



Quantifying the difference between phylogenetic diversity and diversity indices

Magnus Bordewich¹ · Charles Semple²

Received: 20 April 2023 / Revised: 15 October 2023 / Accepted: 5 February 2024

© The Author(s) 2024

Abstract

Phylogenetic diversity is a popular measure for quantifying the biodiversity of a collection Y of species, while phylogenetic diversity indices provide a way to apportion phylogenetic diversity to individual species. Typically, for some specific diversity index, the phylogenetic diversity of Y is not equal to the sum of the diversity indices of the species in Y . In this paper, we investigate the extent of this difference for two commonly-used indices: Fair Proportion and Equal Splits. In particular, we determine the maximum value of this difference under various instances including when the associated rooted phylogenetic tree is allowed to vary across all rooted phylogenetic trees with the same leaf set and whose edge lengths are constrained by either their total sum or their maximum value.

Keywords Phylogenetic tree · Phylogenetic diversity · Diversity indices · Fair Proportion index · Equal Splits index

Mathematics Subject Classification 05C85 · 68R10

1 Introduction

Phylogenetic diversity (PD) is prominent measure in evolutionary biology to quantify the biological diversity of a collection of species. Intuitively, the phylogenetic

Charles Semple, was supported by the New Zealand Marsden Fund.

✉ Charles Semple
charles.semple@canterbury.ac.nz

Magnus Bordewich
m.j.r.bordewich@durham.ac.uk

¹ Department of Computer Science, Durham University, Durham, UK

² School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

diversity of such a collection quantifies how much of the ‘Tree of Life’ is spanned by the species in the collection. Introduced by Faith in 1992 (Faith 1992), PD has been analysed and applied in a variety of contexts for various taxa including plants, bacteria, and mammals (Cadotte et al. 2008; Lozupone and Knight 2007; Safi et al. 2011). On the other hand, (phylogenetic) diversity indices, also called evolutionary distinctiveness measures, quantify an individual species’ contribution to overall phylogenetic diversity, thus providing a convenient way to rank species as, for example, in conservation planning (Redding et al. 2008, 2014). Intuitively, these indices assign species with fewer and more distant relatives higher values than those species with many more and closer relatives.

Diversity indices have been proposed and used as an alternative to PD in conservation prioritisation (Isaac et al. 2007; Redding et al. 2010; Stein et al. 2018). However, although the sum of the diversity indices across *all* species equates to the total phylogenetic diversity, in general, the sum of the diversity indices of a *subset* of the species differs from the subset’s phylogenetic diversity. If a biologist or conservationist is switching from the ideal measure of full-tree PD to an approximation by a simple diversity index, it is important for them to understand the scale of the possible difference that this switch might involve.

Using empirical experiments, the authors in Redding et al. (2008) studied how well different diversity indices ‘captured’ PD with respect to tree shape. Their experiments indicate that diversity indices captured more PD on “more unbalanced trees and on trees with many splits near the present” and concluded that “new measures... may provide a transparent alternative to more complicated full-tree approaches.”

In this paper, we rigorously investigate the extent of this difference for two natural and well-known diversity indices, Fair Proportion and Equal Splits. Both of these indices were included in the empirical study in Redding et al. (2008) and are used, for example, in the conservation initiative ‘EDGE of Existence Programme’ established by the Zoological Society of London (Isaac et al. 2007). As well as determining the maximum difference under various scenarios (see below for details), which will inform the practitioner as to the possible extent to which an index can fail to approximate PD, we also characterised the shape and structure of the phylogenetic trees that realised these differences. Such information is potentially useful to the practitioner in that the more their phylogenetic trees of interest resemble an extremal instance, the more attention they may need to pay to the real possibility of the diversity index failing to accurately capture PD.

A version of the investigation in this paper was posed in earlier work (Haake et al. 2008) but, apart from the related simulation study (Redding et al. 2008), it has not been explored further. Furthermore, the work in this paper is related to a recent study in Wicke and Steel (2020) in which the authors consider the extent of the difference between the Fair Proportion and Equal Splits indices of a species. We next formalise the investigation in this paper, ending the introduction with a high-level overview of the main results.

Throughout the paper, X is a non-empty finite set. A *rooted phylogenetic X -tree* \mathcal{T} is a rooted tree with leaf set X whose non-leaf vertices have out-degree at least two. For technical reasons, if $|X| = 1$, we additionally allow a rooted phylogenetic tree to consist of the single vertex in X . If all non-leaf vertices have out-degree exactly

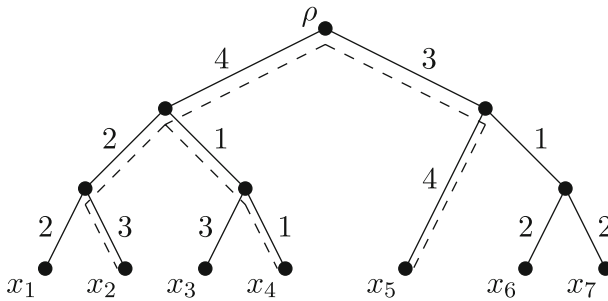


Fig. 1 An edge-weighted phylogenetic X -tree (\mathcal{T}, ℓ) . The minimal subtree of \mathcal{T} connecting the vertices in $\{x_2, x_4, x_5\} \cup \{\rho\}$ is indicated with dashed edges

two, then \mathcal{T} is *binary*. For ease of reading, as all phylogenetic trees in this paper are rooted and binary, we will refer to a “rooted binary phylogenetic tree” as simply a “phylogenetic tree”.

Let \mathcal{T} be a phylogenetic X -tree with root ρ , and consider a map $\ell : E \rightarrow \mathbb{R}^{\geq 0}$ from the set E of edges of \mathcal{T} to the non-negative reals. Collectively, we denote \mathcal{T} and ℓ by the ordered pair (\mathcal{T}, ℓ) and refer to (\mathcal{T}, ℓ) as an *edge-weighted phylogenetic X -tree*. Furthermore, we use $L(\mathcal{T}, \ell)$ to denote the total sum of the edge lengths of (\mathcal{T}, ℓ) . The *phylogenetic diversity* of a subset Y of X on (\mathcal{T}, ℓ) , denoted $PD_{(\mathcal{T}, \ell)}(Y)$, is the sum of the weights of the edges in the (unique) minimal subtree of \mathcal{T} connecting the vertices in $Y \cup \{\rho\}$. Observe that $PD_{(\mathcal{T}, \ell)}(X) = L(\mathcal{T}, \ell)$. To illustrate, consider the edge-weighted phylogenetic X -tree shown in Fig. 1, where $X = \{x_1, x_2, \dots, x_7\}$. The minimal subtree of \mathcal{T} connecting the vertices in $\{x_2, x_4, x_5\} \cup \{\rho\}$ is indicated by dashed edges, and so PD of $\{x_2, x_4, x_5\}$ is 18. The PD of $\{x_5, x_7\}$ is 10.

A *phylogenetic diversity index* for an edge-weighted phylogenetic X -tree (\mathcal{T}, ℓ) is a function $\varphi_{(\mathcal{T}, \ell)} : X \rightarrow \mathbb{R}^{\geq 0}$ that assigns a score to each leaf of \mathcal{T} such that

$$\sum_{x \in X} \varphi_{(\mathcal{T}, \ell)}(x) = PD_{(\mathcal{T}, \ell)}(X) = L(\mathcal{T}, \ell).$$

Furthermore, if, for each $x \in X$, we can write $\varphi_{(\mathcal{T}, \ell)}(x)$ as a linear function of the edge lengths of \mathcal{T} , that is,

$$\varphi_{(\mathcal{T}, \ell)}(x) = \sum_{e \in E} \gamma_{(\mathcal{T}, \ell)}(x, e) \cdot \ell(e) \tag{1}$$

for some constants $\gamma_{(\mathcal{T}, \ell)}(x, e)$ that are independent of $\ell(e)$, we say $\varphi_{(\mathcal{T}, \ell)}$ is a *linear diversity index*. It is easily checked that an arbitrary function $\varphi_{(\mathcal{T}, \ell)}$ of the form shown in (1) is a phylogenetic diversity index if and only if, for each edge e of \mathcal{T} , we have

$$\sum_{x \in X} \gamma_{(\mathcal{T}, \ell)}(x, e) = 1. \tag{2}$$

Two well-studied linear diversity indices underlie the results in this paper. Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree. The *Fair Proportion index* (FP) for a leaf $x \in X$, denoted $FP_{(\mathcal{T}, \ell)}(x)$, is the value

$$FP_{(\mathcal{T}, \ell)}(x) = \sum_{e \in P(\mathcal{T}; \rho, x)} \frac{1}{n(e)} \ell(e),$$

where $P(\mathcal{T}; \rho, x)$ denotes the (unique) path in \mathcal{T} from the root ρ to x , and $n(e)$ denotes the number of leaves that are at the end of a directed path starting at the root and traversing e . Intuitively, the FP index distributes the length of e evenly amongst its descendant leaves. The Fair Proportion index is also called ‘evolutionary distinctiveness’ (Isaac et al. 2007). The second index is the *Equal Splits index* (ES) which for a leaf $x \in X$ is the value

$$ES_{(\mathcal{T}, \ell)}(x) = \sum_{e \in P(\mathcal{T}; \rho, x)} \frac{1}{\Pi(e, x)} \ell(e),$$

where $\Pi(e, x) = 1$ if e is the pendant edge of \mathcal{T} incident with x , and if $e = (u, v)$ is a non-pendant edge of \mathcal{T} , then $\Pi(e, x)$ is the product of the out-degrees of the non-leaf vertices (including v) on the directed path from v to x . Since all phylogenetic trees in this paper are binary, $\Pi(e, x)$ is always a power of 2. In particular, if $e = (u, v)$ is a non-pendant edge of \mathcal{T} , then $\Pi(e, x)$ is 2^m , where m is the number of edges from v to x . The Fair Proportion and Equal Splits indices were introduced in Redding (2003) and Redding and Mooers (2006), respectively, and a direct comparison of these two indices was investigated recently in Wicke and Steel (2020).

As examples of the Fair Proportion and Equal Splits indices, consider the edge-weighted phylogenetic X -tree in Fig. 1. The FP indices for x_5 and x_7 are

$$FP_{(\mathcal{T}, \ell)}(x_5) = \frac{1}{3} \cdot 3 + 1 \cdot 4 = 5$$

and

$$FP_{(\mathcal{T}, \ell)}(x_7) = \frac{1}{3} \cdot 3 + \frac{1}{2} \cdot 1 + 1 \cdot 2 = 3\frac{1}{2}.$$

The ES indices for x_5 and x_7 are

$$ES_{(\mathcal{T}, \ell)}(x_5) = \frac{1}{2} \cdot 3 + 1 \cdot 4 = 5\frac{1}{2}$$

and

$$ES_{(\mathcal{T}, \ell)}(x_7) = \frac{1}{4} \cdot 3 + \frac{1}{2} \cdot 1 + 1 \cdot 2 = 3\frac{1}{4}.$$

For an edge-weighted phylogenetic X -tree (\mathcal{T}, ℓ) and diversity index φ , although the sum of the diversity indices across all taxa in X equates to $PD_{(\mathcal{T}, \ell)}(X)$, the sum of the diversity indices across all taxa in a proper subset Y of X will typically not equal $PD_{(\mathcal{T}, \ell)}(Y)$. For example, in Fig. 1, $PD_{(\mathcal{T}, \ell)}(\{x_5, x_7\}) = 10$, but

$$\sum_{x \in \{x_5, x_7\}} FP_{(\mathcal{T}, \ell)}(x) = 5 + 3\frac{1}{2} = 8\frac{1}{2}.$$

To quantify this, set

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = PD_{(\mathcal{T}, \ell)}(Y) - \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x). \tag{3}$$

We refer to $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$ as the *diversity difference of Y* (relative to (\mathcal{T}, ℓ) and φ). In this paper, we are interested in the extent of this difference for the Fair Proportion and Equal Splits indices.

The following lemma shows that the diversity difference is always non-negative provided φ satisfies a natural distribution property. This property says that, for an edge-weighted phylogenetic X -tree (\mathcal{T}, ℓ) and all $x \in X$, the value $\varphi_{(\mathcal{T}, \ell)}(x)$ can be written as a non-negative linear function of the lengths of edges in the path from the root ρ of \mathcal{T} to x .

Lemma 1.1 *Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree with root ρ , and let $\varphi_{(\mathcal{T}, \ell)}$ be a linear diversity index such that, for each $x \in X$,*

$$\varphi_{(\mathcal{T}, \ell)}(x) = \sum_{e \in P(\mathcal{T}; \rho, x)} \gamma_{(\mathcal{T}, \ell)}(x, e) \cdot \ell(e) \tag{4}$$

for some non-negative constants $\gamma_{(\mathcal{T}, \ell)}(x, e)$ that are independent of $\ell(e)$. If $Y \subseteq X$, then

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) \geq 0.$$

Proof Let $Y \subseteq X$. Then

$$PD_{(\mathcal{T}, \ell)}(Y) = \sum_{e \in \mathcal{T}(Y \cup \{\rho\})} \ell(e)$$

and

$$\sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x) = \sum_{e \in \mathcal{T}(Y \cup \{\rho\})} \left(\sum_{x \in Y} \gamma_{(\mathcal{T}, \ell)}(x, e) \right) \cdot \ell(e),$$

where $\mathcal{T}(Y \cup \{\rho\})$ denotes the minimal subtree of \mathcal{T} connecting the leaves in Y and ρ . Since $\gamma_{(\mathcal{T}, \ell)}(x, e)$ is non-negative for all $x \in X$ and edges e of \mathcal{T} , it follows by (2) that $\sum_{x \in Y} \gamma_{(\mathcal{T}, \ell)}(x, e) \leq 1$. Hence $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) \geq 0$ as required. \square

We call a phylogenetic diversity index that satisfies (4) a *descendant diversity index*. It is easily checked that the Fair Proportion and Equal Splits indices are examples of descendant diversity indices.

In this paper, for each of the Fair Proportion and Equal Splits indices, we will determine, for all positive integers k , the maximum value of the diversity difference over all subsets of X of size k under the following instances:

- (i) a fixed phylogenetic X -tree whose edge lengths are fixed;
- (ii) a fixed phylogenetic X -tree whose edge lengths are constrained by either their (total) sum or their maximum value; and
- (iii) across all phylogenetic X -trees whose edge lengths are constrained by either their (total) sum or their maximum value.

In particular, for (i) and (ii), we give polynomial-time algorithms for finding these maximum values for an arbitrary descendant diversity index while, for (iii), we characterise the edge-weight phylogenetic X -trees and subsets of size k that realise this maximum value for FP and ES. For (iii), it turns out that, in the case that the edge lengths are constrained by their sum, the class of phylogenetic trees that maximise the diversity difference under ES is always a subclass of the phylogenetic trees that maximise the diversity difference under FP. The corresponding characterisations are stated as Theorems 4.2 and 4.1, respectively. However, in the case that the edge lengths are constrained by their maximum value, the class of phylogenetic trees that maximise the diversity difference under FP and ES coincide (Theorem 4.7).

The paper is organised as follows. The next section contains some preliminaries that are used throughout the paper. Section 3 considers maximising the diversity difference on a fixed phylogenetic tree, while Sect. 4 considers maximising the diversity difference across all phylogenetic trees. The last section, Sect. 5, consists of a brief discussion.

2 Preliminaries

Let \mathcal{T} be a phylogenetic X -tree, and let t and w be vertices of \mathcal{T} . If P is a (directed) path in \mathcal{T} from t to w , the length of P , denoted $|P|$, is the number of edges in P . We denote this path by $P(\mathcal{T}; t, w)$. Furthermore, we sometimes refer to a path in \mathcal{T} from an edge $e = (u, v)$ to w , in which case, we mean from v to w , and denote this path by $P(\mathcal{T}; e, w)$. Note that if (u, v) is an edge of \mathcal{T} , then u is ancestral to v .

Subtrees. Let \mathcal{T} be a phylogenetic X -tree with root ρ , and let X' be a subset of $X \cup \{\rho\}$. The minimal subtree of \mathcal{T} connecting the vertices in X' is denoted by $\mathcal{T}(X')$. Furthermore, if $\rho \notin X'$, the *restriction* of \mathcal{T} to X' , denoted by $\mathcal{T}|X'$, is the phylogenetic X' -tree obtained from $\mathcal{T}(X')$ by suppressing all non-root vertices of degree two.

Let \mathcal{T} be a phylogenetic X -tree and let $e = (u, v)$ be an edge of \mathcal{T} . The subset of leaves that are descendants of v is called a *cluster* of \mathcal{T} and is denoted by $C(v)$. The set X as well as each of the singleton subsets of X are clusters of every phylogenetic X -tree, and so such clusters are called the *trivial* clusters of \mathcal{T} . Hence a cluster X' of \mathcal{T} is *non-trivial* if $2 \leq |X'| < |X|$. Thus if \mathcal{T} has a non-trivial cluster, then $|X| \geq 3$.

Furthermore, the rooted subtree obtained from \mathcal{T} by deleting e and whose root is v is a *pendant subtree* of \mathcal{T} .

Cherries and chains. A 2-element subset of X , say $\{a, b\}$, is a *cherry* of a phylogenetic X -tree \mathcal{T} if a and b have the same parent. Now let $X' = \{a_1, a_2, \dots, a_m\}$ be a subset of X such that $m \geq 2$ and, for all $i \in \{1, 2, \dots, m\}$, let p_i denote the parent of a_i . We call X' a *chain* of \mathcal{T} if there is an ordering of the elements of X' , say (a_1, a_2, \dots, a_m) , such that

$$p_m, p_{m-1}, \dots, p_1$$

is a (directed) path in \mathcal{T} , in which case, p_m is the *first parent* of the chain and p_1 is the *last parent* of the chain. Note that we still refer to X' as a chain of \mathcal{T} if $\{p_1, p_2\}$ is a cherry and $p_m, p_{m+1}, \dots, p_2 = p_1$ is a path in \mathcal{T} . The *edge set* of X' consists of the pendant edges incident with a leaf in X' as well as the edges in the path p_m, p_{m-1}, \dots, p_1 .

Subtree prune and regraft. Let \mathcal{T} be a phylogenetic X -tree. For the purposes of the upcoming operation, view the root ρ of \mathcal{T} adjoined to the original root via a pendant edge. Let (u, v) be an edge of \mathcal{T} such that $u \neq \rho$. Let \mathcal{T}' be the phylogenetic X -tree obtained from \mathcal{T} by deleting (u, v) and suppressing u , and then reattaching the pendant subtree $\mathcal{T}(C(v))$ by subdividing an edge, f say, in the component of $\mathcal{T} \setminus (u, v)$ containing ρ with a new vertex u' and adjoining $\mathcal{T}(C(v))$ with a new edge (u', v) . We say that \mathcal{T}' has been obtained from \mathcal{T} by a *rooted subtree prune and regraft* operation. More specifically, in this operation, we have *pruned* $C(v)$ and *regrafted* it to f . In the special case f is the pendant edge incident with the root of \mathcal{T} , we say that $C(v)$ is pruned and *regrafted to* ρ . (Effectively, this corresponds to the subtree being regrafted above the root.) At the completion of this operation, we no longer view the root ρ as being joined via a pendant edge, and so the root of \mathcal{T}' (labelled ρ) is the unique vertex of \mathcal{T}' of in-degree zero after deleting the temporary root and its incident edge.

Diversity indices. Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree with edge set E , and let φ be a descendant diversity index. Then, for each $x \in X$,

$$\varphi_{(\mathcal{T}, \ell)}(x) = \sum_{e \in P(\mathcal{T}; \rho, x)} \gamma_{(\mathcal{T}, \ell)}(x, e) \cdot \ell(e)$$

for some non-negative constants $\gamma_{(\mathcal{T}, \ell)}(x, e)$ that are independent of $\ell(e)$. For a subset Y of X , we denote the contribution of an edge $e = (u, v)$ of \mathcal{T} to $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$ by $\lambda_{(\mathcal{T}, \ell)}(e)$, that is,

$$\lambda_{(\mathcal{T}, \ell)}(e) = \ell(e) - \sum_{y \in Y \cap C(v)} \gamma_{(\mathcal{T}, \ell)}(y, e) \cdot \ell(e) = \ell(e) \left(1 - \sum_{y \in Y \cap C(v)} \gamma_{(\mathcal{T}, \ell)}(y, e) \right)$$

if $Y \cap C(v)$ is non-empty, and $\lambda_{(\mathcal{T}, \ell)}(e) = 0$ otherwise. Thus

$$\sum_{e \in E} \lambda_{(\mathcal{T}, \ell)}(e) = \Delta_{(\varphi, \mathcal{T}, \ell)}(Y).$$

It will be clear in the context which subset of X and which descendant diversity index $\lambda_{(\mathcal{T}, \ell)}(e)$ is referring to. Observe that $\lambda_{(\mathcal{T}, \ell)}(e) \geq 0$. Moreover, if either $C(v) \subseteq Y$ or $C(v) \subseteq X - Y$, then $\lambda_{(\mathcal{T}, \ell)}(e) = 0$.

3 Maximising the diversity difference on a fixed tree

In this section, we consider the problem of determining the maximum value of the diversity difference for a fixed phylogenetic tree for when φ is an arbitrary descendant diversity index.

3.1 Fixed edge lengths

We first consider the instance of a fixed phylogenetic tree with fixed edge lengths. Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree, and let φ be an arbitrary linear diversity index. Now let (\mathcal{T}, ℓ') be the edge-weighted phylogenetic X -tree obtained from (\mathcal{T}, ℓ) by adding, for each $x \in X$, the value

$$L(\mathcal{T}, \ell) - \varphi_{(\mathcal{T}, \ell)}(x)$$

to the length of the pendant edge incident with x . For all non-pendant edges e , we have $\ell'(e) = \ell(e)$. Note that the use of the value $L(\mathcal{T}, \ell)$ is simply to have a constant large enough so that all pendant edges have non-negative lengths.

Lemma 3.1 *Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree, let φ be a linear diversity index, let k be a non-negative integer, and let Y be a subset of X of size k . Then*

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = PD_{(\mathcal{T}, \ell')}(Y) - k \cdot L(\mathcal{T}, \ell).$$

Proof Now,

$$\begin{aligned} \Delta_{(\varphi, \mathcal{T}, \ell)}(Y) &= PD_{(\mathcal{T}, \ell)}(Y) - \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x) \\ &= \left(PD_{(\mathcal{T}, \ell')}(Y) - \sum_{x \in Y} (L(\mathcal{T}, \ell) - \varphi_{(\mathcal{T}, \ell)}(x)) \right) - \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x) \\ &= PD_{(\mathcal{T}, \ell')}(Y) - k \cdot L(\mathcal{T}, \ell) + \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x) - \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x) \\ &= PD_{(\mathcal{T}, \ell')}(Y) - k \cdot L(\mathcal{T}, \ell). \end{aligned}$$

□

Lemma 3.1 is the basis of the following algorithm which finds the maximum value of the diversity difference under a linear diversity index for a given edge-weighted phylogenetic tree. It is well known that Step 2 of the algorithm, that is, finding a subset Y of X of size k such that $PD_{(\mathcal{T}, \ell)}(Y) \geq PD_{(\mathcal{T}, \ell)}(Y')$ for all subsets Y' of X of size k , takes quadratic time in the size of X (Pardi and Goldman 2005; Steel 2005).

MAXDIVERSITYDIFF

Input: An edge-weighted phylogenetic X -tree (\mathcal{T}, ℓ) , a linear diversity index φ , and a non-negative integer k .

Output: A subset Y of X of size k that maximises $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$ across all subsets of X of size k .

1. Let $\ell' : E(\mathcal{T}) \rightarrow \mathbb{R}^{\geq 0}$ be the map defined by setting $\ell'(e) = L(\mathcal{T}, \ell) - \varphi_{(\mathcal{T}, \ell)}(x)$ for all pendant edges e , where x is the leaf incident with e , and $\ell'(e) = \ell(e)$ for all non-pendant edges. Construct (\mathcal{T}, ℓ') .
2. Find a subset Y of X of size k such that $PD_{(\mathcal{T}, \ell')}(Y) \geq PD_{(\mathcal{T}, \ell')}(Y')$ for all subsets Y' of X of size k .
3. Return Y and $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = PD_{(\mathcal{T}, \ell')}(Y) - k \cdot L(\mathcal{T}, \ell)$.

The next theorem is an immediate consequence of the discussion prior to the description of MAXDIVERSITYDIFF and the fact that the construction of (\mathcal{T}, ℓ') from (\mathcal{T}, ℓ) takes linear time (in the size of X).

Theorem 3.2 *Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree, let k be a positive integer, and let φ be a linear diversity index. Then applying MAXDIVERSITYDIFF to (\mathcal{T}, ℓ) , φ , and k correctly returns a subset Y of X of size k that maximises*

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = PD_{(\mathcal{T}, \ell)}(Y) - \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x)$$

across all subsets of X of size k in time $O(|X|^2)$.

3.2 Maximum sum of edge lengths

Now consider the problem of maximising the diversity difference on a fixed phylogenetic tree whose edge lengths are constrained by their (total) sum. More particularly, we consider the following optimisation problem:

FIXED TREE AND TOTAL WEIGHT $(k, \mathcal{T}, m, \varphi)$

Instance: A positive integer k , a phylogenetic X -tree \mathcal{T} , a positive real number m , and a descendant diversity index φ .

Question: Find a subset Y of X of size k and a map $\ell : E \rightarrow \mathbb{R}^{\geq 0}$ on the set E of edges of \mathcal{T} with $L(\mathcal{T}, \ell) = m$ that maximise

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = PD_{(\mathcal{T}, \ell)}(Y) - \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x).$$

We will make use of the next lemma in this subsection as well as in Sect. 4.

Lemma 3.3 *Let $(k, \mathcal{T}, m, \varphi)$ be an instance of FIXED TREE AND TOTAL WEIGHT, and let Y^* and ℓ^* be a solution to this problem. Let $\ell : E \rightarrow \mathbb{R}^{\geq 0}$ be a map on the set E of edges of \mathcal{T} with $L(\mathcal{T}, \ell) = m$ and let Y be a subset of X of size k such that*

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = \Delta_{(\varphi, \mathcal{T}, \ell^*)}(Y^*).$$

If $e_1 = (u_1, v_1)$ and $e_2 = (u_2, v_2)$ are distinct edges of \mathcal{T} such that $\ell(e_1) > 0$ and $\ell(e_2) > 0$, then

- (i) $\sum_{y \in Y \cap C(v_1)} \gamma_{(\mathcal{T}, \ell)}(y, e_1) = \sum_{y \in Y \cap C(v_2)} \gamma_{(\mathcal{T}, \ell)}(y, e_2)$, and
- (ii) $\Delta_{(\varphi, \mathcal{T}, \ell')} (Y) = \Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$, where $\ell'(e_1) = \ell(e_1) + \ell(e_2)$, $\ell'(e_2) = 0$, and $\ell'(f) = \ell(f)$ for all $f \in E - \{e_1, e_2\}$.

Proof Let $\ell : E \rightarrow \mathbb{R}^{\geq 0}$ be such a map. For each $i \in \{1, 2\}$, let $e_i = (u_i, v_i)$ and, for each edge e of \mathcal{T} , define $\lambda_{(\mathcal{T}, \ell)}(e)$ and $\lambda_{(\mathcal{T}, \ell')}(e)$ to be the contribution of e to $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$ and $\Delta_{(\varphi, \mathcal{T}, \ell')}(Y)$, respectively, where ℓ' is as defined in (ii). To prove the lemma, it suffices to show that

$$\lambda_{(\mathcal{T}, \ell)}(e_1) + \lambda_{(\mathcal{T}, \ell)}(e_2) = \lambda_{(\mathcal{T}, \ell')}(e_1) + \lambda_{(\mathcal{T}, \ell')}(e_2).$$

Since $\ell'(e_2) = 0$, it follows that $\lambda_{(\mathcal{T}, \ell')}(e_2) = 0$. Now, for each $i \in \{1, 2\}$,

$$\lambda_{(\mathcal{T}, \ell)}(e_i) = \ell(e_i) - \sum_{y \in Y \cap C(v_i)} \gamma_{(\mathcal{T}, \ell)}(y, e_i) \cdot \ell(e_i).$$

If

$$\sum_{y \in Y \cap C(v_1)} \gamma_{(\mathcal{T}, \ell)}(y, e_1) < \sum_{y \in Y \cap C(v_2)} \gamma_{(\mathcal{T}, \ell)}(y, e_2),$$

then, by linearity,

$$\begin{aligned} \lambda_{(\mathcal{T}, \ell)}(e_1) + \lambda_{(\mathcal{T}, \ell)}(e_2) &< \ell(e_1) + \ell(e_2) - \sum_{y \in Y \cap C(v_1)} \gamma_{(\mathcal{T}, \ell)}(y, e_1) \cdot (\ell(e_1) + \ell(e_2)) \\ &= \lambda_{(\mathcal{T}, \ell')}(e_1) + \lambda_{(\mathcal{T}, \ell')}(e_2), \end{aligned}$$

a contradiction to the maximality of ℓ . Using a symmetric argument, it follows that

$$\sum_{y \in Y \cap C(v_1)} \gamma_{(\mathcal{T}, \ell)}(y, e_1) = \sum_{y \in Y \cap C(v_2)} \gamma_{(\mathcal{T}, \ell)}(y, e_2).$$

Hence

$$\begin{aligned} \lambda_{(\mathcal{T}, \ell)}(e_1) + \lambda_{(\mathcal{T}, \ell)}(e_2) &= \ell(e_1) + \ell(e_2) - \sum_{y \in Y \cap C(v_1)} \gamma_{(\mathcal{T}, \ell)}(y, e_1) \cdot (\ell(e_1) + \ell(e_2)) \\ &= \lambda_{(\mathcal{T}, \ell')}(e_1). \end{aligned}$$

This completes the proof of the lemma. □

By repeated applications of Lemma 3.3(ii), we obtain the following corollary.

Corollary 3.4 *Let $(k, \mathcal{T}, m, \varphi)$ be an instance of FIXED TREE AND TOTAL WEIGHT, and let Y^* and ℓ^* be a solution to this problem. Then there exists a weighting $\ell : E \rightarrow \mathbb{R}^{\geq 0}$ of the edge set E of \mathcal{T} with $\ell(e) = m$ for some edge $e \in E$ and a subset Y of X of size k such that*

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = \Delta_{(\varphi, \mathcal{T}, \ell^*)}(Y^*).$$

It follows by Corollary 3.4 that, given $\mathcal{T}, k, m,$ and φ as in its statement, we can find an edge weighting $\ell : E \rightarrow \mathbb{R}^{\geq 0}$ that maximises $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$ across all subsets of X of size k as follows. For all $e \in E,$ let (\mathcal{T}, ℓ_e) denote the edge-weighted phylogenetic X -tree, where ℓ_e is the edge weighting of \mathcal{T} in which $\ell_e(e) = m$ and $\ell_e(f) = 0$ for all $f \in E - \{e\}.$ Now, for each $e \in E,$ apply Theorem 3.2 and, more particularly, MAXDIVERSITYDIFF to $(\mathcal{T}, \ell_e), k,$ and $\varphi.$ The maximum of the values returned by these applications gives the desired value (as well as a subset of X of size k realising this value). Since the total number of edges in \mathcal{T} is $2|X| - 2,$ we have the next theorem.

Theorem 3.5 *The problem FIXED TREE AND TOTAL WEIGHT $(k, \mathcal{T}, m, \varphi)$ can be solved in $O(|X|^3)$ time, where $|X|$ is the size of the leaf set of $\mathcal{T}.$*

3.3 Maximum edge length

Next consider the problem of maximising the diversity difference on a fixed phylogenetic X -tree \mathcal{T} whose edge lengths are constrained by some maximum value, say $\kappa.$ We begin with a lemma that reduces the problem to an earlier problem.

Lemma 3.6 *Let \mathcal{T} be a phylogenetic X -tree, let $Y \subseteq X,$ and let κ be a non-negative real. If $\ell : E \rightarrow \mathbb{R}^{\geq 0}$ is a map on the set E of edges of \mathcal{T} such that $\ell(e) \leq \kappa$ for all $e \in E,$ and φ is a descendant diversity index, then*

$$\Delta_{(\varphi, \mathcal{T}, \ell_{\max})}(Y) \geq \Delta_{(\varphi, \mathcal{T}, \ell)}(Y),$$

where ℓ_{\max} is the map $\ell_{\max} : E \rightarrow \mathbb{R}^{\geq 0}$ defined by $\ell_{\max}(e) = \kappa$ for all $e \in E.$

Proof Let $e = (u, v) \in E,$ and let $\lambda_{(\mathcal{T}, \ell)}(e)$ and $\lambda_{(\mathcal{T}, \ell_{\max})}(e)$ denote the contribution of e to $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$ and $\Delta_{(\varphi, \mathcal{T}, \ell_{\max})}(Y),$ respectively. To prove the lemma, it suffices to show that $\lambda_{(\mathcal{T}, \ell)}(e) \leq \lambda_{(\mathcal{T}, \ell_{\max})}(e).$

Since $\ell(e) \leq \kappa$ and φ is linear,

$$\begin{aligned} \lambda_{(\mathcal{T}, \ell)}(e) &= \ell(e) - \sum_{y \in Y \cap C(v)} \gamma_{(\mathcal{T}, \ell)}(y, e) \cdot \ell(e) \\ &= \ell(e) \cdot \left(1 - \sum_{y \in Y \cap C(v)} \gamma_{(\mathcal{T}, \ell)}(y, e) \right) \end{aligned}$$

$$\begin{aligned} &\leq \kappa \cdot \left(1 - \sum_{y \in Y \cap C(v)} \gamma_{(\mathcal{T}, \ell)}(y, e) \right) \\ &= \lambda_{(\mathcal{T}, \ell_{\max})}(e), \end{aligned}$$

thereby completing the proof of the lemma. □

By Lemma 3.6, we may assume that all edges of \mathcal{T} have length κ . Thus the problem is reduced to maximising the diversity difference on a fixed phylogenetic tree with fixed edge lengths. Therefore, we can find the maximum value in time $O(|X|^2)$ by applying Theorem 3.2 and, in particular, MAXDIVERSITYDIFF, to $(\mathcal{T}, \ell_{\max})$, a non-negative integer k , and a descendant diversity index φ .

4 Maximising the diversity difference across all trees

In contrast to the computational results in last section, in this section we characterise the edge-weighted phylogenetic X -trees that maximise the diversity difference for FP and ES. If the size of the subset of interest is $|X|$, the diversity difference is zero for all edge-weighted phylogenetic X -trees. Thus, throughout this section, we will impose the condition that $k \leq |X| - 1$.

4.1 Maximum sum of edge lengths

We first consider the problem of maximising the diversity difference across all edge-weighted phylogenetic trees whose edge lengths are constrained by their (total) sum: ALL TREES AND TOTAL WEIGHT (k, X, m, φ)

Instance: A positive integer $k \leq |X| - 1$, a positive real number m , and a diversity index.

Question: Find an edge-weighted phylogenetic X -tree (\mathcal{T}, ℓ) with $L(\mathcal{T}, \ell) = m$ and a subset Y of X that maximises

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = PD_{(\mathcal{T}, \ell)}(Y) - \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x).$$

We begin with the Fair Proportion index.

Theorem 4.1 *Let (k, X, m, FP) be an instance of ALL TREES AND TOTAL WEIGHT. Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree with $L(\mathcal{T}, \ell) = m$ and let Y be a subset of X of size k . Suppose that (\mathcal{T}^*, ℓ^*) and Y^* is a solution to this instance of ALL TREES AND TOTAL WEIGHT. Then*

$$\Delta_{(FP, \mathcal{T}, \ell)}(Y) = \Delta_{(FP, \mathcal{T}^*, \ell^*)}(Y^*)$$

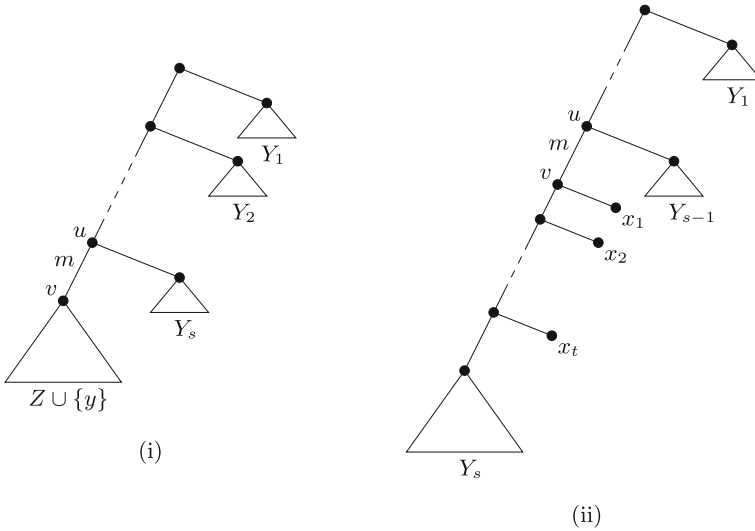


Fig. 2 (i) An illustration of Theorem 4.1, where $X - Y = Z, Z \cup \{y\} = C(v)$, and $Y = Y_1 \cup Y_2 \cup \dots \cup Y_s \cup \{y\}$. All edges have weight zero except the edge (u, v) which has weight m . The “triangles” represent pendant subtrees whose leaf sets are $Y_1, Y_2, \dots, Y_s, Z \cup \{y\}$. (ii) An illustration of Theorem 4.2, where $X - Y = \{x_1, x_2, \dots, x_t\}$, $Y = Y_1 \cup Y_2 \cup \dots \cup Y_s$, and Y_s is non-empty. All edges have weight zero except the edge (u, v) which has weight m . The “triangles” represent pendant subsets whose leaf sets are Y_1, Y_2, \dots, Y_s

if and only if (\mathcal{T}, ℓ) has an edge $e = (u, v)$ such that $\ell(e) = m$, $|Y \cap C(v)| = 1$, and $X - Y \subseteq C(v)$, in which case,

$$\Delta_{(FP, \mathcal{T}, \ell)}(Y) = m \left(1 - \frac{1}{n - k + 1} \right),$$

where $n = |X|$.

To illustrate Theorem 4.1, a generic edge-weighted phylogenetic X -tree optimising $\Delta_{(FP, \mathcal{T}, \ell)}(Y)$ as in the statement of this theorem is shown in Fig. 2(i).

Proof of Theorem 4.1 First note that if (\mathcal{T}, ℓ) and Y satisfy the conditions described in the statement of the theorem, then it follows that

$$\Delta_{(FP, \mathcal{T}, \ell)}(Y) = m \left(1 - \frac{1}{|X| - |Y| + 1} \right) = m \left(1 - \frac{1}{n - k + 1} \right).$$

To show that the only if direction holds, let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree, and suppose that (\mathcal{T}, ℓ) together with a subset Y of size k maximises $\Delta_{(FP, \mathcal{T}, \ell)}(Y)$ across all edge-weighted phylogenetic X -trees (\mathcal{T}', ℓ') with $L(\mathcal{T}', \ell') = m$ and subsets of X of size k . For each edge $e = (u, v)$ of \mathcal{T} , let $Y_e = Y \cap C(v)$ and, for convenience, $X_e = C(v) = X \cap C(v)$. Thus

$$\lambda_{(\mathcal{T}, \ell)}(e) = \ell(e) \left(1 - \frac{|Y_e|}{|X_e|} \right)$$

if $|Y_e| \neq 0$ and $\lambda_{(\mathcal{T}, \ell)}(e) = 0$ if $|Y_e| = 0$. Let e_1, e_2, \dots, e_t denote the edges of \mathcal{T} in which $\lambda_{(\mathcal{T}, \ell)}(e_i) > 0$. Note that, for such edges, $X_e - Y_e$ is non-empty. We next show that $t = 1$ by showing that if $t \geq 2$, then

$$\sum_{i \in \{1, 2, \dots, t\}} \lambda_{(\mathcal{T}, \ell)}(e_i) < m \left(1 - \frac{1}{|X| - |Y| + 1} \right),$$

contradicting the maximality of $\Delta_{(FP, \mathcal{T}, \ell)}(Y)$.

Suppose $t \geq 2$. Since FP is descendant, it follows by Lemma 3.3(i) that

$$\frac{|Y_{e_i}|}{|X_{e_i}|} = \frac{|Y_{e_j}|}{|X_{e_j}|}$$

for all $i, j \in \{1, 2, \dots, t\}$. Therefore

$$\begin{aligned} \sum_{i \in \{1, 2, \dots, t\}} \lambda_{(\mathcal{T}, \ell)}(e_i) &= \ell(e_1) \left(1 - \frac{|Y_{e_1}|}{|X_{e_1}|} \right) + \dots + \ell(e_t) \left(1 - \frac{|Y_{e_t}|}{|X_{e_t}|} \right) \\ &= \ell(e_1) \left(1 - \frac{|Y_{e_1}|}{|X_{e_1}|} \right) + \dots + \ell(e_t) \left(1 - \frac{|Y_{e_1}|}{|X_{e_1}|} \right) \\ &\leq m \left(1 - \frac{|Y_{e_1}|}{|X_{e_1}|} \right) \\ &< m \left(1 - \frac{|Y_{e_1}|}{|X| - |Y| + |Y_{e_1}|} \right) \\ &\leq m \left(1 - \frac{1}{|X| - |Y| + 1} \right), \end{aligned}$$

where the second-to-last inequality holds as the size of X_{e_1} can be no more than the sum of the sizes of $X - Y$ and $Y \cap Y_{e_1} = Y_{e_1}$. Moreover, it is strict when $t \geq 2$ since (i) if there is no directed path containing e_1 and e_2 , then X_{e_1} and X_{e_2} are disjoint and, in particular, $X - (X_{e_1} \cup Y)$ is non-empty and (ii) if, without loss of generality, there is a directed path from e_2 to e_1 , then $X_{e_1} \subset X_{e_2}$ and, as

$$\frac{|Y_{e_1}|}{|X_{e_1}|} = \frac{|Y_{e_2}|}{|X_{e_2}|},$$

again $X - (X_{e_1} \cup Y)$ is non-empty.

Hence $t = 1$ and so, to maximise the diversity difference, $\ell(e_1) = m$, $|Y_{e_1}| = 1$, and $X - Y \subseteq X_{e_1}$, in which case,

$$\Delta_{(FP, \mathcal{T}, \ell)}(Y) = m \left(1 - \frac{1}{n - k + 1} \right).$$

This completes the proof of the theorem. □

The next theorem is the analogue of Theorem 4.1 for the Equal Splits index.

Theorem 4.2 *Let (k, X, m, ES) be an instance of ALL TREES AND TOTAL WEIGHT. Let (\mathcal{T}, ℓ) be an edge-weight phylogenetic X -tree (\mathcal{T}, ℓ) with $L(\mathcal{T}, \ell) = m$ and let Y be a subset of X of size k , and suppose that (\mathcal{T}^*, ℓ^*) and Y^* is a solution to this instance of ALL TREES AND TOTAL WEIGHT. Then*

$$\Delta_{(ES, \mathcal{T}, \ell)}(Y) = \Delta_{(ES, \mathcal{T}^*, \ell^*)}(Y^*)$$

if and only if (\mathcal{T}, ℓ) has an edge $e = (u, v)$ such that $\ell(e) = m$, $X - Y$ is a chain whose first parent is v , and $Y \cap C(v)$ is non-empty, in which case,

$$\Delta_{(ES, \mathcal{T}, \ell)}(Y) = m \left(1 - \frac{1}{2^{n-k}} \right),$$

where $n = |X|$.

A generic edge-weighted phylogenetic X -tree optimising $\Delta_{(ES, \mathcal{T}, \ell)}(Y)$ as in the statement of Theorem 4.2 is shown in Fig. 2(ii). The proof of Theorem 4.2 takes a similar approach to that of Theorem 4.1 but is slightly more involved. We begin with a lemma.

Lemma 4.3 *Let \mathcal{T} be a phylogenetic X -tree, and let $Y \subseteq X$. Let $e = (u, v)$ be an edge of \mathcal{T} , and suppose that $(X - Y) \cap C(v)$ and $Y \cap C(v)$ are both non-empty. Let \mathcal{T}' be the phylogenetic X -tree obtained from \mathcal{T} by replacing the pendant subtree $\mathcal{T}|C(v)$ with a phylogenetic tree on leaf set $C(v)$, where $C(v) - Y$ is a chain whose first parent is v and $C(v) \cap Y$ is a pendant subtree below the chain. If $C(v) - Y$ is not a chain whose first parent is v in \mathcal{T} , then*

$$\frac{1}{2^{n_1-k_1}} = \sum_{y \in Y \cap C(v)} \frac{1}{2^{|P(\mathcal{T}'; v, y)|}} < \sum_{y \in Y \cap C(v)} \frac{1}{2^{|P(\mathcal{T}; v, y)|}},$$

where $n_1 = |C(v)|$ and $k_1 = |Y \cap C(v)|$.

Proof Suppose that \mathcal{T} is not of the same form as \mathcal{T}' . Then there is a $z \in C(v) - Y$ such that either z is in a cherry $\{z, z'\}$ of \mathcal{T} , where $z' \notin Y$ or z is a leaf of a pendant subtree of \mathcal{T} whose two maximal subtrees each contain an element of Y . Let \mathcal{T}'' be the phylogenetic X -tree that is obtained from \mathcal{T} by pruning z and regrafting to the edge e . Relabel the vertex v as v' and the newly created vertex in the subdivided edge v (so that $C(v)$ is unchanged). It is now easily checked that if $y \in Y \cap C(v)$, then $|P(\mathcal{T}; v, y)| \leq |P(\mathcal{T}''; v, y)|$. Moreover, for some $y' \in Y \cap C(v)$, we have $|P(\mathcal{T}; v, y')| < |P(\mathcal{T}''; v, y')|$. In particular,

$$\sum_{y \in Y \cap C(v)} \frac{1}{2^{|P(\mathcal{T}''; v, y)|}} < \sum_{y \in Y \cap C(v)} \frac{1}{2^{|P(\mathcal{T}; v, y)|}}.$$

The lemma now follows by repeating this process until we have constructed a phylogenetic tree in the same form as \mathcal{T}' . □

Proof of Theorem 4.2 If (\mathcal{T}, ℓ) and Y satisfy the conditions described in the statement of Theorem 4.2, then it is easily checked that

$$\Delta_{(ES, \mathcal{T}, \ell)}(Y) = m \left(1 - \frac{1}{2^{n-k}} \right).$$

To complete the proof, let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree, and suppose that (\mathcal{T}, ℓ) together with a subset Y of X of size k maximises $\Delta_{(ES, \mathcal{T}, \ell)}(Y)$ across all edge-weighted phylogenetic X -trees (\mathcal{T}', ℓ') with $L(\mathcal{T}', \ell') = m$ and subsets of X of size k . For each edge $e = (u, v)$ of \mathcal{T} , define $\lambda_{(\mathcal{T}, \ell)}(e)$ to be the contribution of e to $\Delta_{(ES, \mathcal{T}, \ell)}(Y)$. Thus

$$\lambda_{(\mathcal{T}, \ell)}(e) = \ell(e) \left(1 - \sum_{y \in Y \cap C(v)} \frac{1}{2^{|P(\mathcal{T}; v, y)|}} \right)$$

if $|Y \cap C(v)| \neq 0$ and $\lambda_{(\mathcal{T}, \ell)}(e) = 0$ if $|Y \cap C(v)| = 0$. Note that, if $\lambda_{(\mathcal{T}, \ell)}(e) > 0$, then $C(v) - Y$ is non-empty. Let $e_1 = (u_1, v_1), e_2 = (u_2, v_2), \dots, e_t = (u_t, v_t)$ denote the edges of \mathcal{T} in which $\lambda_{(\mathcal{T}, \ell)}(e_i) > 0$. We next show that $t = 1$ by showing that if $t > 1$ then

$$m \left(1 - \frac{1}{2^{n-k}} \right) > \sum_{i \in \{1, 2, \dots, t\}} \lambda_{(\mathcal{T}, \ell)}(e_i).$$

Say $t \geq 2$. Since ES is descendant, it follows by Lemma 3.3(i) that

$$\sum_{y \in Y \cap C(v_i)} \frac{1}{2^{|P(\mathcal{T}; v_i, y)|}} = \sum_{y \in Y \cap C(v_j)} \frac{1}{2^{|P(\mathcal{T}; v_j, y)|}} \tag{5}$$

for all $i, j \in \{1, 2, \dots, t\}$. Hence

$$\begin{aligned} \sum_{i \in \{1, 2, \dots, t\}} \lambda_{(\mathcal{T}, \ell)}(e_i) &= \ell(e_1) \left(1 - \sum_{y \in Y \cap C(v_1)} \frac{1}{2^{|P(\mathcal{T}; v_1, y)|}} \right) + \dots \\ &\quad \dots + \ell(e_t) \left(1 - \sum_{y \in Y \cap C(v_t)} \frac{1}{2^{|P(\mathcal{T}; v_t, y)|}} \right) \\ &= \ell(e_1) \left(1 - \sum_{y \in Y \cap C(v_1)} \frac{1}{2^{|P(\mathcal{T}; v_1, y)|}} \right) + \dots \\ &\quad \dots + \ell(e_t) \left(1 - \sum_{y \in Y \cap C(v_1)} \frac{1}{2^{|P(\mathcal{T}; v_1, y)|}} \right) \end{aligned}$$

$$\begin{aligned} &\leq m \left(1 - \sum_{y \in Y \cap C(v_1)} \frac{1}{2^{|P(\mathcal{T}; v_1, y)|}} \right) \\ &\leq m \left(1 - \frac{1}{2^{n_1 - k_1}} \right), \end{aligned}$$

where $n_1 = |C(v_1)|$ and $k_1 = |Y \cap C(v_1)|$, and the last inequality holds by Lemma 4.3. Since $t \geq 2$, it follows by (5) that $X - (C(v_1) \cup Y)$ is non-empty, so $n - n_1 > 0$. That is,

$$\sum_{i \in \{1, 2, \dots, t\}} \lambda_{(\mathcal{T}, \ell)}(e_i) \leq m \left(1 - \frac{1}{2^{n_1 - k_1}} \right) < m \left(1 - \frac{1}{2^{n - k_1}} \right).$$

Hence $t = 1$, and so to maximise the diversity difference, a single edge $e = (u, v)$ say of (\mathcal{T}, ℓ) has length $\ell(e) = m$ and, by Lemma 4.3, $X - Y$ is a chain with first parent v , and $Y \cap C(v)$ is non-empty, in which case,

$$\Delta_{(ES, \mathcal{T}, \ell)}(Y) = m \left(1 - \frac{1}{2^{n - k}} \right).$$

This completes the proof of the theorem. □

4.2 Maximum edge length

We next consider the problem of maximising the diversity difference across all phylogenetic trees whose edge lengths are constrained by some maximum value. By Lemma 3.6, we may assume all edges of \mathcal{T} have the same weight. Thus, without loss of generality, we will assume that all edges have weight 1 and so, for simplicity, we write \mathcal{T} for (\mathcal{T}, ℓ) . To this end, for a descendant diversity index φ , the contribution of an edge e of \mathcal{T} to $\Delta_{(\varphi, \mathcal{T})}(Y)$ is denoted by $\lambda_{\mathcal{T}}(e)$. We begin with two lemmas and a corollary.

Lemma 4.4 *Let \mathcal{T} be a phylogenetic X -tree, let $Y \subseteq X$ such that $|Y| \leq |X| - 1$, and let $\varphi \in \{FP, ES\}$. If C is a non-trivial cluster of \mathcal{T} such that either $C \subseteq Y$ or $C \subseteq X - Y$, then there exists a phylogenetic X -tree \mathcal{T}' such that*

$$\Delta_{(\varphi, \mathcal{T}')} > \Delta_{(\varphi, \mathcal{T})}.$$

Proof Let v be a non-leaf vertex of \mathcal{T} , and suppose that $C(v) \subseteq Y$ or $C(v) \subseteq (X - Y)$. Without loss of generality, we may assume that $C(v)$ is maximal with this property. Let (u, v) be the edge of \mathcal{T} directed into v and note that, for (u, v) and all edges on a path from v to a leaf, the contribution of these edges under either FP or ES to $\Delta_{(\varphi, \mathcal{T})}(Y)$ is zero regardless of whether $C(v) \subseteq Y$ or $C(v) \subseteq (X - Y)$. Let w be the child vertex of u that is not v . By maximality and $|Y| \neq |X|$, if $C(v) \subseteq Y$, then $(X - Y) \cap C(w) \neq \emptyset$ and, if $C(v) \subseteq (X - Y)$, then $Y \cap C(w) \neq \emptyset$. Let E denote the edge set of \mathcal{T} , let P denote the set of edges of \mathcal{T} on the path from its root ρ to u , and

let $a \in C(v)$. Depending on whether $C(v) \subseteq Y$ or $C(v) \subseteq (X - Y)$, we will compare \mathcal{T} with two other phylogenetic X -trees.

Let \mathcal{T}_1 denote the phylogenetic X -tree obtained from \mathcal{T} by pruning a and regrafting it to ρ . Let E_1 denote the set of edges of \mathcal{T}_1 , let ρ_1 denote the root of \mathcal{T}_1 , let q denote the child of ρ_1 that is not a , and let P_1 denote the set of edges of \mathcal{T}_1 on the path from q to u . Let \mathcal{T}_2 denote the phylogenetic X -tree obtained from \mathcal{T} by pruning a and regrafting it to (u, w) . Let E_2 denote the set of edges of \mathcal{T}_2 , let p_a denote the parent of a in \mathcal{T}_2 , and let P_2 denote the set of edges of \mathcal{T}_2 on the path from its root to p_a .

First suppose that φ is FP. If $C(v) \subseteq Y$, then

$$\sum_{e \in E-P} \lambda_{\mathcal{T}}(e) = \sum_{e_1 \in E_1 - (P_1 \cup \{(\rho_1, q)\})} \lambda_{\mathcal{T}_1}(e_1).$$

Furthermore, as $a \in Y$, and $C_{\mathcal{T}_1}(u) \cap Y$ and $C_{\mathcal{T}_1}(u) \cap (X - Y)$ are both non-empty,

$$\sum_{e \in P} \lambda_{\mathcal{T}}(e) < \sum_{e_1 \in P_1} \lambda_{\mathcal{T}_1}(e_1)$$

and $\lambda_{\mathcal{T}_1}((\rho_1, q)) > 0$. Hence $\Delta_{(FP, \mathcal{T}_1)}(Y) > \Delta_{(FP, \mathcal{T})}(Y)$. On the other hand, if $C(v) \subseteq (X - Y)$, then it is easily checked that

$$\sum_{e \in E} \lambda_{\mathcal{T}}(e) = \sum_{e_2 \in E_2 - \{(u, p_a)\}} \lambda_{\mathcal{T}_2}(e_2)$$

and

$$\lambda_{\mathcal{T}_2}((u, p_a)) > 0$$

as $C_{\mathcal{T}_2}(p_a)$ has a non-empty intersection with Y and $X - Y$. Thus $\Delta_{(FP, \mathcal{T}_2)}(Y) > \Delta_{(FP, \mathcal{T})}(Y)$, and so the lemma holds if φ is FP.

Now suppose that φ is ES. If $C(v) \subseteq Y$, then

$$\sum_{e \in E} \lambda_{\mathcal{T}}(e) = \sum_{e_1 \in E_1 - \{(\rho_1, q)\}} \lambda_{\mathcal{T}_1}(e_1).$$

Since $C_{\mathcal{T}_1}(q) \cap Y$ and $C_{\mathcal{T}_1}(q) \cap (X - Y)$ are both non-empty, $\lambda_{\mathcal{T}_1}((\rho_1, q)) > 0$, and so $\Delta_{(ES, \mathcal{T}_1)}(Y) > \Delta_{(ES, \mathcal{T})}(Y)$ if φ is ES. Furthermore, if $C(v) \subseteq (X - Y)$, then

$$\sum_{e \in E-P} \lambda_{\mathcal{T}}(e) = \sum_{e_2 \in E_2 - P_2} \lambda_{\mathcal{T}_2}(e_2).$$

Hence, as $a \in (X - Y)$,

$$\Delta_{(ES, \mathcal{T}_2)}(Y) - \Delta_{(ES, \mathcal{T})}(Y)$$

$$\begin{aligned}
 &= \sum_{e_2 \in P_2} \lambda_{\mathcal{T}_2}(e_2) - \sum_{e \in P} \lambda_{\mathcal{T}}(e) \\
 &= \sum_{e_2 \in P_2 - \{(u, p_a)\}} \left(1 - \sum_{y \in Y \cap C(w)} \frac{1}{2^{|P(\mathcal{T}_2; e_2, y)|}} \right) + \lambda_{\mathcal{T}_2}((u, p_a)) \\
 &\quad - \sum_{e \in P} \left(1 - \sum_{y \in Y \cap C(w)} \frac{1}{2^{|P(\mathcal{T}; e, y)|}} \right) \\
 &= -\frac{1}{2} \cdot \left(\sum_{e \in P} \left(\sum_{y \in Y \cap C(w)} \frac{1}{2^{|P(\mathcal{T}; e, y)|}} \right) \right) + \lambda_{\mathcal{T}_2}((u, p_a)) \\
 &\quad + \sum_{e \in P} \left(\sum_{y \in Y \cap C(w)} \frac{1}{2^{|P(\mathcal{T}; e, y)|}} \right) \\
 &= \frac{1}{2} \cdot \left(\sum_{y \in Y \cap C(w)} \frac{1}{2^{|P(\mathcal{T}; e, y)|}} \right) + \lambda_{\mathcal{T}_2}((u, p_a)) \\
 &> 0
 \end{aligned}$$

as $\lambda_{\mathcal{T}_2}((u, p_a)) > 0$ since $C_{\mathcal{T}_2}(p_a) \cap Y$ and $C_{\mathcal{T}_2}(p_a) \cap (X - Y)$ are both non-empty. Therefore $\Delta_{(ES, \mathcal{T}_2)}(Y) > \Delta_{(ES, \mathcal{T})}(Y)$. This completes the proof of the lemma. \square

An immediate consequence of Lemma 4.4 is the next corollary.

Corollary 4.5 *Let \mathcal{T} be a phylogenetic X -tree, let k be a positive integer such that $k \leq |X| - 1$, let Y be a subset of X of size k , and let $\varphi \in \{FP, ES\}$. Suppose that \mathcal{T} and Y maximises*

$$\Delta_{(\varphi, \mathcal{T})}(Y) = PD_{\mathcal{T}}(Y) - \sum_{x \in Y} \varphi_{\mathcal{T}}(x)$$

across all phylogenetic X -trees and all subsets of X of size k . If $\{a, b\}$ is a cherry of \mathcal{T} , then $|\{a, b\} \cap Y| = 1$. In particular, \mathcal{T} has no non-trivial cluster C such that either $C \subseteq Y$ or $C \subseteq (X - Y)$.

Lemma 4.6 *Let \mathcal{T} be a phylogenetic X -tree, let k be a positive integer such that $k \leq |X| - 1$, let Y be a subset of X of size k , and let $\varphi \in \{FP, ES\}$. Suppose that \mathcal{T} and Y maximises*

$$\Delta_{(\varphi, \mathcal{T})}(Y) = PD_{\mathcal{T}}(Y) - \sum_{x \in Y} \varphi_{\mathcal{T}}(x)$$

across all phylogenetic X -trees and all subsets of X of size k . If (a_1, a_2, \dots, a_s) is a chain of \mathcal{T} , in which $\{a_1, a_2\}$ is a cherry, then

- (i) $|\{a_1, a_2\} \cap Y| = 1$ and

(ii) for some $p \geq 3$, we have $\{a_1, a_2, \dots, a_{p-1}\} \cap Y = 1$ and $\{a_p, a_{p+1}, \dots, a_s\} \subseteq Y$.

Proof By Corollary 4.5, (a_1, a_2, \dots, a_s) satisfies (i). Suppose that (a_1, a_2, \dots, a_s) does not satisfy (ii). Let $i \geq 3$ be the least index such $a_i \in Y$ but $a_{i+1} \in X - Y$. Let \mathcal{T}' be the phylogenetic X -tree obtained from \mathcal{T} by interchanging a_i and a_{i+1} , that is, pruning a_i and regrafting it to the edge directed into the parent of a_{i+1} so that $(a_1, a_2, \dots, a_{i-1}, a_{i+1}, a_i, a_{i+2}, \dots, a_s)$ is a chain of \mathcal{T}' . Let p_i and p_{i+1} denote the parents of a_i and a_{i+1} in \mathcal{T} , respectively, and let p'_i and p'_{i+1} denote the parents of a_i and a_{i+1} in \mathcal{T}' , respectively.

If φ is the Fair Proportion index, then

$$\begin{aligned} \Delta_{(FP, \mathcal{T}')} (Y) - \Delta_{(FP, \mathcal{T})} (Y) &= \lambda_{\mathcal{T}'}((p'_i, p'_{i+1})) - \lambda_{\mathcal{T}}((p_{i+1}, p_i)) \\ &= \left(1 - \frac{1}{i}\right) - \left(1 - \frac{2}{i}\right) \\ &> 0. \end{aligned}$$

This contradiction to the maximality of \mathcal{T} and Y implies that the lemma holds when φ is FP.

Now suppose that φ is the Equal Splits index. Let P denote the path in \mathcal{T} from the root to p_i and let P' denote the path in \mathcal{T}' from the root to p'_i . Then

$$\begin{aligned} \Delta_{(ES, \mathcal{T}')} (Y) - \Delta_{(ES, \mathcal{T})} (Y) &= -\varphi_{\mathcal{T}'}(a_i) + \varphi_{\mathcal{T}}(a_i) \\ &= -\sum_{e' \in P'} \frac{1}{2^{|P(\mathcal{T}'; e', a_i)|}} + \sum_{e \in P} \frac{1}{2^{|P(\mathcal{T}; e, a_i)|}}. \end{aligned}$$

For all $e \in P - \{(p_{i+1}, p_i)\}$, we have $|P(\mathcal{T}'; e, a_i)| = |P(\mathcal{T}; e, a_i)| - 1$, and the contribution of (p_{i+1}, p_i) to $ES_{\mathcal{T}}(a_i)$ is $\frac{1}{2}$. So

$$\begin{aligned} \Delta_{(ES, \mathcal{T}')} (Y) - \Delta_{(ES, \mathcal{T})} (Y) &= -2 \sum_{e \in P - \{(p_{i+1}, p_i)\}} \frac{1}{2^{|P(\mathcal{T}; e, a_i)|}} \\ &\quad + \left(\frac{1}{2} + \sum_{e \in P - \{(p_{i+1}, p_i)\}} \frac{1}{2^{|P(\mathcal{T}; e, a_i)|}} \right) \\ &= \frac{1}{2} - \sum_{e \in P - \{(p_{i+1}, p_i)\}} \frac{1}{2^{|P(\mathcal{T}; e, a_i)|}} \\ &\geq \frac{1}{2} - \left(\frac{1}{4} + \frac{1}{8} + \dots + \frac{1}{2^{|P|}} \right) \\ &> 0. \end{aligned}$$

This contradiction to the maximality of \mathcal{T} and Y implies that the lemma holds when φ is ES, thereby completing the proof of the lemma. \square

The next theorem is illustrated in Fig. 3. In particular, an edge-weighted phylogenetic X -tree optimising $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$ as in the statement of this theorem is shown in

Fig. 3 An illustration of Theorem 4.7, where $Y = \{x_1, x_i, x_{i+1}, \dots, x_n\}$, so $|\{x_1, x_2\} \cap Y| = 1$, and $X - Y = \{x_2, x_3, \dots, x_{i-1}\}$

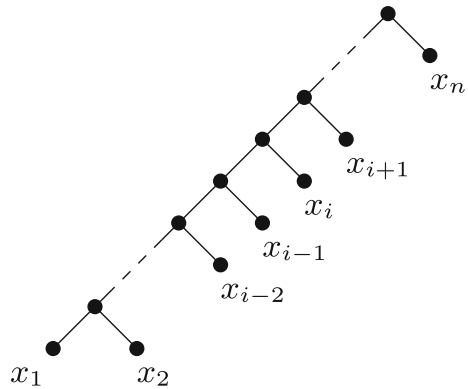


Fig. 3, where $\varphi \in \{FP, ES\}$. It is interesting to note that the outcome of Theorem 4.7 is the same for FP and ES, although the proof of the theorem requires FP and ES to be considered separately.

Theorem 4.7 *Let k be a positive integer such that $k \leq |X| - 1$, let κ be a positive real, and let $\varphi \in \{FP, ES\}$. Then an edge-weighted phylogenetic X -tree (\mathcal{T}, ℓ) with $\ell(e) \leq \kappa$ for all $e \in E(\mathcal{T})$ and a subset Y of X of size k maximises*

$$\Delta_{(\varphi, \mathcal{T}, \ell)}(Y) = PD_{(\mathcal{T}, \ell)}(Y) - \sum_{x \in Y} \varphi_{(\mathcal{T}, \ell)}(x)$$

across all phylogenetic X -trees (\mathcal{T}', ℓ') with $\ell'(e) \leq \kappa$ for all $e \in E(\mathcal{T}')$ and all subsets of X of size k if and only if $\ell(e) = \kappa$ for all $e \in E(\mathcal{T})$ and X is a chain (x_1, x_2, \dots, x_n) such that $|\{x_1, x_2\} \cap Y| = 1$, and, for some $i \geq 3$, $|\{x_1, x_2, \dots, x_{i-1}\} \cap Y| = 1$ and $\{x_i, x_{i+1}, \dots, x_n\} \subseteq Y$.

Proof Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree with $\ell(e) \leq \kappa$ for all $e \in E(\mathcal{T})$ and let Y be a subset of X of size k . Suppose that (\mathcal{T}, ℓ) and Y maximise $\Delta_{(\varphi, \mathcal{T}, \ell)}(Y)$ across all edge-weighted phylogenetic X -trees (\mathcal{T}', ℓ') with $\ell'(e) \leq \kappa$ for all $e \in E(\mathcal{T}')$ and all subsets of X of size k . Recall that, by Lemma 3.6, we may assume that all edges have weight $\kappa = 1$.

We first show that \mathcal{T} does not have two distinct cherries. Suppose that \mathcal{T} has two such cherries. Then \mathcal{T} has a vertex v in which $C(v)$ contains exactly two cherries. By Lemma 4.6 and the maximality of (\mathcal{T}, ℓ) and Y , the set $C(v)$ consists of two disjoint chains (a_1, a_2, \dots, a_s) and (b_1, b_2, \dots, b_t) each of which satisfies properties (i) and (ii) in the statement of Lemma 4.6. Let $A = \{a_1, a_2, \dots, a_s\}$ and let $B = \{b_1, b_2, \dots, b_t\}$.

First suppose that φ is FP. Without loss of generality, we may assume that

$$\frac{|Y \cap A|}{s} \geq \frac{|Y \cap B|}{t}.$$

Say $s \geq 3$ and $a_s \in Y$. Let \mathcal{T}' be the phylogenetic X -tree obtained from \mathcal{T} by pruning a_s and regrafting it to the edge (u, v) directed into v . This operation is shown in Fig. 4,

where \mathcal{T} is drawn above \mathcal{T}' . Let p_s and p'_s denote the parents of a_s in \mathcal{T} and \mathcal{T}' , respectively. Then it is easily seen that

$$\sum_{e \in E(\mathcal{T}) - \{(v, p_s)\}} \lambda_{\mathcal{T}}(e) = \sum_{e' \in E(\mathcal{T}') - \{(p'_s, v)\}} \lambda_{\mathcal{T}'}(e').$$

Furthermore, as $|Y \cap A|/s \geq |Y \cap B|/t$,

$$\begin{aligned} \lambda_{\mathcal{T}}((v, p_s)) &= 1 - \frac{|Y \cap A|}{s} \\ &\leq 1 - \frac{|Y \cap (A \cup B)|}{s + t} \\ &< 1 - \frac{|Y \cap (A \cup B)| - 1}{s + t - 1} \\ &= \lambda_{\mathcal{T}'}((p'_s, v)). \end{aligned}$$

Thus $\Delta_{(FP, \mathcal{T})}(Y) < \Delta_{(FP, \mathcal{T}')}(Y)$. It follows that $|Y \cap A| = 1$, and so we may assume that $Y \cap A = \{a_1\}$.

Say $t \geq 3$ and $b_t \in Y$. Now let \mathcal{T}' be the phylogenetic X -tree obtained from \mathcal{T} by pruning b_t and regrafting it to (u, v) . let p_t and p'_t denote the parents of b_t in \mathcal{T} and \mathcal{T}' , respectively. Then

$$\sum_{e \in E(\mathcal{T}) - \{(v, p_t)\}} \lambda_{\mathcal{T}}(e) = \sum_{e' \in E(\mathcal{T}') - \{(p'_t, v)\}} \lambda_{\mathcal{T}'}(e').$$

Also, as $|Y \cap A| = 1$ and $|Y \cap B| \geq 2$,

$$\begin{aligned} \lambda_{\mathcal{T}}((v, p_t)) &= 1 - \frac{|Y \cap B|}{t} \\ &< 1 - \frac{|Y \cap B|}{s + t} \\ &\leq 1 - \frac{|Y \cap (A \cup B)| - 1}{s + t - 1} \\ &= \lambda_{\mathcal{T}'}((p'_t, v)). \end{aligned}$$

Thus $\Delta_{(FP, \mathcal{T})}(Y) < \Delta_{(FP, \mathcal{T}')}(Y)$, contradicting the maximality of (\mathcal{T}, ℓ) and Y . Therefore $|Y \cap B| = 1$, and so we may assume that $Y \cap (A \cup B) = \{a_1, b_1\}$.

Without loss of generality, we may now assume that $t \geq s$. Let p_s and q_t denote the parents of a_s and b_t in \mathcal{T} , respectively. Let \mathcal{T}'' be the phylogenetic X -tree obtained from \mathcal{T} by pruning a_s and regrafting it to the edge (u, q_t) . Let p''_s denote the parent of a_s in \mathcal{T}'' . Then

$$\sum_{e \in E(\mathcal{T}) - \{(v, p_s)\}} \lambda_{\mathcal{T}}(e) = \sum_{e'' \in E(\mathcal{T}'') - \{(v, p''_s)\}} \lambda_{\mathcal{T}''}(e'')$$

and

$$\begin{aligned} \lambda_{\mathcal{T}}((u, p_s)) &= 1 - \frac{1}{s} \\ &< 1 - \frac{1}{t+1} \\ &= \lambda_{\mathcal{T}''}((u, p_s'')) \end{aligned}$$

as $t \geq s$. Repeating this process for each of $a_{s-1}, a_{s-2}, \dots, a_2$, we obtain a phylogenetic tree X -tree \mathcal{T}_1 with one less cherry than \mathcal{T} , and

$$\Delta_{(FP, \mathcal{T})}(Y) < \Delta_{(FP, \mathcal{T}_1)}(Y),$$

a contradiction to maximality. We deduce that \mathcal{T} has exactly one cherry if φ is FP.

Now suppose that φ is ES. Without loss of generality, we may assume that $t \geq s$ and $a_1, b_1 \in Y$. Let \mathcal{T}' be the phylogenetic X -tree obtained from \mathcal{T} by pruning a_2 and regrafting it to the edge (q_2, b_1) , where q_2 is the parent of b_2 in \mathcal{T} (as depicted in Fig. 5). Then, noting that a_2 is not in Y ,

$$\sum_{y \in (Y - \{a_1, b_1\})} ES_{\mathcal{T}}(y) = \sum_{y \in (Y - \{a_1, b_1\})} ES_{\mathcal{T}'}(y).$$

Furthermore, since the length of the each path from an edge to a_1 is reduced by 1 in forming \mathcal{T}' as edge (p_3, p_2) is suppressed,

$$\sum_{e \in P(\mathcal{T}; \rho, a_1)} \frac{1}{2^{|P(\mathcal{T}; e, a_1)|}} - \sum_{e' \in P(\mathcal{T}'; \rho, a_1)} \frac{1}{2^{|P(\mathcal{T}'; e', a_1)|}} = \frac{1}{2^{|P(\mathcal{T}; \rho, a_1)|-1}}$$

and, since the length of the each path from an edge to b_1 is increased by 1 in forming \mathcal{T}' as edge (q_2, p_2') is added,

$$\sum_{e' \in P(\mathcal{T}'; \rho, b_1)} \frac{1}{2^{|P(\mathcal{T}'; e', b_1)|}} - \sum_{e \in P(\mathcal{T}; \rho, b_1)} \frac{1}{2^{|P(\mathcal{T}; e, b_1)|}} = \frac{1}{2^{|P(\mathcal{T}'; \rho, b_1)|-1}}.$$

But, as $t \geq s$,

$$\frac{1}{2^{|P(\mathcal{T}'; \rho, b_1)|-1}} < \frac{1}{2^{|P(\mathcal{T}; \rho_1, a_1)|-1}}.$$

Therefore

$$\begin{aligned} \sum_{y \in \{a_1, b_1\}} ES_{\mathcal{T}}(y) - \sum_{y \in \{a_1, b_1\}} ES_{\mathcal{T}'}(y) &= \frac{1}{2^{|P(\mathcal{T}; \rho, a_1)|-1}} - \frac{1}{2^{|P(\mathcal{T}'; \rho, b_1)|-1}} \\ &> 0. \end{aligned}$$

It now follows that $\Delta_{(ES, \mathcal{T})}(Y) < \Delta_{(ES, \mathcal{T}')} (Y)$. This contradiction to maximality implies that \mathcal{T} has exactly one cherry if φ is ES. It now follows by Lemma 4.6 that,

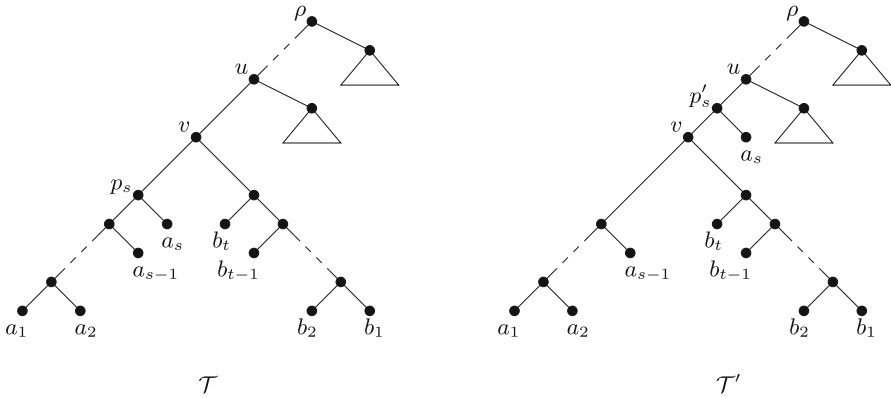


Fig. 4 Illustrating the pruning and regrafting of the leaf a_s in the proof of Theorem 4.7 for FP

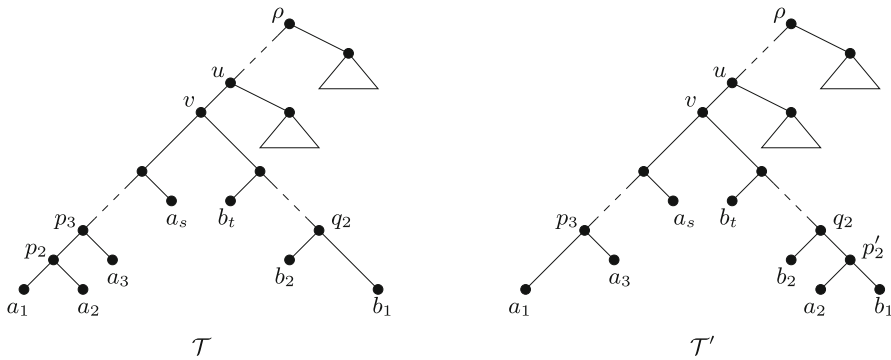


Fig. 5 Illustrating the pruning and regrafting of the leaf a_2 in the proof of Theorem 4.7 for ES

if $\varphi \in \{FP, ES\}$, then to maximise $\Delta_{(\varphi, \mathcal{T})}(Y)$ we have that X is a chain and Y is a subset of X of size k as described in the statement of the theorem. \square

5 Discussion

The results for Fair Proportion and Equal Splits in the last section are strikingly similar. Indeed, they are exactly the same when considering the outcomes of Theorem 4.7. While the proof of Theorem 4.7 eventually required separating into two parts to independently consider FP and ES, the two approaches taken were alike. This suggests that there is probably a natural class of phylogenetic diversity indices which these results are representative of. If so, what is this class? The class of descendant diversity indices is unlikely to be sufficient. However, what if we additionally required the following property? Let (\mathcal{T}, ℓ) be an edge-weighted phylogenetic X -tree and let φ be a descendant diversity index. Let $e = (u, v)$ be an edge of \mathcal{T} , and let $\{x, y\} \subseteq X$ such that $\{x, y\} \subseteq C(v)$. If the number of edges from v to y is at most the number of edges

from v to x , then the contribution of $\ell(e)$ to $\varphi_{(\mathcal{T}, \ell)}(x)$ is at most the contribution of $\ell(e)$ to $\varphi_{(\mathcal{T}, \ell)}(y)$.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Cadotte MW, Cardinale BJ, Oakley TH (2008) Evolutionary history and the effect of biodiversity on plant productivity. *Proc Natl Acad Sci* 105:17012–17017
- Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biol Conserv* 61:1–10
- Haake C-J, Kashiwada A, Su FE (2008) The Shapely value of phylogenetic trees. *J Math Biol* 56:479–497
- Isaac NJB, Turvey ST, Collen B, Waterman C, Baillie JEM (2007) Mammals on the EDGE: Conservation priorities based on threat and phylogeny. *PLoS One* 2:e296
- Lozupone CA, Knight R (2007) Global patterns in bacterial diversity. *Proc Natl Acad Sci* 104:11436–11440
- Pardi F, Goldman N (2005) Species choice for comparative genomics: being greedy works. *PLoS Genet* 1:e71
- Redding DW (2003) Incorporating genetic distinctness and reserve occupancy into a conservation prioritisation approach. Master's thesis. University of East Anglia, Norwich
- Redding DW, DeWolff CV, Mooers AO (2010) Evolutionary distinctiveness, threat status, and ecological oddity in primates. *Conserv Biol* 24:1052–1058
- Redding DW, Hartmann K, Mimoto A, Bokal D, DeVos M, Mooers AO (2008) Evolutionary distinctive species often capture more phylogenetic diversity than expected. *J Theor Biol* 251:606–615
- Redding DW, Mazel F, Mooers AO (2014) Measuring evolutionary isolation for conservation. *PLoS One* 9:e113490
- Redding DW, Mooers AO (2006) Incorporating evolutionary measures into conservation prioritization. *Conserv Biol* 20:1670–1678
- Safi K, Cianciaruso MV, Loyola RD, Brito D, Armour-Marshall K, Diniz-Filho JAF (2011) Understanding global patterns of mammalian functional and phylogenetic diversity. *Phil Trans R Soc B* 366:2536–2544
- Steel M (2005) Phylogenetic diversity and the greedy algorithm. *Syst Biol* 54:527–529
- Stein RW, Mull CG, Kuhn TS, Aschliman NC, Davidson LNK, Joy JB, Smith GJ, Dulvy NK, Mooers AO (2018) Global priorities for conserving the evolutionary history of sharks, rays and chimaeras. *Nat Ecol Evol* 2:288–298
- Wicke K, Steel M (2020) Combinatorial properties of phylogenetic diversity indices. *J Math Biol* 80:687–715

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.