Measuring Tree Balance with Normalized Tree Area

T. Araújo Lima · Flavia M. D. Marquitti · Marcus A. M. de Aguiar

Abstract The study of species organization and their clustering by genetic or phenotypic similarity is carried out with the tools of phylogenetic trees. An important structural property of phylogenetic trees is the balance, which measures how taxa are distributed among clades. Tree balance can be measured using indices such as the Sackin ($S$) and the Total Cophenetic ($\Phi$), which are based on the distance between nodes of the tree and its root. Here, we propose a new metric for tree balance, $d$, the Area per Pair (APP) of the tree, which is a re-scaled version of the so-called tree area. We compute $\bar{d}$ for the rooted caterpillar and maximally balanced trees and we also obtain exact formulas for its expected value and variance under the Yule model. The variance of APP for Yule trees has the remarkable property of converging to an asymptotic constant value for large trees. We compare the Sackin, Total Cophenetic and APP indices for hundreds of empirical phylogenies and show that APP represents the observed distribution of tree balances better than the two other metrics.

Acknowledgements This work was partly funded by the Brazilian agencies: TAL was supported by grant #158002/2018-0 (CNPq). FMMD was supported by grant #2015/11985-3 (FAPESP) and Capes. MAMA was supported by grant #301082/2019-7 (CNPq), #2016/01343-7 (ICTP-SAIFR FAPESP) and #2019/20271 (FAPESP).

T. Araújo Lima (✉)
E-mail: tiagoapl@ifi.unicamp.br
Flavia M. D. Marquitti
E-mail: flamarquitti@gmail.com
Marcus A. M. de Aguiar
E-mail: aguiar@ifi.unicamp.br

Instituto de Física Gleb Wataghin, Universidade Estadual de Campinas, Unicamp, 13083-970 Campinas, SP, Brazil
Keywords Phylogenetic binary tree · Sackin index · Total Cophenet index · Yule Model · Asymptotic variance value

Mathematics Subject Classification (2000) 92D15 · 92C42 · 92B10

1 Introduction

Several analytical and computational tools have been recently developed to understand and characterize phylogenetic trees (Felsenstein 2004). These studies are motivated by the belief that structural properties of trees reflect their evolutionary history and were shaped by processes such as geographic modes of diversification and by rates of speciation and extinction events (Nee et al 1992; Mooers and Heard 1997; Morlon 2014; Cabral et al 2017; Costa et al 2019). Extracting evolutionary information from a tree, however, is not a simple task. Here we consider rooted binary tree graphs where the leaves (tips) denote living (extant) species and the internal nodes represent the point in time when a speciation event occurred. It is appropriate to call the internal nodes ancient species that branched into the living ones. If \( n \) is the number of leaves, then the total number of species (living and ancient) in the tree is \( N = 2n - 1 \).

An important feature of binary trees is their balance. In simple terms, a tree is balanced if it has equal numbers of leaves emanating from both branches of its bifurcations (internal nodes). By contrast, imbalance is the opposite property, unequal numbers of leaves arising from each branch (Shao and Sokal 1990). As the numbers of species in studies of phylogenetic trees has grown, so has the interest in using the shapes of these trees to test hypothesis about evolution. For example, if a few lineages give rise to most of the descendant species because they have acquired an important adaptation, this should be visible in the imbalance of the resulting tree (Felsenstein 2004). Simulations and comparisons with real phylogenies were performed by Costa et al (2019), showing that in a neutral model of evolution balanced trees evolve more frequently if speciation is sympatric (when individuals inhabit the same geographic region). Populations displaying geographical structures (parapatry and allopatry), on the other hand, give rise to more unbalanced trees.

Measures such as the index of Sackin (Sackin 1972; Shao and Sokal 1990; Fischer 2019), Colless (Colless 1982; Coronado et al 2020), and Total Cophenetic indices (Mir et al 2013; Cardona et al 2013) quantify tree balance. Here we focus on the Sackin and Total Cophenetic, as the relate directly to the new index we propose. The index of Sackin is calculated as the sum of all distances between the leaves \( (n \) terms) to the root node of the binary tree. The Total Cophenetic index takes all the distances from the most recent common ancestor (MRCA) of each pair of leaves to the root node and sums them up \( ((n \choose 2) \) terms). MRCA is the node that represents the closest internal node from
whom the pair of leaves descend directly.

In this paper, we propose a new topological balance metric based on the distances between tips, which we call $d_n$, the Area Per Pair (APP). Distances between tips are natural quantities, as they can be related to other distance measures, such as genetic and phenotypic differences between species or individuals. In sec. 2 we define the APP index and relate it with the Sackin and Total Cophenetic indices. Next we calculate the APP for the special cases of maximally balanced and fully unbalanced trees. In sec. 4 we obtain exact formulas for its expected value and variance under the Yule model and in sec. 5 we compare Sackin, Total Cophenetic and APP indices for a large set of empirical trees. Sec. 6 summarizes our conclusions.

2 The Area Per Pair (APP) Index

In phylogenetic trees, leaves (or tips) correspond to the living (or extant) species. In this paper we shall consider only rooted binary trees, where each node is either a leaf, an internal node (giving rise to two new branches), or the root. The distance between leaves $i$ and $j$ is defined as the number of edges that connect the corresponding tips in the tree, also called genealogic distance (Steel and McKenzie 2001; Mulder 2011):

$$d_{i,j} = d_{i,\rho} + d_{j,\rho} - 2\phi_{i,j},$$  \hspace{1cm} (1)

where $\rho$ is the root node, $d_{i,\rho}$ is the distance between $i$ and the root and $\phi_{i,j}$ is the cophenetic value of $(i, j)$, the distance between their most recent common ancestor (MRCA) and the root. Figure 1 shows a simple example illustrating these quantities. The Area Per Pair Index (APP), $\bar{d}_n$, is defined as the average of distances between all pairs of tips in the binary tree:

$$\bar{d}_n = \frac{2}{n(n-1)} \sum_{(i,j)} d_{i,j},$$  \hspace{1cm} (2)

where $n$ is the number of leaves in the tree. APP is a re-scaled version of the total tree area (Mulder 2011; Mir et al 2013):

$$D_n = \sum_{(i,j)} d_{i,j} = \frac{n(n-1)}{2} d_n$$  \hspace{1cm} (3)

The total area, in turn, is related to the Sackin index $-S_n$ (Sackin 1972; Shao and Sokal 1990; Fischer 2019) and Total Cophenetic value, $\Phi_n$, defined by Mir et al (2013):

$$S_n = \sum_{i=1}^{n} d_{i,\rho},$$  \hspace{1cm} (4)
and

$$\Phi_n = \sum_{i,j} \phi_{i,j}. \quad (5)$$

Summing over all pairs of leaves in eq. (1) we obtain:

$$\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{i,j} = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (d_{i,\rho} + d_{j,\rho}) - 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \phi_{i,j}.$$ 

or

$$D_n = (n-1)S_n - 2\Phi_n. \quad (6)$$

Comparing eqs. (3) and (6) we can write $d_n$ as function of $S_n$ and $\Phi_n$:

$$d_n = \frac{2}{n}S_n - \frac{4}{n(n-1)}\Phi_n. \quad (7)$$

In the next section we will calculate the APP index for two special cases of binary trees, namely, the fully unbalanced (rooted caterpillar) and maximally balanced trees. For the Sackin and Total Cophenetic indices these cases correspond to the maximum and minimum values attained for rooted binary trees [Fischer 2019; Mir et al 2013].

![Distances $d_{i,j}$ and $\phi_{i,j}$ in a hypothetical binary tree. Check metrics information in section 1. The green triangle represents the root $\rho$, nodes labeled from 1 to 4 are tips (extant species in phylogenetic trees) and blue squares are internal nodes, representing bifurcations (speciation events in phylogenetic trees).](image)

3 APP Index for some Special Trees

3.1 Fully Unbalanced Trees

Fully unbalanced trees are called caterpillars, characterized by all bifurcations occurring in only one branch, from the same leaf and internal node. Figure 2 shows examples of these trees. Exact expressions for the Sackin ($S_n$) and Total Cophenetic ($\Phi_n$) indices are known for these cases [Fischer 2019; Mir et al 2013]:

$$S_n^{(unb)} = \frac{(n-1)(n+2)}{2}, \quad (8)$$
Measuring Tree Balance with Normalized Tree Area

Fig. 2 Fully unbalanced trees with 2 to 4 leaves. The green triangle represents the root node \( \rho \), nodes labeled with numbers are extant species (tips) and blue squares are internal nodes.

\[
\phi_n^{(unb)} = \frac{n}{6} (n-1)(n-2). \tag{9}
\]

Using these results we can write eq. (7) as

\[
d_n^{(unb)} = \frac{2}{n} \frac{(n-1)(n+2)}{2} - \frac{4}{n(n-1)} \frac{n}{6} (n-1)(n-2) \\
= \frac{n+7}{3} - \frac{2}{n}, \quad \text{for } n > 1, \tag{10}
\]

where \( d_1^{(unb)} = 0 \). For fixed tree size \(- n \), the Sackin and Total Cophenetic indices have maximal values for binary caterpillar trees. For the APP index the value of \( d_n^{(unb)} \) is not always maximal, as we will discuss later at the end of section 3.2. We note that, for large \( n \), APP grows linearly with tree size, whereas the Sackin index grows quadratically and the Total Cophenetic index cubically.

3.2 Maximally Balanced Trees

A tree can be fully balanced only if number of leaves \( n \), is such that \( n = 2^s \), for \( s = 0, 1, 2, ... \). For \( n \neq 2^s \), however, there is at least one maximally balanced tree for any value of \( n \), as illustrated in Figure 3. Every fully balanced tree is also maximally balanced but the opposite is not true. Maximally balanced trees can be constructed following the algorithm in Fischer (2019). Starting with a fully balanced tree of size \( n' = 2^s \), leaves are removed sequentially so as to obtain a maximally balanced tree of size \( n \in (2^{s-1}, 2^s) \). In each step of the process, the cherry with the pair of leaves \((u, v)\) (such that \( d_u,\rho = d_v,\rho \), where \( \rho \) is the root, and \( d_u,v = 2 \)) is suppressed, and its parent node becomes a leaf. Figure 3 illustrates the process for \( n = 6 \), starting with the tree of size \( n = 8 \) and, in two steps, cherries (3, 4) and (5, 6) are deleted to form the maximally
balanced tree. **Exact expressions for** \( S_n \) **and** \( \Phi_n \) **for maximally balanced trees are given by** (Fischer 2019; Mir et al 2013):

\[
S_n^{(bal)} = n(\lceil \log_2 n \rceil + 1) - 2^{\lceil \log_2 n \rceil}, \quad (11)
\]

\[
\Phi_n^{(bal)} = \Phi_{n/2}^{(bal)} + \Phi_{n/2}^{(bal)} + \left( \frac{\lceil n/2 \rceil}{2} + \frac{\lfloor n/2 \rfloor}{2} \right), \quad \text{for } n > 2, \quad (12)
\]

where \( \Phi_1^{(bal)} = \Phi_2^{(bal)} = 0 \). Symbols \( \lfloor x \rfloor \) and \( \lceil x \rceil \) represent the floor and ceiling of \( x \). The APP index can now be computed from eqs. (7), (11) and (12).

For the fully balanced case \( (n = 2^s) \) we can also write closed formulas for all these expressions. For simplicity we write \( S_2^{(bal)} = S_s \) and \( \Phi_2^{(bal)} = \Phi_s \):

\[
S_s = s 2^s, \quad (13)
\]

\[
\Phi_s = 2\Phi_{s-1} + 2^{s-1}(2^{s-1} - 1), \quad \text{for } s > 1, \quad (14)
\]

with \( \Phi_{s=0} = \Phi_{s=1} = 0 \). Solving this recursive relation we also obtain

\[
\Phi_s = 2^{s-1}(2^s - s - 1), \quad (15)
\]

We finally obtain the expression for the APP index. For simplicity we write \( \bar{d}_s^{(bal)} = \bar{d}_s \).

\[
\bar{d}_s = \frac{S_s}{2^{s-1}} = \frac{\Phi_s}{2^{s-1}(2^s - 1)},
\]

\[
= \frac{s 2^s}{2^{s-1}} \frac{2^{s-1}(2^s - s - 1)}{2^{s-2}(2^s - 1)},
\]

\[
= \frac{2}{2^s - 1}[1 + 2^s(s - 1)], \quad \text{for } s > 0, \quad (16)
\]

with \( \bar{d}_{s=0} = 0 \). Using eq. (16) we can make approximations for maximally balanced trees which do not have \( n = 2^s \), but \( n \approx 2^s \):

\[
\bar{d}_n^{(bal)} \approx \frac{2}{n - 1}[1 + n(\log_2 n - 1)], \quad \text{for } n > 1, \quad (17)
\]
Table 1 Comparison between exact (=) and approximated (∼) formulas to APP index for maximally balanced trees. The values highlighted(*) are for fully balanced trees, where the approximation becomes exact.

| n  | 2*  | 3       | 4*  | 5       | 6       | 7       | 8*  | 9       |
|----|-----|---------|-----|---------|---------|---------|-----|---------|
| =  | 2   | 2.66667 | 3.33333 | 3.8   | 4.26667 | 4.57143 | 4.85714 | 5.16667 |
| ∼  | 2   | 2.75489 | 3.33333 | 3.80482 | 4.20391 | 4.55049 | 4.85714 | 5.13233 |

where $d_1^{(bal)} = 0$. In Table 1 we show the accuracy of this approximation for several values of $n$. For large $n$ the APP index for maximally balanced trees grows logarithmically with tree size ($d_n^{(bal)} \sim \log n$), which is slower than the Sackin ($S_n^{(bal)} \sim n \log n$) and Total Cophenetic indices ($\Phi_n^{(bal)} \sim n^2$).

For the Sackin and Total Cophenetic indices, the maximum and minimum values are reached for fully unbalanced and maximally balanced trees with size $n$, respectively (Fischer 2019; Mir et al 2013). Figure 4 shows that for $n > 10$ $d_n$ is larger for fully unbalanced trees than for maximally balanced ones. This indicates that APP index should be a good metric for measuring balance only for large $n$. In the next section, we present exact formulas for $d_n$ under the Yule Model.

### 4 APP Index under the Yule Model

#### 4.1 Expected Value

The Yule Model is a simple algorithm to generate stochastic binary trees (Harding 1971; Kirkpatrick and Slatko 1993; Steel and McKenzie 2001; Cardona et al 2013). The Yule algorithm starts with a cherry (a root and two leaves) and at each step of the process one tip is selected with uniform probability and is replaced by a cherry with two new tips. In this process different trees with the same number of tips are generated with different frequencies. The Yule model describes ensembles of random trees, with no specific elements affecting the evolutionary process. The expected value and variance for the Sackin ($S_n^{(bal)} \sim n \log n$) and Total Cophenetic indices ($\Phi_n^{(bal)} \sim n^2$) can be computed directly for such ensembles.

Let $d_n$ be a random variable associated with a tree of size $n$ created under the Yule model and whose value is the APP index of that tree. To calculate its expected value $E_Y[d_n]$ we will resort to expressions demonstrated previously in Cardona et al (2013) for the expected values of the Sackin and Total Cophenetic indices:

$$E_Y[S_n] = 2n \left[ H_n^{(1)} - 1 \right],$$

$$E_Y[\Phi_n] = n(n - 1) - 2n \left[ H_n^{(1)} - 1 \right],$$

where $H_n^{(1)}$ is the harmonic number of order 1.
Fig. 4 APP index for fully unbalanced and maximally balanced trees for some number of leaves – $n$. For $1 \leq n \leq 3$, the index coincides for the two cases since the trees are the same. For $4 \leq n \leq 9$, the index for balanced trees is larger than for unbalanced. For $n = 10$, the values coincide again, while for $n > 10$, the index for unbalanced trees is larger than for balanced trees.

where $H_n^{(r)} = \sum_{i=1}^{n} 1/i^r$ is the $n$th Harmonic Number of order $r$. Substituting the formulas above in eq. (7) we obtain the expected value of $\bar{d}_n$:

$$\mathbb{E}_Y[\bar{d}_n] = \frac{2}{n} \mathbb{E}_Y[S_n] - \frac{4}{n(n-1)} \mathbb{E}_Y[\Phi_n],$$

$$= \frac{2}{n} \left( 2n \left[ H_n^{(1)} - 1 \right] \right) - \frac{4}{n(n-1)} \left( n(n-1) - 2n \left[ H_n^{(1)} - 1 \right] \right),$$

$$= 4 \left\{ \left[ H_n^{(1)} - 1 \right] \left( \frac{n+1}{n-1} \right) - 1 \right\}, \text{ for } n > 1, \quad (20)$$

where $\mathbb{E}_Y[d_1] = 0$. For large $n$ the asymptotic behavior of the expected value of APP index under the Yule model has positive growth with tree size – $n$ (\(\mathbb{E}_Y[\bar{d}_n] \sim \log n\)). The growth with tree size is slower than the expected values for Sackin (\(\mathbb{E}_Y[S_n] \sim n \log n\)) and for Total Cophenetic indices (\(\mathbb{E}_Y[\Phi_n] \sim n^2\)).

Figure 5 shows the behavior of these expected values as a function of $n$. 
4.2 Variance

Let \( I_n \) be a balance index for a phylogenetic tree with \( n \) leaves and \( \mathbb{E}_Y[I_n] \) and \( \sigma^2_Y[I_n] \) the expected value and variance of this index under the Yule Model respectively. In order to compare trees with different number of leaves it is common to define a normalized index \( \text{Blum et al. [2006]} \) as:

\[
I_n^{(\text{norm})} = \frac{I_n - \mathbb{E}_Y[I_n]}{\sqrt{\sigma^2_Y[I_n]}}. \tag{21}
\]

Clearly \( \mathbb{E}_Y[I_n^{(\text{norm})}] = 0 \) and \( \sigma^2_Y[I_n^{(\text{norm})}] = 1 \). To construct a normalized index for \( \bar{d}_n \) we must calculate its variance under the Yule Model, \( \sigma^2_Y[\bar{d}_n] \). The following property of variance will be useful in our development:

\[
\sigma^2[AX + BY] = A^2\sigma^2[X] + B^2\sigma^2[Y] + 2AB\text{Cov}[X,Y], \tag{22}
\]

where \( X \) and \( Y \) are random variables, \( A \) and \( B \) are constants and \( \text{Cov}[X,Y] \) is the covariance of \( X \) and \( Y \). Applying this property in eq. (7) we get

\[
\sigma^2_Y[\bar{d}_n] = \frac{4}{n^2}\sigma^2_Y[S_n] + \frac{16}{n^2(n-1)^2}\sigma^2_Y[\Phi_n] - \frac{16}{n^2(n-1)}\text{Cov}_Y[S_n, \Phi_n]. \tag{23}
\]

The quantities \( \sigma^2_Y[S_n], \sigma^2_Y[\Phi_n] \) and \( \text{Cov}_Y[S_n, \Phi_n] \) were elegantly calculated using the fact that these indices satisfy recursive relations by Cardona et al. (2013). The final expressions are:

\[
\sigma^2_Y[S_n] = n^2 \left[ 7 - 4H_n^{(2)} \right] - n \left[ 1 + 2H_n^{(1)} \right], \tag{24}
\]

\[
\sigma^2_Y[\Phi_n] = \frac{n}{12} (n^3 - 10n^2 + 131n - 2) - 2n \left[ 2nH_n^{(2)} + 3H_n^{(1)} \right], \tag{25}
\]

\[
\text{Cov}[S_n, \Phi_n] = \frac{n}{6} (n^2 - 51n + 2) + 4n \left[ nH_n^{(2)} + H_n^{(1)} \right]. \tag{26}
\]
Finally, the variance under the Yule Model of the APP index \( \bar{d}_n \) is
\[
\sigma^2_Y[\bar{d}_n] = \frac{4}{3n(n-1)^2} \left[ 20n^3 + 49n^2 + 52n - 1 - 6(n + 1)(n + 5)H_n^{(1)} - 12n(n + 1)^2H_n^{(2)} \right]. \tag{27}
\]

Figure 6 shows the behavior of \( \sigma^2_Y[\bar{d}_n] \) in comparison with the variance of Sackin and the Total Cophenetic indices. For \( n \gg 1 \), this variance converges to an asymptotic value:
\[
\sigma^2_Y[\bar{d}_n] \simeq \frac{4}{3n^3} \left[ 20n^3 - 12n^3\frac{\pi^2}{6} \right] = \frac{8}{3}(10 - \pi^2) \simeq 0.3477, \tag{28}
\]
where we used \( H_n^{(2)} \simeq \zeta(2) = \frac{\pi^2}{6} \) for \( n \gg 1 \) and \( \zeta(r) \) is the Riemman Zeta Function. This asymptotic behavior is the main difference between the APP index and the other two metrics. For large trees the variance of Sackin and Total Cophenetic indices are
\[
\sigma^2_Y[S_n] \simeq n^2 \left[ 7 - 4\frac{\pi^2}{6} \right] \simeq 0.4203n^2 \tag{29}
\]
and
\[
\sigma^2_Y[\Phi_n] \simeq \frac{n^4}{12} \simeq 0.0833n^4. \tag{30}
\]

Figure 7 shows the variances for \( 1 \leq n \leq 256 \). Convergence of the APP index variance is very slow; for \( n = 6000 \), for instance, \( \sigma^2_Y[\bar{d}_n] \simeq 0.3402 \). Figure 8 shows results for an ensemble of 50,000 trees generated with the Yule Model for several values of \( n \) where the Sackin, Total Cophenetic, and APP indices were computed. In the left panels (Sackin – \( S_n \) and Total Cophenetic – \( \Phi_n \)), the dispersion of the distribution for each \( n \) increases with \( n \), more prominently for \( \Phi_n \) in the bottom panel. In contrast, on the right panel (APP index \( \bar{d}_n \)), the dispersion is nearly constant, increasing on slightly with \( n \).

Although there is a remarkable difference between the variance of the different metrics, their normalized versions (according to eq. (21)) erase all relevant distinctions. For \( n \geq 10 \), we plot the maximum (unbalanced trees) and minimum (balanced trees) value of each normalized index. As expected, the spread between the maximum and minimum values is similar for the three indices, Figure 9.

5 Empiric Trees

In order to understand how the distribution of the different balance metrics represents the distribution of real phylogenetic trees, we analyzed empiric phylogenetic trees available at the TreeBASE [Piel et al. 2000; Vos et al. 2012]. We generated a subset among the available trees using three filters: in the field kind
we used “Species Tree” (no “Gene Tree” and “Barcode Tree” were included), and in the field type we used “Single” (no consensus tree were included). With this procedure, we end up with 9805 phylogenetic trees. Among these trees, we created a subset based on the number of tips, by defining the ntaxa field with trees above ten species and below 900 species. After this filtering, we ended up with 9307 trees. Of this subset, only 8999 trees were used, because 308 trees had issues when reading the files from the database for calculating the metrics. We used R (R Core Team 2017) for reading and calculating the balance metrics for the empiric trees. We show the number of tips distribution of the empiric trees in the Figure 10. Figures 11 and 12 show the distribution of the balance indices for the empiric trees. Our index proposal (APP) shows a larger concentration of trees in low values of the index, while the Sackin and
Fig. 8 Black dots: values of indices for the fully unbalanced and maximally balanced trees. Gray dots: values for an ensemble of 50,000 trees obtained from the Yule Model. Left Upper Panel: Sackin Index. Left Bottom Panel: Total Cophenetic index. Right Panel: APP index. Orange, blue and green regions represent the interval of one standard deviation around the mean for the Sackin, Total Cophenetic and APP indices respectively.

Fig. 9 Left Upper Panel: Maximum (unbalanced trees) and minimum (balanced trees) for the normalized Sackin index. Left Bottom Panel: same for the normalized Total Cophenetic index. Right Panel: same for the Normalized APP index. For all indices, the unbalanced and balanced cases depart from zero in very similar ways.

Total Cophenetic indices present a larger concentration in intermediate values. After the normalization procedure described in eq. (21), the three normalized indices investigated tends to show very similar distributions.

We compare the three indices for this set of empirical trees. Figure 13 shows a larger concentration of trees near the region delimited by the interval \((E_Y[d_n] - \sigma_Y[d_n], E_Y[d_n] + \sigma_Y[d_n])\) than for \(S_n\) and \(\Phi_n\). These intervals are represented by the colored areas. The green area for the APP index captures
Fig. 10 Distribution of the number of tips (leaves) in the set of real phylogenies used for comparison. After filtering the original set, regarding only species trees with $n \in [10, 900]$, we analyzed the three indices in a set of 8999 empiric trees.

Fig. 11 Left Upper Panel: distribution of trees with Sackin Index $S_n$. Left Bottom Panel: same for the Total Cophenic index $\Phi_n$. Right Panel: same for the APP index $\bar{d}_n$. Our proposal shows a larger concentration of trees in low values, while the previous indices present a larger concentration in intermediate values.
the distribution of empirical trees better than the Sackin and Total Cophenetic indices. Figure 12 shows the same data in terms of the respective normalized indices, where the colored intervals collapse to a narrow window delimited by $(-1, 1)$, since in this case all expected values are null and the variance is unitary. In this case all indices seem to perform equally well.

\[ \text{Fig. 12} \quad \text{Left Upper Panel: distribution of trees with normalized Sackin Index } S_{n}^{(\text{norm})}. \text{ Left Bottom Panel: same for the Total Cophenetic index } \Phi_{n}^{(\text{norm})}. \text{ Right Panel: same for the APP index } \bar{d}_{n}^{(\text{norm})}. \text{ The normalization procedure described in eq. (21) tends to equalize the distribution of the three indices.} \]

\[ \text{Fig. 13} \quad \text{Left Upper Panel: Sackin index for the set of empirical trees. Left Bottom Panel: same for the Total Cophenetic index. Right Panel: same for the APP index.} \]
6 Conclusions

The organization of species according to their genetic or phenotypic similarity is usually represented by phylogenetic trees. The shape, branch lengths and distribution of bifurcation points on the tree allows us to reconstruct the evolutionary history of the group. Because all living species have a single common ancestor, that lived over a billion years ago (Steel and Penny 2010; Theobald 2010), the phylogenetic construction can be applied to all species, from small groups that speciated recently to the entire biome. Understanding the structure of these trees is, therefore, key to unravel how life evolved in our planet.

In this paper, we have proposed a new index to measure tree balance based on the mean distance between extant species, the Area Per Pair (APP) index. This index has the remarkable property that its variance converges under the Yule Model to a constant value for large phylogenies (28). Using previously derived expressions for the Sackin and the Total Cophenetic indices, we calculated exact formulas for the APP index for fully unbalanced (10) and fully balanced (16) trees.

For ensembles of trees generated with the Yule model, both the expected value and variance were calculated exactly for the APP index, eqs. (20) and (27). The average of APP index grows slower with tree size when compared to the other two indices. The variance of APP index presents the remarkable property of reaching a constant asymptotic value for large trees – eq. (28). For the Sackin and Cophenetic indices, on the other hand, the variance always grows with $n$ – eqs. (29) and (30). As the variance of APP index tends to a
constant, large trees fall in a narrow range of possible values (Figure 8).

Finally, we calculated the non-normalized and normalized versions of the indices for a large set of empirical phylogenetic trees. We observed a higher concentration of data near the region delimited by the interval \( (\bar{d}_n - \sigma_Y, \bar{d}_n + \sigma_Y) \) for the APP index (Figures 13 and 14). These results suggest that APP is a more suitable index for comparing balance between different (and large) tree sizes.

References

Blum MG, François O, Janson S, et al (2006) The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. The Annals of Applied Probability 16(4):2195–2214

Cabrál JS, Valente L, Hartig F (2017) Mechanistic simulation models in macroecology and biogeography: state-of-art and prospects. Ecography 40(2):267 – 280

Cardona G, Mír A, Rosselló F (2013) Exact formulas for the variance of several balance indices under the yule model. Journal of Mathematical Biology 67(1):1833 – 1846

Colless DH (1982) Review of phylogenetics: The theory and practice of phylogenetic systems. Systematic Zoology 31(1):100 – 104

Coronado TM, Fischer M, Herbst L, Rossell F, Wicke K (2020) On the minimum value of the colles index and the bifurcating trees that achieve it. arXiv

Costa CLN, Lemos-Costa P, Marquitti FMD, Fernandes LD, Ramos MF, Schneider DM, Martins AB, de Aguiar MAM (2019) Signatures of microevolutionary processes in phylogenetic patterns. Systematic Biology 68(1):131 – 144

Felsenstein J (2004) Inferring phylogenies. Sinauer Associates

Fischer M (2019) Extremal values of the sackin balance index for rooted binary trees. arXiv

Harding EF (1971) The probabilities of rooted tree-shapes generated by random bifurcation. Advances in Applied Probability 3(1):44 – 77

Kirkpatrick M, Slatko B (1993) Search for evolutionary patterns in the shape of phylogenetic trees. Evolution 47:1171 – 1181

Mír A, Rosselló F, Rotger L (2013) A new balance for phylogenetic trees. Mathematical Biosciences 241(1):125 – 136

Mooers AO, Heard SB (1997) Inferring evolutionary process from phylogenetic tree shape. The quarterly review of Biology 72(1):31–54

Morlon H (2014) Phylogenetic approaches for studying diversification. Ecology Letters 17(4):508 – 525

Mulder WH (2011) Probability distributions of ancestries and genealogical distances on stochastically generated rooted binary trees. Journal of Theoretical Biology 280(1):139 – 145

Nee S, Mooers AO, Harvey PH (1992) Tempo and mode of evolution revealed from molecular phylogenies. PNAS 89(17):8322 – 8326

Piel WH, Donoghue M, Sanderson M, Netherlands L (2000) Treebase: a database of phylogenetic information. In: Proceedings of the 2nd International Workshop of Species 2000

R Core Team (2017) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL https://www.R-project.org/

Sackin M (1972) good and bad phenograms. Systematic Biology 21(2):225–226

Shao KT, Sokal RR (1990) Tree balance. Systematic Zoology 39(3):266 – 276

Steel M, McKenzie A (2001) Properties of phylogenetic trees generated by yule-type speciation models. Mathematical Biosciences 170(1):91 – 112

Steel M, Penny D (2010) Common ancestry put to the test. Nature 465:168 – 169

Theobald DL (2010) A formal test of the theory of universal common ancestry. Nature 465(7295):219 – 222
Vos RA, Balhoff JP, Caravas JA, Holder MT, Lapp H, Maddison WP, Midford PE, Priyam A, Sukumaran J, Xia X, et al (2012) Nexml: rich, extensible, and verifiable representation of comparative data and metadata. Systematic Biology 61(4):675 – 689