Theoretical Population Biology 108 (2016) 13–23

Contents lists available at ScienceDirect

Theoretical Population Biology

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

On joint subtree distributions under two evolutionary models

Taoyang Wu [a,*], Kwok Pui Choi [b,c]

[a] School of Computing Sciences, University of East Anglia, Norwich, United Kingdom
[b] Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore
[c] Department of Mathematics, National University of Singapore, Singapore 117546, Singapore

ARTICLE INFO

Article history:
Received 13 August 2015
Available online 29 November 2015

Keywords:
Phylogenetic tree
Subtree distribution
Yule–Harding–Kingman model
PDA model
Tree indices
Joint distribution

ABSTRACT

In population and evolutionary biology, hypotheses about micro-evolutionary and macro-evolutionary processes are commonly tested by comparing the shape indices of empirical evolutionary trees with those predicted by neutral models. A key ingredient in this approach is the ability to compute and quantify distributions of various tree shape indices under random models of interest. As a step to meet this challenge, in this paper we investigate the joint distribution of cherries and pitchforks (that is, subtrees with two and three leaves) under two widely used null models: the Yule–Harding–Kingman (YHK) model and the proportional to distinguishable arrangements (PDA) model. Based on two novel recursive formulae, we propose a dynamic approach to numerically compute the exact joint distribution (and hence the marginal distributions) for trees of any size. We also obtained insights into the statistical properties of trees generated under these two models, including a constant correlation between the cherry and the pitchfork distributions under the YHK model, and the log-concavity and unimodality of the cherry distributions under both models. In addition, we show that there exists a unique change point for the cherry distributions between these two models.

1. Introduction

Phylogenetic tree shapes have been utilised to test evolutionary processes (see, e.g. Mooers and Heard, 1997; Nordborg, 2001; Blum and François, 2006; Purvis et al., 2011; Stadler, 2013), and more recently, to resolve disease transmission patterns (see, e.g. Colijn and Gardy, 2014). One challenge in these approaches is the ability to compute the distributions of various tree shape indices under the models of interest, which is needed in statistical testing for calculating the p-value of the empirical shape statistics or constructing a confidential interval. Even for some relatively simple null models, this can still be a challenging task. Many current approaches are based on approximating techniques, such as Monte Carlo sampling (see, e.g. Blum and François, 2006) or Gaussian approximation (see, e.g. McKenzie and Steel, 2000), which could be computationally intensive or restricting the tests to the trees above a certain size. Therefore it is desirable to explore efficient ways of computing these distributions exactly.

Two widely used null models for generating random trees in population and evolutionary biology are the Yule–Harding–Kingman (YHK) model (Harding, 1971; Yule, 1925; Kingman, 1982) and the proportional to different arrangements (PDA) model (Aldous, 2001). Under the PDA model all rooted binary trees of the same size are chosen with the same probability (Aldous, 2001) whilst under the YHK model each tree is chosen with a probability proportional to the number of total orderings that can be assigned to its internal nodes so that the relative partial ordering derived from the tree topology is preserved.

In this paper, we are interested in the exact computation of the joint distribution for the number of subtrees under the YHK and PDA model. Here a subtree, also known as a fringe subtree in Aldous (1991), consists of a node and all its descendants. More specifically, we study the distributions of the number of cherries, subtrees with two leaves, and that of pitchforks, subtrees with three leaves. Note that this is equivalent to study the joint distributions of 2-pronged and 3-pronged nodes as defined in Rosenberg (2006), as well as the joint distributions of clades of size two and three as defined in Zhu et al. (2011).

We now describe the contents of the rest of this paper. In the next section we gather some necessary notation and background. In particular, we present a random tree generating process for realising both the YHK and PDA models as described in McKenzie

* Corresponding author.
E-mail addresses: taoyang.wu@uea.ac.uk, taoyang.wu@gmail.com (T. Wu), stackp@nus.edu.sg (K.P. Choi).

http://dx.doi.org/10.1016/j.tpb.2015.11.004
0040-5809/

and Steel (2000). In contrast to the splitting model that were used in several previous studies concerning the asymptotical distributions of subtrees (see, e.g. Chang and Fuchs, 2010), the process used here is based on iteratively attaching leaves. We therefore also collect some observations on the change of the numbers of cherries and pitchforks in a tree when an additional leaf is attached.

In Sections 3 and 4 we study subtree distributions under the YHK and the PDA models, respectively. Our main results include two novel recursive formulae on the joint distributions of cherries and pitchforks; see Theorem 1 for the one under the YHK model and Theorem 4 for the one under the PDA model. These recursions enable us to develop a dynamic approach to numerically compute the joint distributions, and hence also their marginal distributions, for trees of any size.

Rewritten in functional forms, the recursions also provide a way to compute the covariance and correlation of the joint distributions under these two models. Somewhat surprisingly, we find that under the YHK model the correlation between the cherry and the pitchfork distributions is a constant $-\sqrt{14/69}$, which is independent of the number of leaves (see, e.g., Corollary 3). In contrast to currently methods developed respectively for the two models (see, e.g. Rosenberg, 2006; Chang and Fuchs, 2010), the recursions also lead to an alternative and arguably more unified approach to compute the moments of the cherry and the pitchfork distributions, and we demonstrate this by reaffirming several results obtained in previous studies.

Using the recursions on the cherry distribution derived from the joint distribution, we obtain in Theorem 6 the exact formula for the cherry distribution under the PDA model, and derive some interesting properties for cherry distributions, including that they are log-concave and hence unimodal under both models (see Theorems 3 and 7).

In Section 5 we present a comparative study of cherry and pitchfork distributions under the YHK and PDA models. We first compare the mean and the variance of these two distributions under these two models. Then we show in Theorem 8 that there exists a unique change point when comparing cherry distributions, that is, there exists a critical value τ_n for each $n \geq 4$ such that the probability that a random tree with n leaves generated under the YHK model contains k cherries is lower than that under the PDA model if $1 < k < \tau_n$, and higher if $\tau_n < k \leq n/2$. Finally, we conclude in Section 6 with discussions and some open problems.

2. Preliminaries

For later use, we present in this section some basic notation and results concerning phylogenetic trees. Throughout this article, X denotes a finite set with $|X| = n \geq 2$.

Phylogenetic trees. A *phylogenetic tree* $T = (V(T), E(T))$ on X is a rooted tree with leaf set $L(T) = X$ such that the root has one child whilst all other vertices have either zero or two children (see Fig. 1 for an example). Note that in this paper phylogenetic trees are rooted, with their edges directed away from the root. In addition, for technical simplicity we assume without loss of generality that the root has one child (also referred to as planted phylogenetic trees by Baroni et al. (2005)). Let $E^*(T)$ be the set of pendant edges in T, i.e., those edges incident with a leaf. Then we have $|E(T)| = 2n - 1$ and $|E^*(T)| = n$.

Let e be an edge in a phylogenetic tree T. The tree consisting of e and all edges below e is called a *subtree* of T, and is denoted by $T(e)$. In particular, a *cherry* is a subtree with two leaves, and a *pitchfork* is a subtree with three leaves. The number of cherries and pitchforks contained in T are denoted by $C(T)$ and $A(T)$, respectively. Note first that we always have $1 \leq C(T) \leq n/2$ and $0 \leq A(T) \leq n/3$. Moreover, in our definition a cherry contains three edges and a pitchfork contains five edges. As an example, for the tree T depicted in Fig. 1 we have $C(T) = 2$ and $A(T) = 1$. In addition, $T(e_8)$ is a pitchfork with edge set $\{e_1, e_3, e_5, e_7, e_8\}$, and $T(e_7)$ is a cherry with edge set $\{e_1, e_5, e_7\}$. Finally, $C(T)$ and $A(T)$ are respectively equal to the number of 2-pronged nodes and 3-pronged nodes contained in T (see Rosenberg (2006) for the definitions of r-pronged nodes).

Given an edge e in a phylogenetic tree T and a taxon $x_0 \notin L(T)$, let $T[e; x_0]$ be the phylogenetic tree obtained from T by attaching a new leaf labelled with x_0 to the edge e. Formally, let $e = \{u, v\}$ and let w be a vertex not contained in $V(T)$, then $T[e; x_0]$ has vertex set $V(T) \cup \{x_0, w\}$ and edge set $\big(E(T) \setminus \{e\}\big) \cup \{(u, w), (v, w), (w, x_0)\}$ (see Fig. 1 for an illustration of this construction). When the labelling of the new leaf is clear from the context, $T[e; x_0]$ is abbreviated to $T[e]$.

The YHK and the PDA model. In this subsection, we present a formal definition of the two null models investigated in this paper: the *proportional to distinguishable arrangements* (PDA) model and the *Yule–Harding–Kingman* (YHK) model. In contrast to the splitting process used by Aldous (2001) to accommodate the two models, the random process used here is based on iteratively attaching leaves.

Under the Yule–Harding model (Harding, 1971; Yule, 1925), a rooted phylogenetic tree on X is generated as follows. Beginning with the tree with two leaves, we "grow" it by repeatedly uniformly sampling a pendant edge e in the current tree T_{cur} and replace T_{cur} by $T_{cur}[e]$. This process continues until a binary tree with n leaves is obtained. Finally, we label each of its leaves with a label sampled randomly uniformly (without replacement) from $\{x_1, \ldots, x_n\}$. When branch lengths are ignored, the Yule–Harding model is shown by Aldous (1996) to be equivalent to the trees generated by the coalescent process, a backward tree generating process that is widely used in population genetics (Kingman, 1982), and so we call it the YHK model. The probability of generating a tree T under this model is denoted by $\mathbb{P}_y(T)$.

Let $\mathcal{T}_n$ be the set of phylogenetic trees with leaf set $\{x_1, \ldots, x_n\}$. It is well known that the number of trees contained in $\mathcal{T}_n$ is $\varphi(n) := (2n-3)!! = 1 \times 3 \times \cdots \times (2n-3)$ (see e.g. Semple and Steel, 2003). Here we adopt the convention that $\varphi(1) = 1$. Under the PDA model, each tree has the same probability, that is, $1/\varphi(n)$, to be generated. Alternatively, a tree can be generated under the PDA model using a Markov process similar to the one used in the YHK model; the only difference is that the edge e is uniformly sampled from $E(T)$, instead of $E^*(T)$ (see, e.g., McKenzie and Steel, 2000). We use $\mathbb{E}_y$, $\mathbb{V}_y$, Cov_y and ρ_y to denote respectively the expectation, variance, covariance and correlation taken with respect to the probability measure $\mathbb{P}_y$ under the YHK model. Similarly, $\mathbb{E}_u$, $\mathbb{V}_u$, Cov_u and ρ_u are defined with respect to the probability $\mathbb{P}_u$ under the PDA model.

For $n \geq 2$, let A_n (resp. C_n) be the random variable $A(T)$ (resp. $C(T)$) for a random tree T in $\mathcal{T}_n$. In this paper, we are interested in the joint distributions and the marginal properties of A_n and C_n under the YHK and the PDA models.

Subtree pattern. For later use, we present in this subsection several technical results concerning the change of the numbers of cherries and pitchforks when a new leaf is attached to a phylogenetic tree.

We begin with the following notation. Given a phylogenetic tree T, let $E_1(T)$ be the set of pendant edges that are contained in a pitchfork but not a cherry; $E_2(T)$ the set of edges in T that are contained in a cherry but not in a pitchfork (note that in our notation a cherry contains three leaves); $E_3(T)$ the set of pendant edges that are contained in neither a cherry nor a pitchfork; and $E_4(T) = E(T) \setminus (E_1(T) \cup E_2(T) \cup E_3(T))$. For instance, for the tree T depicted in Fig. 1, we have $E_1(T) = \{e_3\}$, $E_2(T) = \{e_2, e_4, e_9\}$, $E_3(T) = \{e_6\}$ and $E_4(T) = \{e_0, e_1, e_5, e_7, e_8, e_{10}\}$. In addition, $E(T)$ can be decomposed into the disjoint union of these four sets of edges. The