	11.15	139
Cisticolidae	23.75	141

Figure 8: Distribution of crown ages and numbers of species of 14 phylogenetic trees.

becomes huge for most empirical trees. We first show that the non-homogenous sampler presented here can provide accurate approximations.

Note that,

$$\begin{aligned}\mathbb{E}\left[f(x_{obs}|\theta)\right] &= \int_{x\in\mathcal{X}(x_{obs})} f(x|\theta)dx \\ &= \int_{x\in\mathcal{X}(x_{obs})} \frac{f(x|\theta)}{f_m(x|\theta,x_{obs})} f_m(x|\theta,x_{obs})dx \\ &\approx \frac{1}{M}\sum_{x_i\sim f_m(x|\theta,x_{obs})} \frac{f(x_i|\theta)}{f_m(x_i|\theta,x_{obs})} \end{aligned} \tag{14}$$

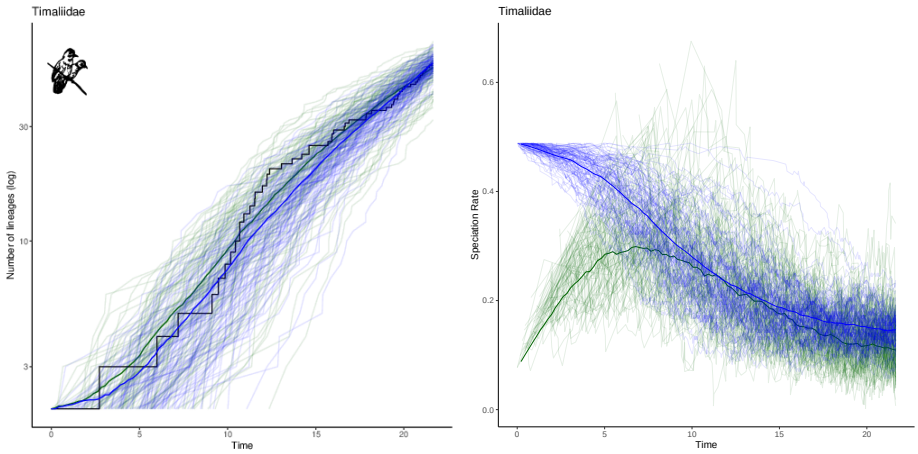
so we can use the Monte-Carlo sampling to approximate the likelihood for every parameter. To assess how well our importance sampler does as a function of MC sample size, we compare MC estimations for the 14 phylogenies for the LDD model, for which an existing solution exists. For each phylogeny, we calculate the MLE for the LDD model with the DDD R package. With these parameters, we perform Monte-Carlo sampling and approximate the expectation (14) with 4 different MC sampling sizes. In Table 1 we show the MC estimations and, in the last column, the analytical solution.

	10^2	10^3	10^4	10^5	10^6	Analytical
Indicatoridae	-42.04(1.1e-01)	-42.39(1.1e-01)	-41.97(4.9e-02)	-42.01(4.1e-02)	-41.95(4.8e-02)	-41.89
Spheniscidae	-50.22(3.4e-01)	-50.64(1.8e-01)	-50.49(1.1e-01)	-50.23(8.1e-02)	-50.35(4.3e-02)	-50.23
Megapodiidae	-68.98(6.1e-02)	-68.77(4.6e-02)	-68.3(1.9e-02)	-68.1(2.1e-02)	-68.09(1.0e-02)	-68
Sittidae	-64.72(3.8e-01)	-64.19(2.3e-01)	-64.42(1.0e-01)	-64.38(7.6e-02)	-64.32(3.0e-02)	-64.31
dendroica	-38.97(3.5e-01)	-39.2(2.3e-01)	-39.04(1.1e-01)	-38.97(6.0e-02)	-39.03(4.3e-02)	-38.91
Tenrecidae	-89.68(1.7e-01)	-89.12(6.0e-02)	-88.84(3.6e-02)	-89.05(2.1e-02)	-88.42(4.4e-02)	-88.4
foraminifera	-118.48(5.9e-02)	-117.7(4.3e-02)	-116.48(2.4e-02)	-116.34(1.9e-02)	-115.75(1.1e-02)	-115.73
Phalacrocoracidae	-79.92(4.6e-02)	-81.06(9.5e-02)	-80.42(5.7e-02)	-80.46(4.0e-02)	-80.4(2.0e-02)	-80.25
Ctenomyidae	-122.66(1.4e-01)	-120.8(1.1e-01)	-120.61(6.1e-02)	-120.7(4.0e-02)	-120.63(5.6e-02)	-120.65
Timaliidae	-154.23(4.7e-03)	-154.93(5.1e-03)	-153.75(3.0e-03)	-153.91(2.2e-03)	-153.56(9.8e-04)	-153.48
Laridae	-232.12(5.9e-03)	-232.15(3.1e-03)	-231.52(2.0e-03)	-231.23(1.1e-03)	-231.14(8.7e-04)	-231.07
Nectariniidae	-416.2(6.3e-04)	-410.81(2.7e-05)	-404.61(2.0e-06)	-402.19(7.4e-07)	-400.74(3.3e-07)	-399.04
Estrildidae	-321.31(5.4e-03)	-316.31(1.7e-04)	-311.26(2.7e-05)	-312.22(2.8e-05)	-310.93(1.4e-05)	-309.35
Cisticolidae	-410.16(2.5e-04)	-403.73(2.3e-05)	-402.29(1.0e-05)	-402.06(7.3e-06)	-400.6(4.0e-06)	-397.94

Table 1: Monte-Carlo approximation of the loglikelihood for each tree at its corresponding MLE for the diversification process generated under the LDD model, for different sample sizes. The last column contains the analytical value obtained with the R package DDD.

If the difference between the analytical loglikelihood and the MC approximated loglikelihood is less than 1, we will conclude that the estimation is good enough, following the AIC principle when comparing a model with three parameters

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.



(a) Lineages through time (b) Speciation rate through time

Figure 9: Evolution of extant species richness (LTT-plot) and evolution of global speciation rates for 13 clades under LDD (blue) and LPD (green) models.

agains a model with four parameters. Under this assumption, we found that a sample size of 1000 is good enough for phylogenies up to approximately 70 species. With a sample size of 10^6 we have very accurate estimations for all clades with the exceptions of Nectariniidae, Estrildidae and Cisticolidae. These are the larger trees with more than 100 species. Table 1 contains detailed estimation for 4 different sample sizes for each phylogeny.

4.2 Estimation and model selection

In Table 2 we report the parameter estimations for the two models of interest for the 14 case-study phylogenies. Looking at the LTT statistics, we observe that there is only a slight improvement of the LPDD model over the LDD model in most of the cases. This is confirmed by the loglikelihood values, where the improvement is not large enough to justify preferring the LPDD model, which is confirmed with the AIC weights which always prefer the LDD model.

Note that the AIC weights are based on a Monte-Carlo approximation of the likelihood which will slightly underestimate models with more parameters. As a result, the AIC is more conservative than a test where the AIC values are calculated using the true value of the loglikelihood. Note that in some cases the loglikelihood for the LPDD model is still smaller than the loglikelihood of the LDD model, which cannot be correct because the LDD model is nested within the LPDD model, and hence the LPDD likelihood should always be smaller than the LPDD likelihood. We argue that this is because the Monte-Carlo approximation is not good enough yet. These computational issues suggest that hypothesis testing with AIC might not be an appropriate tool for model selection. Significance tests, instead, do not depend on the approximation of the likelihood but on the approximation of the Hessian of the likelihood (see equation **??**); because the likelihood is asymptotically quadratic near its maximum (hence the second derivative is constant), the approximation of the Hessian should not present the computational issues that the approximation of the likelihood presents, and hence significance tests seem more reliable. Based on the significance test results, we conclude that the phylodiversity-dependent diversification model provides an alternative/better explanation to/than the diversity-dependent diversification model, at least in some of our clades.

In Figure 9 we see an example of the expected lineages-through-time plot for each model in comparison with the observed lineage through time plot corresponding to the Timaliidae phylogeny, and the speciation rates through time plot. We can see that both models agree that speciation happened roughly at a rate of 0.2 species per million rate during the last 10 million years; however, they diverge on the estimates for the period between 20 and 10 million years ago. Including phylogenetic diversity involves a fluctuating speciation rate around 0.2 spe/Mye reaching its maximum around 15 years ago while the LDD model assumes a monotonously decreasing speciation rate. In general, the difference between the two models is not large and this pattern is present across all the 14 phylogenies.

Finally, in Table 3 we report the loglikelihood estimates for the LPDD model. We see that for most of the cases the sample size was large enough, but for larger trees the convergence performs much slower for the LPDD case than for the LDD case.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Clade	Age	Tips	Model	AICw	LTT	PTT	loglikelihood	μ_0	λ_0	β_N	β_P
Indicatoridae	17.07	16	LDD	0.73	0.49	0.49	-42.01	0.22	1.62	-0.085	0
			LPDD	0.27	0.51	0.51	-41.99 (6e-02)	0.21 (4e-03)	1.61 (5e-04)	-0.085 (5e-05)	-0.001 (3e-04)
Spheniscidae	22.89	17	LDD	0.83	0.41	0.52	-50.23	0.2	1.61	-0.081	0
			LPDD	0.17	0.59	0.48	-50.79 (3e-02)	0.16 (2e-03)	1.49 (1e-02)	-0.079 (6e-04)	0.004 (7e-04)
Megapodiidae	35.01	20	LDD	0.78	0.45	0.48	-68.1	0.1	0.83	-0.036	0
			LPDD	0.22	0.55	0.52	-68.37 (3e-02)	0.09 (1e-03)	0.83 (6e-04)	-0.036 (4e-05)	0 (1e-04)
Sittidae	17.06	23	LDD	0.79	0.5	0.46	-64.38	0.15	0.58	-0.018	0
			LPDD	0.21	0.5	0.54	-64.72 (1e-01)	0.12 (9e-04)	0.4 (1e-03)	-0.021 (3e-04)	0.039 (1e-03)
Dendroica	5.00	24	LDD	0.77	0.46	0.53	-38.97	0.16	3.05	-0.117	0
			LPDD	0.23	0.54	0.47	-39.16 (8e-02)	0.14 (1e-03)	2.99 (2e-02)	-0.118 (1e-03)	0.007 (3e-03)
Tenrecidae	57.20	24	LDD	0.74	0.46	0.47	-89.05	0.11	0.59	-0.02	0
			LPDD	0.26	0.54	0.53	-89.1 (3e-02)	0.09 (6e-04)	0.59 (4e-04)	-0.02 (8e-05)	0.001 (2e-04)
Foraminifera	64.95	31	LDD	0.84	0.39	0.42	-116.34	0.1	1.18	-0.034	0
			LPDD	0.16	0.61	0.58	-117.01 (4e-03)	0.08 (4e-04)	1.17 (1e-04)	-0.034 (2e-06)	0 (4e-06)
Phalacrocoracidae	11.01	32	LDD	0.71	0.41	0.48	-80.46	0.24	1.67	-0.044	0
			LPDD	0.29	0.59	0.52	-80.34 (4e-02)	0.24 (3e-03)	1.59 (2e-02)	-0.038 (3e-04)	-0.027 (3e-03)
Ctenomyidae	9.59	50	LDD	0.74	0.46	0.42	-120.7	0.16	1.15	-0.02	0
			LPDD	0.26	0.54	0.58	-120.75 (6e-02)	0.14 (1e-03)	1.08 (6e-03)	-0.019 (2e-04)	0.007 (2e-03)
Timaliidae	21.65	54	LDD	1	0.48	0.53	-153.91	0.14	0.5	-0.006	0
			LPDD	0	0.52	0.47	-158.63 (7e-03)	0.07 (1e-03)	0.22 (8e-03)	-0.006 (2e-04)	0.034 (2e-03)
Laridae	18.26	98	LDD	0.68	0.19	0.5	-231.23	0.13	0.32	0	0
			LPDD	0.32	0.81	0.5	-231.01 (6e-02)	0.02 (1e-03)	0.46 (2e-03)	0.001 (2e-05)	-0.061 (7e-04)
Nectariniidae	41.06	126	LDD	0.66	0.54	0.83	-402.19	0.14	0.32	-0.001	0
			LPDD	0.34	0.46	0.17	-401.83 (4e-02)	0.02 (4e-04)	0.25 (1e-03)	0 (2e-05)	-0.019 (3e-04)
Estrildidae	11.15	139	LDD	0.88	0.5	0.77	-312.22	0.28	1.05	-0.005	0
			LPDD	0.12	0.5	0.23	-313.21 (6e-03)	0.12 (2e-03)	0.42 (7e-03)	-0.006 (3e-04)	0.176 (1e-02)
Cisticolidae	23.75	141	LDD	0.51	0.48	0.76	-402.06	0.16	0.48	-0.002	0
			LPDD	0.49	0.52	0.24	-401.11 (2e-02)	0.05 (1e-03)	0.37 (4e-03)	0 (5e-05)	-0.04 (1e-03)

Table 2: Parameter estimations for LDD and LPDD model for 14 phylogenies. The fifth column shows the AIC weights for the comparison of these two models. The sixth column is the normalised LTT statistic. The last four columns represent the parameter estimates. Between parentheses we report the standard deviation of the Monte-Carlo approximation.

	10^2	10^3	10^4	10^5	10^6
Indicatoridae	-42.49(1.3e-01)	-42.15(1.0e-01)	-42.26(6.5e-02)	-41.99(6.1e-02)	-42.02(2.7e-02)
Spheniscidae	-51.66(2.4e-01)	-50.68(1.1e-01)	-50.9(6.4e-02)	-50.79(3.0e-02)	-50.75(1.6e-02)
Megapodiidae	-68.97(2.1e-01)	-68.19(8.1e-02)	-68.42(6.2e-02)	-68.37(3.5e-02)	-68.26(4.0e-02)
Sittidae	-65.49(2.7e-01)	-64.49(2.0e-01)	-64.83(1.2e-01)	-64.72(1.2e-01)	-64.73(2.4e-02)
dendroica	-39.25(2.9e-01)	-39.33(1.9e-01)	-39.22(9.4e-02)	-39.16(8.0e-02)	-39.14(3.2e-02)
Tenrecidae	-89.06(8.6e-02)	-88.65(6.0e-02)	-89.61(3.9e-02)	-89.1(3.4e-02)	-89.12(1.5e-02)
foraminifera	-119.3(1.2e-02)	-117.71(5.1e-03)	-117.94(3.4e-03)	-117.01(4.5e-03)	-117.3(2.3e-03)
Phalacrocoracidae	-79.61(1.2e-01)	-81.02(8.5e-02)	-80.54(6.5e-02)	-80.34(4.1e-02)	-80.56(2.1e-02)
Ctenomyidae	-119.95(1.8e-01)	-121.08(2.0e-01)	-120.87(8.2e-02)	-120.75(5.6e-02)	-120.77(7.6e-02)
Timaliidae	-159.57(4.8e-02)	-158.95(2.0e-02)	-158.37(1.4e-02)	-158.63(6.8e-03)	-158.16(6.9e-03)
Laridae	-230.98(6.3e-01)	-230.86(2.2e-01)	-231.03(1.0e-01)	-231.01(5.6e-02)	-231(3.8e-02)
Nectariniidae	-402.09(1.9e-01)	-402.01(9.6e-02)	-401.81(6.9e-02)	-401.83(3.9e-02)	-401.87(3.0e-02)
Estrildidae	-315.33(3.0e-02)	-313.25(1.3e-02)	-313.13(1.1e-02)	-313.21(6.0e-03)	-313.28(3.7e-03)
Cisticolidae	-403.78(5.9e-02)	-401.9(3.6e-02)	-401.52(2.4e-02)	-401.11(1.7e-02)	-401(8.2e-03)

Table 3: Loglikelihood approximations of the LPD model at its MLE value for the 14 phylogenies.

5 Discussion

Diversity-dependent diversification models have been developed during the last decade in order to understand and quantify the existence and impact of ecological limits to macroevolutionary dynamics. At the moment, only models with a dependence of diversification rates on species richness have been implemented, but these models ignore other facets of diversity, such as phylodiversity.

Here, we have completed the statistical methodology introduced in Richter et al. [2020], with the design of a data augmentation scheme that provides an efficient importance sampler. This is a substantial improvement in comparison to the uniform importance sampler considered in Richter et al. [2020], as it enables applying the method to a large number of empirical phylogenies.

In the application to 14 example phylogenies, we studied the LPDD model, i.e., a model with a linear effect of phylodiversity on speciation. We found that including phylodiversity does not provide a substantial improvement in comparison with richness-dependent diversification models. However, phylodiversity does provide an alternative and slightly more complete explanation to speciation dynamics; the LTT statistic and the PTT statistic provide insights and, most of the times, reflect that trees generated by the LPDD model are closer to real phylogenies than trees generated

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

under the LDD model. While the model with fewer parameters is preferred using AIC, the phylodiversity component is statistically significant, suggesting that it should not be ignored.

This may not be the final word because there are some technical improvements to be made. In particular, we did not condition the likelihood on non-extinction of the clade; even though this is generally recommended [Etienne et al., 2016, Stadler, 2013].

Our method is not limited to phylogenetic diversity-dependent diversification models, but allows inference of a general class of species diversification models, considering time, traits, climate, functional diversity, just to name a few. With the data augmentation described her we have provided a general tool that can be potentially used to quantify and test a large number of hypotheses in macroevolutionary diversification.

6 Acknowledgements

This publication is part of the project *Killing two birds with one stone: simultaneous estimation and selection of species diversification models* (with project number 657.014.005) of the research programme *Mathematics for planet Earth (MPE)* which is (partly) financed by the Dutch Research Council (NWO).

This article is based upon work from COST Action "European Cooperation for Statistics of Network Data Science" (CA15109), supported by COST (European Cooperation in Science and Technology).

We also acknowledge funding from the Swiss National Science Foundation, project 200021_188534 entitled Sparse inference of complex networks.

References

Timothy D Walker and James W Valentine. Equilibrium models of evolutionary species diversity and the number of empty niches. *The American Naturalist*, 124(6):887–899, 1984.

Stephen Jay Gould, David M Raup, J John Sepkoski Jr, Thomas JM Schopf, and Daniel S Simberloff. The shape of evolution: a comparison of real and random clades. *Paleobiology*, pages 23–40, 1977.

Daniel L Rabosky. Ecological limits on clade diversification in higher taxa. *The American Naturalist*, 173(5):662–674, 2009.

Fabien L Condamine. Limited by the roof of the world: mountain radiations of apollo swallowtails controlled by diversity-dependence processes. *Biology letters*, 14(3):20170622, 2018.

Gillian C Gibb, Fabien L Condamine, Melanie Kuch, Jacob Enk, Nadia Moraes-Barros, Mariella Superina, Hendrik N Poinar, and Frédéric Delsuc. Shotgun mitogenomics provides a reference phylogenetic framework and timescale for living xenarthrans. *Molecular Biology and evolution*, 33(3):621–642, 2016.

Regina L Cunha, Cláudia Patrão, and Rita Castilho. Different diversity-dependent declines in speciation rate unbalances species richness in terrestrial slugs. *Scientific reports*, 7(1):16198, 2017.

Charles Pouchon, Angel Fernández, Jafet M Nassar, Frédéric Boyer, Serge Aubert, Sébastien Lavergne, and Jesús Mavárez. Phylogenomic analysis of the explosive adaptive radiation of the espeletia complex (asteraceae) in the tropical andes. *Systematic Biology*, 67(6):1041–1060, 2018.

Xin Chen, Alan R Lemmon, Emily Moriarty Lemmon, R Alexander Pyron, and Frank T Burbrink. Using phylogenomics to understand the link between biogeographic origins and regional diversification in ratsnakes. *Molecular phylogenetics and evolution*, 111:206–218, 2017.

Jesús N Pinto-Ledezma, Lorena Mendes Simon, José Alexandre F Diniz-Filho, and Fabricio Villalobos. The geographical diversification of furnariides: the role of forest versus open habitats in driving species richness gradients. *Journal of biogeography*, 44(8):1683–1693, 2017.

Jimmy A McGuire, Christopher C Witt, JV Remsen Jr, Ammon Corl, Daniel L Rabosky, Douglas L Altshuler, and Robert Dudley. Molecular phylogenetics and the diversification of hummingbirds. *Current Biology*, 24(8):910–916, 2014.

R Alexander Pyron and John J Wiens. Large-scale phylogenetic analyses reveal the causes of high tropical amphibian diversity. *Proceedings of the Royal Society B: Biological Sciences*, 280(1770):20131622, 2013.

Liang Xu and Rampal S Etienne. Detecting local diversity-dependence in diversification. *Evolution*, 72(6):1294–1305, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Rampal S Etienne, Alex L Pigot, and Albert B Phillimore. How reliably can we infer diversity-dependent diversification from phylogenies? *Methods in Ecology and Evolution*, 7(9):1092–1099, 2016.

Lee Hsiang Liow, Tiago B Quental, and Charles R Marshall. When can decreasing diversification rates be detected with molecular phylogenies and the fossil record? *Systematic Biology*, 59(6):646–659, 2010.

Leonel Herrera-Alsina, Alex L Pigot, Hanno Hildenbrandt, and Rampal S Etienne. The influence of ecological and geographic limits on the evolution of species distributions and diversity. *Evolution*, 72(10):1978–1991, 2018.

Hélène Morlon. Phylogenetic approaches for studying diversification. *Ecology letters*, 17(4):508–525, 2014.

Daniel L Rabosky and Allen H Hurlbert. Species richness at continental scales is dominated by ecological limits. *The American Naturalist*, 185(5):572–583, 2015.

Knud A Jønsson, Pierre-Henri Fabre, Susanne A Fritz, Rampal S Etienne, Robert E Ricklefs, Tobias B Jørgensen, Jon Fjeldså, Carsten Rahbek, Per GP Ericson, Friederike Woog, et al. Ecological and evolutionary determinants for the adaptive radiation of the madagascan vangas. *Proceedings of the National Academy of Sciences*, 109(17):6620–6625, 2012.

Rampal S Etienne, Bart Haegeman, Tanja Stadler, Tracy Aze, Paul N Pearson, Andy Purvis, and Albert B Phillimore. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proc. R. Soc. B*, page rspb20111439, 2012a.

Matthew M Kling, Brent D Mishler, Andrew H Thornhill, Bruce G Baldwin, and David D Ackerly. Facets of phylodiversity: evolutionary diversification, divergence and survival as conservation targets. *Philosophical Transactions of the Royal Society B*, 374(1763):20170397, 2018.

Samuel Scheiner, Evsey Kosman, Steven Presley, and Michael Willig. The components of biodiversity, with a particular focus on phylogenetic information. *Ecology and Evolution*, 7:6444–6454, 07 2017. doi:10.1002/ece3.3199.

Tania Laity, Shawn W Laffan, Carlos E González-Orozco, Daniel P Faith, Dan F Rosauer, Margaret Byrne, Joseph T Miller, Darren Crayn, Craig Costion, Craig C Moritz, et al. Phylodiversity to inform conservation policy: An australian example. *Science of the Total Environment*, 534:131–143, 2015.

Daniel P Faith and Andrew M Baker. Phylogenetic diversity (pd) and biodiversity conservation: some bioinformatics challenges. *Evolutionary bioinformatics*, 2:117693430600200007, 2006.

JL Cantalapiedra, T Aze, MW Cadotte, GV Dalla Riva, D Huang, F Mazel, MW Pennell, M Ríos, and AØ Mooers. Conserving evolutionary history does not result in greater diversity over geological time scales. *Proceedings of the Royal Society B*, 286(1904):20182896, 2019.

Florent Mazel, Matthew W Pennell, Marc W Cadotte, Sandra Diaz, Giulio Valentino Dalla Riva, Richard Grenyer, Fabien Leprieur, Arne O Mooers, David Mouillot, Caroline M Tucker, et al. Prioritizing phylogenetic diversity captures functional diversity unreliably. *Nature communications*, 9(1):2888, 2018.

Jutta Stadler, Stefan Klotz, Roland Brandl, and Sonja Knapp. Species richness and phylogenetic structure in plant communities: 20 years of succession. *Web Ecology*, 17(2):37–46, 2017.

Caroline Tucker, Marc Cadotte, Sílvia Carvalho, Jonathan Davies, Simon Ferrier, Susanne Fritz, Rich Grenyer, Matthew Helmus, Lanna Jin, Arne Mooers, Sandrine Pavoine, Oliver Purschke, David Redding, Dan Rosauer, Marten Winter, and Florent Mazel. A guide to phylogenetic metrics for conservation, community ecology and macroecology. *Biological Reviews*, 92:n/a–n/a, 01 2016. doi:10.1111/brv.12252.

Campbell O Webb, Gregory S Gilbert, and Michael J Donoghue. Phylodiversity-dependent seedling mortality, size structure, and disease in a bornean rain forest. *Ecology*, 87(sp7):S123–S131, 2006.

Cyrille Violle, Diana R Nemergut, Zhichao Pu, and Lin Jiang. Phylogenetic limiting similarity and competitive exclusion. *Ecology letters*, 14(8):782–787, 2011.

Gabriel C Costa, Daniel O Mesquita, Guarino R Colli, and Laurie J Vitt. Niche expansion and the niche variation hypothesis: does the degree of individual variation increase in depauperate assemblages? *The American Naturalist*, 172(6):868–877, 2008.

Bradford C Lister. The nature of niche expansion in West Indian Anolis lizards I: ecological consequences of reduced competition. *Evolution*, pages 659–676, 1976.

Janne Soininen, Jani Heino, Jyrki Lappalainen, and Risto Virtanen. Expanding the ecological niche approach: Relationships between variability in niche position and species richness. *Ecological Complexity*, 8(1):130–137, 2011.

Charles R Marshall and Tiago B Quental. The uncertain role of diversity dependence in species diversification and the need to incorporate time-varying carrying capacities. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1691):20150217, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Giovanni Laudanno, Bart Haegeman, and Rampal S Etienne. Additional analytical support for a new method to compute the likelihood of diversification models. *Bulletin of mathematical biology*, 82(2):22, 2020.

Daniel P Faith. Conservation evaluation and phylogenetic diversity. *Biological conservation*, 61(1):1–10, 1992.

Francisco Richter, Bart Haegeman, Rampal S. Etienne, and Ernst C. Wit. Introducing a general class of species diversification models for phylogenetic trees. *Statistica Neerlandica*, n/a(n/a):1–14, 2020.

Sean Nee, Robert M May, and Paul H Harvey. The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 344(1309):305–311, 1994.

Oskar Hagen, Klaas Hartmann, Mike Steel, and Tanja Stadler. Age-dependent speciation can explain the shape of empirical phylogenies. *Systematic biology*, 64(3):432–440, 2015.

Patrice Descombes, Theo Gaboriau, Camille Albouy, Christian Heine, Fabien Leprieur, and Lo\"\ic Pellissier. Linking species diversification to palaeo-environmental changes: A process-based modelling approach. *Global Ecology and Biogeography*, 27(2):233–244, 2018.

Emma E. Goldberg, Lesley T. Lancaster, and Richard H. Ree. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Systematic Biology*, 60(4):451–465, 2011. ISSN 10635157. doi:10.1093/sysbio/syr046.

Fabien L Condamine, Jonathan Rolland, and Hélène Morlon. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. *Ecology letters*, 22(11):1900–1912, 2019.

Rampal S Etienne, Bart Haegeman, Tanja Stadler, Tracy Aze, Paul N Pearson, Andy Purvis, and Albert B Phillimore. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society B: Biological Sciences*, 279(1732):1300–1309, 2012b.

Rampal S Etienne and Bart Haegeman. A Conceptual and Statistical Framework for Adaptive Radiations with a Key Role for Diversity Dependence. 180(4), 2012. doi:10.1086/667574.

Michael Foote, Roger A Cooper, James S Crampton, Peter M Sadler, and Michael Foote. Diversity-dependent evolutionary rates in early Palaeozoic zooplankton. (iii):11–14, 2018.

Philippe Jarne, Michel Loreau, Nicolas Mouquet, Patrice David, and Vincent Calcagno. Diversity spurs diversification in ecological communities. *Nature Communications*, 8(May):1–9, 2017. doi:10.1038/ncomms15810. URL `http://dx.doi.org/10.1038/ncomms15810`.

Marcus J Hamilton, Robert S Walker, and Christopher P Kempes. Diversity begets diversity in mammal species and human cultures. *Scientific reports*, 10(1):1–11, 2020.

Paschalia Kapli, Ziheng Yang, and Maximilian J Telford. Phylogenetic tree building in the genomic age. *Nature Reviews Genetics*, pages 1–17, 2020.

Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007.

Annette J Dobson and Adrian Barnett. *An introduction to generalized linear models*. CRC press, 2008.

A.P. Arthur P Dempster, Nan M N.M. Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 39(1):1–38, 1977. ISSN 00359246. doi:http://dx.doi.org/10.2307/2984875. URL `http://www.jstor.org/stable/10.2307/2984875`.

Martin A Tanner and Wing Hung Wong. The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82(398):528–540, 1987.

KS Chan and Johannes Ledolter. Monte carlo em estimation for time series models involving counts. *Journal of the American Statistical Association*, 90(429):242–252, 1995.

Peter W Glynn and Donald L Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.

Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

David A Van Dyk and Xiao-Li Meng. The art of data augmentation. *Journal of Computational and Graphical Statistics*, 10(1):1–50, 2001.

Jerome H Friedman. On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77, 1997.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.01.450729; this version posted July 2, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

David G. Kendall. On the Generalized "Birth-and-Death" Process. *The Annals of Mathematical Statistics*, 19(1):1–15, 1948. ISSN 0003-4851. doi:10.1214/aoms/1177730285.

Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434, 1976.

Le Minh Kieu. Analytical modelling of point process and application to transportation. *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, pages 385–408, 2018.

Bernard Delyon, Marc Lavielle, Eric Moulines, et al. Convergence of a stochastic approximation version of the em algorithm. *The Annals of Statistics*, 27(1):94–128, 1999.

Emanouil Atanassov and Ivan T Dimov. What monte carlo models can do and cannot do efficiently? *Applied Mathematical Modelling*, 32(8):1477–1500, 2008.

Gilles Celeux, Didier Chauveau, and Jean Diebolt. On stochastic versions of the EM algorithm. 1995.

Tobias Rydén et al. Em versus markov chain monte carlo for estimation of hidden markov models: A computational perspective. *Bayesian Analysis*, 3(4):659–688, 2008.

Jing Wang. Em algorithms for nonlinear mixed effects models. *Computational statistics & data analysis*, 51(6): 3244–3256, 2007.

Estelle Kuhn and Marc Lavielle. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131, 2004.

Ernst Wit, Edwin van den Heuvel, and Jan-Willem Romeijn. 'all models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica*, 66(3):217–236, 2012.

Thijs Janzen, Sebastian Höhna, and Rampal S Etienne. Approximate bayesian computation of diversification rates from molecular phylogenies: introducing a new efficient summary statistic, the nltt. *Methods in Ecology and Evolution*, 6 (5):566–575, 2015.

Tanja Stadler. How can we improve accuracy of macroevolutionary rate estimates? *Systematic Biology*, 62(2):321–329, 2013. ISSN 10635157. doi:10.1093/sysbio/sys073.