Swapping Birth and Death: Symmetries and Transformations in Phylodynamic Models

Tanja Stadler[1,*] Joseph Felsenstein[2], and Mike Steel[3]

[1] *Department for Biosystems Science and Engineering, ETH Zürich, Switzerland*

[2] *Department of Genome Sciences, University of Washington, USA*

[3] *Biomathematics Research Centre, University of Canterbury, New Zealand*

**tanja.stadler@bsse.ethz.ch*

December 12, 2018

Abstract

Stochastic birth–death models provide the foundation for studying and simulating evolutionary trees in phylodynamics. A curious feature of such models is that they exhibit fundamental symmetries when the birth and death rates are interchanged. In this paper, we explain and formally derive these transformational symmetries. We also show that these transformational symmetries (encoded in algebraic identities) are preserved even when taxa at the present are sampled with some probability. However, these extended symmetries require the death rate parameter to sometimes take a negative value. In the last part of this paper, we describe the relevance of these transformations and their application to computational phylodynamics, particularly to maximum likelihood and Bayesian inference methods, as well as to model selection. Phylodynamics, phylogenetics, speciation-extinction models, birth-death models, algebraic symmetries, maximum likelihood, Bayesian inference

1 Introduction

Linear birth–death models play a pivotal role in phylodynamics. These stochastic models provide a prior distribution on evolutionary trees (both the shape and edge length distribution) for Bayesian inference methods [20, 17]. Moreover, these models allow biologists to estimate key parameters of macroevolution (such as speciation rates corresponding to birth rates, and extinction rates corresponding to death rates) from reconstructed phylogenetic trees which were dated by fossil (or other time-sampled) evidence [10].

The study of such models dates back to some classical papers from the early to mid-20th century [21, 5, 6], and the application of these models to phylogenetics and phylodynamics flourished from the 1990s onwards [10, 11]. Further in-depth mathematical analysis [1, 8, 2, 9, 7] has extended our understanding of the properties of these models and extensions that allow more complex processes of birth and death.

In this paper, we identify and explore curious symmetries in fundamental birth–death model probability distributions when the birth and death rates (λ and μ) are swapped. We will start the paper by providing an intuitive account of this symmetry that seems at first a little surprising. We extend this to the more general setting where a third parameter is introduced — the sampling probability ρ of taxa sampled at the present — and show how analogous symmetries can be derived by a transformation that reduces these three parameters to just two (λ', μ'). One can view these as 'corrected' birth and death rates, except for the caveat that this new death rate μ' can now take negative values. A major advantage of working with the transformed pair of parameters (λ', μ') is that it captures the correct dimensionality of the process (namely 2), thereby avoiding the inherent redundancy present in the 3-dimensional parameterization that uses the triple (λ, μ, ρ). This viewpoint has implications for phylogenetic and phylodynamic inferences, both in the maximum likelihood and Bayesian settings, and we explore these implications in the latter part

of our paper.

2 Birth–death symmetries

Consider a phylogenetic tree that evolves from a single ancestral taxon according to a birth–death process, with a constant birth rate $\lambda \geq 0$ and a constant death rate $\mu \geq 0$. Suppose that at some time point in the tree, there are n taxa present. Let $p_{n,m}(t \mid \lambda, \mu)$ be the probability that at time t later, there will be m taxa present. These transition probabilities are classical and provide a foundation for phylodynamic models. However, the starting point for this paper is the following curious symmetry:

$$p_{11}(t \mid \lambda, \mu) = p_{11}(t \mid \mu, \lambda). \tag{1}$$

This equation states the surprising result that the probability of one individual having one surviving descendant after time t remains the same if we swap the birth rate (λ) and the death rate (μ). Thus a process with a birth rate of, say, 100 and a death rate of, say, 1 — a scenario with a very fast-growing population — has the same probability of having one surviving descendant as a process with a birth rate of 1 and a death rate of 100, a scenario where we know that the process eventually leads to extinction.

We will see that identities such as Eqn. (1) fall out from an algebraic analysis of birth–death models (provided later). Our aim in the meantime is to provide an intuitively transparent (but still rigorous) argument for Eqn. (1), as well as the following more general identity, namely that the probability of n individuals having n surviving descendants after time t is the same for birth rate λ and death rate μ or birth rate μ and death rate λ, for all $n \geq 1$.

Proposition 2.1. *For any non-negative value of λ, μ and any value of $n \geq 1$:*

$$p_{n,n}(t \mid \lambda, \mu) = p_{n,n}(t \mid \mu, \lambda).$$

We now provide a direct and intuitively-transparent proof of Proposition 2.1. We deal in detail with the case $n = 1$ (i.e. Eqn. (1)); however, the result for $n \geq 1$ follows by essentially applying the same idea. We start a birth–death process with one individual. The waiting time between 'events' (a birth event or death event) is $\exp(n(\lambda + \mu))$, where n is the number of individuals at the considered time point. Let $p = \frac{\lambda}{\lambda+\mu}$, and consider two different scenarios (one proceeds forward in time, the other backward):

- Scenario 1: The process starts at time 0 and is stopped at time $t > 0$. At an event, with probability p, we add an individual and, with probability $1 - p$. we remove an individual. Scenario 1 is a classic forward-in-time birth–death process.

- Scenario 2: The process starts at time $t > 0$ and is stopped at time 0. At an event, with probability $1 - p$ we add an individual and, with probability p, we remove an individual. Scenario 2 is a birth–death process in reversed time with the birth and death rates being interchanged compared with Scenario 1.

Intuitively, the result of the time-reversed process with birth and death being interchanged is analogous to the forward-in-time process. However, we justify this intuition by a formal argument showing that the probability of observing one individual after time t is the same under Scenario 1 and Scenario 2.

Consider some population size trajectory X that starts at time 0 with one individual and ends with one individual after time t. At each event, X can grow or decrease by one. Let the number of growth events be k, which therefore also equals

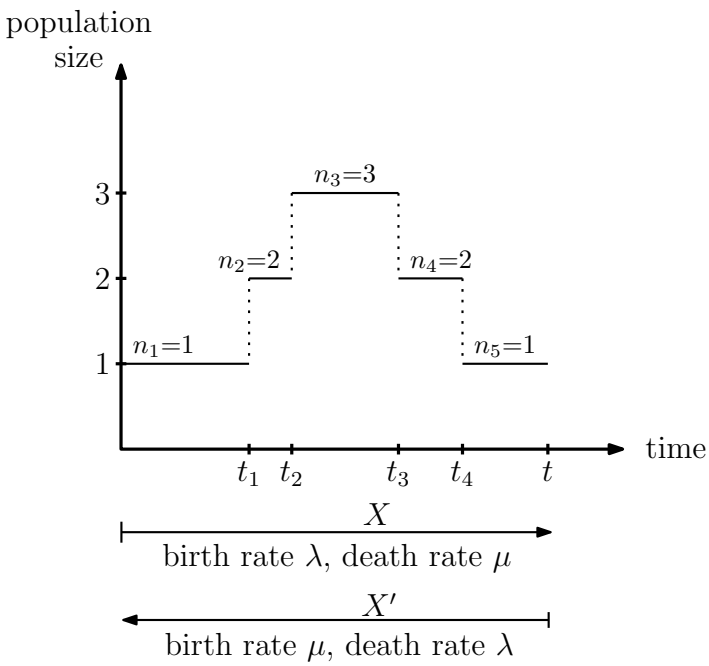

Figure 1: The forward-in-time birth–death process with realization X and the equivalent time-reversed process with interchanged rates and realization X'.

the number of death events. Denote the time of these $2k$ events by $t_1, t_2, \ldots t_{2k}$, and define $t_0 = 0$ and $t_{2k+1} = t$. See Figure 2 for an example with $k = 2$.

The probability density of X under Scenario 1, $L_1(X)$, is a product of the probability for the birth events, p^k, for the death events $(1-p)^k$, and the waiting times between events, $\prod_{i=1}^{2k}(\lambda+\mu)n_i e^{-(\lambda+\mu)n_i(t_i-t_{i-1})}$, where n_i is the number of individuals prior to the event at time t_i. Finally, the term $e^{-(\lambda+\mu)(t-t_{2k})}$ stipulates that no subsequent event happens after the event at time t_{2k}. In summary, the probability density of X under Scenario 1 for $k > 0$ is:

$$L_1(X) = p^k(1-p)^k(\lambda+\mu)e^{-(\lambda+\mu)((t_1-t_0)+(t_{2k+1}-t_{2k}))}\prod_{i=2}^{2k}(\lambda+\mu)n_i e^{-(\lambda+\mu)n_i(t_i-t_{i-1})}.$$

For $k = 0$, we have

$$L_1(X) = e^{-(\lambda+\mu)(t_{2k+1}-t_0)}.$$

Now we reverse time in the realization X and call it X'. Thus, X' starts where X ends, and X' ends where X starts. The probability density of X' under Scenario

2 is then $L_2(X')$. We establish $L_2(X')$ analogous to the procedure above, with the birth events in X being death events in X' and vice versa. Thus, the same p and $(1-p)$ factors are multiplied when calculating the probability density of X' under Scenario 2, compared to the probability density of X under Scenario 1. Furthermore, the waiting time contributions are the same for Scenario 1 and Scenario 2, and thus $L_1(X) = L_2(X')$.

Note that $p_1(t \mid \lambda, \mu)$ is the integral over all realizations X under Scenario 1, $p_1(t \mid \lambda, \mu) = \sum_{k=0}^{\infty} \int_{\tau} L_1(X_{\tau,k}) d\tau$, where $X_{\tau,k}$ is a realization with k birth events according to an event time vector $\tau = (t_1, t_2, \ldots, t_{2k})$.

Analogously, $p_1(t \mid \mu, \lambda) = \sum_{k=0}^{\infty} \int_{\tau} L_2(X'_{\tau,k}) d\tau$. Since $L_1(X_{\tau,k}) = L_2(X'_{\tau,k})$, each component in this integration has the same probability density and thus we have $p_1(t \mid \lambda, \mu) = p_1(t \mid \mu, \lambda)$.

One can directly extend this argument to establish Proposition 2.1 for any value of $n \geq 1$ by considering the associated forward-in-time and backward-in-time processes. However, as we will derive this equation later from an algebraic identity, we do not describe this further here.

3 General symmetries under incomplete sampling

We continue to study a birth–death model with constant and non-negative birth and death rates λ and μ. However, we now allow each of the individuals present at time t to be sampled (independently) with probability $\rho \in (0, 1]$.

Let us first suppose that we start with one individual at time t_0, and let $p_i(t \mid \lambda, \mu, \rho)$ be the probability that i sampled descendants are observed (i.e. extant and sampled) at time $t_0 + t$. Exact expressions for $p_i(t) = p_i(t \mid \lambda, \mu, \rho)$ are provided by the following formulae (all proofs of theorems and corollaries are found

in the Appendix). Let

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

Theorem 3.1. *For $\lambda \neq \mu$, we have:*

$$p_n(t) = \begin{cases} 1 - \frac{\rho(\lambda-\mu)}{\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t}}, & \text{if } n = 0; \\ \frac{\rho(\lambda-\mu)^2 e^{-(\lambda-\mu)t}}{(\rho\lambda+(\lambda(1-\rho)-\mu)e^{-(\lambda-\mu)t})^2}, & \text{if } n = 1; \\ p_1(t)(\lambda q(t))^{n-1}, & \text{if } n > 1. \end{cases}$$

Note that $\lambda q(t) = p_2(t)/p_1(t)$, and for $\rho = 1$ and $\mu > 0$, we have $q(t) = \frac{1}{\mu}p_0(t)$.

Corollary 3.2. *For the critical case $\lambda = \mu$, let $q(t) = q(t \mid \lambda = \mu, \rho) = \frac{\rho t}{1+\rho\lambda t}$. We then have:*

$$p_n(t) = \begin{cases} 1 - \frac{\rho}{1+\rho\lambda t}, & \text{if } n = 0; \\ \frac{\rho}{(1+\rho\lambda t)^2}, & \text{if } n = 1; \\ p_1(t)(\lambda q(t))^{n-1}, & \text{if } n > 1. \end{cases}$$

We investigate the expressions for $p_i(t \mid \lambda, \mu, \rho)$ in detail, and identify symmetries with respect to λ and μ. We begin with the case where all extant taxa are sampled.

Theorem 3.3. *In the case of complete sampling (i.e. $\rho = 1$), we have:*

$$\begin{aligned} \lambda p_0(t \mid \lambda, \mu) &= \mu p_0(t \mid \mu, \lambda), \\ 1 - p_0(t \mid \lambda, \mu) &= e^{(\lambda-\mu)t}(1 - p_0(t \mid \mu, \lambda)), \\ \mu^{n-1} p_n(t \mid \lambda, \mu) &= \lambda^{n-1} p_n(t \mid \mu, \lambda), n \geq 1, \\ q(t \mid \lambda, \mu) &= q(t \mid \mu, \lambda). \end{aligned}$$

Next, consider the probability $p_{n,m}(t \mid \lambda, \mu)$ of having m individuals at time t,

given that n are present at time 0 (as described in the introduction).

Corollary 3.4. *For $m \geq n \geq 1$, we have: $\mu^{m-n} p_{n,m}(t \mid \lambda, \mu) = \lambda^{m-n} p_{n,m}(t \mid \mu, \lambda)$. In particular, $p_{n,n}(t \mid \lambda, \mu) = p_{n,n}(t \mid \mu, \lambda)$, for all $n \geq 1$ (as in Proposition 2.1). For $0 \leq m < n$, we have: $\lambda^{n-m} p_{n,m}(t \mid \lambda, \mu) = \mu^{n-m} p_{n,m}(t \mid \mu, \lambda)$.*

3.1 Negative 'death rates' in the case of incomplete sampling

We now investigate the case $\rho \leq 1$. We introduce two new variables λ' and μ', which will play a key role in the remainder of the paper. They are defined by λ, μ and ρ according to the following transformation:

$$\lambda' = \rho\lambda \text{ and } \mu' = \mu - \lambda(1 - \rho).$$

Note that when $\rho = 1$, we have $\lambda' = \lambda$. Further, for all vales of ρ we have $\lambda' - \mu' = \lambda - \mu$ (thus $\lambda' \neq \mu'$ if and only if $\lambda \neq \mu$). Note also that $\mu' < 0$ is entirely possible (for example, when $\lambda = 4\mu$ and $\rho = 0.5$, we obtain $\mu' = -\mu$). In this case, μ' can not easily be viewed as a death rate (nor as a birth rate); however, allowing μ' to take any real value (positive or negative) means that all parameter triplets (λ, μ, ρ) have a transformation to (λ', μ').

The following lemma is straightforward to verify using simple algebra.

Lemma 3.5. *For all $\lambda, \mu \geq 0$ and $0 < \rho < 1$, the four functions*

$$\lambda q(t \mid \lambda, \mu, \rho),\ \lambda(1 - p_0(t \mid \lambda, \mu, \rho)),\ \lambda p_1(t \mid \lambda, \mu, \rho),\ \text{and } \lambda p_n(t \mid \lambda, \mu, \rho)$$

can be written as functions of only two parameters (λ' and μ') when $\lambda \neq \mu$ (rather than the three parameters λ, μ, ρ). When $\lambda = \mu$, these four functions can be written as functions of the single parameter λ'.

In order to investigate symmetries, we define the following functions, which only depend on λ', μ', and t (rather than the four parameters λ, μ, ρ and t). Let:

$$\begin{aligned}
\tilde{p}_0(t \mid \lambda', \mu') &:= \frac{1}{\rho}(1 - p_0(t \mid \lambda, \mu, \rho)), \\
\tilde{q}(t \mid \lambda', \mu') &:= \frac{1}{\rho} q(t \mid \lambda, \mu, \rho), \\
\tilde{p}_1(t \mid \lambda', \mu') &:= \frac{1}{\rho} p_1(t \mid \lambda, \mu, \rho), \\
\tilde{p}_n(t \mid \lambda', \mu') &:= \tilde{p}_1(t \mid \lambda', \mu') \tilde{q}(t \mid \lambda', \mu')^{n-1}.
\end{aligned}$$

For $\lambda \neq \mu$, these equations are,

$$\begin{aligned}
\tilde{p}_0(t \mid \lambda', \mu') &= \frac{\lambda' - \mu'}{\lambda' - \mu' e^{-(\lambda' - \mu')t}}, \\
\tilde{p}_1(t \mid \lambda', \mu') &= \frac{(\lambda' - \mu')^2 e^{-(\lambda' - \mu')t}}{(\lambda' - \mu' e^{-(\lambda' - \mu')t})^2}, \\
\tilde{p}_n(t \mid \lambda', \mu') &= \frac{1}{\rho^n} p_1(t \mid \lambda, \mu, \rho)(q(t \mid \lambda, \mu, \rho))^{n-1}, \\
\tilde{q}(t \mid \lambda', \mu') &= \frac{1 - e^{-(\lambda' - \mu')t}}{\lambda' - \mu' e^{-(\lambda' - \mu')t}}.
\end{aligned}$$

In particular, we have: $\tilde{p}_1(t \mid \lambda', \mu') = p_1(t \mid \lambda, \mu, \rho = 1)$. This leads to the following symmetries with respect to λ' and μ'.

Theorem 3.6. *For $\mu' \geq 0$, the following symmetries hold:*

$$\begin{aligned}
\tilde{p}_0(t \mid \lambda', \mu') &= \tilde{p}_0(t \mid \mu', \lambda') e^{(\lambda' - \mu')t}, \\
\tilde{q}(t \mid \lambda', \mu') &= \tilde{q}(t \mid \mu', \lambda'),
\end{aligned}$$

and for all $n \geq 1$:

$$\tilde{p}_n(t \mid \lambda', \mu') = \tilde{p}_n(t \mid \mu', \lambda').$$

4 Tree probability densities

Let $\mathcal{T}$ be a phylogenetic tree generated by a birth–death process starting with one taxon and being stopped after time t_0. Each individual alive after time t_0 is sampled with probability ρ. In this tree, all extinct lineages are pruned, and only the lineages leading to the sampled tips are kept. Such a tree is also called the *reconstructed tree* [10], as indicated by the red lines in Fig. 2. Let this tree have n sampled tips and the branching times $t_1 > t_2, \ldots > t_{n-1}$, where time is measured from the present time 0. Let $L(t)$ be the number of co-existing lineages of tree $\mathcal{T}$ at time t (see Fig. 2).

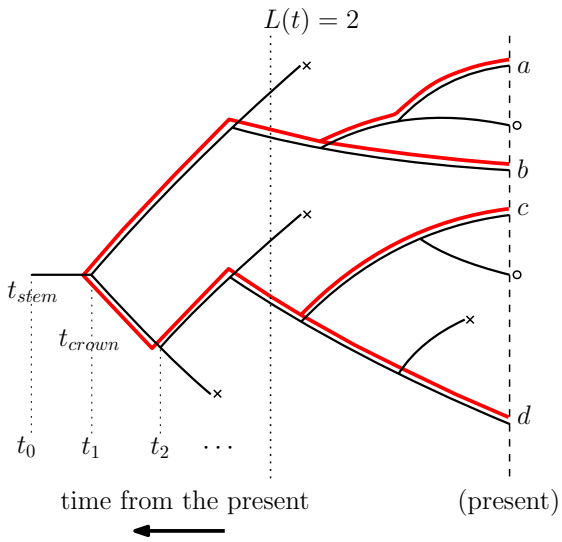

Figure 2: A phylogenetic tree $\mathcal{T}$ that evolves under under a birth–death process with rates λ, μ and with sampling at the present with probability ρ. Lineages ending in a death (extinction) are marked by $\times$ whereas lineages at the present that are not sampled are marked by o. The reconstructed tree on the sampled extant taxa is shown in red.

Let $f(\mathcal{T} \mid L(t_0) = 1)$ be the probability density of the tree $\mathcal{T}$, and let $f(\mathcal{T} \mid t_0 = t_{stem})$ be the probability density of the tree $\mathcal{T}$, given that at least one individual is sampled at present. Thus t_0 is the stem age (t_{stem}) of the process. For $\rho = 1$, this corresponds to conditioning on non-extinction of the process. Let $f(\mathcal{T} \mid t_0 = t_{stem}, L_s(0) = n)$ denote the probability density of the tree $\mathcal{T}$, given that we sample exactly n tips at present (denoted by $L_s(0) = n$).

The tree $\mathcal{T}$ in these formulations was a tree starting with one individual, leading to two lineages at time t_1 in the past. Alternatively, a tree $\mathcal{T}$ may start with two lineages at time t_1 ago; the probability of such a tree is $f(\mathcal{T}|L(t_1) = 2)$. Let $f(\mathcal{T}|t_1 = t_{crown})$ be the probability density of the tree $\mathcal{T}$ conditioning on sampling at least one descendant individual from both initial lineages. Note that when conditioning on sampling, the time t_1 is the crown age of the clade (t_{crown}). Furthermore, let $f(\mathcal{T} \,|\, t_1 = t_{crown}, L_s(0) = n)$ be the probability density of the tree $\mathcal{T}$ conditioned on sampling exactly n tips at present. Finally, in the setting where t_0 is chosen uniformly at random from $(0, \infty)$, then a tree $\mathcal{T}$ conditioned on n tips and integrated over all possible t_0 has probability density $f(\mathcal{T} \mid L_s(0) = n)$.

In what follows, we assume $\lambda > 0$ and thus $\lambda' > 0$; otherwise, we cannot obtain a tree with $n > 1$.

Theorem 4.1. *The tree probability densities can be expressed as functions of* $p_0(t \,|\, \lambda, \mu, \rho), p_1(t \,|\, \lambda, \mu, \rho)$ *and* $q(t \,|\, \lambda, \mu, \rho)$*, or* $\hat{p}_0(t \,|\, \lambda', \mu'), \hat{p}_1(t \,|\, \lambda', \mu')$ *and* $\hat{q}(t \,|\, \lambda', \mu')$*. Omitting the parameters* $\lambda, \mu, \rho, \lambda'$ *and* μ' *in these functions for easier reading, the*

expressions are given in the following table:

Tree probability densities	(λ, μ, ρ) − parameters	(λ', μ') − parameters
Unconditioned		
$f(\mathcal{T} \mid L(t_0) = 1)$	$p_1(t_0) \prod_{i=1}^{n-1} \lambda p_1(t_i)$	$\rho \tilde{p}_1(t_0) \prod_{i=1}^{n-1} \lambda' \tilde{p}_1(t_i)$
$f(\mathcal{T} \mid L(t_1) = 2)$	$p_1(t_1)^2 \prod_{i=2}^{n-1} \lambda p_1(t_i)$	$(\rho \tilde{p}_1(t_0))^2 \prod_{i=2}^{n-1} \lambda' \tilde{p}_1(t_i)$
Conditioned		
$f(\mathcal{T} \mid t_0 = t_{stem})$	$\frac{p_1(t_0)}{1-p_0(t_0)} \prod_{i=1}^{n-1} \lambda p_1(t_i)$	$\frac{\tilde{p}_1(t_0)}{\tilde{p}_0(t_0)} \prod_{i=1}^{n-1} \lambda' \tilde{p}_1(t_i)$
$f(\mathcal{T} \mid t_1 = t_{crown})$	$\left(\frac{p_1(t_1)}{1-p_0(t_1)}\right)^2 \prod_{i=2}^{n-1} \lambda p_1(t_i)$	$\left(\frac{\tilde{p}_1(t_0)}{\tilde{p}_0(t_0)}\right)^2 \prod_{i=2}^{n-1} \lambda' \tilde{p}_1(t_i)$
$f(\mathcal{T} \mid L_s(0) = n)$	$n \frac{p_1(t_1)}{1-p_0(t_1)} \prod_{i=1}^{n-1} \lambda p_1(t_i)$	$n \frac{\tilde{p}_1(t_0)}{\tilde{p}_0(t_0)} \prod_{i=1}^{n-1} \lambda' \tilde{p}_1(t_i)$
$f(\mathcal{T} \mid t_0 = t_{stem}, L_s(0) = n)$	$\prod_{i=1}^{n-1} \frac{p_1(t_i)}{q(t_0)}$	$\prod_{i=1}^{n-1} \frac{\tilde{p}_1(t_i)}{\tilde{q}(t_0)}$
$f(\mathcal{T} \mid t_1 = t_{crown}, L_s(0) = n)$	$\frac{1}{(n-1)} \prod_{i=2}^{n-1} \frac{p_1(t_i)}{q(t_0)}$	$\frac{1}{(n-1)} \prod_{i=2}^{n-1} \frac{\tilde{p}_1(t_i)}{\tilde{q}(t_0)}$

We note that the expressions in the middle column have been presented in [12] [Eq. 1-7], highlighting that $f(\mathcal{T} \mid L(t_1) = 2)$ goes back to [19] for $\rho = 1$, $f(\mathcal{T} \mid t_1 = t_{crown})$ was derived in [10], and $f(\mathcal{T} \mid t_1 = t_{crown}, L_s(0) = n)$ in [20] (both for $\rho \in (0, 1]$). Furthermore, $f(\mathcal{T} \mid t_0 = t_{stem}, L_s(0) = n)$ is stated for $\rho = 1$ in [4]. The idea of parameter transformation (right column) has been introduced for $f(\mathcal{T} \mid L_s(0) = n)$ in [13].

Remark 4.2. Only the expressions for the unconditioned tree probability densities (i.e. the equations not conditioning on observing at least one sample) depend on all

three parameters λ, μ and ρ. The remaining five expressions (the conditioned tree probability densities) only depend on two parameters (λ', μ'), meaning only two out of the three birth–death parameters λ, μ, ρ can be inferred from the phylogenetic tree. Furthermore, the expressions for $f(\mathcal{T} \mid t_0 = t_{stem}, L_s(0) = n)$ and $f(\mathcal{T} \mid t_1 = t_{crown}, L_s(0) = n)$ (i.e. the expressions where we condition on both the age of the process and the number of sampled tips) give the same result for λ', μ' and for when the parameters are swapped to μ', λ'. Note that $\mu' \leq 0$ is possible, whereas $\lambda' > 0$, thus the switching is only well-defined if $\mu' > 0$.

5 Mapping from (λ', μ') to the birth–death model parameters (λ, μ, ρ) with consequences for maximum likelihood and Bayesian inference

When using the tree probability densities in a maximum likelihood inference framework, the expressions are maximized over the parameters for a given tree. Based on the five conditioned tree probability density equations, we should optimize over λ' and μ', with $\lambda' \in (0, \infty)$ and $\mu' \in (-\infty, \infty)$, instead of maximizing over the three parameters λ, μ and ρ, as the latter parameterization induces a ridge in the likelihood surface and thus optimization is problematic. This is equivalent to optimizing when assuming complete sampling (and allowing the 'death rate' μ' to be negative) and, in a second step, assuming a sampling probability ρ and transforming from (λ', μ') to (λ, μ). We next investigate for which chosen values of ρ we can transform λ', μ' to λ, μ.

Theorem 5.1. *Let P denote the conditioned tree probability density for an arbitrary tree $\mathcal{T}$ given $\lambda' \in (0, \infty)$ and $\mu' \in (-\infty, \infty)$. The expression for P is given in the right column of Theorem 4.1. Each (λ', μ') has corresponding birth–death parameters $(\lambda \in (0, \infty), \mu \in [0, \infty), \rho \in (0, 1])$, namely:*

- *Given $\mu' \geq 0$, we obtain the same tree probability density P using the expression in the middle column of Theorem 4.1 with parameters $(\lambda = \lambda'/\rho, \mu = \mu' - \lambda' + \lambda'/\rho)$, where ρ is any value in $\rho \in (0, 1]$.*

- *Given $\mu' < 0$, we obtain the same tree probability density P using the expression in the middle column of Theorem 4.1 with parameters $(\lambda = \lambda'/\rho, \mu = \mu' - \lambda' + \lambda'/\rho)$, where ρ is any value in $\rho \in (0, \frac{1}{1-\mu'/\lambda'}]$.*

Given the correlations among λ, μ and ρ, one may decide to perform a Bayesian Markov chain Monte Carlo analysis on $\lambda' \in (0, \infty), \mu' \in (-\infty, \infty)$. Care has to be taken though regarding the priors, since these priors play out in non-straightforward ways. Assume, for example, that the analysis is performed by sampling λ', μ'. For each sampled parameter pair, one might pick a $\rho \in (0, 1]$ uniformly at random. Given that $\mu' \geq 0$, this would yield a uniform distribution on the chosen ρ. However, given that some sampled parameter pairs reveal $\mu' < 0$, it follows that only a small ρ, namely $\rho \in (0, \frac{1}{1-\mu'/\lambda'}]$ is possible, meaning that overall, the samples on ρ would be non-uniform, with a preference for small values of ρ. Thus, in the Bayesian setting, it is advantageous to estimate λ, μ, ρ in order to have control over their priors.

6 Mappings between birth–death model parameters (λ, μ, ρ) and $(\hat{\lambda}, \hat{\mu}, \hat{\rho})$

Next we characterize all birth–death parameters that are transformations of λ, μ, ρ.

Theorem 6.1. *Let (λ, μ, ρ) be birth–death parameters with the corresponding (λ', μ'). There exists parameters $\hat{\lambda} > 0, \hat{\mu} \geq 0$, and $\hat{\rho} \in (0, 1]$ with*

$$\lambda\rho = \hat{\lambda}\hat{\rho} = \lambda' \text{ and } \mu - \lambda(1-\rho) = \hat{\mu} - \hat{\lambda}(1-\hat{\rho}) = \mu'$$

if $\mu/\lambda \geq 1$ (for all $\hat{\rho} \in (0, 1]$) and if $\mu/\lambda < 1$ (for all $0 < \hat{\rho} \leq \rho/(1 - \frac{\mu}{\lambda})$).

Note that the parameters (λ, μ, ρ) and $(\hat{\lambda}, \hat{\mu}, \hat{\rho})$ thus give thus rise to the same tree probability density.

Corollary 6.2. *With $\frac{\mu}{\lambda} < 1$ (and thus $\hat{\rho} \leq \rho/(1-\frac{\mu}{\lambda})$) a transformation always exists for $\hat{\rho} < \rho$. However, a parameter transformation may not be possible for $\hat{\rho} > \rho$ (for example, if $\frac{\mu}{\lambda} = 0$, we cannot transform to $\rho' > \rho$).*

Next we consider $\hat{\rho} = 1$ (i.e. the transformation to the case of complete sampling).

Corollary 6.3. *Let (λ, μ, ρ) be birth–death parameters with the corresponding (λ', μ'). There exists a transformation to $(\hat{\lambda} > 0, \hat{\mu} \geq 0, \hat{\rho} = 1)$ if $\frac{\mu}{\lambda} \geq 1 - \rho$. If $0 \leq \frac{\mu}{\lambda} < 1 - \rho$, no transformation exists.*

6.1 Consequences for the birth–death tree distribution

Sometimes, proofs of the properties of the conditioned tree distribution are carried out for complete sampling (i.e. for parameters $\hat{\lambda}, \hat{\mu}, \hat{\rho} = 1$). Such properties also hold for incomplete sampling if $\frac{\mu}{\lambda} \geq 1$ or if $\frac{\mu}{\lambda} \geq 1 - \rho$. To include the parameter space, $0 \leq \frac{\mu}{\lambda} < 1 - \rho$ the proof needs to be done with explicitly acknowledging incomplete sampling. This was noticed already in [18].

6.2 Consequences for model choice regarding complete sampling

For a given phylogenetic tree, it is tempting to ask if a model with $\rho = 1$ or $\rho = \hat{\rho} < 1$ fits the data better. However, for every parameter combination $(\lambda, \mu, 1)$, we also find a parameter combination $(\hat{\lambda}, \hat{\mu}, \hat{\rho})$ with both parameter triples having the same conditioned tree probability density. Moreover, there are parameter combinations $(\hat{\lambda}, \hat{\mu}, \hat{\rho})$ without a corresponding triplet where $\rho = 1$ (see Corollary 6.2). Thus, the model with $\rho < 1$ always gets more support than the model with $\rho = 1$. In summary, such a test is meaningless because of the parameter correlations.

7 Discussion

Birth–death models have been studied for almost 100 years [21, 5]. However, surprising properties are still being uncovered. Here, we presented some unexpected symmetries in birth–death models, namely, fundamental birth–death probability distributions are invariant towards swapping the birth and the death rate. We explained this surprising observation in a special case, by using an argument that is both intuitive and precise, then derived more general symmetries algebraically. Second, we showed that a birth–death process with incomplete taxon sampling can be described phylogenetically through two parameters instead of three parameters due to parameter correlations, and that the two-parameter description again reveals symmetries.

Such correlations have important consequences for using birth–death models in phylogenetic and phylodynamic inference. In particular, the likelihood surface of the three birth–death parameters λ, μ and ρ for a given tree has a ridge casued by the correlations, and we can therefore only estimate two of the three parameters. Maximum likelihood estimation should thus be done over the two parameters. On the other hand, in Bayesian analysis, using the two-parameter description of the process would not allow us to use all prior information on the three original parameters and therefore using the original parameterization is advantageous.

Furthermore, we showed that for some of the parameter triplets (λ, μ, ρ), their two-parameter description is, in fact, equivalent to a birth–death process with complete sampling. However, in some cases, the resulting 'death' rate is negative, and thus the transformed parameters cannot always be considered as a birth–death process with complete-sampling. This means that we cannot simply prove properties of phylogenetic trees for complete sampling and then extrapolate to incomplete sampling, as we then miss some birth–death parameter combinations (namely the ones leading to a negative 'death' rate). Furthermore, testing whether the data are com-

pletely sampled ($\rho = 1$) or not ($\rho < 1$) is not informative, as the models with $\rho < 1$ always have more support: parameter triplets for incomplete sampling may only have corresponding complete sampling parameters with a negative 'death' rate, whereas birth and death rates under complete sampling have a corresponding triplet for all $\rho \in (0, 1]$.

The birth–death model presented here is the simplest model for speciation and extinction, or for transmission and recovery. However, it has limitations for explaining the data, as it assumes exponential growth of the population, although populations cannot have unlimited growth, and it assumes that all individuals are dynamically equivalent. There has been considerable work on extending the birth–death model to address such limitations [8, 9, 15, 3, 16], but no symmetries and only very special parameter correlations have been observed [17]. It will be interesting to explore in the future whether the observed symmetries and correlations in our simple model are also present in these more complex models.

8 Acknowledgements

JF thanks Nicolas Salamin for sharing with him his numerical observation that swapping λ and μ in the the birth-death likelihood conditioned on the number of tips leaves the likelihood invariant. This observation inspired JF to investigate analytically whether there was a corresponding symmetry in $p_{11}(t \,|\, \lambda, \mu)$. TS is supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 335529).

9 Appendix: Proofs

Proof of Theorem 3.1. For $\lambda > 0$ and $\rho > 0$, the expressions are provided in [14], based on earlier work by [10, 20]. In fact, [14] requires $\lambda > \mu$, but the proof is identical for $\lambda < \mu$.

These expresions also hold for $\lambda = 0$ (and thus $\mu > 0$). To see this, observe first that the probability $p_1(t \mid \lambda = 0, \mu, \rho)$ is the product of the probability of no death $e^{-\mu t}$ with the sampling probability ρ. Indeed this equation simplifies to:

$$p_1(t \mid \lambda = 0, \mu, \rho) = \frac{\rho\mu^2}{\mu^2 e^{\mu t}} = \rho e^{-\mu t}.$$

Second, notice that the probability $p_n(t \mid \lambda = 0, \mu, \rho)$ for $n > 1$ is 0, as no birth event may occur. Indeed, by using the expressions above, we get $q(t \mid \lambda = 0, \mu, \rho) = \frac{\rho}{\mu}(1 - e^{-\mu t})$ and thus $p_n(t \mid \lambda = 0, \mu, \rho) = 0$ for $n > 1$.

Finally, the probability $p_0(t \mid \lambda = 0, \mu, \rho)$ is $1 - p_1(t \mid \lambda = 0, \mu, \rho) = 1 - \rho e^{-\mu t}$. Again, the above equation simplifies:

$$p_0(t \mid \lambda = 0, \mu, \rho) = 1 - \frac{-\rho\mu}{-\mu e^{\mu t}} = 1 - \rho e^{-\mu t}.$$

□

Proof of Corollary 3.2. Note that these equations can be derived from the supercritical case $\lambda > \mu$ by setting $\lambda - \mu = \epsilon$ and using the property $e^{-\epsilon} \sim 1 - \epsilon$ as $\epsilon \to 0$. In particular, for the expression in the denominator, we obtain:

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

from which we directly get the expressions above. □

Now we first prove Theorem 3.6 and then provide the proofs for the special case of complete sampling.

Proof of Theorem 3.6. The first three equations can be directly observed. For the last equation, observe that $\tilde{p}_0(t \mid \lambda', \mu') = \frac{\lambda'-\mu'}{\lambda'-\mu' e^{-(\lambda'-\mu')t}}$. When swapping λ' and μ', we obtain:

$$\tilde{p}_0(t \mid \mu', \lambda') = \frac{-\lambda' + \mu'}{\mu' - \lambda' e^{(\lambda'-\mu')t}} = \frac{(\lambda' - \mu')e^{-(\lambda'-\mu')t}}{\lambda' - \mu' e^{-(\lambda'-\mu')t}} = \tilde{p}_0(t \mid \lambda', \mu')e^{-(\lambda'-\mu')t}.$$

□

Proof of Theorem 3.3. For $p_0(t \mid \lambda, \mu) = \frac{\mu(1-e^{-(\lambda-\mu)t})}{\lambda-\mu e^{-(\lambda-\mu)t}}$, we obtain,

$$\begin{aligned} \lambda p_0(t \mid \lambda, \mu) &= \mu \frac{\lambda(1 - e^{-(\lambda-\mu)t})}{\lambda - \mu e^{-(\lambda-\mu)t}} \\ &= \mu \frac{\lambda(1 - e^{(\lambda-\mu)t})}{\mu - \lambda e^{(\lambda-\mu)t}} \\ &= \mu p_0(t \mid \mu, \lambda). \end{aligned}$$

The remaining four equations are directly observed as special cases of Theorem 3.6. Alternatively, they can be established through simple algebraic rearrangements. □

Proof of Theorem 3.4. First, we assume that both λ and μ are different from 0. For $m = 0$, we have $\lambda^n p_{n,0}(t \mid \lambda, \mu) = (\lambda p_0(t \mid \lambda, \mu))^n = (\mu p_0(t \mid \mu, \lambda))^n = \mu^n p_{n,0}(t \mid \mu, \lambda)$ where the second equality follows from Thm. 3.3. For $m > 0$, we use a generating function argument. Let

$$P(x) = \sum_{i \geq 0} p_t(t \mid \lambda, \mu)x^i \text{ and } \tilde{P}(x) = \sum_{i \geq 0} p_t(t \mid \mu, \lambda)x^i.$$

Theorem 3.3 gives:

$$P(x) = \frac{\mu}{\lambda} \cdot \tilde{P}\left(\frac{\lambda}{\mu}x\right).$$

Now $p_{n,m}(t \mid \lambda, \mu)$ is the coefficient of x^m in $P(x)^n$, which, by the previous equation, equals $(\frac{\mu}{\lambda})^n$ multiplied by the coefficient of x^m in $\tilde{P}\left(\frac{\lambda}{\mu}x\right)^n$. The latter coefficient is just $p_{n,m}(t \mid \lambda, \mu)$ times $\left(\frac{\lambda}{\mu}\right)^m$. Thus $p_{n,m}(t \mid \lambda, \mu) = \left(\frac{\mu}{\lambda}\right)^n \cdot p_{n,m}(t \mid \lambda, \mu) \cdot \left(\frac{\lambda}{\mu}\right)^m$, which

leads to the claimed identities in the cases where $m > 0$ and $\lambda, \mu \neq 0$.

Finally, we prove the case where λ=0; the case where $\mu = 0$ is then analogous. For the case $m = n$, $p_{n,n}(t \mid 0, \mu)$ is the probability of no event happening within t, and thus $p_{n,n}(t \mid 0, \mu) = e^{-\mu t} = p_{n,n}(t \mid \mu, 0)$. If $m > n$, $p_{n,m}(t \mid \lambda, \mu) = 0$ and thus the equation in the corollary is true. If $m < n$, then $p_{n,m}(t \mid \mu, \lambda) = 0$ and again the equation in the corollary is true. □

Proof of Theorem 5.1. We can always transform $\lambda' \in (0, \infty)$ and $\rho \in (0, 1]$ to $\lambda \in (0, \infty)$ via $\lambda = \lambda'/\rho$. Second, since $\mu' = \mu - \lambda(1 - \rho)$ and $\mu \geq 0$, we have $\mu = \mu' - \lambda' + \lambda = \mu' - \lambda' + \lambda'/\rho \geq 0$, and thus $\lambda' - \mu' \leq \lambda'/\rho$. Thus, we can only transform μ' to μ if this last inequality is fulfilled. This constrains our choices for $\rho \in (0, 1]$:

- For $\lambda' - \mu' > 0$, the constraint is $0 < \rho \leq \frac{1}{1-\mu'/\lambda'}$.
  - If $\mu' \geq 0$, then $\frac{1}{1-\mu'/\lambda'} \geq 1$, and thus there exists a transformation from λ', μ' to λ, μ, ρ for all $\rho \in (0, 1]$.
  - If $\mu' < 0$, we have $0 < \frac{1}{1-\mu'/\lambda'} < 1$ and thus we require $0 < \rho \leq \frac{1}{1-\mu'/\lambda'}$ for a transformation to λ, μ.
- For $\lambda' - \mu' < 0$ (implying $\mu'/\lambda' > 1$ and $\mu' > 0$), our constraint is, $1 \geq \rho \geq \frac{1}{1-\mu'/\lambda'}$. Since $\frac{1}{1-\mu'/\lambda'} < 0$, this means that a transformation for all $\rho \in (0, 1]$ exists.
- For $\lambda' - \mu' = 0$ (implying $\mu' > 0$), we require $0 \leq \lambda'/\rho$ which is fulfilled for all $\rho \in (0, 1]$.

□

Proof of Theorem 6.1. We can always transform $\lambda \in (0, \infty)$ and $\rho \in (0, 1]$ to $\hat{\lambda} \in (0, \infty)$ and $\hat{\rho} \in (0, 1]$ via $\hat{\lambda} = \lambda\rho/\hat{\rho}$. Second, since $\hat{\mu} = \mu - \lambda(1 - \rho/\hat{\rho})$ with $\hat{\mu} \geq 0$, we need to determine for which $\hat{\rho}$ we have $\hat{\mu} = \mu - \lambda(1 - \rho/\hat{\rho}) \geq 0$.

- For $\lambda = \mu$, we have $\hat{\mu} = \lambda'/\hat{\rho} > 0$ for all $\hat{\rho} \in (0, 1]$.
- For $\lambda \neq \mu$, we obtain,

$$\mu - \lambda(1 - \rho/\hat{\rho}) > 0 \Rightarrow \frac{\mu}{\lambda} \geq 1 - \frac{\rho}{\hat{\rho}} \Rightarrow \frac{\rho}{\hat{\rho}} \geq 1 - \frac{\mu}{\lambda}.$$

  - For $\frac{\mu}{\lambda} > 1$, we have $0 > 1 - \frac{\mu}{\lambda}$ and thus we have $\hat{\mu} = \mu - \lambda(1 - \rho/\hat{\rho}) \geq 0$ for all $\hat{\rho} \in (0, 1]$.
  - For $\frac{\mu}{\lambda} < 1$, $\hat{\rho}$ needs to fulfil $\frac{\rho}{1-\frac{\mu}{\lambda}} \geq \rho'$ such that $\hat{\mu} \geq 0$.

□

References

[1] David Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.*, 16(1):23–34, 2001.

[2] David Aldous, Maxim Krikun, and Lea Popovic. Five statistical questions about the tree of life. *Syst. Biol.*, 60(3):318–328, 2009.

[3] Rampal S Etienne, Bart Haegeman, Tanja Stadler, Tracy Aze, Paul N Pearson, Andy Purvis, and Albert B Phillimore. Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proc. Roy. Soc. Lond. B.*, 279(1732):1300–1309, 2012.

[4] J. Felsenstein. Inferring phylogenies. *Sinauer Associates, Sunderland, Massachusetts*, 8:8–5, 2004.

[5] David G. Kendall. On some modes of population growth leading to r. a. fisher's logarithmic series distribution. *Biometrika*, 35(1/2):6–15, 1948.

[6] David G. Kendall. On the generalized "birth-and-death" process. *Ann. Math. Statist.*, 19(1):1–15, 1948.

[7] A. Lambert and T. Stadler. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theor. Pop. Biol.*, 90:113–128, 2013.

[8] W.P. Maddison. Estimating a binary character's effect on speciation and extinction. *Syst. Biol.*, 56(5):701–710, 2007.

[9] Hélène Morlon, Todd L Parsons, and Joshua B Plotkin. Reconciling molecular phylogenies with the fossil record. *Proc. Natl. Acad. Sci. USA*, 1081(39):6327–6332, 2011.

[10] S. C. Nee, R. M. May, and P.H. Harvey. The reconstructed evolutionary process. *Phil. Trans. Roy. Soc. Ser B.*, 344:305–311, 1994.

[11] B. Rannala and Z. Yang. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.*, 43:304–311, 1996.

[12] T Stadler. How can we improve accuracy of macroevolutionary rate estimates? *Systematic biology*, 62(2):321, 2013.

[13] Tanja Stadler. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. *J. Theor. Biol.*, 261(1):58–66, 2009.

[14] Tanja Stadler. Sampling-through-time in birth–death trees. *J. Theor. Biol.*, 267(3):396–404, 2010.

[15] Tanja Stadler. Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl. Acad. Sci. USA*, 108(15):6187–6192, 2011.

[16] Tanja Stadler and Sebastian Bonhoeffer. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. Roy. Soc. Ser. B.*, 368(1614):20120198, 2013.

[17] Tanja Stadler, Denise Kühnert, Sebastian Bonhoeffer, and Alexei J Drummond. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV)). *Proc. Natl. Acad. Sci. USA*, 110(1):228–233, 2013.

[18] Tanja Stadler and Mike Steel. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. *J. Theor. Biol.*, 297:33–40, 2012.

[19] E. A. Thompson. *Human evolutionary trees.* Cambridge University Press, 1975.

[20] Z. Yang and B. Rannala. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Mol. Biol. Evol.*, 17(7):717–724, 1997.

[21] G. U. Yule. A mathematical theory of evolution: based on the conclusions of Dr. J.C. Willis. *Phil. Trans. Roy. Soc. Ser. B.*, 213:21–87, 1924.