Journal of Theoretical Biology 430 (2017) 207–214

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Comparing the rankings obtained from two biodiversity indices: the Fair Proportion Index and the Shapley Value

Kristina Wicke, Mareike Fischer*

Department of Mathematics and Computer Science, University Greifswald, Greifswald, Germany

ARTICLE INFO

Article history:
Received 20 March 2017
Revised 10 July 2017
Accepted 13 July 2017
Available online 15 July 2017

Keywords:
Phylogenetic diversity
Shapley Value
Fair Proportion Index
Ranking order
Ultrametric
Computation

ABSTRACT

The Shapley Value and the Fair Proportion Index of phylogenetic trees have been frequently discussed as prioritization tools in conservation biology. Both indices rank species according to their contribution to total phylogenetic diversity, allowing for a simple conservation criterion. While both indices have their specific advantages and drawbacks, it has recently been shown that both values are closely related. However, as different authors use different definitions of the Shapley Value, the specific degree of relatedness depends on the specific version of the Shapley Value – it ranges from a high correlation index to equality of the indices. In this note, we first give an overview of the different indices. Then we turn our attention to the mere ranking order provided by either of the indices. We compare the rankings obtained from different versions of the Shapley Value for a phylogenetic tree of European amphibians and illustrate their differences. We then undertake further analyses on simulated data and show that even though the chance of two rankings being exactly identical (when obtained from different versions of the Shapley Value) decreases with an increasing number of taxa, the distance between the two rankings converges to zero, i.e., the rankings are becoming more and more alike. Moreover, we introduce our freely available software package FairShapley, which was implemented in Perl and with which all calculations have been performed.

1. Introduction

Due to limited financial means, biodiversity conservation programs often need to prioritize the species to conserve. Two indices used in this matter are the Shapley Value and the Fair Proportion Index. Both are based on phylogenetic trees and rank species according to their contribution to overall biodiversity.

The Shapley Value was first introduced by Haake et al. (2007) for unrooted trees and reflects the average biodiversity contribution of a species. The Fair Proportion Index, on the other hand, lacks a biological link to conservation, but is significantly easier to calculate and has been preferred in practice. Under a different name (ED for *Evolutionary Distinctiveness*) the Fair Proportion Index has for example been used in the 'EDGE of Existence' Project, established by the *Zoological Society of London* in 2007 (see Isaac et al., 2007). However, Hartmann (2013) observed a strong correlation between the Shapley Value and the Fair Proportion Index on rooted trees, where the Shapley Value was calculated for the unrooted version of the tree by suppressing the root vertex. Very recently, Fuchs and Jin (2015) have extended the concept of the Shapley Value to rooted trees and have shown that the two indices are identical for these trees. They also introduced a slightly modified version of the Shapley Value, which again is highly correlated to the Fair Proportion Index.

In this note we first give an overview of the various versions of the Shapley Value and their respective relatedness with the Fair Proportion Index, before we focus on the mere ranking order of taxa obtained from different versions of the indices. Although the indices are highly correlated, they can result in different ranking orders, especially when the trees become large. We will show with a simulation study based on random trees that in fact, despite the increasing correlation as the number of species grows, different ranking orders are still more likely than equal ones. Therefore, in order to demonstrate what the correlation really implies, we treat the ranking lists as vectors and use the so-called Manhattan distance to measure the difference between two rankings suggested by different indices. We then show that the distance between these rankings tends to 0 as the number of species grows.

All calculations in this manuscript were performed using our new software tool FairShapley, which has been made publicly available at http://www.mareikefischer.de/Software/FairShapley.zip. This tool, which was implemented in Perl, is able to calculate all

* Corresponding author.
E-mail addresses: kristina.wicke@uni-greifswald.de (K. Wicke), email@mareikefischer.de (M. Fischer).

http://dx.doi.org/10.1016/j.jtbi.2017.07.010
0022-5193/

versions of the Shapley Value as well as the Fair Proportion Index as explained in this paper.

2. Preliminaries

Before we can present our results, we need to introduce some notation and definitions. Recall that a phylogenetic tree is a connected, acyclic graph, where the leaves are bijectively labelled by some set X of species, which are also often called taxa. A *rooted* phylogenetic tree is a phylogenetic tree with a designated root node ρ. In biology, *binary* phylogenetic trees are of particular importance. A phylogenetic tree is called *unrooted binary* if all internal nodes have degree 3. It is called *rooted binary* if all internal nodes have degree 3 except for one specified root node ρ of degree 2. Throughout this paper, we always specify whether we are referring to rooted or unrooted trees. When we write T^u, this notation refers to an unrooted phylogenetic tree, whereas T^r always refers to a rooted phylogenetic tree. In both cases, when we refer to the size of a tree, we mean the number $n = |X|$ of taxa, i.e., the number of leaves of the tree under consideration. Note that a rooted tree can also be turned into an unrooted tree by abolishing the designation of a specified root node. In case of binary phylogenetic trees, a rooted tree can be turned into an unrooted tree by suppressing the root node ρ, i.e., by deleting ρ and the two edges adjacent to it and re-connecting the two resulting degree-2 vertices with a new edge. We subsequently elaborate how turning a rooted tree into an unrooted one can change the various diversity indices.

In biodiversity conservation, the *phylogenetic diversity* of a set of species plays an important role. This concept captures how diverse or different a set of species is. Mathematically, this requires the trees under consideration to come with edge lengths (e.g., representing evolutionary time since the last common ancestor or substitution rate). Therefore, we assume all edges in the trees to have positive edge lengths assigned to them, and we denote the length of an edge e as λ_e. Moreover, recall that a rooted tree is called *ultrametric* if the path lengths from all leaves to the root are identical. Here, the path lengths are calculated as the sum of all edge lengths on the path from a leaf to the root. The concept of ultrametric trees is also often referred to as the *molecular clock hypothesis* in biology. Note, however, that throughout this paper we do not assume ultrametricity unless stated otherwise.

We are now in the position to formally define phylogenetic diversity, or *PD* for short.

Definition 1. The phylogenetic diversity (*PD*) of a phylogenetic tree is defined as follows:

1. For a rooted phylogenetic tree T^r with leaf set X, the PD^r of a subset $S \subseteq X$ of taxa is calculated by summing up the edge lengths of the phylogenetic subtree of T^r containing S and the root (i.e., the sum of branch lengths in the smallest spanning tree in T^r containing S and the root). Thus, the *PD* of a single taxon is the length of the path from the root to the leaf representing this taxon.
2. In case of an unrooted phylogenetic tree T^u, the unrooted phylogenetic diversity, PD^u, of a subset $S \subseteq X$ of taxa is defined as the sum of edge lengths in the minimal spanning tree in T^u connecting those leaves. The *PD* of a single taxon is defined as 0.

Note that in an ultrametric tree, all taxa have the same PD^r, and note that if one considers the unrooted version T^u of a rooted tree T^r, the *PD* may decrease due to the different definitions.

Example 1. Consider Fig. 1, which depicts trees T^r and T^u on taxon set $X = \{A, B, C, D\}$. Note that here, T^u is the tree you get by suppressing the root of T^r. Now consider the highlighted subset $S = \{A, B\} \subseteq X$. The phylogenetic diversity of S can be calculated as follows: $PD^r(S) = 1 + 1 + 1 + 1 = 4$, and $PD^u(S) = 1 + 1 = 2$. The difference between the two definitions of diversity can be explained by the path of length 2 connecting S with the root, which is disregarded in the unrooted case.

One more concept we need before we can turn our attention to diversity prioritization indices is the concept of a *ranking*. Here, a ranking r is just an assignment of ranking numbers to the elements of X, where for any pair of taxa $x, y \in X$, x either receives a higher or lower ranking number than y or the ranking numbers of x and y are equal (we then call x and y tied). We say that a function $f : X \to \mathbb{R}$ induces a ranking r_f if the ranking number of x in r_f is smaller than the ranking number of y precisely if $f(x) > f(y)$. If $f(x) = f(y)$ for some $x \neq y$, x and y receive the same ranking number.

Example 2. Let $X = \{A, B, C, D\}$. Let $f(A) = 0.5$, $f(B) = 3$, $f(C) = 0.2$ and $f(D) = 1.5$. Then the induced ranking is $r_f(A, B, C, D) = (3, 1, 4, 2)$. Now let $g(A) = 0.5$, $g(B) = 0.5$, $g(C) = 0.2$ and $g(D) = 1.5$. Then we retrieve the induced ranking $r_g(A, B, C, D) = (2, 2, 4, 1)$, where A and B are tied.

Next, recall that the so-called Manhattan distance d_1 (also known as L_1 distance or l_1 metric) between two vectors $r, s \in \mathbb{R}^n$ is defined as follows:

$$d_1(r, s) = \|r - s\| = \sum_{i=1}^{n} |r(i) - s(i)|.$$

We will later on use the Manhattan distance to measure the difference between two rankings induced by different biodiversity indices. Notice that for comparing rankings, often the so-called Kendall tau distance is used. The Kendall tau distance counts the number of pairwise disagreements between two rankings, but can only deal with total rankings, i.e. rankings without ties. As rankings obtained by different biodiversity indices may include ties, we use the Manhattan distance instead (Comparisons where the Kendall tau distance is used by breaking ties arbitrarily can be found in the supporting information (S1 Text)).

However, since we want to observe the behavior of the different prioritization indices for increasing numbers of taxa, we need to normalize the calculated distances. This is due to the fact that whenever the number of taxa increases, even small differences between two rankings have a higher impact on the distance. So we need to normalize in order to take into account that whenever the number of taxa increases, the maximum possible Manhattan distance increases, too. So we divide exactly by this factor. Thus, we define the *normalized Manhattan distance* $d_1^*(r_1, r_2)$ for two rankings r_1 and r_2 with associated ranking vectors v_{r_1} and v_{r_2} as follows:

$$d_1^*(r_1, r_2) := \frac{d_1(v_{r_1}, v_{r_2})}{\max\limits_{r', s'} d_1(v_{r'}, v_{s'})}.$$

Note that the maximum in the denominator is obtained when $r' = (1, 2, \ldots, n)$ and $s' = (n, n-1, \ldots, 1)$.

Now we are in a position to introduce the biodiversity indices, which we will analyze in the following.

2.1. Various indices for biodiversity conservation

In this section, we will present and analyze some indices for biodiversity conservation, which have recently been discussed in the literature. All of these indices turn out to be related, but as different authors use different definitions of these indices, their results sometimes differ. We will therefore give an overview about the relationships of the various definitions.