Mathematical Biosciences 342 (2021) 108688

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Letter to the Editor

A simple derivation of the mean of the Sackin index of tree balance under the uniform model on rooted binary labeled trees

In recent years, *Mathematical Biosciences* has reported several studies of the properties of tree balance statistics, measures that report the extent to which nodes of an evolutionary tree are "balanced" in having subtrees with similar numbers of descendants [1–3]. For trees with a fixed number of leaves, these and related studies [4–6] have examined features such as the minima and maxima of two of the earliest and most popular tree balance statistics—the Sackin [7] and Colless [8] indices—and the means and variances of these statistics under two of the most popular probability models for evolutionary tree shapes, the Yule–Harding and uniform models.

The Sackin index sums path lengths from leaves to the root, considering all leaves [9, p. 53]. For binary trees, the Colless index sums the absolute difference between the numbers of leaves in the two subtrees descended from internal nodes, considering all internal nodes [9, p. 53]. In the Yule–Harding model for n taxa, beginning with a single species and proceeding iteratively, species are all equally likely to be the next to speciate, inducing a particular probability distribution on rooted binary labeled trees with n leaves [9, p. 43]. In the uniform model, all rooted binary labeled trees with n leaves are equiprobable [9, p. 50].

Mathematical properties of the Sackin and Colless indices under the Yule–Harding and uniform models have long been of interest, with an initial derivation of the mean of the Colless index under the Yule–Harding model [10] and several subsequent studies [11–16] preceding the recent wave of investigations. Hence, in returning to these indices, with all the attention that had previously been focused on them, Mir et al. [1] were evidently surprised to discover that one of their formulas—a fundamental result that solves one of the most natural problems that might be posed concerning a tree balance index—appeared to be novel. They wrote *"we obtain an exact formula for the expected value of the Sackin index under the uniform model, a result that seems to be new in the literature."*

Formally, for $n \geq 2$, consider the set $\mathrm{RB}(X)$ of rooted binary labeled trees whose leaves are bijectively labeled by n distinct labels in a set X [9, p. 12]. $\mathrm{RB}(X)$ contains $\mathrm{rb}(n) = (2n-3)!! = (2n-3)(2n-5)\cdots(5)(3)(1) = (2n-2)!/[2^{n-1}(n-1)!]$ trees [9, eq. 2.2].

Definition 1. Consider a rooted binary labeled tree $T \in \mathrm{RB}(X)$. For each leaf $x \in X$, let $\ell(x)$ give the length in edges of the directed path from the root ρ of T to leaf x. The **Sackin index** for T is

$$S_T = \sum_{x \in X} \ell(x).$$

Definition 2. For $n \geq 2$, given a probability distribution θ on the set of rooted binary labeled trees $\mathrm{RB}(X)$ with $|X| = n$, let S_n denote the random variable obtained by randomly choosing $T \in \mathrm{RB}(X)$ according to θ and computing the value S_T.

An asymptotic large-n expectation under the uniform model, $\mathbb{E}_U[S_n] \sim \sqrt{\pi} n^{3/2}$, was studied by Blum et al. [16] (see also the table on p. 14 of [17], in which the uniform model corresponds to the $\beta = -1.5$ case of the more general beta-splitting model, and the table entry $\bar{\rho}(-1.5) = \sqrt{\pi} n^{1/2}$ gives the asymptotic mean path length to the root for a node chosen at random in a tree under the uniform model).

The exact $\mathbb{E}_U[S_n]$ was only recently obtained by Mir et al. [1].

Theorem 3 (*[1, Theorem 22]*). *The expectation of the Sackin index under the uniform model on rooted binary labeled trees of n leaves is*

$$\mathbb{E}_U[S_n] = n\left[\frac{(2n-2)!!}{(2n-3)!!} - 1\right] = \frac{n!\,(n-1)!}{(2n-2)!}\left[2^{2n-2} - \binom{2n-2}{n-1}\right].$$

The proof by Mir et al. [1] of Theorem 3 begins with enumerations of classes of trees defined by the path length to the root from a leaf node with a specified label. The enumerations give rise to a sum that is solved after much algebra with the help of three sums evaluated by automatic summation algorithms.

The purpose of this note is to produce a new, simple proof of Theorem 3 characterizing the expected value of the Sackin index under the uniform model. To provide the new proof, we use the fact that taxa are exchangeable in the uniform model [9, p. 52], so that the probability under the model of a rooted binary labeled tree T in $\mathrm{RB}(X)$ can be computed from the shape of T, disregarding the labels.

Definition 4. A probability distribution θ on $\mathrm{RB}(X)$ satisfies the **exchangeability property** if for each rooted binary labeled tree T in $\mathrm{RB}(X)$ and each permutation σ of its leaf labels, $\mathbb{P}_\theta(T) = \mathbb{P}_\theta(\sigma(T))$.

The exchangeability of the uniform model enables use of a result from Than & Rosenberg [18]. A subset A of the label set X is said to represent a *cluster* in a labeled tree T if for some node v of T, the leaves descended from v are bijectively labeled by the elements of A [9, p. 18].

Proposition 5 (*[18, Lemma 6], [9, Proposition 3.5]*). *If a probability distribution θ on $\mathrm{RB}(X)$ satisfies the exchangeability property, then*

$$\mathbb{E}_\theta[S_n] = \sum_{k=1}^{n-1} \binom{n}{k} k p_n(k),$$

where $n = |X|$, and $p_n(k)$ is the probability that a given subset $A \subseteq X$ with $|A| = k$, $1 \leq k \leq n$, is a cluster of a tree of n leaves sampled from $\mathrm{RB}(X)$ according to θ.

This result is obtained by noting that for $T \in \mathrm{RB}(X)$, the sum $S_T = \sum_{x \in X} \ell(x)$ that computes the Sackin index, proceeding over leaves

https://doi.org/10.1016/j.mbs.2021.108688
3 June 2021
Available online 16 September 2021

of T, can be converted to a sum over edges of T. In particular, for each leaf v of T and each edge e ancestral to v, v appears in the cluster of T immediately descended from e. Each edge of T contributes a count to S_T equal to the size of the subtree rooted below the edge. Hence, $\sum_{x\in X}\ell(x) = \sum_{k=1}^{n-1} kL_k$ where L_k counts the number of clusters in T with size k leaves. Taking the expectation of L_k over all trees sampled from RB(X) and using $\mathbb{E}_\theta[L_k] = \binom{n}{k}p_n(k)$, the proposition follows.

The probabilities $p_n(1)$ and $p_n(n)$ satisfy $p_n(1) = p_n(n) = 1$, as each leaf ($|A| = 1$) is a cluster of each tree in RB(X), as is the full set of leaves ($|A| = n$). For $2 \le k \le n-1$, $0 < p_n(k) < 1$. In particular, the number of rooted binary labeled trees for the k leaves in a cluster A with $|A| = k$ is $\mathrm{rb}(k)$; treating the cluster A as a node, the number of rooted binary labeled trees that contain the remaining $n-k$ leaves and the cluster A is $\mathrm{rb}(n-k+1)$. As each rooted binary labeled tree of n leaves has probability $1/\mathrm{rb}(n)$ under the uniform model, we therefore have the following result.

Proposition 6 ([18, eq. 10], [9, eq. 3.4]). *Under the uniform model, for* $1 \le k \le n$,

$$p_n(k) = \frac{\mathrm{rb}(k)\,\mathrm{rb}(n-k+1)}{\mathrm{rb}(n)} = \frac{\binom{n-1}{k-1}}{\binom{2n-2}{2k-2}}.$$

We now provide the proof of Theorem 3.

Proof of Theorem 3. By Propositions 5 and 6,

$$\begin{aligned}\mathbb{E}_U[S_n] &= \sum_{k=1}^{n-1}\binom{n}{k}kp_n(k)\\ &= \sum_{k=1}^{n-1}\binom{n}{k}k\frac{\binom{n-1}{k-1}}{\binom{2n-2}{2k-2}}\\ &= \frac{n!\,(n-1)!}{(2n-2)!}\sum_{k=0}^{n-2}\binom{2k}{k}\binom{2(n-1)-2k}{(n-1)-k}. \end{aligned} \tag{1}$$

This expression is simplified by a "remarkable property of the 'middle' elements of Pascal's triangle" [19, eq. 5.39], the identity

$$4^m = \sum_{j=0}^{m}\binom{2j}{j}\binom{2m-2j}{m-j}. \tag{2}$$

Adding a term for $k = n-1$ to the sum in Eq. (1), we take $m = n-1$ in the "remarkable" Eq. (2), obtaining

$$\mathbb{E}_U[S_n] = \frac{n!\,(n-1)!}{(2n-2)!}\left[4^{n-1} - \binom{2n-2}{n-1}\binom{0}{0}\right]. \quad \square$$

The identity in Eq. (2) is quickly obtained by expressing coefficients of the series expansion for $f(z) = (1-4z)^{-1}$ in two different ways. Trivially, $[z^m]f(z) = 4^m$. We also have $f(z) = g(z)^2$ for $g(z) = (1-4z)^{-1/2}$. Taking the series expansion of $g(z)$, we have $[z^j]g(z) = \binom{2j}{j}$, so that the identity follows from $[z^m]f(z) = \sum_{j=0}^{m}\left([z^j]g(z)\right)\left([z^{m-j}]g(z)\right)$.

The asymptotic mean of the Sackin index can be computed by application of Stirling's formula to the expression in Theorem 3; we can also quickly deduce the asymptotic mean by rewriting the expression in Theorem 3 in terms of a Catalan number. Recalling that the Catalan number C_n satisfies $C_n = \frac{1}{n+1}\binom{2n}{n}$, we obtain the following alternate formula.

Corollary 7. *The expectation of the Sackin index under the uniform model on rooted binary labeled trees of* n *leaves is*

$$\mathbb{E}_U[S_n] = \frac{4^{n-1}}{C_{n-1}} - n.$$

We compute the asymptotic expression for the mean Sackin index from the asymptotic expression for the Catalan numbers, $C_n \sim 4^n/(n^{3/2}\sqrt{\pi})$.

Corollary 8. *As* $n \to \infty$, *the expectation of the Sackin index under the uniform model on rooted binary labeled trees of* n *leaves satisfies*

$$\mathbb{E}_U[S_n] \sim \sqrt{\pi}n^{3/2}.$$

With the considerable attention devoted to the Sackin index in nearly 50 years since its introduction, we can add to the surprise of Mir et al. [1] in finding that their result on its expectation under the uniform model has a simple proof.

Interestingly, two reviewers pointed us to additional proofs. Coronado et al. [6] obtained a closed form for a class of recurrences that includes a recurrence for the mean Sackin index. The Sackin index satisfies a stochastic recurrence $S_n = S_k + S_{n-k} + n$ [9, eq. 3.12]. Under the uniform model, the probability that the "left" subtree of a rooted binary tree with n leaves has k leaves, $1 \le k \le n-1$, is [17, eq. 5],

$$q_n(k) = q_n(n-k) = \frac{1}{2}\binom{n}{k}\frac{\mathrm{rb}(k)\,\mathrm{rb}(n-k)}{\mathrm{rb}(n)} = \frac{C_{k-1}\,C_{n-k-1}}{C_{n-1}}.$$

The expected Sackin index then has recurrence

$$\mathbb{E}_U[S_n] = \left(\sum_{k=1}^{n-1} q_n(k)\,\mathbb{E}_U[S_k] + q_n(n-k)\,\mathbb{E}_U[S_{n-k}]\right) + n, \tag{3}$$

with $\mathbb{E}_U[S_1] = 0$. Noting that $\sum_{k=1}^{n-1} q_n(k)\,\mathbb{E}_U[S_k] + q_n(n-k)\,\mathbb{E}_U[S_{n-k}] = 2\sum_{k=1}^{n-1} q_n(k)\,\mathbb{E}_U[S_k]$, the recurrence is solved by the special case of Proposition 6 of Coronado et al. [6] with $X_1 = 0$, $a_1 = 1$, $a_\ell = 0$ for $2 \le \ell \le n-1$, and $b_\ell = 0$ for all ℓ, recovering Theorem 3.

Fuchs & Jin [20] reported the mean depth of a leaf chosen at random under the uniform model on rooted binary labeled trees—or $\mathbb{E}_U[S_n]/n$. They exploited an oft-noted mapping [16,17,21] that connects the rooted binary labeled trees and the Catalan trees, a class of rooted binary unlabeled trees in which left and right descendants are distinguished and internal nodes have either a left descendant node, a right descendant node, or both. In a uniform probability model on the Catalan trees, quantities related to the Sackin index have long been studied [22,23]; Fuchs & Jin [20] obtained and solved a version of the recurrence in Eq. (3), reporting their Theorem 4 in a form similar to that of our Corollary 7.

Note that our new proof provides a method for evaluating the expected Sackin index for any probability model for which the quantity $p_n(k)$ can be calculated. For the Yule–Harding model, $p_n(k) = 2n/[k(k+1)\binom{n}{k}]$ [9, eq. 3.5]. As was noted by Steel [9, eq. 3.11], the approach in our proof of Theorem 3 provides a computation of the mean Sackin index under the Yule-Harding model as well, $E_{YH}[S_n] = 2n\sum_{k=2}^{n}\frac{1}{k}$.

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 reviewers who pointed us to additional proofs of Corollary 7 and to the quick proof of Eq. (2) by use of generating functions. We thank Mike Steel for suggesting a close look at [17]. Support was provided by National Institutes of Health grant R01 GM131404.

References

[1] A. Mir, F. Rosselló, L. Rotger, A new balance index for phylogenetic trees, Math. Biosci. 241 (2013) 125–136.
[2] G. Cardona, A. Mir, F. Rosselló, L. Rotger, The expected value of the squared cophrenetic metric under the Yule and the uniform models, Math. Biosci. 295 (2018) 73–85.
[3] K. Bartoszek, T.M. Coronado, A. Mir, F. Rosselló, Squaring within the Colless index yields a better balance index, Math. Biosci. 331 (2021) 108503.
[4] G. Cardona, A. Mir, F. Rosselló, Exact formulas for the variances of several balance indices under the Yule model, J. Math. Biol. 67 (2013) 1833–1846.

[5] T.M. Coronado, M. Fischer, L. Herbst, F. Rosselló, K. Wicke, On the minimum value of the Colless index and the bifurcating trees that achieve it, J. Math. Biol. 80 (2020) 1993–2054.
[6] T.M. Coronado, A. Mir, F. Rosselló, L. Rotger, On Sackin's original proposal: the variance of the leaves' depths as a phylogenetic balance index, BMC Bioinformatics 21 (2020) 154.
[7] M. Sackin, 'Good' and 'bad' phenograms, Syst. Zool. 21 (1972) 225–226.
[8] D. Colless, Review of "Phylogenetics: the theory and practice of phylogenetic systematics", Syst. Zool. 31 (1982) 100–104.
[9] M. Steel, Phylogeny: Discrete and Random Processes in Evolution, Society for Industrial and Applied Mathematics, Philadelphia, 2016.
[10] S.B. Heard, Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees, Evolution 46 (1992) 1818–1826.
[11] M. Kirkpatrick, M. Slatkin, Searching for evolutionary patterns in the shape of a phylogenetic tree, Evolution 47 (1993) 1171–1181.
[12] J.S. Rogers, Response of Colless's tree imbalance to number of terminal taxa, Syst. Biol. 42 (1993) 102–105.
[13] J.S. Rogers, Central moments and probability distribution of Colless's coefficient of tree imbalance, Evolution 48 (1994) 2026–2036.
[14] J.S. Rogers, Central moments and probability distributions of three measures of phylogenetic tree imbalance, Syst. Biol. 45 (1996) 99–110.
[15] M.G.B. Blum, O. François, On statistical tests of phylogenetic tree imbalance: the Sackin and other indices revisited, Math. Biosci. 195 (2005) 141–153.
[16] M.G.B. Blum, O. François, S. Janson, The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance, Ann. Appl. Probab. 16 (2006) 2195–2214.
[17] D. Aldous, Probability distributions on cladograms, in: D. Aldous, R. Pemantle (Eds.), Random Discrete Structures, Springer-Verlag, New York, 1996, pp. 1–18.
[18] C.V. Than, N.A. Rosenberg, Mean deep coalescence cost under exchangeable probability distributions, Discrete Appl. Math. 174 (2014) 11–26.
[19] R.L. Graham, D.E. Knuth, O. Patashnik, Concrete Mathematics, second ed., Addison-Wesley, Boston, 1994.
[20] M. Fuchs, E.Y. Jin, Equality of Shapley value and fair proportion index in phylogenetic trees, J. Math. Biol. 71 (2015) 1133–1147.
[21] H. Chang, M. Fuchs, Limit theorems for patterns in phylogenetic trees, J. Math. Biol. 60 (2012) 481–512.
[22] J.A. Fill, N. Kapur, Limiting distributions for additive functionals on Catalan trees, Theoret. Comput. Sci. 326 (2004) 69–102.
[23] R. Sedgewick, P. Flajolet, An Introduction To the Analysis of Algorithms, second ed., Addison-Wesley, Upper Saddle River, NJ, 2013.

Matthew C. King, Noah A. Rosenberg *
Department of Biology, Stanford University, Stanford, CA 94305, United States of America
E-mail address: noahr@stanford.edu (N.A. Rosenberg).

* Corresponding author.