The probability distribution of the reconstructed phylogenetic tree with occurrence data

Ankit Gupta, Marc Manceau, Timothy Vaughan, Mustafa Khammash*, Tanja Stadler*

Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland.

Abstract

We consider a homogeneous birth-death process with incomplete sampling. Three successive sampling schemes are considered. First, individuals can be sampled through time and included in the tree. Second, they can be occurrences which are sampled through time and not included in the tree. Third, individuals reaching present day can be sampled and included in the tree. Upon sampling, individuals are removed (i.e. die).

The outcome of the process is thus composed of the reconstructed evolutionary tree spanning all individuals sampled and included in the tree, and a timeline of occurrence events which are not placed along the tree. We derive a formula allowing one to compute the joint probability density of these, which can readily be used to perform maximum likelihood or Bayesian estimation of the parameters of the model.

In the context of epidemiology, our probability density allows us to estimate transmission rates through a joint analysis of epidemiological case count data and phylogenetic trees reconstructed from pathogen sequences. Within macroevolution, our equations are the basis for taking into account fossil occurrences from paleontological databases together with extant species phylogenies for estimating speciation and extinction rates. Thus, we provide the theoretical framework for bridging not only the gap between phylogenetics and epidemiology, but also the gap between phylogenetics and paleontology.

Keywords: birth-death process, epidemiology, macroevolution, phylogenetics, phylodynamics

1. Introduction

Birth-death processes are used extensively in both epidemiology and macroevolution, to model the underlying population dynamics of, respectively, infected individuals and species. For simplicity, we will refer to the atomic particles of the process as *individuals* in this paper. In its most simple form, a birth-death

*Corresponding authors, mustafa.khammash@bsse.ethz.ch, tanja.stadler@bsse.ethz.ch

process describes the population dynamics of a set of independent individuals, each of which can give birth to another individual with a constant birth rate λ, or die with a constant death rate μ.

Seminal results on this process have been derived by Kendall (1948), who already evoked potential applications to epidemiology. Much more recently, Nee et al. (1994) and Nee and May (1997) have made important developments to the theory, showing how to compute the probability density of the *reconstructed evolutionary tree*, i.e. the tree obtained by first tracing all genealogical relationships between individuals, before erasing all branches that do not reach present. This has paved the way to extensive use of birth-death processes in modern phylogenetics, with many refinements: birth and death rates have been proposed to vary through time (Morlon et al., 2011; Stadler, 2011), to vary across lineages in the tree (Alfaro et al., 2009), to vary depending on the *type* of individuals (Maddison et al., 2007), or to depend on the number of individuals (Etienne et al., 2012; Leventhal et al., 2013).

In order to fit various applications, the sampling scheme of individuals has also been put under careful scrutiny. Nee et al. (1994) initially suggested that individuals could be sampled at present with a given probability ρ; this uniform sampling is also called *field of bullets* sampling. Yet, trees that were reconstructed using this sampling scheme were always ultrametric trees describing the genealogical relationships between present-day individuals only. This assumption was relaxed by Stadler (2010), who additionally modeled the sampling of individuals throughout the process, with a fixed per-individual sampling rate ψ. When sampled, an individual is displayed along the tree and the reconstructed (non-necessarily ultrametric) tree corresponds to the genealogical relationships between all sampled-through-time and sampled-at-present individuals. This model opened the way to numerous new applications in epidemiology and macroevolution, allowing to take into account non-synchronous data. In phylogenetics, this process is commonly used as a prior on the genealogical relationships between individuals sampled through time (Stadler et al., 2011). In macroevolution, it allows one to use molecular and paleontological evidence together, by simultaneously considering present-day species and the subset of fossil taxa which can be placed along a tree using morphological characters (Zhang et al., 2015; Gavryushkina et al., 2017).

One step further towards considering even more data jointly in one analysis has been performed by Vaughan et al. (2019), who introduce two types of sampling through time. The first one, which is named *sampling and sequencing*, is the same as previously described. The term *sequencing* here referring to the fact that the individual is placed unambiguously along the reconstructed tree using its genetic sequence. Alternatively, an individual could be *sampled and not sequenced*, in which case its existence is only recorded as a time point occurrence along a timeline. This approach is very promising, for it enables one to consider jointly data from case count epidemiological studies together with trees reconstructed from pathogens sequenced during an outbreak. Alternatively, in the context of macroevolution, it allows one to use poorly

preserved fossil occurrences, which could not be placed along the reconstructed tree. Yet, in its current form, the inference framework proposed by Vaughan et al. (2019) relies on computer intensive Monte-Carlo simulations within a particle filtering approach which prevents its use on large datasets. Similarly, (Heath et al., 2014) proposes placing occurrences on a fixed tree using Markov chain Monte Carlo methodology. Here the drawback is again the computer intensive approach as well as relying on a fixed tree rather than sequences.

In this study, we derive a closed form formula for the joint probability density of a reconstructed tree with individuals sampled through time and a record of occurrences, i.e. sampling times for individuals not included in the tree. The underlying model is a birth-death process with sampling through time and at present, where upon sampling individuals are removed (i.e. die). The density can readily be included within phylodynamic tools as a prior in a Bayesian inference framework based on sequences and occurrences, or can be used for maximum likelihood parameter estimation based on a tree and occurrences. Its computational efficiency opens the way to analyse large datasets available for either epidemiology or macroevolution studies.

The aim of the paper is to provide a mathematical study of the birth-death model giving rise to serially sampled reconstructed trees and occurrences. We first introduce model notations and some preliminary observations. Then, we provide an alternative derivation of the probability density of the reconstructed evolutionary tree under the birth-death process with sampling through time (Stadler, 2010), assuming that an individual is removed upon sampling. Finally, we show how to extend this derivation to account for non-sequenced occurrences, and provide an analytical formula to compute the probability density of the reconstructed tree and occurrences. We finally discuss the use of this density to perform inferences in epidemiology and macroevolution in a maximum likelihood or Bayesian setting.

2. Model and notations

We consider a constant rate birth-death process with incomplete sampling. We allow the possibility that for some sampled individuals, the attachment times within the tree are known, whereas for others only their sampling times is known but their attachment times within the tree are unknown. The latter sampling events are called *occurrences*.

We assume that an individual gives birth at rate λ, and its death rate is $(\mu+\psi+\omega)$. Here μ is the death rate *without sampling*, ψ is the death rate *with sampling and tree-inclusion* and ω is the death rate *with sampling but without tree-inclusion* (i.e. an occurrence). Additionally we assume that extant individuals, alive at present time, are sampled and included in the reconstructed tree with probability ρ. For convenience, we will refer to the five types of events in this model using their parameter names. In other words, λ-events refer to the creation of a new individual, μ-events refer to death events without sampling, ψ-events refer to

death events with sampling and tree inclusion, ω-events refer to death events with sampling but without tree inclusion, and finally ρ-events refer to the sampling of extant lineages at present time.

Time is assumed to be 0 at present, and increases going into the past. The process starts at time $t_{or} > 0$ in the past, with one infected lineage and it generates a birth-death lineage tree from which the reconstructed tree is derived. Our data consists of both the reconstructed tree $\mathcal{T}$ and the occurrence set $\mathcal{O}$. The reconstructed tree $\mathcal{T}$ is generated by all the ψ- and ρ-events giving rise to leaves, together with all λ-events subtending those leaves. The occurrence set $\mathcal{O}$ consists of the times of all the ω-events, all belonging to the interval $(0, t_{or})$; see Figure 1 for an example.

We will be interested in the probability density of the joint observation of $(\mathcal{O}, \mathcal{T})$, which will ultimately depend on the times at which observed λ-, ψ-, and ω-events happen, but not on the tree topology (see e.g. (Aldous et al., 2001; Stadler, 2010)). All these times pooled together are denoted as $t_0, t_1, ..., t_n$, starting with $t_0 = 0$ and ending with $t_n = t_{or}$, and the number of observed lineages on the interval (t_h, t_{h+1}) is denoted k_h.

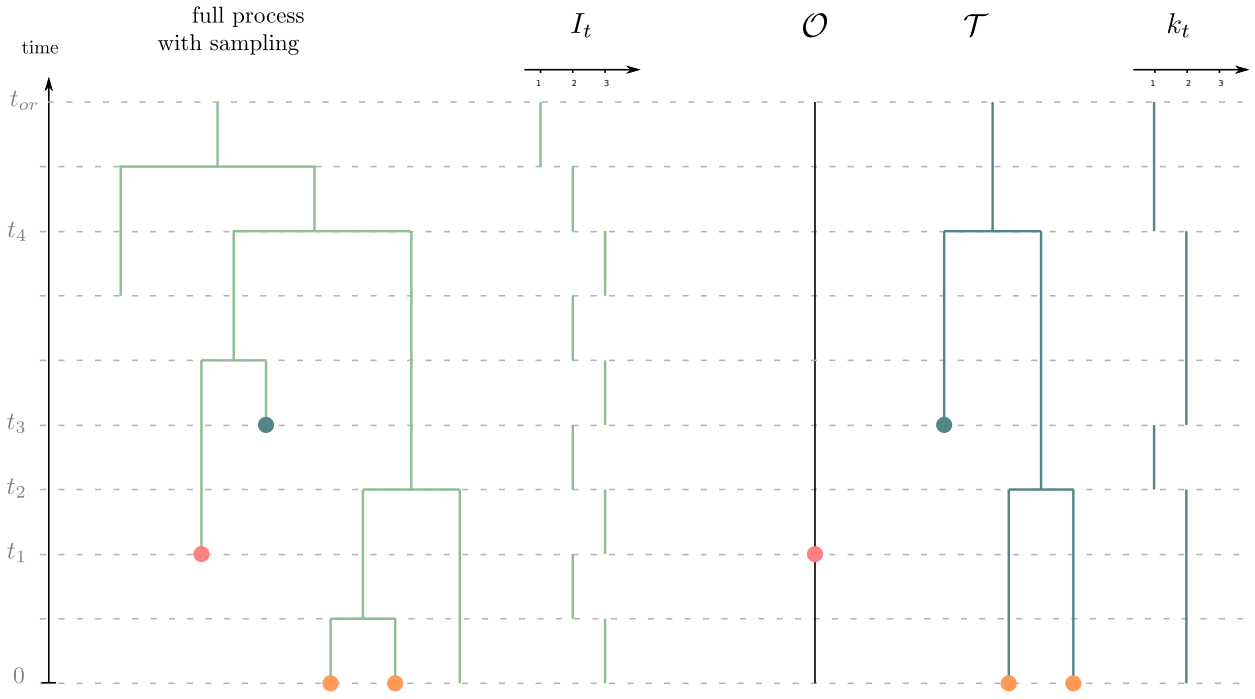

Figure 1: General setting of the method. On the left in green, the full process with sampling is shown. Red dots correspond to ω-sampling (sampling through time without sequencing), blue dots correspond to ψ-sampling (sampling through time with sequencing) and orange dots correspond to ρ-sampling at present. On the right, the observations are shown: a sampled-through-time reconstructed tree $\mathcal{T}$ together with sequential observations $\mathcal{O}$ of sampled individuals along a timeline.

2.1. *Introducing useful quantities*

Let $u(t)$ be the probability that an individual alive at time t before today has no sampled extinct or extant descendant lineages. Also let $p(t)$ be the probability that a lineage alive at time t before today has precisely one sampled extant lineage and no sampled extinct descendant lineages. We can write the Master Equations for these two probabilities as

$$\frac{du}{dt} = \mu - (\lambda + \mu + \omega + \psi)u(t) + \lambda u(t)^2 \tag{2.1}$$

$$\frac{dp}{dt} = -(\lambda + \mu + \omega + \psi)p(t) + 2\lambda u(t)p(t), \tag{2.2}$$

with initial condition $(u(0), p(0)) = (1-\rho, \rho)$. Defining

$$c_1 = \sqrt{(\lambda - \mu - \omega - \psi)^2 + 4\lambda(\psi + \omega)} \quad \text{and} \quad c_2 = -\frac{\lambda - \mu - \omega - \psi - 2\lambda\rho}{c_1}, \tag{2.3}$$

the solution of these Master Equations (see Theorem 3.1 in Stadler (2010)) is given by

$$u(t) = \frac{1}{2\lambda}\left[\lambda + \mu + \omega + \psi + c_1 \frac{e^{-c_1 t}(1-c_2) - (1+c_2)}{e^{-c_1 t}(1-c_2) + (1+c_2)}\right], \tag{2.4}$$

$$p(t) = \frac{4\rho}{2(1-c_2^2) + e^{-c_1 t}(1-c_2)^2 + e^{c_1 t}(1+c_2)^2}. \tag{2.5}$$

2.2. *Key first observations*

In order to understand better future calculations, we must examine the birth-death dynamics of the *number of hidden lineages* in greater detail. Suppose $(X_s)_{s\geq 0}$ be the *forward in time* stochastic process describing the number of hidden lineages in some time-interval $[s_1, s_2] \subset [0, t_{or}]$, where $s_1 < s_2$ and we use this new letter s because time is oriented (only in this section of the manuscript) from the origin 0 towards the present t_{or}. The number of observed lineages in $\mathcal{T}$ is fixed and equal to k in this interval and also there is no occurrence event in this interval. This stochastic process $(X_s)_{s\geq 0}$ is a birth-death process over nonnegative integers $\mathbb{N}_0$ and it has the following characteristics:

- When the state is i, the rate of birth is $\lambda(k+i)$ and if this event happens, the state moves to $(i+1)$ with probability

$$\phi_{ik} := 1 - \frac{\frac{k!}{(k-2)!}}{\frac{(k+i+1)!}{(k+i-1)!}} = 1 - \frac{k(k-1)}{(k+i)(k+i+1)}$$

and with probability $1 - \phi_{ik}$ the state moves to some *absorbing state* Δ outside the state space $\mathbb{N}_0 = \{0, 1, \dots\}$. Note that as the number of infected lineages increases from $k+i$ to $(k+i+1)$ upon a birth event ϕ_{ik} is simply the probability that when we "look backwards", the two coalescing lineages

are not both among the sampled lineages. In case the two coalescing lineages are among the sampled lineages then the number of observed lineages will not be fixed in the time-interval $[s_1, s_2]$ and hence this birth-death trajectory for total number of infected lineages becomes infeasible, which is equivalent to saying that it gets absorbed at state Δ.

- When the number of hidden lineages is i, the rate of death is $(\mu + \psi + \omega)(k + i)$ and if this event happens for some state $i > 0$ then the state moves to $(i - 1)$ with probability

$$\kappa = \frac{\mu}{\mu + \psi + \omega}$$

and it moves to the absorbing state Δ with probability $(1 - \kappa)$. Note that $(1 - \kappa)$ is simply the probability of a death event either being a ω-event or a ψ-event. Both such events will violate our assumption that there is no occurrence event and the number of observed lineages is fixed in the time-interval $[s_1, s_2]$. Moreover if $i = 0$ and a death event happens then the birth-death trajectory again becomes infeasible and so it gets absorbed at state Δ.

It is clear that due to the presence of an absorbing state outside the state space, the process $(X_s)_{s\geq 0}$ is *non-conservative*. The rate of change of the distribution of this process can be specified by its generator (see Chapter 3 in Ethier and Kurtz (1986)). We can define this operator on test functions $f : \mathbb{R}_+ \times \mathbb{N}_0 \to \mathbb{R}$ by

$$\begin{aligned}\mathbb{A}_k f(s,i) = \frac{\partial f(s,i)}{\partial t} &+ \lambda(k+i)\left[\phi_{ik} f(s,i+1) - f(s,i)\right] \\ &+ (\mu+\psi+\omega)(k+i)\left[\mathbb{1}_{\{i>0\}}\kappa f(s,i-1) - f(s,i)\right].\end{aligned} \tag{2.6}$$

Here we implicitly assume that function f is continuously differentiable in the first coordinate. We now come to a very important proposition on which our whole analysis depends.

Proposition 2.1. *Let $u(t)$ be given by (2.4) and define the function $f_k : \mathbb{R}_+ \times \mathbb{N}_0 \to \mathbb{R}$ by*

$$f_k(s,i) = \frac{(k+i)!}{i!} u(t_{or} - s)^i.$$

Then the action of generator $\mathbb{A}_k$ on function f_k simplifies to

$$\mathbb{A}_k f_k(s,i) = k\left(2\lambda u(t_{or} - s) - (\lambda + \mu + \psi + \omega)\right) f_k(s,i).$$

Moreover the following is a positive martingale in the interval $[s_1, s_2]$ w.r.t. the filtration generated by process X,

$$M_s = p(t_{or} - s)^k f_k(s, X_s), \tag{2.7}$$

Remark 2.2. *The function f_k is a time-varying eigenfunction for the operator $\mathbb{A}_k$ and the corresponding eigenvalue is $k\left(2\lambda u(t_{or}-s)-(\lambda+\mu+\psi+\omega)\right)$.*

Proof See Appendix. □

Note that the fact that M_s is a martingale implies that

$$\mathbb{E}(M_{s_2}) = \mathbb{E}(M_{s_1})$$

which yields

$$\mathbb{E}(f_k(s_2, X_{s_2})) = \left(\frac{p(t_{or}-s_1)}{p(t_{or}-s_2)}\right)^k \mathbb{E}(f_k(s_1, X_{s_1})). \tag{2.8}$$

Reversing the direction of time and letting $t_1 = (t_{or}-s_2)$ and $t_2 = (t_{or}-s_1)$, conditioning on $X_{s_1} = i$, and exploiting the time-homogeneity of the Markov process $(X_s)_{s\geq 0}$ we can express (2.8) as

$$\sum_{j\geq 0} \frac{(k+j)!}{j!}\, u_{t_1}^j\, \mathbb{P}_i(X(t_2-t_1)=j) = \left(\frac{p_{t_2}}{p_{t_1}}\right)^k \frac{(k+i)!}{i!}\, u_{t_2}^i, \tag{2.9}$$

where the subscript i denotes that the initial state is $X_0 = i$. This formula will help us provide an alternative derivation of the probability density of a reconstructed tree without occurrence data originally introduced by Stadler (2010).

3. Revisiting the probability density of the reconstructed tree with samples through time

We assume in this section that $\omega = 0$, and there is no occurrence data. We wish to offer an alternative derivation of the probability density of the reconstructed tree spanning both ρ-sampled and ψ-sampled individuals as in Stadler (2010).

Let $t_0 < t_1 < t_2 < \cdots < t_n$ be the ordered set of times at which the tree events occur backward in time starting from the present time $t_0 = 0$ and culminating at $t_n = t_{or}$. For any $t \in [0, t_{or}]$, call k_t the number of observed lineages at time t in $\mathcal{T}$ and let $L_t^{(i)} = \mathbb{P}\left(\mathcal{T}_t^{\downarrow} \mid I_t = k_t + i\right)$ be the probability for the observed tree *below* time t (i.e. in the time-interval $[0,t]$) when the total number of lineages is $k_t + i$ at time t. Clearly the probability density of the observed tree is $L_{t_{or}}^{(0)} = \mathbb{P}\left(\mathcal{T}_{t_{or}}^{\downarrow} \mid I_{t_{or}} = 1\right)$, and to obtain this quantity we would like to study how L_t evolves with time t and how it depends on the state i.

We now introduce an *ansatz* for the form of $L_t^{(i)}$, given by

$$L_t^{(i)} = \frac{(k+i)!}{i!}\, u_t^i\, W(t), \tag{3.10}$$

where $W(t)$ is some real-valued *weight* function that only depends on time t but not on i or k.

Let us consider the interval (t_{h-1}, t_h) for some $h \geq 1$. In this interval the number of observed lineages is constant and equal to k. Using ansatz (3.10) and applying the Markov property we obtain

$$\begin{aligned}\frac{(k+i)!}{i!}\, u_{t_h}^i W(t_h^-) &= L_{t_h^-}^{(i)} \\ &= \sum_{j\geq 0} \mathbb{P}_i(X(t_h - t_{h-1}) = j)\; L_{t_{h-1}^+}^{(j)} \\ &= \sum_{j\geq 0} \mathbb{P}_i(X(t_h - t_{h-1}) = j)\; \frac{(k+j)!}{j!}\; u_{t_{h-1}}^j\; W(t_{h-1}^+) \\ &= \left(\frac{p(t_h)}{p(t_{h-1})}\right)^k\; \frac{(k+i)!}{i!}\; u_{t_h}^i\; W(t_{h-1}^+),\end{aligned}$$

where the last equality follows from (2.9). This proves that under ansatz (3.10)

$$W(t_h^-) = \left(\frac{p(t_h)}{p(t_{h-1})}\right)^k W(t_{h-1}^+). \tag{3.11}$$

We now examine how $W(t^+)$ and $W(t^-)$ are related, for any time t at which we observe an event. Suppose first that the event at time t is a ψ-event. Then $k^+ = k^- + 1$ and if the total number of lineages is $k^+ + i$ at time t_h^+ then the rate at which this event happens is $\psi(k^+ + i)$ and subsequently the total number of lineages falls to $(k^- + i)$. This shows that

$$\begin{aligned}\frac{(k^+ + i)!}{i!}\; u_t^i\; W(t^+) &= L_{t^+}^{(i)} \\ &= \psi(k^+ + i)\; L_{t^-}^{(i)} \\ &= \psi(k^- + 1 + i)\; \frac{(k^- + i)!}{i!}\; u_t^i\; W(t^-) \\ &= \psi \frac{(k^+ + i)!}{i!}\; u_t^i\; W(t^-),\end{aligned}$$

and hence

$$W(t^+) = \psi W(t^-). \tag{3.12}$$

Now suppose the event at time t is a λ-event. Then $k^+ = k^- - 1$ and if the total number of lineages is $(k^+ + i)$ at time t^+ then the rate at which this event happens is $\lambda(k^+ + i)$ and with probability $1/((k^+ + i)(k^+ + 1 + i))$ this event is a coalescent event between two observed ordered lineages. Once this event happens the total number of lineages increments to $(k^- + i)$. Therefore

$$\begin{aligned}\frac{(k^+ + i)!}{i!}\; u_t^i\; W(t^+) &= L_{t^+}^{(i)} \\ &= \lambda(k^+ + i) \left(\frac{1}{(k^+ + i)(k^+ + 1 + i)}\right) L_{t^-}^{(i)} \\ &= \frac{\lambda}{k^+ + 1 + i} \frac{(k^- + i)!}{i!}\; u_t^i\; W(t^-)\end{aligned}$$

$$= \lambda \frac{(k^+ + i)!}{i!}\, u_t^i\, W(t^-),$$

which proves that

$$W(t^+) = \lambda W(t^-). \tag{3.13}$$

Let k_0 be the number of extant lineages at time $t_0 = 0$. Using relations (3.11), (3.12) and (3.13) we can propagate the weight function $W(t)$ backward in time starting from $W(t_0^+) = \rho^{k_0}$ and ending at $W(t_n^-)$ which is equal to the tree density $\mathcal{L}(\mathcal{T})$. This backward propagation scheme is described in Algorithm 1 and it yields a closed-form formula very similar to theorem 3.5 in Stadler (2010). Since individuals are here removed upon sampling, our result can be derived from theorem 3.5 by setting $k = 0$ and dropping $p_0(y_i)$ factors corresponding to the death of ψ-sampled leaves.

Algorithm 1 Computes the probability density $\mathcal{L}(\mathcal{T})$.

Input: Observed tree $\mathcal{T}$, and parameters $t_{or}, \lambda, \mu, \psi, \rho$.
Output: The density value $\mathcal{L}(\mathcal{T})$.

Set $W(t_0^+) = \rho^{k_0}$
for h = 1,. . . ,n **do**
    Set

$$W(t_h^-) = \left(\frac{p(t_h)}{p(t_{h-1})}\right)^{k_h} W(t_{h-1}^+)$$

    if h = n **then**
        return $W(t_n^-)$
    else if t_h is a ψ-event **then**
        Set $W(t_h^+) = \psi W(t_h^-)$
    else t_h is a λ-event
        Set $W(t_h^+) = \lambda W(t_h^-)$
    end if
end for

4. The density of the reconstructed tree and case count record

We now consider the scenario where $\omega \neq 0$ and we have occurrence data $\mathcal{O}$ along with the observed lineage tree $\mathcal{T}$. As in the previous section, let $t_0 < t_1 < t_2 < \cdots < t_n$ be the ordered set of times at which the events occur backward in time starting from the present time $t_0 = 0$ and culminating at $t_n = t_{or}$. For any $t \in [0, t_{or}]$, let $L_t^{(i)} = \mathbb{P}\left(\mathcal{T}_t^{\downarrow}, \mathcal{O}_t^{\downarrow} \mid I_t = k_t + i\right)$ be the probability for the observed lineage tree and occurrence data *below* time t (i.e. in the time-interval $[0, t]$) when the number of observed lineages in $\mathcal{T}$ is k_t at time t and the number of hidden lineages is i. Clearly the probability density of the observed tree and occurrence data is again $L_{t_{or}}^{(0)} = \mathbb{P}\left(\mathcal{T}_{t_{or}}^{\downarrow}, \mathcal{O}_{t_{or}}^{\downarrow} \mid I_{t_{or}} = 1\right)$, and to obtain this quantity we would like to study

how L_t evolves with time t and how it depends on the state i. It will become evident that ansatz (3.10) will not work for this probability, and so we generalize this ansatz as

$$L_t^{(i)} = \sum_{l=0}^{q} \frac{(k+i)!}{(i-l)!}\, u_t^{i-l}\, W_l(t), \tag{4.14}$$

where q is a nonnegative integer, k is the number of observed lineages at time t, and $W(t) = (W_0(t), \ldots, W_q(t))$ is some vector-valued *weight* function that only depends on time t but not on i or k. Note that for $q = 0$, this ansatz becomes the previous ansatz (3.10).

4.1. Backward evolution at punctual events

We first examine how $W(t^+)$ and $W(t^-)$ are related when t is an event time. Suppose first that the event at time t is a ψ-event. Then $k^+ = k^- + 1$ and if the total number of lineages is $k^+ + i$ at time t^+ then the rate at which this event happens is $\psi(k^+ + i)$ and subsequently the total number of lineages falls to $(k^- + i)$. This shows that

$$\begin{aligned}
\sum_{l=0}^{q} \frac{(k^+ + i)!}{(i-l)!}\, u_{t^+}^{i-l}\, W_l(t^+) &= L_{t^+}^{(i)} \\
&= \psi(k^+ + i) L_{t^-}^{(i)} \\
&= \psi(k^+ + i) \sum_{l=0}^{q} \frac{(k^- + i)!}{(i-l)!}\, u_t^{i-l}\, W_l(t^-) \\
&= \psi \sum_{l=0}^{q} \frac{(k^+ + i)!}{(i-l)!}\, u_t^{i-l}\, W_l(t^-),
\end{aligned}$$

and hence

$$W(t^+) = \psi W(t^-). \tag{4.15}$$

Now suppose that the event at time t is a λ-event. Then $k^+ = k^- - 1$ and if the total number of lineages is $(k^+ + i)$ at time t^+ then the rate at which this event happens is $\lambda(k^+ + i)$ and with probability $1/((k^+ + i)(k^+ + i + 1))$ this event is a coalescent event between two observed ordered lineages. Once this event happens the total number of lineages increments to $(k^- + i)$. Therefore

$$\begin{aligned}
\sum_{l=0}^{q} \frac{(k^+ + i)!}{(i-l)!}\, u_t^{i-l}\, W_l(t^+) &= L_{t^+}^{(i)} \\
&= \lambda(k^+ + i) \left(\frac{1}{(k^+ + i)(k^+ + 1 + i)} \right) L_{t^-}^{(i)} \\
&= \frac{\lambda}{k^+ + 1 + i} \sum_{l=0}^{q} \frac{(k^- + i)!}{(i-l)!}\, u_t^{i-l}\, W_l(t^-) \\
&= \lambda \sum_{l=0}^{q} \frac{(k^+ + i)!}{(i-l)!}\, u_t^{i-l}\, W_l(t^-),
\end{aligned}$$

which proves that

$$W(t^+) = \lambda W(t^-). \tag{4.16}$$

So far the transition conditions (4.15) and (4.16) are identical to what we encountered in the previous section. However the difference comes for ω-events as we now discuss. Suppose t is a ω-event. Then $k^+ = k^-$ and if the total number of lineages is $(k^+ + i)$ at time t^+ then the rate at which this event happens is $\omega(k^+ + i)$ and subsequently the total number of lineages falls to $(k^+ + i - 1)$. This shows that

$$\begin{aligned}
\sum_{l=0}^{q} \frac{(k^+ + i)!}{(i-l)!}\, u_t^{i-l}\, W_l(t^+) &= L_{t^+}^{(i)} \\
&= \omega(k^+ + i) L_{t^-}^{(i-1)} \\
&= \omega(k^+ + i) \sum_{l=0}^{q} \frac{(k^+ + i - 1)!}{(i-1-l)!}\, u_t^{i-1-l}\, W_l(t^-) \\
&= \omega \sum_{l=0}^{q} \frac{(k^+ + i)!}{(i-(l+1))!}\, u_t^{i-(l+1)}\, W_l(t^-),
\end{aligned}$$

and hence

$$W(t^+) = \omega \mathbb{S} W(t^-), \tag{4.17}$$

where $\mathbb{S}$ is the shift-operator defined by

$$\mathbb{S}(v_1, \ldots, v_n) = (0, v_1, \ldots, v_n) \quad \text{for any} \quad (v_1, \ldots, v_n) \in \mathbb{R}^n. \tag{4.18}$$

Setting $q = 0$ in this calculation shows that the simpler ansatz (3.10), in which $W(t)$ is a scalar function instead of a vector-valued function, is not compatible with the requirement that $W(t)$ is not a function of i or k.

4.2. Backward evolution on a time interval without punctual events

Now that we have the transition conditions, (4.15), (4.16) and (4.15), we can propagate the weight function $W(t)$ backward in time, provided we can evaluate how it evolves in a time-interval (t_{h-1}, t_h) for any $h \geq 1$, and this backward evolution preserves our ansatz (4.14). In order to study this backward evolution we need some new notation and a simple lemma, which we now provide.

For any time $t \geq 0$ and $\theta \in (0, 1)$, define

$$c_2(\theta) = \frac{\lambda + \mu + \omega + \psi - 2\lambda\theta}{c_1}, \tag{4.19}$$

$$u(t, \theta) = \frac{1}{2\lambda} \left[\lambda + \mu + \omega + \psi + c_1 \frac{e^{-c_1 t}(1 - c_2(\theta)) - (1 + c_2(\theta))}{e^{-c_1 t}(1 - c_2(\theta)) + (1 + c_2(\theta))} \right], \tag{4.20}$$

$$\text{and} \quad p(t,\theta) = \frac{4(1-\theta)}{2(1-c_2(\theta)^2) + e^{-c_1 t}(1-c_2(\theta))^2 + e^{c_1 t}(1+c_2(\theta))^2} \tag{4.21}$$

where c_1 is as given in (2.3). Notice that if we set $\theta = 1-\rho$ then $c_2(\theta)$, $u(t,\theta)$ and $p(t,\theta)$ become identical to c_2, $u(t)$ and $p(t)$ defined in Section 2. Henceforth for any $k \in \mathbb{N}_0$ we also define the ratio

$$R_k(t,\theta) = \left(\frac{p(t,\theta)}{p(0,\theta)}\right)^k, \tag{4.22}$$

and when $k=1$, we drop the subscript and refer to $R_k(t,\theta)$ as $R(t,\theta)$. To ease further algebra, we also name here the function corresponding to the denominator of $p(t,\theta)$, namely $q(t,\theta) = 4(1-\theta)/p(t,\theta)$. The next lemma gives us analytical expressions for the higher order partial derivatives of $R_k(t,\theta)$ and $u(t,\theta)$ w.r.t. θ.

Lemma 4.1. *Let $u(t,\theta)$ and $R_k(t,\theta)$ be as defined by (4.20) and (4.22) respectively. Then for any $n = 1,2,\ldots$ we have the following:*

(A) Let $\mathcal{J}_n = \{(j_1,j_2) \in \mathbb{N}_0^2 : j_1 + 2j_2 = n\}$. Then $\partial_\theta^n R_k(t,\theta)$ is given by

$$\partial_\theta^n R_k(t,\theta) = n! R_k(t,\theta) \sum_{(j_1,j_2)\in\mathcal{J}_n} \frac{(-1)^{j_1+j_2}}{2^{j_2}} \frac{(j_1+j_2+k-1)!}{j_1! j_2! (k-1)!} \left(\frac{\partial_\theta q(t,\theta)}{q(t,\theta)}\right)^{j_1} \left(\frac{\partial_\theta^2 q(t,\theta)}{q(t,\theta)}\right)^{j_2}.$$

(B) The quantity $\partial_\theta^n u(t,\theta)$ is given by

$$\begin{aligned}\partial_\theta^n u(t,\theta) = \Big[&-\frac{c_1}{8\lambda}\left(e^{c_1 t}(1+c_2(\theta))^2 - e^{-c_1 t}(1-c_2(\theta))^2\right)\partial_\theta^n R(t,\theta) \\ &+\frac{n}{2}\left(e^{c_1 t} + e^{-c_1 t} + c_2(\theta)(e^{c_1 t} - e^{-c_1 t})\right)\partial_\theta^{n-1} R(t,\theta) \\ &-\frac{\lambda n(n-1)}{2c_1}\left(e^{c_1 t} - e^{-c_1 t}\right)\partial_\theta^{n-2} R(t,\theta)\Big].\end{aligned}$$

Proof See Appendix. □

Recall the formula (2.9), for the probability evolution on a time-interval $[t_1,t_2]$ on which the number of observed lineages remains fixed at k, and call now $\theta = u(t_1, 1-\rho)$. Appealing to the semi-group property of solutions to ODEs (2.1)-(2.2) we get

$$u(t_2,1-\rho) = u(t_2-t_1, u(t_1,1-\rho)) = u(t_2-t_1,\theta) \quad \text{and} \quad \frac{p(t_2,1-\rho)}{p(t_1,1-\rho)} = \frac{p(t_2-t_1,\theta)}{p(0,\theta)}.$$

Hence we can rewrite (2.9) as

$$\sum_{j\geq 0} \frac{(k+j)!}{j!}\,\theta^j \mathbb{P}_i(X(t)=j) = \frac{(k+i)!}{i!}\,u(t,\theta)^i\; R_k(t,\theta), \tag{4.23}$$

where $t = (t_2 - t_1)$. Note that the probability $\mathbb{P}_i(X(t)=j)$ does not depend on θ. Differentiating (4.23) l times w.r.t. θ we obtain

$$\sum_{j\geq 0} \frac{(k+j)!}{j!}\frac{j!}{(j-l)!}\,\theta^{j-l}\;\mathbb{P}_i(X(t)=j) = \frac{(k+i)!}{i!}\sum_{m=0}^{l}\binom{l}{m}\partial_\theta^m\left[u(t,\theta)^i\right]\partial_\theta^{l-m} R_k(t,\theta). \tag{4.24}$$

Applying the *Faà di Bruno's* formula (see Fraenkel (1978)), for any $m = 1, 2, \dots$ yields

$$\partial_\theta^m \left[u(t,\theta)^i\right] = \sum_{n=0}^{m} \frac{i!}{(i-n)!} u(t,\theta)^{i-n} \mathcal{B}_{m,n}(\partial_\theta u(t,\theta), \partial_\theta^2 u(t,\theta), \dots, \partial_\theta^{m-n+1} u(t,\theta)), \tag{4.25}$$

where $\mathcal{B}m, n(x_1, x_2, \dots, x_{m-n+1})$ is the incomplete Bell polynomial. Such polynomials can be computed efficiently via a recurrence relation

$$\mathcal{B}_{m,n}(x_1, x_2, \dots, x_{m-n+1}) = \sum_{i=0}^{m-n+1} \binom{m-1}{n-1} x_i \mathcal{B}_{m-i,n-1}(x_1, \dots, x_{m-n-i}),$$

where $\mathcal{B}_{0,0} = 1$ and $\mathcal{B}_{m,0} = \mathcal{B}_{0,m} = 0$ for each $m \geq 1$. Henceforth we denote

$$\mathbf{B}_{mn}(t,\theta) = \mathcal{B}_{m,n}(\partial_\theta u(t,\theta), \partial_\theta^2 u(t,\theta), \dots, \partial_\theta^{m-n+1} u(t,\theta)).$$

Substituting (4.25) in (4.24) we obtain

$$\begin{aligned}
&\sum_{j\geq 0} \frac{(k+j)!}{j!} \frac{j!}{(j-l)!} \, \theta^{j-l} \, \mathbb{P}_i(X(t) = j) \\
&= \frac{(k+i)!}{i!} \sum_{m=0}^{l} \sum_{n=0}^{m} \binom{l}{m} \frac{i!}{(i-n)!} \, u(t,\theta)^{i-n} \, \mathbf{B}_{mn}(t,\theta) \partial_\theta^{l-m} R_k(t,\theta) \\
&= \frac{(k+i)!}{i!} \sum_{n=0}^{l} \frac{i!}{(i-n)!} \, u(t,\theta)^{i-n} \sum_{m=n}^{l} \binom{l}{m} \mathbf{B}_{mn}(t,\theta) \partial_\theta^{l-m} R_k(t,\theta).
\end{aligned} \tag{4.26}$$

Letting

$$\mathbf{C}_{ln}(t,\theta,k) = \sum_{m=n}^{l} \binom{l}{m} \mathbf{B}_{mn}(t,\theta) \partial_\theta^{l-m} R_k(t,\theta), \tag{4.27}$$

we can express (4.26) as

$$\sum_{j\geq 0} \frac{(k+j)!}{(j-l)!} \, \theta^{j-l} \, \mathbb{P}_i(X(t) = j) = \sum_{n=0}^{l} \frac{(k+i)!}{(i-n)!} \, u(t,\theta)^{i-n} \, \mathbf{C}_{ln}(t,\theta,k). \tag{4.28}$$

This formula will play a critical role in determining the backward propagation of the vector-valued weight function $W(t) = (W_0(t), \dots, W_q(t))$ between the transition points.

Indeed, let us consider the interval (t_{h-1}, t_h) for some $h \geq 1$. In this interval the number of observed lineages is constant and we assume that it is equal to k_h. Also let $\theta_h = u(t_h, 1-\rho)$ for each h. Using ansatz (4.14) and applying the Markov property we obtain

$$\begin{aligned}
\sum_{l=0}^{q} \frac{(k_h+i)!}{(i-l)!} \, \theta_h^{i-l} \, W_l(t_h^-) &= L^{(i)}_{t_h^-} \\
&= \sum_{j\geq 0} \mathbb{P}_i(X(t_h - t_{h-1}) = j) L^{(i)}_{t_{h-1}^+}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{l=0}^{q} W_l(t_{h-1}^+) \sum_{j\geq 0} \mathbb{P}_i(X(t_h - t_{h-1}) = j) \frac{(k_h + j)!}{(j-l)!}\ \theta_{h-1}^{j-l} \\
&= \sum_{l=0}^{q} \sum_{n=0}^{l} W_l(t_{h-1}^+) \frac{(k_h + i)!}{(i-n)!}\ \theta_h^{i-n} \mathbf{C}_{ln}(t_h - t_{h-1}, \theta_{h-1}, k_h) \\
&= \sum_{n=0}^{q} \frac{(k_h + i)!}{(i-n)!}\ \theta_h^{i-n} \ \sum_{l=n}^{q} W_l(t_{h-1}^+) \mathbf{C}_{ln}(t_h - t_{h-1}, \theta_{h-1}, k_h),
\end{aligned}
$$

where the second-last equality is due to formula (4.28). This calculation proves that ansatz (4.14) is preserved by the backward evolution of the weight vector $W(t)$ and

$$
W(t_h^-) = W(t_{h-1}^+) \mathbf{C}(t_h - t_{h-1}, \theta_{h-1}, k_h), \tag{4.29}
$$

where $W(t_h^-) = (W_0(t_h^-), \ldots, W_q(t_h^-))$ and $W(t_{h-1}^+) = (W_0(t_{h-1}^+), \ldots, W_q(t_{h-1}^+))$ are $(q+1)$-dimensional row vectors and $\mathbf{C}(t_h - t_{h-1}, \theta_{h-1}, y_h)$ is the $(q+1) \times (q+1)$ lower-triangular matrix matrix whose entries are given by $\mathbf{C}_{ln}(t_h - t_{h-1}, \theta_{h-1}, k_h)$ for $l \geq n$ and 0 for $l < n$. Note that by using Lemma 4.1, this matrix can be analytically computed.

4.3. Summary of the likelihood computation

Let k_0 be the number of extant and sampled lineages at time $t_0 = 0$. Using relations (4.29), (4.15), (4.16) and (4.17) we can propagate the vector-valued weight function $W(t)$ backward in time starting from $W(t_0^+)$ whose entries are all equal to ρ^{k_0} and ending at $W(t_{or}^-)$, whose first component $W_1(t_{or}^-)$ is equal to the target probability density $\mathcal{L}(\mathcal{T}, \mathcal{O})$. Note that due to the presence of the shift operator $\mathbb{S}$ in (4.17) the dimensional of $W(t)$ is $(q+1)$ where q is the number of occurrence events in the time-interval $[0, t]$. This backward propagation scheme is described in Algorithm 2 and it simplifies to Algorithm 1 in the absence of occurrence data.

Algorithm 2 Computes the probability density $\mathcal{L}(\mathcal{O}, \mathcal{T})$.

Input: Observed $(\mathcal{O}, \mathcal{T})$, and parameters $t_{or}, \lambda, \mu, \psi, \omega, \rho$.
Output: The probability density $\mathcal{L}(\mathcal{T}, \mathcal{O})$.

Set $q = 0$, $\theta = (1 - \rho)$ and $W(t_0^+) = \rho^{k_0}$.
for h $= 1, \ldots,$ n **do**
    Compute the $(q + 1) \times (q + 1)$ lower-triangular matrix $C(t_h - t_{h-1}, \theta, k_h)$ whose entry $C_{ln}(t_h - t_{h-1}, \theta, k_h)$ for $l \geq n$ is given by (4.27).
    Viewing $W(t_{h-1}^+) = (W_0(t_{h-1}^+), \ldots, W_q(t_{h-1}^+))$ as a $1 \times (q + 1)$ vector, set

$$W(t_h^-) = W(t_{h-1}^+)C(t_h - t_{h-1}, \theta, k_h).$$

    if i = n **then**
        return $W_1(t_n^-)$ which is the first component of the weight vector $W(t_n^-)$.
    else if t_h is a ψ-event **then**
        Set $W(t_h^+) = \psi W(t_h^-)$
    else if t_h is a λ-event **then**
        Set $W(t_h^+) = \lambda W(t_h^-)$
    elset_h is a ω-event
        Set $W(t_h^+) = \omega \mathbb{S} W(t_h^-)$, where $\mathbb{S}$ is the shift operator (4.18).
        Set $q = q + 1$.
    end if
    Set $\theta = u(t_h - t_{h-1}, \theta)$.
end for

4.4. Numerical implementation and sanity check

The main algorithm to compute the probability density of $(\mathcal{O}, \mathcal{T})$ has been implemented numerically and is available on GitHub: https://github.com/ankitgupta83/tree-and-occurrences/. As a check that the method leads to correct values, we compared the results obtained with our analytical calculation (Algorithm 2) with the ones obtained using Monte Carlo simulations as described in Vaughan et al. (2019) and implemented here in C++.

We consider two toy examples depicted in Figure 2(A) for which we calculate the probability density. The comparisons are shown in Figure 2(B) and Figure 2(C), and one can see that there is a close match between numerical values computed analytically and those estimated with simulations.

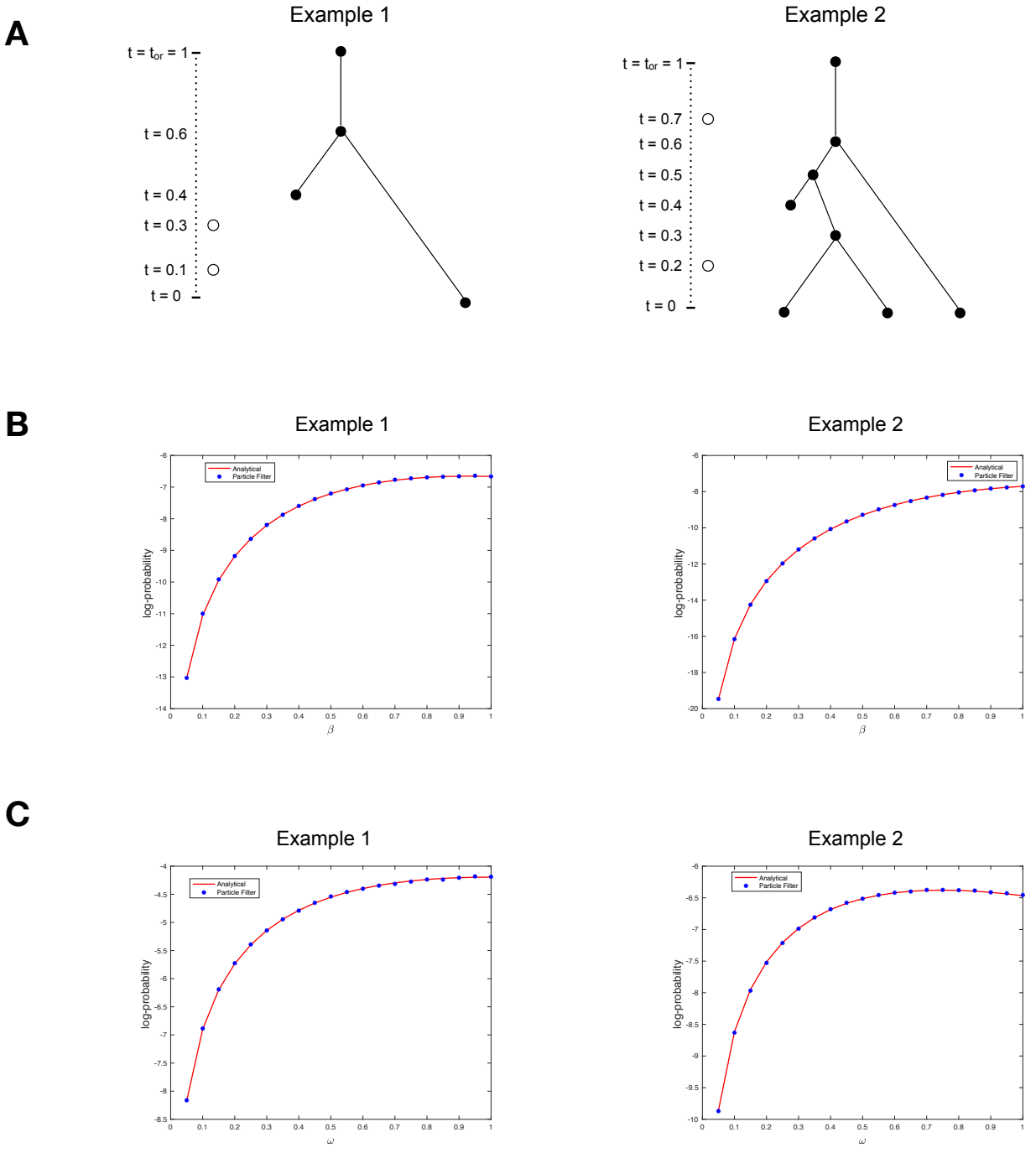

Figure 2: (A) Two toy examples of reconstructed trees along with occurrence data (open circles). (B) Comparison between probability densities obtained analytically with Algorithm 2 and the corresponding probability densities obtained with simulations as described in Vaughan et al. (2019). The parameter λ varies between 0 and 2, while the other parameters are set to $\psi = 0.1$, $\mu = 0$, $\omega = 0.1$ and $\rho = 0.7$. (C) Same as (B) except now the parameter ω varies between 0 and 1, while the other parameters are set to $\psi = 0.1$, $\mu = 0$, $\lambda = 1.25$ and $\rho = 0.7$.

5. Applications

Our results pave the way to different applications that we present in this section.

5.1. Different flavors of the likelihood

Following Stadler (2010), we first point out that our main result, Algorithm 2, allows us to compute the density of the observations, given a model with fixed time of origin t_{or}. This quantity, that we call $\mathcal{L}$ below, can be seen as an extension of Theorem 3.5 in Stadler (2010) (while assuming that samples correspond to death events), when occurrences can be observed.

Depending on the analysis that one wants to perform, it might be desirable to condition this density on various events, including,

1. the number of ρ-sampled leaves at present,

$$\mathcal{L}(\mathcal{O}, \mathcal{T} \mid \{\rho\text{-sampling } n \text{ individuals } \}) = \frac{\mathcal{L}}{\mathbb{P}(\{\rho\text{-sampling } n \text{ individuals } \})}$$

   where the denominator is a well-known quantity which can, e.g., be found in Theorem 3.3 in Stadler (2010), replacing μ by $\mu + \psi + \omega$ in our case.

2. the survival of the process up to the present,

$$\mathcal{L}(\mathcal{O}, \mathcal{T} \mid \{ \text{ survival of at least one individual at present } \}) = \frac{\mathcal{L}}{1 - u_{t_{or}}}$$

3. the survival of two lineages starting at time t_{mrca},

$$\mathcal{L}(\mathcal{O}, \mathcal{T} \mid \{ \text{ survival of two lineages originating at } t_{mrca}\}) = \frac{\mathcal{L}}{(1 - u_{t_{or}})^2} \frac{p(t_{mrca})}{\lambda p(t_{or})}$$

In each case, since the event we condition on is in the observed data, one only needs to divide by the probability of the event. These are only three possibilities that one could use. In the context of phylodynamics, conditioning on the survival of the process, or on the survival of two lineages starting at t_{mrca} is common practice, since we study an evolutionary process precisely because it survived.
We also stress that the probability density provided in this manuscript applies to an oriented and unlabelled tree. If one wishes to compare the probability densities of trees under different generating models, then it might be useful to add a combinatorial factor for dealing, e.g. with unoriented and labelled trees, as described in Stadler (2010).

5.2. Maximum likelihood estimators

The previous probability densities can readily be used to estimate parameters of the model from reconstructed trees and occurrences. Taken as a function of parameters, the density is indeed called likelihood, and

maximum likelihood estimators can be obtained by maximizing the likelihood function over the parameters. Since it does not appear straight-forward to identify optimal parameter configurations analytically, we suggest relying on numerical optimizers instead. In this context, computing quickly the likelihood value using Algorithm 2 in place of the more computationally intensive Monte-Carlo algorithm proposed by Vaughan et al. (2019), can prove essential, since the optimization requires many calls to the function.

5.3. Bayesian analysis

This density is also a key component when using this model in a Bayesian framework, where one can estimate the parameters of the model directly from the occurrences and the sequencing data rather than fixing a tree. Suppose we observe the sampling times of occurrences $\mathcal{O}$ (individuals without any measurement data), the sampling times and genotypic or phenotypic measurements of m ψ-sampled individuals, and the genotypic or phenotypic measurements of n ρ-sampled individuals. We call $\mathcal{A}$ the sequencing data, summarizing all the genotypic or phenotypic measurements (e.g. a nucleotide alignment of pathogens, or a collection of morphological traits for fossil species, or any combination of those), and θ all parameters relating to the model of character evolution along a tree.

Loosely defining f to represent all probability densities involved, and relying on the name of the random variable to know which one we refer to, one is generally interested in sampling from,

$$f(\mathcal{T}, \theta, \lambda, \mu, \rho, \psi, \omega, t_{or} \mid \mathcal{O}, \mathcal{A}) \propto f(\mathcal{A} \mid \mathcal{T}, \theta) f(\mathcal{O}, \mathcal{T} \mid \lambda, \mu, \rho, \psi, \omega, t_{or}) f(\lambda, \mu, \rho, \psi, \omega, \theta, t_{or}),$$

with $f(\lambda, \mu, \rho, \psi, \omega, \theta, t_{or})$ being the prior distribution on the model parameters. Standard MCMC (Markov-Chain Monte-Carlo) algorithms can be used to sample from this posterior distribution. As a result, from sequencing and occurrence data, we can directly obtain the marginal distribution of birth-death parameters integrated over the posterior distribution of trees.

5.4. Simulations of the process

While the raw process is quite easy to simulate forward-in-time, it can be more difficult to simulate it under a different conditioning event $\mathcal{C}$, relating for example to the number of ρ-, ψ-, or ω-samples. Naive rejection sampling is not an efficient option if one wishes to condition on an event happening with very low probability. Adapting the particle filtering algorithm developed by Vaughan et al. (2019) would be a much better option.

Another option is to use the results of this paper and directly sample from the density $f(\mathcal{O}, \mathcal{T} \mid \lambda, \mu, \rho, \psi, \omega, t_{or}, \mathcal{C})$, using our ability to evaluate $f(\mathcal{O}, \mathcal{T} \mid \lambda, \mu, \rho, \psi, \omega, t_{or})$ quickly, within an MCMC algorithm with any reasonable movement proposal changing the reconstructed tree and possibly occurrence times or numbers while satisfying the constraint $\mathcal{C}$.

6. Discussion

In this study, we derived a closed-form probability density formula of a reconstructed tree and an occurrence record (i.e. a record of when cases occurred) under a linear birth-death model with sampling. This can readily be used for statistical purposes, to infer the parameters of the model in a maximum likelihood or a Bayesian framework.

In the context of epidemiology, this study offers a way to improve the accuracy of statistical estimates of key epidemiological parameters, such as the transmission and recovery rate (λ and $(\mu + \psi + \omega)$), and thus the basic reproductive number $R = \lambda/(\mu + \psi + \omega)$, using jointly the occurrence record and the phylogenetic tree reconstructed from pathogen sequences. This should be of use to health policy makers as well as epidemiologists, enabling them to jointly use the epidemiological data (occurrence record) as well as molecular data (genetic sequences), to recover the dynamic of an outbreak.

In the context of macroevolution, the present work contributes to the ongoing effort towards bridging the gap between inferences made from the fossil record and inferences made from contemporary data. It can be seen as an extension of the birth-death process with sampling through time (Stadler, 2010), allowing one to take into account fossil occurrences which evolutionary relationships to other taxa are not well resolved. However, we need to point out here that we assume removal of a sampled individual from the population, while a species continues to exist if a specimen is preserved in and observed from the fossil record. Numerical rather than our analytic treatment of the model can deal with non-removal upon sampling (Vaughan et al., 2019).

Many extensions of this model are possible. One of the simplest would be to consider time-varying rate parameters $\lambda_t, \mu_t, \psi_t, \omega_t$. The most widely used approach to do so in the recent literature has been to consider so-called *skyline* versions of birth-death processes (Stadler et al., 2013), meaning that the rate parameters are piecewise constant functions. Because we expressed all our results based on two key functions u_t and p_t, whose analytical expressions can easily be obtained considering piecewise constant rates, our central result would still hold. The challenge, however, would be to use such a model with an appropriate number of rate shifts so as not to overfit the data, although in a Bayesian context the number of rate shifts could also be estimated as part of a reversible jump MCMC scheme (Green, 1995).

Another important extension of the model would be to allow for the possibility not to remove individuals upon sampling. In fact, the model considered in Stadler (2010) does assume that ψ-sampled individuals keep living in the process. However, it is central for the derivation of our result that ω-sampled individuals are removed upon sampling. Our choice to remove all individuals upon sampling also seems reasonable for a number of applications in epidemiology, since sampled individuals are generally those that have sought medical attention, and - even if not quarantined - might be much less likely to spread the disease (but see

also (Gavryushkina et al., 2014)).

Recently, Stadler et al. (2018) introduced an other extension of the birth-death model with sampling through time to account for data on the lifetime of some individuals. It assumes that individuals can be sampled multiple times, and that the first and last sampling event are recorded along the tree. This extension of the model could be transposed in our setting, where the record of occurrences would become a record of sampled lifetime intervals. Such an approach would allow to use infection interval data or stratigraphic range data.

In summary, we present a way to analytically calculate the likelihood of phylogenetic trees together with occurrence data. So far, this likelihood could only be calculated analytically for the phylogenetic tree or for the occurrence data. Thus, we view this work as a major step towards coherent statistical analysis of different data sources under the birth-death model with sampling through time.

Acknowledgements

The authors are very grateful to Rachel Warnock for helpful comments on potential applications of the model.

References

Aldous, D.J., et al., 2001. Stochastic models and descriptive statistics for phylogenetic trees, from yule to today. Statistical Science 16, 23–34.

Alfaro, M.E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D.L., Carnevale, G., Harmon, L.J., 2009. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. P. Natl. Acad. Sci. USA 106, 13410–13414.

Ethier, S.N., Kurtz, T.G., 1986. Markov processes. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York. Characterization and convergence.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A., Phillimore, A.B., 2012. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. Proceedings of the Royal Society B: Biological Sciences 279, 1300–1309.

Fraenkel, L., 1978. Formulae for high derivatives of composite functions, in: Mathematical Proceedings of the Cambridge Philosophical Society, Cambridge University Press. pp. 159–165.

Gavryushkina, A., Heath, T.A., Ksepka, D.T., Stadler, T., Welch, D., Drummond, A.J., 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Systematic biology 66, 57–73.

Gavryushkina, A., Welch, D., Stadler, T., Drummond, A.J., 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLoS computational biology 10, e1003919.

Green, P., 1995. Reversible jump markov chain monte carlo computation and bayesian model determination. Biometrika 82, 711–732. doi:10.1093/biomet/82.4.711.

Heath, T.A., Huelsenbeck, J.P., Stadler, T., 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111, –2957.

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

Leventhal, G.E., Günthard, H.F., Bonhoeffer, S., Stadler, T., 2013. Using an epidemiological model for phylogenetic inference reveals density dependence in hiv transmission. Molecular biology and evolution 31, 6–17.

Maddison, W.P., Midford, P.E., Otto, S.P., 2007. Estimating a binary character's effect on speciation and extinction. Systematic Biology 56, 701–710. doi:10.1080/10635150701607033.

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

Nee, S., May, R.M., 1997. Extinction and the loss of evolutionary history. Science 278, 692–694.

Nee, S., May, R.M., Harvey, P.H., 1994. The reconstructed evolutionary process. Philosophical Transactions of the Royal Society of London B: Biological Sciences 344, 305–311.

Stadler, T., 2010. Sampling-through-time in birth–death trees. Journal of theoretical biology 267, 396–404.

Stadler, T., 2011. Mammalian phylogeny reveals recent diversification rate shifts. P. Natl. Acad. Sci. USA 108, 6187–6192.

Stadler, T., Gavryushkina, A., Warnock, R.C., Drummond, A.J., Heath, T.A., 2018. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. Journal of theoretical biology 447, 41–55.

Stadler, T., Kouyos, R., von Wyl, V., Yerly, S., Böni, J., Bürgisser, P., Klimkait, T., Joos, B., Rieder, P., Xie, D., et al., 2011. Estimating the basic reproductive number from viral sequence data. Molecular biology and evolution 29, 347–357.

Stadler, T., Kühnert, D., Bonhoeffer, S., Drummond, A.J., 2013. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences 110, 228–233.

Vaughan, T.G., Leventhal, G.E., Rasmussen, D.A., Drummond, A.J., Welch, D., Stadler, T., 2019. Estimating epidemic incidence and prevalence from genomic data. Molecular Biology and Evolution doi:10.1093/molbev/msz106.

Zhang, C., Stadler, T., Klopfstein, S., Heath, T.A., Ronquist, F., 2015. Total-evidence dating under the fossilized birth–death process. Systematic biology 65, 228–249.

Appendix

Proof of Proposition 2.1

Note that reversing the direction of time in (2.1) we can write the ODE for $u(t_{or}-s)$ as

$$\frac{du(t_{or}-s)}{dt} = -\mu + (\lambda+\mu+\omega+\psi)u(t_{or}-s) - \lambda u(t_{or}-s)^2 \tag{A.1}$$

Applying the generator $\mathbb{A}_k$ on function $f_k(s,i)$ we obtain

$$\begin{aligned}
\mathbb{A}_k f_k(s,i) = {} & \frac{(k+i)!}{i!} i u(t_{or}-s)^{i-1} \frac{du(t_{or}-s)}{dt} \\
& + \lambda(k+i)\left(\phi_{ik}\frac{(k+i+1)!}{(i+1)!}u(t_{or}-s)^{i+1} - \frac{(k+i)!}{i!}u(t_{or}-s)^i\right) \\
& + (\mu+\omega+\psi)(k+i)\left(\mathbb{1}_{\{i>0\}}\kappa\frac{(k+i-1)!}{(i-1)!}u(t_{or}-s)^{i-1} - \frac{(k+i)!}{i!}u(t_{or}-s)^i\right).
\end{aligned}$$

Since $i \geq 0$, noting that

$$(k+i)\frac{(k+i-1)!}{(i-1)!} = \frac{(k+i)!}{i!}i \qquad \text{and} \qquad (k+i)\phi_{ik}\frac{(k+i+1)!}{(i+1)!} = (2k+i)\frac{(k+i)!}{i!},$$

and using (A.1) we get

$$\begin{aligned}
\mathbb{A}_k f_k(s,i) = {} & -\mu\frac{(k+i)!}{i!}iu(t_{or}-s)^{i-1} \\
& + (\lambda+\mu+\omega+\psi)\frac{(k+i)!}{i!}iu(t_{or}-s)^i \\
& - \lambda\frac{(k+i)!}{i!}iu(t_{or}-s)^{i+1} + \lambda(2k+i)\frac{(k+i)!}{i!}u(t_{or}-s)^{i+1} \\
& - \lambda(k+i)\frac{(k+i)!}{i!}u(t_{or}-s)^i + \mu i\frac{(k+i)!}{i!}u(t_{or}-s)^{i-1} \\
& - (\mu+\omega+\psi)(k+i)\frac{(k+i)!}{i!}u(t_{or}-s)^i \\
= {} & -k(\lambda+\mu+\omega+\psi)\frac{(k+i)!}{i!}u(t_{or}-s)^i + 2k\lambda\frac{(k+i)!}{i!}u(t_{or}-s)^{i+1} \\
= {} & k\left(2\lambda u(t_{or}-s) - (\lambda+\mu+\psi+\omega)\right) f_k(s,k+i).
\end{aligned}$$

We now prove that M_s defined by (2.7) is a martingale w.r.t. the filtration $\mathcal{F}_s$ generated by process X. Note that since this process has generator $\mathbb{A}_k$ the following is a $\mathcal{F}_s$-martingale (see Chapter 4 in Ethier and Kurtz (1986))

$$m_s = f_k(s, X(s)) - f_k(s_1, X(s_1)) - \int_{s_1}^{s} k(2\lambda u(t_{or} - z) - (\lambda + \mu + \psi + \omega)) f_k(z, X(z)) dz.$$

Writing this equation in differential form we obtain

$$dm_s = df_k(s, X(s)) - k\left(2\lambda u(t_{or} - s) - (\lambda + \mu + \psi + \omega)\right) f_k(s, X(s)) ds.$$

Multiplying both sides by the integrating factor

$$J_s = \exp\left(- \int_{s_1}^{s} k\left(2\lambda u(t_{or} - z) - (\lambda + \mu + \psi + \omega)\right) dz \right)$$

we get

$$J_s dm_s = J_s df_k(s, X(s)) + f_k(s, X(s)) dJ_s = dJ_s f_k(s, X(s)).$$

Upon integration we see that $\int_{s_1}^{s} J_z dm_z$ is a martingale which implies that $J_s f_k(s, X(s))$ is a positive $\mathcal{F}_s$-martingale in the time-interval $[s_1, s_2]$. Note that J_s satisfies the ODE

$$dJ_s = -k\left(2\lambda u(t_{or} - s) - (\lambda + \mu + \psi + \omega)\right) J_s$$

which is similar to (2.2). Exploiting this similarity allows us to write

$$J_s = \left(\frac{p(t_{or} - s)}{p(t_{or} - s_1)} \right)^k$$

Now the fact that $J_s f_k(s, X(s))$ is a positive $\mathcal{F}_s$-martingale proves that M_s is also such a martingale. This completes the proof of this proposition. □

Proof of Lemma 4.1

Observe that

$$\partial_\theta c_2(\theta) = -\frac{2\lambda}{c_1}$$

and all higher order derivatives of $c_2(\theta)$ are zero, i.e. $\partial_\theta^m c_2(\theta) = 0$ for all $m \geq 2$. As $q(t, \theta)$ is a simple quadratic function of $c_2(\theta)$ we get from chain-rule for derivatives that

$$\begin{aligned}
\partial_\theta q(t, \theta) &= -\frac{4\lambda}{c_1} \left(e^{c_1 t} - e^{-c_1 t} + c_2(\theta)(e^{c_1 t} + e^{-c_1 t} - 2) \right) \\
\partial_\theta^2 q(t, \theta) &= \frac{8\lambda^2}{c_1^2} \left(e^{c_1 t} + e^{-c_1 t} - 2 \right), \\
\text{and} \quad \partial_\theta^m q(t, \theta) &= 0 \qquad \text{for} \qquad m = 3, 4, \ldots .
\end{aligned}$$

As $q(0,\theta) = 4$ is independent of θ we can express the derivatives of $R_k(t,\theta)$ as

$$\partial_\theta^n R_k(t,\theta) = 4^k \partial_\theta^n \left[\frac{1}{(q(t,\theta))^k}\right].$$

Using the derivatives of $q(t,\theta)$ computed above and the *Faà di Bruno's* formula (see Fraenkel (1978)) we obtain the expression for $\partial_\theta^n R_k(t,\theta)$ reported in part (A).

For part (B) observe that we can write $u(t,\theta)$ as

$$u(t,\theta) = \frac{\lambda+\mu+\omega+\psi}{2\lambda} - \frac{c_1}{8\lambda} r(\theta) R(t,\theta) \tag{A.2}$$

where

$$r(\theta) = e^{c_1 t}(1+c_2(\theta))^2 - e^{-c_1 t}(1-c_2(\theta))^2.$$

The derivatives of $r(\theta)$ can be obtained as

$$\partial_\theta r(\theta) = -\frac{4\lambda}{c_1}\left[e^{c_1 t} + e^{-c_1 t} + c_2(\theta)(e^{c_1 t} - e^{-c_1 t})\right],$$
$$\partial_\theta^2 r(\theta) = \frac{8\lambda^2}{c_1^2}\left[e^{c_1 t} - e^{-c_1 t}\right]$$
$$\text{and} \quad \partial_\theta^m r(\theta) = 0 \qquad \text{for} \qquad m = 3, 4, \ldots.$$

Applying the generalized product-rule for derivatives to formula (A.2) we can express $\partial_\theta^n u(t,\theta)$ as

$$\partial_\theta^n u(t,\theta) = -\frac{c_1}{8\lambda} \sum_{m=0}^{2} \binom{n}{m} \partial_\theta^m r(\theta) \partial_\theta^{n-m} R(t,\theta).$$

Substituting the values of $\partial_\theta^m r(\theta)$ and simplifying, we obtain the expression for $\partial_\theta^n u(t,\theta)$ reported in part (B). This completes the proof of this lemma. □