Two Results about the Sackin and Colless Indices for Phylogenetic Trees and Their Shapes

Gary Goh[1], Michael Fuchs[2] and Louxin Zhang[1*]

[1*]Department of Mathematics, National University of Singapore,
10 Lower Kent Ridge Road, Singapore, 119076, Singapore.
[2]Department of Mathematical Sciences, National Chengchi
University, Taipei, 116, Taiwan.

*Corresponding author(s). E-mail(s): matzlx@nus.edu.sg;
Contributing authors: e0148664@u.nus.edu; mfuchs@nccu.edu.tw;

Abstract

The Sackin and Colless indices are two widely-used metrics for measuring the balance of trees and for testing evolutionary models in phylogenetics. This short paper contributes two results about the Sackin and Colless indices of trees. One result is the asymptotic analysis of the expected Sackin and Colless indices of tree shapes (which are full binary rooted unlabelled trees) under the uniform model where tree shapes are sampled with equal probability. Another is a short direct proof of the closed formula for the expected Sackin index of phylogenetic trees (which are full binary rooted trees with leaves being labelled with taxa) under the uniform model.

Keywords: Phylogenetics, tree balance, Sackin index, Colless index, asymptotic analysis

MSC Classification: 05A16 , 05C30 , 92D15

1 Introduction

The Sackin [1, 2] and Colless [3] indices are two widely-used metrics for measuring the balance of phylogenetic trees and testing evolutionary models [4–9].

Phylogenetic trees are binary rooted trees in which each internal node has two children and only the leaves are labelled one-to-one with taxa. For a phylogenetic tree, its Sackin index is defined as the sum over its internal nodes of the number of leaves below that node, whereas its Colless index is defined as the sum over its internal nodes of the balance of that node, where the balance of a node is defined to be the difference in the number of leaves below the two children of that node. Because of their wide applications, the two tree balance metrics have been studied in the past decades (see the recent comprehensive survey [10]).

The Sackin and Colless indices of a random phylogenetic tree have been investigated under the Yule-Harding model (where ordered tree shapes, i.e., tree shapes with the children of each non-leaf node having a left-to-right order, are generated using a birth process, the leaves of these ordered tree shapes are labeled according to a permutation on the taxa which is chosen uniformly at random, and then the left-to-right order of the children of non-leaf nodes is disregarded) and the uniform model (where trees are sampled with equal probability) [9, 11–13]. The expected Sackin and Colless indices of a phylogenetic tree are proved to be asymptotic to $\sqrt{\pi}n^{3/2}$ under the uniform model and $n \log n$ under the Yule-Harding model [12, 13]. Recently, Mir et al. [14] discovered surprisingly that the expected Sackin index of a phylogenetic tree is simply

$$\frac{4^{n-1}n!(n-1)!}{(2n-2)!} - n \tag{1}$$

under the uniform model. An alternative proof of this closed formula was given by King and Rosenberg [15]. Both asymptotic and exact results on the variances of the Sackin and Colless indices have also been reported [9, 12, 13, 16].

It is not hard to see that the Sackin index of a binary tree is actually equal to the sum of the depths of all its leaves [17]. Therefore, the Sackin index and the tree height have also been studied for other types of trees in the combinatorics and theoretical computer science literature [18–21].

In this paper, we focus on two questions about the Sackin and Colless indices. The first question is what the expected Sackin and Colless indices of a random binary tree shape are under the uniform model [22]. Here, tree shapes (also called Otter or Polya trees) are binary rooted trees with unlabeled leaves where each internal node has two children. Although there is increasing interest in tree balance indices for tree shapes in the study of phylodynamic problems [23, 24], to the best of our knowledge, the statistical properties of these two indices and other tree balance indices have not been formally studied for tree shapes [10]. Here, we prove that the expected Sackin and Colless indices of a tree shape with n leaves are asymptotic to $\sqrt{\pi}\lambda^{-1}n^{3/2}$ under the uniform model, where $\lambda \approx 1.1300337163$.

Given that the closed formula (1) for the expected Sackin index of a phylogenetic tree under the uniform model is rather simple, the second question is whether a direct proof exists for the formula or not. We answer this question by using a simple recurrence for the Sackin index that is derived using the fact that all the phylogenetic trees on n taxa can be enumerated by inserting the n-th taxon into every edge of the phylogenetic trees on $n-1$ taxa [25]. Recently, this technique was used by Zhang for computing the sum over all nodes of the number of the descendants of that node and counting the number of tree-child networks with one reticulation [26].

2 Basic definitions and notation

2.1 Phylogenetic trees and shapes

A tree shape is a full binary rooted tree in which all nodes are unlabeled. A phylogenetic tree on n taxa is a full binary rooted tree with n leaves in which its leaves are uniquely labeled with a taxon and each of the $n-1$ non-leaf nodes has two children.

Let T be a phylogenetic tree on n taxa or a tree shape. We use $V_0(T)$ to denote the set of all non-leaf nodes of T and $V(T)$ to denote the set of all nodes. A leaf x is said to be *below* a node u in T if the unique path from the root to x passes through u. We use $\ell_T(u)$ to denote the number of leaves below u in T. Also, we set $\ell_T(u) := 1$ if u is a leaf.

Let $u \in V_0(T)$. The *balance* of u is defined to be $|\ell_T(v) - \ell_T(w)|$, where v and w are the two children of u. We use $\delta_T(u)$ to denote the balance of u.

For each non-root $u \in V(T)$, we use $p(u)$ to denote the parent of u in T.

2.2 Sackin and Colless indices

Definition 1 The Sackin index of a tree shape or a phylogenetic tree T is defined to be $\sum_{u\in V_0(T)} \ell_T(u)$, and denoted by $S(T)$.

Definition 2 The Colless index of a tree shape or a phylogenetic tree T is defined to be $\sum_{u\in V_0(T)} \delta_T(u)$, and denoted by $C(T)$.

The expected Sackin and Colless indices of a tree shape under the uniform model are respectively defined as:

$$\mathrm{ESI}_{sh}(n) := \frac{1}{b_n} \sum_{T\in\mathcal{T}(n)} \mathrm{S}(T)$$

and

$$\mathrm{ECI}_{sh}(n) := \frac{1}{b_n} \sum_{T\in\mathcal{T}(n)} C(T),$$

where $\mathcal{T}(n)$ denotes the set of all tree shapes with n leaves and $b_n := |\mathcal{T}(n)|$. Although there does not exist a closed formula for b_n, b_n can be computed using the following recurrence formulas for $n > 1$ (A001190 in the On-Line Encyclopedia of Integer Sequences[1]):

$$b_n = \sum_{1 \le k < n/2} b_k b_{n-k} + \begin{cases} 0, & \text{if } n \text{ is odd;} \\ \frac{1}{2} b_{n/2}(b_{n/2}+1), & \text{if } n \text{ is even.} \end{cases} \tag{2}$$

Equivalently, the generating function $B(z) := \sum_i b_i z^i$ satisfies the following equation:

$$B(z) = z + \frac{1}{2}\left(B(z)^2 + B(z^2)\right). \tag{3}$$

The expected Sackin index of a phylogenetic tree under the uniform model is defined similarly, that is,

$$\mathrm{ESI}_p(n) := \frac{1}{a_n} \sum_{P \in \mathcal{P}(n)} S(P),$$

where $\mathcal{P}(n)$ denotes the set of all phylogenetic trees on n taxa and $a_n := |\mathcal{P}(n)| = \frac{(2n-2)!}{2^{n-1}(n-1)!}$ (see [17]).

3 Asymptotic analysis of the expected Sackin and Colless indices for tree shapes

Recall that $\mathcal{T}(n)$ denotes the set of all possible tree shapes with n leaves. Let $S_n := \sum_{T \in \mathcal{T}(n)} S(T)$, which is the sum of the Sackin index over all tree shapes with n leaves. Obviously, $S_1 = 0$ and $S_2 = 2$.

For $n > 2$, $\mathcal{T}(n)$ can be obtained by combining every pair of tree shapes $T' \in \mathcal{T}(k)$ and $T'' \in \mathcal{T}(n-k)$, where k can range from 1 to $\lfloor n/2 \rfloor$. For a specific integer $k \le n/2$, $T \in \mathcal{T}(k)$ and $T' \in \mathcal{T}(n-k)$, $S(T) = n + S(T') + S(T'')$ for the tree shape T obtained by combining T' and T'', as there are n leaves below the root of T.

Using the facts mentioned in the previous paragraph and Eqn. (2), we obtain that:

$$\begin{aligned} S_n &= \sum_{1 \le k < n/2} \left(\sum_{T \in \mathcal{T}(k)} \sum_{T' \in \mathcal{T}(n-k)} (n + S(T) + S(T')) \right) \\ &= \sum_{1 \le k < n/2} \left(n b_k b_{n-k} + \sum_{T \in \mathcal{T}(k)} \sum_{T' \in \mathcal{T}(n-k)} (S(T) + S(T')) \right) \\ &= n \sum_{1 \le k < n/2} b_k b_{n-k} \end{aligned}$$

[1] https://oeis.org/

$$\begin{aligned}
&+\sum_{1\le k<n/2}\left(\sum_{T\in\mathcal{T}(k)}\sum_{T'\in\mathcal{T}(n-k)}S(T)+\sum_{T\in\mathcal{T}(k)}\sum_{T'\in\mathcal{T}(n-k)}S(T')\right)\\
&= nb_n+\sum_{1\le k<n/2}(b_{n-k}S_k+b_kS_{n-k})\\
&= nb_n+\sum_{1\le k<n}S_kb_{n-k}, \qquad (4)
\end{aligned}$$

for odd n and

$$\begin{aligned}
S_n &= nb_n+\sum_{1\le k<n/2}\left(\sum_{T\in\mathcal{T}(k)}\sum_{T'\in\mathcal{T}(n-k)}(S(T)+S(T'))\right)\\
&\quad+\sum_{T,T'\in\mathcal{T}(n/2):T\neq T'}(S(T)+S(T'))+\sum_{T\in\mathcal{T}(n/2)}2S(T)\\
&= nb_n+\sum_{1\le k<n/2}(b_{n-k}S_k+b_kS_{n-k})+\left(\sum_{T\in\mathcal{T}(n/2)}(b_{n/2}-1)S(T)\right)+2S_{n/2}\\
&= nb_n+S_{n/2}+\sum_{1\le k<n}S_kb_{n-k} \qquad (5)
\end{aligned}$$

for even n.

3.1 The asymptotic value of $\mathrm{ESI}_{sh}(n)$

It is unknown whether or not one can derive a closed formula for S_n from Eqn. (4)-(5). However, an asymptotic analysis of S_n follows from the classical asymptotic analysis of b_n from Eqn. (2). In order to recall the latter, we need the notion of Δ-analyticity. First, a Δ-domain with parameters δ and ϕ is a domain in the complex plane of the form:

$$\Delta := \{z\in\mathbb{C}\ :\ |z|<1+\delta,\ |\arg(z-1)|>\phi\}$$

with $\delta>0$ and $0<\phi<\pi/2$; see Definition VI.1 in [27]. A function is called Δ-analytic if it is analytic in such a Δ-domain.

Lemma 1 *([19]) The convergence radius ρ of the generating function $B(z)$ of b_n in Eqn. (3) satisfies $1/4\le\rho\le1/2$, where $\rho+B(\rho^2)/2=1/2$. Moreover, $B(z)$ is Δ-analytic and satisfies as $z\to\rho$ in a Δ-domain:*

$$B(z)=1-\lambda\sqrt{1-z/\rho}+\mathcal{O}(1-z/\rho), \qquad \lambda:=\sqrt{2\rho+2\rho^2B'(\rho^2)}. \qquad (6)$$

Thus,

$$b_n\sim\frac{\lambda}{2\sqrt{\pi}n^{3/2}\rho^n}, \qquad (n\to\infty). \qquad (7)$$

Remark 1 ρ and λ can be computed up to very high precision, e.g.,

$$\rho = 0.40269750367\cdots \qquad \text{and} \qquad \lambda = 1.1300337163\cdots.$$

The computation is done as follows: first, Eqn. (2) is used to compute a finite number of terms $\tilde{b}_n$ of b_n that are used to set up a polynomial $\tilde{B}(z)$ which approximates $B(z)$; then, find $\tilde{\rho}$ such that $\tilde{\rho} + \tilde{B}(\rho^2)/2 = 1/2$. Clearly, $\tilde{\rho}$ approximates ρ and this approximation can be made arbitrarily precise; also an approximation of λ can be derived from it via Eqn. (6).

Remark 2 The asymptotic expansion in Eqn. (7) follows from the *singularity expansion* in Eqn. (6) by the *transfer theorems* (see Theorem VI.3 and Corollary VI.1 in [27]) which assert that if $A(z)$ is Δ-analytic with $A(z) \sim c(1-z/\rho)^{-\alpha}$, where $c, \rho \in \mathbb{R} \setminus \{0\}$ and $\alpha \in \mathbb{C} \setminus \{0, -1, -2, \ldots\}$, then $[z^n]A(z) \sim [z^n]c(1-z/\rho)^{-\alpha} \sim c\rho^{-n}n^{\alpha-1}/\Gamma(\alpha)$, where $[z^n]f(z)$ denotes the n-th coefficient in the Maclaurin series of $f(z)$ and $\Gamma(z)$ is the gamma function. Indeed, set $A(z) := B(z) - 1$. Then, by (6), we have $A(z) \sim -\lambda\sqrt{1-z/\rho}$ and thus

$$b_n = [z^n]A(z) \sim \frac{-\lambda}{\Gamma(-1/2)n^{3/2}\rho^n} = \frac{\lambda}{2\sqrt{\pi}n^{3/2}\rho^n}$$

which is (7). (In the last step, we used that $\Gamma(-1/2) = -2\sqrt{\pi}$.)

Remark 3 More generally, the process of showing that a genearting function $A(z)$ is Δ-analytic, deriving the expansion $A(z) \sim c(1-z/\rho)^{-\alpha}$ as $z \to \rho$ and then using the transfer theorems to obtain the asymptotics of $[z^n]A(z)$ is called *singularity analysis*; see Chapter VI in [27].

Remark 4 Singularity analysis is closed under several operations on functions; see Section VI.10 in [27]. For instance, if singularity analysis can be applied to $A(z)$, it can also be applied to $A'(z)$, where the singularity expansion of $A'(z)$ is obtained from the one of $A(z)$ by term-by-term differentiation. E.g., $B'(z)$ from the previous lemma is also Δ-analytic with singularity expansion as $z \to \rho$

$$B'(z) \sim \frac{\lambda}{2\rho} \cdot \frac{1}{\sqrt{1-z/\rho}}, \tag{8}$$

from which the asymptotic expansion of $[z^n]B'(z)$ follows by the transfer theorems. (Of course, since $[z^n]B'(z) = (n+1)[z^{n+1}]B(z)$, this expansion is just the expansion in Eqn. (7) multiplied by n/ρ.)

Theorem 2 *Under the uniform model, the expected Sackin index of a tree shape with n leaves, $\mathrm{ESI}_{sh}(n)$, is asymptotic to $\pi^{1/2}\lambda^{-1}n^{3/2}$, where λ is given in Eqn. (6).*

Proof The recurrence formulas in Eqn. (4)-(5) translate into the following equation for the generating function $S(z) := \sum_i S_i z^i$ of S_n:

$$S(z) = zB'(z) + S(z)B(z) + S(z^2) \tag{9}$$

since the generating function of $\sum_{1\le k<n} S_k b_{n-k}$ is the product $S(z)B(z)$ and

$$\sum_{n\ge1} n b_n z^n = zB'(z), \qquad \sum_{n \text{ even}} S_{n/2} z^n = S(z^2).$$

Indeed,

$$\begin{aligned} S(z) = \sum_{n\ge1} S_n z^n &= \sum_{n\ge1} n b_n z^n + \sum_{n\ge1}\left(\sum_{1\le k<n} S_k b_{n-k}\right) z^n + \sum_{n \text{ even}} S_{n/2} z^n \\ &= zB'(z) + S(z)B(z) + S(z^2) \end{aligned}$$

which gives (9).

Next, by rewriting Eqn. (9) into

$$S(z) = \frac{zB'(z) + S(z^2)}{1 - B(z)},$$

we see that the radius of convergence of $S(z)$ is equal to ρ. (Because $B(z)$ and $B'(z)$ both have radius of convergence equal to ρ and $S(z^2)$ is analytic at ρ since $0 < \rho < 1$). Moreover, from Eqn. (6) and the closure properties of singularity analysis (Remark 4 above), we obtain that $S(z)$ is Δ-analytic and satisfies as $z \to \rho$ in a Δ-domain:

$$S(z) \sim \frac{\rho(\lambda/2\rho)(1 - z/\rho)^{-1/2} + S(\rho^2)}{\lambda\sqrt{1 - z/\rho} + \mathcal{O}(1 - z/\rho)} \sim \frac{1}{2} \cdot \frac{1}{1 - z/\rho},$$

where we used Eqn. (8) and that $S(z^2)$ is analytic at ρ.

By the transfer theorems (see Remark 2), we obtain:

$$S_n \sim \frac{1}{2}[z^n](1 - z/\rho)^{-1} \sim \frac{1}{2\rho^n}, \qquad (n \to \infty) \tag{10}$$

and thus

$$\mathrm{ESI}_{sh}(n) = \frac{S_n}{b_n} \sim \frac{1/(2\rho^n)}{\lambda/\left(2\sqrt{\pi} n^{3/2} \rho^n\right)} = \sqrt{\pi}\lambda^{-1} n^{3/2}, \qquad (n \to \infty)$$

using Eqn. (7). This proves the claim. □

3.2 The asymptotic value of $\mathbf{ECI}_{sh}(n)$

Next, we derive the asymptotic value of $\mathrm{ECI}_{sh}(n)$. First, for each internal node u of a tree, we use $c_1(u)$ and $c_2(u)$ to denote the two children of u. We have that $\ell(u) = \ell(c_1(u)) + \ell(c_2(u))$ and thus $\delta(u) = |\ell(c_1(u)) - \ell(c_2(u))| = \ell(u) - 2\min(\ell(c_1(u)), \ell(c_2(u))$. From this, it follows that for each tree shape T, $D(T) := S(T) - C(T) = 2\sum_{u\in V_0(T)} \min(\ell(c_1(u)), \ell(c_2(u))$.

Defining

$$D_n := \frac{1}{2}\sum_{T\in\mathcal{T}(n)} D(T),$$

we obtain:

$$C_n = \sum_{T\in\mathcal{T}(n)} C(T) = S_n - 2D_n. \tag{11}$$

Since the integer subsets $\{k : 1 \le k < n/2\}$ and $\{k : 1 \le k \le n/2\}$ are the same and $\sum_{1\le k<n/2} b_{n-k} D_k = \sum_{n/2\le n-k<n} b_{n-k} D_{n-(n-k)} = \sum_{n/2\le k<n} b_k D_{n-k}$ for odd n, we have the following recurrence formula:

$$
\begin{aligned}
D_n &= \sum_{1\le k<n/2} \left(\sum_{T\in\mathcal{T}(k)} \sum_{T'\in\mathcal{T}(n-k)} (D(T)+D(T')+k) \right) \\
&= \sum_{1\le k<n/2} kb_kb_{n-k} + \sum_{1\le k<n/2} (b_{n-k}D_k + b_kD_{n-k}) \\
&= \sum_{1\le k\le n/2} kb_kb_{n-k} + \sum_{1\le k<n} D_kb_{n-k}, \text{ for odd } n. \qquad (12)
\end{aligned}
$$

We now consider the case when n is even. Since each shape with $n/2$ leaves can form exactly $b_{n/2}-1$ shape pairs with all $b_{n/2}-1$ other shapes, we have that

$$
\sum_{T,T'\in\mathcal{T}(n/2):T\ne T'} (D(T)+D(T')) = \sum_{T\in\mathcal{T}(n/2)} (b_{n/2}-1)D(T) = (b_{n/2}-1)D(n/2).
$$

We also have that, for even n,

$$
\sum_{1\le k<n/2} (b_{n-k}D_k + b_kD_{n-k}) + b_{n/2}D_{n/2} = \sum_{1\le k<n} b_{n-k}D_k.
$$

Therefore,

$$
\begin{aligned}
D_n = &\sum_{1\le k<n/2} \left(\sum_{T\in\mathcal{T}(k)} \sum_{T'\in\mathcal{T}(n-k)} (D(T)+D(T')+k) \right) \\
&+ \sum_{T,T'\in\mathcal{T}(n/2):T\ne T'} (D(T)+D(T')+n/2) + \sum_{T\in\mathcal{T}(n/2)} (2D(T)+n/2) \\
= &\sum_{1\le k<n/2} kb_kb_{n-k} + \sum_{1\le k<n/2} (b_{n-k}D_k + b_kD_{n-k}) \\
&+ \left((b_{n/2}-1)D_{n/2} + \binom{b_{n/2}}{2}\frac{n}{2} \right) + 2D_{n/2} + \frac{n}{2}b_{n/2} \\
= &\sum_{1\le k\le n/2} kb_kb_{n-k} + \sum_{1\le k<n} D_kb_{n-k} - \frac{n}{2}\binom{b_{n/2}}{2} + D_{n/2}, \text{ for even } n.
\end{aligned}
\qquad (13)
$$

We first need a technical lemma for:

$$
F_n := \sum_{1\le k\le n/2} kb_kb_{n-k} + \begin{cases} 0, & \text{if } n \text{ is odd;} \\ -\frac{n}{2}\binom{b_{n/2}}{2}, & \text{if } n \text{ is even.} \end{cases}
$$

Lemma 3 *We have* $F_n = \mathcal{O}\left(n^{-1}\rho^{-n}\right)$.

Proof By using Eqn. (7),

$$\begin{aligned} F_n &= \mathcal{O}\left(\rho^{-n}\sum_{1\le k\le n/2} k^{-1/2}(n-k)^{-3/2} + n^{-2}\rho^{-n}\right) \\ &= \mathcal{O}\left(n^{-1}\rho^{-n}\int_0^{1/2} x^{-1/2}(1-x)^{-3/2}\mathrm{d}x + n^{-2}\rho^{-n}\right) \\ &= \mathcal{O}\left(n^{-1}\rho^{-n} + n^{-2}\rho^{-n}\right) = \mathcal{O}\left(n^{-1}\rho^{-n}\right), \end{aligned}$$

where in the second step, we approximated the sum by an integral. □

Now, define:

$$\tilde{D}_n := Kn^{-1}\rho^{-n} + \sum_{1\le k<n}\tilde{D}_k b_{n-k} + \begin{cases} 0, & \text{for } n \text{ is odd;} \\ \tilde{D}_{n/2}, & \text{for } n \text{ is even,} \end{cases} \tag{14}$$

where K is the implied $\mathcal{O}$-constant from the last lemma. The reason for considering this sequence is that it (a) majorizes D_n, namely, $D_n \le \tilde{D}_n$ (which is easily proved by induction) and (b) its asymptotics can derived with similar tools as used in the proof of Theorem 5.

Lemma 4 *We have,*

$$\tilde{D}_n \sim \frac{K}{\lambda\sqrt{\pi}}n^{-1/2}(\log n)\rho^{-n}, \qquad (n\to\infty).$$

Consequently, $D_n = \mathcal{O}\left(n^{-1/2}(\log n)\rho^{-n}\right)$.

Proof Let $\tilde{D}(z) := \sum_i \tilde{D}_i z^i$ be the generating function of $\tilde{D}_n$. Then, the recurrence in Eqn. (14) translates into

$$\tilde{D}(z) = K\log\frac{1}{1-z/\rho} + \tilde{D}(z)B(z) + \tilde{D}(z^2) \tag{15}$$

since

$$\sum_{n\ge1} Kn^{-1}\rho^{-n}z^n = K\log\frac{1}{1-z/\rho}$$

and the rest of the terms in (15) are explained as in the derivation of Eqn. (9). Solving for $D(z)$ gives:

$$\tilde{D}(z) = \frac{K\log\frac{1}{1-z/\rho} + \tilde{D}(z^2)}{1-B(z)}.$$

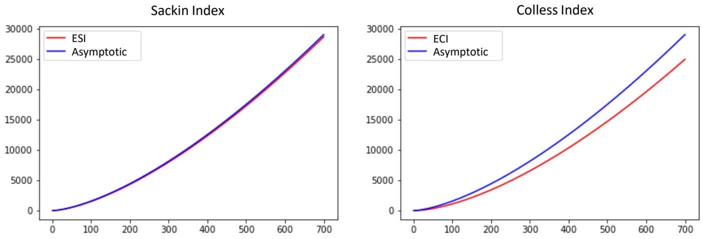

Fig. 1 The exact and asymptotic values of the expected Sackin (left) and Colless (right) indices.

Thus, from Eqn. (6), $\tilde{D}(z)$ satisfies as $z \to \rho$ in a Δ-domain:

$$\tilde{D}(z) \sim \frac{K \log \dfrac{1}{1-z/\rho} + \tilde{D}(\rho^2)}{\lambda\sqrt{1-z/\rho} + \mathcal{O}(1-z/\rho)} \sim \frac{K}{\lambda} \cdot \frac{\log \dfrac{1}{1-z/\rho}}{\sqrt{1-z/\rho}}$$

from which the claimed result follows by the transfer theorems (which also work with log-factors; see Theorem VI.3 in [27]). □

Now from Eqn. (10), Eqn. (11) and Lemma 4, we have the following result.

Theorem 5 *Under the uniform model, the expected Colless index of a tree shape with n leaves,* $\mathrm{ECI}_{sh}(n)$*, is asymptotic to* $\pi^{1/2}\lambda^{-1}n^{3/2}$.

3.3 Visualization on the asymptotic analyses

The exact and asymptotic values of $\mathrm{ESI}_{sh}(n)$ and $\mathrm{ECI}_{sh}(n)$ were computed and compared for n up to 700 (Figure 1). Here, For $\mathrm{ESI}_{sh}(n)$, the exact values were computed using the formulas in Eqn. (4)-(5) and the asymptotic values were computed using the formula in Theorem 2. For $\mathrm{ECI}_{sh}(n)$, the exact values were computed using the formulas in Eqn. (11)-(13) and the asymptotic values were computed using Theorem 5. The comparison indicates that the asymptotic value $\sqrt{\pi}\lambda^{-1}n^{3/2}$ is a very good approximation to the Sackin index even for a small number n. However, the asymptotic value overestimates the Colless index with a relatively large margin. The large margin is due to the fact that $\mathrm{ESI}_{sh}(n) - \mathrm{ECI}_{sh}(n) = 2D(n)/b(n)$ is of the order $n \log n$ according to our proof; however, the relative error will tend to 0 with a speed of at least $\log n/\sqrt{n}$.

4 The expected Sackin index for phylogenetic trees

Mir et al. discovered the following simple closed formula for the expected Sackin index for a phylogenetic tree under the uniform model.

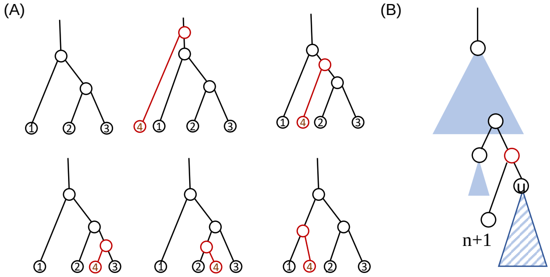

Fig. 2 (A) Illustration of the process of generating phylogenetic trees on $n+1$ taxa through inserting Leaf $n+1$ in each edge of a phylogenetic tree on taxa $\{1, 2, \cdots, n\}$ for $n = 3$. (B) After Leaf $n+1$ is attached onto the edge entering the node u, the number of leaves below the parent of $n+1$ in the obtained tree Q is equal to 1 plus that of u in the original tree P, i.e., $1 + \ell_P(u)$

.

Theorem 6 *([14]) For any n,* $\mathrm{ESI}_p(n) = \frac{4^{n-1} n!(n-1)!}{(2n-2)!} - n$*.*

An alternative proof was presented in [15] recently. Here, we will present a short direct proof using the following enumeration of phylogenetic trees (see [25] for example):

> Assume that there is an open edge entering the root of each phylogenetic tree. $\mathcal{P}(n+1)$ can be obtained from $\mathcal{P}(n)$ by attaching Leaf $n+1$ on each of the $2n-1$ edges of every tree of $\mathcal{P}(n)$ (Figure 2.A).

Let $S_n^{(p)} := \sum_{P \in \mathcal{P}(n)} S(P)$. Note that $S_n^{(p)} = \mathrm{ESI}_p(n) \times a_n$, where $a_n = |\mathcal{P}(n)|$. For each $P \in \mathcal{P}(n)$, we use $\mathcal{A}(P)$ to denote the set of $2n-1$ phylogenetic trees on $n+1$ taxa that are obtained from P by attaching Leaf $n+1$ on each of the $2n-1$ tree edges of P. Then,

$$S_{n+1}^{(p)} = \sum_{P \in \mathcal{P}(n)} \sum_{Q \in \mathcal{A}(P)} S(Q). \tag{16}$$

Consider a tree $Q \in \mathcal{A}(P)$. Note that Leaf $n+1$ and its parent are the only nodes of Q that are not found in P. Assume that Q is obtained by attaching Leaf $n+1$ to the edge e that enters u in P. The number of leaves below the parent of Leaf $n+1$ is $1 + \ell_P(u)$ in Q (Figure 2.B). Therefore, the amount contributed by the parents of Leaf $n+1$ to the sum $\sum_{Q \in \mathcal{A}(P)} S(Q)$ is:

$$\begin{aligned}\sum_{u \in V(P)} (1 + \ell_P(u)) &= (2n-1) + \sum_{u \in V(P)} \ell_P(u) \\ &= (2n-1) + n + \sum_{u \in V_0(P)} \ell_P(u) \\ &= 3n - 1 + S(P), \end{aligned} \tag{17}$$

where the n in the second expression is the sum of $\ell_P(u)$ (which is 1) over all the n leaves u in P and $V(P)$ and $V_0(P)$ are the set of nodes and non-leaf nodes, respectively, in P; see Section 2.1.

For $w \in V(P)$, we have either $\ell_P(w) = \ell_Q(w)$ or $\ell_P(w) = \ell_Q(w) + 1$. Furthermore, the latter holds if and only if Q is obtained by attaching Leaf $n+1$ to an edge below w in P. Since there are $2\ell_P(w) - 2$ edges below w in P, thus $\ell_Q(w) = \ell_P(w) + 1$ for exactly $2\ell_P(w) - 2$ trees Q of $\mathcal{A}(P)$. Therefore,

$$\begin{aligned}\sum_{Q\in\mathcal{A}(P)} S(Q) &= (2n-1)S(P) + (S(P) + (3n-1)) + \sum_{w\in V_0(P)} (2\ell_P(w) - 2)\\ &= 2nS(P) + (3n-1) + 2S(P) - 2|V_0(P)|\\ &= 2(n+1)S(P) + (n+1).\end{aligned}$$

Adding $n+1$ to each term in the left-hand side of the above equality, which can be considered as the contribution of the $n+1$ leaves, we further have:

$$\begin{aligned}\sum_{Q\in\mathcal{A}(P)} (S(Q) + (n+1)) &= 2(n+1)S(P) + (n+1) + (2n-1)(n+1)\\ &= 2(n+1)\left(S(P) + n\right).\end{aligned}$$

By Eqn. (16), we obtain the following simple recurrence formula:

$$\begin{aligned}S^{(p)}_{n+1} + (n+1)a_{n+1} &= \sum_{P\in\mathcal{P}(n)} \sum_{Q\in\mathcal{A}(P)} (S(Q) + (n+1))\\ &= \sum_{P\in\mathcal{P}(n)} 2(n+1)\left(S(P) + n\right)\\ &= 2(n+1)\left(S^{(p)}_n + na_n\right). \qquad (18)\end{aligned}$$

Since $S^{(p)}_2 = 2$ and $a_2 = 1$, Eqn. (18) implies that $S^{(p)}_n = 2^{n-1}n! - na_n$ and

$$\mathrm{ESI}_p(n) = \frac{S^{(p)}_n}{a_n} = \frac{4^{n-1}n!(n-1)!}{(2n-2)!} - n$$

Theorem 6 is proved.

5 Conclusion

In this short paper, we contributed two results to the study of the Sackin and Colless indices. We have proved that the asymptotic value of Sackin and Colless indices are the same for tree shapes under the uniform model. The same phenomenon was also observed for phylogenetic trees under the uniform model; see [13]. Thus, our result is expected since tree shapes under the uniform model are known to behave similar to phylogenetic trees under the uniform model; see the discussion in the introduction of [19]. In particular, the average height of phylogenetic trees and binary tree shapes with n leaves are both asymptotically equal to $2\lambda^{-1}\sqrt{\pi n}$ (see [18] and [19]).

We also presented a short direct proof of the closed formula for the expected Sackin index of phylogenetic trees under the uniform model. The proof is based on a tree enumeration approach that is different from one used in [14] and [15]. This technique was also used by Goh [28] to derive a short proof of the closed formula for the expected total cophenetic index of a phylogenetic tree under the uniform model that was introduced in [14] (see also [10]). It is an interesting problem whether or not the proof technique in Section 4 can be used to investigate other tree balance indices (such as those given in the survey paper [10]).

CRediT authorship contribution statement

G. Goh: Recurrence formulas; L. Zhang: Recurrence formulas, writing; M. Fuchs: Asymptotic analysis, writing.

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

The authors thanks the two anonymous reviewers for useful suggestions and comments for preparing the final version of this paper. LZ was supported by MOE Tier 1 grant R-146-000-318-114; MF was supported by MOST-109-2115-M-004-003-MY2.

References

[1] Sackin, M.J.: “Good” and “Bad” Phenograms. Systematic Biology **21**(2), 225–226 (1972). https://doi.org/10.1093/sysbio/21.2.225

[2] Shao, K.-T., Sokal, R.R.: Tree balance. Systematic Zoology **39**(3), 266–276 (1990)

[3] Colless, D.H.: Review of “phylogenetics: the theory and practice of phylogenetic systematics”. Systematic Zoology **31**(1), 100–104 (1982)

[4] Avino, M., Ng, G.T., He, Y., Renaud, M.S., Jones, B.R., Poon, A.F.: Tree shape-based approaches for the comparative study of cophylogeny. Ecology and Evolution **9**(12), 6756–6771 (2019)

[5] Xue, C., Liu, Z., Goldenfeld, N.: Scale-invariant topology and bursty branching of evolutionary trees emerge from niche construction. Proceedings of the National Academy of Sciences **117**(14), 7879–7887 (2020)

[6] Mooers, A.O., Heard, S.B.: Inferring evolutionary process from phylogenetic tree shape. The Quarterly Review of Biology **72**(1), 31–54 (1997)

[7] Scott, J.G., Maini, P.K., Anderson, A.R., Fletcher, A.G.: Inferring tumor proliferative organization from phylogenetic tree measures in a computational model. Systematic Biology **69**(4), 623–637 (2020)

[8] Blum, M.G.B., Heyer, E., François, O., Austerlitz, F.: Matrilineal fertility inheritance detected in hunter–gatherer populations using the imbalance of gene genealogies. PLoS Genetics **2**(8), 122 (2006)

[9] Kirkpatrick, M., Slatkin, M.: Searching for evolutionary patterns in the shape of a phylogenetic tree. Evolution **47**(4), 1171–1181 (1993)

[10] Fischer, M., Herbst, L., Kersting, S., Kühn, L., Wicke, K.: Tree balance indices: a comprehensive survey. arXiv preprint arXiv:2109.12281 (2021)

[11] Heard, S.B.: Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. Evolution **46**(6), 1818–1826 (1992)

[12] Blum, M.G., François, O.: On statistical tests of phylogenetic tree imbalance: the sackin and other indices revisited. Mathematical Biosciences **195**(2), 141–153 (2005)

[13] Blum, M.G., François, O., Janson, S.: The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. The Annals of Applied Probability **16**(4), 2195–2214 (2006)

[14] Mir, A., Rosselló, F., *et al.*: A new balance index for phylogenetic trees. Mathematical Biosciences **241**(1), 125–136 (2013)

[15] King, M.C., Rosenberg, N.A.: A simple derivation of the mean of the sackin index of tree balance under the uniform model on rooted binary labeled trees. Mathematical Biosciences **342**, 108688 (2021)

[16] Coronado, T.M., Mir, A., Rosselló, F., Rotger, L.: On sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index. BMC Bioinformatics **21**(1), 1–17 (2020)

[17] Steel, M.: Phylogeny: discrete and random processes in evolution. SIAM (2016)

[18] Flajolet, P., Odlyzko, A.: The average height of binary trees and other

simple trees. Journal of Computer and System Sciences **25**(2), 171–213 (1982)

[19] Broutin, N., Flajolet, P.: The distribution of height and diameter in random non-plane binary trees. Random Structures & Algorithms **41**(2), 215–252 (2012)

[20] Fill, J.A., Kapur, N.: Limiting distributions for additive functionals on catalan trees. Theoretical Computer Science **326**(1-3), 69–102 (2004)

[21] Fuchs, M., Jin, E.Y.: Equality of shapley value and fair proportion index in phylogenetic trees. Journal of Mathematical Biology **71**(5), 1133–1147 (2015)

[22] Rogers, J.S.: Central moments and probability distributions of three measures of phylogenetic tree imbalance. Systematic Biology **45**(1), 99–110 (1996)

[23] Colijn, C., Plazzotta, G.: A metric on phylogenetic tree shapes. Systematic Biology **67**(1), 113–126 (2018)

[24] Kim, J., Rosenberg, N.A., Palacios, J.A.: Distance metrics for ranked evolutionary trees. Proceedings of the National Academy of Sciences **117**(46), 28876–28886 (2020)

[25] Felsenstein, J.: Inferring Phylogenies. Sunderland, MA, USA: Sinauer Assoc Inc (2004)

[26] Zhang, L.: Generating normal networks via leaf insertion and nearest neighbor interchange. BMC Bioinformatics **20**(20), 1–9 (2019)

[27] Flajolet, P., Sedgewick, R.: Analytic Combinatorics. Cambridge University Press (2009)

[28] Goh, G.: Metrics for Measuring the Shape of Phylogenetic Trees. Honors Thesis, National University of Singapore (2022)