PHYLOGENETIC DIVERSITY AND THE GREEDY ALGORITHM

M Steel

Department of Mathematics and Statistics
University of Canterbury
Private Bag 4800
Christchurch, New Zealand

Report Number: UCDMS2004/14 JUNE 2004

PHYLOGENETIC DIVERSITY AND THE GREEDY ALGORITHM

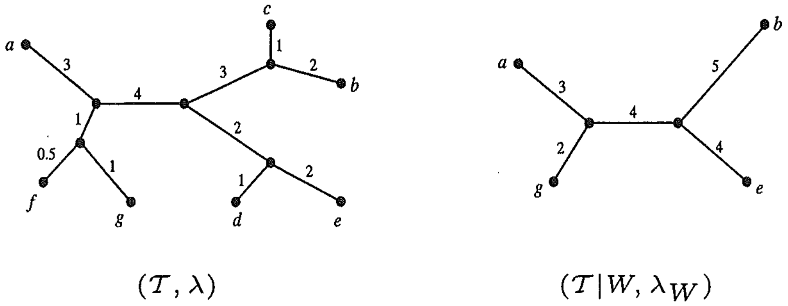
MIKE STEEL

ABSTRACT. Given a phylogenetic tree with leaves labelled by a collection of species, and with weighted edges, the 'phylogenetic diversity' of any subset of the species is the sum of the edge weights of the minimal subtree connecting the species. This measure is relevant in biodiversity conservation where one may wish to compare different subsets of species according to how much evolutionary variation they encompass. In this note we show that phylogenetic diversity has an attractive mathematical property that ensures that we can solve the following problem easily by the greedy algorithm: find a subset of the species of any given size k of maximal phylogenetic diversity. We also describe an extension of this result that also allows weights to be assigned to species.

Date: 3 February 2004.
Key words and phrases. phylogenetic diversity, greedy algorithm, biodiversity conservation.
Thanks to the New Zealand Marsden Fund (UOC310) for supporting this research.

1. Introduction

Let $\mathcal{T}$ be a phylogenetic X–tree, that is, a tree whose leaves comprise the set X of species (or populations) under study, and whose remaining vertices are of degree at least 3. Let λ be a weighting of the edges of $\mathcal{T}$ by positive real numbers (often called 'branch lengths'). Given a subset W of X, we can consider the induced phylogenetic W–tree, denoted $\mathcal{T}|W$ that connects just those species in W and its associated edge weighting λ_W which assigns to each edge e of $\mathcal{T}|W$ the sum of the $\lambda(e)$ values over those edges of $\mathcal{T}$ in the path that corresponds to e. An example is illustrated in Fig. 1; formal definitions of these concepts are provided in [6].



$(\mathcal{T}, \lambda)$

$(\mathcal{T}|W, \lambda_W)$

FIGURE 1. *Left:* An edge-weighted tree. *Right:* The induced edge weighted tree on the subset $W = \{a, b, g, e\}$, which is the (unique) set of four species of largest PD score amongst all sets of 4 species.

The *phylogenetic diversity* of W, denoted $PD(W)$, is defined as

$$PD(W) := \sum_e \lambda_W(e)$$

where the summation is over all edges e in the tree $\mathcal{T}|W$.

For example, in Fig. 1, for $W = \{a, b, g, e\}$ we have $PD(W) = 3+2+4+5+4 = 18$. Note that $PD(W)$ also depends on (T, λ) but we will think of these as fixed. Also, in case $|W| = 1$ we set $PD(W) = 0$. Later we will consider an extension of PD that allows each species to also have a weight assigned to it.

The PD score was formally introduced by Dan Faith in 1992 [2] and has been subsequently applied in areas of biodiversity conservation (see eg. [1] and the references therein). The score provides some indication of how much genetic variation each possible subset W contains in relation to the entire variation in the tree (by comparing $PD(W)$ to $\sum_e \lambda(e)$) and so may therefore be useful for determining which subsets of species might be best to conserve when it is not possible to conserve them all. The PD score has also recently been investigated mathematically by [4] but for a different purpose - namely to extend the classic result on tree reconstruction from pairwise distances to m–wise values.

In this note we show how to efficiently find (and characterise) subsets of X of given size that have maximal PD score. Clearly to examine all subsets of X of given size k is not feasible if k is large, however we show that the sets of any given size that have maximal PD score are precisely the ones that can be built up using a greedy approach starting with the maximal PD sets of size 2. This greedy approach was suggested in Faith's original paper ([2], p.5, column 1, paragraph 2) though no claim (or proof) was made that this was anything more than a locally maximal procedure; as we will see it also leads to a global maximization of PD. Using the greedy algorithm described below, it would be straightforward to find a subset of species of size (say) 100 having maximal phylogenetic diversity for that number of species, in an edge-weighted tree on (say) 1000 species. The algorithm is valid for both binary (i.e. 'fully resolved') phylogenetic trees and non-binary trees containing polytomies.

2. Main result

To state our results we require some further definitions. For $k \geq 1$ let

$$pd_k := \max\{PD(W) : W \subseteq X, |W| = k\}$$

and let

$$PD_k := \{W \subseteq X : |W| = k \text{ and } PD(W) = pd_k\}.$$

Thus pd_k is the largest possible phylogenetic diversity value across all subsets of species of size k, while PD_k is the set of all collections of k species that realize this maximal phylogenetic diversity.

Theorem 2.1. *PD_k consists precisely of those subsets of X of size k that can be built up as follows: Select any pair of species that are maximally far apart (in the edge-weighted tree $(\mathcal{T}, \lambda)$) and then sequentially add elements of X so as to maximize at each step the increase in PD score.*

Example. For the pair $(\mathcal{T}, \lambda)$ in Fig. 1 the greedy algorithm will start by selecting $\{a, b\}$ which is the unique element of PD_2, and then add leaf e and then g to obtain the set $W = \{a, b, e, g\}$ which is the unique element in PD_4. Thus for this example, $pd_4 = 18$.

Proof of Theorem 2.1

The proof relies on first establishing the following fundamental mathematical property of PD. Suppose we are given a pair $(\mathcal{T}, \lambda)$ and subsets W, W' of X with $2 \leq |W'| < |W|$. Then there always exists some species $x \in W - W'$ so that:

$$PD(W - \{x\}) + PD(W' \cup \{x\}) \geq PD(W) + PD(W'). \tag{1}$$

To establish this property, we first introduce some notation: we refer to the edges of a phylogenetic tree as *exterior* if they are incident with a leaf (i.e. a species in X) otherwise we say that the edge is *interior.*

Consider the tree $\mathcal{T}|W$. We can view the tree $\mathcal{T}$ as being obtained from $\mathcal{T}|W$ by attaching a set $\mathcal{F}$ of subtrees of $\mathcal{T}$ to certain vertices and (subdivisions of) edges of $\mathcal{T}|W$. For example, in Fig. 1 $\mathcal{T}$ is obtained from $\mathcal{T}|W$ by attaching the three (one-vertex) subtrees c, d and f to single vertex subdivisions (i.e. the 'midpoints') of the exterior edges of $\mathcal{T}|W$ incident with b, e and g, respectively.

Since $|W'| < |W|$ and $|W| > 2$ there exists at least one exterior edge e of $\mathcal{T}|W$ with the property that any subtree of $\mathcal{F}$ that attaches to e does not contain any species in W'. Let $x \in W - W'$ be the leaf of $\mathcal{T}$ incident with e. We have

$$PD(W) = PD(W - \{x\}) + \lambda_W(e). \tag{2}$$

Now, since e was chosen in $\mathcal{T}|W$ so that any subtree in $\mathcal{F}$ that contains a species in W' must attach either to some edge of $\mathcal{T}|W$ that is different to e, or to a vertex of $\mathcal{T}|W$, we have

$$PD(W' \cup \{x\}) \geq PD(W') + \lambda_W(e). \tag{3}$$

Combining (2) and (3) gives (1).

Now suppose that $W \in PD_k$ and $W' \in PD_{k-1}$. Select a species $x \in W - W'$ to satisfy (1). Then:

$$W' \cup \{x\} \in PD_k \text{ and } W - \{x\} \in PD_{k-1} \tag{4}$$

since $PD(W - \{x\}) \leq PD(W')$ with equality precisely if $W - \{x\} \in PD_{k-1}$ and $PD(W' \cup \{x\}) \leq PD(W)$ with equality precisely if $W \cup \{x\} \in PD_k$, which, combined with (1), gives (4).

Theorem 2.1 now follows easily from (4) by standard arguments from 'greedoid' theory ([3]). Specifically, (4) shows that any element of PD_k (for $k \geq 3$) is obtained from any element of PD_{k-1} by adding a single new element of X, which must by necessity be an element that maximizes the increase in PD score. □

2.1. **An extension.** Suppose we have a function $f : X \to \mathbb{R}$ and we let

$$PD_f(W) := PD(W) + \sum_{x \in W} f(x).$$

For example, in biodiversity conservation a negative score $f(x) < 0$ might be the cost associated with conserving species x; alternatively a positive score $f(x) > 0$ might allow for additional incentives to preserve x (for example if it is globally endangered). Finding a subset W of size k to maximize $PD_f(W)$ is easy for any function f, by applying the same greedy approach (starting by finding a set W of size 2 to maximize $PD_f(W)$, and sequentially add new species so as to maximize the PD_f score). That is, Theorem 2.1 applies if we replace PD by PD_f for any choice of f. This follows from the observation that for any $W, W' \subseteq X$ with $x \in W - W'$ the difference

$$\Delta(f, x) := PD_f(W - \{x\}) + PD_f(W' \cup \{x\}) - (PD_f(W) + PD_f(W'))$$

is independent of f and so, by property (1), x can be chosen so that $\Delta(f, x)$ is non-negative (since $PD = PD_0$ for the zero function $0(x) = 0$ for all $x \in X$).

In particular this extension allows us to determine easily whether or not any given species x in X is in every set (or in any set) of maximal phylogenetic diversity amongst all subsets of X of size k, without having to check them all exhaustively.

To achieve this, let $pd_k(f)$ denote the maximal value of $PD_f(W)$ over all subsets W of X of size k and let us compare pd_k, $pd_k(+\delta_x)$ and $pd_k(-\delta_x)$ where

$$\pm\delta_x(x') = \begin{cases} \pm 1, & \text{if } x = x'; \\ 0, & \text{otherwise.} \end{cases}$$

It is easily checked that $pd_k(+\delta_x) - pd_k = 1$ iff x lies in at least one set in PD_k and $pd_k(-\delta_x) - pd_k = -1$ iff x is in every set in PD_k.

More generally, for any function $f : X \rightarrow \mathbb{R}$ we can also determine easily whether or not any given species x in X is in every (or in any) set of size k of maximal PD_f score by comparing $pd_k(f)$ and $pd_k(f \pm \delta_x)$.

3. Remarks

An alternative representation of phylogenetic diversity is given as follows:

$$PD(W) = \sum_{x,y \subseteq W} \mu_{\mathcal{T}}(x,y) d_{(\mathcal{T},\lambda)}(x,y) \tag{5}$$

where $\mu_{\mathcal{T}}(x,y) = \prod_{v \in p(\mathcal{T}|W,x,y)} (d_{\mathcal{T}|W}(v) - 1)^{-1}$ (here $p(\mathcal{T}|W,x,y)$ is the set of non-leaf vertices of $\mathcal{T}|W$ that lie on the path connecting x and y and $d_{\mathcal{T}|W}(v)$ is the number of edges of $\mathcal{T}|W$ incident with v) and $d_{(\mathcal{T},\lambda)}(x,y)$ is the sum of the edge weights across the path in $\mathcal{T}$ connecting x and y. This follows from a representation of $PD(W)$ when $W = X$ that was described for binary phylogenetic trees by Pauplin [5], and generalized to arbitrary phylogenetic trees in [7]. Equation (5) may be useful in ecological studies where a phylogenetic tree is well-established, but its edge weights are not. The representation (5) separates the topological features of the tree (the term $\mu_{\mathcal{T}}(x,y)$) from the metric properties $d_{(\mathcal{T},\lambda)}$ so the well established topology of the tree suffices to determine the $\mu_{\mathcal{T}}(x,y)$ values. If reasonable estimates of evolutionary distance between pairs of species are known (but not exactly, otherwise the branch lengths could be accurately recovered) then these can be used as estimates of the $d_{(\mathcal{T},\lambda)}(x,y)$ values.

4. Acknowledgements

I thank Gary Barker and Dan Faith for helpful discussion on the biological aspects of phylogenetic diversity during a NZ Landcare workshop (March 2004), and to Andreas Dress for some helpful background on the greedy algorithm.

References

[1] Barker, G. M. (2002). Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biol. J. Linnean Soc.* **76**, 165–194.

[2] Faith, D.P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61**, 1–10.

[3] Korte, B., L. Lovász and R. Schrader. Greedoids, algorithms and combinatorics (Springer Berlin) 1991.

[4] Pachter, L. and Speyer, D. (2004). Reconstructing trees from subtree weights. *Applied Mathematics Letters* (in press).

[5] Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, **51**, 41–47.

[6] Semple, C. and Steel, M. (2003). Phylogenetics. Oxford University Press.

[7] Semple, C. and Steel, M. (2004). Cyclic permutations and evolutionary trees. *Adv. Appl. Math.* **32(4)**, 669–680.

ALLAN WILSON CENTRE FOR MOLECULAR ECOLOGY AND EVOLUTION, BIOMATHEMATICS RESEARCH CENTRE, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

E-mail address: `m.steel@math.canterbury.ac.nz`