ELSEVIER

Mathematical Biosciences 170 (2001) 91–112

Mathematical Biosciences

www.elsevier.com/locate/mbs

Properties of phylogenetic trees generated by Yule-type speciation models ☆

Mike Steel *, Andy McKenzie

Department of Mathematics and Statistics, Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Received 12 April 2000; received in revised form 7 September 2000; accepted 7 November 2000

Abstract

We investigate some discrete structural properties of evolutionary trees generated under simple null models of speciation, such as the Yule model. These models have been used as priors in Bayesian approaches to phylogenetic analysis, and also to test hypotheses concerning the speciation process. In this paper we describe new results for three properties of trees generated under such models. Firstly, for a rooted tree generated by the Yule model we describe the probability distribution on the depth (number of edges from the root) of the most recent common ancestor of a random subset of k species. Next we show that, for trees generated under the Yule model, the approximate position of the root can be estimated from the associated unrooted tree, even for trees with a large number of leaves. Finally, we analyse a biologically motivated extension of the Yule model and describe its distribution on tree shapes when speciation occurs in rapid bursts.

Keywords: Trees; Phylogeny; Speciation; Yule model; Maximum likelihood

1. Introduction

Phylogenetic trees are widely used in biology to represent evolutionary relationships between species. In these trees the leaves represent extant species, and the internal vertices represent hypothesised speciation events. There is much interest in the process of speciation, and the extent and manner in which the distribution of phylogenetic tree shapes can be modelled by a

☆ This research was supported by the New Zealand Marsden Fund (UOC-MIS-003).

* Corresponding author: Tel.: +64-3 366 7001 ext. 7688; fax: +64-3 364 2587.
E-mail address: m.steel@math.canterbury.ac.nz (M. Steel).

0025-5564/01/$ - see front matter
PII: S0025-5564(00)00061-4

random process. Several simple stochastic models of speciation have been proposed and several investigators have aimed to test or refine such models by comparing their predictions with published phylogenetic trees [1–9]. These models make predictions about the shape of the phylogenetic tree connecting the extant species. These models can provide prior probabilities for phylogenetic trees in Bayesian approaches to tree reconstruction [10–12], and they are also used as a basis for calculating the probability of certain configurations under random speciation [13]. These probabilities may then be useful in testing hypotheses concerning the speciation process.

In this paper we will consider just the model's predictions regarding the discrete underlying tree structure, without regard to the lengths of the edges. While such an approach may neglect some informative characteristics of the tree, the approach has two motivations – firstly, the predictions regarding the discrete tree remain valid under a much wider class of models (they are insensitive to underlying parameters) and, secondly, we are interested in isolating out the information that is conveyed solely by the discrete tree shape.

In this paper we consider some properties of the Yule model, which is perhaps the simplest stochastic model for speciation. We then define and investigate an extension of this model. We begin by introducing some basic terminology for phylogenetic trees (Section 2). The Yule model is then introduced, and some of its properties are described (Section 3). We then consider the probability distribution on the number of edges separating the root of a tree from the most recent common ancestor of a randomly selected subset of size k (Section 4). Next, a maximum likelihood approach to edge-rooting an unrooted tree is presented, and simulation is used to show that even for large unrooted trees the approximate location of the root can be identified with high probability (Section 5). Following this a modification of the Yule model is considered in which the rate of speciation of a lineage is dependent on the time back to the last speciation event on that lineage (Section 6). We show that this modified model reduces to the uniform model under the condition of 'explosive radiation'.

2. Terminology

Evolutionary relationships are generally represented by rooted or unrooted *binary* (*phylogenetic*) *trees* [14]. Such trees consist of uniquely labeled *vertices* of degree 1 called *leaves* and unlabeled *internal* vertices of degree 3 (also, in case the tree is rooted, it contains an additional *root vertex* of degree 2 – in this way every vertex can be regarded as having exactly two descendants). We say a vertex v is a *descendant* of another vertex w, if w lies on the path between v and the root vertex. Edges adjacent to a leaf are called *pendant edges*, while all other edges are *internal*. A (tree) *shape* is the unlabeled tree obtained by dropping the labeling of the leaves of a binary phylogenetic tree. For further clarification of these terms see Fig. 1.

Throughout this paper we will use T to denote a phylogenetic tree, and τ to denote a tree shape. We will frequently use the asymptotic expression $f(n) \sim g(n)$ to denote $\lim_{n\to\infty} f(n)/g(n) = 1$. As usual, $\mathbb{P}[A]$ (resp. $\mathbb{P}[A|B]$) denotes the probability of event A (resp. the conditional probability of event A given B), and $\mathbb{E}[X]$ (resp. $\mathbb{E}[X|Y]$) denotes the expectation of random variable X (resp. the conditional expectation of X given Y).

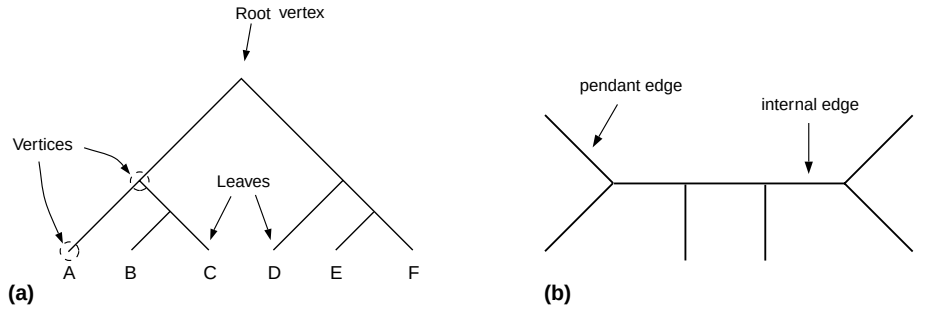

Fig. 1. Some terminology for trees: (a) a rooted binary phylogenetic tree with six leaves; (b) an unrooted binary tree shape with six leaves.

3. The Yule model

A simple model of speciation is to assume the exchangeability condition that, at any given time, each of the then-extant species are equally likely to give rise to one new species. The 'rate' of speciation may vary with time, or with the present and past number of species. Also we may allow extinctions (or random sampling of extant taxa) provided that a similar exchangeability criterion applies – that is, whenever an extinction event occurs each of the then-extant species is equally likely to go extinct. Depending on how the various parameters are set in such a model, we obtain various probability densities over all edge-weighted trees that connect a group of extant species. However, if we simply regard these trees as unlabeled discrete graphs without edge length (tree shapes) then the underlying parameters and details do not affect the resulting discrete probability distribution, provided the exchangeability criteria still apply (see Ref. [1]). This distribution on tree shapes is often called the *Yule model* and it has been widely studied [3,15–17].

We can reformulate this model in the discrete setting, by evolving a (discrete) tree shape under the following rule. We start with the rooted tree on two leaves and repeat the following procedure until the tree has n leaves:

> *For the tree shape so far constructed, select a leaf randomly and uniformly, and make it the direct ancestor of two new descendent leaves.*

Alternatively, we may attach an edge added uniformly and randomly to a pendant edge at each step. This process is illustrated in Fig. 2.

This process provides a probability distribution on rooted tree shapes and also on unrooted tree shapes (by suppressing the root). Also if species are assigned to the leaves in random order we also obtain probability distributions on rooted and unrooted phylogenetic trees [3].

The Yule model arises in a number of seemingly different ways. For example, in the context of population genetics, one has the *coalescent* model [1,18,19]. In this model one starts with n objects, then picks two at random to coalesce, giving $n-1$ objects. This process is repeated until there is only a single object left. If this process is reversed, starting with one object to give n objects, then it is equivalent to the Yule model. Note that in the coalescent model there is commonly a probability distribution for the times of coalescences, but in the Yule model we ignore this element.

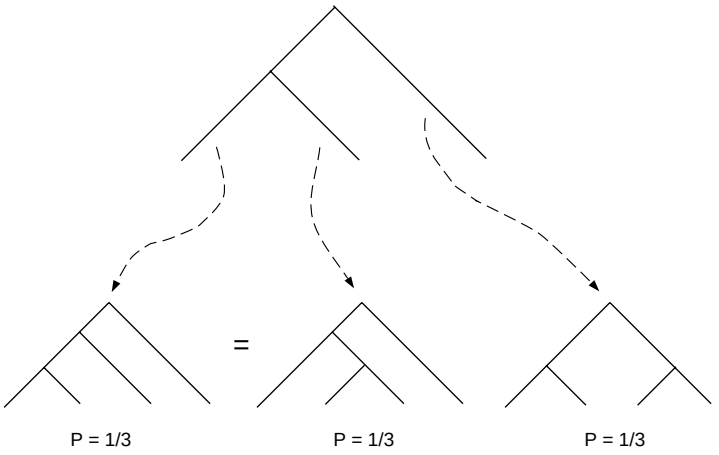

Fig. 2. The Yule model probabilities for shapes with four leaves. A shape on four leaves is formed by the splitting of one of the pendant edges of the shape on three leaves. Each pendant edge has the same probability of splitting, so for the shape on three leaves each pendant edge has a probability of $1/3$ of splitting. The resulting symmetric shape on four leaves has a probability of $1/3$. The other two shapes on four leaves are the same (up to rotation about internal vertices), and so the probability of this shape is $2/3$.

Another closely related realisation of the Yule model is obtained as follows. Given a rooted binary phylogenetic tree T, let $\mathring{V}$ denote the set of internal vertices of T. A *ranking* of T is a function r that associates to each vertex $v \in \mathring{V}$ of T a unique element from the set $\{1, 2, \ldots, |\mathring{V}|\}$ in such a way that $r(v_1), r(v_2), \ldots$ is strictly increasing along any sequence $v_1, v_2, \ldots$ of vertices directed away from the root. Thus we might regard r as describing the order of the speciation events that are represented by the internal vertices of T. Observe that a phylogenetic tree having the shape of the right-most tree in Fig. 2 has exactly two possible rankings, while for the left-most tree there is just one possible ranking. The pair (T, r) is sometimes called a *labeled history*. If we now select a labeled history (T, r) on n species uniformly and consider just T, then this once again leads to the Yule distribution on rooted binary phylogenetic trees. Furthermore, if we consider just the shape of T we obtain the Yule model on rooted tree shapes.

This connection with labeled histories provides a convenient tool for describing the probability distribution of a tree shape τ, since it is possible to count the number of labeled histories, and rankings on a given tree. For a vertex v of a rooted binary phylogenetic tree, let $\delta(v)$ denote the number of internal vertices (including v) that are descendants of v (v' is a descendant of v if the path from v' to the root includes v). Note that $\delta(v)$ is equal to one less than the number of leaves of the tree that are descendants of v. The following lemma has been proved before using Young tableaux [20]; here we give another proof using poset theory.

Lemma 1. *For a rooted binary phylogenetic tree with n leaves, the number of associated labeled histories is precisely*

$$\frac{(n-1)!}{\prod_{v \in \mathring{V}} \delta(v)},$$

where $\mathring{V}$ denotes the set of internal vertices of the tree.

Proof. If we regard the internal vertices of T as forming a partially ordered set by directing all edges away from the root then, in the parlance of poset theory, we are counting the number of linear extensions of this partially ordered set, which is a well-studied problem. In general, given an m-element partially ordered set P, if we let $\lambda_x = \{y \in P : y \geqslant x\}$, then the number of linear extensions of P equals

$$\frac{m!}{\prod_{x\in P} \lambda_x} \tag{1}$$

when the 'Hasse diagram' of P is a rooted tree, whose root is a minimal element in the poset [21]. In the current setting, this applies with $m = n - 1$ and $\lambda_x = \delta(x)$. □

We next recall a well-known and elegant expression for the total number of labeled histories on n species. Each labeled history on n species is generated in a unique way by the coalescent process that starts with the n leaves and works back up to the root, repeatedly combining some pair of species to form a new amalgamated species that represents that pair. Since there are $\binom{i}{2}$ choices for the pair to be amalgamated if there are i objects (species and amalgamated species) present, one obtains the following result [22].

Lemma 2. *The total number of labeled histories on n species is*

$$\frac{n!(n-1)!}{2^{n-1}}. \tag{2}$$

For a rooted phylogenetic tree T let $P_{\mathrm{Y}}(T)$ denote the probability of generating T under the Yule model. The following result was established by Brown [2] using induction. Here we provide an alternative proof.

Proposition 1.

$$P_{\mathrm{Y}}(T) = \frac{2^{n-1}}{n!} \prod_{v\in \mathring{V}} \delta(v)^{-1}$$

where $\mathring{V}$ is the set of internal vertices of T.

Proof. Under the Yule distribution (realised via the uniform distribution on labeled histories) the probability of generating T is simply number of rankings for T divided by the total number of ranked binary phylogenetic trees. Now T has exactly $(n-1)!/(\prod_{v\in \mathring{V}} \delta(v))$ possible rankings by Lemma 1. Dividing by the total number of ranked binary phylogenetic trees given by Lemma 2 we obtain the result. □

We describe two further properties of the Yule model that will be useful later. Suppose we generate a rooted phylogenetic tree under the Yule model, and randomly select (with equal probability) one of the two subtrees incident with the root of this tree. Then the number of leaves

in this subtree is uniformly distributed between 1 and $n-1$ [13]. That is, if we let $p_Y(i)$ denote the probability that the number of leaves in this subtree is exactly i, then

$$p_Y(i) = \frac{1}{n-1} \tag{3}$$

for $i = 1, 2, \ldots, n-1$.

Suppose now we let s_1 denote a particular species from our set of n species. Let $p^*_Y(i)$ denote the probability that the subtree of T incident with the root that contains species s_1 has exactly i leaves. Then,

Lemma 3.

$$p^*_Y(i) = \frac{2i}{n(n-1)}$$

for $i = 1, 2, \ldots, n-1$.

Proof. Randomly select (with equal probability) one of the two subtrees of T incident with the root and let A denote the set of species that appear as leaves in this subtree. Then, $p^*_Y(i) = \mathbb{P}[|A| = i \mid s_1 \in A]$. By Bayes' theorem,

$$\mathbb{P}[|A| = i \mid s_1 \in A] = \frac{\mathbb{P}[s_1 \in A \mid |A| = i]\mathbb{P}[|A| = i]}{\mathbb{P}[s_1 \in A]} \tag{4}$$

and $\mathbb{P}[s_1 \in A \mid |A| = i] = i/n$, $\mathbb{P}[s_1 \in A] = 1/2$, and, by (1), $\mathbb{P}[|A| = i] = 1/(n-1)$. The result now follows. □

A further important property of the Yule model is that it satisfies the following *hereditary* property. Let us generate a rooted binary tree T according to the Yule model, and let t_1, t_2 denote the two subtrees of T incident with the root. Let $S(t_1)$ denote the subset of species that label t_1, and let S denote a fixed subset of species. Then, conditional on the event that S is the set of species labeling the leaves of t_1 the probability distribution on t_1 is also the Yule distribution. This property follows from a particular case of the *group elimination property*, described by Aldous [1].

4. Depth of a most recent common ancestor

Suppose we evolve a rooted phylogenetic tree T on n extant species under the Yule model, and we select a random subset S of k extant species. Let $X_{n,k}$ denote the number of edges separating the root of T from the vertex in T that corresponds to the most recent common ancestor (MRCA) of S. In this section we investigate the probability distribution of $X_{n,k}$ for various values of k, particularly in the limit as n becomes large. Some of the reasons why a biologist might be interested in such questions are discussed by Sanderson [23]; the cases $k = 1$ and $k = 2$ are also of some independent interest as we will see.

Note that although we will regard S as a random subset of the n species, our results would apply even if we regard S as a fixed set of species, since we are investigating properties of S in a tree that is generated by a model that assigns equal probability to all possible labelings of the leaves by the n species. Also, whenever we talk about the *distance* between two vertices in a tree, we are referring to the number of edges separating the two vertices (also called the *graph distance*).

4.1. Distance of MRCA from root

The case $k = 1$ corresponds to the distance of a randomly selected leaf from the root, and has been analysed before [24,25]. In the following theorem $c(n,q)$ denotes the unsigned Stirling number of the first kind, which is the number of permutations on n elements that have exactly q cycles [26].

Theorem 1 [24,25]. *Let P^q_{n+1} be the probability that a randomly chosen leaf from a tree on $n+1$ leaves has distance q from the root. Under the Yule model we have*

$$P^q_{n+1} = \frac{2^q c(n,q)}{(n+1)!}.$$

Furthermore, the mean (μ_n) and variance (σ_n^2) of this distribution are given by

$$\mu_n = 2\sum_{j=2}^{n} \frac{1}{j}; \quad \sigma_n^2 = 2\sum_{j=2}^{n} \frac{1}{j} - 4\sum_{j=2}^{n} \frac{1}{j^2},$$

where $\mu_1 = 0$ and $\sigma_1^2 = 0$.

For $k > 1$, the asymptotic probability $\mathbb{P}[X_{n,k} = 0]$ has already been determined by Sanderson [23] who showed that

$$\lim_{n\to\infty} \mathbb{P}[X_{n,k} = 0] = 1 - \frac{2}{k+1}.$$

We generalise this result by (i) providing an exact, closed-form expression for $\mathbb{P}[X_{n,k} = 0]$ and (ii) showing that, as n becomes large, $X_{n,k}$ has a geometric distribution with parameter $2/(k+1)$.

Theorem 2.
1. $\mathbb{P}[X_{n,k} = 0] = 1 - 2/(k+1) \times (n-k)/(n-1)$.
2. *For $k > 1$, $r \geqslant 0$, $\lim_{n\to\infty} \mathbb{P}[X_{n,k} \geqslant r] = (2/(k+1))^r$.*

Proof. By exchangeability we may assume that S consists of a fixed species s_1 together with $k-1$ species randomly selected from the remaining $n-1$ species. Generate a tree T on n species under the Yule model, and let $v_0, v_1, \ldots, v_q$ denote the vertices on path in T from the root to the leaf v_q labeled by s_1, and let N_i denote the total number of leaves of T descendant from vertex v_i for $i = 0, \ldots, q$. Thus, N_i is a strictly decreasing sequence, with $N_0 = n$, $N_q = 1$.

Now, $X_{n,k} \geqslant r$ precisely if all the elements of S are descendants of v_r. Provided $q > r$, species s_1 is a descendant of v_r, and since the remaining $k-1$ species in S are randomly selected from the

remaining $n-1$ species then, conditional on N_r, the probability that they are all descendants of v_r is exactly

$$\frac{\binom{N_r-1}{k-1}}{\binom{n-1}{k-1}}.$$

Thus, if we adopt the convention that $\binom{a}{b}=0$ for $a<b$, and extend the sequence $N_0,\ldots,N_q$ by setting $N_i=1$ for all $i>q$, then we can write,

$$\mathbb{P}[X_{n,k}\geqslant r \mid N_r=i]=\frac{\binom{i-1}{k-1}}{\binom{n-1}{k-1}}=\frac{(i-1)\cdots(i-k+1)}{(n-1)\cdots(n-k+1)}. \tag{5}$$

Now, for $j>1, s>0$,

$$\mathbb{P}[N_s=i \mid N_{s-1}=j]=\frac{2i}{j(j-1)}$$

by Lemma 3, and the hereditary property of the Yule model (described at the end of Section 3). Consequently, using the identities

$$i(i-1)\cdots(i-k+1)=\binom{i}{k}k!$$

and

$$\sum_{i=1}^{j-1}\binom{i}{k}=\binom{j}{k+1}$$

we obtain,

$$\mathbb{P}[X_{n,k}\geqslant r \mid N_{r-1}=j]=\frac{2}{j(j-1)}\sum_{i=1}^{j-1}\frac{i(i-1)\cdots(i-k+1)}{(n-1)\cdots(n-k+1)}=\frac{2}{k+1}\,\frac{(j-2)\cdots(j-k)}{(n-1)\cdots(n-k+1)}.$$

Taking $r=1$, and noting that $N_0=n$ with probability 1, we obtain the first part of the theorem.

For part two we use the identity

$$(j-2)\cdots(j-k)=(j-1)\cdots(j-k+1)-(k-1)\times(j-2)\cdots(j-k+1)$$

(arising from $\binom{a}{b}+\binom{a}{b-1}=\binom{a+1}{b}$) to write

$$\mathbb{P}[X_{n,k}\geqslant r \mid N_{r-1}=j]=\frac{2}{k+1}\,\frac{(j-1)\cdots(j-k+1)}{(n-1)\cdots(n-k+1)}+\mathrm{O}(n^{-1}).$$

We now observe that the first term on the right-hand side of this equation is the same as the term on the right-hand side of (5), except with i replaced by j, and a multiplicative factor of $2/(k+1)$. Consequently, by repeating this step $r-1$ times we have

$$\mathbb{P}[X_{n,k} \geqslant r] = \mathbb{P}[X_{n,k} \geqslant r \mid N_0 = n] = \left(\frac{2}{k+1}\right)^r + \mathrm{O}(2^r \times n^{-1})$$

as required. □

4.2. The expected distance between two leaves

The case $k = 2$ allows us to obtain an expression for the expected distance between two randomly selected leaves in a rooted binary tree with n leaves generated by the Yule model. Recall that the distance $d(v_1, v_2)$ between a pair of vertices v_1, v_2 of T is being measured by the number of edges separating them. Let $v_{ij} \in \mathring{V}$ denote the most recent common ancestor of i and j in the tree, and let ρ denote the root of the tree. Then for leaves i, j of T,

$$d(i,j) = d(i,\rho) + d(j,\rho) - 2d(v_{ij},\rho)$$

and so

$$\mathbb{E}[d(i,j)] = \mathbb{E}[d(i,\rho)] + \mathbb{E}[(j,\rho)] - 2\mathbb{E}[d(v_{ij},\rho)].$$

Now, $\mathbb{E}[d(i,\rho)] = \mathbb{E}[d(j,\rho)] = \mu_n$ (see Theorem 1) while $\mathbb{E}[d(v_{ij},\rho)] = \mathbb{E}[X_{n,2}]$ and so

$$\mathbb{E}[d(i,j)] = 2\mu_n - 2\mathbb{E}[X_{n,2}]. \tag{6}$$

By Theorem 2, $\lim_{n\to\infty} \mathbb{E}[X_{n,2}] = 2$ and so if μ_n^* denotes the expected distance between two randomly selected leaves, then

$$\lim_{n\to\infty} 2\mu_n - \mu_n^* \sim 4. \tag{7}$$

Actually, it is possible to derive an exact expression for $\mathbb{E}[X_{n,2}]$, namely

$$\mathbb{E}[X_{n,2}] = 2\left(1 - \frac{\mu_n}{n-1}\right)$$

which provides an exact expression for $\mathbb{E}[d(i,j)]$.

5. Rooting an unrooted tree

Typically, construction of an evolutionary tree for a set of species is a two stage process. In the first stage, using biological data of some sort, an unrooted tree is constructed. In the next stage, the unrooted tree is rooted at some point. Commonly this is done by outgroup comparison, or using some auxiliary data (for example embryological or fossil data) [27].

However, in some circumstances an outgroup is not available, or the auxiliary data is unclear. Furthermore, the choice of outgroup can strongly influence the accuracy of tree reconstruction [28]. In these circumstances heuristic methods provide an alternative way to root the tree. For example, in the *midpoint method*, the root is located at the point halfway between the two leaves that are the furtherest distance apart [29,30]. In another approach the root is located at a point where the mean distance to the species on either side is the same (for example, the program TREECON [31] uses this method). Here we explore a third alternative, based on the structure of the trees under the Yule model.

Before proceeding further we introduce some terminology. For a rooted binary tree T' the associated unrooted binary tree T is obtained from T' by suppressing the root and identifying the two edges incident with the root to form a single edge e – we call this edge the *root edge* of T. Given T and its root edge, one can easily recover the rooted tree by *subdividing* e – that is, by placing a new (root) vertex at the midpoint of the root edge.

In applications, one is typically given just the unrooted tree T and one would like to estimate which edge is the root edge, or at least find a small subset of edges that contains the root edge with high probability.

5.1. Maximum likelihood estimation of the root edge

Suppose we have a stochastic model (such as the Yule model) for the generation of rooted binary phylogenetic trees. Given an unrooted binary tree, T, and an edge, e, let $\mathbb{P}[T, e]$ denote the probability of generating the rooted binary tree obtained by subdividing edge e of T. Let $\mathbb{P}[T] = \sum_e \mathbb{P}[T, e]$ which is the probability of generating a rooted binary tree which produces T when the root is suppressed (this provides a probability distribution on unrooted binary phylogenetic trees). Finally, let

$$\mathbb{P}[e \mid T] = \frac{\mathbb{P}[T, e]}{\mathbb{P}[T]}. \tag{8}$$

Note that $\mathbb{P}[e \mid T]$ is the probability that edge e is the root edge of T, given that T is the unrooted tree obtained by suppressing the root.

For example, consider a labeled unrooted tree on four leaves (Fig. 3). The probability of this tree ($\mathbb{P}[T]$) is 1/3. For the internal edge, the probability of the corresponding labeled rooted tree is 1/9, thus the conditional probability ($\mathbb{P}[e \mid T]$) for the internal edge is 1/3. For the pendant edges the probability of the corresponding labeled rooted tree is 1/18, thus the conditional probability ($\mathbb{P}[e \mid T]$) for each pendant edge is 1/6.

Given an unrooted binary tree T, the *method of maximum likelihood* selects as its estimate of the root edge any edge e that maximizes $\mathbb{P}[e \mid T]$. We let $E_{\max}(T)$ denote the set of edges of T that maximize $\mathbb{P}[e \mid T]$, and we let $e_{\max}(T)$ denote any edge in $E_{\max}(T)$. It is possible, for example when symmetry is present, that $|E_{\max}(T)| > 1$, but we will show below that, for the Yule model, $|E_{\max}(T)| \leqslant 3$.

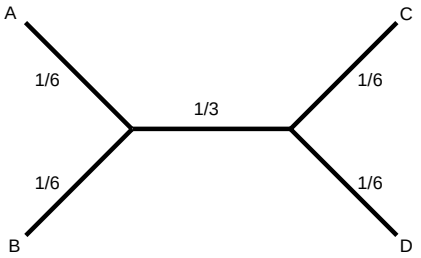

Fig. 3. Conditional probabilities ($\mathbb{P}[e \mid T]$) for the edges of a labeled unrooted tree on four leaves.

5.2. Probability of locating the root edge

Suppose we generate a rooted binary tree T' on n leaves according to the Yule distribution, and we let $u(T')$ denote the unrooted binary tree obtained from T' by suppressing the root. Let $\epsilon(n)$ denote the probability that a particular maximum likelihood edge ($e_{\max}$) of $u(T')$ is the root edge of $u(T')$. By the law of total probability,

$$\epsilon(n) = \sum_T \mathbb{P}[e_{\max} \text{ is the root edge of } T \mid u(T') = T]\mathbb{P}[u(T') = T],$$

where the summation is over all unrooted binary trees on the set of n species and, hence,

$$\epsilon(n) = \sum_T \mathbb{P}[e_{\max}(T) \mid T]\mathbb{P}[T]. \tag{9}$$

One might expect that $\epsilon(n)$ would converge to 0 as n tends to infinity, since the number of edges (and so possible root edges) grows without bound. Indeed we will show that $\mathbb{P}[e_{\max}(T) \mid T]$ can converge to 0 for certain ('caterpillar') trees as the number of leaves grows.

However we will show that $\epsilon(n)$ has a non-zero limit. This parallels similar non-zero asymptotic behaviour for an analogous model, the Yule–Furry model, in which edges are added at random to vertices [32]. Furthermore, although the limit $\epsilon(n)$ is small (about 0.15) the fact that it is non-zero suggests that one should be able to locate the root edge to within a small (edge) distance of $e_{\max}(T)$ with high probability, and this is confirmed by simulations.

For n small, $\epsilon(n)$ can be explicitly calculated, but for larger values $\epsilon(n)$ was approximated by simulation. Simulated values were calculated by the formula

$$\epsilon(n) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}[e_{\max}(T_i) \mid T_i] = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{P}[T_i, e_{\max}(T_i)]}{\mathbb{P}[T_i]}, \tag{10}$$

where T_i is a labeled unrooted binary tree on n leaves obtained by generating a rooted tree according to the Yule process, then unrooting it, and where N is the number of trees generated.

The simulation results suggest that $\lim_{n\to\infty} \epsilon(n) \approx 0.15$ (Fig. 4(a)). The five edges with the largest conditional probabilities for a tree were always an internal edge and the four edges adjacent to it. Let $\epsilon_5(n)$ denote the mean value for the sum of the five largest conditional probabilities for a tree. The simulations suggest that $\lim_{n\to\infty} \epsilon_5(n) \approx 0.58$ (Fig. 4(b)). Thus, even for a large unrooted tree, the location of the root may be narrowed down to a small cluster of five edges, of which one is more likely than not to be the true root. Progressively extending the radius further it appears from simulations that the limiting expected probability that the root edge is within a given (edge) distance d from $e_{\max}(T)$ continues to increase towards 1. For example, when $d = 3$ the limiting probability appears to be close to 0.9.

5.3. Exact asymptotic value of $\epsilon(n)$

In order to calculate the exact limiting value of $\epsilon(n)$ we need some preliminary results. Given an edge e of an unrooted phylogenetic tree, let $H(e)$ denote the number of labeled histories associated to the rooted tree that arises from T by subdividing edge e.

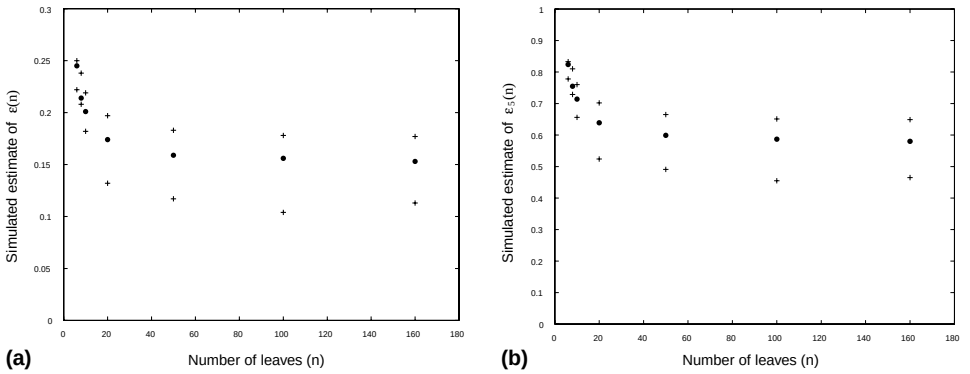

Fig. 4. Simulation results for the conditional probability of edges. Two hundred unrooted trees were randomly generated for different values of n. The trees were produced by unrooting the rooted tree generated by a Yule process. The minimum and maximum probabilities for each simulation are represented by crosses (+). (a) Estimate of the mean probability that $e_{\max}(T)$ contains the true root. (b) Estimate of the mean value for the sum of the five largest conditional probabilities for a tree.

Lemma 4. *Let edge e be an internal edge of an unrooted binary phylogenetic tree T. Denote the four subtrees of T adjacent to e by A,B,C,D, and let a,b,c,d respectively denote the number of leaves in these trees* (*Fig.* 5(b)). *Then* $H(e) \geqslant H(e')$ *for each of the four edges* e' *incident with e precisely if both the following two inequalities hold*:

$$a + b \geqslant \max\{c, d\}; \quad c + d \geqslant \max\{a, b\}. \tag{11}$$

Furthermore, $H(e) > H(e')$ *for all* e' *precisely if these two inequalities hold as strict inequalities.*

Proof. Without loss of generality we may represent e and e' as in Fig. 5(a). From Lemma 1 we have

$$H(e) = \frac{(n-1)!}{(n-1)(c+d-1)\prod_{v\in\mathring{C}}\delta(v)\prod_{v\in\mathring{D}}\delta(v)\prod_{v\in\mathring{F}}\delta(v)}. \tag{12}$$

If the tree is rooted at the adjacent edge e' the number of histories is

$$H(e') = \frac{(n-1)!}{(n-1)(f+d-1)\prod_{v\in\mathring{C}}\delta(v)\prod_{v\in\mathring{D}}\delta(v)\prod_{v\in\mathring{F}}\delta(v)}. \tag{13}$$

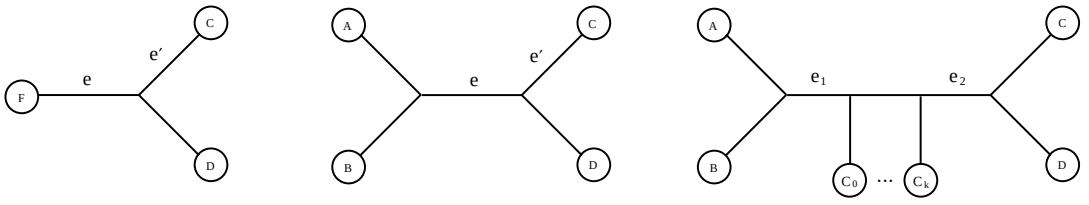

Fig. 5. Generic unrooted binary trees with subtrees A, B, C, D, F with a, b, c, d, f leaves, respectively: (a) with three distinguished edges; (b) with four distinguished edges; (c) a hypothetical tree with two edges e_1, e_2 in $E_{\max}(T)$ that are separated by $k \geqslant 1$ edges. $C_0, \ldots, C_k$ denotes subtrees with $c_0, \ldots, c_k$ leaves, respectively.

Therefore, $H(e) \geqslant H(e')$ precisely if

$$f \geqslant c. \tag{14}$$

Now let the tree F be split into two subtrees A and B (Fig. 5(b)). Applying (14) to the edge labeled e, and then labeling in turn each adjacent edge as e', leads to the two 4-branch inequalities. If both of the 4-branch inequalities are strict, then $H(e)$ is strictly larger than $H(e')$. □

Lemma 5. *Any two edges in $E_{\max}(T)$ are adjacent.*

Proof. We will derive a contradiction by supposing that there exists two non-adjacent edges e_1, e_2 in $E_{\max}(T)$. Under this assumption we can represent T as in Fig. 5(c), where $k \geqslant 1$, and $c_0, c_1, \ldots, c_k$ are all positive. For edge e_1 to be in $E_{\max}(T)$ we must have, from Lemma 4,

$$a + b \geqslant c + d + c_1 + \cdots + c_k. \tag{15}$$

Likewise for edge e_2 we must have

$$c + d \geqslant a + b + c_0 + \cdots + c_{k-1}. \tag{16}$$

Adding (15) and (16) we get

$$a + b + c + d \geqslant a + b + c + d + c_0 + 2(c_1 + \cdots c_{k-1}) + c_k. \tag{17}$$

This implies that $c_0 = c_1 = \cdots = c_k = 0$, contradicting our original supposition, thus any two edges in $E_{\max}(T)$ must be adjacent. □

As we are dealing with binary trees we have the following straightforward consequence of Lemma 5.

Corollary 1. $|E_{\max}(T)| \leqslant 3$. *Furthermore, if both the inequalities in* (11) *are strict, then* $|E_{\max}(T)| = 1$.

We now calculate the exact asymptotic value of $\epsilon(n)$. We do this by embedding the discrete process of rooting a tree into a continuous analogue involving 'stick breaking'. The asymptotic properties of this process have been previously analysed in Ref. [33]; here we are concerned only with comparisons involving the first two breaks of the stick.

Theorem 3. *Generate a rooted binary tree with n leaves randomly under the Yule model, and let T denote the tree obtained by suppressing the root. The probability that edge $e_{\max}(T)$ is unique and equal to the root edge of T convergeces to the value* $4\ln(4/3) - 1$ (≈ 0.15) *as* $n \to \infty$.

Proof. Let us generate a rooted binary tree T' on n leaves under the Yule model, and let t_1, t_2 denote the two rooted subtrees of T' incident with the root. Let n_i ($i = 1, 2$) denote the number of leaves in t_i and let n_{i1}, n_{i2} denote the number of leaves in the two subtrees of t_i incident with its root. Thus, $n_{i1} + n_{i2} = n_i$. Let T denote the associated unrooted tree obtained from T' by suppressing the root of T'. By Corollary 1 the probability that the edge $e_{\max}(T)$ is unique and equals the root edge of T is the probability that

$$n_1 > \max\{n_{21}, n_{22}\} \tag{18}$$

and

$$n_2 > \max\{n_{11}, n_{12}\}. \tag{19}$$

Now, by (3), n_1 is uniformly distributed between 1 and $n-1$. Furthermore, conditional on n_i, n_{i1} is also uniformly distributed between 1 and n_i (by the hereditary property of the Yule model described at the end of Section 3). Note that if $n_1 \leqslant n_2$, then inequality (19) is trivially satisfied, while if $n_2 \leqslant n_1$ inequality (18) is satisfied. Thus, we can determine the asymptotic probability that inequalities (18) and (19) hold by the following 'stick-breaking' process.

Take a unit interval, and break it randomly and uniformly at some point along its length. Let X be a random variable representing the length of the shorter section. Take the longer section from this first split and uniformly randomly break it again. Let Y be a random variable representing the length of the longer section for this second split. In the present setting X will represent $\min\{n_1, n_2\}$ and Y will represent the term on the right-hand side of inequality (18) (if $n_1 \leqslant n_2$) or inequality (19) (if $n_2 \leqslant n_1$). Thus, the probability we wish to calculate in the asymptotic limit is that $\mathbb{P}(Y < X)$. We have

$$\mathbb{P}(Y < X) = \int_0^1 \mathbb{P}(Y < X \mid X = x) f(x)\, \mathrm{d}x, \tag{20}$$

where $f(x)$ is the density function for X. Since X is distributed uniformly on $[0, \frac{1}{2}]$ we have

$$f(x) = \begin{cases} 2 & \text{if } 0 \leqslant x \leqslant \frac{1}{2}, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

If $X < 1/3$, then $Y > 1/3$ and so the probability that $Y < X$ is zero. If $x \geqslant 1/3$, then conditional on $X = x$, Y is distributed uniformly on the interval $[(1/2)(1-x), 1-x]$. Thus, if $g_x(y)$ is the density function for Y conditional on the event $X = x$, then

$$g_x(y) = \begin{cases} \frac{2}{1-x} & \text{if } x \in [0, \frac{1}{2}],\ y \in [\frac{1}{2}(1-x), 1-x], \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

Consequently,

$$\mathbb{P}(Y < X \mid X = x) = \begin{cases} 0 & \text{if } x < \frac{1}{3}, \\ \int_{(1/2)(1-x)}^{x} g_x(y)\, \mathrm{d}y = -3 + \frac{2}{1-x} & \text{if } x \geqslant \frac{1}{3}. \end{cases} \tag{23}$$

Substituting (21) and (23) back into (20) we obtain

$$\mathbb{P}(Y < X) = 2 \int_{1/3}^{1/2} \left[-3 + \frac{2}{1-x} \right] \mathrm{d}x = 4 \ln\left(\frac{4}{3}\right) - 1,$$

which completes the proof. □

5.4. A family of trees for which $\mathbb{P}[e_{\max}(T) \mid T] \to 0$

A *caterpillar tree* is any unrooted binary phylogenetic tree that reduces to a path (a tree having vertices of degree 1 or 2) once the pendant edges and leaves are deleted. The simulation results

suggest that caterpillar trees are the trees for which $\mathbb{P}[e_{\max}(T) \mid T]$ is smallest. For the caterpillar tree $\mathbb{P}[e_{\max}(T) \mid T]$ may be calculated exactly, and we show that asymptotically this probability converges to 0.

Theorem 4. *Let C_n denote a caterpillar tree on n leaves. Then,*

$$\mathbb{P}[e_{\max}(C_n) \mid C_n] = \begin{cases} \frac{8}{3} \dfrac{(n-2)!}{2^n((n-1)/2)!((n-3)/2)!}, & n \text{ odd}, \\ \frac{8}{3} \dfrac{(n-2)!}{2^n\{((n-2)/2)!\}^2}, & n \text{ even}. \end{cases} \tag{24}$$

Asymptotically, as $n \to \infty$,

$$\mathbb{P}[e_{\max}(C_n) \mid C_n] \sim \frac{2}{3}\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}} \to 0. \tag{25}$$

Proof. For the tree C_n the determination of $E_{\max}(C_n)$ may easily be found using the 4-branch inequalities of Lemma 4. For n odd there are two edges in $E_{\max}(C_n)$, located at the two edges where there are $(n-1)/2$ leaves on one side and $(n+1)/2$ leaves on the other side. For n even there is a single edge in $E_{\max}(C_n)$ located at the edge where there is $n/2$ leaves on each side. The probability that $e_{\max}(T)$ is the root edge of C_n is, in either case,

$$\mathbb{P}[e_{\max}(C_n) \mid C_n] = \frac{\mathbb{P}[C_n, e_{\max}(C_n)]}{\mathbb{P}[C_n]}. \tag{26}$$

Consider, firstly, the numerator of this equation. From Proposition 1,

$$\mathbb{P}[C_n, e_{\max}(C_n)] = \begin{cases} \dfrac{2^{n-1}}{n!} \dfrac{1}{(n-1)((n-1)/2)!((n-3)/2)!}, & n \text{ odd}, \\ \dfrac{2^{n-1}}{n!} \dfrac{1}{(n-1)\{((n-2)/2)!\}^2}, & n \text{ even}. \end{cases} \tag{27}$$

Now consider the denominator of (26). We may calculate $\mathbb{P}[C_n]$ by summing $\mathbb{P}[C_n, e]$ over all edges e of C_n (and using Proposition 1, together with a binomial identity). Alternatively, we can compute $\mathbb{P}[C_n]$ by first computing the probability of generating a tree having the caterpillar shape c_n on n leaves – this satisfies the recursion

$$\mathbb{P}[c_n] = \frac{4}{n-1}\mathbb{P}[c_{n-1}], \quad \text{where } \mathbb{P}[c_5] = 1, \tag{28}$$

since at each stage there are four possible edges that the next leaf can be attached to. We then need to divide $\mathbb{P}[c_n]$ by the number of phylogenetic trees on n leaves having shape c_n, and this number is $n!/8$. Using either approach to compute $\mathbb{P}[C_n]$ we obtain

$$\mathbb{P}[C_n] = \frac{3 \times 4^{n-2}}{n!(n-1)!}. \tag{29}$$

Combining the numerator and denominator terms gives (24). The second part of the theorem follows from the asymptotic equation $(1/2^n)\binom{n}{n/2} \sim \sqrt{2/(\pi n)}$. □

6. An extension of the Yule model

In the Yule model, at any time each existing species has the same probability of giving rise to a new species, and all lineages are treated exchangeably. Here we consider a simple modification of this model, in which the rate of speciation events on a given lineage is a function of the time back to the last speciation event on that lineage.

More precisely, we suppose that at time $t = 0$ there is just one species, labeled s_0, subject to a 2-state Markov process on state space $\{1, 2\}$. Under this process, s_0 is initially in state 1, and state 2 corresponds to a 'speciation event', that is, the replacement of the original species by two species (either two new species, or the original species plus one new one, and we will not distinguish here between these two possibilities). Let $s(t)$ denote the rate of change from state 1 to state 2 at time t, we call this the speciation rate. Once a speciation event occurs (say at time Λ) the two species are again assumed to be independently subject to the same Markov process, with time reset to 0 (that is, with rate function $s(t - \Lambda)$). Continuing in this way, we obtain a probability distribution on the trees of descent of species starting from s_0 up to some fixed time t which we can assume (by re-scaling s if necessary) lies in the range $[0, 1]$.

The biological motivation for this model is that a recently evolved species, or the species that it has split off from, are often colonising new regions or niches, and so may be more likely to give rise to further new species (in a given short time period) than a species that has existed for a very long time without giving rise to any new species (thus we are thinking of s being a monotone decreasing function). It would also be interesting and useful to build extinctions into such a model, however we do not pursue this here.

Kubo and Iwasa [6] consider a rate-varying model of speciation, however in their case, the speciation rate is a function of (absolute) time, rather than the lineage-specific time back to the last speciation event. Our model has more similarity to that discussed by Heard [5] who used computer simulation rather than analytical techniques in his analysis. Our general approach, which encompasses more than one model in a single analytical framework, is akin to that taken by Aldous [9]. We are interested in the probability distribution that this model induces on the tree that describes the species descendent from s_0. Since we are only interested in the 'shape' of these speciation trees, we will mostly deal with trees in which the vertices are unlabeled.

6.1. Terminology

In this section the following additional terminology for rooted trees is necessary.

- For $n \geqslant 1$, let $UB(n)$ denote the (finite) set of *u*nlabeled *b*inary trees consisting of n leaves together with an additional leaf, the *root leaf* s_0 (where the root leaf is the top-most vertex), and whose remaining internal vertices are all of degree 3 (Fig. 6(a)).
- For $n \geqslant 2$, let $EUB(n)$ denote the (finite) set of *e*dge-rooted *u*nlabeled *b*inary trees obtained from $UB(n)$ by deleting from each tree the root leaf and its incident edge. If $\tau \in UB(n)$, then we will let τ^* denote the associated tree in $EUB(n)$ (Fig. 6(b)).
- For $\tau \in UB(n)$, let $L(\tau)$ be the set of distinct trees that can be obtained by assigning the (species) labels $\{1, \ldots, n\}$ bijectively to the n non-root leaves of τ. Let $LB(n) := \bigcup_{\tau \in UB(n)} L(\tau)$, the set of *l*abeled *b*inary trees (Fig. 6(c)).

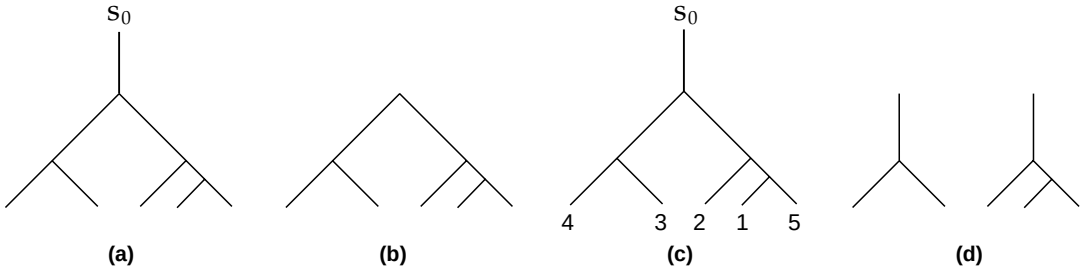

Fig. 6. Rooted tree types. (a) An unlabeled binary tree on five leaves; $\tau \in UB(5)$. The root vertex is labeled s_0. (b) The edge-rooted unlabeled binary tree on five leaves obtained by removing the root leaf and its incident edge; $\tau^* \in EUB(5)$. (c) A labeled binary tree on five leaves; $\tau \in LB(5)$. (d) The two subtrees, τ^1 and τ^2 of the tree τ in (a).

6.2. Formulae

For the model described above, the speciation tree at time $t \in [0,1]$, $T(t)$, is the unlabeled tree of descent of the species that have evolved up to time t from the root leaf s_0. For $0 \leqslant t \leq 1$ and $\tau \in UB(n)$, consider the following (absolute and conditional) probabilities

$$f(\tau,t) := \mathbb{P}[T(t) = \tau]; \quad p(\tau) := \mathbb{P}[T(1) = \tau \mid T(1) \text{ has } n \text{ leaves}]. \tag{30}$$

Let $\Lambda(s_0)$ denote the time until speciation of s_0, and set

$$S(x) := \mathbb{P}[\Lambda(s_0) \geqslant x]; \quad \sigma(x) := s(x)S(x), \tag{31}$$

where $s(x)$ is, as previously, the speciation rate at moment x.

Since the speciation of s_0 is a time-dependent Poisson process we have, from Ref. [34],

$$\mathbb{P}[\Lambda(s_0) \geqslant x] = \exp\left[-\int_0^x s(\lambda)\,\mathrm{d}\lambda\right]. \tag{32}$$

Thus, $\sigma(x) = \lim_{\delta \to 0+}(\mathbb{P}[\Lambda(s_0) \in (x, x+\delta)])/\delta$ and so, by the assumptions that define the model, we have the following fundamental recursion:

$$f(\tau,t) = 2^{\delta(\tau)} \int_0^t f(\tau^1, t-x) f(\tau^2, t-x)\sigma(x)\,\mathrm{d}x, \tag{33}$$

where τ^1 and τ^2 denote the two subtrees of τ whose two vertex sets (i) intersect precisely on v and (ii) cover all vertices of τ except s_0 (Fig. 6(d)) and where

$$\delta(\tau) = \begin{cases} 1 & \text{if } \tau^1 \neq \tau^2, \\ 0 & \text{otherwise.} \end{cases} \tag{34}$$

Let $N(t)$ denote the total number of species existing at time $t \in [0,1]$, and let

$$v(k,t) := \mathbb{P}[N(t) = k].$$

For $\tau \in UB(n)$ we wish to calculate the conditional probability

$$p(\tau) = \mathbb{P}[T(1) = \tau \mid N(1) = n] = \frac{f(\tau,1)}{v(n,1)}. \tag{35}$$

The number $v(n,1)$ appearing in Eq. (35) is given by

$$v(n,1) = \sum_{\tau \in UB(n)} f(\tau, 1).$$

However, the number of terms in this summation grows exponentially with n. Thus, we also give a simple recursion for computing the functions $v(1,t), \ldots, v(k,t)$ and thereby the number $v(n,1)$, as follows:

$$v(1,t) = S(t)$$

$$v(k,t) = \sum_{i=1}^{k-1} \int_0^t v(i, t-x) v(k-i, t-x) \sigma(x)\, \mathrm{d}x.$$

We may also wish to compute the probability of the induced edge-rooted tree. Thus, given $\tau \in UB(n)$ and its associated tree $\tau^* \in EUB(n)$. Let

$$p(\tau^*) := \lim_{\epsilon \to 0+} \mathbb{P}[T(1) = \tau \mid N(1) = n;\ \Lambda(s_0) < \epsilon]. \tag{36}$$

The motivation for considering $p(\tau^*)$ is that one is frequently interested in the distribution on edge-rooted trees, and we can simplify matters by supposing that the first speciation event happened at time 0. We have the recursion

$$p(\tau^*) = 2^{\delta(\tau)} p(\tau^1) p(\tau^2) \tag{37}$$

with τ^1, τ^2 as in Eq. (33).

Note that if we wish to compare the probability ratios of two trees, then we can dispense with the function v altogether, since $p(\tau)/p(\tau') = f(\tau,1)/f(\tau',1)$. For $i \in \{1,2,3\}$, there is just one tree in $UB(i)$ so we write $UB(i) = \{\tau_i\}$. We have

$$f(\tau_1, t) = S(t),$$

$$f(\tau_2, t) = \int_0^t S(t-x)^2 \sigma(x)\, \mathrm{d}x,$$

$$f(\tau_3, t) = 2 \int_0^t S(t-x) f(\tau_2, t-x) \sigma(x)\, \mathrm{d}x.$$

For the (only) two trees $\tau_{1,3}$ and $\tau_{2,2}$ in $UB(4)$, as shown in Fig. 7, we have

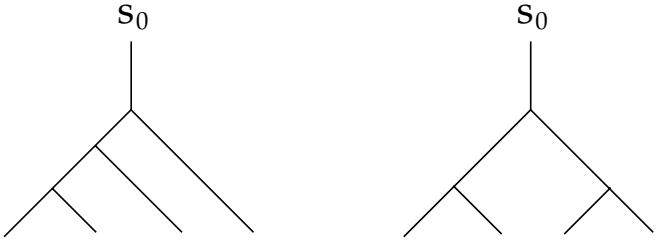

Fig. 7. The two tree shapes on four leaves ($\tau_{1,3}$ and $\tau_{2,2}$).

$$f(\tau_{1,3},t)=2\int_0^t S(t-x)f(\tau_3,t-x)\sigma(x)\,\mathrm{d}x$$

and

$$f(\tau_{2,2},t)=\int_0^t f(\tau_2,t-x)^2\sigma(x)\,\mathrm{d}x.$$

Thus we can obtain an explicit expression for the ratio of the probabilities of $\tau_{2,2}$ and $\tau_{1,3}$, and even simpler formulae for corresponding rooted trees, by a further application of Eqs. (33) and (37). This is summarized in Theorem 5.

Theorem 5. *Under the rate varying speciation model the tree shapes on four leaves satisfy*

$$\frac{p(\tau_{2,2})}{p(\tau_{1,3})}=\frac{\int_0^1\{\int_0^{1-x}S(1-x-s)^2\sigma(s)\,\mathrm{d}s\}^2\sigma(x)\,\mathrm{d}x}{4\int_0^1 S(1-x)\sigma(x)\{\int_0^{1-x}S(1-x-s)\sigma(s)\{\int_0^{1-x-s}S(1-x-s-r)^2\sigma(r)\,\mathrm{d}r\}\,\mathrm{d}s\}\,\mathrm{d}x},$$

$$\frac{p(\tau^*_{2,2})}{p(\tau^*_{1,3})}=\frac{\{\int_0^1 S(1-x)^2\sigma(x)\,\mathrm{d}x\}^2}{4S(1)\int_0^1 S(1-x)\sigma(x)\{\int_0^{1-x}S(1-x-r)^2\sigma(r)\,\mathrm{d}r\}\,\mathrm{d}x}.$$

6.3. Two classes of models

1. The simplest model has $s(x)=s>0$, constant. This gives the Yule model as described in Section 3. In this case, $\sigma(x)=s\mathrm{e}^{-sx}$ and $N(t)$ models a pure birth process, so $v(k,t)=\mathrm{e}^{-st}(1-\mathrm{e}^{-st})^{k-1}$. Under this model, $p(\tau_{2,2})=p(\tau^*_{2,2})=1/3$, and, more generally, by Proposition 1, $p(\tau)=p(\tau^*)=2^{u(\tau)}\prod_{i>2}(i-1)^{-d_i(\tau)}$, where $d_i(\tau)$ denotes the number of internal vertices of τ which have exactly i descendant leaves, and $u(\tau)$ is the number of *unbalanced* internal vertices of τ – that is, internal vertices for which the two descendant subtrees are not identical.

2. The models of a second class are those which satisfy the condition

$$s(x)=0 \quad \text{for } x>\epsilon,$$

which we will call 'explosive radiation' models. In these models, unless a species has undergone a speciation event within the last ϵ time interval, it will never do so. Thus, in these models, speciation events would tend to be clustered close together. We now analyse this model, and show that, provided epsilon is sufficiently small, then this model is precisely that induced by a *uniform distribution* on leaf-labeled trees.

Under the uniform model, on rooted trees, a tree is selected uniformly from $LB(n)$, and then it is viewed as an unlabeled tree $\tau\in UB(n)$. The probability of the tree shape τ is calculated as $p_{\mathrm{unif}}(\tau)=|L(\tau)|/|LB(n)|$, where $L(\tau)=\{T\in LB(n):T \text{ is a leaf-labeling of } \tau\}$. Fortunately, the numerator and denominator of this ratio can both be evaluated exactly, and so we get an explicit formula for $p_{\mathrm{unif}}(\tau)$ as follows. We have $|L(\tau)|=n!2^{-b(\tau)}$, where $b(\tau)$ is the number of *balanced* internal vertices of τ – that is, internal vertices for which the two descendant subtrees are identical. Now, from Ref. [35],

$$|LB(n)| = (2n-3)!! = (2n-3) \times (2n-5) \times \cdots \times 3 \times 1 = \frac{(2n-2)!}{(n-1)!2^{n-1}}. \tag{38}$$

Therefore, under the uniform model on rooted trees,

$$p_{\text{unif}}(\tau) = \frac{|L(\tau)|}{|LB(n)|} = \binom{2n-2}{n-1}^{-1} 2^{u(\tau)}, \tag{39}$$

where $u(\tau)$ is the number of *unbalanced* internal vertices (and so $b(\tau) + u(\tau) = n-1$).

Theorem 6. *Under an explosive radiation model, with $\epsilon < 1/n$, the probability distribution on trees is precisely that induced by the uniform model. That is,*

$$p(\tau) = p_{\text{unif}}(\tau) \quad \forall \tau \in UB(n).$$

Proof. We use induction on n to establish the following:

$$\text{CLAIM : if } \tau \in UB(n), \quad \text{then } f(\tau, t) = c(n)2^{u(\tau)} \quad \text{for } t > n\epsilon,$$

where $c(n) = \mathrm{e}^{-n\int_0^\epsilon s(\lambda)\,\mathrm{d}\lambda}\left(1 - \mathrm{e}^{-\int_0^\epsilon s(\lambda)\,\mathrm{d}\lambda}\right)^{n-1}$.

The claim clearly holds for $n = 1$, since in this case, if $t > \epsilon$,

$$f(\tau, t) = S(t) = \mathrm{e}^{-\int_0^t s(\lambda)\,\mathrm{d}\lambda} = \mathrm{e}^{-\int_0^\epsilon s(\lambda)\,\mathrm{d}\lambda}.$$

Now suppose the result holds for $n = k \geqslant 1$, and let $\tau \in UB(k+1)$. Then, from Eq. (33) and the fact that $s(x)$ is 0 for $x > \epsilon$, we have, for $t > \epsilon$,

$$f(\tau, t) = 2^{\delta(\tau)} \int_0^\epsilon f(\tau^1, t-x) f(\tau^2, t-x)\sigma(x)\,\mathrm{d}x.$$

For $i = 1, 2$, let k_i denote the number of leaves of τ^i (thus, $k_1 + k_2 = k+1$). If $t > (k+1)\epsilon$, and $x < \epsilon$, we have $t - x > k\epsilon \geqslant k_i\epsilon$ (since $k_1, k_2 \leqslant k$). Thus we may apply the induction hypothesis to $f(\tau^1, t-x)$ and $f(\tau^2, t-x)$ over the range of integration and deduce that, for $t > (k+1)\epsilon$,

$$f(\tau, t) = 2^{\delta(\tau)} \int_0^\epsilon c(k_1)2^{u(\tau^1)} c(k_2)2^{u(\tau^2)}\sigma(x)\,\mathrm{d}x = 2^{u(\tau)} c(k_1)c(k_2) \int_0^\epsilon \sigma(x)\,\mathrm{d}x = 2^{u(\tau)} c(k+1)$$

by the definition of the function c.

By Eq. (35), $p(\tau) = f(\tau, 1)/v(n, 1)$ and therefore, since $\epsilon < 1/n$, we can apply the above claim to deduce that $p(\tau) = c^*(n)2^{u(\tau)}$ for a function c^* that depends only on n and perhaps the function s. However, it is easy to show that c^* does not depend on s at all, and that it must equal $c^\dagger(n) := n\binom{2n-2}{n-1}^{-1}$, since we have, from (39)

$$c^\dagger(n) \sum_{\tau \in UB(n)} 2^{u(\tau)} = \sum_{\tau \in UB(n)} p_{\text{unif}}(\tau) = 1 = \sum_{\tau \in UB(n)} p(\tau) = c^*(n) \sum_{\tau \in UB(n)} 2^{u(\tau)}, \tag{40}$$

and thus $c^\dagger(n) = c^*(n) = 1/\sum_{\tau \in UB(n)} 2^{u(\tau)}$. Hence, $p(\tau) = p_{\text{unif}}(\tau)$, as claimed. □

Acknowledgements

We thank Charles Semple and the anonymous referees for some helpful comments on an earlier version of this paper.

References

[1] D. Aldous, Probability distributions on cladograms, in: D. Aldous, R. Permantle (Eds.), Random Structures, vol. 76, Springer, Berlin, 1996, p. 1.
[2] J. Brown, Probabilities of evolutionary trees, Syst. Biol. 43 (1) (1994) 78.
[3] E. Harding, The probabilities of rooted tree-shapes generated by random bifurcation, Adv. Appl. Prob. 3 (1971) 44.
[4] S. Heard, Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees, Evolution 46 (6) (1992) 1818.
[5] S. Heard, Patterns in phylogenetic tree balance with variable and evolving speciation rates, Evolution 50 (6) (1996) 2141.
[6] T. Kubo, Y. Iwasa, Inferring the rates of branching and extinction from molecular phylogenies, Evolution 49 (1995) 694.
[7] A. Mooers, S. Heard, Inferring evolutionary process from phylogenetic tree shape, Quart. Rev. Biol. 72 (1) (1997) 31.
[8] A. McKenzie, M. Steel, Distributions of cherries for two models of trees, Math. Biosci. 164 (1) (2000) 81.
[9] D. Aldous, Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today [online], Available: http://www.stat.berkeley.edu/users/aldous/ bibliog.html[Accessed: August 25], (2000).
[10] B. Rannala, Z. Yang, Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference, Molecular Evolution 43 (1996) 304.
[11] B. Mau, M. Newton, B. Larget, Bayesian phylogenetic inference via Markov chain Monte Carlo methods, Biometrics 55 (1999) 1.
[12] S. Li, D.K. Pearl, H. Doss, Phylogenetic tree construction using Markov chain Monte Carlo, J. Am. Statist. Assoc. 95 (450) (2000) 493.
[13] J. Slowinski, Probabilities of n-trees under two models: a demonstration that asymmetrical interior nodes are not improbable, Syst. Zool. 39 (1) (1990) 89.
[14] R. Page, E. Holmes, Molecular Evolution: A Phylogenetic Approach, Blackwell Science, Oxford, 1998, p. 11 ((Chapter 2)).
[15] J. Slowinski, C. Guyer, Testing the stochasticity of patterns of organismal diversity: an improved null model, Am. Nat. 134 (6) (1989) 907.
[16] S. Nee, R. May, P. Harvey, The reconstructed evolutionary process, Philos. Trans. R. Soc. London B 344 (1994) 305.
[17] J. Losos, F. Adler, Stumped by trees? A generalized null model for patterns of organismal diversity, Am. Nat. 145 (3) (1995) 329.
[18] J. Kingman, On the genealogy of large populations, J. Appl. Prob. 19A (1982) 27.
[19] F. Tajima, Evolutionary relationships of DNA sequences in finite populations, Genetics 105 (1983) 437.
[20] D. Knuth, The Art of Computer Programming, vol. 3, Addison-Wesley, Reading, MA, 1997, p. 67 (Chapter 5, Problem 20).
[21] R. Stanley, Enumerative Combinatorics: Cambridge Studies in Advanced Mathematics, vol. 1, no. 49, Cambridge University, Cambridge, 1997, p. 312.
[22] A. Edwards, Estimation of the branch points of a branching diffusion process, J. R. Stat. Soc. Ser. B 32 (1970) 155.
[23] M. Sanderson, How many taxa must be sampled to identify the the root node of a large clade?, Syst. Biol. 45 (2) (1996) 168.
[24] W. Lynch, More combinatorial properties of certain trees, Comput. J. 71 (1965) 299.

[25] H. Mahmoud, Evolution of Random Search Trees, Wiley, New York, 1992, p. 72.
[26] R. Stanley, Enumerative Combinatorics, in: Cambridge Studies in Advanced Mathematics, vol. 1, no. 49, Cambridge University, Cambridge, 1997, p. 18.
[27] M. Ridley, Evolution, Evolution, Blackwell Science, Cambridge, MA, USA, 1993, p. 460 ((Chapter 17)).
[28] A. Smith, Rooting molecular trees: problems and strategies, Biol. J. Linnean Soc. 51 (1994) 279.
[29] J. Farris, Estimating phylogenetic trees from distance matrices, Am. Nat. 106 (1972) 646.
[30] D.L. Swofford, G.J. Olsen, P.J. Waddell, D.M. Hillis, Phylogenetic inference, in: D.M. Hillis, C. Moritz, B.K. Mable (Eds.), Molecular Systmatics, Sinauer Associates Inc., Sunderland, MA, USA, 1996, p. 488.
[31] Y.V. de Peer, R.D. Wacher, Treecon: a software package for the construction and drawing of evolutionary trees, Comput. Applic. Biosci. 9 (1993) 177.
[32] J. Haigh, The recovery of the root of a tree, J. Appl. Prob. 7 (1970) 79.
[33] W. v. Zwet, A proof of Kakutani's conjecture on random subdivision of longest intervals, Annal. Prob. 6 (1) (1978) 133.
[34] J. Medhi, Stochastic Processes, Wiley, New York, 1982, p. 119 ((Chapter 4)).
[35] L. Cavalli-Sforza, A. Edwards, Phylogenetic analysis, Evolution 21 (1967) 550.