Journal of Theoretical Biology 584 (2024) 111794

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Distributions of 4-subtree patterns for uniform random unrooted phylogenetic trees

Kwok Pui Choi [a], Gursharn Kaur [b], Ariadne Thompson [c], Taoyang Wu [c,*]

[a] Department of Statistics and Data Sciences, National University of Singapore, Singapore 117546, Singapore
[b] Biocomplexity Institute, University of Virginia, Charlottesville, 22911, USA
[c] School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK

ARTICLE INFO

Keywords:
Tree shape
Joint subtree distributions
Pólya urn model
Limit distributions
One branch at a time model
PDA model

ABSTRACT

Tree shape statistics based on peripheral structures have been utilized to study evolutionary mechanisms and inference methods. Partially motivated by a recent study by Pouryahya and Sankoff on modeling the accumulation of subgenomes in the evolution of polyploids, we present the distribution of subtree patterns with four or fewer leaves for the unrooted Proportional to Distinguishable Arrangements (PDA) model. We derive a recursive formula for computing the joint distributions, as well as a Strong Law of Large Numbers and a Central Limit Theorem for the joint distributions. This enables us to confirm several conjectures proposed by Pouryahya and Sankoff, as well as provide some theoretical insights into their observations. Based on their empirical datasets, we demonstrate that the statistical test based on the joint distribution could be more sensitive than those based on one individual subtree pattern to detect the existence of evolutionary forces such as whole genome duplication.

1. Introduction

Tree shape indices have been providing a useful tool for statistical analysis in phylogenetic studies, including testing evolutionary models, assessing the impact of selection, understanding tumor evolution (see, e.g. a recent book by Fischer et al. (2023) and the references therein). One general approach is to compare the distribution of a given tree shape index calculated from real datasets with that computed from random tree generating models, typically as a null model associated with the evolutionary processes under investigation. For instance, in a recent study Pouryahya and Sankoff (2022) adopted this approach to decipher the evolution history of multiple polyploidizations of flowering plants, with a focus on understanding the sequence of polyploidization events leading to the accumulation of current genomes. To this end, the authors proposed a null hypothesis based on the following two postulates: first, each increment in genome ploidy from $N-1$ to N is the result of adding a complement set of chromosomes to the $N-1$ copies already present; secondly, this subgenome addition is modeled as attaching a new edge to a randomly chosen branch of the existing phylogeny, which will be referred to here as the "one-at-a-time" model on unrooted trees. To test this hypothesis, they developed several statistical tests based on shape indices derived from counting peripheral subtree patterns, partially motivated by the rationale that these peripheral structures might represent events that occurred more recently and hence less subject to obscuration imposed by subsequent evolutionary processes.

Although various distributional properties are known for subtree patterns of rooted phylogenetic trees in which a specific node is designated as the root (see, e.g. McKenzie and Steel, 2000; Rosenberg, 2006; Chang and Fuchs, 2010; Disanto and Wiehe, 2013; Hagen et al., 2015; Wu and Choi, 2016; Plazzotta and Colijn, 2016), relatively less is known for unrooted trees. One subtle difference here is that a rooted tree inherits an intrinsic temporal direction derived from the root, and hence can be analyzed in a recursive manner, which is not available for unrooted trees. Furthermore, the differences between distributions of shape statistics from rooted trees and those from unrooted trees could also be intrinsically related to the tree generating models (Choi et al., 2020).

The starting point of this paper is to notice that the "one-at-a-time" model used by Pouryahya and Sankoff is closely related to the well-known uniform random tree model, which is also known as the Proportional to Distinguishable Arrangements (in short, PDA) model in phylogenetic analysis. Indeed, both models induce the same probability distribution on the tree shapes of a given number of leaves, when the taxon labels of trees sampled from the PDA model are ignored.

* Corresponding author.
E-mail addresses: stackp@nus.edu.sg (K.P. Choi), gursharn@virginia.edu (G. Kaur), Ariadne.Thompson@uea.ac.uk (A. Thompson), taoyang.wu@uea.ac.uk (T. Wu).

https://doi.org/10.1016/j.jtbi.2024.111794
Received 28 October 2023; Received in revised form 10 March 2024; Accepted 13 March 2024
Available online 16 March 2024

Furthermore, the peripheral subtree patterns they studied are those with four or fewer leaves. For two leaves, there is a unique pattern termed by Pouryahya and Sankoff as paired terminals which is also commonly known as cherries. For three leaves, there is also a unique pattern termed as triples and often known as pitchforks. However, there are two types of subtree patterns with four leaves: one contains two cherries which was termed as quadruples of type I and will be referred to as balanced quartet subtree here; the other contains precisely one cherry which was called quadruples of type II and will be referred to as asymmetric quartet subtree here. Using the first letters from their names, we let A_n, B_n, P_n and C_n denote, respectively, the numbers of **a**symmetric quartets, **b**alanced quartets, **p**itchforks, and **c**herries in a random unrooted tree with n leaves sampled from the PDA model. Since these numbers depend only on tree shapes, but not on the taxon labels, they have the same probability distributions as the subtree patterns in the one-at-a-time model. Based on recurrence relations, Pouryahya and Sankoff (2022) computed the probabilities of these random variables for small n which enabled them to conduct their tests, which also motivated them to make several conjectures regarding the asymptotic mean and variance of these random variables, see Conjectures 2.2, 2.3, 3.2, 3.3, 4.2, 4.3, 4.5 and 4.6 in their paper.

Instead of focusing on the behavior of random variables for the frequency of individual subtree patterns, here we are interested in the random vector $\mathbf{q}_n = (A_n, B_n, P_n, C_n)$ which will be referred to as the 4-subtree vector. By coupling this vector with an extended Pólya urn model associated with edge types of uniform random trees, we derive a recurrence relation of the joint probability mass function (pmf) of these vectors, which enables us to deduce the exact formula for their mean and variance–covariance matrices, and hence show that all the conjectures mentioned in the last paragraph are true. Furthermore, based on some recent results on the limiting distributions of extended Pólya urn models, we also derive a law of large numbers and a central limit theorem for the random vector $\mathbf{q}_n$: the first can be regarded as a stronger version of some conjectures mentioned above, the later explains the bell-shape curves of the four figures observed in Pouryahya and Sankoff (2022). Finally, we also demonstrate that statistical tests based on distributions of this random vector could be more sensitive than those based on its individual components, such as the detection of the possible existence of the whole genome duplication.

The rest of this paper is organized as follows. In the next section, we present some definitions concerning phylogenetic trees and the PDA processes. In Section 3, we describe an urn model associated with the PDA process. This enables us to obtain the exact distribution of the 4-subtree vector $\mathbf{q}_n$ in Section 4. We next derive the central limit theorem and the law of large numbers for $\mathbf{q}_n$ in Section 5. We conclude this paper in the final section with a discussion on statistical tests and some open problems.

2. Preliminaries

In this section, we present some basic notation and background concerning phylogenetic trees and the PDA random tree model. Throughout this paper, n is a positive integer greater than or equal to eight unless stated otherwise.

2.1. Phylogenetic trees

A *tree* $T = (V(T), E(T))$ is a connected acyclic graph with vertex set $V(T)$ and edge set $E(T)$. A vertex is referred to as a *leaf* if it has degree one, and an *interior vertex* otherwise. An edge incident to a leaf is called a *pendant edge*. All trees considered in this paper are unrooted and *binary*, that is, each interior vertex has precisely three neighbors.

In this paper, a *phylogenetic tree* on a finite set X is an unrooted binary tree with leaves bijectively labeled by the elements of X. The set of phylogenetic trees on $\{1, 2, \ldots, n\}$ is denoted by $\mathcal{T}_n$. See Fig. 1 for examples of trees in $\mathcal{T}_8$ and $\mathcal{T}_9$. Given an edge e in a phylogenetic tree T on X and a taxon $x' \notin X$, let $T[e; x']$ be the phylogenetic tree on $X \cup \{x'\}$ obtained by attaching a new leaf with label x' to the edge e. For instance, in Fig. 1, tree $T_2 = T_1[e_{12}; 9]$ is obtained from T_1 by attaching leaf 9 to the edge e_{12}. We simply use $T[e]$ instead of $T[e; x']$ when the taxon name x' is not essential.

Removing an edge in a phylogenetic tree T results in two connected components; each of which is referred to as a subtree of T, also commonly known as a fringe subtree or a peripheral subtree. A subtree is called a *cherry* if it has two leaves, and a *pitchfork* if it has three leaves. Furthermore, a subtree with four leaves is referred to as a *quartet* subtree and there are two types of quartets: an asymmetric quartet contains one cherry while a balanced quartet contains two cherries. For instance, the two subtrees of T_1 resulting from removing e_{11} in Fig. 1 are both quartets: the one with leaf set $\{1, 4, 5, 8\}$ is a balanced quartet while the other one is an asymmetric quartet. Furthermore, a cherry is referred to as *independent* if it is not contained in any pitchfork or quartet subtree. Similarly, a pitchfork is called *independent* if it is not contained in any quartet subtree. Finally, let $A(T)$, $B(T)$, $P(T)$ and $C(T)$ denote, respectively, the number of asymmetric quartets, the number of balanced quartets, the number of pitchforks, and the number of cherries for a given phylogenetic tree T.

2.2. The PDA processes

Let $\mathcal{T}_n$ be the set of unrooted phylogenetic trees with n leaves. In this subsection, we introduce the Proportional to Distinguishable Arrangements (PDA) process.

Under the PDA process, a random phylogenetic tree T_n in $\mathcal{T}_n$ is generated as follows.

(i) Start with a uniformly chosen unrooted tree with eight leaves which are labeled with the taxon set $\{1, 2, \ldots, 8\}$;
(ii) for $8 \le k \le n$, uniformly choose a random edge e in T_k and let $T_{k+1} = T_k[e; k+1]$.

Generally, the PDA process is initialized with an unrooted tree with three or four leaves, but for the purpose this paper it is easier to start with eight leaves, which does not change the distribution of sampled trees in $\mathcal{T}_n$ for $n \ge 8$. Furthermore, the above process could be used to generate unlabeled trees, also referred to as tree shapes, when the leaf labels of the resulting phylogenetic tree are discarded. In this context, the PDA process is equivalent to the 'one-branch-at-a-time' model as described in Pouryahya and Sankoff (2022) in that both models give the same probability distribution of binary unrooted unlabeled trees with $n \ge 8$ leaves. In particular, there are only four tree shapes with eight leaves, as illustrated in Fig. 2.

For $n \ge 8$, let $\mathbf{q}_n = (A_n, B_n, P_n, C_n)$ be the random vector consisting of random variables $A(T)$, $B(T)$, $P(T)$, and $C(T)$ for a random tree T in $\mathcal{T}_n$ sampled from the PDA process, which will be referred to as the *4-subtree vector* in this paper. The probability distributions of $\mathbf{q}_n$ are referred to as 4-subtree distributions (for nontrivial subtrees with four, three or two leaves). In this paper, we are interested in the distributional properties of $\mathbf{q}_n$, such as the mean $\mathbb{E}(A_n)$ and the variance $\mathbb{V}(A_n)$ of A_n, and the covariance $Cov(A_n, B_n)$ between A_n and B_n etc.

3. An urn model from edge types

To study the 4-subtree vector (A_n, B_n, P_n, C_n) in this section, we briefly describe a Pólya urn model with 8 colors that is associated with the PDA model (see, e.g. Choi et al., 2021; Kaur et al., 2023 for more details about a general urn model). Our starting point is the following edge typing scheme (as illustrated in Fig. 3) which assigns a type $j \in \{1, \ldots, 8\}$ for each edge in a phylogenetic tree with eight or more leaves. Specifically,

(E1): pendant edges in unbalanced quartet subtrees that are not contained in a pitchfork;

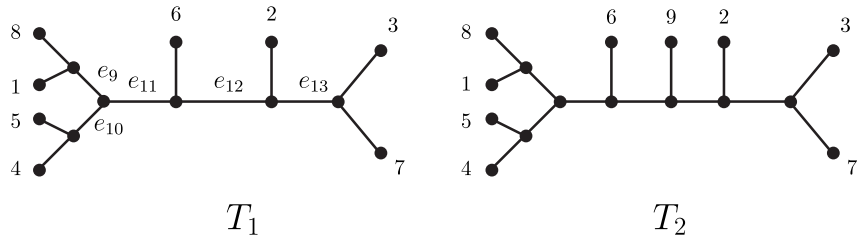

Fig. 1. Examples of phylogenetic trees. T_1 is an (unrooted) phylogenetic tree on $\{1,\ldots,8\}$; $T_2 = T_1[e_{12}]$ is a phylogenetic tree on $X = \{1,\ldots,9\}$ obtained from T_1 by attaching a new leaf labeled 9 to the edge e_{12}. The shape of a tree when the labeling of the leaves are ignored is also referred to as an unlabeled tree.

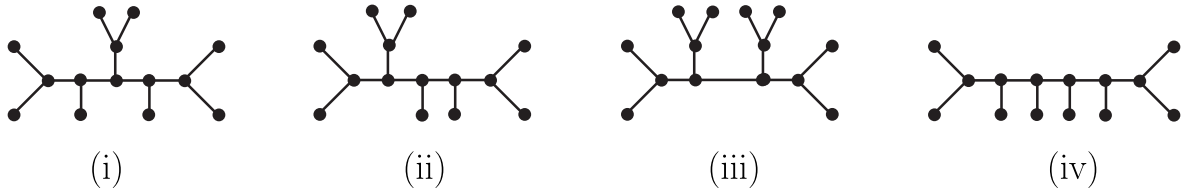

Fig. 2. The four tree shapes with eight leaves. Note that (ii) is the tree shape of T_1 in Fig. 1 obtained by removing all leave lables. Furthermore, two distinct phylogenetic trees may have the same tree shape.

Table 1
Frequency vectors and distributions for the four tree shapes with eight leaves as presented in Fig. 2. The column below $\mathbf{q}(T)$ contains the 4-subtree vectors for the four tree shapes. The probability distribution below $\mathbb{P}(T)$ are computed under the PDA model. The columns below PS and ASTRAL are, respectively, the counts of the four tree shapes among 10 total empirical tree shapes obtained by Pouryahya and Sankoff (2022) in their Table 3 and Table 4, which will be further discussed here in Section 6.

	$\mathbf{q}(T)$	$\mathbb{P}(T)$	PS	ASTRAL
(i)	(0,0,2,3)	8/33	6	5
(ii)	(1,1,1,3)	8/33	1	0
(iii)	(0,2,0,4)	1/33	3	3
(iv)	(2,0,2,2)	16/33	0	2

(E2): pendant edges in unbalanced quartet subtrees that are not in a cherry and not of type 1;
(E3): edges in balanced quartet subtrees;
(E4): pendant edges in independent pitchforks that are not contained in a cherry;
(E5): edges contained in or incident with independent pitchforks that are not of type 4;
(E6): edges contained in or incident with independent cherries;
(E7): pendant edges that are contained in neither a cherry nor a pitchfork;
(E8): edges that are not of type 1–7.

For $1 \le j \le 8$, let $E_j(T)$ be the set of edges of type j. Then each edge e in T belongs to one and only one $E_j(T)$ when T has eight or more leaves. Furthermore, consider the *characteristic map* χ that maps each edge in e in T to the jth canonical row vector with 8 dimensions if e is contained in $E_j(T)$. For instance, for each edge e contained in $E_2(T)$, we have $\chi(e) = (0,1,0,0,0,0,0,0)$. Finally, let $\beta(T) = \left(|E_1(T)|,\ldots,|E_8(T)|\right)$ be the type vector associated with T, where $|E_j(T)|$ counts the number of type j edges in T. Then, we have the following observation.

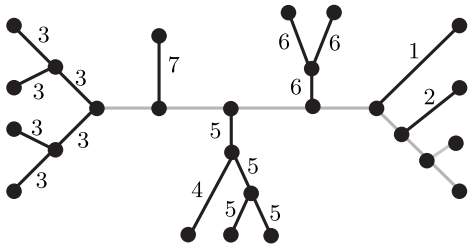

Fig. 3. A phylogenetic tree with edge types indicated. The type of an edge is indicated by the number next to it. For the eight edges in gray representing type 8, the number 8 is omitted for simplicity.

Lemma 1. *Suppose that T is a phylogenetic tree with $n \ge 8$ leaves with its associated 4-subtree vector $\mathbf{q}(T) = (a,b,p,c)$. Then, we have*

$$\beta(T) = (a, a, 6b, p-a, 4(p-a), 3(c-p-2b), n-2c-p-a, n+4a-p-c-3). \tag{1}$$

Proof. Each asymmetric quartet subtree contains precisely one edge of type 2 and one edge of type 3, so $|E_1(T)| = |E_2(T)| = a$. As each balanced quartet subtree contains six edges, we have $|E_3(T)| = 6b$. Next, a pitchfork is not independent if and only if it is contained in an asymmetric quartet subtree. Hence, there are $p-a$ independent pitchforks, each of which contributes one type 4 edge and four type 5 edges. Thus we have $|E_4(T)| = p-a$ and $|E_5(T)| = 4(p-a)$. Next, each cherry that is not independent is contained in either a pitchfork or a balanced quartet subtree. This implies $|E_6(T)| = 3(c-p-2b)$ as each independent cherry contributes to three type 6 edges. Furthermore, there are n pendant edges, which are independent if it is not contained in a cherry or a pitchfork or an asymmetric quartet subtree. This implies $|E_7(T)| = n-2c-p-a$. Finally, since there are $2n-3$ edges in T, we have

$$|E_8(T)| = (2n-3) - \sum_{1\le j\le 7} |E_j(T)| = n+4a-p-c-3. \quad \square$$

Now consider a PDA process starting with an unrooted tree T_8 with eight leaves. Then, we associate it with an urn $\{Z_m\}_{m\ge 0}$ with balls of 8

Table 2
The difference $\mathbf{q}(T[e]) - \mathbf{q}(T)$ of the 4-subtree vectors according to the type of e. For each type i of e, this difference is given by the corresponding column below i in the head row. For instance, if e is of type 1, then the difference is $(-1, 0, 0, 1)$.

	1	2	3	4	5	6	7	8
a	−1	−1	0	0	1	0	0	0
b	0	1	−1	1	0	0	0	0
p	0	−1	1	−1	0	1	0	0
c	1	1	0	1	0	0	1	0

different colors containing $Z_{0,j} = |E_j(T_8)|$ balls of color $j \in \{1, 2, \dots, 8\}$ at time 0. In the associated urn, at each time step, a ball is drawn uniformly at random and returned with some extra balls, depending on the color selected and the replacement scheme R as below:

$$R = \begin{bmatrix} -1 & -1 & 0 & 1 & 4 & 3 & 0 & -4 \\ -1 & -1 & 6 & 0 & 0 & 0 & 1 & -3 \\ 0 & 0 & -6 & 1 & 4 & 3 & 0 & 0 \\ 0 & 0 & 6 & -1 & -4 & 0 & 0 & 1 \\ 1 & 1 & 0 & -1 & -4 & 0 & 0 & 5 \\ 0 & 0 & 0 & 1 & 4 & -3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}. \tag{2}$$

More precisely, if a ball with color i is drawn, then we return the selected ball along with additional R_{ij} balls of color j, for every $j \in \{1, 2, \dots, 8\}$. For instance, when a ball of color 3 is drawn, we return the selected ball, add one ball of color 4, four balls of color 5, three balls of color 6, and remove six balls of color 3. Note that at time step k, the number $Z_{k,j}$ of color j balls in our urn counts the number of type j edges in the tree T_{8+k}. Furthermore, the number of type j edges in tree $T[e]$, which is obtained from $T = T_{8+k}$ by attaching an extra leaf to edge e of type i, is precisely $Z_{k,j} + R_{i,j}$. In other words, the urn model is 'coupled' with the PDA model in the sense that the dynamic of the edge count vector $\beta(T_n)$ associated with the PDA process $\{T_n\}_{n\geq 8}$ is precisely described by the associated urn model $\{Z_{k,j}\}_{k\geq 0, 1\leq j\leq 8}$. Moreover, we have the following observation, which can be verified directly by checking each type of the selected edge, and hence the proof is omitted.

Lemma 2. *Suppose that T is a phylogenetic tree with eight or more leaves. Then we have*

$$\beta(T[e]) = \beta(T) + \chi(e)R. \tag{3}$$

Furthermore, note that the urn models constructed here are *tenable*, that is, at each step it is always possible to add or remove balls according to the replacement matrix R.

4. Exact distributions

In this section, we study the exact distribution of the 4-subtree vector $\mathbf{q}_n = (A_n, B_n, P_n, C_n)$, a random vector counting the different types of subtrees with four or fewer leaves. To this end, we start with deriving a recursion for computing the probability distribution of this random vector based on the urn model described in Section 3.

Recall that the edge type vector $\beta(T)$ associated with a tree T is closely related to the 4-subtree vector $\mathbf{q}$. More precisely, the number of asymmetric quartet trees is given by $E_1(T) = E_2(T)$, and that of balanced quartet subtree by $E_3(T)/6$, the pitchfork by $E_1(T) + E_4(T)$, and the cherries by $E_1(T) + E_3(T)/3 + E_4(T) + E_6(T)$. Therefore, from the replacement matrix R in (2) we can construct the following table detailing the changes of 4-subtree vector $\mathbf{q}(T[e]) - \mathbf{q}(T)$ according to the type of e.

Using Table 2, we now derive the following recursive formula on the probability distributions of the random 4-subtree vector under the PDA model. Note that similar formulas have been derived for subsets of the entries in the random vector (see, e.g. Theorem 1 in Choi et al., 2020 and Propositions 4.2 in Pouryahya and Sankoff, 2022).

Theorem 1. *Let $\sigma_n(a, b, p, c) = \mathbb{P}(\mathbf{q}(T_n) = (a, b, p, c))$ under the PDA model. Then for $n \geq 8$ we have*

$$\begin{aligned}\sigma_{n+1}(a, b, p, c) = {} & \frac{a+1}{2n-3}\sigma_n(a+1, b, p, c-1) + \frac{a+1}{2n-3}\sigma_n(a+1, b-1, \\ & p+1, c-1) + \frac{6b+6}{2n-3}\sigma_n(a, b+1, p-1, c) \\ & + \frac{p-a+1}{2n-3}\sigma_n(a, b-1, p+1, c-1) + \frac{4(p-a+1)}{2n-3} \\ & \times \sigma_n(a-1, b, p, c) + \frac{3(c-p-2b+1)}{2n-3}\sigma_n(a, b, p-1, c) \\ & + \frac{n-a-p-2c+2}{2n-3}\sigma_n(a, b, p, c-1) \\ & + \frac{n+4a-p-c-3}{2n-3}\sigma_n(a, b, p, c). \end{aligned} \tag{4}$$

Moreover, $\sigma_8(a, b, p, c)$ is $8/33$ if $(a, b, p, c) = (0, 0, 2, 3)$ or $(a, b, p, c) = (1, 1, 1, 3)$, $16/33$ if $(a, b, p, c) = (2, 0, 2, 2)$, $1/33$ if $(a, b, p, c) = (0, 2, 0, 4)$, and 0 otherwise.

Proof. For $1 \leq i \leq 8$, let ϵ_i be the row vector constructed from the column under i in Table 2 and $\epsilon'_i = (a, b, p, c) - \epsilon_i$. For instance, we have $\epsilon_1 = (-1, 0, 0, 1)$ and $\epsilon'_1 = (a+1, b, p, c-1)$. Let $\{T_k\}_{k\geq 8}$ be a sample path of the unrooted tree sampled from the PDA process. Let e be the edge sampled at time point n, that is, $T_{n+1} = T[e]$. Then we have

$$\begin{aligned}\sigma_{n+1}(a, b, p, c) &= \mathbb{P}\left(\mathbf{q}(T_{n+1}) = (a, b, p, c)\right) = \sum_{i=1}^{8} \mathbb{P}(\mathbf{q}(T_n) \\ &= (a, b, p, c) - \epsilon_i, \chi(e) = \mathbf{e}_i) \\ &= \sum_{i=1}^{8} \mathbb{P}(\chi(e) = \mathbf{e}_i \mid \mathbf{q}(T_n) = \epsilon'_i)\mathbb{P}(\mathbf{q}(T_n) = \epsilon'_i) \\ &= \sum_{i=1}^{8} \frac{|E_i(T_n)|}{|E(T_n)|}\mathbb{P}(\mathbf{q}(T_n) = \epsilon'_i). \end{aligned} \tag{5}$$

When $i = 1$, from Lemma 1 and Table 2 the corresponding summation factor for index 1 in the last summation of (5) is given by

$$\begin{aligned}\frac{|E_1(T)|}{|E(T)|}\mathbb{P}(\mathbf{q}(T_n) = (a, b, p, c) - \epsilon_1) &= \frac{a+1}{2n-3}\mathbb{P}(\mathbf{q}(T_n) = (a+1, b, p, c-1)) \\ &= \frac{a+1}{2n-3}\sigma_n(a+1, b, p, c-1).\end{aligned}$$

Now the theorem follows by using a similar approach to work out each term corresponding to index $i \in \{2, \dots, 8\}$ in the last summation of (5). □

For two positive integer n and k, we define the falling factorial $n^{\underline{k}}$ and the double falling factorial $n^{\underline{\underline{k}}}$ as

$$n^{\underline{k}} = \prod_{i=0}^{k-1}(n-i) = n(n-1)(n-2)\dots(n-k+1), \quad \text{and}$$
$$n^{\underline{\underline{k}}} = \prod_{i=0}^{k-1}(n-2i) = n(n-2)(n-4)\dots(n-2k+2).$$

It is known that under the unrooted PDA model (see, e.g. Choi et al., 2020), for $n \geq 6$ we have

$$\begin{aligned}\mathbb{E}(P_n) &= \frac{n^{\underline{3}}}{2(2n-5)^{\underline{\underline{2}}}}, \quad \mathbb{E}(C_n) = \frac{n^{\underline{2}}}{2(2n-5)}, \\ \mathbb{E}(P_n^2) &= \frac{n(n^5 - 7n^4 - 19n^3 + 229n^2 - 480n + 276)}{4(2n-5)^{\underline{4}}}, \\ \mathbb{E}(P_nC_n) &= \frac{n(n^4 - 6n^3 + 5n^2 + 12n - 12)}{4(2n-5)^{\underline{\underline{3}}}}, \quad \mathbb{E}(C_n^2) = \frac{n^{\underline{2}}(n^2 - n - 8)}{4(2n-5)^{\underline{\underline{2}}}}. \end{aligned} \tag{6}$$

Theorem 2. *Under the PDA model, for $n \geq 8$ the random vector $\mathbf{q}_n = (A_n, B_n, P_n, C_n)$ has the mean*

$$\mathbb{E}(\mathbf{q}_n) = \left(\frac{n^{\underline{4}}}{2(2n-5)^{\underline{3}}}, \frac{n^{\underline{4}}}{8(2n-5)^{\underline{3}}}, \frac{n^{\underline{3}}}{2(2n-5)^{\underline{2}}}, \frac{n^{\underline{2}}}{2(2n-5)} \right), \tag{7}$$

and the following entries in the variance–covariance matrix:

$$\begin{aligned}
\mathbb{V}(A_n) &= \frac{3n^{\underline{4}}(12n^6 - 384n^5 + 5013n^4 - 34006n^3 + 125715n^2 - 238730n + 181125)}{2(2n-5)^{\underline{3}}(2n-5)^{\underline{6}}},\\
Cov(A_n, B_n) &= \frac{-n^{\underline{4}}(28n^6 - 768n^5 + 8401n^4 - 46782n^3 + 139891n^2 - 214170n + 132300)}{8(2n-5)^{\underline{3}}(2n-5)^{\underline{6}}},\\
Cov(A_n, P_n) &= \frac{n^{\underline{4}}(4n^4 - 90n^3 + 736n^2 - 2520n + 2905)}{2(2n-5)^{\underline{2}}(2n-5)^{\underline{5}}},\\
Cov(A_n, C_n) &= \frac{-n^{\underline{4}}(n^2 - 5n - 5)}{2(2n-5)(2n-5)^{\underline{4}}},\\
\mathbb{V}(B_n) &= \frac{3n^{\underline{5}}(76n^5 - 1924n^4 + 18833n^3 - 88641n^2 + 199546n - 171360)}{32(2n-5)^{\underline{3}}(2n-5)^{\underline{6}}},\\
Cov(B_n, P_n) &= \frac{-3n^{\underline{4}}(2n^4 - 33n^3 + 188n^2 - 432n + 350)}{4(2n-5)^{\underline{2}}(2n-5)^{\underline{5}}},\\
Cov(B_n, C_n) &= \frac{3n^{\underline{6}}}{8(2n-5)(2n-5)^{\underline{4}}},\\
\mathbb{V}(P_n) &= \frac{3n^{\underline{3}}(4n^4 - 76n^3 + 527n^2 - 1555n + 1610)}{4(2n-5)^{\underline{2}}(2n-5)^{\underline{4}}},\\
Cov(P_n, C_n) &= \frac{-3n^{\underline{3}}(n-5)}{2(2n-5)(2n-5)^{\underline{3}}},\\
\mathbb{V}(C_n) &= \frac{n^{\underline{2}}(n-4)^{\underline{2}}}{2(2n-5)(2n-5)^{\underline{2}}},
\end{aligned}$$

where $\mathbb{V}(A_n)$ denotes the variance of A_n, and $Cov(A_n, B_n)$ the covariance between A_n and B_n.

The proof of this theorem is rather technical. For better flow of presentation of our results, it is presented in the Appendix. We end this section by noting that this theorem provides affirmative answers to the conjectures (i.e. Conjectures 2.2, 2.3, 3.2, 3.3, 4.2, 4.3, 4.5 and 4.6) of Pouryahya and Sankoff (2022), whose proof is straightforward and hence omitted.

Corollary 1. *Under the PDA model, we have*

$$\lim_{n\to\infty} \left(\frac{\mathbb{E}[A_n]}{n}, \frac{\mathbb{E}[B_n]}{n}, \frac{\mathbb{E}[P_n]}{n}, \frac{\mathbb{E}[C_n]}{n} \right) = \left(\frac{1}{16}, \frac{1}{64}, \frac{1}{8}, \frac{1}{4} \right),$$

and

$$\lim_{n\to\infty} \left(\frac{\mathbb{V}[A_n]}{\mathbb{E}[A_n]}, \frac{\mathbb{V}[B_n]}{\mathbb{E}[B_n]}, \frac{\mathbb{V}[P_n]}{\mathbb{E}[P_n]}, \frac{\mathbb{V}[C_n]}{\mathbb{E}[C_n]} \right) = \left(\frac{9}{16}, \frac{57}{64}, \frac{3}{8}, \frac{1}{4} \right).$$

5. Limiting distributions

Consider the Urn model $Z_n = (Z_{n,1}, \ldots, Z_{n,8})$ associated with the PDA model as discussed in Section 3. Then the limiting distribution of the 4-subtree counting vector can be deduced from the limiting distribution of the urn model. To this end, we recall some standard notation in linear algebra. That is, for a vector Z, let $Z^\top$ denote the transpose of Z; for a matrix U is a matrix, let U^{-1} denote the inverse of U. Next, we notice that the replacement matrix R as in (2) has the following eigenvalues (counted with multiplicity)

$$\lambda_1 = 2, \quad \lambda_2 = \lambda_3 = \lambda_4 = 0, \quad \lambda_5 = -2, \quad \lambda_6 = -4, \quad \lambda_7 = \lambda_8 = -6, \tag{8}$$

where $\lambda_1 = 2$ will be referred to as the principal eigenvalue. Furthermore, the replacement matrix R also satisfies the following four technical conditions:

(A1) *Tenable:* It is always possible to draw balls and follow the replacement rule, that is, we never get stuck in following the rules (see, e.g. Mahmoud, 2009, p.46).

(A2) *Small:* That is, by the eigenvalues in (8) it follows that all eight eigenvalues of R are real; the maximal eigenvalue $\lambda_1 = 2$ is positive and $\lambda_1 > 2\lambda$ holds for each of the other seven eigenvalues λ of R.

(A3) *Strictly Balanced:* The column vector $(1, 1, \ldots, 1)^\top$ (whose entries are all one) is a right eigenvector of R corresponding to λ_1 and $\mathbf{v}_1 = \frac{1}{64}(3, 2, 2, 2, 8, 9, 10, 28)$ is a left eigenvector corresponding to λ_1 is a stochastic vector.

(A4) *Diagonalizable:* R is diagonalizable over real numbers.

To see that condition (A4) holds, consider the diagonal matrix $\Lambda = \mathrm{diag}(2, 0, 0, 0, -2, -4, -6, -6)$ whose non-diagonal elements are 0 and diagonal elements are the eigenvalues $2, 0, 0, 0, -2, -4, -6, -6$. Then R is diagonalizable as

$$U^{-1} R U = \Lambda$$

holds with

$$U = \begin{bmatrix}
1 & -18 & 0 & -1 & -3 & -5 & -6 & -1\\
1 & 0 & 0 & 1 & -3 & -9 & -34 & -5\\
1 & -1 & 0 & 0 & 1 & 5 & 29 & 4\\
1 & -3 & -4 & 0 & -3 & -9 & -35 & -4\\
1 & 0 & 1 & 0 & 1 & 1 & 0 & 1\\
1 & -1 & 0 & 0 & 1 & 5 & \frac{35}{3} & 0\\
1 & -3 & 0 & 0 & -3 & -5 & -7 & 0\\
1 & 3 & 0 & 0 & 1 & 1 & 1 & 0
\end{bmatrix}, \tag{9}$$

and

$$U^{-1} = \frac{1}{192} \times \begin{bmatrix}
6 & 6 & 9 & 6 & 24 & 27 & 30 & 84\\
-8 & -8 & -12 & -4 & -16 & 0 & 20 & 28\\
8 & 8 & -36 & -44 & 16 & 0 & 28 & 20\\
-40 & 152 & 180 & 28 & 112 & 0 & -140 & -292\\
36 & 36 & 54 & 0 & 0 & -90 & -120 & 84\\
-24 & -24 & -36 & 12 & 48 & 72 & 36 & -84\\
6 & 6 & 9 & -6 & -24 & -9 & -6 & 24\\
-26 & -26 & 9 & 26 & 104 & -9 & 26 & -104
\end{bmatrix}. \tag{10}$$

The theorem below describes the asymptotic behavior of $\beta(T_n)$, which enables us to deduce the asymptotic properties of the distribution of the 4-subtree vector which consists of the numbers of quartet subtrees, pitchforks and the cherries for the PDA model in Corollary 2. Here $\xrightarrow{a.s.}$ means *almost sure convergence*; while $\xrightarrow{d}$ means *convergence in distribution*, also known as *weak convergence* (see, e.g. Grimmett and Stirzaker, 2001, Section 7.2 for more details on these two modes of convergence).

Theorem 3. *Suppose that T_m is an arbitrary phylogenetic tree with m leaves with $m \geq 2$, and that T_n is a tree with n leaves generated by the PDA process starting with T_m. Then we have*

$$\frac{\beta(T_n)}{n} \xrightarrow{a.s.} 2\mathbf{v}_1 \quad \text{and} \quad \frac{\beta(T_n) - 2n\mathbf{v}_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{11}$$

as $n \to \infty$, where $\mathbf{v}_1 = (3, 2, 2, 2, 8, 9, 10, 28)/64$ and $\mathcal{N}(\mathbf{0}, \Sigma)$ is the multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix

$$\Sigma = \frac{1}{1024} \times \begin{bmatrix} 36 & 36 & -42 & -20 & -80 & -54 & -20 & 144 \\ 36 & 36 & -42 & -20 & -80 & -54 & -20 & 144 \\ -42 & -42 & 513 & -30 & -120 & -81 & -30 & -168 \\ -20 & -20 & -30 & 52 & 208 & -18 & -44 & -128 \\ -80 & -80 & -120 & 208 & 832 & -72 & -176 & -512 \\ -54 & -54 & -81 & -18 & -72 & 657 & -114 & -264 \\ -20 & -20 & -30 & -44 & -176 & -114 & 308 & 96 \\ 144 & 144 & -168 & -128 & -512 & -264 & 96 & 688 \end{bmatrix}. \tag{12}$$

Proof. Consider the PDA process $\{T_n\}_{n\geq 8}$ starting with an initial phylogenetic tree T_8 with eight leaves. Let $Z_k = \beta(T_{k+8})$ for $k \geq 0$. Then $Z_k = (Z_{k,1}, \ldots, Z_{k,8})$, where $Z_{k,i} = |E_i(T_{m+k})|$ for $1 \leq i \leq 8$, is the urn model with 8 colors derived from the edge partition of the PDA process. Therefore, it is a tenable model with $Z_0 = \beta(T_8)$ and replacement matrix R as given in (2).

Since (A1)–(A4) are satisfied by the replacement matrix R, by Theorem 1 in Choi et al. (2021) and the fact that $\lambda_1 = 2$ it follows that

$$\frac{Z_k}{k} \xrightarrow{a.s.} 2\mathbf{v}_1 \quad \text{with} \quad k \to \infty, \quad \text{and hence} \quad \frac{\beta(T_n)}{n} = \frac{(n-8)}{n} \frac{Z_{n-8}}{(n-8)} \xrightarrow{a.s.} \mathbf{v}_1 \quad \text{with} \quad n \to \infty.$$

For $1 \leq j \leq 8$, let $\mathbf{e}_j$ denote the jth canonical row vector whose jth entry is 1 while the other entries are all zero. Considering the matrix U in (9), then $\mathbf{u}_j = U\mathbf{e}_j^\top$ denote the jth column of U, and $\mathbf{v}_j = \mathbf{e}_j U^{-1}$ the jth row of U^{-1}. Then by Theorem 2 in Choi et al. (2021) and $k = n-8$ we have

$$\frac{Z_{n-8} - 2(n-8)\mathbf{v}_1}{\sqrt{n-8}} = \frac{Z_k - 2k\mathbf{v}_1}{\sqrt{k}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma), \tag{13}$$

where

$$\Sigma = \sum_{i,j=2}^{8} \frac{\lambda_i \lambda_j \mathbf{u}_i^\top diag(\mathbf{v}_1)\mathbf{u}_j}{1 - \lambda_i - \lambda_j} \mathbf{v}_i^\top \mathbf{v}_j \tag{14}$$

is presented in (12). Therefore, we have

$$\frac{\beta(T_n) - n\mathbf{v}_1}{\sqrt{n}} = \frac{Z_{n-8} - (n-8)\mathbf{v}_1}{\sqrt{n}} + \frac{8\mathbf{v}_1}{\sqrt{n}} = \frac{\sqrt{n-8}}{\sqrt{n}} \frac{Z_{n-8} - (n-8)\mathbf{v}_1}{\sqrt{n-8}} + \frac{8\mathbf{v}_1}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma).$$

□

By Theorem 3, it is straightforward to obtain the following result on the distribution of the 4-subtree vector. Note that the strong law of large numbers in (15) below proves a stronger version of Conjectures 2.2, 3.2, 4.2 and 4.5 of Pouryahya and Sankoff (2022). Moreover, the central limit theorem in (16) implies that the limiting distribution of $\{(A_n, B_n, P_n, C_n) - \mathbb{E}(A_n, B_n, P_n, C_n)\}/\sqrt{n}$ is multivariate normal, which explains the bell-shape curves of Figures 2, 6, 8 and 10 observed in Pouryahya and Sankoff (2022).

Corollary 2. *Under the PDA model, for the distribution of 4-subtree vector* (A_n, B_n, P_n, C_n)*, we have*

$$\frac{1}{n}(A_n, B_n, P_n, C_n) \xrightarrow{a.s.} \left(\frac{1}{16}, \frac{1}{64}, \frac{1}{8}, \frac{1}{4}\right), \tag{15}$$

and

$$\frac{(A_n, B_n, P_n, C_n) - n\left(\frac{1}{16}, \frac{1}{64}, \frac{1}{8}, \frac{1}{4}\right)}{\sqrt{n}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \frac{1}{4096}\begin{bmatrix} 144 & -28 & 64 & -64 \\ -28 & 57 & -48 & 48 \\ 64 & -48 & 192 & 0 \\ -64 & 48 & 0 & 256 \end{bmatrix}\right), \tag{16}$$

as $n \to \infty$.

Proof. Consider the PDA process $\{T_n\}_{n\geq 8}$ starting with a tree T_8 with eight leaves. Denote the ith entry in $\beta(T_n)$ by $\beta_{n,i}$ for $1 \leq i \leq 8$. Then the corollary follows from Theorem 3 by noting that we have $A_n = \beta_{n,1}$, $B_n = \beta_{n,3}/6$, $P_n = \beta_{n,1} + \beta_{n,4}$, and $C_n = \beta_{n,1} + \beta_{n,4} + (\beta_{n,3} + \beta_{n,6})/3$. □

By Corollary 2, it follows that under the PDA model A_n/B_n, the ratio of the number of asymmetric quartet trees and that of the balanced quartet trees, converges to 4 almost surely as $n \to \infty$. In other words, asymptotically among all the quartet trees contained in a random PDA tree, four fifths are asymmetric. This is in line with the following heuristic reasoning: under the PDA model a quartet tree is formed by adding an edge uniformly to one of the five edges in or incident with a pitchfork, and only in one edge out of these five edges, the quartet tree formed is of balanced type.

6. Discussion

Pouryahya and Sankoff (2022) studied the polyploidization history of the genome of a variety of sugarcane, *Saccharum officinarum*. This genome consists of 80 chromosomes, which can be partitioned into 10 sets of eight 'homeologous' chromosomes, also known as homoeologous chromosomes (see, e.g. Glover et al., 2016) . For each set of n ($n = 8$, in this study) homeologous chromosomes, based on the gene trees inferred from the paralogous genes on each of these chromosomes, the authors proposed a method of constructing a consensus unrooted binary tree having n leaves. In total, 10 consensus trees are obtained via their method, six of which is of Shape (i) in Fig. 2, one of Shape (ii), and three of Shape (iii) (see also Table 1 where the profile is referred to as PS. See also Section 5 in Pouryahya and Sankoff (2022) for details, and their method can be applied to other polyploid genomes). Furthermore, they also obtained another set of tree shapes using the well-known ASTRAL III package (Zhang et al., 2018), which is referred to as ASTRAL in Table 1. Based on the distributions of the number of four subtrees discussed here, we conducted the exact multinomial test using the R-package EMT (R. Core Team, 2021) of the null hypothesis that these consensus trees are generated under the PDA model: one for each of the four types of subtrees and an additional one for the joint distributions. The p-values of these five statistical tests are presented in Table 3. All tests are concordant in concluding to reject the hypothesis using a threshold of p-value of 0.05. This is in line with the conclusion in Pouryahya and Sankoff (2022), that is, the accumulation of subgenomes in *Saccharum officinarum* is unlikely to be "one at a time". Indeed, the p-values based on individual subtree structures in our Table 3 agree well with those obtained in Tables 3 and 4 in Pouryahya and Sankoff (2022). However, our p-value based on the joint distribution has a smaller value than any of the four p-values obtained using the distribution of individual subtrees, indicating that the statistic based on the joint distribution is more sensitive than those based on individual ones.

The results obtained in this paper also naturally lead to several broad questions for future study. First, in this paper we investigated subtree frequencies for unrooted trees under the PDA model. It would be interesting to extend the results obtained here to random rooted trees, and also to other tree generating models, such as Ford's alpha model which includes both the PDA model and the Yule model (see,

Table 3
p-values of five statistical tests using the PDA model as the null hypothesis and the data from Pouryahya and Sankoff (2022). The column below $\mathbf{q}$ contains the p-values based on the distribution of the 4-subtree vector $\mathbf{q}_8 = (A_8, B_8, P_8, C_8)$, and the ones below A, B, P, C contain those based on the distributions of the corresponding components of $\mathbf{q}_8$.

	$\mathbf{q}$	A	B	P	C
PS	0.0000156	0.00004512	0.004341	0.004341	0.00007994
ASTRAL	0.0002672	0.002287	0.001453	0.001453	0.001625

e.g. Coronado et al., 2018; Kaur et al., 2023). Next, in addition to subtree patterns up to four leaves, it is of interest to study more general subtree patterns, such as k-caterpillars and k-pronged nodes as mentioned in Rosenberg (2003), Fuchs (2008). Finally, it would be interesting to study shape statistics for other random structures proposed for modeling evolution, such as distributions of branch lengths by Ferretti et al. (2017), ranked trees by Kim et al. (2020), and shape statistics in phylogenetic networks (see, e.g. Bienvenu et al., 2022; Stufler, 2022; Fuchs et al., 2024).

CRediT authorship contribution statement

Kwok Pui Choi: Conceptualization, Investigation, Writing – original draft. **Gursharn Kaur:** Investigation, Writing – original draft, Writing – review & editing. **Ariadne Thompson:** Investigation, Writing – original draft. **Taoyang Wu:** Conceptualization, Investigation, Project administration, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We are grateful to two anonymous reviewers for their helpful suggestions that improve the presentation of the paper. K.P. Choi acknowledges the support of Singapore Ministry of Education Academic Research Fund, Singapore [Grant number R-155-000-222-114]. G. Kaur acknowledges the support provided by NSF Expeditions in Computing Grant, USA [CCF-1918656 and CCF-191781]. A. Thompson is supported by the UKRI Biotechnology and Biological Sciences Research Council Norwich Research Park Biosciences Doctoral Training Partnership, UK [Grant number BB/M011216/1]. We thank David Sankoff for insightful discussions on modeling polyplidizations. K.P. Choi and T. Wu thank the Institute for Mathematical Sciences, National University of Singapore and the program organizers of *Mathematics of Evolution-Phylogenetic Trees and Networks* where part of this project is completed.

Appendix. Proof of Theorem 2

Proof. Let φ be an arbitrary function $\mathbb{R}^4 \to \mathbb{R}$. By Theorem 1, under the PDA model for $n \geq 8$ we have

$$\begin{aligned}(2n-3)\mathbb{E}(\varphi[\mathbf{q}_{n+1}]) &= \mathbb{E}(A_n\varphi[\mathbf{q}_n + (-1,0,0,1)])\\ &+ \mathbb{E}(A_n\varphi[\mathbf{q}_n + (-1,1,-1,1)]) + 6\mathbb{E}(B_n\varphi[\mathbf{q}_n + (0,-1,1,0)])\\ &+ \mathbb{E}((P_n - A_n)\varphi[\mathbf{q}_n + (0,1,-1,1)])+\\ &+ 4\mathbb{E}((P_n - A_n)\varphi[\mathbf{q}_n + (1,0,0,0)])\\ &+ 3\mathbb{E}((C_n - 2B_n - P_n)\varphi[\mathbf{q}_n + (0,0,1,0)])\\ &+ \mathbb{E}((n - 2C_n - P_n - A_n)\varphi[\mathbf{q}_n + (0,0,0,1)])\\ &+ \mathbb{E}((n + 4A_n - P_n - C_n - 3)\varphi[\mathbf{q}_n]). \qquad (17)\end{aligned}$$

Substituting $\varphi[(a,b,p,c)] = a$ in (17) and using $\mathbb{E}(P_n)$ in (6), we have

$$(2n-3)\mathbb{E}(A_{n+1}) = (2n-9)\mathbb{E}(A_n) + 4\mathbb{E}(P_n) = (2n-9)\mathbb{E}(A_n) + \frac{2n^{\underline{3}}}{(2n-5)^{\underline{2}}}.$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(A_8) = 40/33$ shows that $\mathbb{E}(A_n) = n^{\underline{4}}/[2(2n-5)^{\underline{3}}]$ holds for all $n \geq 8$.

Similarly, substituting $\varphi[(a,b,p,c)] = b$ in (17) leads to

$$(2n-3)\mathbb{E}(B_{n+1}) = (2n-9)\mathbb{E}(B_n) + \mathbb{E}(P_n) = (2n-9)\mathbb{E}(B_n) + \frac{n^{\underline{3}}}{2(2n-5)^{\underline{2}}}.$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(B_8) = 10/33$ shows that $\mathbb{E}(B_n) = n^{\underline{4}}/[8(2n-5)^{\underline{3}}]$ holds for all $n \geq 8$.

To establish the result on the variance–covariance matrix $K(\mathbf{q}_n, \mathbf{q}_n) = \mathbb{E}(\mathbf{q}_n^T\mathbf{q}_n) - \mathbb{E}(\mathbf{q}_n)^T\mathbb{E}(\mathbf{q}_n)$, it suffices to show that $\mathbb{E}(\mathbf{q}_n^T\mathbf{q}_n)$ is equal to (see Eq. (18) given in Box I).

To this end, substituting $\varphi[(a,b,p,c)] = ac$ in (17) and using (6), we have

$$\begin{aligned}(2n-3)\mathbb{E}(A_{n+1}C_{n+1}) &= (2n-11)\mathbb{E}(A_nC_n) + (n-2)\mathbb{E}(A_n) + 4\mathbb{E}(P_nC_n)\\ &= (2n-11)\mathbb{E}(A_nC_n) + \frac{n^{\underline{3}}(3n^2-11n-6)}{2(2n-5)^{\underline{3}}}.\end{aligned}$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(A_8C_8) = 8/3$ shows that for all $n \geq 8$, we have $\mathbb{E}(A_nC_n) = n^{\underline{4}}(n^2-5n-2)/[4(2n-5)^{\underline{4}}]$.

Next, substituting $\varphi[(a,b,p,c)] = ap$ in (17) and using (6), we have

$$\begin{aligned}(2n-3)\mathbb{E}(A_{n+1}P_{n+1}) &= (2n-13)\mathbb{E}(A_nP_n) + 3\mathbb{E}(A_nC_n) + \mathbb{E}(A_n) + 4\mathbb{E}(P_n^2)\\ &= (2n-13)\mathbb{E}(A_nP_n) + \frac{n^{\underline{3}}(n-4)(7n^2-8n-159)}{4(2n-5)^{\underline{4}}}.\end{aligned}$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(A_8P_8) = 24/11$ shows that for all $n \geq 8$, we have $\mathbb{E}(A_nP_n) = n^{\underline{4}}(n^3 - 7n^2 - 22n + 166)/[4(2n-5)^{\underline{5}}]$. Next, substituting $\varphi[(a,b,p,c)] = bc$ in (17) and using (6), we have

$$\begin{aligned}(2n-3)\mathbb{E}(B_{n+1}C_{n+1}) &= (2n-11)\mathbb{E}(B_nC_n) + n\mathbb{E}(B_n) + \mathbb{E}(P_nC_n) + \mathbb{E}(P_n)\\ &= (2n-11)\mathbb{E}(B_nC_n) + \frac{n^{\underline{3}}(3n^2-n-48)}{8(2n-5)^{\underline{3}}}.\end{aligned}$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(B_8C_8) = 32/33$ shows that for all $n \geq 8$, we have $\mathbb{E}(B_nC_n) = n^{\underline{4}}(n^2-n-24)/[16(2n-5)^{\underline{4}}]$.

Next, substituting $\varphi[(a,b,p,c)] = bp$ in (17) and using (6), we have

$$\begin{aligned}(2n-3)\mathbb{E}(B_{n+1}P_{n+1}) &= (2n-13)\mathbb{E}(B_nP_n) + 3\mathbb{E}(B_nC_n) - 6\mathbb{E}(B_n)\\ &\quad + \mathbb{E}(P_n^2) - \mathbb{E}(P_n)\\ &= (2n-13)\mathbb{E}(A_nP_n) + \frac{7n^{\underline{6}}}{16(2n-5)^{\underline{4}}}.\end{aligned}$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(B_8P_8) = 8/33$ shows that for all $n \geq 8$, we have $\mathbb{E}(B_nP_n) = n^{\underline{7}}/[16(2n-5)^{\underline{5}}]$.

Next, substituting $\varphi[(a,b,p,c)] = ab$ in (17) and using (6), we have

$$\begin{aligned}(2n-3)\mathbb{E}(A_{n+1}B_{n+1}) &= (2n-15)\mathbb{E}(A_nB_n) + \mathbb{E}(A_nP_n) + 4\mathbb{E}(B_nP_n) - \mathbb{E}(A_n)\\ &= (2n-15)\mathbb{E}(A_nB_n) + \frac{n^{\underline{7}}}{2(2n-5)^{\underline{5}}}.\end{aligned}$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(A_8B_8) = 8/33$ shows that for all $n \geq 8$, we have $\mathbb{E}(A_nB_n) = n^{\underline{8}}/[16(2n-5)^{\underline{6}}]$. Similarly, substituting $\varphi[(a,b,p,c)] = a^2$ in (17) and using (6), we have

$$\begin{aligned}(2n-3)\mathbb{E}(A_{n+1}^2) &= (2n-15)\mathbb{E}(A_n^2) + 8\mathbb{E}(A_nP_n) - 2\mathbb{E}(A_n) + 4\mathbb{E}(P_n)\\ &= (2n-15)\mathbb{E}(A_n^2) + \frac{n^{\underline{3}}(2n^4-8n^3-206n^2+1613n-3141)}{(2n-5)^{\underline{5}}}.\end{aligned}$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(A_8^2) = 24/11$ shows that for all $n \geq 8$, we have $\mathbb{E}(A_n^2) = n^{\underline{4}}(n^4 - 6n^3 - 133n^2 + 1374n - 3450)/[4(2n-5)^{\underline{6}}]$.

$$\begin{bmatrix} \frac{n^{\underline{4}}(n^4-6n^3-133n^2+1374n-3450)}{4(2n-5)^{\underline{6}}} & \frac{n^{\underline{8}}}{16(2n-5)^{\underline{6}}} & \frac{n^{\underline{4}}(n^3-7n^2-22n+166)}{4(2n-5)^{\underline{5}}} & \frac{n^{\underline{4}}(n^2-5n-2)}{4(2n-5)^{\underline{4}}} \\ \frac{n^{\underline{8}}}{16(2n-5)^{\underline{6}}} & \frac{n^{\underline{4}}(n^4+42n^3-1069n^2+7410n-16320)}{64(2n-5)^{\underline{6}}} & \frac{n^{\underline{7}}}{16(2n-5)^{\underline{5}}} & \frac{n^{\underline{4}}(n^2-n-24)}{16(2n-5)^{\underline{4}}} \\ \frac{n^{\underline{4}}(n^3-7n^2-22n+166)}{4(2n-5)^{\underline{5}}} & \frac{n^{\underline{7}}}{16(2n-5)^{\underline{5}}} & \frac{n(n^5-7n^4-19n^3+229n^2-480n+276)}{4(2n-5)^{\underline{4}}} & \frac{n(n^4-6n^3+5n^2+12n-12)}{4(2n-5)^{\underline{3}}} \\ \frac{n^{\underline{4}}(n^2-5n-2)}{4(2n-5)^{\underline{4}}} & \frac{n^{\underline{4}}(n^2-n-24)}{16(2n-5)^{\underline{4}}} & \frac{n(n^4-6n^3+5n^2+12n-12)}{4(2n-5)^{\underline{3}}} & \frac{n^{\underline{2}}(n^2-n-8)}{4(2n-5)^{\underline{2}}} \end{bmatrix}. \tag{18}$$

Box I.

Finally, substituting $\varphi[(a,b,p,c)] = b^2$ in (17) and using (6), we have

$$\begin{aligned}(2n-3)\mathbb{E}(B_{n+1}^2) &= (2n-15)\mathbb{E}(B_n^2) + 2\mathbb{E}(B_nP_n) + 6\mathbb{E}(B_n) + \mathbb{E}(P_n)\\ &= (2n-15)\mathbb{E}(B_n^2) + \frac{n^{\underline{3}}(n^4+38n^3-769n^2+4252n-7362)}{8(2n-5)^{\underline{5}}}.\end{aligned}$$

Solving the last recurrence equation with the initial condition $\mathbb{E}(B_8^2) = 4/11$ shows that for all $n \geq 8$, we have $\mathbb{E}(B_n^2) = n^{\underline{4}}(n^4+42n^3-1069n^2+7410n-16320)/[64(2n-5)^{\underline{6}}]$.

Together with (6), this concludes the computation of all entries in (18), and hence completes the proof of the theorem. □

References

Bienvenu, F., Lambert, A., Steel, M., 2022. Combinatorial and stochastic properties of ranked tree-child networks. Random Struct. Algorithms 60 (4), 653–689.

Chang, H., Fuchs, M., 2010. Limit theorems for patterns in phylogenetic trees. J. Math. Biol. 60 (4), 481–512.

Choi, K.P., Kaur, G., Wu, T., 2021. On asymptotic joint distributions of cherries and pitchforks for random phylogenetic trees. J. Math. Biol. 83 (4), 40.

Choi, K.P., Thompson, A., Wu, T., 2020. On cherry and pitchfork distributions of random rooted and unrooted phylogenetic trees. Theor. Popul. Biol. 132, 92–104.

Coronado, T.M., Mir, A., Rosselló, F., 2018. The probabilities of trees and cladograms under Ford's α-model. Sci. World J. 2018.

Disanto, F., Wiehe, T., 2013. Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model. Math. Biosci. 242 (2), 195–200.

Ferretti, L., Ledda, A., Wiehe, T., Achaz, G., Ramos-Onsins, S.E., 2017. Decomposing the site frequency spectrum: The impact of tree topology on neutrality tests. Genetics 207 (1), 229–240.

Fischer, M., Herbst, L., Kersting, S., Kühn, L., Wicke, K., 2023. Tree Balance Indices: A Comprehensive Survey. Springer Nature Switzerland AG, Cham.

Fuchs, M., 2008. Subtree sizes in recursive trees and binary search trees: Berry–Esseen bounds and Poisson approximations. Combin. Probab. Comput. 17 (5), 661–680.

Fuchs, M., Liu, H., Yu, T.C., 2024. Limit theorems for patterns in ranked tree-child networks. Random Struct. Algorithms 64 (1), 15–37.

Glover, N.M., Redestig, H., Dessimoz, C., 2016. Homoeologs: What are they and how do we infer them? Trends Plant Sci. 21 (7), 609–621.

Grimmett, G.R., Stirzaker, D.R., 2001. Probability and Random Processes, third ed. Oxford University Press.

Hagen, O., Hartmann, K., Steel, M., Stadler, T., 2015. Age-dependent speciation can explain the shape of empirical phylogenies. Syst. Biol. 64 (3), 432–440.

Kaur, G., Choi, K.P., Wu, T., 2023. Distributions of cherries and pitchforks for the ford model. Theor. Popul. Biol. 149, 27–38.

Kim, J., Rosenberg, N.A., Palacios, J.A., 2020. Distance metrics for ranked evolutionary trees. Proc. Natl. Acad. Sci. 117 (46), 28876–28886.

Mahmoud, H.M., 2009. Pólya urn models. In: Texts in Statistical Science Series, CRC Press, Boca Raton, FL, p. xii+290.

McKenzie, A., Steel, M.A., 2000. Distributions of cherries for two models of trees. Math. Biosci. 164, 81–92.

Plazzotta, G., Colijn, C., 2016. Asymptotic frequency of shapes in supercritical branching trees. J. Appl. Probab. 53 (4), 1143–1155.

Pouryahya, F., Sankoff, D., 2022. Peripheral structures in unlabelled trees and the accumulation of subgenomes in the evolution of polyploids. J. Theoret. Biol. 532, 110924.

R. Core Team, 2021. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/.

Rosenberg, N.A., 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly and polyphyly in a coalescent model.. Evolution 57 (7), 1465–1477.

Rosenberg, N.A., 2006. The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Ann. Comb. 10, 129–146.

Stufler, B., 2022. A branching process approach to level-k phylogenetic networks. Random Struct. Algorithms 61 (2), 397–421.

Wu, T., Choi, K.P., 2016. On joint subtree distributions under two evolutionary models. Theor. Popul. Biol. 108, 13–23.

Zhang, C., Rabiee, M., Sayyari, E., Mirarab, S., 2018. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics 19 (6), 15–30.