

## RESEARCH

## Open Access

# Computing evolutionary distinctiveness indices in large scale analysis

Iain Martyn<sup>1</sup>, Tyler S Kuhn<sup>2</sup>, Arne O Mooers<sup>1\*</sup>, Vincent Moulton<sup>3</sup> and Andreas Spillner<sup>4</sup>

## Abstract

We present optimal linear time algorithms for computing the Shapley values and 'heightened evolutionary distinctiveness' (HED) scores for the set of taxa in a phylogenetic tree. We demonstrate the efficiency of these new algorithms by applying them to a set of 10,000 reasonable 5139-species mammal trees. This is the first time these indices have been computed on such a large taxon and we contrast our finding with an ad-hoc index for mammals, fair proportion (FP), used by the Zoological Society of London's EDGE programme. Our empirical results follow expectations. In particular, the Shapley values are very strongly correlated with the FP scores, but provide a higher weight to the few monotremes that comprise the sister to all other mammals. We also find that the HED score, which measures a species' unique contribution to future subsets as function of the probability that close relatives will go extinct, is very sensitive to the estimated probabilities. When they are low, HED scores are less than FP scores, and approach the simple measure of a species' age. Deviations (like the *Solendron* genus of the West Indies) occur when sister species are both at high risk of extinction and their clade roots deep in the tree. Conversely, when endangered species have higher probabilities of being lost, HED scores can be greater than FP scores and species like the African elephant *Loxodonta africana*, the two solendons and the thumbless bat *Furipterus horrens* can move up the rankings. We suggest that conservation attention be applied to such species that carry genetic responsibility for imperiled close relatives. We also briefly discuss extensions of Shapley values and HED scores that are possible with the algorithms presented here.

## 1 Introduction

A phylogenetic tree is a directed graph that portrays the evolutionary relationships among its leaves. The shape of a phylogenetic tree of species can also be viewed as a measure of the redundant and unique evolutionary information embodied in the species: a species in a large and recently-diverged genus like *Mus* shares much of its evolutionary history with many other species, while the monotypic platypus (*Ornithorhynchus anatinus*) embodies a large amount of mammalian evolutionary information not found elsewhere (as expressed in its peculiar genome [1] and phenotype [2]).

Importantly, many species across the tree of life are at risk of extinction due to human activity. Motivated by the need to objectively prioritize conservation effort in an age of triage [3], the Zoological Society of London (ZSL) is spearheading a conservation campaign that

identifies those species that are at once imperiled with extinction and that are minimally redundant within their taxonomic group (<http://www.edgeofexistence.org>). There are many measures of this evolutionary redundancy [4], but all have the common feature that species with fewer closer relatives are given higher rank. The measure chosen by the ZSL is called 'fair proportion' (FP), and is a weighted sum of the edge lengths along the path from the root of an ultra-metric tree to a leaf, with the weights being 1/number of species that share that edge [5]. FP has the useful property that the sum of the values across the species is the sum of all the edge-lengths of the tree (this sum of edgelengths of a (sub) tree is often called the Phylogenetic Diversity (PD) of the tree [6]). So, across all mammals, the platypus has the highest FP score. Under the EDGE approach, FP scores are then multiplied by the probability of extinction for a species to produce an 'EDGE' score (for Evolutionarily Distinct and Globally Endangered), allowing for a global ranking of species within a higher taxon to help in the allocation of conservation attention.

\* Correspondence: [amooers@sfu.ca](mailto:amooers@sfu.ca)

<sup>1</sup>IRMACS and BioSciences, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6 Canada

Full list of author information is available at the end of the article

Interestingly, FP scores are very highly correlated across simulated trees to the Shapley values [7], which is the expected increase in PD that a focal species brings to unrooted trees representing equiprobable subsets of taxa [8]. This measure was adapted by Steel and colleagues [9] to capture the extra PD a species brings to future unrooted subsets, where subsets are sampled in proportion to their probability of persisting in the future. It is called ‘heightened evolutionary distinctiveness’ (HED) and falls in a family of ‘expected PD’ measures [10]. The HED score correlates with the length of the pendant edge leading to the corresponding leaf of the tree (see below), but also highlights species that will become increasingly distinctive if and when imperiled relatives go extinct.

Collen et al. recently [11] published an updated ranking of mammals based on the FP score, based on 1,000 reasonable 5020-tip mammal trees. The authors suggested that the HED scores would be interesting to compare to the FP scores, but given that the fastest previously known algorithm for computing HED scores [8] has a quadratic run time, they did not compute these scores. Here, we first introduce fast (linear-time) algorithms for computing both Shapley values and HED scores, and then apply and compare these scores with FP scores across an improved sample of 10,000 near-complete (5139 species) mammal trees.

## 2 Methods

Let  $\mathcal{T} = (V, E, \lambda)$  be an unrooted, edge-weighted phylogenetic tree on a set  $X$  with  $n$  taxa. Here  $V$  and  $E$  denote the set of vertices and edges of the tree and  $\lambda$  is a map that assigns to every edge  $e \in E$  a non-negative real number, the *length*  $\lambda(e)$  of this edge. With every edge  $e$  of  $\mathcal{T}$  is associated a *split*  $S_e$  of  $X$ . For any  $x \in X$ , we denote by  $S_e(x)$  that set in  $S_e$  that contains  $x$  and by  $\bar{S}_e(x)$  the other set. In addition, for any subset  $Y \subseteq X$ ,  $PD_{\mathcal{T}}(Y)$  denotes the total length of the smallest subtree of  $\mathcal{T}$  containing the taxa in  $Y$ , also known as the *phylogenetic diversity* of  $Y$  with respect to  $\mathcal{T}$  (see e. g. [6]). In the following we first define the two indices we will focus on in this paper and then present optimal linear time algorithms for computing them.

### 2.1 The Shapley value

In [8] the Shapley value  $\psi_x^{sh}(\mathcal{T})$  of a taxon  $x \in X$  with respect to an unrooted, edge-weighted phylogenetic tree  $\mathcal{T} = (V, E, \lambda)$  is defined as follows:

$$\psi_x^{sh}(\mathcal{T}) = \frac{1}{n!} \sum_{Z \subseteq X, x \in Z} (|Z| - 1)! (|X| - |Z|)! \cdot (PD_{\mathcal{T}}(Z) - PD_{\mathcal{T}}(Z - \{x\}))$$

In [8] it is also shown that the Shapley value of a taxon with respect to a phylogenetic tree  $\mathcal{T}$  is a certain linear combination of the lengths of the edges of  $\mathcal{T}$ . More specifically, fixing any taxon  $x$ , we have:

$$\psi_x^{sh}(\mathcal{T}) = \sum_{e \in E} \frac{|\bar{S}_e(x)|}{|X| |S_e(x)|} \lambda(e)$$

Note that the fact that the coefficients in this linear combination can be computed in polynomial time implies that the Shapley value of  $x$  with respect to  $\mathcal{T}$  can be computed in polynomial time. In fact, an algorithm with run time  $O(n^2)$  is presented in [8]. An implementation of this algorithm is available as part of the Bio::Phylo software package [12].

### 2.2 Heightened evolutionary distinctiveness (HED)

In [9] the index HED was introduced which is defined as follows. Let  $p : X \rightarrow [0, 1]$  be a map that assigns to each  $x \in X$  a real number in the closed interval between 0 and 1. We can interpret  $p(x)$  as the probability that taxon  $x$  will go extinct within a certain amount of time in the future. Then, for any  $x \in X$ , the HED of  $x$  with respect to an unrooted, edge-weighted phylogenetic tree  $\mathcal{T} = (V, E, \lambda)$  is defined as follows:

$$\psi_x^{hed}(\mathcal{T}) = \sum_{Z \subseteq X - \{x\}} \left( \prod_{y \in Z} (1 - p(y)) \prod_{y \in X - (Z \cup \{x\})} p(y) \right) \cdot (PD_{\mathcal{T}}(Z \cup \{x\}) - PD_{\mathcal{T}}(Z))$$

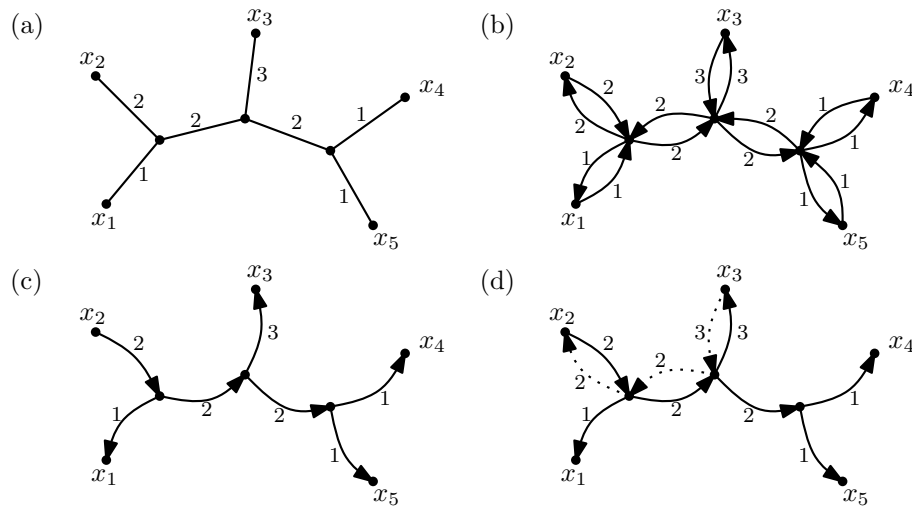
It is shown in [9] that, similarly to the Shapley value, the HED index is a linear combination of the lengths of the edges of  $\mathcal{T}$ , namely:

$$\psi_x^{hed}(\mathcal{T}) = \sum_{e \in E} \left( \prod_{y \in (S_e(x) - \{x\})} p(y) \cdot \left( 1 - \prod_{y \in \bar{S}_e(x)} p(y) \right) \right) \cdot \lambda(e)$$

### 2.3 Linear time algorithms

The basic idea for the design of a linear time algorithm for computing the Shapley values and the HED indices of the taxa with respect to a phylogenetic tree  $\mathcal{T}$  is very similar. It will be convenient to replace, in the given phylogenetic tree  $\mathcal{T}$ , every edge  $e = \{v, w\}$  by a pair of directed arcs  $(v, w)$  and  $(w, v)$  and put  $\lambda(v, w) = \lambda(w, v) = \lambda(e)$  (see Figure 1 for an example).

Let  $A$  denote the set of arcs we have introduced. For every  $x \in X$ , we denote by  $A_x$  the subset of those arcs in  $A$  that are directed away from  $x$  (see Figure 1(c)). Note that, for the Shapley value, the linear combination that yields  $\psi_x^{sh}(\mathcal{T})$  for a taxon  $x \in X$  can be expressed as follows: With each  $a \in A$  is associated the real number



**Figure 1** (a) A phylogenetic tree on  $X = \{x_1, x_2, \dots, x_5\}$ . (b) Replacing each edge by a pair of directed arcs. (c) The arcs in the set  $A_{x_2}$ . (d) The change from  $A_{x_2}$  to  $A_{x_3}$ .

$$\kappa(a) = \frac{|\bar{S}_e(y)|}{|X||S_e(y)|}$$

where  $e$  is the edge that gave rise to arc  $a$  and  $y$  is an arbitrary element in  $X$  such that  $a$  is directed away from  $y$ . Then we have

$$\psi_x^{sh}(\mathcal{T}) = \sum_{a \in A_x} \kappa(a) \cdot \lambda(a).$$

Now, fix an ordering  $x_1, x_2, \dots, x_n$  of the taxa in  $X$  that corresponds to walking around a planar drawing of  $\mathcal{T}$  (see Figure 1(a)). Assume we have already computed the value  $\psi_{x_i}^{sh}(\mathcal{T})$  for some  $i \in \{1, 2, \dots, n-1\}$ . Then it is not hard to see (cf. Figure 1(d)) that we have

$$\psi_{x_{i+1}}^{sh}(\mathcal{T}) = \psi_{x_i}^{sh}(\mathcal{T}) - \left( \sum_{a \in (A_{x_i} - A_{x_{i+1}})} \kappa(a) \cdot \lambda(a) \right) + \left( \sum_{a \in (A_{x_{i+1}} - A_{x_i})} \kappa(a) \cdot \lambda(a) \right),$$

that is, it suffices to consider the arcs that correspond to the edges of  $\mathcal{T}$  that lie on the unique path from  $x_i$  to  $x_{i+1}$ .

Our algorithm for computing the Shapley values for all  $x \in X$  can be summarized as follows:

- (1) Select a suitable ordering  $x_1, x_2, \dots, x_n$  of  $X$ .
- (2) Compute, for each arc  $a \in A$ , the value  $\kappa(a)$ .
- (3) Compute  $\psi_{x_1}^{sh}(\mathcal{T})$ .
- (4) For  $i=1, 2, \dots, n-1$ , compute  $\psi_{x_{i+1}}^{sh}(\mathcal{T})$  from  $\psi_{x_i}^{sh}(\mathcal{T})$ .

It is not hard to see that, if  $\mathcal{T}$  is given in Newick-format [13], steps (1)-(3) have a run time in  $O(n)$ . To establish that also step (4) has a run time in  $O(n)$ , it suffices to note that every edge of  $\mathcal{T}$  is involved in the computation of  $\psi_{x_{i+1}}^{sh}(\mathcal{T})$  from  $\psi_{x_i}^{sh}(\mathcal{T})$  for at most two  $i \in \{1, 2, \dots, n-1\}$ .

Now we turn to the HED index. In analogy to the definition of the values  $\kappa(\cdot)$  above, we put

$$\kappa'(a) = \left( \prod_{y \in S_e(z)} p(y) \right) \cdot \left( 1 - \prod_{y \in \bar{S}_e(z)} p(y) \right)$$

for each  $a \in A$ , where  $e$  is the edge that gave rise to arc  $a$  and  $z$  is an arbitrary element in  $X$  such that  $a$  is directed away from  $z$ . Then, for each  $x \in X$ , we clearly have

$$p(x) \cdot \psi_x^{hed}(\mathcal{T}) = \sum_{a \in A} \kappa'(a) \cdot \lambda(a).$$

So, in a preprocessing step we compute  $\kappa'(a)$  for all  $a \in A$  in linear time. Then we can apply our algorithm above to compute  $p(x) \cdot \psi_x^{hed}(\mathcal{T})$  for each  $x \in X$ , simply replacing the values  $\kappa(a)$  by  $\kappa'(a)$  for all  $a \in A$ . This immediately yields the values  $\psi_x^{hed}(\mathcal{T})$  for those  $x \in X$  with  $p(x) > 0$  in linear time.

It remains to describe how we can deal with those  $x \in X$  with  $p(x) = 0$ . For a subset  $A \subseteq X$ , let  $A^*$  denote the set of those  $x \in A$  with  $p(x) > 0$ . Then we put

$$\tilde{\kappa}'(a) = \begin{cases} \left( \prod_{y \in S_e^*(z)} p(y) \right) & \text{if } |S_e^*(z)| \geq |S_e(z)| - 1 \\ \cdot \left( 1 - \prod_{y \in \bar{S}_e^*(z)} p(y) \right) & \\ 0 & \text{otherwise} \end{cases}$$

for each  $a \in A$ , where, as before,  $e$  is the edge that gave rise to arc  $a$  and  $z$  is an arbitrary element in  $X$  such that  $a$  is directed away from  $z$ . In addition, define

$$\tilde{\psi}_x(\mathcal{T}) = \sum_{a \in A} \tilde{\kappa}'(a) \cdot \lambda(a)$$

for every  $x \in X$ . It is not hard to check that  $\tilde{\psi}_x(\mathcal{T}) = \psi_x^{\text{hed}}(\mathcal{T})$  holds for all  $x \in X$  with  $p(x) = 0$  and, therefore, also these values can be computed in linear time.

### 3 Application

We tested the utility of the new linear time algorithms for the Shapley values and HED scores by applying them to an updated version of the complete mammal tree the ZSL used to generate EDGE scores [11]. We outline the dataset and implementation below.

#### 3.1 Dataset

One issue with producing distinctiveness indices is how tree uncertainty is incorporated [14][11]. The current supertree of mammals is only 50% resolved [15], and the resulting polytomies produce edge lengths that are biased long - in other words, species in polytomies seem older and more distinctive than they should. Collen et al. dealt with this by producing a sample of 1000 trees that each resolve these polytomies via Bayesian methods outlined in Kuhn et al. [15]. To obtain our dataset, we followed this method (given in further detail in the next section) and increased the sample size to 10,000 trees to obtain a better picture of the uncertainty.

A further and more vexing issue with producing evolutionary redundancy indices, especially for large trees, is dealing with taxonomic instability. The current mammal species supertree contains 5020 species [16]. Collen et al. found 396 species from the third and most recent edition of 'Mammal Species of the World' (MSW3) [17] that were not found on this tree (producing an interim taxonomic list of 5416 species). Of these 396 species, however, 75 are well-known to be extinct [18], leaving only 321 problem species.

Collen et al. estimated fair proportion scores for 250 of these remaining species by attributing to each the average of the scores across its relevant genus. For two further species of conservation concern (*Pseudoryx nghentinhensis* and *Laonastes aenigmamus*), molecular estimates of likely time of divergence were used to construct estimates of FP scores. We conducted a literature search and were able to estimate locations for all these 252 species by placing them next to their presumed sister taxon and dividing the edge length in half.

Finally, we needed to deal with taxonomic conflict between the reconciled MSW3 + supertree taxonomy

and the International Union for Conservation (IUCN) taxonomy from which we drew information on imperilment. Collen et al. found that the IUCN lists a further 28 species in the MSW3 taxonomy as extinct, as well as reclassifying 31 as subspecies, and demonstrating that 74 are synonyms for other species already on the supertree. Whereas Collen et al. simply removed all these tips from their final ranking after estimating fair proportion scores, we dropped these taxa from our trees first, leaving a total of 5139 mammal species that reconciles the MSW3 taxonomy with the latest IUCN taxonomic notes, and which leaves 1 percent ((321-252)/5208) mammal species as 'pseudo-extinct'. In passing, we call for a consortium to organize a mammal database to offer a stable single source for taxonomic and biological information for this and other important taxa.

#### 3.2 Implementation

The algorithms were implemented in R v2.12.0 [19] and fashioned so as to read in trees in the Newick format, one of the most common and simplest formats for storing trees [13]. Prior to application the algorithms were tested on different size randomly generated Yule trees and it was confirmed that they ran in linear time (data not shown). All analyses were carried out on a single core of a Quad Core Intel Xeon 2.33 Ghz processor with 16 GB of memory. The R-script is available upon request.

Trees were produced according to the method described in Kuhn et al. [15] with one further modification: uncertainty in individual node ages on the mammal supertree was accommodated by allowing for each node age to be drawn from a prior constrained normal distribution (see also [11]). The mean of each distribution was the best age estimate given by Fritz et al. [16] and each corresponding standard deviation was simply given as (best-worst estimate)/1.96, where the worst estimate is the estimate furthest from the best. The resolved trees were produced using BEAST v.1.6.2 [20]. The set of 10,000 trees represent a combined output from 8 independent runs. For each run trees were recorded every 2,000 steps after a burn-in period of at least 400,000. The burn-in was chosen based on visual examination of log files in Tracer v1.5 [21].

We assigned probability of extinction values to each of the remaining taxa based on the most recent threat level assessment by the IUCN, following the method of Mooers et al. [22]. Briefly, each of the five indicative IUCN categories (critically endangered, endangered, vulnerable, near threatened, and least concern) is assigned a discrete value. However, since there is as of yet no definitive set of values to use, we used two: the 'Isaac' set where the values assigned to categories are 0.4, 0.2, 0.1, 0.05, 0.025 respectively (so, a doubling of risk with

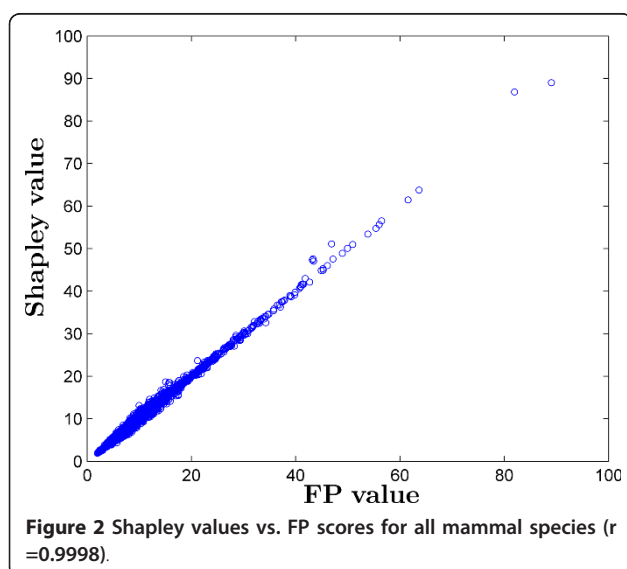
each increase in imperilment rank, and the ‘IUCN 100’ set where values are 0.999, 0.667, 0.1, 0.01, 0.0001 respectively. For the 696 data deficient species that could not be placed into one of the five categories, we assigned the weighted mean of the probability of extinction values of the other species. For the first set this was 0.07 and for the second was 0.11. Thus for the first set we performed two runs, one equating data deficient with near threatened status (probability of extinction = 0.05) and one equating data deficient with vulnerable status (probability of extinction = 0.1). These two analyses were then combined to yield a more realistic final result. For the latter set we could equate data deficient as vulnerable (probability of extinction = 0.1).

## 4 Results and Discussion

### 4.1 FP scores vs Shapley values

In Additional File 1 we report FP scores, Shapley values, and HED scores for all 5139 species, where the scores are the average over the set of 10,000 trees. As we expected, Shapley values and FP scores are very strongly correlated (Figure 2). As proven by Hartmann [7], the Shapley value for a given taxa should approach the FP score as the number of elements in  $X$  tends to infinity. For full details the reader is directed to Hartmann, but this result may be intuited (viewing some interior vertex  $r$  of the given tree  $\mathcal{T}$  as a root) by breaking down the Shapley value as

$$\psi_x^{sh}(\mathcal{T}) = \sum_{e \in s(\mathcal{T}, x, r)} \frac{|\bar{S}_e(x)|}{|X||S_e(x)|} \lambda(e) + \sum_{e \notin s(\mathcal{T}, x, r)} \frac{|\bar{S}_e(x)|}{|X||S_e(x)|} \lambda(e)$$



and noting that the FP score is defined as

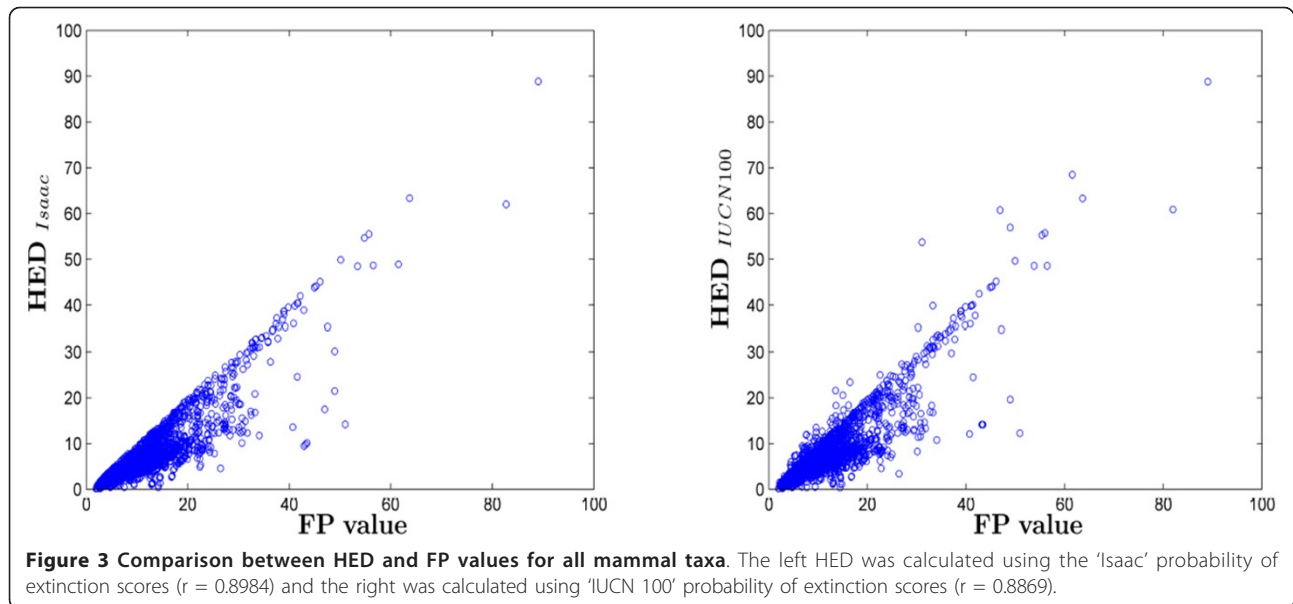
$$\psi_x^{fp}(\mathcal{T}) = \sum_{e \in s(\mathcal{T}, x, r)} \frac{\lambda(e)}{|S_e(x)|}$$

where  $s(\mathcal{T}, x, r)$  denotes the set of edges forming the path in  $\mathcal{T}$  from  $x$  to the root  $r$ . Consider the 5 outliers (*Zaglossus attenboroughi*, *Zaglossus bartoni*, *Zaglossus bruijnii*, *Tachyglossus aculeatus*, and *Ornithorhynchus anatinus*). These taxa form a monophyletic group (Order Monotremata) directly connected to the root of the tree. They thus receive higher Shapley values than FP scores because while the left term in the Shapley value approaches the FP term, the right term, especially for deep connecting branches between this group and the rest of the tree, cannot be approximated as zero. As a result, they gain an additional significant positive contribution to their Shapley value that is not incorporated in their FP value. Another way to intuit this result is to note that the Shapley value is for unrooted trees, and so the additional PD that any monotreme contributes to a possible future subset includes both its distance to the root and the stem age of the placental mammals.

### 4.2 FP scores vs HED scores

Due to the similarity between Shapley values and FP scores, we focussed on comparing HED with FP. As observed in Figure 3, HED generated with both sets of probability of extinction estimates also correlate well with FP ( $r = 0.8984$  for ‘Isaac’,  $r = 0.8869$  for IUCN 100) though there remains significant scatter. We explain this behaviour by noting that HED correlates even more strongly with just the pendant edge of each tip ( $r = 0.9947$ ,  $r = 0.9507$ ), suggesting that the internal tree structure is less relevant. This makes sense, as while for Shapley and FP the contribution for deeper edges can be seen to be approximately proportional to  $1/n$ , for HED the contribution is approximately proportional to  $a^n$ , where most often  $a < 1$ . This of course is a much faster decrease and effectively reduces all edges below the pendant edges to higher order negligible terms. We observe as well that for the ‘Isaac’ set of extinction values it is impossible for these higher order terms to be equal or greater to the equivalent terms for Shapley values and FP scores, even if every taxa on the tree was critically endangered (0.4).

As suggested by Faith [10] this exclusion of deeper edges can oftentimes be justified, as Shapley values and FP scores can overestimate the contribution of an interior branch. For example, an imperiled species belonging to a relatively recent diverging clade with a long interior edge will score highly despite the fact that perhaps its sister taxa are in no danger of extinction. In this case the evolutionary distinctiveness represented by the long



interior edge is in little danger of being lost even if the more endangered species go extinct. This is manifest for *Ornithorhynchus anatinus*, a species we encountered earlier as one of the furthest outliers in Figure 2. The platypus, as its more commonly known, is the oldest member of *Monotremata* and is sister to the Echnidas. While the remaining species are critically endangered, the short-faced echidna *Tachyglossus aculeatus* is listed by the IUCN as least concern. This is enough to, in Faith's words, "secure the interior branch" and thus lower the ranking of any species between *Tachyglossus aculeatus* and the interior branch (here just *Ornithorhynchus anatinus*). Similar situations where a sister species or down-tree species is relatively safe explain the majority of divergence between FP and HED.

Figure 3 also illustrates the effect of changing the probabilities of extinction. When some species are at very high risks of extinction, HED can be greater than FP, because they are expected to contribute large amounts to the tree following extinction of close relatives. This is most intuitively seen with the two remaining elephant species. While it is inconceivable that the Indian elephant *Elephas maximus* would be allowed to go extinct (as it breeds in zoos), the relatively less imperiled African elephant *Loxodonta africana* carries genetic responsibility for its close cousin. The two solendons of the West Indies offer another example. While on remarkably old pendant edges (40 my), they also jointly root very deep in the tree (at 82 mya). Because both are critically endangered, it is not unlikely that one will end up being the sole representative of their (now) shared interior branch.

These examples suggest it may be profitable to take a more dynamic view of how individual species represent

the evolutionary history of their group. The fair proportion metric as used by the EDGE programme is an intuitively compelling measure of evolutionary distinctiveness, and, at the limit, it approaches the well-characterized Shapley value. It may be that considering all future subsets of taxa to be equally likely is a conservative approach to measuring worth (the future is indeed grim for much of biodiversity, and projections based on current imperilment are very imprecise). However, as argued forcefully by Faith [10], it may also make sense to consider future expected PD more explicitly, such that an HED-style metric should be considered. We suggest that, at the very least, some attention be given to species that are relatively cheap to manage (because they are not yet in grave danger, such as the short-faced echidna) and that are also expected to represent large swathes of biodiversity under worst-case scenarios (e.g. if we were to lose all currently-imperiled monotremes).

## 5 Extensions

The fact that the Shapley and HED values are measures of evolutionary distinctiveness on unrooted trees suggests that the above approach to highlighting imperiled and evolutionarily isolated bits of biodiversity could be extended from species on a tree to populations connected via a network on the landscape. Importantly, the algorithms presented here for computing Shapley values and HED scores lend themselves naturally to split networks [23]. The motivation for such an extension comes from the observation that prioritizing populations within species may present policymakers with a useful tool after a species has been legally listed (e.g. through an Endangered Species Act) for conservation management.

Once a species has been awarded protection, and funds are allocated for survival and recovery, an early step in any management plan is to assess how many populations there are, what state each is in, how they are demographically and genetically connected on the landscape, and where genetic diversity lies. As when arguing for a triage approach to species conservation, it may be useful and efficient to highlight those populations of an endangered species that are at once distinctive and that carry genetic responsibility for other populations. Costs and benefits may be easier to compare within than between species, such that objective decisions as to where to invest scarce conservation resources may be more palatable.

## Additional material

**Additional File 1: Complete Scores.** This file contains the average Shapley, HED, and FP scores for all 5139 mammal species across the 10,000 tree distribution, as well the respective standard deviations on this average. It also contains the IUCN threat category for each species at the time of writing.

## 8 Acknowledgements

We thank the Isaac Newton Institute, Cambridge for facilitating this collaboration and NSERC Canada and Simon Fraser University for funding it.

## Author details

<sup>1</sup>IRMACS and BioSciences, Simon Fraser University, 8888 University Drive, Burnaby, V5A 1S6 Canada. <sup>2</sup>Department of Biology, The Pennsylvania State University, 208 Mueller Laboratory, University Park, PA USA 16802. <sup>3</sup>School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ UK. <sup>4</sup>Department of Mathematics and Computer Science, University of Greifswald, Germany.

## Authors' contributions

AS, VM conceived of the algorithm, AS produced the equations, TK created the trees, and IM and AM implemented the algorithm, performed the study, and wrote the first draft of the paper. All authors read and approved the final manuscript.

## Competing interests

The authors declare that they have no competing interests

Received: 13 October 2011 Accepted: 13 April 2012

Published: 13 April 2012

## References

1. Warren W, Hillier L, G J, et al: "Genome analysis of the platypus reveals unique signatures of evolution". *Nature* 2008, **453**:175-183.
2. Redding D, DeWolff C, Mooers A: "Evolutionary distinctiveness, threat status, and ecological oddity in primates". *Conservation Biology* 2010, **24**.
3. Marris E: "Conservation priorities: What to let go". *Nature* 2007, **450(7167)**:152-155.
4. Redding D W, Hartmann K, Mimoto A, Bokal D, DeVos M, Mooers A: "Evolutionarily distinctive species often capture more phylogenetic diversity than expected". *Journal of Theoretical Biology* 2008, **251**:606-615.
5. Redding D: *Incorporating genetic distinctness and reserve occupancy into a conservation prioritisation approach* 2003, Master's thesis, University of East Anglia, Norwich.
6. Faith D: "Conservation evaluation and phylogenetic diversity". *Biological Conservation* 1992, **61**:1-10.

7. Hartmann K: *Biodiversity conservation and evolutionary models* 2008, PhD thesis, University of Canterbury, Christchurch, NZ.
8. Haake C, Kashiwada A, Su F: "The Shapley value of phylogenetic trees". *Journal of Mathematical Biology* 2008, **56**:479-497.
9. Steel M, Mimoto A, Mooers A: "Hedging our bets: the expected contribution of species to future phylogenetic diversity". *Evolutionary Bioinformatics* 2007, **3**:237-244.
10. Faith D: "Threatened species and the potential loss of phylogenetic diversity: conservation scenarios based on estimated extinction probabilities and phylogenetic risk analysis". *Cons Biol* 2008, **22**:1461-1470.
11. Collen B, Turvey S, Waterman C, Meredith H, Kuhn T S, Baillie J, Isaac N: "Investing in evolutionary history: implementing a phylogenetic approach for mammal conservation". *Phil Trans Roy Soc B* 2011, **366**:2611-2622.
12. Vos R, Caravas J, Hartmann K, Jensen M, Miller C: "BIO: Phyloinformatics analysis using perl". *BMC Bioinformatics* 2011, **12**.
13. "The Newick tree format". [http://evolution.genetics.washington.edu/phylip/newicktree.html].
14. Isaac N, Turvey S, Collen B, Waterman C, Baillie J: "Mammals on the edge: Conservation priorities based on threat and phylogeny". *PLoS ONE* 2007, **2**(3).
15. Kuhn T, Mooers A, Thomas G: "A simple polytomy resolver for dated phylogenies". *Methods in Ecology and Evolution* 2011.
16. Fritz S, Olaf O R-E, Purvis A: "Geographical variation in predictors of mammalian extinction risk: big is bad, but only in the tropics". *Ecology Letters* 2009, **12**:538-549.
17. Wilson D, Reeder D: *Mammal species of the world. A taxonomic and geographic reference*. 3 edition. Johns Hopkins University Press; 2008.
18. Mooers A, Goring S, Turvey S, Kuhn T: "Holocene extinctions and the loss of feature diversity". In *Holocene extinctions*. Edited by: Turvey S. Oxford University Press; 2008:279-338.
19. R Development Core Team: *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria; 2008.
20. Drummond A, Rambaut A: "Beast: Bayesian evolutionary analysis by sampling trees". *BMC Evolutionary Biology* 2007, **7**:214.
21. Rambaut A, Drummond A: *Tracer v1.4* 2007.
22. Mooers A, Faith D, Maddison W: "Converting endangered species categories to probabilities of extinction for phylogenetic conservation prioritization". *PLoS ONE* 2008, **3**(11):e3700.
23. Huson D, Bryant D: "Application of phylogenetic networks in evolutionary studies". *Molecular Biology and Evolution* 2006, **23**:254-267.

doi:10.1186/1748-7188-7-6

Cite this article as: Martyn et al.: Computing evolutionary distinctiveness indices in large scale analysis. *Algorithms for Molecular Biology* 2012 **7**:6.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at  
www.biomedcentral.com/submit

