Journal of Theoretical Biology 380 (2015) 321–331

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events

Sebastian Höhna [a,b,c,d,*]

[a] Department of Integrative Biology, University of California, Berkeley, CA 94720, USA
[b] Department of Statistics, University of California, Berkeley, CA 94720, USA
[c] Department of Evolution and Ecology, University of California, Davis, CA 95616, USA
[d] Department of Mathematics, Stockholm University, Stockholm, SE-106 91 Stockholm, Sweden

HIGHLIGHTS

- A general framework for the probability density of a reconstructed phylogeny.
- Derivation of the density under piecewise-linear rates and mass-extinction events.
- A unified notation to simplify the comparison between different birth–death models.
- A compendium of commonly used birth–death models in diversification rate analyses.

ARTICLE INFO

Article history:
Received 19 February 2014
Received in revised form
4 March 2015
Accepted 2 June 2015
Available online 12 June 2015

Keywords:
Birth–death process
Diversification
Incomplete taxon sampling
Probability density function
Likelihood

ABSTRACT

The homogeneous reconstructed evolutionary process is a birth–death process without observed extinct lineages. Each species evolves independently with the same diversification rate—speciation rate, $\lambda(t)$, and extinction rate, $\mu(t)$—that may change over time. The process is commonly applied to model species diversification where the data are reconstructed phylogenies, e.g. trees estimated from present-day molecular data, and used to infer diversification rates.

In the present paper I develop the general probability density of a reconstructed tree under any homogeneous, time-dependent birth–death process. I demonstrate how to adapt this probability density when conditioning on the survival of one or two initial lineages, or on the process realizing n species, and also how to transform between the probability density of a reconstructed tree and the probability density of the speciation times.

I demonstrate the use of the general time-dependent probability density functions by deriving the probability density of a reconstructed tree under a birth–death-shift model with explicit mass-extinction events. I extend these functions to several special cases, including the pure-birth process, the pure-death process, the birth–death process, and the critical-branching process. Thus, I specify equations for the most commonly used birth–death models in a unified framework (e.g. same condition and same data) using a common notation.

1. Introduction

The birth–death process is commonly used to model species diversification and to infer diversification rates (speciation and extinction rates) from reconstructed phylogenies (Nee, 2006). Likelihood-based estimates, whether based on a maximum likelihood or Bayesian framework, require the probability density function of the reconstructed tree under a birth–death process and dominate parameter estimation methods in phylogenetics (Huelsenbeck et al., 2001; Holder and Lewis, 2003). Besides their use in parameter estimation, the probability density functions are crucial for hypothesis testing, e.g. in testing whether rates have been constant or variable over time (Huelsenbeck and Rannala, 1997; Rabosky, 2006).

The probability density function of a reconstructed tree under the reconstructed evolutionary process has been derived under various scenarios by different authors (Yule, 1925; Kendall, 1948; Raup et al., 1973; Thompson, 1975; Bailey, 1990; Hey, 1992; Nee et al., 1994; Rabosky, 2006; Lambert, 2010; Morlon et al., 2011;

---

[*] Correspondence address: Department of Integrative Biology, University of California, Berkeley, CA 94720, USA.
E-mail address: Sebastian.Hoehna@gmail.com

http://dx.doi.org/10.1016/j.jtbi.2015.06.005

Stadler, 2011; Etienne et al., 2012) and has been applied in several studies (for reviews see Ricklefs, 2004; Nee, 2006; Ricklefs, 2007; Pyron and Burbrink, 2013). However, it remains challenging to compare these probability density functions because they differ in their notation, derivation and conditioning, e.g. conditioning on survival of the process or conditioning on obtaining exactly n species (Stadler, 2013). Furthermore, the probability density functions are inconsistently applied to reconstructed trees or speciation times: each requires different combinatorial factors. This inconsistency prevents the use of model-selection methods and so precludes the comparison of candidate models.

The present paper provides a study of the time-dependent homogeneous reconstructed evolutionary process under various time-dependent diversification rate functions and serves as a compendium of probability distribution functions presented in a common notation. I start by deriving the probability density of a reconstructed tree under the time-dependent birth–death process in the general case, e.g. with any diversification rate functions. I then demonstrate how to condition on the survival of one initial lineage (where the process starts at the stem node of the tree), two initial lineages (where the process starts at the crown node of the tree), or on the process realizing n species today. Additionally, I show how any of the derived probability densities can be transformed to apply to reconstructed trees or the speciation times of a reconstructed tree only.

The use of the probability densities presented here lies in their applicability to any diversification rate functions. I demonstrate this flexibility by deriving the explicit probability density of a reconstructed tree under a birth–death-shift model, e.g. piecewise constant diversification rates, with explicit mass-extinction events. I complete this discussion on the time-dependent reconstructed evolutionary process with mass-extinction events by providing the probability densities of a reconstructed tree under a pure-birth process with constant-rate and exponentially decaying rate and a constant-rate birth–death process.

2. The reconstructed evolutionary process

I define the birth–death process with non-constant rates for rooted, strictly bifurcating trees following the notation of Nee et al. (1994). Let $N(t)$ denote the number of species alive at time t. Furthermore, let the process start with a single species at time t_0, such that $N(t_0)=1$. A speciation event increases the number of species by one—i.e. from k to $k+1$ assuming that k species are alive at time t—after an exponentially distributed time with rate $k\lambda(t)$. Similarly, an extinction event decreases the number of species by one after an exponentially distributed time with rate $k\mu(t)$. At a speciation event, one of the k species gives birth to a new species, where the probability of each species giving birth is equally probable. At an extinction event, one species simply dies, where each species has the same probability of going extinct. Commonly, the process is stopped at the present time, denoted T, and the number of extant species is denoted by $N(T)=n$. Fig. 1a depicts a binary tree resulting from a birth–death process, showing both extant and extinct lineages. This is known as a *complete tree*. Fig. 1b shows the same tree but after removing all extinct lineages; this is a *reconstructed tree*. Reconstructed trees are the data (observations) that I consider here.

2.1. Probability density of a reconstructed evolutionary tree

Let Ψ denote a reconstructed evolutionary tree comprising a tree topology τ and the set of branching times $\mathbb{T}$. Let us derive the probability density of all speciation times $\mathbb{T}$ in the reconstructed tree following Thompson (1975, Eq. (3.4.6)) (see also Eq. (20) in Nee et al., 1994). I will use the notation that $P(N(T)=n \mid N(t)=k)$ is the probability of observing n species at time T given k living species at time t. First, let me restate the probability of obtaining exactly 1 species at time T with $T>t$. At any given time s, $t<s<T$, the number of species could be between 1 and infinity, however, all but one species must go extinct until time T (Thompson, 1975):

$$\begin{aligned} P(N(T)=1 \mid N(t)=1) = \sum_{k=1}^{\infty} & k \times P(N(s)=k \mid N(t)=1) \\ & \times P(N(T)=1 \mid N(s)=1) \\ & \times (P(N(T)=0 \mid N(s)=1))^{k-1} \end{aligned} \tag{1}$$

The factor k arises because any of the k currently alive species could survive until the present and we need to consider each possibility.

Next, let us consider each branch of the reconstructed tree (see Fig. 1). There are two types of branches: (1) internal branches that end at a speciation event denoted by $\mathcal{I}$ and (2) terminal branches that end at the sampling time T denoted by $\mathcal{T}$. The probability density of the ith terminal branch, ν_i, that originated at time t_i is simply the probability of starting with one species and observing exactly one species:

$$P(\nu_i \mid \nu_i \in \mathcal{T}) = P(N(T)=1 \mid N(t_i)=1). \tag{2}$$

The probability of the ith internal branch, ν_i, starting at time t_i and ending at a speciation event at time t_j is given by the probability that there were k species at time t_j, one of these speciated with probability $k\lambda(t_j)$, this species survived (was sampled) but all other $k-1$ species went extinct before the present time (were not sampled):

$$\begin{aligned} P(\nu_i \mid \nu_i \in \mathcal{I}) = \sum_{k=1}^{\infty} \Big(& k \times \lambda t_j \times P(N(t_j)=k \mid N(t_i)=1) \\ & \times (P(N(T)=0 \mid N(t_j)=1))^{k-1} \Big). \end{aligned} \tag{3}$$

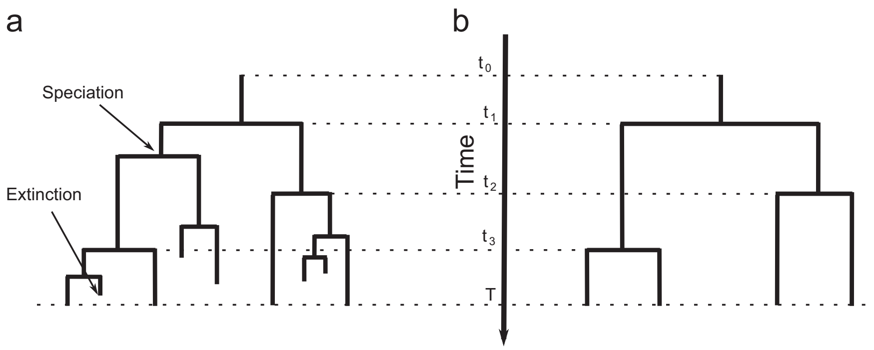

Fig. 1. A simulated birth–death tree starting with a single species at time t_0, thus $N(t_0)=1$. The process was stopped at time T. (a) The complete tree containing both extant and extinct species. (b) The reconstructed tree containing only extant species. Speciation events in the reconstructed tree occurred at times t_1, t_2 and t_3.