Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models

Tanja Stadler [a,*], Mike Steel [b]

[a] Institut für Integrative Biologie, ETH Zürich, Universitätsstr. 16, 8092 Zürich, Switzerland
[b] Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

ABSTRACT

The constant rate birth–death process is a popular null model for speciation and extinction. If one removes extinct and non-sampled lineages, this process induces 'reconstructed trees' which describe the relationship between extant lineages. We derive the probability density of the length of a randomly chosen pendant edge in a reconstructed tree. For the special case of a pure-birth process with complete sampling, we also provide the probability density of the length of an interior edge, of the length of an edge descending from the root, and of the diversity (which is the sum of all edge lengths). We show that the results depend on whether the reconstructed trees are conditioned on the number of leaves, the age, or both.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The constant rate birth–death process is a widely-used null model for speciation and extinction (Mooers and Heard, 1997; Nee, 2001). This model has been used to test the hypothesis of constant macroevolutionary rates and to quantify the rates of speciation (Paradis, 1998; Pybus and Harvey, 2000; Ricklefs et al., 2007; Stadler, 2011). Despite its wide use, the process continues to reveal new and sometimes unexpected results: even the simple Yule model leads to a curious property highlighted in a recent paper (Steel and Mooers, 2010; Mooers et al., in press): the expected length of a randomly chosen edge in a Yule tree is half of the expected waiting time until a speciation event occurs. In order to attribute such "surprises" in empirical data to the null model instead of trying to find further explanations, the null models need to be well-understood. In this paper, we characterize the lengths of pendant edges in birth–death trees on extant species (so-called reconstructed trees) and improve our understanding of interior edge lengths in Yule trees.

We will first explain the concept of a reconstructed tree which was originally introduced in Nee et al. (1994) (see also Fig. 1). The birth–death process is a binary branching process. The process starts with a single species at time x_0 before the present. At all times until the present, each species has a constant rate λ of

speciation and a constant rate μ of extinction (with $0 \leq \mu \leq \lambda$). Such a process induces a birth–death tree. At the present, each extant species is sampled with probability f. Throughout this paper, we prune extinct and non-sampled species in the birth–death tree, i.e. we consider the birth–death tree which is induced by the sampled extant species. The tree without the extinct and non-sampled species is called the reconstructed tree, as empirical data typically infers this reconstructed tree (unless fossil information is included).

We consider three different scenarios for stopping the process (i.e. defining the present):

- Scenario (i): we condition the process on having n extant sampled species, or
- Scenario (ii): we condition the process on having n extant sampled species and age x_1 for the most recent common ancestor of the extant species, or
- Scenario (iii): we condition the process on having age x_1 for the most recent common ancestor of the extant species.

As the start of the process (the time of origin x_0 or the time of most recent common ancestor x_1) is a parameter of the birth–death model, we have to assume a prior distribution for the time of origin when not conditioning the reconstructed trees on its age (Scenario (i)). We make the common assumption that the first species originated at any time x_0 in the past with uniform probability (Aldous and Popovic, 2005). This is also called an improper prior on $(0,\infty)$. Conditioning the resulting reconstructed

* Corresponding author. Tel.: +41 44 632 45 48; fax +41 44 632 12 71.
  E-mail addresses: tanja.stadler@env.ethz.ch (T. Stadler),
mike.steel@canterbury.ac.nz (M. Steel).

tree to have n extant species yields a proper distribution for the time of origin (Gernhard, 2008).

A special case of the birth–death process is the Yule model (Yule, 1924), which is obtained by setting $\mu = 0$ and $f = 1$. Under the Yule model, Scenario (i) is equivalent to stopping the process just before the $(n+1)$th speciation event (Hartmann et al., 2010).

In the following, we derive the probability density of the length of a randomly chosen pendent edge in a reconstructed tree generated by a constant rate birth–death process under Scenarios (i), (ii), and (iii), for all parameters satisfying $1 \geq \mu/\lambda \geq 1-f$. For the special case of a Yule model under Scenario (i), we find that a randomly selected pendant edge (or interior edge) has an exponentially distributed length with parameter 2λ (Corollary 3.2 and Theorem 3.3). This result generalizes (Steel and Mooers, 2010), where the expected length of a pendant edge was calculated to be $1/(2\lambda)$. For the Yule model (under Scenarios (i)–(iii)), we also derive the probability density of the length of an edge descending from the root and the sum of all edges.

2. Preliminaries

We first present some notations and a preliminary result that will be useful later. A reconstructed tree on n extant species has $n-1$ interior vertices at times $x_1 > \cdots > x_{n-1}$ in the past (see Fig. 2). We call the speciation event at time x_k the kth speciation event. We say that a leaf x of a reconstructed tree is *adjacent to* the kth speciation event if (v,x) is an arc of the tree, where v is the vertex that corresponds to the kth speciation event. For example, in Fig. 2 the left-most leaf of the tree is adjacent to the second speciation event. The following result was established in Stadler
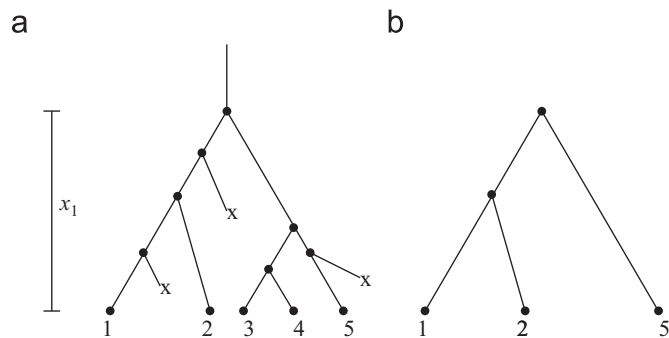
(2008) (and also, an equivalent result, in the context of coalescent theory, was stated in Blum and François, 2005); we will provide a shorter and more direct proof than either of these existing proofs here.

Theorem 2.1. *For a reconstructed Yule or birth–death tree on $n \geq 2$ extant species, the probability under Scenario (i) or (ii) that a randomly-selected leaf is adjacent to the kth speciation event is*

$$v(k) = \frac{2k}{n(n-1)}. \tag{1}$$

Proof. Our proof is based on the equivalence of the probability distributions on tree topologies under three models—the Yule process, the reconstructed birth–death tree and the coalescent tree (if one ignores branch lengths) (Aldous, 2001). Observe that leaf L is adjacent to the kth speciation event vertex in the Yule tree if and only if L fails to coalesce for the first $n-k-1$ coalescence events, but does so on the next coalescence event (when $k+1$ points are available to coalesce and k of these possible $\binom{k+1}{2}$ pairs involve L). The probability of this, under the coalescent process, is precisely

$$\prod_{j=1}^{n-k-1} \left[1 - \frac{n-j}{\binom{n-j+1}{2}} \right] \frac{k}{\binom{k+1}{2}}.$$

Now, expansion and simplication (cancelation) of this product yields $v(k) = 2k/(n(n-1))$. □

The following expressions will be useful later. For $0 \leq \mu \leq \lambda$, we define

$$p_0(s) := \begin{cases} \dfrac{(1-e^{-(\lambda-\mu)s})}{\lambda - \mu e^{-(\lambda-\mu)s}} & \text{if } \mu < \lambda; \\ \dfrac{s}{1+\lambda s} & \text{if } \mu = \lambda \end{cases} \tag{2}$$

and

$$p_1(s) := \begin{cases} \dfrac{(\lambda-\mu)^2 e^{-(\lambda-\mu)s}}{(\lambda - \mu e^{-(\lambda-\mu)s})^2} & \text{if } \mu < \lambda; \\ \dfrac{1}{(1+\lambda s)^2} & \text{if } \mu = \lambda. \end{cases} \tag{3}$$

The significance of these quantities is that, under a birth–death process, the probability of a lineage producing 0 (resp. 1) offspring (both sampled and non-sampled) after time s is $\mu p_0(s)$ (resp. $p_1(s)$) (Kendall, 1949; Stadler, 2010). Note that, for Yule trees, we have

$$p_0(s) = (1-e^{-\lambda s})/\lambda; \quad \text{and} \quad p_1(s) = e^{-\lambda s}.$$

Fig. 1. (a) A birth–death tree starting with a single lineage, and with extinct lineages ending in 'x'. The time from the extant lineages to their most recent common ancestor is x_1. (b) The 'reconstructed tree' obtained by deleting the initial single lineage and all the extinct and two non-sampled extant lineages (namely the two labeled 3 and 4).
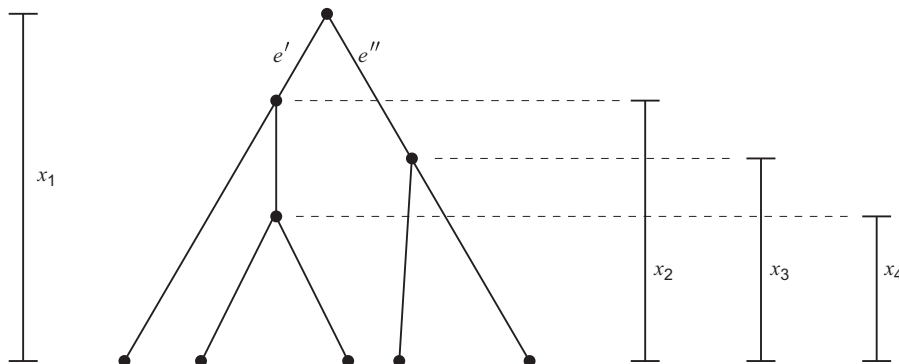

Fig. 2. For $k=1, \ldots, n-1$, x_k denotes the time from the present into the past when the kth speciation event occurred in a reconstructed birth–death tree. The two edges incident with the root e' and e'' (considered in Section 5) are also shown, with e' the shorter of the two.

3. Length of a pendant edge in completely sampled birth–death and Yule trees

In this section, we calculate the probability density function (pdf) of the length of a random pendant edge in a reconstructed tree assuming $f = 1$. For Scenarios (i) and (ii), we use this to derive the expected length of such an edge; moreover for Scenario (i) we calculate the pdf of the length of a random *interior* edge for a Yule model.

3.1. Conditioning on n

For pendant edges, the following result was established in Mooers et al. (in press).

Theorem 3.1. *The length of a randomly selected pendant edge in a reconstructed birth–death tree on n extant species has probability density function*

$$f_P(s|n) = 2\lambda p_1(s)(1 - \lambda p_0(s))$$

and expectation, for $0 < \mu < \lambda$:

$$\mathbb{E}[P|n] = \frac{\mu + (\lambda - \mu)\log(1 - \mu/\lambda)}{\mu^2},$$

for $\mu = \lambda$:

$$\mathbb{E}[P|n] = \frac{1}{\lambda},$$

for $\mu = 0$:

$$\mathbb{E}[P|n] = \frac{1}{2\lambda}.$$

Here we note the following direct consequence of this result for the Yule model (obtained by setting $\mu = 0$ in $f_P(s|n)$ of Theorem 3.1):

Corollary 3.2. *The length of a randomly picked pendant edge in a tree on n leaves under the Yule model with rate λ is exponentially distributed with rate 2λ.*

We now establish a corresponding result for a randomly selected interior edge in a Yule tree (again under Scenario (i)).

Theorem 3.3. *The length of a randomly selected interior edge in a Yule (rate λ) tree on n leaves is exponentially distributed with rate 2λ.*

Proof. We will establish this theorem by induction on n. For $n = 3$, we have one interior edge. The waiting time between the first speciation event (yielding two species) and the second speciation event (yielding three species) is the length of the interior edge. This waiting time is the time until the first of the two species speciates. As each species has a speciation rate λ, this waiting time is an exponential distribution with rate 2λ.

Now assume we established the theorem for $n = k$. For establishing the theorem for $k + 1$, we first note that the Yule tree on $k + 1$ species with the uniform prior for the time of origin corresponds to a Yule process forward in time which is stopped just before the $(k+1)$th speciation event (Hartmann et al., 2010). Now cut off the tree on $k + 1$ leaves $\mathcal{T}_{k+1}$ at time x_k, yielding a Yule tree on k leaves $\mathcal{T}_k$. A randomly selected interior edge $\mathcal{T}_k$ has an exponentially distributed length with rate 2λ (induction assumption). Each interior edge in $\mathcal{T}_k$ is also an interior edge in $\mathcal{T}_{k+1}$. Additionally, in $\mathcal{T}_{k+1}$, one of the pendant edges becomes an interior edge (the edge which speciates at time x_k). As each randomly selected pendant edge in $\mathcal{T}_k$ has an exponentially distributed length with rate 2λ (Corollary 3.2), the new interior

edge in $\mathcal{T}_{k+1}$ also has an exponentially distributed length with rate 2λ. This yields that a randomly picked interior edge in $\mathcal{T}_{k+1}$ has an exponentially distributed length with rate 2λ, and thereby establishes the induction step. □

3.2. Conditioning on n and x_1

In order to derive the edge length distribution, we need the following lemma:

Lemma 3.4. *The probability density function for the time of the kth speciation event in a reconstructed birth–death tree, conditional on having $n \geq 2$ extant species and on the first speciation event being at time x_1 is*

$$f_{n,k}(s|n,x_1) = (n-2)\binom{n-3}{k-2}G(s|x_1)^{n-k-1}(1 - G(s|x_1))^{k-2}g(s|x_1),$$
$$k = 2, \ldots n-1,$$

where

$$g(s|x_1) = \frac{p_1(s)}{p_0(x_1)}, \quad G(s|x_1) = \frac{p_0(s)}{p_0(x_1)},$$

with p_0 and p_1 the functions described in Eqs. (2) and (3).

Proof. In Gernhard (2008), we established that the speciation times $x_2 > \cdots > x_{n-1}$ are the order statistics of $n - 2$ i.i.d. random variables $s_2, \ldots, s_{n-2}$ with probability density $g(s|x_1)$. For the distribution, by integrating $g(s|x_1)$ with respect to s we obtain the function $G(s|x_1)$. Now, x_{k+1} is the kth order statistic (with the first order statistic being the largest and the $(n-2)$th order statistic being the smallest value), thus we have the probability density function for x_k (e.g. Dehling and Haupt , 2003)

$$f_{n,k+1}(s|t) = (n-1-k)\binom{n-2}{n-1-k}G(s|t)^{n-k-2}(1 - G(s|t))^{k-1}g(s|t).$$

Equivalently

$$f_{n,k+1}(s|t) = (n-2)\binom{n-3}{k-1}G(s|t)^{n-k-2}(1 - G(s|t))^{k-1}g(s|t). \quad \square$$

Theorem 3.5. *The length P of a randomly selected pendant edge in a reconstructed birth–death tree on n extant species and age x_1, over the range $0 \leq P < x_1$, has cumulative probability distribution $\int_0^P f_P(s|n,x_1)\,ds$, where*

$$f_P(s|n,x_1) = 2\frac{(n-2)}{n(n-1)}\frac{p_1(s)}{p_0(x_1)}\left((n-1) - (n-3)\frac{p_0(s)}{p_0(x_1)}\right).$$

For the boundary value $P = x_1$, the probability that $P = x_1$ is strictly positive and equals $2/(n(n-1))$.

Proof. The proof is similar to the proof of Theorem 3.1 in Mooers et al. (in press). The case $P = x_1$ corresponds to the pendant edge attaching to the root, an event which has probability $2/(n(n-1))$ (Eq. (1)). For $P < x_1$, we have

$$f_P(s|n,x_1) = \sum_{k=2}^{n-1} v(k)f_{n,k}(s|n,x_1)$$

$$= 2\frac{(n-2)}{n(n-1)}\sum_{k=2}^{n-1} k\binom{n-3}{k-2}G(s|x_1)^{n-k-1}(1 - G(s|x_1))^{k-2}g(s|x_1)$$

$$= 2\frac{(n-2)}{n(n-1)}g(s|x_1)\sum_{k=0}^{n-3}(k+2)\binom{n-3}{k}G(s|x_1)^{n-k-3}(1 - G(s|x_1))^k$$

$$= 2\frac{(n-2)}{n(n-1)}g(s|x_1)\left(2 + \sum_{k=1}^{n-3} k\binom{n-3}{k}G(s|x_1)^{n-k-3}(1 - G(s|x_1))^k\right)$$

$$= 2\frac{(n-2)}{n(n-1)}g(s|x_1)$$

$$\times\left(2+\sum_{k=0}^{n-4}(n-3)\binom{n-4}{k}G(s|x_1)^{n-k-4}(1-G(s|x_1))^{k+1}\right)$$

$$=2\frac{(n-2)}{n(n-1)}g(s|x_1)(2+(n-3)(1-G(s|x_1)))$$

$$=2\frac{(n-2)}{n(n-1)}g(s|x_1)((n-1)-(n-3)G(s|x_1)),$$

which establishes the theorem. $\square$

Evaluating the first moment integral of $f_P(s|n,x_1)$ yields,

Corollary 3.6. *For $0<\mu<\lambda$, the expected length of a pendant edge is*

$$\mathbb{E}[P|n,x_1]=\frac{1}{(n-1)n}\left(2x_1+\frac{(n-2)}{p_0(x_1)(1-e^{-(\lambda-\mu)x_1})}C\right),$$

where

$$C=\frac{(n-3)p_0(x_1)}{\lambda\mu}(\lambda-\mu e^{-(\lambda-\mu)x_1})$$
$$-\frac{x_1 p_1(x_1)(\lambda-e^{-(\lambda-\mu)x_1}\mu)}{\lambda^2}\left(\frac{4}{\lambda-\mu}(\lambda-e^{-(\lambda-\mu)x_1}\mu)-e^{-(\lambda-\mu)x_1}(n+1)\right)$$
$$-\frac{\log(1-\mu p_0(x_1))}{(\lambda\mu)^2}(e^{-(\lambda-\mu)x_1}\mu(-4\mu-(n+1)(\lambda-\mu))$$
$$-\lambda(-4\lambda+(n+1)(\lambda-\mu))).$$

For $\mu=\lambda$

$$\mathbb{E}[P|n,x_1]=\frac{1}{(n-1)n}\left(2x_1+\frac{(n-2)}{x_1 p_0(x_1)\lambda^2}\left((n-7)x_1+(n+1)p_0(x_1)\right.\right.$$
$$\left.\left.+\frac{6-2n+4\lambda x_1}{\lambda}\log[1+\lambda x_1]\right)\right).$$

Setting $\mu=0$ in Corollary 3.6 yields Theorem 2 of Mooers et al. (in press).

3.3. Conditioning on x_1

Lemma 3.7. *The probability of a reconstructed birth–death tree with the first speciation event being at time x_1 having $n\geq2$ extant descendants is*

$$p_n(x_1)=(n-1)\frac{p_1(x_1)^2(\lambda p_0(x_1))^{n-2}}{(1-\mu p_0(x_1))^2}.$$

Proof. The probability of a single individual having n extant and sampled offspring after time x_1, given that it has at least one offspring is (Kendall, 1949)

$$q_n(x_1)=\frac{p_1(x_1)(\lambda p_0(x_1))^{n-1}}{1-\mu p_0(x_1)}.$$

This leads to

$$p_n(x_1)=\sum_{k=1}^{n-1}\frac{p_1(x_1)(\lambda p_0(x_1))^{k-1}}{1-\mu p_0(x_1)}\frac{p_1(x_1)(\lambda p_0(x_1))^{n-k-1}}{1-\mu p_0(x_1)}$$
$$=(n-1)\frac{p_1(x_1)^2(\lambda p_0(x_1))^{n-2}}{(1-\mu p_0(x_1))^2}. \quad\square$$

Theorem 3.8. *The length P of a randomly selected pendant edge in a reconstructed birth–death tree of age x_1, over the range $0\leq P<x_1$, has cumulative probability distribution $\int_0^P f_P(s|x_1)\,ds$, where*

$$f_P(s|x_1)=2\frac{p_1(s)p_1(x_1)^2}{p_0(x_1)(1-\mu p_0(x_1))^2}\left(h(1,x_1)-\frac{p_0(s)}{p_0(x_1)}h(3,x_1)\right)$$

and where

$$h(k,x_1):=\frac{(k+1)\lambda p_0(x_1)-k}{(1-\lambda p_0(x_1))^2}-\frac{2k\log[1-\lambda p_0(x_1)]}{(\lambda p_0(x_1))^2}-\frac{2k}{\lambda p_0(x_1)}.$$

For the boundary value $P=x_1$, the probability that $P=x_1$ is strictly positive and equals

$$2\frac{(1-\lambda p_0(x_1))^2}{(\lambda p_0(x_1))^2}\cdot(-\log(1-\lambda p_0(x_1))-\lambda p_0(x_1)).$$

Proof. Using Theorem 3.5 and Lemma 3.7, for $P<x_1$, we have

$$f_P(s|x_1)=\sum_{n=3}^{\infty}f_P(s|n,x_1)p_n(x_1)$$
$$=\sum_{n=3}^{\infty}2\frac{(n-2)}{n(n-1)}\frac{p_1(s)}{p_0(x_1)}\left((n-1)-(n-3)\frac{p_0(s)}{p_0(x_1)}\right)$$
$$\times(n-1)\frac{p_1(x_1)^2(\lambda p_0(x_1))^{n-2}}{(1-\mu p_0(x_1))^2}$$
$$=2\frac{p_1(s)p_1(x_1)^2}{p_0(x_1)(1-\mu p_0(x_1))^2}\sum_{n=3}^{\infty}\frac{(n-2)}{n}$$
$$\times\left((n-1)-(n-3)\frac{p_0(s)}{p_0(x_1)}\right)(\lambda p_0(x_1))^{n-2}.$$

We also have

$$h(k,x_1):=\sum_{n=3}^{\infty}\frac{n-2}{n}(n-k)(\lambda p_0(x_1))^{n-2}$$
$$=\sum_{n=3}^{\infty}(n-2)(\lambda p_0(x_1))^{n-2}-\sum_{n=3}^{\infty}k\frac{n-2}{n}(\lambda p_0(x_1))^{n-2}$$
$$=\frac{\lambda p_0(x_1)}{(1-\lambda p_0(x_1))^2}-\frac{k}{1-\lambda p_0(x_1)}+k+\frac{2k}{(\lambda p_0(x_1))^2}\sum_{n=3}^{\infty}\frac{(\lambda p_0(x_1))^n}{n}$$
$$=\frac{(k+1)\lambda p_0(x_1)-k}{(1-\lambda p_0(x_1))^2}+k$$
$$+\frac{2k\left(-\log[1-\lambda p_0(x_1)]-\lambda p_0(x_1)-\frac{(\lambda p_0(x_1))^2}{2}\right)}{(\lambda p_0(x_1))^2}$$
$$=\frac{(k+1)\lambda p_0(x_1)-k}{(1-\lambda p_0(x_1))^2}-\frac{2k\log[1-\lambda p_0(x_1)]}{(\lambda p_0(x_1))^2}-\frac{2k}{\lambda p_0(x_1)}.$$

Overall, this yields

$$f_P(s|x_1)=2\frac{p_1(s)p_1(x_1)^2}{p_0(x_1)(1-\mu p_0(x_1))^2}\left(h(1,x_1)-\frac{p_0(s)}{p_0(x_1)}h(3,x_1)\right).$$

Using the identity $p_1(x_1)=(1-\mu p_0(x_1))(1-\lambda p_0(x_1))$ establishes the proof for $P<x_1$. For $P=x_1$, the probability of this event, by Eq. (1), is $\sum_{n\geq2}(2/(n(n-1)))\cdot p_n(x_1)$. Substituting the expression for $p_n(x_1)$ from Lemma 3.7, and applying the identities

$$p_1(x_1)=(1-\mu p_0(x_1))(1-\lambda p_0(x_1))\quad\text{and}\quad\sum_{n\geq2}\frac{y^n}{n}=-\log(1-y)-y,$$

we obtain the last equality claimed in the theorem. This completes the proof. $\square$

Evaluating the first moment integral of $f_P(s|x_1)$, or evaluating $\sum_{n=2}^{\infty}\mathbb{E}[P|n,x_1]p_n(x_1)$ with $\mathbb{E}[P|n,x_1]$ from Corollary 3.6, provides an analytic expression for the expected length of a randomly chosen pendant edge.

4. Length of a pendant edge in incompletely sampled trees

In this section, we generalize the results of Section 3 to the case of incomplete sampling. We do not have a general result for all sampling probabilities $0<f\leq1$, but we require $1\geq\mu/\lambda\geq1-f$. We show that for this parameter restriction, a reconstructed tree

under incomplete sampling is equivalent to a completely sampled tree and modified birth and death rates. Thus, the pendant edge lengths can be calculated assuming complete sampling and using the transformed rates.

In Stadler (2009), it is shown that a birth–death process with parameters λ, μ, f (with $\lambda \geq \mu \geq 0$ and $\mu - \lambda(1-f) \geq 0$), under Scenario (i) or (ii) induces the same distribution on reconstructed trees as a birth–death process with parameters $\hat{\lambda}, \hat{\mu}$ and complete extant species sampling, where

$$\hat{\lambda} = f\lambda, \quad \hat{\mu} = \mu - \lambda(1-f).$$

We note that the parameter restriction $\lambda \geq \mu \geq 0$ and $\mu - \lambda(1-f) \geq 0$ is equivalent to $1 \geq \mu/\lambda \geq 1-f$.

The following theorem establishes the analogue result under Scenario (iii).

Theorem 4.1. *Let T be a reconstructed tree with speciation times $x_1, x_2, \ldots, x_{n-1}$ which evolved under a birth–death process conditioned on the age of the most recent common ancestor being x_1. The probability density of T with parameters λ, μ, f (with $1 \geq \mu/\lambda \geq 1-f$) equals the probability density of T with birth rate $\hat{\lambda}$, death rate $\hat{\mu}$ (where $\hat{\lambda} = f\lambda, \hat{\mu} = \mu - \lambda(1-f)$), and complete sampling.*

Proof. The probability density of x is provided in Stadler (2010), Theorem 3.8, with $m = 0$,

$$f(T|x_1) = \left(\frac{q_1(x_1)}{1-q_0(x_1)}\right)^2 \cdot \prod_{i=2}^{n-1} \lambda q_1(x_i),$$

where

$$q_0(x_1) = 1 - \frac{f(\lambda-\mu)}{f\lambda + (\lambda(1-f)-\mu)e^{-(\lambda-\mu)x_1}},$$

$$q_1(x_1) = \frac{f(\lambda-\mu)^2 e^{-(\lambda-\mu)x_1}}{(f\lambda + (\lambda(1-f)-\mu)e^{-(\lambda-\mu)x_1})^2}.$$

The transformation $\hat{\lambda} = f\lambda, \hat{\mu} = \mu - \lambda(1-f)$ now establishes the theorem. □

Thus under Scenarios (i)–(iii), the reconstructed trees under incomplete sampling can be interpreted as completely sampled trees. We therefore showed in particular that the length of pendant edges of a birth–death tree with parameters λ, μ, f (where $1 \geq \mu/\lambda \geq 1-f$) is given by the equations assuming complete sampling (Section 3), using the birth rate $\hat{\lambda} = f\lambda$ and the death rate $\hat{\mu} = \mu - \lambda(1-f)$.

5. Length of a root edge in a Yule tree

A reconstructed tree has two edges descending from the root, and we denote these as e' and e'' where we may assume that e' is shorter than e'' (see Fig. 2). By a *root edge* we mean the selection of e' or e'' with equal probability. We will calculate the length of the root edge in a Yule tree (i.e. $\mu = 0, f = 1$) under Scenarios (i)–(iii).

We will show that, under Scenario (i), a root edge is longer than a randomly chosen interior edge (which has exponential distribution with parameter 2λ, Theorem 3.3).

5.1. Conditioning on n

Let X_i be an exponentially distributed random variable with parameter $i\lambda$. Then for $k \geq 2$, $I_k = \sum_{i=2}^{k} X_i$ is the hypo-exponential distribution with probability density

$$f_{I_k}(l) = \sum_{i=2}^{k} \lambda i e^{-\lambda i l} \prod_{j=2, j \neq i}^{k} \frac{j}{j-i},$$

which can be transformed to

$$f_{I_k}(l) = \sum_{i=2}^{k} \lambda e^{-\lambda i l}(-1)^{i-2} \frac{k!}{(k-i)!(i-2)!}$$

$$= k(k-1)\sum_{i=2}^{k} \lambda(-e^{-\lambda l})^i \binom{k-2}{i-2}. \tag{4}$$

Theorem 5.1. *The length L of one of the two edges descending from the root (picked uniformly at random) has probability density function*

$$f_L(l|n) = \lambda e^{-\lambda l}(1 - (1-e^{-\lambda l})^{n-2}(1-ne^{-\lambda l})).$$

Proof. In a Yule tree with n extant species, the shorter edge e' of the two edges descending from the root has a length L' that is exponentially distributed with parameter 2λ. The length L'' of the longer edge descending from the root, e'', is calculated as follows. First note that the waiting time, X_i, between the $i-1$th and ith speciation event is an exponential distribution with parameter $i\lambda$. Let e'' terminate at the kth speciation event in the tree. The length of A is then I_k with probability density given in Eq. (4). The probability that A terminates at the k-th speciation event $(2 < k < n)$ is

$$p_k = \frac{1}{k}\prod_{j=3}^{k-1}\left(1-\frac{1}{j}\right) = 1 \Big/ \binom{k}{2}$$

and the probability that A does not terminate until the present is

$$p_n = \prod_{j=3}^{n-1}\left(1-\frac{1}{j}\right) = 2/(n-1).$$

The length of e'' is

$$L'' = \sum_{k=3}^{n} I_k p_k$$

and thus the density function of L'' is

$$f_{L''}(l|n) = \sum_{k=3}^{n} f_{I_k}(l) p_k$$

$$= 2\sum_{k=3}^{n-1}\sum_{i=2}^{k} \lambda(-e^{-\lambda l})^i \binom{k-2}{i-2} + 2n\sum_{i=2}^{n} \lambda(-e^{-\lambda l})^i \binom{n-2}{i-2}$$

$$= 2\sum_{i=3}^{n-1}\sum_{k=i}^{n-1} \lambda(-e^{-\lambda l})^i \binom{k-2}{i-2} + 2\sum_{k=3}^{n-1} \lambda e^{-2\lambda l}$$

$$\quad + 2n\sum_{i=2}^{n} \lambda(-e^{-\lambda l})^i \binom{n-2}{i-2}$$

$$= 2\sum_{i=3}^{n-1} \lambda(-e^{-\lambda l})^i \binom{n-2}{i-1} + 2(n-3)\lambda e^{-2\lambda l}$$

$$\quad + 2n\sum_{i=2}^{n} \lambda(-e^{-\lambda l})^i \binom{n-2}{i-2}$$

$$= 2\sum_{i=3}^{n-1} \lambda(-e^{-\lambda l})^i \left(\binom{n-2}{i-1} + n\binom{n-2}{i-2}\right) + 2(2n-3)\lambda e^{-2\lambda l}$$

$$\quad + 2n\lambda(-e^{-\lambda l})^n$$

$$= 2\sum_{i=3}^{n-1} \lambda(-e^{-\lambda l})^i i\binom{n-1}{i-1}$$

$$\quad + 2(2n-3)\lambda e^{-2\lambda l} + 2n\lambda(-e^{-\lambda l})^n$$

$$= 2\lambda\sum_{i=3}^{n-1} i(-e^{-\lambda l})^i \binom{n-1}{i-1} + 2(2n-3)\lambda e^{-2\lambda l} + 2n\lambda(-e^{-\lambda l})^n$$

$$= 2\lambda e^{-\lambda l}(1 - e^{-\lambda l} - (1-e^{-\lambda l})^{n-2}(1-ne^{-\lambda l})).$$

The density of L is

$$f_L(l|n) = f_{L'}(l)/2 + f_{L''}(l)/2 = \lambda e^{-\lambda l}(1-(1-e^{-\lambda l})^{n-2}(1-ne^{-\lambda l})),$$

which establishes the theorem. □

Calculating the first moment of L yields:

Corollary 5.2. *The expected length L of one of the two edges descending from the root (selected at random) is $\mathbb{E}[L|n] = 1/\lambda$ $(1-(1/n))$.*

Comparing this result with Corollary 3.2 and Theorem 3.3 we see that for a Yule tree conditioned on having n species, the expected length of one of the two edges descending from the root, selected at random, is (asymptotically) *twice* the length of a randomly selected edge (or of a randomly selected pendant edge, or of a randomly selected interior edge).

5.2. Conditioning on x_1

Before considering Scenario (ii), we first consider Scenario (iii), i.e. the time since the root is x_1. In a Yule tree with the root having age x_1, select one of the two edges incident with the root uniformly at random (e.g. by a fair coin toss), and let L denote its length (up to time x_1). Then L has a discontinuous distribution

$$\mathbb{P}(L > l|x_1) = \begin{cases} e^{-\lambda l} & \text{for } 0 < t < x_1; \\ 0 & \text{for } l \geq x_1; \end{cases}$$

which implies that $\mathbb{E}[L|x_1] = (1/\lambda)(1-e^{-\lambda x_1})$, and so, in particular

$$\mathbb{E}[L|x_1] = \frac{1}{\lambda} - o(1),$$

where $o(1)$ is a term that goes to zero exponentially fast with x_1.

Note that the distribution of L should not be confused with the truncated exponential distribution arising from the conditional probability $\mathbb{P}(L > l|L \leq t)$, since this is the probability that a speciation event occurs on this lineage before or at time x_1.

5.3. Conditioning on n and x_1

We first consider a Yule tree starting from a single lineage at time 0 and grown for time t. Let K_t be the number of leaves at time t, and consider the length I of the initial edge up to time t. Let $\mathbb{P}(I > l|K_t = k)$ denote the probability that I is greater than l, conditional on the event that $K_t = k$.

Lemma 5.3.
$$\mathbb{P}(I > l|K_t = k) = \begin{cases} \alpha^{k-1} & \text{if } l < t; \\ 0 & \text{if } l \geq t; \end{cases}$$

where $\alpha = (1-e^{-\lambda(t-l)})/(1-e^{-\lambda t})$.

Proof. Let $f(s|K_t = k)$ be the density of I conditional on $K_t = k$. By Bayes' formula

$$f(s|K_t = k) = \frac{\mathbb{P}(K_t = k|I = s)f(s)}{\mathbb{P}(K_t = k)}, \tag{5}$$

where

$$f(s) = \lambda e^{-\lambda s}$$

is the (unconditional) first branch length. The unconditional distribution of K_t is

$$\mathbb{P}(K_t = k) = (1-e^{-\lambda t})^{k-1}e^{-\lambda t}.$$

Moreover, it is an easy exercise to show that, for $s < t$

$$\mathbb{P}(K_t = k|I = s) = (k-1)(1-e^{-\lambda(t-s)})^{k-2}e^{-2\lambda(t-s)}.$$

Combining these three expressions into (5) gives

$$f(s|K_t = k) = \frac{(k-1)\lambda e^{-\lambda(t-s)}(1-e^{-\lambda(t-s)})^{k-2}}{(1-e^{-\lambda t})^{k-1}}. \tag{6}$$

Now

$$\mathbb{P}(I > l|K_t = k) = \int_l^t f(s|K_t = k) \, ds \tag{7}$$

and so substituting (6), applying the substitution $u = 1-e^{-\lambda t}e^{\lambda s}$ and rearranging, one obtains

$$\mathbb{P}(I > l|K_t = k) = \left[\frac{k-1}{(1-e^{-\lambda t})^{k-1}}\right] \cdot \int_0^{1-e^{-\lambda t}e^{\lambda l}} u^{k-2} \, du, \tag{8}$$

which gives the value α^{k-1} for $l < t$. □

Consider now a Yule tree T and suppose we condition on both n and x_1. Select uniformly at random one of the edges of T that are incident with the root, let L denote its length up to time x_1, and let $\mathbb{P}(L > l|n,x_1)$ denote the probability that $L > l$, conditional on T having n leaves at time x_1.

Theorem 5.4.
$$\mathbb{P}(L > l|n,x_1) = \begin{cases} \dfrac{1}{n-1} \cdot \left(\dfrac{1-\alpha^{n-1}}{1-\alpha}\right) & \text{for } l \leq x_1; \\ 0 & \text{for } l > x_1. \end{cases} \tag{9}$$

Proof. It is a fundamental property of the Yule model that the number of leaves beneath the selected edge has a uniform distribution between 1 and $n-1$. Thus, if $l \leq x_1$ then

$$\mathbb{P}(L > l|n,x_1) = \frac{1}{n-1}\sum_{k=1}^{n-1}\mathbb{P}(I > l|K_{x_1} = k)$$

and the result now follows directly by Lemma 5.3. □

Let us now replace λ by its maximum likelihood (ML) estimate $\lambda_{ML} = \ln(n/2)/x_1$ and evaluate the expression for $\mathbb{P}(L > t|n,x_1)$ as $n \to \infty$. Note that if we let $\lambda = \ln(n/2)/t$ then

$$\alpha = \frac{1-\dfrac{2}{n}e^{\lambda l}}{1-\dfrac{2}{n}}$$

and so

$$\alpha^{n-1} = \frac{\left(1-\dfrac{2}{n}e^{\lambda l}\right)^{n-1}}{\left(1-\dfrac{2}{n}\right)^{n-1}} \sim \frac{\exp(-2e^{\lambda l})}{e^{-2}},$$

where $\lambda = \lambda_{ML} = \ln(n/2)/t$ and where $\sim$ denotes asymptotic equivalence as $n \to \infty$. Moreover

$$\frac{1}{n-1} \cdot \frac{1}{1-\alpha} = \frac{n}{2(n-1)}\frac{1-\dfrac{2}{n}}{e^{\lambda l}-1} \sim \frac{1}{2(e^{\lambda l}-1)}.$$

So, from (9), we obtain

$$\mathbb{P}(L > l|n,x_1) \sim \frac{1-e^{-w}}{w},$$

where $w = 2(e^{\lambda l}-1)$ and where $\sim$ denotes asymptotic equivalence as $n \to \infty$.

To determine the expectation of L conditional on n,x_1 we simply integrate this expression from $L=0$ to $L=x_1$ with respect to l (using the well known identity that $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x) \, dx$ for a non-negative continuous random variable X). Noting that $dl = dw/(\lambda(2+w))$, we obtain the following result which exhibits a different limit to the value $\frac{1}{2}\lambda$ for a randomly selected edge of T

or the limiting value $1/\lambda$ described above when we just condition on x_1.

Corollary 5.5. *The expected length of one of the two randomly selected root edges, conditional on n and x_1, and with λ set equal to its ML value, converges to c/λ as $n \to \infty$ where*

$$c = \int_0^\infty \frac{1-e^{-x}}{x(2+x)}\, dx = 0.8158\ldots$$

It is instructive to compare this result with Corollary 5.2.

6. Diversity in a Yule tree

In this section, we calculate the sum of all edge lengths in a Yule tree which is also called the diversity.

6.1. Conditioning on n

Theorem 6.1. *The sum D of all branches in a Yule tree with $n \geq 2$ leaves has a gamma distribution with density function*

$$f_D(d|n) = \frac{\lambda^{n-1}e^{-\lambda d}d^{n-2}}{(n-2)!}.$$

In particular, D has mean $\mathbb{E}[D|n] = (n-1)/\lambda$ and variance $\mathrm{Var}[D|n] = (n-1)/\lambda^2$ and is asymptotically normally distributed.

Proof. The sum of all edge lengths in a Yule tree with n leaves is $D = \sum_{j=2}^n jX_j$ where X_j has an exponential distribution with parameter λj. Note that jX_j has an exponential distribution with parameter $1/j \cdot \lambda j = \lambda$, and so D is a sum of $n-1$ independent exponential random variables, each having parameter λ, and so D has the claimed gamma distribution. $\square$

Remark 6.2. Theorem 6.1 provides an alternative proof for a random edge in a Yule tree having expected length $1/2\lambda$ (which we derived in Corollary 3.2 and Theorem 3.3): the expected sum of all edges in a Yule tree on n leaves is $(n-1)/\lambda$ (Theorem 6.1). As a tree on n leaves has $2(n-1)$ edges, a random edge has expected length $1/2\lambda$.

6.2. Conditioning on n and x_1

As shown in Gernhard (2008), a Yule tree of age x_1 with $n \geq 3$ leaves, the $n-2$ speciation events $S_2,\ldots,S_{n-1}$ descending from the root are i.i.d. random variables with density

$$g(s) = \frac{\lambda e^{-\lambda s}}{1-e^{-\lambda x_1}}.$$

The sum of all branch lengths is the sum of the $n-2$ speciation times and $2x_1$ (accounting for the two branches descending from the root).

Theorem 6.3. *The sum D of all branches in a Yule tree of age x_1 and n leaves has moment generating function*

$$M_D(s|n,x_1) = e^{2x_1 s} \cdot \left(\frac{\lambda(1-e^{(s-\lambda)x_1})}{(\lambda-s)(1-e^{-\lambda x_1})} \right)^{n-2}.$$

Proof. Let C denote the sum of the $n-2$ speciation times $S_2,\ldots,S_{n-1}$. Then C is a convolution of $n-2$ random variables with probability density $g(s) = \lambda e^{-\lambda s}/(1-e^{-\lambda x_1})$. If $M_C(s|n,x_1)$ denotes the moment generating function for C, then

$$M_C(s|n,x_1) = \mathbb{E}[e^{Cs}] = \mathbb{E}[e^{\sum_{j=2}^{n-1} S_j s}] = \prod_{j=2}^{n-1} \mathbb{E}[e^{S_j s}]$$

$$= \left(\int_{-\infty}^\infty e^{sx}g(x)\, dx \right)^{n-2}$$

$$= \left(\frac{\lambda}{1-e^{-\lambda x_1}} \int_0^{x_1} e^{sx-\lambda x}\, dx \right)^{n-2}$$

$$= \left(\frac{\lambda(1-e^{(s-\lambda)x_1})}{(\lambda-s)(1-e^{-\lambda x_1})} \right)^{n-2}.$$

Now $D = 2x_1 + C$ and so $M_D(s|n,x_1) = e^{2x_1 s}M_C(s|n,x_1)$, which leads directly to the expression claimed. $\square$

6.3. Conditioning on x_1

The sum D of all branches in a Yule tree of age x_1 has probability density function

$$f_D(d|x_1) = \sum_{n=2}^\infty p_n(x_1)f_D(d|n,x_1),$$

where $f_D(d|n,x_1)$ is the density conditional on n and x_1. In particular, noting that $\mathbb{E}[D|n,x_1] = (d/ds)M_D(s|n,x_1)_{|s=0}$, we have

$$\mathbb{E}[D|x_1] = \sum_{n=2}^\infty p_n(x_1)\frac{d}{ds}M_D(s|n,x_1)_{|s=0},$$

which gives

$$\mathbb{E}[D|x_1] = \frac{2}{\lambda}(e^{\lambda x_1}-1),$$

a result that was derived via a different argument in Steel and Mooers (2010).

7. Conclusion

In this paper, we derive the probability density and expectation for the length of a randomly picked pendant edge of a reconstructed birth–death tree. We investigate this under three scenarios in which the resulting reconstructed trees induced by the birth–death process are finite: we either consider trees (i) with a fixed number of leaves, (ii) with a fixed number of leaves and a fixed age of the most recent common ancestor of the extant sampled species, or (iii) with a fixed age of the most recent common ancestor of the extant sampled species.

We noted in Section 4 that under our three Scenarios (i)–(iii), the original process with sampling (parameters λ,μ,f, with $1 \geq \mu/\lambda \geq 1-f$) can be transformed into a birth–death process and complete sampling (with a transformed birth rate $\hat{\lambda} = f\lambda$ and a transformed death rate $\hat{\mu} = \mu - \lambda(1-f)$). Hence, we state all results on the length of pendant edges assuming complete sampling (Section 3). It remains a future task to generalize the results such that we can calculate the pendant edge lengths for $0 \leq \mu/\lambda < 1-f$, i.e. for small turnover μ/λ and sparse sampling.

For the Yule model (i.e. $\mu = 0, f = 1$) we further determine under our three Scenarios (i)–(iii) the probability density for the length of an edge descending from the root and the sum of all edges, which is also known as the diversity. In particular, we show that the pendant as well as the interior edge lengths under Scenario (i) are exponentially distributed with parameter 2λ while an edge descending from the root is asymptotically twice as long compared to a randomly chosen edge. Furthermore, the diversity follows a gamma distribution.

Knowledge of the branch lengths and diversity in a phylogenetic tree is important for conservation strategies. When present-day species become extinct, short pendant edges cause little diversity loss while long pendant edges cause severe diversity

loss. In previous work, the expected loss of diversity was considered under constant rate birth–death models with mild extinction (Mooers et al., in press), and under constant rate birth–death models with severe extinction, namely $\lambda = \mu$ (Nee and May, 1997) (note that (Nee and May, 1997) derived the results under the coalescent with constant population size; in expectation, such coalescent trees with n leaves equal birth–death trees with $\lambda = \mu$ up to scaling of branch lengths; branch lengths are equal in expectation for $\mu = \lambda = n/2$ (Gernhard, 2008)). Characterizing the full distribution of the loss of diversity remains an open task. Our results yielding the branch length and diversity distribution might be a first step towards characterizing the distribution of diversity loss (rather than only the expectation). Knowledge of the distribution will help understand the stochastic variability of diversity loss. An asymptotic normal law for diversity loss was recently established by Faller et al. (2008), but this was established only for certain deterministic classes of trees, rather than for trees generated by a stochastic process.

For future work, it would be interesting to compare the branch lengths and diversity (loss) under the constant rate birth–death process to models under which the rates of speciation and extinction may vary. In particular, speciation and extinction rates may dependent on the number of species that are extant at that moment (i.e. density-dependent speciation) (Rabosky and Lovette, 2008), or on time (i.e. environmental-dependent speciation) (Nee et al., 1994; Stadler, 2011), or on a particular trait (i.e. trait-dependent speciation) (Maddison, 2007).

Acknowledgments

References

Aldous, D., Popovic, L., 2005. A critical branching process model for biodiversity. Adv. Appl. Prob. 37 (4), 1094–1115.

Aldous, D.J., 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. 16 (1), 23–34.

Blum, M., François, O., 2005. Minimal clade size and external branch length under the neutral coalescent. Adv. Appl. Prob. 37 (3), 647–662.

Dehling, H., Haupt, B., 2003. Einfuehrung in die Wahrscheinlichkeitstheorie und Statistik. Springer.

Faller, B., Pardi, F., Steel, M., 2008. Distribution of phylogenetic diversity under random extinction. J. Theor. Biol. 251 (2), 286–296.

Gernhard, T., 2008. The conditioned reconstructed process. J. Theor. Biol. 253 (4), 769–778.

Gernhard, T., 2008. New analytic results for speciation times in neutral models. Bull. Math. Biol. 70 (4), 1082–1097.

Hartmann, K., Wong, D., Stadler, T., 2010. Sampling trees from evolutionary models. Syst. Biol. 59 (4), 465–476.

Kendall, D.G., 1949. Stochastic processes and population growth. J. R. Stat. Soc. Ser. B 11, 230–264.

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

Mooers, A., Gascuel, O., Stadler, T., Li, H., Steel, M. Branch lengths on birth–death trees and the expected loss of phylogenetic diversity. Syst. Biol, doi:10.1093/sysbio/syr090. In press.

Mooers, A.O., Heard, S.B., 1997. Inferring evolutionary process from phylogenetic tree shape. Quart. Rev. Biol. 72 (1), 31–54.

Nee, S.C., 2001. Inferring speciation rates from phylogenies. Evolution 55 (4), 661–668.

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

Nee, S.C., May, R.M., Harvey, P., 1994. The reconstructed evolutionary process. Philos. Trans. R. Soc. Ser. B 344, 305–311.

Paradis, E., 1998. Testing for constant diversification rates using molecular phylogenies: a general approach based on statistical tests for goodness of fit. Mol. Biol. Evol. 15 (4), 476.

Pybus, O.G., Harvey, P.H., 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Ser. B 267 (1459), 2267–2272.

Rabosky, D., Lovette, I., 2008. Density-dependent diversification in North American wood warblers. Proc. R. Soc. Ser. B 275 (1649), 2363.

Ricklefs, R., Losos, J., Townsend, T., 2007. Evolutionary diversification of clades of squamate reptiles. J. Evol. Biol. 20 (5), 1751–1762.

Stadler, T., 2008. Lineages-through-time plots of neutral models for speciation. Math. Biosci. 216, 163–171.

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

Stadler, T., 2010. Sampling-through-time in birth–death trees. J. Theor. Biol. 267 (3), 396–404.

Stadler, T., 2011. Mammalian phylogeny reveals recent diversification rate shifts. Proc. Natl. Acad. Sci. 108 (15), 6187–6192.

Steel, M., Mooers, A., 2010. Expected length of pendant and interior edges of a Yule tree. Appl. Math. Lett. 23 (11), 1315–1319.

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