What fruits can we get from this tree?

Laudanno, Giovanni

DOI:
10.33612/diss.155031292

Chapter 2

Additional analytical support for a new method to compute the likelihood of diversification models

G. Laudanno, B. Haegeman, R. S. Etienne
Bulletin of Mathematical Biology, 2020

Abstract

Molecular phylogenies have been increasingly recognized as an important source of information on species diversification. For many models of macroevolution, analytical likelihood formulas have been derived to infer macroevolutionary parameters from phylogenies. A few years ago, a general framework to numerically compute such likelihood formulas was proposed, which accommodates models that allow speciation and/or extinction rates to depend on diversity. This framework calculates the likelihood as the probability of the diversification process being consistent with the phylogeny from the root to the tips. However, while some readers found the framework presented in Etienne et al. (2012) convincing, others still questioned it (personal communication), despite *numerical* evidence that for special cases the framework yields the same (i.e. within double precision) numerical value for the likelihood as analytical formulas do that were independently derived for these special cases. Here we prove *analytically* that the likelihoods calculated in the new framework are correct for all special cases with known analytical likelihood formula. Our results thus add substantial mathematical support for the overall coherence of the general framework.

2.1 Introduction

One of the major challenges in the field of macro-evolution is understanding the mechanisms underlying patterns of diversity and diversification. A very fruitful approach has been to model macro-evolution as a birth-death process which reduces the problem to the specification of macroevolutionary events (i.e. speciation and extinction). However, providing likelihood expressions for these models given empirical data on speciation and extinction events is quite challenging, for the following reason. While such a likelihood is very easy to derive when full information is available for all events, typically the data involves phylogenetic trees constructed with molecular data collected from extant species alone. Hence, no extinction events and speciation events leading to extinct species are recorded in such phylogenetic trees. For a variety of models this problem can be overcome by considering a reconstructed process, whereby the phylogeny of extant species can be regarded as a pure-birth process with time-dependent speciation rate (Nee, May, and Harvey, 1994). But this approach is not generally valid.

Thus, the methods employed to derive likelihood expressions are usually applicable to a limited set of models. They do not apply to models that assume that speciation and extinction rates depend on the number of species in the system. Hence, potential feedback of diversity itself on diversification rates, due to interspecific competition or niche filling, is completely ignored. The first to incorporate such feedbacks were Rabosky and Lovette (2008), who made rates dependent on the number of species present at every given moment in time, analogously to logistic growth models used in population biology. However, their model had to assume that there is no extinction for mathematical tractability, which stands in stark contrast to the empirical data: the fossil record provides us with many examples of extinct species.

Etienne et al. (2012) presented a framework to compute the likelihood of phylogenetic branching times under a diversity-dependent diversification process that explicitly accounts for the influence of species that are not in the phylogeny, because they have become extinct. We note that diversity-dependence as implemented in the approach of Etienne et al. (2012) does not need to start at the crown of a branching process: it can already act earlier. This feature has already been used in applications to island biogeography (Valente, Phillimore, and Etienne, 2015). Some of our colleagues have doubts that this framework contains a formal argument that the solution of the set of ordinary differential equations that together constitute the framework gives the likelihood of the model for a given phylogenetic tree. Instead, only numerical evidence for a small set of parameter combina-

tions has been provided that the method yields, in the appropriate limit, the known likelihood for the standard diversity-independent (i.e. using constant-rates) birth-death model. This likelihood was first provided by Nee, May, and Harvey (1994), using a breaking-the-tree approach. Later, Lambert and Stadler (2013) used coalescent point process theory to provide an approach to obtain likelihood formulas for a wider set of models. These models did not include diversity-dependence. For example, Lambert, Morlon, and Etienne (2015) applied their framework to the protracted birth-death model (Etienne, Morlon, and Lambert, 2014), which is a generalization of the diversity-independent model where speciation is no longer an instantaneous event (Etienne and Rosindell, 2012). For this model they provided an explicit likelihood expression.

Here we provide an analytical proof that the likelihood of Etienne et al. (2012) reduces to the likelihood of Lambert, Morlon, and Etienne (2015) – and hence to that of Nee, May, and Harvey (1994) – for the case of diversity-independent diversification.

The extant species belonging to a clade are often not all available for sequencing, because some species are difficult to obtain tissue from (either because they are difficult to find/catch, or because they are endangered, or because they have recently become extinct due to anthropogenic rather than natural causes) or because it is difficult to extract their DNA. This means that our data consists of a phylogenetic tree of an incomplete sample of species, and thus of an incomplete set of speciation events, even for those that lead to the species that we observe today. This incomplete sampling has been described by two different random models. The first model assumes that a fixed number of extant species are not represented in the phylogeny. This model might be appropriate for well-described taxonomic groups, such as birds, where we have a good idea of the species that are evolutionarily related, but we are simply missing some data points for the reasons mentioned above. This sampling model is called n-sampling (Lambert, Morlon, and Etienne, 2015). The second model assumes that extant species are represented in the phylogeny with a fixed probability ρ. This sampling scheme is called ρ-sampling (Lambert, Morlon, and Etienne, 2015), but is also referred to as f-sampling (Nee, May, and Harvey, 1994). The framework of Etienne et al. (2012) assumes n-sampling, but in this paper we show that it can also be extended to incorporate ρ-sampling.

In the next section we summarize the framework of Etienne et al. (2012) and we provide the likelihood formula analytically derived by Lambert, Morlon, and Etienne (2015) for the special case of diversity-independent but time-dependent diversification with n-sampling. Then we proceed by showing that the probabil-

ity generating functions of these two likelihoods are identical. We end with a discussion where we point out how the framework of Etienne et al. (2012) can be extended to include ρ-sampling and how it relates to the likelihood formula of Rabosky and Lovette (2008) for the diversity-dependent birth-death model without extinction.

2.2 The diversity-dependent diversification model

Diversification models are birth-death processes in which "birth" and "death" correspond to speciation and extinction events, respectively. In the simplest case, the per-species speciation rate λ and the per-species extinction rate μ are constants. Here we consider diversification models in which the per-species speciation and extinction rates depend on the number of species n present at time t, i.e., diversity-dependent, which we denote by λ_n and μ_n. We also allow the speciation and extinction rates to depend on time t, i.e., $\lambda_n(t)$ and $\mu_n(t)$, although the latter dependence is often not explicit in our notation.

We assume that the diversification process starts at time t_c from a crown, i.e., from two ancestor species. Assuming that at a later time $t > t_c$ the process has n species, the transition probabilities in the infinitesimal time interval $[t, t+dt]$ are

from n to $n+1$ species	with probability $n\lambda_n(t)\,dt$
from n to $n-1$ species	with probability $n\mu_n(t)\,dt$
n does not change	with probability $1-n\lambda_n(t)\,dt-n\mu_n(t)\,dt$.

The diversification process runs until the present time t_p.

We denote by $P_n(t)$ the probability that the process has n species at time t. This probability satisfies the following ordinary differential equation (ODE, called master equation or forward Kolmogorov equation (Bailey, 1990)),

$$\frac{dP_n(t)}{dt} = \mu_{n+1}\,(n+1)P_{n+1}(t) + \lambda_{n-1}\,(n-1)P_{n-1}(t) - (\lambda_n+\mu_n)\,nP_n(t), \quad (2.2.1)$$

where we omit in the notation the time dependence of the speciation and extinction rates.

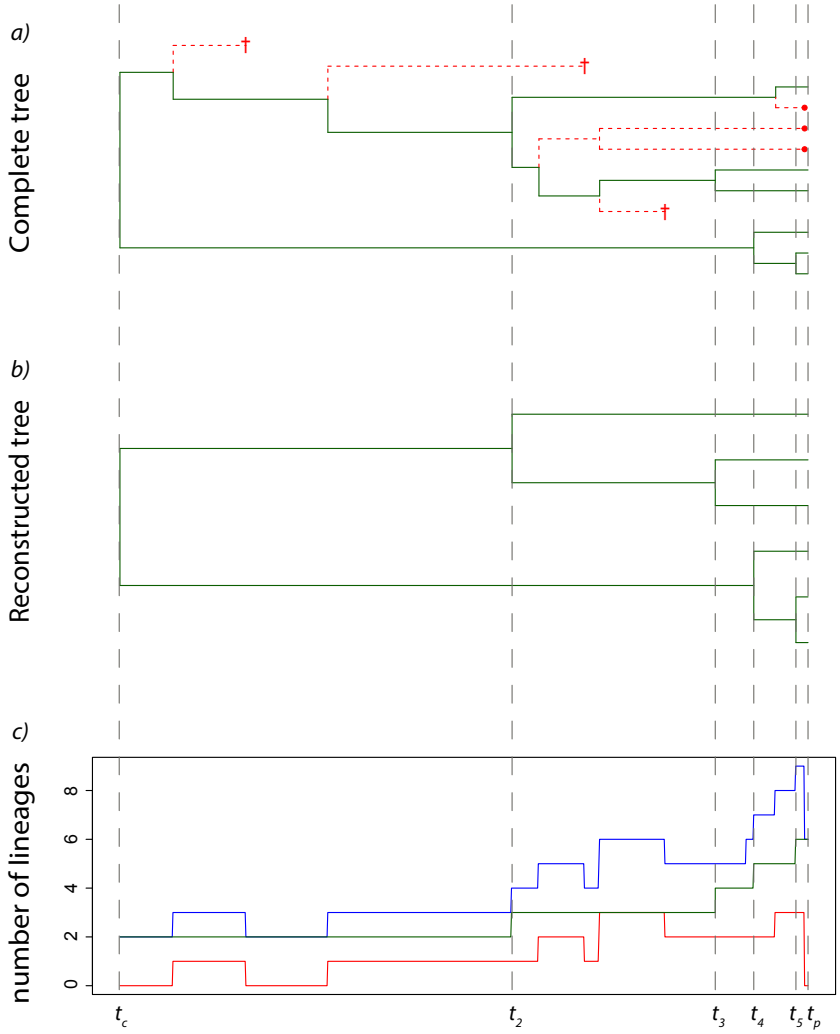

Figure 2.1: a) Full tree where missing species are plotted as red dashed lines: the ones ending in a cross become extinct before the present, whereas the ones ending with a red dot are unsampled species at the present; b) Corresponding reconstructed tree in which only extant species are present. This is the type of tree we usually work with because actual phylogenetic trees are usually obtained from molecular data taken from extant species; c) Lineages-through-time plot: the green line represents the number of lineages leading to extant species (k), the red line represents lineages leading to extinct or unsampled species (m), and the blue line represents the total number of lineages ($n = k + m$).

Sampling models

At the present time t_p a subset of the n extant species are observed and sampled. This sampling process can been modelled in two different ways (see introduction). The first model assumes that a fixed number of species is unsampled, which corresponds to the n-sampling scheme of Lambert, Morlon, and Etienne (2015). That is, the number of extant species at t_p that are not sampled, a number we denote by m_p, is a model parameter. The second model assumes that each extant species at the present time is sampled with a given probability, which has been called f-sampling (Nee, May, and Harvey, 1994) or ρ-sampling (Lambert, Morlon, and Etienne, 2015). In this case the number of unsampled species m_p is a random variable, and the probability with which extant species are sampled is a model parameter, which we denote by f_p.

Reconstructed tree

A realization of the diversification process from t_c to t_p can be represented graphically as a tree, see Figure 2.1. The complete tree shows all the species that have originated in the process (Fig. 2.1, panel a). However, in practice we have only access to the reconstructed tree, i.e., the complete tree from which we remove all the species that became extinct before the present or that were not sampled (Fig. 2.1, panel b). While it would be straightforward to infer information about the diversification process based on the complete tree, this task is much more challenging when only the reconstructed tree is available.

This paper deals with likelihood formulas for a reconstructed phylogenetic tree. The number of tips equals the number of sampled extant species k_p. We assume that also the number of unsampled extant species is known, a number we denote by m_p. The information contained in a phylogenetic tree consists of a topology and a set of branching times. For a large set of diversification models, including the diversity-dependent one, all trees having the same branching times but different topologies are equally probable (Lambert and Stadler, 2013). Hence, rather than computing the likelihood of a specific topology, we present formulas for the likelihood of the vector of branching times. We denote the vector of branching times by $\mathbf{t} = (t_2, t_3, \ldots, t_{k_p-1})$, where t_k is the branching time at which the phylogenetic tree changes from k to $k+1$ branches. It will be convenient to set $t_0 = t_1 = t_c$ and $t_{k_p} = t_p$.

Likelihood conditioning

It is common practice to condition the likelihood on the survival of both ancestor lineages to the present time (Nee, May, and Harvey, 1994). Indeed, we would only do an analysis on trees that have actually survived to the present. To incorporate this fact we need to divide the unconditioned likelihood by the probability for each of the ancestor species at the crown age to have sampled extant descendants. This probability would necessarily depend on the way extant species were sampled, i.e. using either the n-sampling or the f-sampling model. However, for the sake of simplicity, here we apply the same conditioning as presented in the original paper (Etienne et al., 2012), where it is required that the descendants survive to the present, but not that they are sampled. In this way the conditioning becomes independent of the choice of the sampling scheme.

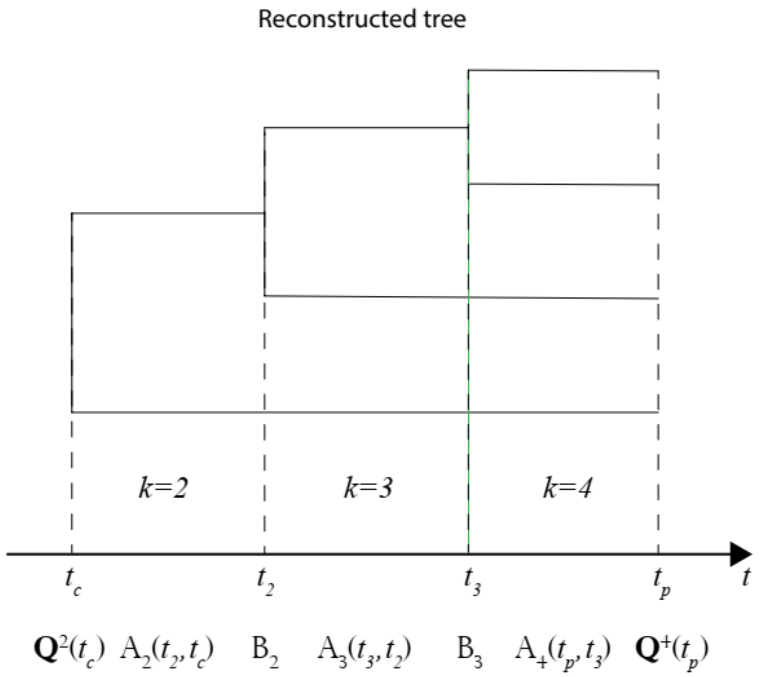

Figure 2.2: An example of how to build a likelihood for a tree with $k_p = 4$ tips. We start with a vector $\mathbf{Q}^2(t_c)$ at the crown age. We use $\mathrm{A}_k(t_k,t_{k-1})$ and $\mathrm{B}_k(t_k)$ to evolve the vector across the entire tree (on branches and nodes, respectively) up to the present time t_p according to $\mathbf{Q}^4(t_4) = A_4(t_4,t_3)B_3(t_3)A_3(t_3,t_2)B_2(t_2)A_2(t_2,t_c)\mathbf{Q}^2(t_c)$. At the present time the likelihood accounting for m_p missing species will be proportional to the m_p-th component of the vector $L_{4,m_p} \propto Q^4_{m_p}(t_p)$.

2.3 The Q-framework

Etienne et al. (2012) presented an approach to compute the likelihood of a phylogeny for the diversity-dependent model. It is based on a new variable, $Q_m^k(t)$, which they described as "the probability that a realization of the diversification process is consistent with the phylogeny up to time t, and has $n = m+k$ species at time t" (Ref. Etienne et al., 2012, Box 1), where k lineages are represented in the phylogenetic tree (because they are ancestral to one of the k_p species extant and sampled at present) and m additional species are present but unobserved (Fig. 2.1, panel c). These species might not be in the phylogenetic tree because they became extinct before the present or because they are either not discovered or not sampled yet (see introduction). From hereon we will refer to these species denoted by m as missing species. We cannot ignore these missing species, because in a diversity-dependent speciation process, they can influence the speciation and extinction rates.

We start by describing the computation of the variable $Q_{m_p}^{k_p}(t_p)$, which proceeds from the crown age t_c to the present time t_p. It is convenient to arrange the values $Q_m^k(t)$, with $m = 0,1,2,\ldots$, into the vector $\mathbf{Q}^k(t)$. The initial vector $\mathbf{Q}^{k=2}(t_c)$ is transformed into the vector $\mathbf{Q}^k(t)$ at a later time t as follows (Ref. Etienne et al., 2012, Appendix S1, Eq. (S1)):

$$\begin{aligned}\mathbf{Q}^k(t) = {} & \mathrm{A}_k(t_{k-1},t)\,\mathrm{B}_{k-1}(t_{k-1})\,\mathrm{A}_{k-1}(t_{k-2},t_{k-1})\ldots \\ & \mathrm{A}_3(t_2,t_3)\,\mathrm{B}_2(t_2)\,\mathrm{A}_2(t_c,t_2)\,\mathbf{Q}^{k=2}(t_c),\end{aligned}$$

with $t_{k-1} \le t \le t_k$. The operators A_k and B_k are infinite-dimensional matrices that operate along the tree, on branches and nodes, respectively (Fig. 2.2). Continuing this computation until the present time t_p, we get

$$\begin{aligned}\mathbf{Q}^{k_p}(t_p) = {} & \mathrm{A}_{k_p}(t_{k_p-1},t_p)\,\mathrm{B}_{k_p-1}(t_{k_p-1})\,\mathrm{A}_{k_p-1}(t_{k_p-2},t_{k_p-1})\ldots \\ & \mathrm{A}_3(t_2,t_3)\,\mathrm{B}_2(t_2)\,\mathrm{A}_2(t_c,t_2)\,\mathbf{Q}^{k=2}(t_c). \qquad (2.3.1)\end{aligned}$$

Note that Eq. (2.3.1) generalizes Eq. (S1) of Ref. Etienne et al., 2012 to the case in which the rates are time-dependent.

We specify the different terms appearing in the right-hand side of Eq. (2.3.1):

- For the initial vector $\mathbf{Q}^{k=2}(t_c)$ we assume that there are no missing species at crown age, that is, $Q_m^{k=2}(t_c) = \delta_{m,0}$.

- The matrix A_k corresponds to the dynamics of $Q_m^k(t)$ in the time interval $[t_{k-1},t_k]$, during which the phylogenetic tree has k branches. Etienne et al.

(2012) argued that these dynamics are given by the following ODE system (Ref. Etienne et al., 2012, Box 1, Eq. (B2)):

$$\begin{aligned}\frac{dQ_m^k(t)}{dt} &= \mu_{k+m+1}(m+1)Q_{m+1}^k(t) + \lambda_{k+m-1}(m-1+2k)Q_{m-1}^k(t) \\ &\quad - (\lambda_{m+k}+\mu_{m+k})(m+k)Q_m^k(t), \quad \forall m>0, \\ \frac{dQ_0^k(t)}{dt} &= \mu_{k+1}Q_1^k(t) - (\lambda_k+\mu_k)kQ_0^k(t), \qquad \text{if } m=0.\end{aligned} \tag{2.3.2}$$

The quantity $Q_m^{k_p}(t_p)$ is proportional to the likelihood of the tree at the present time with m unsampled extant species (see Claim 2.3.1 for the precise statement, including the constant of proportionality). We can collect the coefficients of $Q_m^k(t)$ on the right-hand side of the ODE system in a matrix $\mathrm{V}_k(t)$. If we do so, the system can be rewritten as

$$\frac{d\mathbf{Q}^k(t)}{dt} = \mathrm{V}_k(t)\,\mathbf{Q}^k(t),$$

which has solution

$$\mathbf{Q}^k(t) = \exp\Big(\int_{t_{k-1}}^{t} \mathrm{V}_k(s)\,ds\Big)\,\mathbf{Q}^k(t_{k-1}),$$

so that

$$\mathrm{A}_k(t_{k-1},t) = \exp\Big(\int_{t_{k-1}}^{t} \mathrm{V}_k(s)\,ds\Big). \tag{2.3.3}$$

- The matrix B_k transforms the solution of the ODE system ending at t_k into the initial condition of the ODE system starting at t_k. It is a diagonal matrix with components $k\lambda_{k+m}dt$, so that

$$Q_m^{k+1}(t_k) = \big(\mathrm{B}_k(t_k)\big)_{m,m} Q_m^k(t_k) = k\lambda_{k+m}dt\,Q_m^k(t_k). \tag{2.3.4}$$

The multiplication by $\lambda_{k+m}dt$ corresponds to the probability that a speciation occurs in the time interval $[t_k, t_k+\mathrm{d}t]$. We multiply by a factor k because we do not specify which lineage speciates (recall that we compute the likelihood of a vector of branching times rather than of a specific topology). In the likelihood expressions we will omit the differential (a choice

that is widely adopted across the vast majority of this kind of models in the literature) as it is actually not essential in parameter estimation. Therefore, we will work with a likelihood density, but for simplicity we will refer to it as a likelihood.

We are then ready to formulate the claim made by Etienne et al. (2012) (in particular, from their Appendix S1, see Eqs. (S2) and (S6) to obtain Eq. 2.3.5 and Eqs. (S7-11) to obtain Eq. 2.3.6).

Claim 2.3.1 *Consider the diversity-dependent diversification model, given by speciation rates $\lambda_n(t)$ and extinction rates $\mu_n(t)$. The diversification process starts at crown age t_c with two ancestor species, and ends at the present time t_p, at which a fixed number of species m_p are not sampled. A phylogenetic tree is constructed for the sampled species. Then, the likelihood that the phylogenetic tree has k_p tips and vector of branching times $\mathbf{t} = (t_1, t_2, \ldots, t_{k_p-1})$, conditional on the event that both crown lineages survive until the present, is equal to*

$$L_{k_p,\mathbf{t},m_p} = \frac{Q^{k_p}_{m_p}(t_p)}{\binom{k_p+m_p}{m_p} P_c(t_c,t_p)}. \tag{2.3.5}$$

The term $Q^{k_p}_{m_p}(t_p)$ in the numerator of this expression is obtained from Eq. (2.3.1). The term $P_c(t_c,t_p)$ in the denominator, where the subscript c stands for conditioning, is equal to

$$P_c(t_c,t_p) = \sum_{m=0}^{\infty} \frac{6}{(m+2)(m+3)} \sum_{n=0}^{\infty} \left(\mathrm{A}_2(t_c,t_p)\right)_{m,n} Q^{k=2}_n(t_c), \tag{2.3.6}$$

where $Q^{k=2}_n(t_c) = \delta_{n,0}$.

The structure of the likelihood expression (2.3.5) can be understood intuitively. It is proportional to $Q^{k_p}_{m_p}(t_p)$, which in Etienne et al.'s interpretation is the probability that the diversification process generates the phylogenetic tree with k_p tips and m_p missing species at present time t_p. The combinatorial factor $\binom{k_p+m_p}{m_p}$ accounts for the number of ways to select m_p missing species out of a total pool of $k_p + m_p$ species. The factor $P_c(t_c,t_p)$ is the probability that both ancestor species at crown age t_c have descendant species at the present time t_p. Hence, this factor applies the likelihood conditioning.

Etienne et al. (2012) provided numerical evidence that Claim 2.3.1 is in agreement with the likelihood provided by Nee, May, and Harvey (1994) under the hypothesis of diversity-independent speciation and extinction rates and no missing

species at the present. However, a rigorous analytical proof, even for this specific case, has not yet been given. In this paper we show that Claim 2.3.1 holds (1) for the diversity-independent (but possibly time-dependent) case and (2) for the diversity-dependent case without extinction (i.e., extinction rate $\mu = 0$).

2.4 The likelihood for the diversity-independent case

Claim 2.3.1 proposes a likelihood expression for the case with a known number of unsampled species at the present, i.e., it accounts for n-sampling. For the diversity-independent case, i.e., $\lambda_n(t) = \lambda(t)$ and $\mu_n(t) = \mu(t)$, the likelihood is contained in a more general result established by Lambert, Morlon, and Etienne (2015). In the following proposition we derive an explicit likelihood expression by restricting the result of Lambert et al. to the diversity-independent case.

Proposition 2.4.1 *Consider the diversity-independent diversification model, given by speciation rates $\lambda(t)$ and extinction rate $\mu(t)$. The diversification process starts at crown age t_c with two ancestor species, and ends at the present time t_p, at which a fixed number of species m_p are not sampled. A phylogenetic tree is constructed for the sampled species. Then, the likelihood that the phylogenetic tree has k_p tips and vector of branching times $\mathbf{t} = (t_1, t_2, \ldots, t_{k_p-1})$, conditional on the event that both crown lineages survive until the present, is equal to*

$$L^{(div\text{-}indep)}_{k_p,\mathbf{t},m_p} = \frac{(k_p-1)!}{\binom{k_p+m_p}{m_p}} (1-\eta(t_c,t_p))^2 \prod_{i=2}^{k_p-1} \lambda(t_i)(1-\xi(t_i,t_p))(1-\eta(t_i,t_p)) \sum_{\mathbf{m}|m_p} \prod_{j=0}^{k_p-1} (m_j+1)(\eta(t_j,t_p))^{m_j}, \tag{2.4.1}$$

where we used the convention $t_0 = t_1 = t_c$. The components m_j (with $j = 0, 1, \ldots, k_p - 1$) of the vectors $\mathbf{m}$, in the sum on the second line, are non-negative integers satisfying $\sum_{j=0}^{k_p-1} m_j = m_p$. The functions $\xi(t,t_p)$ and $\eta(t,t_p)$ are given by

$$\begin{aligned}\xi(t,t_p) &= 1 - \frac{1}{\alpha(t,t_p) + \int_t^{t_p} \lambda(s)\,\alpha(t,s)\,ds} \\ &= 1 - \frac{1}{1 + \int_t^{t_p} \mu(s)\,\alpha(t,s)\,ds} \qquad (2.4.2)\\ \eta(t,t_p) &= 1 - \frac{\alpha(t,t_p)}{\alpha(t,t_p) + \int_t^{t_p} \lambda(s)\,\alpha(t,s)\,ds} \\ &= 1 - \frac{\alpha(t,t_p)}{1 + \int_t^{t_p} \mu(s)\,\alpha(t,s)\,ds}, \qquad (2.4.3)\end{aligned}$$

with

$$\alpha(t,s) = \exp\Big(\int_t^s (\mu(s') - \lambda(s'))\,ds'\Big).$$

The functions $\xi(t,t_p)$ and $\eta(t,t_p)$ are those appearing in Kendall's solution of the birth-death model (see Ref. Kendall, 1948b, Eqs. (10–12)), and are useful to describe the process when time-dependent rates are involved. Given the probability $P_n(t,t_p)$ of realizing a process starting with 1 species at time t and ending with n species at time t_p, we have that $\xi(t,t_p) = P_0(t,t_p)$ and $\eta(t,t_p) = \frac{P_{n^\star+1}(t,t_p)}{P_{n^\star}(t,t_p)}$ for any $n^\star > 0$.

Proof The likelihood for n-sampling was originally provided by Ref. (Lambert, Morlon, and Etienne, 2015), Eq. (7), but we start from the explicit version provided in Ref. (Etienne, Morlon, and Lambert, 2014), Eq. (1), see corrigendum in Ref. (Etienne, 2017),

$$\begin{aligned}L^{(\text{div-indep})}_{k_p,\mathbf{t},m_p} &= \frac{(k_p-1)!}{\binom{k_p+m_p}{m_p}} \\ &\quad \times (g(t_c,t_p))^2 \prod_{i=2}^{k_p-1} f(t_i,t_p) \sum_{\mathbf{m}|m_p} \prod_{j=0}^{k_p-1} (m_j+1)(1-g(t_j,t_p))^{m_j}.\end{aligned} \qquad (2.4.4)$$

Etienne, Morlon, and Lambert (2014) and Lambert, Morlon, and Etienne (2015) specify the functions $f(t,t_p)$ and $g(t,t_p)$ as the solution of a system of ODEs for the case of protracted speciation, a model where speciation does not take place instantaneously but is initiated and needs time to complete. The standard

diversification model is then obtained by taking the limit in which the speciation-completion rate tends to infinity. In this limit the four-dimensional system of Etienne, Morlon, and Lambert (2014), Eq. (2), reduces to a two-dimensional system of ODEs,

$$\begin{aligned} f(t,t_p) = \frac{dg(t,t_p)}{dt} &= \lambda(t)\,(1-q(t,t_p))\,g(t,t_p) \\ \frac{dq(t,t_p)}{dt} &= -\mu(t) + (\mu(t)+\lambda(t))\,q(t,t_p) - \lambda(t)\,q^2(t,t_p). \end{aligned}$$

Note that in this paper time t runs from past to present rather than from present to past as in Etienne, Morlon, and Lambert (2014). The conditions at the present time t_p are given by $g(t_p,t_p)=1$ and $q(t_p,t_p)=0$.

The solution of this system of ODEs can be expressed in terms of $\eta(t,t_p)$ and $\xi(t,t_p)$,

$$\begin{aligned} f(t,t_p) &= \lambda(t)\,(1-\xi(t,t_p))\,(1-\eta(t,t_p)) \\ g(t,t_p) &= 1-\eta(t,t_p) \\ q(t,t_p) &= \xi(t,t_p), \end{aligned}$$

which can be checked using the derivatives of the expressions 2.4.3 and 2.4.2

$$\begin{aligned} \frac{\partial\eta(t,t_p)}{\partial t} &= -\lambda(t)\,(1-\xi(t,t_p))\,(1-\eta(t,t_p)) \\ \frac{\partial\xi(t,t_p)}{\partial t} &= -(\mu(t)-\lambda(t)\,\xi(t,t_p))\,(1-\xi(t,t_p)). \end{aligned}$$

Substituting the functions $f(t,t_p)$ and $g(t,t_p)$ into the likelihood expression (2.4.4) concludes the proof. □

The functions $\xi(t,t_p)$ and $\eta(t,t_p)$ are directly related to the functions used by Nee, May, and Harvey (1994). In particular, the functions they denoted by $P(t,t_p)$ and u_t correspond in our notation to $1-\xi(t,t_p)$ and $\eta(t,t_p)$, respectively.

This correspondence allows us to get an intuitive understanding of the likelihood expression (2.4.1). First consider the case without missing species. Setting $m_p=0$, we get

$$L^{(\text{div-indep})}_{k_p,\mathbf{t},0} = (k_p-1)!\,(1-\eta(t_c,t_p))^2 \prod_{i=2}^{k_p-1} \lambda(t_i)(1-\xi(t_i,t_p))(1-\eta(t_i,t_p)),$$

which is identical to the breaking-the-tree likelihood of Nee et al. (1994, Eq. (20)). In the latter approach the phylogenetic tree is broken into single branches: two for

the interval $[t_c, t_p]$ and one for each interval $[t_i, t_p]$ with $i = 2, 3, \ldots, k_p - 1$. Each branch contributes a factor $(1 - \xi(t_i, t_p))(1 - \eta(t_i, t_p))$, equal to the probability that the branch starting at t_i has a single descendant species at t_p. For the two branches originating at t_c, the factor $(1 - \xi(t_i, t_p))$, equal to the probability of having (one or more) descendant species, drops due to the conditioning. For the other branches, there is an additional factor $\lambda(t_i)$ for the speciation events.

Next consider the case with missing species. Each of the branches resulting from breaking the tree can contribute species to the pool of m_p missing species. For the branch over the interval $[t_j, t_p]$, there are m_j such species, each contributing a factor $\eta(t_j, t_p)$ to the likelihood. Indeed, $(1-\xi(t_j,t_p))(1-\eta(t_j,t_p))(\eta(t_j,t_p))^{m_j}$ is equal to probability of having exactly $m_j + 1$ descendant species at the present time. One of these species is represented in the phylogenetic tree, justifying the combinatorial factor $(m_j + 1)$ in the second line of Eq. (2.4.1).

Finally, we recall the expressions for the functions $\xi(t, t_p)$ and $\eta(t, t_p)$ in the case of constant rates, $\lambda(t) = \lambda$ and $\mu(t) = \mu$,

$$\xi(t,t_p) = \frac{\mu\left(1 - \mathrm{e}^{-(\lambda-\mu)(t_p-t)}\right)}{\lambda - \mu\,\mathrm{e}^{-(\lambda-\mu)(t_p-t)}}$$
$$\eta(t,t_p) = \frac{\lambda\left(1 - \mathrm{e}^{-(\lambda-\mu)(t_p-t)}\right)}{\lambda - \mu\,\mathrm{e}^{-(\lambda-\mu)(t_p-t)}}.$$

2.5 Equivalence for the diversity-independent case

Likelihood formula (2.4.1) allows speciation and extinction rates to be arbitrary functions of time, $\lambda(t)$ and $\mu(t)$. Here we show that, for the diversity-independent case, we find the same likelihood formula with the approach of Etienne et al. (2012). From now on, we will use the short-hand notation ∂_x for the partial derivative with respect to the generic variable x.

Theorem 2.5.1 *Claim 2.3.1 holds for the diversity-independent case.*

Proof The proof relies heavily on generating functions. First, we introduce the generating function for the variables $Q_m^k(t)$,

$$F_k(z,t) = \sum_{m=0}^{\infty} z^m Q_m^k(t). \tag{2.5.1}$$

The set of ODEs satisfied by $Q_m^k(t)$, Eq. (2.3.2), transforms into a partial differential equation (PDE) for the generating function $F_k(z,t)$,

$$\begin{aligned}\partial_t F_k(z,t) &= (\mu(t) - z\lambda(t))(1-z)\partial_z F_k(z,t) \\ &\quad + k(2z\lambda(t) - \lambda(t) - \mu(t))F_k(z,t) \\ &= c(z,t)\partial_z F_k(z,t) + k\partial_z c(z,t)F_k(z,t),\end{aligned} \tag{2.5.2}$$

with

$$c(z,t) = (\mu(t) - z\lambda(t))(1-z).$$

Note that the number of branches k changes at each branching time, so that the PDE for $F_k(z,t)$ is valid only for $t_{k-1} \leq t \leq t_k$ (corresponding to the operator A_k). At branching time t_k, the solution $F_k(z,t_k)$ has to be transformed to provide the initial condition for the PDE for $F_{k+1}(z,t)$ at time t_k (corresponding to the operator B_k). Using Eq. (2.3.4) and dropping the differential, we get

$$F_{k+1}(z,t_k) = k\lambda(t_k)F_k(z,t_k). \tag{2.5.3}$$

The initial condition at crown age is $F_2(z,t_c) = 1$ because $Q_m^{k=2}(t_c) = \delta_{m,0}$.

Next, we define $P_n(s,t)$ as the probability that the birth-death process that started with one species at time s has n species at time t. The corresponding generating function is defined as,

$$G(z,s,t) = \sum_{n=0}^{\infty} z^n P_n(s,t). \tag{2.5.4}$$

The set of ODEs satisfied by $P_n(s,t)$, Eq. (2.2.1), transforms into a PDE,

$$\partial_t G(z,s,t) = c(z,t)\partial_z G(z,s,t). \tag{2.5.5}$$

Its solution was given by Kendall (1948, Eq. (9)),

$$G(z,s,t) = \frac{\xi(s,t) + (1 - \xi(s,t) - \eta(s,t))z}{1 - z\eta(s,t)}, \tag{2.5.6}$$

where $\xi(s,t)$ and $\eta(s,t)$ are given in Eqs. (2.4.2) and (2.4.3).

The generating function $F_k(z,t)$ can be expressed in terms of the generating function $G(z,s,t)$, as shown in the following lemma.

Lemma 1 *The generating function $F_k(z,t)$ of the variables $Q_m^k(t)$ is given by*

$$F_k(z,t) = H^2(z,t_c,t)\prod_{j=2}^{k-1} j\lambda_j(t_j)H(z,t_j,t) \tag{2.5.7}$$

with

$$H(z,s,t) = \partial_z G(z,s,t) = \frac{(1-\xi(s,t))\,(1-\eta(s,t))}{(1-z\,\eta(s,t))^2}. \tag{2.5.8}$$

To prove the lemma, let us suppose that the solution of Eq. (2.5.2) is of the form,

$$F_k(z,t) = C_k(\mathbf{t})\prod_{j=0}^{k-1}\partial_z G(z,t_j,t) \tag{2.5.9}$$

where we used the convention $t_0 = t_1 = t_c$ and $C_k(\mathbf{t})$ is a constant depending on the branching times. This expression can be rewritten as,

$$F_k(z,t) = C_k(\mathbf{t})\,\frac{1}{k}\sum_{i=0}^{k-1}\partial_z G(z,t_i,t)\prod_{j\neq i,j=0}^{k-1}\partial_z G(z,t_j,t).$$

The partial derivatives of F_k can now be computed,

$$\begin{aligned}
\partial_z F_k &= C_k(\mathbf{t})\sum_{i=0}^{k-1}\partial_z^2 G(z,t_i,t)\prod_{j\neq i,j=0}^{k-1}\partial_z G(z,t_j,t)\\
\partial_t F_k &= C_k(\mathbf{t})\sum_{i=0}^{k-1}\partial_t\partial_z G(z,t_i,t)\prod_{j\neq i,j=0}^{k-1}\partial_z G(z,t_j,t).
\end{aligned}$$

We substitute these expressions into the PDE, Eq. (2.5.2),

$$\begin{aligned}
&\sum_{i=0}^{k-1}\partial_t\partial_z G(z,t_i,t)\prod_{j\neq i,j=0}^{k-1}\partial_z G(z,t_j,t)\\
&\quad = c(z,t)\sum_{i=0}^{k-1}\partial_z^2 G(z,t_i,t)\prod_{j\neq i,j=0}^{k-1}\partial_z G(z,t_j,t)\\
&\qquad + k\,\partial_z c(z,t)\,\frac{1}{k}\sum_{i=0}^{k-1}\partial_z G(z,t_i,t)\prod_{j\neq i,j=0}^{k-1}\partial_z G(z,t_j,t).
\end{aligned}$$

This equation is satisfied if the following equation is satisfied for every $i = 0, 1, \ldots, k$,

$$\begin{aligned}
&\partial_t \partial_z G(z,t_i,t) \prod_{j\neq i, j=0}^{k-1} \partial_z G(z,t_j,t) \\
&\quad = c(z,t)\, \partial_z^2 G(z,t_i,t) \prod_{j\neq i, j=0}^{k-1} \partial_z G(z,t_j,t) \\
&\qquad + \partial_z c(z,t)\, \partial_z G(z,t_i,t) \prod_{j\neq i, j=0}^{k-1} \partial_z G(z,t_j,t).
\end{aligned}$$

This is the case if

$$\partial_t \partial_z G(z,t_i,t) = c(z,t)\, \partial_z^2 G(z,t_i,t) + \partial_z c(z,t)\, \partial_z G(z,t_i,t),$$

or, equivalently, if

$$\partial_z \big[\partial_t G(z,t_i,t)\big] = \partial_z \big[c(z,t)\, \partial_z G(z,t_i,t)\big].$$

This is an identity because $G(z,t_i,t)$ satisfies Eq. (2.5.5).

Next, we verify that the constants $C_k(\mathbf{t})$ can be determined such that initial conditions (2.5.3) are satisfied. This is indeed the case if we take

$$C_k(\mathbf{t}) = \prod_{j=2}^{k-1} j\lambda(t_j).$$

Introducing the function $H(z,s,t)$ and using $t_0 = t_1 = t_c$ complete the proof of the lemma.

Next, we use Eq. (2.5.7) to derive an explicit expression for the likelihood (2.3.5) of Claim 2.3.1. It will be useful to have explicit expressions for derivatives of the function $H(z,s,t)$. It follows from Eq. (2.5.8) that

$$\frac{1}{a!}\partial_z^a \big[H^b(z,t_j,t)\big] = \binom{a+2b-1}{a} H^b(z,t_j,t) \left(\frac{\eta(t_j,t)}{1-z\,\eta(t_j,t)}\right)^a, \quad (2.5.10)$$

where a and b are positive integers.

To evaluate the numerator of Eq. (2.3.5), we have to extract $Q_{m_p}^{k_p}(t_p)$ from the

generating function $F_{k_p}(z,t_p)$. Using Leibniz' rule,

$$
\begin{aligned}
Q_{m_p}^{k_p}(t_p) &= \frac{1}{m_p!}\partial_z^{m_p}\left[F_{k_p}(z,t_p)\right]_{z=0} \\
&= \frac{C_{k_p}(\mathbf{t})}{m_p!}\,\partial_z^{m_p}\left[\prod_{j=0}^{k_p-1} H(z,t_j,t_p)\right]_{z=0} \\
&= \frac{C_{k_p}(\mathbf{t})}{m_p!}\sum_{\mathbf{m}|m_p}\binom{m_p}{m_0,m_1,\ldots,m_{k_p-1}}\prod_{j=0}^{k_p-1}\partial_z^{m_j}\left[H(z,t_j,t_p)\right]_{z=0} \\
&= \frac{C_{k_p}(\mathbf{t})}{m_p!}\sum_{\mathbf{m}|m_p}\frac{m_p!}{\prod_i m_i!}\prod_{j=0}^{k_p-1}(m_j+1)! \\
&\times\left[H(z,t_j,t_p)\left(\frac{\eta(t_j,t_p)}{1-z\,\eta(t_j,t_p)}\right)^{m_j}\right]_{z=0} \\
&= C_{k_p}(\mathbf{t})\prod_{j=0}^{k_p-1}H(0,t_j,t_p)\sum_{\mathbf{m}|m_p}\frac{1}{\prod_{i=0}^{k_p-1}m_i!}\prod_{j=0}^{k_p-1}(m_j+1)!\,\eta^{m_j}(t_j,t_p) \\
&= \prod_{j=2}^{k_p-1} j\lambda(t_j)\prod_{j=0}^{k_p-1}(1-\xi(t_j,t_p))(1-\eta(t_j,t_p)) \\
&\times\sum_{\mathbf{m}|m_p}\prod_{j=0}^{k_p-1}(m_j+1)\,\eta^{m_j}(t_j,t_p) \\
&= (k_p-1)!(1-\xi(t_c,t_p))^2(1-\eta(t_c,t_p))^2 \\
&\times\prod_{j=2}^{k_p-1}\lambda(t_j)\,(1-\xi(t_j,t_p))(1-\eta(t_j,t_p)) \\
&\times\sum_{\mathbf{m}|m_p}\prod_{j=0}^{k_p-1}(m_j+1)\,\eta^{m_j}(t_j,t_p). \qquad (2.5.11)
\end{aligned}
$$

To evaluate the denominator of Eq. (2.3.5), we have to extract $Q_m^{k=2}(t_p)$ from the generating function,

$$
Q_m^{k=2}(t_p) = \frac{1}{m!}\partial_z^m\left[F_2(z,t_p)\right]_{z=0} = \frac{1}{m!}\partial_z^m\left[H^2(z,t_c,t_p)\right]_{z=0}.
$$

Substituting into Eq. (2.3.6) and using Eq. (2.5.10), we get

$$\begin{aligned} P_c(t_c,t_p) &= \sum_{m=0}^{\infty} \frac{6}{(m+2)(m+3)} \frac{1}{m!} \partial_z^m \left[H^2(z,t_c,t_p)\right]_{z=0} \\ &= H^2(0,t_c,t_p) \sum_{m=0}^{\infty} (m+1)\, \eta^m(t_c,t_p) \\ &= (1-\xi(t_c,t_p))^2. \end{aligned} \tag{2.5.12}$$

Finally, substituting Eqs. (2.5.11) and (2.5.12) into the likelihood formula (2.3.5) of Claim 2.3.1,

$$\begin{aligned} L_{k_p,\mathbf{t},m_p} = \frac{(k_p-1)!}{\binom{k_p+m_p}{m_p}} (1-\eta(t_1,t_p))^2 \prod_{j=2}^{k_p-1} \lambda(t_j)(1-\xi(t_j,t_p))(1-\eta(t_j,t_p)) \\ \sum_{\mathbf{m}|m_p} \prod_{j=0}^{k_p-1} (m_j+1)\, \eta^{m_j}(t_j,t_p), \end{aligned} \tag{2.5.13}$$

which is identical to likelihood formula (2.4.1). This concludes the proof of the theorem. □

2.6 A note on sampling a fraction of extant species

Nee, May, and Harvey (1994) noted that one way to model the sampling of extant species is equivalent to a mass extinction just before the present. This sampling model corresponds to sampling each extant species with a given probability f_p, which has also been called ρ-sampling (Lambert, Morlon, and Etienne, 2015). We use the link with mass extinction to extend the previous formula for n-sampling to the case of ρ-sampling.

First, we formulate the ρ-sampling version of Claim 2.3.1.

Claim 2.6.1 *Consider the diversity-dependent diversification model, given by speciation rates $\lambda_n(t)$ and extinction rates $\mu_n(t)$. The diversification process starts at crown age t_c with two ancestor species, and ends at the present time t_p, at which extant species are sampled with probability f_p. Then, the likelihood of a phylogenetic tree with k_p tips and branching times $\mathbf{t}$, conditional on the event that both crown lineages survive until the present, is equal to*

$$L_{k_p,\mathbf{t}} = \frac{P_s(t_c,\mathbf{t},t_p,f_p)}{P_c(t_c,t_p)}. \tag{2.6.1}$$

The term $P_s(t_c,\mathbf{t},t_p,f_p)$ in the numerator, where the subscript s stands for sampling, is equal to

$$P_s(t_c,\mathbf{t},t_p,f_p) = \sum_{m=0}^{\infty} f_p^{k_p}\,(1-f_p)^m\,Q_m^{k_p}(t_p), \tag{2.6.2}$$

where $Q_m^{k_p}(t_p)$ is obtained from Eq. (2.3.1). The term $P_c(t_c,t_p)$ in the denominator, where the subscript c stands for conditioning, is equal to

$$P_c(t_c,t_p) = \sum_{m=0}^{\infty} \frac{6}{(m+2)(m+3)}\,Q_m^{k=2}(t_p), \tag{2.6.3}$$

where $Q_m^{k=2}(t_p)$ is again obtained from Eq. (2.3.1).

Next, we establish as a reference the likelihood formula for ρ-sampling in the diversity-independent case.

Proposition 2.6.1 *Consider the diversity-independent diversification model, given by speciation rates $\lambda(t)$ and extinction rates $\mu(t)$. The diversification process starts at crown age t_c with two ancestor species, and ends at the present time t_p, at which extant species are sampled with probability f_p. Then, the likelihood of a phylogenetic tree with k_p tips and branching times $\mathbf{t}$, conditional on the event that both crown lineages survive until the present, is equal to*

$$L_{k_p,\mathbf{t}}^{(div\text{-}indep)} = (k_p-1)!\,(1-\widetilde{\eta}(t_c,t_p))^2 \prod_{i=2}^{k_p-1} \lambda(t_i)(1-\widetilde{\xi}(t_i,t_p))(1-\widetilde{\eta}(t_i,t_p)). \tag{2.6.4}$$

The functions $\widetilde{\xi}(t,t_p)$ and $\widetilde{\eta}(t,t_p)$ are given by

$$\begin{aligned}
\widetilde{\xi}(t,t_p) &= 1 - \frac{f_p}{\alpha(t,t_p) + f_p \int_t^{t_p} \lambda(s)\,\alpha(t,s)\,ds} \\
&= 1 - \frac{f_p}{f_p + (1-f_p)\,\alpha(t,t_p) + f_p \int_t^{t_p} \mu(s)\,\alpha(t,s)\,ds} \qquad (2.6.5)\\
\widetilde{\eta}(t,t_p) &= 1 - \frac{\alpha(t,t_p)}{\alpha(t,t_p) + f_p \int_t^{t_p} \lambda(s)\,\alpha(t,s)\,ds} \\
&= 1 - \frac{\alpha(t,t_p)}{f_p + (1-f_p)\,\alpha(t,t_p) + f_p \int_t^{t_p} \mu(s)\,\alpha(t,s)\,ds}, \qquad (2.6.6)
\end{aligned}$$

with

$$\alpha(t,s) = \exp\Big(\int_t^s (\mu(s') - \lambda(s'))\, ds'\Big).$$

Proof We use the equivalence between ρ-sampling and a mass extinction, see Ref. (Nee, May, and Harvey, 1994), Eq. (31). We introduce a modified extinction rate $\mu(t)$ containing a delta function just before the present,

$$\widetilde{\mu}(t) = \mu(t) - \ln f_p\ \delta(t - t_p). \tag{2.6.7}$$

The likelihood formula is then obtained by setting $m_p = 0$ in Eq. (2.4.1), while evaluating the functions $\xi(t,t_p)$ and $\eta(t,t_p)$ with the modified extinction rate $\widetilde{\mu}(t,t_p)$. This establishes Eq. (2.6.4); it remains to be proven that the modified functions $\widetilde{\xi}(t,t_p)$ and $\widetilde{\mu}(t,t_p)$ are given by Eqs. (2.6.5) and (2.6.6). This follows by noting that the modified version $\widetilde{\alpha}(t,t_p)$ of the function $\alpha(t,t_p)$ appearing in Eqs. (2.4.2) and (2.4.3) satisfies

$$\begin{aligned}\widetilde{\alpha}(t,t_p) &= \exp\Big(\int_t^{t_p} (\widetilde{\mu}(s) - \lambda(s))\, ds\Big) \\ &= \frac{1}{f_p} \exp\Big(\int_t^{t_p} (\mu(s) - \lambda(s))\, ds\Big) = \frac{1}{f_p}\, \alpha(t,t_p),\end{aligned}$$

while $\widetilde{\alpha}(t,s) = \alpha(t,s)$ if $s < t_p$. □

We are then ready to establish the following result.

Theorem 2.6.1 *Claim 2.6.1 holds for the diversity-independent case.*

Proof We use again the equivalence between ρ-sampling and a mass extinction, see Eq. (2.6.7). Due to Theorem 2.5.1, likelihood formula (2.3.5) is valid for the diversity-independent case. Hence, we can derive the corresponding likelihood formula for ρ-sampling by introducing the modified extinction rate $\widetilde{\mu}(t)$, and setting $m_p = 0$ in the likelihood formula for n-sampling.

The introduction of the modified extinction rate $\widetilde{\mu}(t,t_p)$ corresponds to applying an additional operator to the vector $\mathbf{Q}^{k_p}(t_p)$ at the present time. In particular, the modified vector $\widetilde{\mathbf{Q}}^{k_p}(t_p)$ is given by

$$\widetilde{\mathbf{Q}}^{k_p}(t_p) = \mathrm{C}(f_p)\, \mathbf{Q}^{k_p}(t_p), \tag{2.6.8}$$

where the operator $\mathrm{C}(f_p)$ corresponds to the following ODE, acting in a small time interval $[t_p-\varepsilon, t_p]$ before the present,

$$\frac{d\widetilde{Q}_m^{k_p}(t)}{dt} = \big(\mu - \frac{1}{\varepsilon}\ln f_p\big)(m+1)\widetilde{Q}_{m+1}^{k_p}(t) + \lambda\,(m-1+2k_p)\widetilde{Q}_{m-1}^{k_p}(t) - \big(\lambda + \big(\mu - \frac{1}{\varepsilon}\ln f_p\big)\big)(m+k_p)\widetilde{Q}_m^{k_p}(t),$$

where we added a delta peak to the extinction rate, Eq. (2.6.7), in the ODE satisfied by $\mathbf{Q}^k(t)$, Eq. (2.3.2). In the limit $\varepsilon \to 0$ the terms in $\frac{1}{\varepsilon}$ dominate, so that

$$\frac{d\widetilde{Q}_m^{k_p}(t)}{dt} = -\frac{1}{\varepsilon}\ln f_p\,(m+1)\widetilde{Q}_{m+1}^{k_p}(t) + \frac{1}{\varepsilon}\ln f_p\,(m+k_p)\widetilde{Q}_m^{k_p}(t).$$

This can be rewritten in matrix form as

$$\frac{d\widetilde{\mathbf{Q}}^{k_p}(t)}{dt} = \frac{1}{\varepsilon}\mathrm{W}(f_p)\,\widetilde{\mathbf{Q}}^{k_p},$$

where the operator $\mathrm{W}(f_p)$ is an infinite-dimensional matrix with components

$$\mathrm{W}_{m,n}(f_p) = \begin{cases} \ln f_p\,(m+k_p) & \text{if } m = n \\ -\ln f_p\,(m+1) & \text{if } m = n-1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence, the operator $\mathrm{C}(f_p)$, which is also an infinite-dimensional matrix, is equal to

$$\mathrm{C}(f_p) = \exp\Big(\int_{t_p-\varepsilon}^{t_p}\frac{1}{\varepsilon}\mathrm{W}(f_p)\,ds\Big) = \exp\big(\mathrm{W}(f_p)\big).$$

We need the row $m=0$ to evaluate the likelihood, which is equal to

$$\mathrm{C}_{m=0,n}(f_p) = f_p^{k_p}\,(1-f_p)^n.$$

We are then ready to evaluate likelihood formula (2.3.5) with the modified extinction rate. Setting $m_p = 0$, we get

$$L_{k_p,\mathbf{t}} = \frac{\widetilde{Q}_0^{k_p}(t_p)}{P_c(t_c,t_p)}.$$

Recall that the conditioning probability $P_c(t_c,t_p)$ is not affected by the process of sampling extant species. We get

$$\begin{aligned} L_{k_p,\mathbf{t}} &= \frac{\left(\mathrm{C}(f_p)\,\mathbf{Q}^{k_p}(t_p)\right)_{m=0}}{P_c(t_c,t_p)} \\ &= \frac{\sum_{n=0}^{\infty}\mathrm{C}_{m=0,n}(f_p)\,Q_n^{k_p}(t_p)}{P_c(t_c,t_p)} \\ &= \frac{\sum_{n=0}^{\infty}f_p^{k_p}\,(1-f_p)^n\,Q_n^{k_p}(t_p)}{P_c(t_c,t_p)} \\ &= \frac{P_s(t_c,\mathbf{t},t_p,f_p)}{P_c(t_c,t_p)}, \end{aligned}$$

which is identical to Eq. (2.6.4). This ends the proof. □

Note that Eq. 2.6.2 is equal to $f_p^{k_p}F_{k_p}(1-f_p,t_p)$ which provides an alternative route to prove Claim 2.6.1 (Manceau et al., 2019).

Finally, we give the expressions for the functions $\widetilde{\xi}(t,t_p)$ and $\widetilde{\eta}(t,t_p)$ in the case of constant rates, $\lambda(t)=\lambda$ and $\mu(t)=\mu$,

$$\begin{aligned} \widetilde{\xi}(t,t_p) &= \frac{f_p\,\mu+((1-f_p)\lambda-\mu)\,\mathrm{e}^{-(\lambda-\mu)(t_p-t)}}{f_p\,\lambda+((1-f_p)\lambda-\mu)\,\mathrm{e}^{-(\lambda-\mu)(t_p-t)}} \\ \widetilde{\eta}(t,t_p) &= \frac{f_p\,\lambda\left(1-\mathrm{e}^{-(\lambda-\mu)(t_p-t)}\right)}{f_p\,\lambda+((1-f_p)\lambda-\mu)\,\mathrm{e}^{-(\lambda-\mu)(t_p-t)}}, \end{aligned}$$

which are identical to Eqs. (4) and (5) in the paper by Stadler (2012).

2.7 The diversity-dependent case without extinction

Rabosky and Lovette (2008) derived the likelihood for a particular instance of the diversity-dependent diversification model, namely, when there is no extinction. This is the only case for which a diversity-dependent likelihood formula is available. Here we show that this case is dealt with correctly in the approach of Etienne et al. (2012).

We start by reformulating the result of Rabosky and Lovette (2008) in our notation.

Proposition 2.7.1 *Consider the diversity-dependent model without extinction, given by speciation rates $\lambda_n(t)$. The diversification process starts at crown age t_c with*

two ancestor species, and ends at the present time t_p, at which all extant species are sampled. Then, the likelihood of a phylogenetic tree with k_p tips and branching times $\mathbf{t}$ *is equal to*

$$L^{(no\text{-}extinct)}_{k_p,\mathbf{t},0} = (k_p-1)! \prod_{i=2}^{k_p-1} \lambda_i(t_i) \prod_{j=2}^{k_p} \exp\Big(-j\int_{t_{j-1}}^{t_j} \lambda_j(s)\,ds\Big), \tag{2.7.1}$$

where we used the convention $t_1 = t_c$ and $t_{k_p} = t_p$.

Proof Eq. (2.7.1) follows from Eqs. (2.4) and (2.5) in Ref. Rabosky and Lovette (2008), by noting that ξ_i in their notation corresponds to

$$\exp\Big(-\sum_{j=i}^{k_p}\int_{t_{j-1}}^{t_j} \lambda_j(s)\,ds\Big)$$

in our notation. □

Note that in the case without extinction likelihood conditioning has no effect.

Theorem 2.7.1 *Claim 2.3.1 holds for the diversity-dependent case without extinction.*

Proof To evaluate likelihood expression (2.3.5), we have to solve the ODE for $Q^k_m(t)$, Eq. (2.3.2). Because species cannot become extinct and because all extant species are sampled, every species created during the process is represented in the phylogeny, i.e., there are no missing species. Hence, only the $m = 0$ component of $\mathbf{Q}^k(t)$ is different from zero. The ODE simplifies to

$$\frac{dQ^k_0(t)}{dt} = -k\lambda_k(t)\,Q^k_0(t),$$

where t belongs to $[t_{k-1}, t_k]$. Note that in this time interval there are exactly k species. Given the initial condition $Q^k_0(t_{k-1})$ at t_{k-1} , the solution is

$$Q^k_0(t) = Q^k_0(t_{k-1}) \exp\Big(-k\int_{t_{k-1}}^{t} \lambda_k(s)\,ds\Big).$$

At branching time t_k, variable $Q^k_0(t_k)$ is transformed into variable $Q^{k+1}_0(t_k)$,

$$Q^{k+1}_0(t_k) = k\lambda_k(t_k)\,Q^k_0(t_k).$$

Using the initial condition at crown age t_c, $Q^{k=2}_0(t_c) = 1$, we get

$$Q^{k_p}_0(t_p) = \prod_{i=2}^{k_p-1} i\lambda_i(t_i) \prod_{j=2}^{k_p} \exp\Big(-j\int_{t_{j-1}}^{t_j} \lambda_j(s)\,ds\Big).$$

Substituting into Eq. (2.3.5) yields the desired result. □

2.8 Concluding remarks

We have shown here that for the diversity-independent, but time-dependent birth-death model with n-sampling, the framework of Etienne et al. (2012) yields the same likelihood derived by Lambert, Morlon, and Etienne (2015) (also presented in a more explicit form in Etienne, 2017 and Etienne, Morlon, and Lambert (2014)). This provides strong support for the correctness of this framework, but does not prove that it is also correct for the case of diversity-dependence. We have thus far not been able to provide alternative evidence for this framework, apart from the fact that parameter estimations on simulations of this model provide reasonable, although sometimes biased, estimates (Etienne et al., 2012). We hope that our analysis here will suggest directions for a further substantiating of the framework. The approach taken by Manceau et al. (2019) may be promising, as it also provides numerical evidence for the correctness of the framework in the diversity-dependent case.

Most existing macroevolutionary models rely on the hypothesis that the subcomponents of trees do not interact (and one can thus apply a breaking-the-tree approach, as in Nee, May, and Harvey (1994), pag. 308), therefore letting the likelihood be a factorization of terms that comes independently from the tree's edges and nodes. However, such a hypothesis is not always valid. The diversification process likely also depends on properties of other lineages than the lineage under consideration. The analytical treatment of Etienne et al. (2012) arguments presented in this work suggests a direction towards deriving the likelihood for much more complicated models with "interacting branches", with the arguably simplest case being diversity dependence, i.e. dependence only on the total number of lineages present at any time. Our work, showing analytically that Etienne et al.'s model agrees with existing formulas for likelihoods of simple diversification models, suggests that future models that aim to deal with interacting branches should consider such a structure as a reference point, in the same fashion as models dealing with "breakable" trees often refer to Nee, May, and Harvey (1994) paradigm.

In this article we have proved that the framework to compute a likelihood for diversity-dependent processes by Etienne et al. (2012) agrees with analytical results obtained for diversity-independent diversification models. This suggests that the framework is valid for more general models that take into account the effect of diversity of speciation and extinction rates while still being able to deal with unsampled species in the phylogeny, when this number is known. Our results can thus improve the understanding of the general architecture of macroevolutionary

diversification models providing useful tools for the development of new models.