The 'shape' of phylogenies under simple random speciation models

Mike Steel and Andy McKenzie

Biomathematics Research Centre, University of Canterbury,
Christchurch, New Zealand

Abstract. We describe 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.

Here we describe new results for four 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. Secondly, for trees generated under the Yule and uniform models, we describe the induced distribution they generate on the number C_n of *cherries* in the tree, where a cherry is a pair of leaves each of which is adjacent to a common ancestor. 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.

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 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 [1–8]. These models make predictions about the shape of the phylogenetic tree connecting the extant species, and they are also used as a basis for calculating the probability of certain configurations under random speciation [9]. These probabilities may then be useful in testing hypotheses concerning the speciation process. Speciation models can also provide prior probabilities for phylogenetic trees in Bayesian approaches to tree reconstruction [10–12].

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

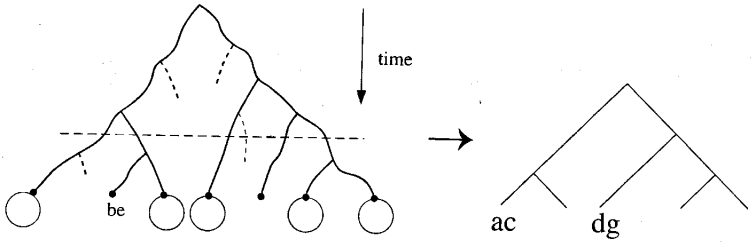


Fig. 1. A hypothetical sequence of speciation and extinction events is illustrated on the left. The descending (dashed) lineages become extinct, while seven species survive to the present. At the time corresponding to the horizontal dashed line there are six extant species. The circled present-day species (a, c, d, g, f) are those that are selected by the investigator for study. The induced phylogenetic tree on these five selected species is shown at right.

underlying parameters) and, secondly, we are interested in isolating out the information that is conveyed solely by the discrete tree shape. One may also make a random selection of species from the set of extant species, to obtain the (induced) phylogenetic tree (as illustrated in Fig. 1) and for certain models this does not alter the induced distribution on (discrete) tree shapes.

We will initially concentrate on some properties of the Yule model, which is perhaps the simplest stochastic model for speciation. In this model whenever a speciation event occurs, each of the species existing at that moment is equally likely to give rise to the new species. However the rate of speciation (and extinction) may vary arbitrarily with time. For example, in Fig. 1 each of the six species that exists at the time indicated by the horizontal dashed line has an equal rate of speciation at that moment.

Despite the generality of such a model it leads to a well-defined probability distribution on the (discrete) shapes of phylogenetic trees. In this paper we investigate certain predictions that this (and other) models make regarding some of the properties of phylogenetic trees that can be determined simply by considering their shape.

One of the problems we will consider is how to recover the position of the root of such a tree, when one is only given the associated unrooted tree. This is particularly relevant in phylogenetic analysis, since most tree reconstruction methods return an unrooted tree. Somewhat surprisingly, it is possible to estimate the approximate position of the root fairly accurately even for very large trees generated by the Yule model (a similar result for a related, but different, process was established by [13]).

It is useful to contrast the Yule distribution with (i) the distribution on tree shapes obtained by selecting a phylogenetic tree uniformly at random (the "uniform distribution") and (ii) the shapes of published phylogenetic trees that have been reconstructed from biological data. Several studies (see for example

[14]), have suggested that published trees tend, on average, to be less balanced than the Yule model would predict, yet more balanced than the uniform model would allow. Of course a published tree is only an estimate of the true species tree. Thus, it may also be important to determine biases in tree shape that arise due to particular tree reconstruction methods, and other factors, such as the non-random selection of species by the investigator [8].

Nevertheless, with a view to obtaining less balanced trees than those generated by the Yule model, we will consider a simple modification to this model. In this modified model a species that has recently speciated is more likely to speciate again than one that has existed without undergoing speciation for a long time. Such a model appears to lead to less balanced trees, and we prove that in a sufficiently extreme case it leads precisely to the uniform model (which, incidentally, provides a natural way in which the uniform model may be regarded as a speciation model). It would be interesting, for future work, to evaluate precisely how well these types of models can account for the shapes of published trees, and also to explore the effect of extinctions in such models.

The structure of the paper is as follows. We begin by introducing some basic terminology for phylogenetic trees (Sect. 2). The Yule model is then introduced, and some of its properties are described (Sect. 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 (Sect. 4). Following this we investigate the induced probability distribution for the number of cherries under the Yule and uniform models (Sect. 5). Exact formulae are given for small trees, and the asymptotic distribution is found to be normal. Next, a maximum likelihood approach to edge-rooting an unrooted tree is presented, and it is shown that even for large unrooted trees the approximate location of the root can be identified with high probability (Sect. 6). 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 (Sect. 7). We show that this modified model reduces to the uniform model under the condition of "explosive radiation". The results we describe here are mostly based on [7] for Sect. 5 and [15] for Sects. 4,6,7 where full proofs and details may be found.

2 Terminology

Evolutionary relationships are often represented by rooted or unrooted *binary (phylogenetic) trees* [16]. Such trees consist of labeled *nodes* of degree 1 called *leaves* and unlabelled *internal* nodes of degree 3 (also, in case the tree is rooted, it contains an additional *root node* of degree 2, so that every node can be regarded as having exactly two descendants). A pair of leaves adjacent to a common node is called a *cherry*. 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. 2. The number of *rooted* leaf-labeled binary trees on n leaves

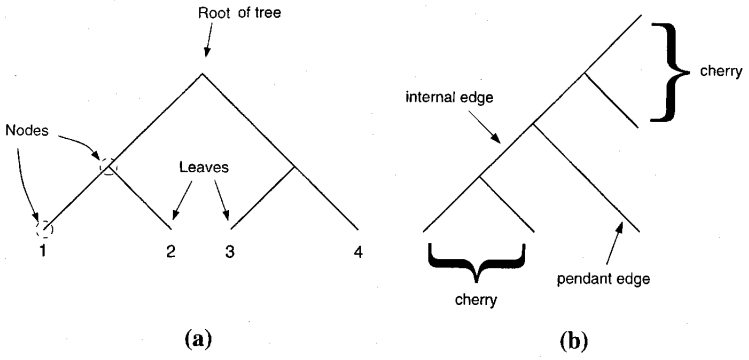


Fig. 2. **(a)** A labeled rooted tree with 4 leaves. **(b)** An unrooted tree shape with 5 leaves

is given by $(2n-3)!! = 1 \times 3 \times 5 \cdots \times (2n-3)$, and the number of *unrooted* leaf-labeled binary trees on n leaves is given by $(2n-5)!! = 1 \times 3 \times 5 \cdots \times (2n-5)$ [17,18].

Throughout 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 $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1$. 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).

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 in the tree. 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 [1]). This distribution on tree shapes is often called the *Yule model* and it has been widely studied [3,19–21].

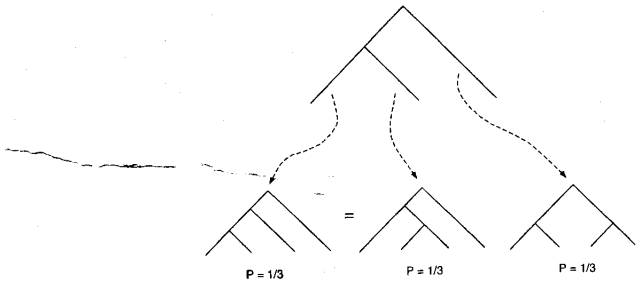


Fig. 3. The Yule model probabilities for shapes with 4 leaves. A shape on 4 leaves is formed by the splitting of one of the pendant edges of the shape on 3 leaves. Each pendant edge has the same probability of splitting, so for the shape on 3 leaves each pendant edge has a probability of 1/3 of splitting. The resulting symmetric shape on 4 leaves has a probability of 1/3. The other two shapes on 4 leaves are the same (up to rotation about internal vertices), and so the probability of this shape is 2/3

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. 3.

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 [1,22,23]. 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.

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.

Then, for a rooted binary phylogenetic tree with n leaves, the number of associated labeled histories is precisely

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

where $\mathring{V}$ denotes the set of internal vertices of the tree. This result is from [2,24]. It is also possible to give an exact expression for the total number of labeled histories on a set of species. From [25] the number of labeled histories on n species is

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

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 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 [1].

4 Depth of a most recent common ancestor (MRCA)

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 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 [26]; 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*). The following summarises results from [15].

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 [27,28]. 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 [29].

Theorem 1. *[27,28] 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^2_n) of this distribution are given by

$$\mu_n = 2\sum_{j=2}^{n} \frac{1}{j}; \qquad \sigma^2_n = 2\sum_{j=2}^{n} \frac{1}{j} - 4\sum_{j=2}^{n} \frac{1}{j^2}$$

where $\mu_1 = 0$ and $\sigma^2_1 = 0$.

For $k > 1$, the asymptotic probability $\mathbb{P}[X_{n,k} = 0]$ was determined by Sanderson [26] who showed that

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

Here we provide an exact, closed-form expression for $\mathbb{P}[X_{n,k} = 0]$. In addition, as n becomes large, $X_{n,k}$ has a geometric distribution with parameter $2/(k+1)$.

Theorem 2. *[15]*

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

2. *For $k > 1, r \geq 0$,*
$$\lim_{n\to\infty} \mathbb{P}[X_{n,k} \geq r] = \left(\frac{2}{k+1}\right)^r$$

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 $\mathbb{E}[d(v_{ij},\rho)] = \mathbb{E}[X_{n,2}]$, so we have $\mathbb{E}[d(i,j)] = 2\mu_n - 2\mathbb{E}[X_{n,2}]$. 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

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

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 Probability distribution for cherries

A *cherry* is a pair of leaves, each of which is adjacent to a shared vertex. In this section we describe the distribution of the number of cherries for a tree that evolves under the Yule model. We also compare this to the distribution of the number of cherries for a tree selected uniformly at random from the set of all (unrooted) leaf-labelled trees. Note that when we suppress the root of a rooted tree (and thereby convert an rooted tree into an unrooted one) the number of cherries is either unchanged or it increases by at most 1, and this difference has a negligible effect on the asymptotic results described.

Under either model we will let the random variable C_n denote the number of cherries in the randomly generated tree. By realizing the process of cherry formation in these two models by extended Polya urn models it can be shown that C_n is asymptotically normal. We also give exact formulas for the mean and standard deviation of C_n in these two models. This section is based on [7], where proofs of the main results may be found.

5.1 Yule Model

Theorem 3. *[7,30] Let μ_n be the mean number of cherries for a rooted binary tree on n leaves, and σ_n^2 be the variance for the number of cherries. Under the Yule distribution we have the recursions, for $n \geq 2$:*

$$\mu_{n+1} = 1 + \mu_n(1 - \frac{2}{n}); \quad \sigma_{n+1}^2 = \sigma_n^2(1 - \frac{4}{n}) + \frac{2}{n}\mu_n(1 - \frac{2}{n}\mu_n)$$

which may be solved exactly to give

$$\mu_n = \frac{n}{3} \quad (n \geq 3); \quad \sigma_n^2 = \frac{2n}{45} \quad (n \geq 5)\,.$$

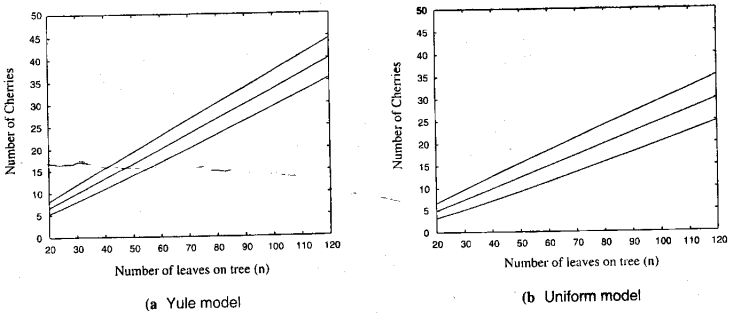


Fig. 4. Rejection limits for large n of the Yule and uniform null hypotheses at the $\alpha = 0.05$ level. The solid line represents the mean number of cherries, while the dashed lines are the lower and upper limits for rejection of the null hypotheses. The rejection limits are based upon a normal approximation which is valid for $n \gtrsim 20$. **(a)** Yule model **(b)** Uniform model

Asymptotically we have

$$\frac{C_n - n/3}{\sqrt{2n/45}} \to \mathcal{N}(0,1)\,.$$

The Yule model can be used as a simple null hypothesis to explore patterns in phylogenetic trees. A simple two-tailed test of the Yule null hypothesis, for a given tree, can be made based on the number of cherries in the tree. If the number of cherries is below some lower critical value, or above some upper critical value, then the Yule null hypothesis is rejected.

For small n, the rejection limits may be calculated exactly using a recursive formula for the probabilities. For larger values of n ($n \gtrsim 20$) a normal approximation is valid. In this case, based on the asymptotic result in Theorem 3, the rejection region for a two-sided test at the α level is given by

$$C_n < \frac{n}{3} - Z_{\frac{\alpha}{2}}\sqrt{\frac{2n}{45}} \quad \text{and} \quad C_n > \frac{n}{3} + Z_{\frac{\alpha}{2}}\sqrt{\frac{2n}{45}}\,.$$

The lower and upper critical values for rejection at an $\alpha = 0.05$ level are shown in Fig. 4. If the Yule model is rejected then this implies that one or more of the assumptions upon which it is based is invalid. Often it is assumed that the assumption of equal probability of speciation is the invalid assumption, but this need not be the case [21].

5.2 Uniform model

In the uniform model equal probability is assigned to each possible leaf-labeled binary tree on n leaves. Thus the uniform model distribution may be used to

model the frequency of outcomes that would occur if the process of tree reconstruction did no better than random selection from the set of possible binary trees on n leaves.

Theorem 4. *[7,30–32] Let μ_n be the mean value of C_n for an unrooted binary tree on n leaves, and σ_n^2 be the variance for C_n. Under the uniform model, for $n \geq 4$,*

(a)

$$\mathbb{P}[C_n = k] = \frac{n!(n-2)!(n-4)!2^{n-2k}}{(n-2k)!(2n-4)!k!(k-2)!}, \quad k \geq 2$$

(b)

$$\mu_n = \frac{n(n-1)}{2(2n-5)} \sim \frac{n}{4}; \quad \sigma_n^2 = \frac{n(n-1)(n-4)(n-5)}{2(2n-5)^2(2n-7)} \sim \frac{n}{16}$$

Asymptotically we have

$$\frac{C_n - n/4}{\sqrt{n/16}} \to \mathcal{N}(0,1)\,.$$

A test of the uniform model null hypothesis may be constructed based on the number of cherries in a tree. For small n the probability distribution given in Theorem 4 may be used to calculate the rejection limits. For larger n ($n \gtrsim 20$) a analysis similar to that for the Yule model, but based on the asymptotic result in Theorem 4, gives as the rejection region:

$$C_n < \frac{n}{4} - Z_{\frac{\alpha}{2}}\sqrt{\frac{n}{16}} \quad \text{and} \quad C_n > \frac{n}{4} + Z_{\frac{\alpha}{2}}\sqrt{\frac{n}{16}}\,.$$

The lower and upper critical values for rejection at an $\alpha = 0.05$ level are shown in Fig. 4.

5.3 An example

Figure 1 in [33] is a rooted phylogenetic tree for 34 species of *eureptantic nemerteans* (ribbon worms). This tree has 7 cherries (rooted or unrooted). For the Yule model null hypothesis test at the $\alpha = 0.05$ level the lower rejection limit is 8 cherries or less, and the upper rejection limit is 15 cherries or more. So for the ribbon worm tree the Yule model null hypothesis is rejected. For the uniform model null hypothesis test at the $\alpha = 0.05$ level the lower rejection limit is 5 cherries or less, and the upper rejection limit is 13 cherries or more, and so the test does not reject the uniform model null hypothesis. In any hypothesis test, however, it is important to note that a reconstructed tree is only an estimate of the underlying species tree. Consequently a more refined analysis would take into account the uncertainty and possible biases in phylogeny reconstruction [8,34].

6 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) [35].

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 [36]. 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 [37,38]. 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 described in [39] 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. For detailed proofs of the results that follow see [15].

6.1 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

$$\mathbb{P}[e \mid T] = \frac{\mathbb{P}[T, e]}{\mathbb{P}[T]} \,. \tag{3}$$

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 4 leaves (Fig. 5). 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.

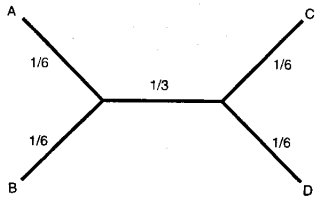


Fig. 5. Conditional probabilities ($\mathbb{P}[e \mid T]$) for the edges of a labeled unrooted tree on 4 leaves

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 $|E_{\max}(T)| > 1$. However, we will describe below that, for the Yule model, $|E_{\max}(T)| \leq 3$.

6.2 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 $\varepsilon(n)$ 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,

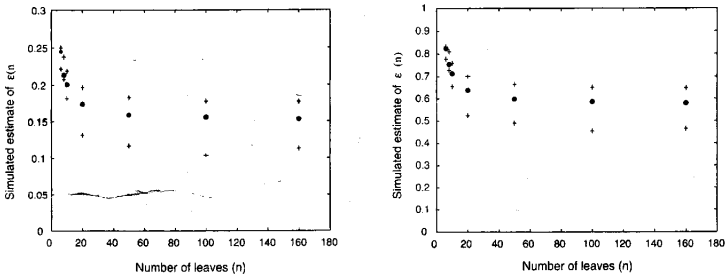
$$\varepsilon(n) = \sum_T \mathbb{P}[e_{\max} \text{ is the root edge of } T | u(T') = T]\mathbb{P}[u(T') = T],$$

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

$$\varepsilon(n) = \sum_T \mathbb{P}[e_{\max}(T) \mid T]\ \mathbb{P}[T]\ . \tag{4}$$

One might expect that $\varepsilon(n)$ 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 see 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 see that $\varepsilon(n)$ 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 [13]. Furthermore, although the limit $\varepsilon(n)$ 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.



(a (b

Fig. 6. 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

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

$$\varepsilon(n) \approx \frac{1}{N}\sum_{i=1}^{N} \mathbb{P}[e_{\max}(T_i) \mid T_i] = \frac{1}{N}\sum_{i=1}^{N} \frac{\mathbb{P}[T_i, e_{\max}(T_i)]}{\mathbb{P}[T_i]}\,, \tag{5}$$

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 $\lim_{n\to\infty} \varepsilon(n) \approx 0.15$ (Fig. 6a). The five edges with the largest conditional probabilities for a tree were always an internal edge and the four edges adjacent to it. Let $\varepsilon_5(n)$ denote the mean value for the sum of the five largest conditional probabilities for a tree. The simulations suggest that $\lim_{n\to\infty} \varepsilon_5(n) \approx 0.58$ (Fig. 6b). 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.

6.3 Exact asymptotic value of $\varepsilon(n)$

Given an edge e of an unrooted phylogenetic tree, let $H(e)$ denote the number of labeled histories associated with the rooted tree that arises from T by subdividing edge e. 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. Then $H(e) \geq H(e')$ for each of the four edges e' incident with e precisely if both the following two inequalities hold:

$$a + b \geq max\{c, d\}; \quad c + d \geq max\{a, b\} . \tag{6}$$

Furthermore, $H(e) > H(e')$ for all e' precisely if these two inequalities hold as strict inequalities. It follows that any two edges in $E_{\max}(T)$ are adjacent, and consequently, $|E_{\max}(T)| \leq 3$. Furthermore, if both the inequalities in (6) are strict, then $|E_{\max}(T)| = 1$.

We now describe the exact asymptotic value of $\varepsilon(n)$. This value was calculated by embedding the discrete process of rooting a tree into a continuous analogue involving 'stick breaking'.

Theorem 5. *[15] 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 converges to the value $4\ln(4/3) - 1$ (≈ 0.15) as $n \to \infty$.*

We end this section by noting that $\mathbb{P}[e_{\max}(T) \mid T]$ may be arbitrarily close to zero, for a tree T with a sufficiently large number of leaves. For example, consider a *caterpillar tree*, which 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 asymptotically, as $n \to \infty$,

$$\mathbb{P}\left[e_{\max}(C_n) \mid C_n\right] \sim \frac{2}{3}\sqrt{\frac{2}{\pi}}\frac{1}{\sqrt{n}} \to 0 . \tag{7}$$

7 Extending 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

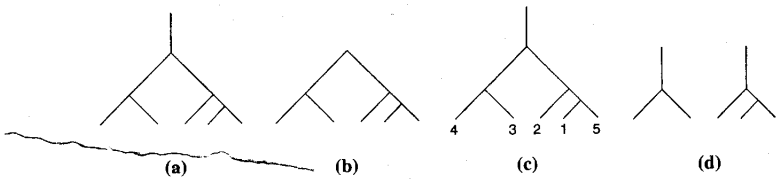


Fig. 7. Rooted tree types **(a)** An unlabeled binary tree on 5 leaves ; $\tau \in \mathrm{UB}(5)$. The root vertex is labeled s_0. **(b)** The edge-rooted unlabeled binary tree on 5 leaves obtained by removing the root leaf and its incident edge ; $\tau^* \in EUB(5)$ **(c)** A labeled binary tree on 5 leaves ; $\tau \in \mathrm{LB}(5)$ **(d)** The two subtrees, τ^1 and τ^2 of the tree τ in **(a)**

$s(t - \Lambda)$). 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 rescaling 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 colonizing 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 [14]. 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. More detail of the work that follows may be found in [15].

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

- For $n \geq 1$, 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. 7a).
- For $n \geq 2$, 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 $\tau \in UB(n)$ we will let τ^* denote the associated tree in $EUB(n)$ (Fig. 7b).

- 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 $LB(n) := \cup_{\tau \in UB(n)} L(\tau)$, the set of *l*abeled *b*inary trees (Fig. 7c).

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 $0 \leq t \leq 1$ and $\tau \in UB(n)$, consider the following (absolute and conditional) probabilities

$$f(\tau, t) := \mathbb{P}[T(t) = \tau]; \qquad p(\tau) := \mathbb{P}[T(1) = \tau | T(1) \text{ has } n \text{ leaves}] . \tag{8}$$

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

$$S(x) := \mathbb{P}[\Lambda(s_0) \geq x]; \qquad \sigma(x) := s(x)S(x) , \tag{9}$$

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 [40]

$$\mathbb{P}[\Lambda(s_0) \geq x] = exp[-\int_0^x s(\lambda) d\lambda] . \tag{10}$$

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

$$f(\tau, t) = 2^{\delta(\tau)} \int_0^t f(\tau^1, t-x) f(\tau^2, t-x) \sigma(x) dx \tag{11}$$

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. 7d), and where

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

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

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

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

$$p(\tau) = \mathbb{P}[T(1) = \tau | N(1) = n] = \frac{f(\tau, 1)}{\nu(n, 1)} . \tag{12}$$

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

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

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

$$p(\tau^*) = 2^{\delta(\tau)} p(\tau^1) p(\tau^2) , \tag{13}$$

with τ^1, τ^2 as in Eq. (11).

7.1 Two classes of models

1. The simplest model has $s(x) = s > 0$, constant. This gives the Yule model as described in Sect. 3. In this case, $\sigma(x) = se^{-sx}$ and $N(t)$ models a pure birth process, so $\nu(k,t) = e^{-st}(1-e^{-st})^{k-1}$. Under this model $p(\tau) = p(\tau^*) = 2^{u(\tau)}\prod_{i>2}(i-1)^{-d_i(\tau)}$, 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. A second class of models are those which satisfy the condition:

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

which we will call "explosive radiation" models. In these model, unless a species has undergone a speciation event within the last ε time interval, it will never do so. Thus, in this model, speciation events would tend to be clustered close together. Our last result (Theorem 6) shows 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 $p_{unif}(\tau) = |L(\tau)|\,/\,|LB(n)|$, 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 $p_{unif}(\tau)$ 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 [17], $|LB(n)| = \frac{(2n-2)!}{(n-1)!2^{n-1}}$ Therefore, under the uniform model on rooted trees,

$$p_{unif}(\tau) = \frac{|L(\tau)|}{|LB(n)|} = \binom{2n-2}{n-1}^{-1} 2^{u(\tau)}\,, \tag{14}$$

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

Theorem 6. *[15] Under an explosive radiation model, with $\varepsilon < 1/n$, the probability distribution on trees is precisely that induced by the uniform model. That is,*

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

Acknowledgement

We thank Charles Semple for some helpful comments on an earlier version of this manuscript. This research was supported by the New Zealand Marsden Fund (UOC-MIS-003).

References

1. D. Aldous: Probability distributions on cladograms, in *Random Structures*, ed. by D. Aldous, R. Permantle, Vol. 76 (Springer, 1996), pp. 1–18
2. J. Brown: Syst. Biol. **43**(1), 78–90 (1994)
3. E. Harding: Adv. Appl. Prob. **3**, 44–77 (1971)
4. S. Heard: Evolution **46**(6), 1818–1826 (1992)
5. S. Heard: Evolution **50**(6), 2141–2148 (1996)
6. T. Kubo, Y. Iwasa: Evolution **49**, 694–704 (1995)
7. A. McKenzie, M. Steel: Mathematical Biosciences **164**(1), 81–92 (2000)
8. A. Mooers, S. Heard: The Quarterly Review of Biology **72**(1), 31–54 (1997)
9. J. Slowinski: Syst. Zool. **39**(1), 89–94 (1990)
10. B. Rannala, Z. Yang: Molecular Evolution **43**, 304–311 (1996)
11. B. Mau, M. Newton, B. Larget: Biometrics **55**, 1–12 (1999)
12. S. Li, D.K. Pearl, H. Doss: Journal of the American Statistical Association **95**(450), 493–508 (2000)
13. J. Haigh: J. Appl. Prob. **7**, 79–88 (1970)
14. D. Aldous: (2000), 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]
15. M. Steel, A. McKenzie: Math. Biosci. **170**(1), 91–112 (2001)
16. R. Page, E. Holmes: *Molecular Evolution: a phylogenetic approach* (Blackwell Science, 1998), Chap. 2, pp. 11–36
17. L. Cavalli-Sforza, A. Edwards: Evolution **21**, 550–570 (1967)
18. A. Edwards, L. Cavalli-Sforza: Reconstruction of evolutionary trees, in *Phenetic and phylogenetic classification*, ed. by W. Heywood, J. McNeill (1964), no. 6, pp. 67–76
19. J. Slowinski, C. Guyer: The American Naturalist **134**(6), 907–921 (1989)
20. S. Nee, R. May, P. Harvey: Phil. Trans. R. Soc. Lond. B **344**, 305–311 (1994)
21. J. Losos, F. Adler: The American Naturalist **145**(3), 329–342 (1995)
22. J. Kingman: J. Appl. Prob. **19A**, 27–43 (1982)
23. F. Tajima: Genetics **105**, 437–460 (1983)
24. D. Knuth: *The Art of Comuter Programming*, Vol. 3 (Addison-Wesley, Reading, Massachusetts, 1997), Chap. 5, p. 67, problem 20
25. A. Edwards: J. R. Stat. Soc. Ser. B **32**, 155–174 (1970)
26. M. Sanderson: Syst. Biol. **45**(2), 168–173 (1996)
27. W. Lynch: The Computer Journal **71**, 299–302 (1965)
28. H. Mahmoud: *Evolution of random search trees* (John Wiley and Sons Ltd, New York, 1992), p. 72
29. R. Stanley: *Enumerative Combinatorics: Vol. 1*, no. 49 in Cambridge Studies in Advanced Mathematics (Cambridge University Press, 1997), pp. 18–19
30. M. Steel, D. Penny: Syst. Biol. **42**(2), 126–141 (1993)
31. M.D. Hendy, D. Penny: Math. Biosci. **59**, 277–290 (1982)
32. M.A. Steel: SIAM J. Discr. Math. **1**(4), 541–551 (1988)
33. M. Härlin: Biological Journal of the Linnean Society **58**, 325–342 (1996)
34. J.P. Huelsenbeck, M. Kirkpatrick: Evolution **50**(4), 1418–1424 (1996)
35. M. Ridley: *Evolution* (Blackwell Science,Inc., Cambridge, Massachusetts, USA, 1993), Chap. 17, pp. 460–466
36. A. Smith: Biological Journal of the Linnean Society **51**, 279–292 (1994)
37. J. Farris: American Naturalist **106**, 646–668 (1972)

38. D.L. Swofford, G.J. Olsen, P.J. Waddell, D.M. Hillis: Phylogenetic inference, in *Molecular Systmatics*, ed. by D.M. Hillis, C. Moritz, B.K. Mable (Sinauer Associates, Inc., Sunderland, Massachusetts, U.S.A, 1996), p. 488
39. Y.V. de Peer, R.D. Wacher: Comput. Applic. Biosci. **9**, 177–182 (1993)
40. J. Medhi: *Stochastic Processes* (John Wiley and Sons Ltd, New York, 1982), Chap. 4, p. 119