Hindawi
The Scientific World Journal
Volume 2018, Article ID 1916094, 7 pages
https://doi.org/10.1155/2018/1916094

Research Article

The Probabilities of Trees and Cladograms under Ford's α-Model

Tomás M. Coronado, Arnau Mir, and Francesc Rosselló

Balearic Islands Health Research Institute (IdISBa) and Department of Mathematics and Computer Science, University of the Balearic Islands, 07122 Palma, Spain

Correspondence should be addressed to Francesc Rosselló; cesc.rossello@uib.es

Received 1 February 2018; Accepted 8 March 2018; Published 18 April 2018

Academic Editor: Béla Tóthmérész

Ford's α-model is one of the most popular random parametric models of bifurcating phylogenetic tree growth, having as specific instances both the uniform and the Yule models. Its general properties have been used to study the behavior of phylogenetic tree shape indices under the probability distribution it defines. But the explicit formulas provided by Ford for the probabilities of unlabeled trees and phylogenetic trees fail in some cases. In this paper we give correct explicit formulas for these probabilities.

1. Introduction

The study of random growth models of rooted phylogenetic trees and the statistical properties of the shapes of the phylogenetic trees they produce was initiated almost one century ago by Yule [1] and it has gained momentum in the last 20 years: see, for instance, [2–8]. The final goal of this line of research is to understand the relationship between the forces that drive evolution and the topological properties of "real-life" phylogenetic trees [3, 9]; see also [10, Chapter 33]. One of the most popular such models is Ford's α-model for rooted bifurcating phylogenetic trees or *cladograms* [4]. It consists of a parametric model that generalizes both the uniform model (where new leaves are added equiprobably to any arc, giving rise to the uniform probability distribution on the sets of cladograms with a fixed set of taxa) and Yule's model (where new leaves are added equiprobably only to *pendant* arcs, i.e., to arcs ending in leaves) by allocating a possibly different probability (that depends on a parameter α and hence its name, "α-model") to the addition of the new leaves to pendant arcs or to internal arcs.

When models like Ford's model are used to contrast topological properties of phylogenetic trees contained in databases like TreeBase (`https://treebase.org`), only their general properties (moments, asymptotic behavior) are employed. But, in the course of a research where we have needed to compute the probabilities of several specific cladograms under this model [11], we have noticed that the explicit formulas that Ford gives in [4, §3.5] for the probabilities of cladograms and of *tree shapes* (unlabeled rooted bifurcating trees) are wrong, failing for some trees with $n \geqslant 8$ leaves; see Propositions 29 and 32 in [4], with the definition of $\widehat{q}$ given in page 30 therein, for Ford's formulas.

So, to help the future user of Ford's model, in this paper we give the correct explicit formulas for these probabilities. This paper is accompanied by the GitHub page `https://github.com/biocom-uib/prob-alpha` where the interested reader can find a SageMath [12] module to compute these probabilities and their explicit values on the sets $\mathcal{T}_n$ of cladograms with n leaves labeled $1, \ldots, n$, for every n from 2 to 8.

2. Preliminaries

2.1. Definitions, Notations, and Conventions. Throughout this paper, by a *tree* T, we mean a rooted bifurcating tree. As it is customary, we understand T as a directed graph, with its arcs pointing away from the root, which we shall denote by r_T. Then, all nodes in T have out-degree either 0 (its *leaves*, which form the set $L(T)$) or 2 (its *internal nodes*, which form the set $V_{\text{int}}(T)$). The *children* of an internal node v are those

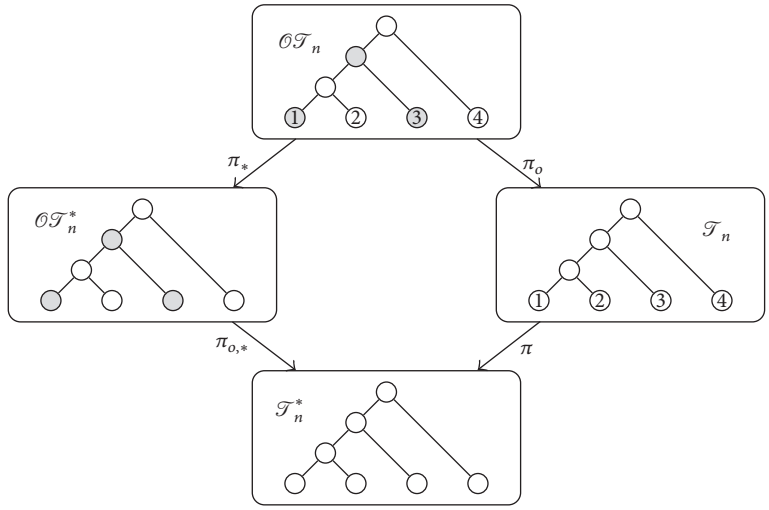


Figure 1: An example of images under the forgetful mappings between (ordered and unordered) cladograms and tree shapes. In the ordered objects, the ordering is represented by the nodes' colors: gray $\prec$ white.

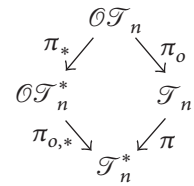
nodes u such that (v, u) is an arc in T, and they form the set child(v). A node x is a *descendant* of a node v when there exists a directed path from v to x in T. For every node v, the *subtree* T_v *of* T *rooted at* v is the subgraph of T induced on the set of descendants of v.

A tree T is *ordered* when it is endowed with an *ordering* $\prec_v$ on every set child(v). A *cladogram* (resp., an *ordered cladogram*) on a set of taxa Σ is a tree (resp., an ordered tree) with its leaves bijectively labeled in Σ. Whenever we want to stress the fact that a tree is not a cladogram, that is, it is an unlabeled tree, we shall use the term *tree shape*.

It is important to point out that although ordered trees have no practical interest from the phylogenetic point of view, because the orderings on the sets of children of internal nodes do not carry any phylogenetic information, they are useful from the mathematical point of view, because the existence of the orderings allows one to easily prove certain extra properties that can later be translated to the unordered setting (cf. Proposition 1).

An *isomorphism* of ordered trees is an isomorphism of rooted trees that moreover preserves these orderings. An *isomorphism* of cladograms (resp., of ordered cladograms) is an isomorphism of trees (resp., of ordered trees) that preserves the leaves' labels. We shall always identify a tree shape, an ordered tree shape, a cladogram, or an ordered cladogram, with its isomorphism class, and in particular we shall make henceforth the abuse of language of saying that two of these objects, T, T', *are the same*, in symbols $T = T'$, when they are (only) isomorphic. We shall denote by $\mathcal{T}_n^*$ and $\mathcal{OT}_n^*$, respectively, the sets of tree shapes and of ordered tree shapes with n leaves. Given any finite set of taxa Σ, we shall denote by $\mathcal{T}_\Sigma$ and $\mathcal{OT}_\Sigma$, respectively, the sets of cladograms and of ordered cladograms on Σ. When the specific set Σ is unrelevant and only its cardinal matters, we shall write $\mathcal{T}_n$ and $\mathcal{OT}_n$ (with $n = |\Sigma|$) instead of $\mathcal{T}_\Sigma$ and $\mathcal{OT}_\Sigma$, and then we shall understand that Σ is $[n] = \{1, 2, \ldots, n\}$.

There exist natural isomorphism-preserving forgetful mappings



that "forget" the orderings or the labels of the trees. In particular, we shall call the image of a cladogram under π its *shape*. Figure 1 depicts an example of images under these forgetful mappings.

Let us introduce some more notations. For every node v in a tree T, $\kappa_T(v)$ is its number of descendant leaves. For every internal node v in an ordered tree T, with children $v_1 \prec_v v_2$, its *numerical split* is the ordered pair $\mathrm{NS}_T(v) = (\kappa_T(v_1), \kappa_T(v_2))$. If, instead, T is unordered and if child(v) $= \{v_1, v_2\}$ with $\kappa_T(v_1) \leqslant \kappa_T(v_2)$, then $\mathrm{NS}_T(v) = (\kappa_T(v_1), \kappa_T(v_2))$. In both cases, the *multiset of numerical splits* of T is $\mathrm{NS}(T) = \{\mathrm{NS}_T(v) \mid v \in V_{\mathrm{int}}(T)\}$. For instance, if T is the cladogram depicted in Figure 2, then

$$\mathrm{NS}(T) = \{(1,1), (1,1), (1,1), (2,2), (1,4), (2,5)\}. \quad (1)$$

A *symmetric branch point* in a tree T is an internal node v such that if v_1 and v_2 are its children, then the subtrees T_{v_1} and T_{v_2} of T rooted at them have the same shape. For instance, the symmetric branch points in the cladogram depicted in Figure 2 are those filled in black.

Given two cladograms T and T' on Σ and Σ', respectively, with $\Sigma \cap \Sigma' = \emptyset$, their *root join* is the cladogram $T \star T'$ on $\Sigma \cup \Sigma'$ obtained by connecting the roots of T and T' to a (new) common root r; see Figure 3. If T, T' are ordered cladograms, $T \star T'$ is ordered by inheriting the orderings on T and T' and ordering the children of the new root r as $r_T \prec_r r_{T'}$. If T and T' are tree shapes, a similar construction yields a tree shape

$T \star T'$; if they are moreover ordered, then $T \star T'$ becomes an ordered tree shape as explained above.

2.2. The α-Model. Ford's α-model [4] defines, for every $n \geqslant 1$, a family of probability density functions $P^{(*)}_{\alpha,n}$ on $\mathcal{T}^*_n$ that depends on one parameter $\alpha \in [0,1]$, and then it translates this family into three other families of probability density functions $P_{\alpha,n}$ on $\mathcal{T}_n$, $P^{(o,*)}_{\alpha,n}$ on $\mathcal{OT}^*_n$, and $P^{(o)}_{\alpha,n}$ on $\mathcal{OT}_n$, by imposing that the probability of a tree shape is equally distributed among its preimages under π, $\pi_{o,*}$, and $\pi \circ \pi_o = \pi_{o,*} \circ \pi_*$, respectively.

It is well known [13] that every $T \in \mathcal{T}_n$ can be obtained in a unique way by adding recurrently to a single node labeled 1 new leaves labeled $2, \ldots, n$ to arcs (i.e., splitting an arc (u,v) into two arcs (u,w) and (w,v) and then adding a new arc from the inserted node w to a new leaf) or to a new root (i.e., adding a new root w and new arcs from w to the old root and to a new leaf). The value of $P^{(*)}_{\alpha,n}(T^*)$ for $T^* \in \mathcal{T}^*_n$ is determined through all possible ways of constructing cladograms with shape T^* in this way. More specifically,

(1) if T_1 and T_2 denote, respectively, the only cladograms in $\mathcal{T}_1$ and $\mathcal{T}_2$, let $P'_{\alpha,1}(T_1) = P'_{\alpha,2}(T_2) = 1$;

(2) for every $m = 3, \ldots, n$, let $T_m \in \mathcal{T}_m$ be obtained by adding a new leaf labeled m to T_{m-1}. Then

$$P'_{\alpha,m}(T_m) = \begin{cases} \dfrac{\alpha}{m-1-\alpha} \cdot P'_{\alpha,m-1}(T_{m-1}) & \text{if the new leaf is added to an internal arc or to a new root} \\ \dfrac{1-\alpha}{m-1-\alpha} \cdot P'_{\alpha,m-1}(T_{m-1}) & \text{if the new leaf is added to a pendant arc;} \end{cases} \tag{2}$$

(3) When the desired number n of leaves is reached, the probability of every tree shape $T^*_n \in \mathcal{T}^*_n$ is defined as

$$P^{(*)}_{\alpha,n}(T^*_n) = \sum_{\pi(T_n)=T^*_n} P'_{\alpha,n}(T_n). \tag{3}$$

For instance, Figure 4 shows the construction of two cladograms in $\mathcal{T}_5$ with the same shape and how their probability $P'_{\alpha,5}$ is built using the recursion in Step (2). If we generate all cladograms in $\mathcal{T}_5$ with this shape, we compute their probabilities $P'_{\alpha,5}$, and then we add up all these probabilities, we obtain the probability $P^{(*)}_{\alpha,5}$ of this shape, which turns out to be $2(1-\alpha)/(4-\alpha)$; cf. [4, Figure 23].

Once $P^{(*)}_{\alpha,n}$ is defined on $\mathcal{T}^*_n$, it is transported to $\mathcal{T}_n$, $\mathcal{OT}^*_n$, and $\mathcal{OT}_n$ by defining the probability of an object in one of these sets as the probability of its image in $\mathcal{T}^*_n$ divided by the number of preimages of this image:

(i) For every $T \in \mathcal{T}_n$, if $\pi(T) = T^* \in \mathcal{T}^*_n$ and it has k symmetric branch points, then

$$P_{\alpha,n}(T) = \frac{2^k}{n!} \cdot P^{(*)}_{\alpha,n}(T^*), \tag{4}$$

because $|\pi^{-1}(T^*)| = n!/2^k$ (see, e.g., [4, Lemma 31]).

(ii) For every $T_o \in \mathcal{OT}_n$, if $\pi_o(T_o) = T \in \mathcal{T}_n$, then

$$P^{(o)}_{\alpha,n}(T_o) = \frac{1}{2^{n-1}} \cdot P_{\alpha,n}(T), \tag{5}$$

because $|\pi_o^{-1}(T)| = 2^{n-1}$ (T has 2^{n-1} different preimages under π_o, obtained by taking all possible different combinations of orderings on the $n-1$ sets child(v), $v \in V_{\text{int}}(T^*)$).

(iii) For every $T^*_o \in \mathcal{OT}^*_n$, if $\pi_{o,*}(T^*_o) = T^* \in \mathcal{T}^*_n$ and it has k symmetric branch points, then

$$P^{(o,*)}_{\alpha,n}(T^*_o) = \frac{1}{2^{n-k-1}} \cdot P^{(*)}_{\alpha,n}(T^*), \tag{6}$$

because $|\pi^{-1}_{o,*}(T^*)| = 2^{n-1-k}$ (from the 2^{n-1} possible preimages of T^* under $\pi_{o,*}$, defined by all possible different combinations of orderings on the $n-1$ sets child(v), $v \in V_{\text{int}}(T^*)$, those differing only on the orderings on the children of the k symmetric branch points are actually the same ordered tree shape).

The family $(P^{(o,*)}_{\alpha,n})_n$ of probabilities of ordered tree shapes satisfies the useful Markov branching recurrence (in the sense of [2, §4]) given by the following proposition. In it and in the sequel, let, for every $a, b \in \mathbb{Z}^+$,

$$q_\alpha(a,b) = \frac{\Gamma_\alpha(a)\,\Gamma_\alpha(b)}{\Gamma_\alpha(a+b)} \cdot \varphi_\alpha(a,b), \tag{7}$$

where

$$\varphi_\alpha(a,b) = \frac{\alpha}{2}\binom{a+b}{a} + (1-2\alpha)\binom{a+b-2}{a-1} \tag{8}$$

and $\Gamma_\alpha : \mathbb{Z}^+ \to \mathbb{R}$ is the mapping defined by $\Gamma_\alpha(1) = 1$ and, for every $n \geqslant 2$, $\Gamma_\alpha(n) = (n-1-\alpha) \cdot \Gamma_\alpha(n-1)$.

Proposition 1. *For every $0 < m < n$ and for every $T^*_m \in \mathcal{OT}^*_m$ and $T^*_{n-m} \in \mathcal{OT}^*_{n-m}$,*

$$\begin{aligned} &P^{(o,*)}_{\alpha,n}(T^*_m \star T^*_{n-m}) \\ &\quad = q_\alpha(m, n-m)\, P^{(o,*)}_{\alpha,m}(T^*_m)\, P^{(o,*)}_{\alpha,n-m}(T^*_{n-m}). \end{aligned} \tag{9}$$

This recurrence, together with the fact that $P^{(o,*)}_{\alpha,1}$ of a single node is 1, implies that, for every $T^*_o \in \mathcal{OT}^*_n$,

$$P^{(o,*)}_{\alpha,n}(T^*_o) = \prod_{(a,b)\in \text{NS}(T^*_o)} q_\alpha(a,b). \tag{10}$$

For proofs of Proposition 1 and (10), see Lemma 27 and Proposition 28 in [4], respectively.

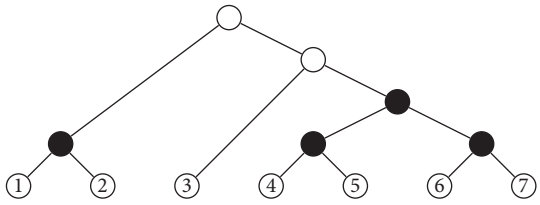


Figure 2: A cladogram in $\mathscr{T}_7$. The black nodes are its symmetric branch points.

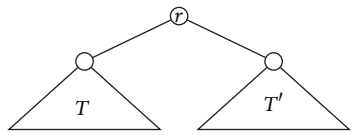


Figure 3: The root join $T \star T'$.

3. Main Results

Our first result is an explicit formula for $P_{\alpha,n}(T)$, for every $n \geqslant 1$ and $T \in \mathscr{T}_n$.

Proposition 2. *For every $T \in \mathscr{T}_n$, its probability under the α-model is*

$$P_{\alpha,n}(T) = \frac{2^{n-1}}{n! \cdot \Gamma_\alpha(n)} \prod_{(a,b)\in \mathrm{NS}(T)} \varphi_\alpha(a,b). \tag{11}$$

Proof. Given $T \in \mathscr{T}_n$, let T_o be any ordered cladogram such that $\pi_o(T_o) = T$, and let $T_o^* = \pi_*(T_o) \in \mathscr{OT}_n^*$ and $T^* = \pi(T) = \pi_{o,*}(T_o^*)$. If T^* has k symmetric branch points, then, by (4), (6), and (10),

$$\begin{aligned} P_{\alpha,n}(T) &= \frac{2^k}{n!} \cdot P_{\alpha,n}^{(*)}(T^*) = \frac{2^k}{n!} \cdot 2^{n-k-1} \cdot P_{\alpha,n}^{(o,*)}(T_o^*) \\ &= \frac{2^{n-1}}{n!} \prod_{(a,b)\in \mathrm{NS}(T_o^*)} q_\alpha(a,b). \end{aligned} \tag{12}$$

Now, on the one hand, it is easy to check that

$$\begin{aligned} &\mathrm{NS}(T) \\ &\quad = \left\{(\min\{a,b\}, \max\{a,b\}) \mid (a,b) \in \mathrm{NS}(T_0^*)\right\}, \end{aligned} \tag{13}$$

and therefore, since q_α is symmetric,

$$P_{\alpha,n}(T) = \frac{2^{n-1}}{n!} \prod_{(a,b)\in \mathrm{NS}(T)} q_\alpha(a,b). \tag{14}$$

It remains to simplify this product. If, for every $v \in V_{\mathrm{int}}(T)$, we denote its children by v_1 and v_2, then

$$\begin{aligned} &\prod_{(a,b)\in \mathrm{NS}(T)} q_\alpha(a,b) \\ &\quad = \prod_{v\in V_{\mathrm{int}}(T)} \frac{\Gamma_\alpha(\kappa_T(v_1))\,\Gamma_\alpha(\kappa_T(v_2))}{\Gamma_\alpha(\kappa_T(v))} \varphi_\alpha(\mathrm{NS}(v)). \end{aligned} \tag{15}$$

For every $v \in V_{\mathrm{int}}(T) \setminus \{r_T\}$, the term $\Gamma_\alpha(\kappa_T(v))$ appears twice in this product: in the denominator of the factor corresponding to v itself and in the numerator of the factor corresponding to its parent. Therefore, all terms $\Gamma_\alpha(\kappa_T(v))$ in this product vanish except $\Gamma_\alpha(\kappa_T(r_T)) = \Gamma_\alpha(n)$ (that appears in the denominator of its factor) and every $\Gamma_\alpha(\kappa_T(v)) = \Gamma_\alpha(1) = 1$ with v, a leaf. Thus,

$$P_{\alpha,n}(T) = \frac{2^{n-1}}{n!} \cdot \frac{1}{\Gamma_\alpha(n)} \cdot \prod_{v\in V_{\mathrm{int}}(T)} \varphi_\alpha(\mathrm{NS}(v)) \tag{16}$$

as we claimed. □

Remark 3. Ford states (see [4, Proposition 32 and page 30]) that if $T \in \mathscr{T}_n$, then

$$P_{\alpha,n}(T) = \frac{2^k}{n!} \prod_{(a,b)\in \mathrm{NS}(T)} \widehat{q}_\alpha(a,b), \tag{17}$$

where k is the number of symmetric branching points in T and

$$\widehat{q}_\alpha(a,b) = \begin{cases} 2q_\alpha(a,b) & \text{if } a \neq b \\ q_\alpha(a,b) & \text{if } a = b. \end{cases} \tag{18}$$

If we simplify $\prod_{(a,b)\in \mathrm{NS}(T)} \widehat{q}_\alpha(a,b)$ as in the proof of Proposition 2, this formula for $P_{\alpha,n}(T)$ becomes

$$P_{\alpha,n}(T) = \frac{2^{k+m}}{n! \cdot \Gamma_\alpha(n)} \cdot \prod_{(a,b)\in \mathrm{NS}(T)} \varphi_\alpha(a,b), \tag{19}$$

where m is the number of internal nodes whose children have different numbers of descendant leaves. This formula does not agree with the one given in Proposition 2 above, because

$$\begin{aligned} k + m = n - 1 - \Big|\Big\{ & v \in V_{\mathrm{int}}(T) \mid \mathrm{child}(v) \\ & = \{v_1, v_2\},\ \kappa_T(v_1) = \kappa_T(v_2) \text{ but } \pi\left(T_{v_1}\right) \\ & \neq \pi\left(T_{v_2}\right)\Big\}\Big| \end{aligned} \tag{20}$$

and, hence, it may happen that $k + m < n - 1$. The first example of a cladogram with this property (and the only one, up to relabeling, with at most 8 leaves) is the cladogram $\widetilde{T} \in \mathscr{T}_8$ depicted in Figure 5. For it, our formula gives (see (8.22) in the document `ProblsAlpha.pdf` in `https://github.com/biocom-uib/prob-alpha`)

$$P_{\alpha,8}\left(\widetilde{T}\right) = \frac{(1-\alpha)^2(2-\alpha)}{126(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)} \tag{21}$$

while expression (19) assigns to $\widetilde{T}$ a probability of half this value:

$$\frac{(1-\alpha)^2(2-\alpha)}{252(7-\alpha)(6-\alpha)(5-\alpha)(3-\alpha)}. \tag{22}$$

This last value cannot be right, for several reasons. Firstly, by [4, §3.12], when $\alpha = 1/2$, Ford's model is equivalent to the uniform model, where every cladogram in $\mathscr{T}_n$ has the same probability

$$\frac{1}{|BT_n|} = \frac{1}{(2n-3)!!} \tag{23}$$

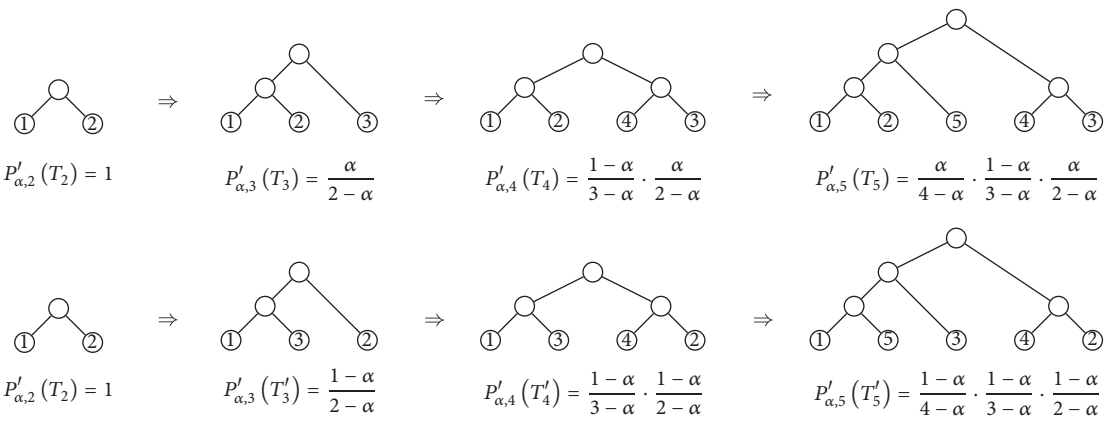


FIGURE 4: Two examples of computations of the probability $P'_{\alpha,n}$ of a cladogram through its construction in Step (2) of the definition of the α-model.

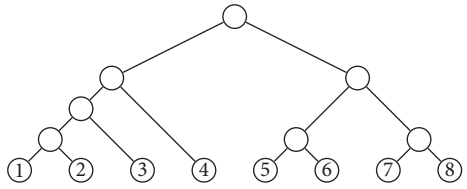


FIGURE 5: The cladogram $\widetilde{T} \in \mathscr{T}_8$ used in Remark 3.

and when $\alpha = 0$, Ford's model gives rise to the Yule model [1, 14], where the probability of every $T \in \mathscr{T}_n$ is

$$P_Y(T) = \frac{2^{n-1}}{n!} \prod_{v \in V_{\text{int}}(T)} \frac{1}{\kappa_T(v) - 1}. \tag{24}$$

In particular, $P_{1/2,8}(\widetilde{T})$ should be equal to 1/135135 and $P_{0,8}(\widetilde{T})$ should be equal to 1/19845. Both values are consistent with our formula, while expression (22) yields half these values.

As a second reason, which can be checked using a symbolic computation program, let us mention that if we take expression (22) as the probability of $\widetilde{T}$ and hence of all other cladograms with its shape, and we assign to all other cladograms in $\mathscr{T}_8$ the probabilities computed with Proposition 2, which agree on them with the values given by (19) (they are also provided in the aforementioned document `ProblsAlpha.pdf`), these probabilities do not add up 1.

Combining Proposition 2 and (4) we obtain the following result.

Corollary 4. *For every $T^* \in \mathscr{T}_n^*$ with k symmetric branch points,*

$$P_{\alpha,n}^{(*)}(T^*) = \frac{2^{n-k-1}}{\Gamma_\alpha(n)} \prod_{(a,b) \in \text{NS}(T^*)} \varphi_\alpha(a,b). \tag{25}$$

This formula does not agree, either, with the one given in [4, Proposition 29]: the difference lies again in the same factor of 2 to the power of the number of internal nodes that are not symmetric branch points but whose children have the same number of descendant leaves.

The family of density mappings $(P_{\alpha,n})_n$ satisfies the following Markov branching recurrence.

Corollary 5. *For every $0 < m < n$ and for every $T_m \in \mathscr{T}_m$ and $T_{n-m} \in \mathscr{T}_{n-m}$,*

$$P_{\alpha,n}(T_m \star T_{n-m}) = \frac{2q_\alpha(m, n-m)}{\binom{n}{m}} P_{\alpha,m}(T_m) P_{\alpha,n-m}(T_{n-m}). \tag{26}$$

Proof. If $T_m \in \mathscr{T}_m$ and $T_{n-m} \in \mathscr{T}_{n-m}$, then

$$P_{\alpha,m}(T_m) = \frac{2^{m-1}}{m!\Gamma_\alpha(m)} \prod_{(a,b) \in \text{NS}(T_m)} \varphi_\alpha(a,b)$$

$$P_{\alpha,n-m}(T_{n-m}) = \frac{2^{n-m-1}}{(n-m)!\Gamma_\alpha(n-m)} \cdot \prod_{(a,b) \in \text{NS}(T_{n-m})} \varphi_\alpha(a,b),$$

$$P_{\alpha,n}(T_m \star T_{n-m}) = \frac{2^{n-1}}{n!\Gamma_\alpha(n)} \prod_{(a,b) \in \text{NS}(T_m \star T_{n-m})} \varphi_\alpha(a,b)$$

$$= \frac{2^{n-1}}{n!\Gamma_\alpha(n)} \varphi_\alpha(m, n-m) \left(\prod_{(a,b) \in \text{NS}(T_m)} \varphi_\alpha(a,b) \right) \cdot \left(\prod_{(a,b) \in \text{NS}(T_{n-m})} \varphi_\alpha(a,b) \right) = \frac{2^{n-1}}{n!\Gamma_\alpha(n)} \varphi_\alpha(m, n-m) \cdot \frac{m!\Gamma_\alpha(m)}{2^{m-1}} P_{\alpha,m}(T_m) \cdot \frac{(n-m)!\Gamma_\alpha(n-m)}{2^{n-m-1}} \cdot P_{\alpha,n-m}(T_{n-m}) = \frac{2q_\alpha(m, n-m)}{\binom{n}{m}} P_{\alpha,m}(T_m) \cdot P_{\alpha,n-m}(T_{n-m}) \tag{27}$$

as we claimed. □

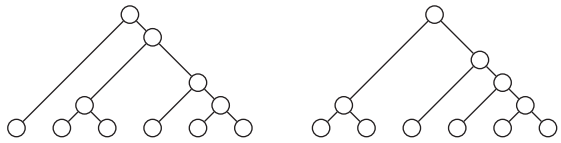


Figure 6: The tree shapes in $\mathscr{T}_6^*$ mentioned in Remark 6.

Remark 6. Against what is stated in [4], $(P_{\alpha,n}^{(*)})_n$ does not satisfy any Markov branching recurrence; that is, there does not exist any symmetric mapping $Q : \mathbb{Z}^+ \times \mathbb{Z}^+ \to \mathbb{R}$ such that, for every $k, l \geqslant 1$ and for every $T_k \in \mathscr{T}_k^*$ and $T_l \in \mathscr{T}_l^*$,

$$P_{\alpha,k+l}^{(*)}\left(T_k \star T_l\right) = Q\left(k,l\right) \cdot P_{\alpha,k}^{(*)}\left(T_k\right) \cdot P_{\alpha,l}^{(*)}\left(T_l\right). \tag{28}$$

Indeed, let $T_m^*, \widehat{T}_m^* \in \mathscr{T}_m^*$ be any two different tree shapes, both with m leaves and k symmetric branch points, for instance, the tree shapes in $\mathscr{T}_6^*$ depicted in Figure 6. Then,

$$\begin{aligned} P_{\alpha,m}^{(*)}\left(T_m^*\right) &= \frac{2^{m-k-1}}{\Gamma_\alpha\left(m\right)} \prod_{(a,b)\in \mathrm{NS}(T_m^*)} \varphi_\alpha\left(a,b\right), \\ P_{\alpha,m}^{(*)}\left(\widehat{T}_m^*\right) &= \frac{2^{m-k-1}}{\Gamma_\alpha\left(m\right)} \prod_{(a,b)\in \mathrm{NS}(\widehat{T}_m^*)} \varphi_\alpha\left(a,b\right). \end{aligned} \tag{29}$$

In this case, $T_m^* \star T_m^* \in \mathscr{T}_{2m}^*$ has $2k+1$ symmetric branch points and therefore

$$\begin{aligned} P_{\alpha,2m}^{(*)}\left(T_m^* \star T_m^*\right) &= \frac{2^{2m-2k-2}}{\Gamma_\alpha\left(2m\right)} \prod_{(a,b)\in \mathrm{NS}(T_m^* \star T_m^*)} \varphi_\alpha\left(a,b\right) \\ &= \frac{2^{2m-2k-2}}{\Gamma_\alpha\left(2m\right)} \varphi_\alpha\left(m,m\right) \left(\prod_{(a,b)\in \mathrm{NS}(T_m^*)} \varphi_\alpha\left(a,b\right) \right)^2 \\ &= \frac{2^{2m-2k-2}}{\Gamma_\alpha\left(2m\right)} \varphi_\alpha\left(m,m\right) \left(\frac{\Gamma_\alpha\left(m\right)}{2^{m-k-1}} P_{\alpha,m}^{(*)}\left(T_m^*\right) \right)^2 \\ &= q_\alpha\left(m,m\right) P_{\alpha,m}^{(*)}\left(T_m^*\right) P_{\alpha,m}^{(*)}\left(T_m^*\right) \end{aligned} \tag{30}$$

while $T_m^* \star \widehat{T}_m^* \in \mathscr{T}_{2m}^*$ has $2k$ symmetric branch points and therefore

$$\begin{aligned} P_{\alpha,2m}^{(*)}\left(T_m^* \star \widehat{T}_m^*\right) &= \frac{2^{2m-2k-1}}{\Gamma_\alpha\left(2m\right)} \prod_{(a,b)\in \mathrm{NS}(T_m^* \star \widehat{T}_m^*)} \varphi_\alpha\left(a,b\right) \\ &= \frac{2^{2m-2k-1}}{\Gamma_\alpha\left(2m\right)} \varphi_\alpha\left(m,m\right) \left(\prod_{(a,b)\in \mathrm{NS}(T_m^*)} \varphi_\alpha\left(a,b\right) \right) \\ &\cdot \left(\prod_{(a,b)\in \mathrm{NS}(\widehat{T}_m^*)} \varphi_\alpha\left(a,b\right) \right) = \frac{2^{2m-2k-1}}{\Gamma_\alpha\left(2m\right)} \varphi_\alpha\left(m,m\right) \\ &\cdot \frac{\Gamma_\alpha\left(m\right)}{2^{m-k-1}} P_{\alpha,m}^{(*)}\left(T_m^*\right) \cdot \frac{\Gamma_\alpha\left(m\right)}{2^{m-k-1}} P_{\alpha,m}^{(*)}\left(\widehat{T}_m^*\right) \\ &= 2q_\alpha\left(m,m\right) P_{\alpha,m}^{(*)}\left(T_m^*\right) P_{\alpha,m}^{(*)}\left(\widehat{T}_m^*\right) \end{aligned} \tag{31}$$

and $q_\alpha(m,m) \neq 2q_\alpha(m,m)$. This shows that there does not exist any well-defined, single real number $Q(m,m)$ such that

$$\begin{aligned} P_{\alpha,2m}^{(*)}\left(T_{1,m}^* \star T_{2,m}^*\right) &= Q\left(m,m\right) \cdot P_{\alpha,m}^{(*)}\left(T_{1,m}^*\right) \\ &\cdot P_{\alpha,m}^{(*)}\left(T_{2,m}^*\right) \end{aligned} \tag{32}$$

for every $T_{1,m}^*, T_{2,m}^* \in \mathscr{T}_m^*$.

Data Availability

The data used to support the findings of this study are available at the Github page that accompanies this paper.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by Spanish Ministry of Economy and Competitiveness and European Regional Development Fund Project DPI2015-67082-P (MINECO/FEDER). The authors thank G. Cardona and G. Riera for several comments on the SageMath module that accompanies this paper.

References

[1] G. U. Yule, "A Mathematical Theory of Evolution, Based on the Conclusions of Dr. J. C. Willis, F.R.S.," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 213, no. 402-410, pp. 21–87, 1925.

[2] D. Aldous, "Probability distributions on cladograms," in *Random discrete structures*, D. Aldous and R. Pemantle, Eds., vol. 76, pp. 1–18, Springer, New York, NY, USA, 1996.

[3] M. G. B. Blum and O. François, "Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance," *Systematic Biology*, vol. 55, no. 4, pp. 685–691, 2006.

[4] D. J. Ford, "Probabilities on cladograms: Introduction to the alpha model," PhD Thesis (Stanford University), https://arxiv.org/abs/math/0511246.

[5] S. Keller-Schmidt, M. Tuğrul, V. M. Eguíluz, E. Hernández-García, and K. Klemm, "Anomalous scaling in an age-dependent branching model," *Physical Review E: Statistical, Nonlinear, and Soft Matter Physics*, vol. 91, no. 2, Article ID 022803, 2015.

[6] M. Kirkpatrick and M. Slatkin, "Searching for Evolutionary Patterns in the Shape of a Phylogenetic Tree," *Evolution*, vol. 47, no. 4, pp. 1171–1181, 1993.

[7] L. Popovic and M. Rivas, "Topology and inference for Yule trees with multiple states," *Journal of Mathematical Biology*, vol. 73, no. 5, pp. 1251–1291, 2016.

[8] R. Sainudiin and A. Véber, "A Beta-splitting model for evolutionary trees," *Royal Society Open Science*, vol. 3, no. 5, Article ID 160016, 2016.

[9] A. O. Mooers and S. B. Heard, "Inferring evolutionary process from phylogenetic tree shape," *The Quarterly Review of Biology*, vol. 72, no. 1, pp. 31–54, 1997.

[10] J. Felsenstein, *Inferring Phylogenies*, Sinauer Associates Inc., 2004.

[11] T. M. Coronado, A. Mir, F. Rosselló, and G. Valiente, "A balance index for phylogenetic trees based on quartets," https://arxiv.org/abs/1801.05411.

[12] SageMath, the Sage Mathematics Software System (Version 7.6), The Sage Developers (2017) http://www.sagemath.org.

[13] L. L. Cavalli-Sforza and A. W. Edwards, "Phylogenetic Analysis: Models and Estimation Procedures," *Evolution*, vol. 21, no. 3, pp. 550–570, 1967.

[14] E. F. Harding, "The probabilities of rooted tree-shapes generated by random bifurcation," *Advances in Applied Probability*, vol. 3, pp. 44–77, 1971.