bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Robust, Universal Tree Balance Indices

Jeanne Lemant[1,2,3], Cécile Le Sueur[1] Veselin Manojlović,[4], and Robert Noble[1,4,*]

[1] *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland*

[2] *Current address: Swiss Tropical and Public Health Institute, Basel, Switzerland*

[3] *Current address: University of Basel, Basel, Switzerland*

[4] *Department of Mathematics, City, University of London, London, UK*

**robert.noble@city.ac.uk*

Abstract

Balance indices that quantify the symmetry of branching events and the compactness of trees are widely used to compare evolutionary processes or tree-generating algorithms. Yet existing indices have important shortcomings, including that they are unsuited to the tree types commonly used to describe the evolution of tumours, microbial populations, and cell lines. The contributions of this article are twofold. First, we define a new class of robust, universal tree balance indices. These indices take a form similar to Colless' index but account for node sizes, are defined for trees with any degree distribution, and enable more meaningful comparison of trees with different numbers of leaves. Second, we show that for bifurcating and all other full m-ary cladograms (in which every internal node has the same out-degree), one such Colless-like index is equivalent to the normalised reciprocal of Sackin's index. Hence we both unify and generalise the two most popular existing tree balance indices. Our indices are intrinsically normalised and can be computed in linear time. We conclude that these more widely applicable indices have potential to supersede those in current use.

Tree balance indices – most notably those credited to Sackin (1972) and Colless (1982) – are widely used to describe speciation processes, compare cladograms, and assert the correctness of tree reconstruction methods (Shao and Sokal, 1990; Mooers and Heard, 1997). These indices have recently been introduced to oncology (Chkhaidze et al., 2019; Scott et al., 2020) because methods for determining and classifying modes of tumour evolution have clinical value (Maley et al., 2017). A problem here is that the trees that best describe tumour evolution are clone trees in which node sizes are informative and which frequently contain linear sections; indeed, developing methods to distinguish linear from branching tumour evolution is an important area of ongoing research (Davis et al., 2017). Existing tree balance indices are unsuited to these topologies and take no account of node size. Moreover, even when applied only to bifurcating cladograms, existing indices are unreliable for comparing trees with different numbers of leaves.

Here we develop a new class of robust, universal tree balance indices. Our definitions not only extend the tree balance concept and open up new applications but also unify the two main approaches to quantifying balance as proposed by Sackin and Colless. We describe several general advantages of our indices compared to those in current use.

Materials and Methods

Rooted trees

We consider exclusively rooted trees in which all edges are oriented away from the root (which will be topmost in our figures). This orientation defines a natural order on the tree, from top to bottom: all edges are assumed to extend from the root to the other *internal nodes* and finally to the terminal nodes or *leaves*. The *out-degree* of a node i, written $d^+(i)$, is the number of direct descendants, ignoring any descendant branches in which all nodes have zero size. Internal nodes have out-degree at least one, whereas leaves have out-degree zero.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

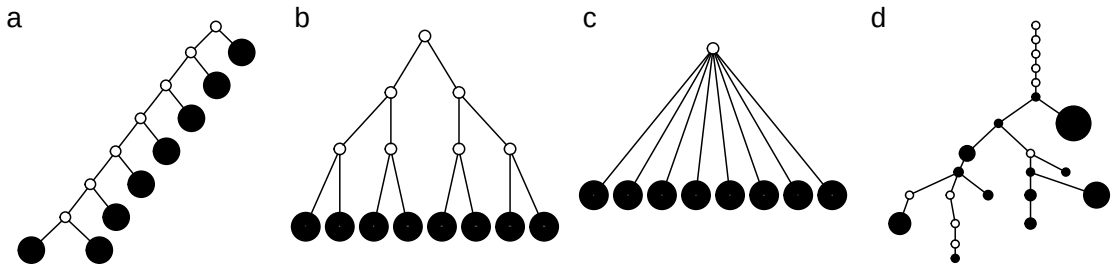

Figure 1: Contrasting trees. **a**: Caterpillar tree with $I_S = 35$, $I_{S,norm} = 1$, $I_C = 21$, $I_{C,norm} = 1$, $I_\Phi = 56$, $I_{\Phi,norm} = 1$. **b**: Fully symmetric bifurcating tree with $I_S = 24$, $I_{S,norm} \approx 0.59$, $I_C = I_{C,norm} = 0$, $I_\Phi = 16$, $I_{\Phi,norm} \approx 0.29$. **c**: Star tree with $I_S = 8$, $I_{S,norm} = 0$, I_C and $I_{C,norm}$ undefined, $I_\Phi = I_{\Phi,norm} = 0$. **d**: Clone tree of the lung tumour CRUK0065 in the TRACERx cohort (Jamal-Hanjani et al., 2017). In the clone tree, nodes represented by empty circles correspond to extinct clones, and the diameters of other nodes are proportional to the corresponding clone population sizes.

Some tree types have particular names. A *caterpillar tree* (Fig. 1a) is a bifurcating tree in which each internal node has one leaf. A *fully symmetric* tree (Fig. 1b) is such that every internal node with the same depth has the same degree or, equivalently, for each internal node i all the subtrees rooted at i are identical. A *star tree* (Fig. 1c) is a tree whose leaves are all attached to the root, which is the only internal node.

Cladograms, species trees and clone trees

Cladograms are trees that represent relationships between extant biological taxa (leaves) via edges linking them to hypothetical extinct ancestors (internal nodes). A common conception is that only bifurcating cladograms can be considered fully resolved and linear parts are inadmissible. However, linear sections in cladograms are appropriate for representing anagenesis (in which a descendant replaces its ancestor), while budding (in which an ancestor produces a descendant and remains extant) can give rise to cladogram nodes with out-degree greater than two (Podani, 2013). An extant ancestor is represented in a cladogram by a leaf stemming from the internal ancestor node (so the two nodes represent the same taxon).

An alternative way to represent extant ancestors is as internal nodes (like in a genealogy with overlapping generations). Such diagrams are known to organismal biologists as species trees and to oncologists as clone trees. In a clone tree, each node represents a clone (a set of cells that share alterations of interest due to common descent) and edges represent the chronology of alterations. Clone tree nodes can have any out-degree, including $d^+ = 1$, and each node – including internal nodes – can be associated with a non-negative size, related to the clone population size at the time of observation (as in Figure 1d). The size of a tree or subtree can then be defined as the sum of its node sizes.

When nodes are associated with sizes, the addition or removal of even vanishingly small terminal branches can change leaves into internal nodes or vice versa and so substantially change the value of existing tree balance indices. This behaviour is unsatisfactory because these small branches typically represent either newly-created types that have yet to experience evolutionary forces or types on the verge of extinction, and in either case their relative sizes convey negligible information about the mode of evolution. Data sets may also omit rare types due to sampling error or because genetic sequencing methods have imperfect sensitivity (Turajlic et al., 2018).

The change due to the addition of terminal nodes is greater when the tree is a cladogram rather than a species or clone tree. For example, when a three-node, two-leaf tree (Fig. 2a) is augmented by adding a node j to a leaf i (Fig. 2b), the three original nodes retain their positions in the species or clone tree (middle column of Figure 2), but in the cladogram (right column) node i becomes two nodes (i_1 and i_2), the larger of which is now further from the root. As the size of the new node j is continuously reduced to zero, the species or clone tree changes continuously, whereas the cladogram undergoes an abrupt change of topology when the size of node j reaches zero. We conclude that

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

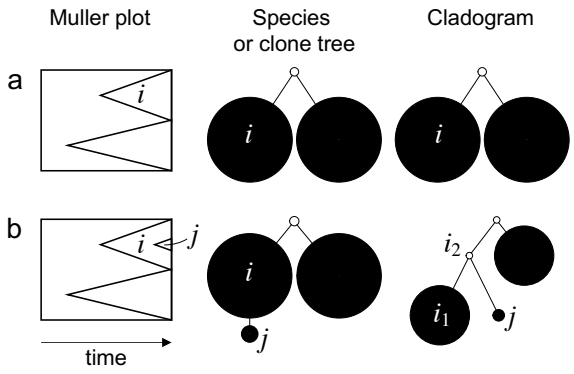

Figure 2: Muller plots (left column), species or clone trees (middle column), and cladograms (right column) representing evolution by splitting only (a) and both splitting and budding (b). Nodes represented by empty circles correspond to extinct types.

the species or clone tree representation is more robust than the cladogram representation in the general case in which nodes are associated with sizes and ancestors can be extant. Also an index that accounts for non-zero internal node sizes can be made more robust than one that does not.

Existing tree balance indices

The most widely used tree balance indices are in fact imbalance indices, such that more balanced trees are assigned smaller values. These indices were introduced to study cladograms and take no account of node size. The most popular are Sackin's index and Colless' index.

Sackin's index

Let T be a tree with set of leaves $L(T)$. For a leaf $l \in L(T)$, let ν_l be the number of internal nodes between l and the root, which is included in the count. Then the index credited to Sackin (1972) is

$$I_S(T) = \sum_{l \in L(T)} \nu_l.$$

For two bifurcating trees on the same number of leaves, a less balanced tree has higher values of ν as the tree is in a sense less compact (compare trees **a** and **b** in Figure 1).

Since the value tends to increase with the number of nodes, Shao and Sokal (1990) proposed normalising I_S with respect to trees on $n > 2$ leaves by subtracting its minimum possible value for such trees and then dividing by the difference between the maximum and minimum possible values. The minimal I_S is reached on the star tree, such as tree **c** in Figure 1, and hence $\min_n(I_S) = n$. The maximum is attained on the caterpillar tree, such as tree **a**:

$$\max_n(I_S) = n - 1 + \sum_{\nu=1}^{n-1} \nu = n - 1 + n(n-1)/2 = (n-1)(n+2)/2.$$

The normalised index is then

$$I_{S,norm}(T) = \frac{I_S(T) - n}{(n+2)(n-1)/2 - n}.$$

This normalised index is not very satisfactory as a balance index because it fails to capture an intuitive notion of balance. For example, it is not obvious why fully symmetric tree **b** should be considered less balanced than star tree **c** in Figure 1, yet its $I_{S,norm}$ value is much larger. To address this issue, Shao and Sokal (1990) further suggested normalising I_S relative to its extremal values among trees with the same number of internal nodes as well as the same number of leaves. But even then the index remains unreliable for comparing trees with different numbers of leaves. For example, the index is 1 for every caterpillar tree, yet long caterpillar trees are intuitively less balanced than

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

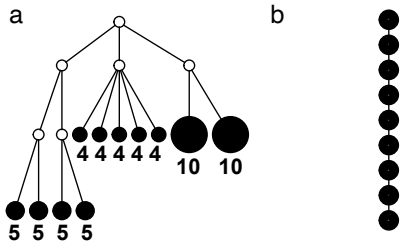

Figure 3: Trees with extremal J values. Numbers shown below nodes are node sizes. Empty nodes have null size. **a**: A tree in which each internal node has null size and splits its descendants into subtrees of equal size, and hence $J = 1$. This tree can be considered balanced only according to an index that accounts for node size. **b**: A linear tree, for which $J = 0$.

short ones. The conventional I_S normalisations are not defined for trees containing linear parts. Moreover, since I_S doesn't account for node size, it is highly sensitive to the addition or removal of relatively tiny terminal branches. Hence Sackin's index is neither universal nor robust.

Colless' index

For an internal node i of a bifurcating tree T, define n_{i_1} as the number of leaves of the left branch of the subtree rooted at i, and n_{i_2} as the number of leaves of the right branch. Then the index defined by Colless (1982) is

$$I_C(T) = \sum_{i \in \widetilde{V}(T)} |n_{i_1} - n_{i_2}|,$$

where $\widetilde{V}(T)$ is the set of internal nodes. The index can be normalised for the set of trees on $n > 2$ leaves by dividing by its maximal value, $\binom{n-1}{2}$, which is reached on the caterpillar tree (as in Figure 1a).

To generalise Colless' index to multifurcating trees, Mir et al. (2018) recently introduced a family of Colless-like balance indices, including I_C as a special case. Each of these indices $\mathfrak{C}_{D,f}$ is determined by a weight function f, which assigns a size to each subtree as a function of its out-degree, and a dissimilarity function D. By definition of D, Colless-like indices are zero if and only if each internal node divides its descendants into subtrees of equal size according to f. But since these indices are normalised by dividing by the maximal value for trees on the same number of leaves, they are unreliable for comparing trees with different numbers of leaves. In common with Sackin's index, the total cophenetic index I_Φ (Mir et al., 2013) (see Appendix), and other existing indices, the Colless-like indices so far defined are neither universal nor robust.

Desirable properties of a universal, robust tree balance index

Our aim is to derive a tree balance index J that is useful for classifying and comparing rooted trees that can have any distributions of node degrees and node sizes. Here we specify five desirable properties that such an index should have. The first two axioms relate to extrema and universality, in the sense of an index being defined for trees with any degree distribution. The other three axioms are concerned with robustness and are relevant only when nodes can have arbitrary sizes.

Conventionally, a tree is considered maximally balanced only if every internal node splits its descendants into subtrees on the same number of leaves (Shao and Sokal, 1990). We generalise this concept by requiring that every internal node splits its descendants into at least two subtrees of equal size, as in Figure 3a. We term this the *equal splits* property. We then set necessary and sufficient conditions for maximal balance:

Axiom 0.1 (Maximum value)**.** *$J(T) \leq 1$ for all trees T, and $J(T) = 1$ only if T has equal splits. Furthermore, if T has equal splits and every internal node of T has null size (or equivalently represents an extinct taxon) then $J(T) = 1$.*

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Another convention is that narrow trees with relatively many internal nodes are considered highly imbalanced. Linear trees (that is, trees in which every node i has $d^+(i) \leq 1$, as in Figure 3b) are even narrower than caterpillar trees. Also the most unequal binary split is one that assigns all descendants to one branch and none to the other. Hence our second desirable property:

Axiom 0.2 (Minimum value)**.** *$J(T) \geq 0$ for all trees T, and $J(T) = 0$ if and only if T is a linear tree.*

Our third desirable property is that our index should be insensitive to the presence of uninformative terminal branches:

Axiom 0.3 (Leaf limit)**.** *Let T be a tree with finitely many nodes and l be a leaf of T. Suppose we create a new tree T' by adding to T a subtree T_l with finitely many nodes, rooted at l. As the size of T_l excluding its root approaches zero, so $J(T') \to J(T)$.*

Our fourth desirable property ensures that a linear section of a tree is regarded as a maximally unequal split:

Axiom 0.4 (Linear limit)**.** *Let j be a node of a tree T with $d^+(j) = 1$. Suppose we create a new tree T' by adding to T a subtree with finitely many nodes, rooted at j. As the size of T_j excluding its root approaches zero, so $J(T') \to J(T)$.*

Lastly, we require continuity with respect to varying node size:

Axiom 0.5 (Continuity)**.** *If the population of any node of any tree T varies continuously in $\mathbb{R}_{>0}$, then $J(T)$ varies continuously.*

The wording of Axiom 0.1 raises an important question: Can trees with non-zero-sized internal nodes be considered maximally balanced? The following proposition provides the answer.

Proposition 0.6. *Axioms 0.3 and 0.4 each imply that equal splits are not sufficient for maximal balance.*

Proof. Suppose that equal splits are sufficient for maximal balance. First consider a one-node tree T. If we add a vanishingly small linear subtree to T then the new tree T' will have $J(T') = 0$. But if we instead add two vanishingly small subtrees of equal size to T then we obtain $J(T') = 1$. This implies that whatever value we assign to $J(T)$, we cannot satisfy Axiom 0.3. Second, consider a linear tree T in which the sum of the non-root node sizes is δ. Then $J(T) = 0$. But if we add another subtree to the root, also of size δ, then the new tree T' will have $J(T') = 1$, even as $\delta \to 0$. This contradicts Axiom 0.4. □

We therefore face a choice: either weaken Axioms 0.3 and 0.4 or accept that equal splits are not sufficient for maximal balance. We choose the second option (and as a corollary obtain $J = 0$ for the single-node tree) because we want our indices to be not only universal but also highly robust when applied to real, imperfect data. We will further argue that this choice is appropriate from a biological viewpoint and is consistent with the ideas underlying previous tree balance indices.

Results

General definition of universal, robust tree balance indices

Before defining a new class of balance indices we need to introduce some more notation. For a tree T, we will use $V(T)$ to denote the set of all nodes of T, which we will abbreviate to V when the identity of the tree is unambiguous. Let $f(v) \geq 0$ denote the size of node v (not necessarily a function of the out-degree). Then S_i denotes the size of the subtree T_i rooted at i, and S_i^* is the size of T_i excluding its root:

$$S_i := \sum_{v \in V(T_i)} f(v); \quad S_i^* := \sum_{\substack{v \in V(T_i) \\ v \neq i}} f(v) = S_i - f(i).$$

We will use $\widetilde{V}(T)$ or simply $\widetilde{V}$ to denote the set of all internal nodes such that $\{i \in \widetilde{V}\} := \{i \in V : S_i^* > 0\}$.

We then introduce three continuous functions of subtree sizes:

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- An *importance* factor $g : \mathbb{R}_{>0} \to \mathbb{R}_{>0}$ with $g(x) \to 0$ as $x \to 0$;
- A *non-root dominance* factor $h : \mathbb{R}_{>0} \times \mathbb{R}_{>0} \to (0, 1]$ with $h(x_1, x_2) \to 0$ as $x_1 \to 0$, and $h(x_1, x_2) = 1$ if and only if $x_1 = x_2$;
- A *balance score* W that assigns $W_i \in [0, 1]$ to each internal node i such that $W_i = 0$ if and only if $d^+(i) = 1$, and $W_i = 1$ if and only if i splits its descendants into at least two equally sized subtrees.

To allow us to define W more rigourously, let $\mathcal{S}$ denote the set of vectors with positive components that sum to unity:

$$\mathcal{S} \coloneqq \cup_{k\geq 1}\{(x_1, \ldots, x_k)|x_1, \ldots, x_k > 0, x_1 + \ldots + x_k = 1\}.$$

Then $W \colon \mathcal{S} \to [0, 1]$ is such that, for all $(x_1, \ldots, x_k) \in \mathcal{S}$:

- For every permutation π, $W(x_1, \ldots, x_k) = W(x_{\pi(1)}, \ldots, x_{\pi(k)})$;
- $W(x_1, \ldots, x_k) = 1$ if and only if $k > 1$ and $x_1 = \ldots = x_k$;
- $W = 0$ if and only if $\max(x_1, \ldots, x_k) = 1$;
- W is a continuous function with respect to each of its arguments.

We then define a balance index in terms of subtree sizes as

$$J \coloneqq \frac{1}{\sum_{k\in\widetilde{V}} g(S_k^*)} \sum_{i\in\widetilde{V}} g(S_i^*)h(S_i^*, S_i)W_i, \tag{1}$$

where $W_i = W(S_{i_1}/S_i^*, \ldots, S_{i_p}/S_i^*)$ and $i_1, \ldots, i_p$ are the children of node i. A short proof that this type of index satisfies our five axioms for robustness and universality is presented in the Appendix.

Interpretation of factors W, g and h

The balance score W in our general definition (Equation 1) measures the extent to which an internal node splits its descendants into equally sized subtrees. The importance factor g assigns more weight to nodes that are the roots of large subtrees. In biological terms, this means giving more weight to types that have more descendants. The continuous function h quantifies the extent to which a node should be considered a leaf (which doesn't contribute to determining tree balance in Colless-like indices) as opposed to an internal node (which does). From a biological point of view, nodes that are large relative to their descendants represent extant populations whose evolutionary fate remains largely undetermined.

Factors g and h together ensure that our indices consider a tree imbalanced unless there is strong evidence to the contrary. For example, the one-node tree provides no evidence and is considered maximally imbalanced. The inclusion of h also means that a tree can achieve maximal balance only if all its internal nodes have zero size, which is equivalent to all ancestors being extinct, as in a cladogram. This requirement can be removed simply by omitting h from the definition, but then, as per Proposition 0.6, our robustness Axioms 0.3 and 0.4 will not be satisfied.

Sackin's and Colless' indices similarly assign more weight to nodes that have more descendant leaves or are closer to the root. As Mooers and Heard (1997) have remarked, it is reasonable to put more weight on nodes deeper within the tree because "those nodes are the most informative, as the subclades they define are older and therefore sample longer periods of evolutionary time."

A specific index based on the Shannon entropy

In defining a specific index, we start by opting for the simplest choices of importance and non-root dominance factors:

$$g(x) = x, \quad h(x_1, x_2) = x_1/x_2.$$

The role of the balance score function W is to quantify the extent to which a set of objects (specifically subtrees) have equal size. A well-known index that satisfies the necessary conditions is the normalised Shannon entropy.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

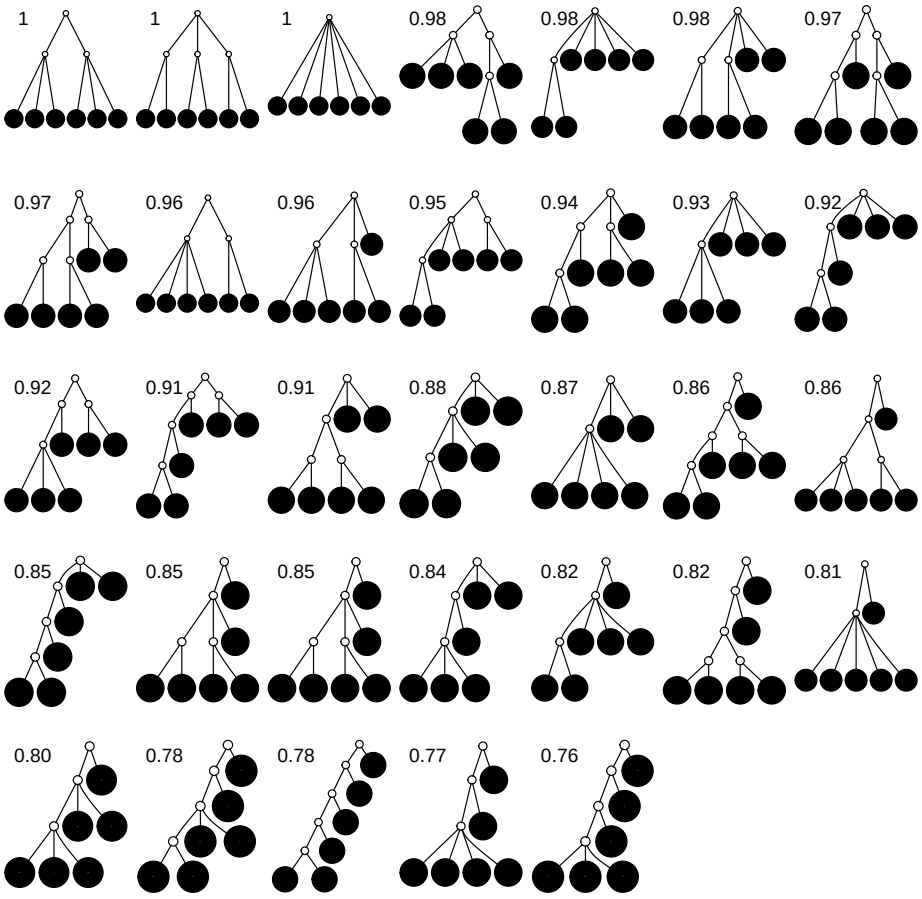

Figure 4: All multifurcating cladograms on six leaves without linear parts and with equally sized leaves, sorted and labelled by J^1 value.

Assume a population is partitioned into $n \in \mathbb{N}$ types, with each type i accounting for a proportion p_i. Then the Shannon entropy with base b is defined as ${}^1H_b := -\sum_{i=1}^{n} p_i \log_b p_i$. If all types have equal frequencies $p_i = 1/n$ then ${}^1H_b = \log_b n$. If the types have unequal sizes then ${}^1H_b < \log_b n$. And if the abundance is mostly concentrated on one type j, such that $p_j \to 1$, then ${}^1H_b \to 0$.

Let $C(i)$ denote the set of children (immediate descendants) of a node i, and for $j \in C(i)$ let $p_{ij} := S_j/S_i^*$ denote the relative size of subtree T_j compared to all subtrees attached to i. A balance score based on the normalised Shannon entropy is then

$$W_i^1 = \sum_{j \in C(i)} W_{ij}^1, \quad \text{with } W_{ij}^1 = \begin{cases} -p_{ij} \log_{d^+(i)} p_{ij} & \text{if } p_{ij} > 0 \text{ and } d^+(i) \geq 2, \\ 0 \quad \text{otherwise.} \end{cases} \tag{2}$$

From aforementioned properties of the Shannon entropy, it follows that $W_i^1 \in [0, 1]$, with $W_i^1 = 0$ if and only if $d^+(i) = 1$, and $W_i^1 = 1$ if and only if i splits its descendants into at least two equally sized subtrees. Therefore the following specific balance index satisfies our robustness and universality axioms:

$$J^1 := \frac{1}{\sum_{k \in \widetilde{V}} S_k^*} \sum_{i \in \widetilde{V}} S_i^* \frac{S_i^*}{S_i} W_i^1. \tag{3}$$

The definition simplifies when we restrict the domain to the set of multifurcating cladograms in which all leaves have equal size f_0 (corresponding to equally important extant types) and internal nodes have zero size (representing extinct ancestors). For all internal nodes i in such trees, $S_i^* = S_i = f_0 n_i$, where n_i is the number of leaves of the subtree rooted at i. The general definition of

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Equation 1 then becomes a weighted average of node balance scores:

$$J = \frac{1}{\sum_{k \in \widetilde{V}} n_k} \sum_{i \in \widetilde{V}} n_i W_i, \tag{4}$$

and the specific definition of Equation 3 becomes

$$J^1 = \frac{-1}{\sum_{k \in \widetilde{V}} n_k} \sum_{i \in \widetilde{V}} \sum_{j \in C(i)} n_j \log_{d^+(i)} \frac{n_j}{n_i}. \tag{5}$$

For example, Figure 4 shows the J^1 values of all cladograms on six leaves without linear parts. Unlike I_S and I_C, J^1 does not consider the caterpillar tree the least balanced of these cladograms.

There are of course many alternative options for W. Since the Shannon entropy belongs to a family of generalised entropies ${}^qH$ (Chao et al., 2014), the above reasoning can be generalised to define a balance score W^q, and hence a robust, universal balance index J^q, for every $q > 0$ (see Appendix). Other candidates for W include one minus the variance of the proportional subtree sizes, or one minus the mean deviation from the median (Mir et al., 2018). We prefer W^1 mostly because, as we shall show, it is the only function for which Equation 4 is a generalisation of the normalised inverse Sackin index.

Relationship with Colless' index

Like Colless' index and Colless-like indices as previously defined, our new family of tree balance indices is based on the intuitive idea of assigning a value to each internal node, summing these values, and then normalising the sum. A Colless-like index in the sense of Mir et al. (2018) depends on a function $f : \mathbb{N} \to \mathbb{R}_{\geq 0}$, which assigns node sizes, and a dissimilarity score $D : \mathcal{R} \to \mathbb{R}_{\geq 0}$, where $\mathcal{R}$ is the set of non-null real vectors. Before normalisation, such an index has the form

$$\mathfrak{C}_{D,f} = \sum_{i \in \widetilde{V}} D(\delta_f(T_{i_1}), \ldots, \delta_f(T_{i_k})),$$

where $\{i_1, \ldots, i_k\}$ are the children of node i. The function δ_f assigns a size to each subtree by summing the node sizes: $\delta_f(T) = \sum_{j \in V(T)} f(d^+(j))$. Neglecting the initial normalising factor, our general definition (Equation 1) has a similar form and can be considered Colless-like in only a slightly broader sense. Our definition nevertheless differs in three important ways.

First, whereas the unbounded dissimilarity index D measures both the relative imbalance of subtrees and their combined size, and is undefined for nodes with out-degree one, we split these two roles into a normalised balance score W and an unbounded importance factor g and – crucially – we assign a W value (specifically zero) to nodes with out-degree one. This difference enables us to extend the balance index definition to trees with any degree distribution. It also makes it easy to normalise our index for any tree, simply by dividing by the sum of the importance factors. Furthermore, our normalisation is universal, rather than being based on comparison with other trees with the same number of leaves. For example, our index judges long caterpillar trees less balanced than short ones (Fig. 5a), whereas Sackin's index, Colless' index, and the total cophenetic index consider all caterpillar trees on more than two leaves equally imbalanced.

Second, we multiply the balance score by an additional non-root dominance factor h. This factor makes the balance index robust when internal nodes can have non-zero size, which blurs the distinction between internal nodes and leaves. Non-root dominance plays no role if all internal nodes have null size, as in cladograms (because then $h \equiv 1$).

Third, instead of assigning a size to each node as a function of its out-degree, we associate a node's size with the size of the biological population it represents. This ensures that our indices are reliably robust when applied to real data.

Relationship with Sackin's index

The sum $\sum_{k \in \widetilde{V}} n_k$ is just another way of expressing Sackin's index (summing over internal nodes instead of leaves). Therefore J in Equation 4 is essentially a weighted Sackin index (with each term in the sum weighted by the balance score W) divided by the unweighted Sackin index. In the special,

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

important case of full m-ary cladograms, the weighted sum in J^1 (Equation 5) simplifies yet further. Let $\mathcal{T}_{n,m}$ denote the set of all trees on n leaves such that all internal nodes have the same out-degree $m > 1$, every internal node has null size, and all leaf sizes are equal. Then we obtain a remarkably simple relationship between J^1 and Sackin's index:

Proposition 0.7. *Let T be a tree on n leaves with $d^+(i) = m > 1$ and $f(i) = 0$ for every internal node i. Then*

$$J^1(T) = \frac{{}^1H_m(T)S(T)}{\sum_{k\in\widetilde{V}} S_k},$$

where ${}^1H_m(T)$ is the Shannon entropy (base m) of the proportional node sizes, and $S(T)$ is the size of T. If additionally all leaves of T have the same size (so $T \in \mathcal{T}_{n,m}$) then

$$J^1(T) = \frac{\min_{n,m} I_S}{I_S(T)} = \frac{n\log_m n}{I_S(T)}, \tag{6}$$

where $\min_{n,m} I_S$ is the minimum I_S value of trees in $\mathcal{T}_{n,m}$.

The above result is somewhat surprising as it unifies our Colless-like index, which can be viewed as a weighted average of internal node balance scores, and Sackin's index, which is the sum of all leaf depths. A short proof of Proposition 0.7 is presented in the Appendix. The converse result, which is also proved in the Appendix, justifies our choice of W^1 instead of alternative balance score functions:

Proposition 0.8. *Let J be a tree balance index such that*

$$J(T) = \frac{1}{\sum_{k\in\widetilde{V}} n_k} \sum_{i\in\widetilde{V}} n_i W\left(\frac{n_{i_1}}{n_i}, \ldots, \frac{n_{i_{p(i)}}}{n_i}\right),$$

where $i_1, \ldots, i_{p(i)}$ are the children of node i, and W is a balance score satisfying the conditions stated before Equation 1. Suppose that for all trees $T \in \mathcal{T}_{n,m}$, $J(T) = n\log_m n/I_S(T)$. Then $W = W^1$.

The right-hand side of Equation 6 incidentally provides an alternative way of normalising Sackin's index on full m-ary cladograms, including the bifurcating cladograms on which the index was originally defined. This normalised inverse Sackin index, which we can define as $J_S := n\log_m n/I_S$, provides a more satisfactory way of comparing trees that differ in their node degrees or leaf counts. $J_S = 1$ if and only if the tree has minimal depth given m, which is equivalent to being fully symmetric (because in this case $\log_m n = \nu_l$ for every leaf l). Hence J_S is a *sound* tree balance index in the sense defined by Mir et al. (2018). For $m > 1$, we have $J_S > 0$ but $\min J_S \to 0$ as $n \to \infty$, which makes sense because trees with more leaves can be made less balanced. In particular, when T is a caterpillar tree on $n \geq 2$ leaves,

$$J_S(T) = \frac{\min_{n,2} I_S}{\max_{n,2} I_S} = \frac{2n\log_2 n}{(n-1)(n+2)},$$

as illustrated in Figure 5a. The definition of J_S can be naturally extended to the case $m \leq 1$ by setting $J_S(T) := 0$ if T is linear or has only one node. From this point of view, J^1 (a Colless-like index) is a generalisation of J_S (the normalised reciprocal of Sackin's index) to the domain of trees with arbitrary degree distributions and arbitrary node sizes.

Distributions under the Yule and uniform models

An immediate corollary of Proposition 0.7 is that J^1 can be used to test whether a set of full m-ary cladograms is consistent with a particular tree-generating model, with exactly the same sensitivity as Sackin's index. For example, Figures 5a and 5b show J^1 distributions for random trees generated from the Yule and uniform models, which generate bifurcating cladograms. These two distributions have insignificant overlap when the trees have at least a few dozen leaves.

Kirkpatrick and Slatkin (1993) showed that the expectation of I_S for the Yule model is

$$\mathbb{E}_{Yule}(I_S) = 2n\sum_{i=2}^{n}\frac{1}{i} = 2n\ln n + (2\gamma - 2)n + o(n),$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

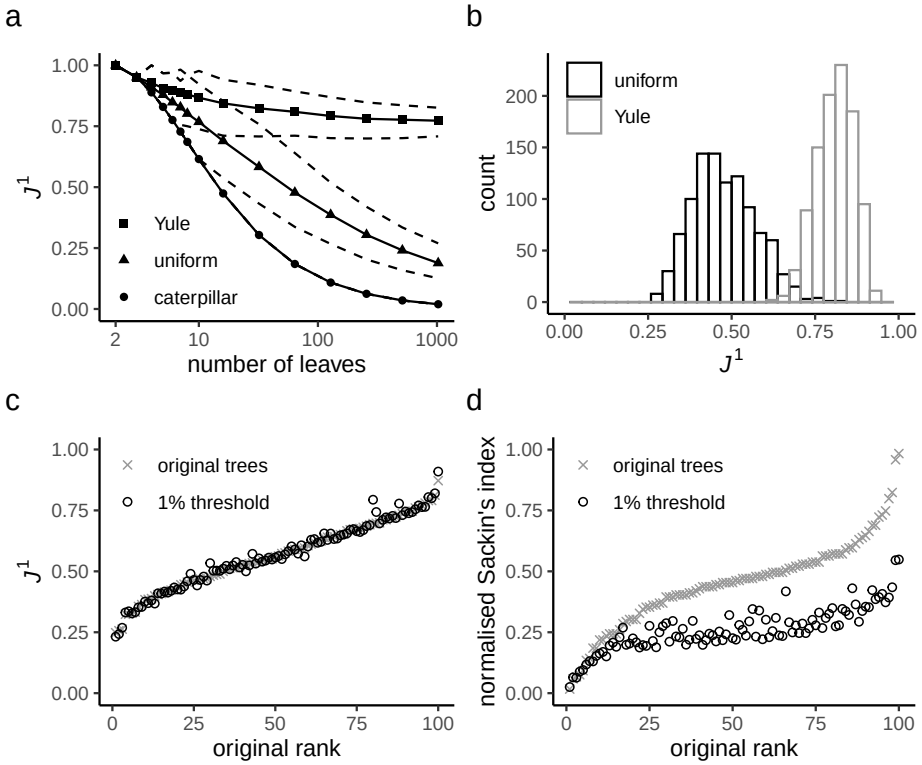

Figure 5: **a**: J_1 values for caterpillar trees and random trees generated from the Yule and uniform models (1,000 trees per data point). All internal nodes have null size and all leaves have equal size. Solid curves are the means and dashed curves are the 5th and 95th percentiles. **b**: J^1 distributions for random trees on 64 leaves generated from the Yule and uniform models (1,000 trees per model). **c**: J^1 values for 100 random trees on 16 leaves, before and after applying a 1% sensitivity threshold. These random trees were generated from the alpha-gamma model with $\alpha \sim \text{Unif}(0, 1)$ and $\gamma \sim \text{Unif}(0, \alpha)$. **d**: $I_{S,norm}$ values for the same set of random trees.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

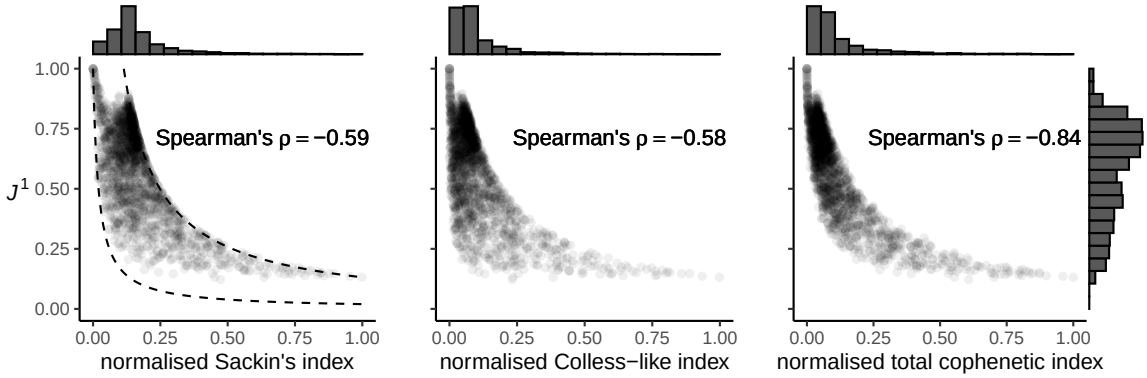

Figure 6: Scatter plots of J^1 versus normalised Sackin's, Colless-like, and total cophenetic indices for 2,000 random cladograms on 100 leaves. Histograms in the margins show the marginal distributions. Dashed reference curves in the first panel are obtained by substituting $I_{S,norm}$ into Equation 6 with $n = 100$ and $m = 2$ (upper curve) or $m = 100$ (lower curve). We use the Colless-like index with $f(n) = \ln(n + e)$ and D the mean deviation from the median, as recommended by Mir et al. (2018). Normalisation of each index other than J^1 depends only on the number of leaves and so does not affect correlations. Trees were generated from the alpha-gamma model with $\alpha \sim \text{Unif}(0, 1)$ and $\gamma \sim \text{Unif}(0, \alpha)$.

where γ is Euler's constant and n is the number of leaves. We find that a good approximation to the J^1 mean for the Yule model is $n \log_2 n / \mathbb{E}_{Yule}(I_S)$, which approaches $1/(2 \ln 2) \approx 0.72$ as $n \to \infty$.

The expectation of I_S for the uniform model approaches $\sqrt{\pi} n^{3/2}$ as the number of leaves $n \to \infty$ (Blum et al., 2006). Via trial and improvement, we find that a very good approximation to the J^1 mean for the uniform model is $n \log_2 n / (1.7 n^{3/2} - n - 0.808)$, which approaches zero as $n \to \infty$.

Robustness when applied to random trees

To test the robustness of J^1, we generated random multifurcating trees with node sizes drawn from a continuous uniform distribution, and then compared J^1 values for these trees before and after applying a 1% sensitivity threshold. In the latter case, whenever the combined frequency of a clone and its descendants was below 1%, we merged the corresponding subtree with the clone's parent, to simulate imperfect detection of rare types. As expected, the J^1 values for the two sets of trees were highly similar, with a median difference of only 0.8% (Figure 5c). In contrast, the median difference in Sackin's index for the same two sets of trees (after resolving any linear parts in the manner of Figure 2) was 20% (Figure 5d), confirming that J^1 is much more robust than Sackin's index to the omission of rare types.

Correlations with preexisting indices

To compare J^1 to Sackin's index, a Colless-like index, and the total cophenetic index (defined in the Appendix) on a diverse set of trees, we generated 2,000 random multifurcating cladograms on 100 leaves using the alpha-gamma model (Chen et al., 2009) via the R package *CollessLike* (Mir et al., 2018). As shown in Figure 6, our new balance index correlates negatively with the previously defined imbalance indices on this set of random trees, indicating that it captures a similar notion of balance. The strongest correlation is between J^1 and the total cophenetic index (Spearman's $\rho = -0.84$ for all trees, and $\rho = -0.97$ for trees with mean out-degree greater than 3).

Discontinuities

Although our indices are robust to the addition of uninformative nodes, the addition of informative nodes – however small – can create a discontinuity. Consider a node that splits its descendants into several subtrees of similar size. The addition of a new, relatively small subtree to this node will

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

create imbalance even as – in fact especially as – the size of this new subtree approaches zero. Our J^q indices are sensitive to this case.

As a more precise example, consider a star tree with $l > 1$ leaves each of size $f_0 > 0$. Suppose we add to the root another $n - l$ leaves each of size x with $0 \leq x \leq f_0$. The root is assumed to have population 0, so that $S_1 = S_1^* = lf_0 + (n-l)x$ and $J^1 = W_1^1$. If $x = 0$ then $W_1^1(x) = 1$ since all l leaves have the same size. If $x > 0$ then

$$W_1^1(x) = -\left[l \frac{f_0}{lf_0 + (n-l)x} \log_n \left(\frac{f_0}{lf_0 + (n-l)x} \right) + (n-l) \frac{x}{lf_0 + (n-l)x} \log_n \left(\frac{x}{lf_0 + (n-l)x} \right) \right].$$

As $x \to 0$, so $W_1^1(x) \to \log_n l$. This implies that adding infinitesimally small leaves reduces the balance score from 1 to $\log_n l$, to account for the abrupt loss of balance. The size of the jump is at most $1 - \log_3 2 \approx 0.37$, and it approaches zero as $l/n \to 1$.

Implementation and algorithmic complexity

Assuming the identity of the root is known, our new indices can be computed from an adjacency matrix in $\mathcal{O}(N)$ time, where N is the number of nodes. Subtree sizes are computed via depth-first search, which takes linear time, and the computation of the balance index takes at most $\sum_{i=1}^{N} |\mathrm{Adj}(i)| = N - 1$ steps, where $\mathrm{Adj}(i)$ is the adjacency list of node i. Efficient R code for calculating J^q is shared in an online repository (Noble and Lemant, 2021).

Discussion

Here we have defined a new class of tree balance index that unifies, generalises, and in various ways improves upon existing definitions. These indices are applicable to a wider set of trees and enable important new applications.

A challenge in comparing simulated phylogenies and trees inferred from data is that the former are exact, whereas the latter are often incomplete (Scott et al., 2020). In oncology, for example, it has been shown that whether or not a rare tumour clone is detected depends on both methodology and chance (Turajlic et al., 2018). Our balance indices largely solve this problem as they are robust to the omission of rare types, as demonstrated briefly here and more comprehensively in a companion paper (Noble et al., accepted for publication). Besides tumour evolution, our indices are especially well suited to the study of microbial evolution and any other system in which population sizes matter or linear evolution can occur.

In generalising conventional indices we also obviate their shortcomings. Even when restricted to the tree types on which previous indices are defined, our indices enable more meaningful comparison of trees with different degree distributions or different numbers of leaves. This advantage might make our indices preferable to other options more generally.

Because of its unique relationship with Sackin's index, we especially recommend J^1 – a weighted average of the normalised entropies of the internal nodes – as defined in general by Equation 3 and more simply for cladograms by Equation 5. Given that Sackin's index has been well studied, it is convenient that J^1 inherits some of the properties of that index when applied to full m-ary cladograms, including its relatively high sensitivity in distinguishing between alternative tree-generating models (Kirkpatrick and Slatkin, 1993; Agapow and Purvis, 2002). Within our framework, Sackin's index is seen not as a general balance index but rather as a normalising factor, which works as a balance index only in the special case of full m-ary cladograms (for which the numerator of J^1 is independent of tree topology).

Proposition 0.7 implies that determining the precise moments of J^1 for a model that generates full m-ary cladograms is equivalent to determining the moments of the reciprocal of Sackin's index. Figure 6 suggests that J^1 has interesting relationships with other indices such as the total cophenetic index. These are promising areas for further investigation.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Funding

This work was supported by the National Cancer Institute at the National Institutes of Health (grant number U54CA217376) to RN and VM. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgements

We thank Niko Beerenwinkel, Laura Keller, Francesco Marass, Lisa Lamberti, Jack Kuipers and Katharina Jahn for helpful conversations.

Author contributions

RN conceived the project. JL and RN developed the balance indices with helpful input from CLS. JL and RN obtained mathematical results with helpful input from VM. JL and RN wrote the paper. All authors have read and approved this manuscript.

References

Paul Michael Agapow and Andy Purvis. Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis. *Systematic Biology*, 51 (6):866–872, 2002.

Michael G. B. Blum, Olivier François, and Svante Janson. The mean, variance and limiting distribution of two statistics sensitive to phylogenetic tree balance. *The Annals of Applied Probability*, 16(4):2195–2214, 2006.

Anne Chao, Chun-Huo Chiu, and Lou Jost. Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annual Review of Ecology, Evolution, and Systematics*, 45(1):297–324, 2014.

Bo Chen, Daniel Ford, and Matthias Winkel. A new family of Markov branching trees: The alpha-gamma model. *Electronic Journal of Probability*, 14:400–430, 2009.

Ketevan Chkhaidze, Timon Heide, Benjamin Werner, Marc J. Williams, Weini Huang, Giulio Caravagna, Trevor A. Graham, and Andrea Sottoriva. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. *PLOS Computational Biology*, 15(7):e1007243, 2019.

Donald H. Colless. Review of phylogenetics: the theory and practice of phylogenetic systematics. *Systematic Zoology*, 31(1):100–104, 1982.

Alexander Davis, Ruli Gao, and Nicholas Navin. Tumor evolution: Linear, branching, neutral or punctuated? *Biochimica et Biophysica Acta - Reviews on Cancer*, 1867(2):151–161, 2017.

Mariam Jamal-Hanjani, Gareth A. Wilson, Nicholas McGranahan, Nicolai J. Birkbak, Thomas B.K. Watkins, Selvaraju Veeriah, Seema Shafi, Diana H. Johnson, Richard Mitter, Rachel Rosenthal, et al. Tracking the evolution of non–small-cell lung cancer. *New England Journal of Medicine*, 376(22):2109–2121, 2017.

Mark Kirkpatrick and Montgomery Slatkin. Searching for evolutionary patterns in the shape of a phylogenetic tree. *Evolution*, 47(4):1171–1181, 1993.

Carlo C. Maley, Athena Aktipis, Trevor A. Graham, Andrea Sottoriva, Amy M. Boddy, Michalina Janiszewska, Ariosto S. Silva, Marco Gerlinger, Yinyin Yuan, Kenneth J. Pienta, Karen S. Anderson, Robert Gatenby, Charles Swanton, David Posada, Chung-I Wu, Joshua D. Schiffman, E. Shelley Hwang, Kornelia Polyak, Alexander R. A. Anderson, Joel S. Brown, Mel Greaves, and Darryl Shibata. Classifying the evolutionary and ecological features of neoplasms. *Nature Reviews Cancer*, 17(10):605–619, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Arnau Mir, Francesc Rosselló, et al. A new balance index for phylogenetic trees. *Mathematical biosciences*, 241(1):125–136, 2013.

Arnau Mir, Lucía Rotger, and Francesc Rosselló. Sound Colless-like balance indices for multifurcating trees. *PloS one*, 13(9), 2018.

Arne O. Mooers and Stephen B. Heard. Inferring Evolutionary Process from Phylogenetic Tree Shape. *The Quarterly Review of Biology*, 72(1):31–54, 1997.

Robert Noble and Jeanne Lemant. *RUtreebalance: Robust, universal tree balance indices*, 2021. URL `https://github.com/robjohnnoble/RUtreebalance`.

Robert Noble, Dominik Burri, Cécile Le Sueur, Jeanne Lemant, Yannick Viossat, Jakob Nikolas Kather, and Niko Beerenwinkel. Spatial structure governs the mode of tumour evolution. *Nature Ecology & Evolution*, accepted for publication.

János Podani. Tree thinking, time and topology: comments on the interpretation of tree diagrams in evolutionary/phylogenetic systematics. *Cladistics*, 29(3):315–327, 2013.

M.J. Sackin. "Good" and "bad" phenograms. *Systematic Biology*, 21(2):225–226, 1972.

Jacob G Scott, Philip K Maini, Alexander RA A Anderson, and Alexander G Fletcher. Inferring Tumor Proliferative Organization from Phylogenetic Tree Measures in a Computational Model. *Systematic Biology*, 69(4):623–637, 2020.

Kwang-Tsao Shao and Robert R Sokal. Tree Balance. *Systematic Zoology*, 39(3):266, 1990.

Samra Turajlic, Hang Xu, Kevin Litchfield, Andrew Rowan, Stuart Horswell, Tim Chambers, Tim O'Brien, Jose I. Lopez, Thomas B.K. Watkins, David Nicol, Mark Stares, Ben Challacombe, Steve Hazell, Ashish Chandra, Thomas J. Mitchell, Lewis Au, Claudia Eichler-Jonsson, Faiz Jabbar, Aspasia Soultati, Simon Chowdhury, Sarah Rudman, Joanna Lynch, Archana Fernando, Gordon Stamp, Emma Nye, Aengus Stewart, Wei Xing, Jonathan C. Smith, Mickael Escudero, Adam Huffman, Nik Matthews, Greg Elgar, Ben Phillimore, Marta Costa, Sharmin Begum, Sophia Ward, Max Salm, Stefan Boeing, Rosalie Fisher, Lavinia Spain, Carolina Navas, Eva Grönroos, Sebastijan Hobor, Sarkhara Sharma, Ismaeel Aurangzeb, Sharanpreet Lall, Alexander Polson, Mary Varia, Catherine Horsfield, Nicos Fotiadis, Lisa Pickering, Roland F. Schwarz, Bruno Silva, Javier Herrero, Nick M. Luscombe, Mariam Jamal-Hanjani, Rachel Rosenthal, Nicolai J. Birkbak, Gareth A. Wilson, Orsolya Pipek, Dezso Ribli, Marcin Krzystanek, Istvan Csabai, Zoltan Szallasi, Martin Gore, Nicholas McGranahan, Peter Van Loo, Peter Campbell, James Larkin, and Charles Swanton. Deterministic Evolutionary Trajectories Influence Primary Tumor Growth: TRACERx Renal. *Cell*, 173(3):595–610.e11, 2018.

Appendix

Definition of the total cophenetic index

The cophenetic value $\phi(k, l)$ of a pair of leaves (k, l) is the depth of their lowest common ancestor. The total cophenetic index (Mir et al., 2013) is then the sum of the cophenetic values over all pairs of leaves:

$$I_\Phi(T) = \sum_{N-n+1 \leq k < l \leq n} \phi(k, l),$$

where N is the number of nodes and n the number of leaves. As in Sackin's index, the principle is that an unbalanced tree stretches more than a balanced tree. Being explicitly defined for all multifurcating trees, the total cophenetic index permits meaningful comparison of any two multifurcating trees on the same number of leaves.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For trees on $n > 2$ leaves, the minimum of the total cophenetic index is reached on the star tree, with $\min_n(I_\Phi) = 0$. The maximum is attained on the caterpillar tree:

$$\max_n(I_\Phi) = \sum_{k=2}^{n-1}\sum_{l=1}^{k-1} m = \sum_{k=2}^{n-1} \frac{1}{2}k(k-1) = \frac{1}{2}\left(\frac{(n-1)n(2n-1)}{6} - \frac{n(n-1)}{2}\right)$$
$$= \frac{n(n-1)(n-2)}{6} = \binom{n}{3}.$$

Hence a normalised version of the total cophenetic index is $I_{\Phi,norm}(T) = I_\Phi(T)/\binom{n}{3}$. This normalised imbalance index is not minimal for all fully symmetric trees. For example, the cophenetic value of the two leftmost leaves of fully symmetric tree **b** in Figure 1 is two, and so both the unnormalised and normalised cophenetic indices of tree **b** will be nonzero.

Proof that the index of Equation 1 satisfies our five axioms

Proof. Axiom 0.1 (Maximum value): We have $J \leq 1$ since h and W lie between zero and one by definition. Also if any internal node j of tree T doesn't split its descendants into at least two equally sized subtrees then $W_j < 1$ by definition and so

$$\sum_{i\in\widetilde{V}} g(S_i^*)h(S_i^*, S_i)W_i < \sum_{k\in\widetilde{V}} g(S_k^*) \implies J(T) < 1.$$

Finally, let T be a tree such that every internal node splits its descendants into at least two equally sized subtrees. Then $W_i = 1$ for all $i \in \widetilde{V}$ by definition. And if every internal node has null population then $S_i^* = S_i$, which implies $h(S_i^*, S_i) = 1$ for all $i \in \widetilde{V}$ by definition. Hence

$$J(T) = \frac{1}{\sum_{k\in\widetilde{V}} g(S_k^*)} \sum_{i\in\widetilde{V}} g(S_i^*) = 1.$$

Axiom 0.2 (Minimum value): We have $J \geq 0$ since g, h and W are always non-negative by definition. Also if T is a linear tree then $W_i = 0$ for all $i \in \widetilde{V}$ by definition, and hence $J(T) = 0$. Conversely, if some internal node j has $d^+(i) > 1$ then $W_j > 0$ by definition and, because g and h are always positive by definition, we must have $J(T) > 0$.

Axiom 0.3 (Leaf limit): Adding a subtree to a leaf l changes the tree balance value via the contributions of three sets of nodes: the newly added nodes, the former leaf l, and all other internal nodes. First, for each internal node $i \in \widetilde{V}(T_l)$ with $i \neq l$, as $S_l^* \to 0$ so also $S_i^* \to 0$ (because $S_i^* \leq S_l^*$), which implies $g(S_i^*) \to 0$ by definition, and hence the first contribution approaches zero. Next, as $S_l^* \to 0$, so $h(S_l^*, S_l) \to 0$ by definition, which implies that the second contribution approaches zero. Lastly, the third contribution approaches zero because g, h and W are continuous by definition.

Axiom 0.4 (Linear limit): Adding a subtree to a node j, with previously $d^+(j) = 1$, changes the tree balance value via the contributions of the newly added nodes and of node j. The first contribution approaches zero for the same reason as in the leaf limit proof. Now without loss of generality let j_1 denote the original child of j, and $j_2, \ldots, j_k$ denote the newly added children of j. As $S_{j_2} + \ldots + S_{j_k} \to 0$ there are two possibilities. If we also have $S_{j_1} \to 0$ then $S_j^* = S_{j_1} + S_{j_2} + \ldots + S_{j_k} \to 0$, which implies $h(S_j^*, S_j) \to 0$ by definition. Otherwise, $\max(S_{j_2}, \ldots, S_{j_k})/S_{j_1} \to 0$, which implies $W(S_{j_1}, S_{j_2}, \ldots, S_{j_k}) \to 0$ by definition. In either case the second contribution approaches zero.

Axiom 0.5 (Continuity): The continuity of J follows immediately from the continuity of g, h and W. □

Other balance indices based on generalised entropies

As defined by Chao et al. (2014), generalised entropies for $q \geq 0, q \neq 1$ are

$${}^qH := \frac{1}{q-1}\left(1 - \sum_{i=1}^{P} p_i^q\right).$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Parameter q determines the sensitivity to the type frequencies. ${}^0H$ is simply the richness (minus 1) of the population, which corresponds to ignoring the frequencies and just counting the types. For $0 < q < 1$, rare types are given more weight than implied by their proportion, whereas for $q > 1$ abundant types matter more. ${}^2H$ is the Gini-Simpson coefficient. In the limit $q \to 1$ we recover the Shannon entropy ${}^1H_e$.

For $q > 0$, ${}^qH$ attains its maximum value if and only if all types have equal frequency $p_i = 1/m$:

$$\max({}^qH) = \frac{1}{q-1}\left(1 - \frac{1}{m^{q-1}}\right) = \frac{m^{q-1}-1}{m^{q-1}(q-1)}.$$

We can therefore define a normalised balance score W_i^q for $q > 0, q \neq 1$ and $i \in \widetilde{V}$:

$$W_i^q := \begin{cases} \dfrac{d^+(i)^{q-1}}{d^+(i)^{q-1}-1}\left(1 - \sum\limits_{j\in C(i)} p_{ij}^q\right) & \text{if } d^+(i) \geq 2 \\ 0 \quad \text{otherwise.} \end{cases}$$

A balance index J^q satisfying our axioms is then

$$J^q := \frac{1}{\sum_{k\in\widetilde{V}} S_k^*} \sum_{i\in\widetilde{V}} S_i^* \frac{S_i^*}{S_i} W_i^q,$$

for any $q > 0$. In the limit $q \to 1$, $J^q \to J^1$.

Proof of Proposition 0.7

Proof. By definition of J^1, if T is a tree on n leaves with $d^+(i) = m > 1$ and $f(i) = 0$ for every internal node i then

$$J^1(T) = \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in\widetilde{V}} \sum_{j\in C(i)} S_j \log_m \frac{S_j}{S_i}.$$

The number of distinct subtrees that contain a given leaf l is equal to its number of ancestors, which is the same as ν_l, the depth of l. So the sum of subtree sizes over the set of all internal nodes is equal to the sum of ν_l multiplied by leaf size over the set of all leaves:

$$\sum_{k\in\widetilde{V}} S_k = \sum_{k\in L} \nu_k f(k).$$

Summing first over the internal nodes and then over their children gives the same result:

$$\sum_{i\in\widetilde{V}} \sum_{j\in C(i)} S_j = \sum_{i\in\widetilde{V}} S_i = \sum_{i\in L} \nu_i f(i) = \sum_{i\in L} f(i) \sum_{j=1}^{\nu_i} 1.$$

Let $a(i,j)$ denote the ancestor of node i at distance j, with $a(i,0) = i$ and $a(i,\nu_i) = r$ (the root) for all i. Then by extension,

$$\sum_{i\in\widetilde{V}} \sum_{j\in C(i)} S_j \theta(S_i, S_j) = \sum_{i\in L} f(i) \sum_{j=1}^{\nu_i} \theta(S_{a(i,j)}, S_{a(i,j-1)}),$$

for any function θ. In particular, we have

$$\sum_{i\in\widetilde{V}} \sum_{j\in C(i)} S_j \log_m \frac{S_j}{S_i} = \sum_{i\in L} f(i) \sum_{j=1}^{\nu_i} \log_m \frac{S_{a(i,j-1)}}{S_{a(i,j)}}.$$

Substituting this result into the expression for J^1 we find

$$\begin{aligned} J^1(T) &= \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} \sum_{j=1}^{\nu_i} f(i) \log_m \frac{S_{a(i,j-1)}}{S_{a(i,j)}} \\ &= \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} f(i) \sum_{j=1}^{\nu_i} (\log_m S_{a(i,j-1)} - \log_m S_{a(i,j)}). \end{aligned}$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.25.457695; this version posted September 20, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The right-hand sum is a telescoping series that collapses to give

$$J^1(T) = \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} f(i)(\log_m S_{a(i,0)} - \log_m S_{a(i,\nu_i)}).$$

Now since i is a leaf, $\log_m S_{a(i,0)} = \log_m S_i = \log_m f(i)$. Also $\log_m S_{a(i,\nu_i)} = \log_m S_r = \log_m S(T)$. Hence

$$\begin{aligned} J^1(T) &= \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} f(i)(\log_m f(i) - \log_m S(T)) \\ &= \frac{-1}{\sum_{k\in\widetilde{V}} S_k} \sum_{i\in L} f(i) \log_m \frac{f(i)}{S(T)} = \frac{{}^1H_m(T)S(T)}{\sum_{k\in\widetilde{V}} S_k}. \end{aligned}$$

If additionally all leaves i of T have the same size $f(i) = f_0$ then $S(T) = nf_0$, ${}^1H_m(T) = \log_m n$, and $\sum_{k\in\widetilde{V}} S_k = f_0 I_S(T)$, which implies $J^1(T) = n \log_m n / I_S(T)$. □

Proof of Proposition 0.8

Proof. Since $\sum_{k\in\widetilde{V}} n_k = I_S(T)$, the conditions are equivalent to

$$I_S(T)J(T) = \sum_{i\in\widetilde{V}} n_i W_i = n \log_m n, \qquad \text{with } W_i = W\left(\frac{n_{i_1}}{n_i}, \ldots, \frac{n_{i_{p(i)}}}{n_i}\right),$$

where $n_{i_1}, \ldots n_{p(i)}$ are the children of i. Let T be a tree in $\mathcal{T}_{n,m}$ and i be an internal node of T. Then $T_i \in \mathcal{T}_{n_i,m}$ and $T_j \in \mathcal{T}_{n_j,m}$ for every child j of i. Therefore

$$I_S(T_i)J(T_i) = n_i W_i + \sum_{j\in C(i)} J(T_j) = n_i W_i + \sum_{j\in C(i)} n_j \log_m n_j.$$

Also, $I_S(T_i)J(T_i) = n_i \log_m n_i$, so we have

$$\begin{aligned} & n_i W_i + \sum_{j\in C(i)} n_j \log_m n_j = n_i \log_m n_i \\ \implies & W_i = \log_m n_i - \sum_{j\in C(i)} \frac{n_j}{n_i} \log_m n_j. \end{aligned}$$

Since $\sum_{j\in C(i)} n_j = n_i$, this implies

$$W_i = \sum_{k\in C(i)} \frac{n_k}{n_i} \log_m n_i - \sum_{j\in C(i)} \frac{n_j}{n_i} \log_m n_j = -\sum_{j\in C(i)} \frac{n_j}{n_i} \log_m \frac{n_j}{n_i} = W_i^1.$$

□