Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 246 (2007) 186–194
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Optimizing phylogenetic diversity under constraints

Vincent Moulton[a], Charles Semple[b,*], Mike Steel[b]

[a]School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, UK
[b]Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

Received 17 October 2006; received in revised form 14 December 2006; accepted 15 December 2006
Available online 22 December 2006

Abstract

Phylogenetic diversity (PD) is a measure of the extent to which different subsets of taxa span an evolutionary tree, and provides a quantitative tool for studying biodiversity conservation. Recently, it was shown that the problem of finding subsets of taxa of given size to maximize PD can be efficiently solved by a greedy algorithm. In this paper, we extend this earlier work, beginning with a more explicit description of the underlying combinatorial structure of the problem and its connection to greedoid theory. Next we show that an extension of the PD optimization problem to a phylogeographic setting is NP-hard, although a special case has a polynomial-time solution based on the greedy algorithm. We also show how the greedy algorithm can be used to solve some special cases of the PD optimization problem when the sets that are restricted to are ecologically 'viable'. Finally, we show that three measures related to PD fail to be optimized by a greedy algorithm.

Keywords: Phylogenetic diversity; Greedy algorithm

1. Introduction

A central question in conservation biology is how to measure, predict, and preserve biodiversity as species face extinction. One unifying approach to this question is to measure the 'biodiversity' of a collection of species in terms of the evolutionary diversity that those species span in a 'tree of life'—a measure often referred to as 'phylogenetic diversity' (PD) (Barker, 2002; Faith, 1992; Faith and Baker, 2006; Hartmann and Steel, 2007). Loosely speaking, if $\mathscr{T}$ is a (phylogenetic) tree whose leaf labels comprise a set X of species, and whose edges have non-negative real-valued lengths, then for a subset Y of X, the PD score of Y is the sum of the lengths of the edges of the minimal subtree of $\mathscr{T}$ that connects Y (in the case that $\mathscr{T}$ is rooted, the root vertex must also be connected). Depending on how the edge lengths are assigned, PD can measure either the genetic diversity or the total evolutionary time spanned by the subset of species. PD is also relevant to other problems in bioinformatics such as prioritizing species for sequencing in genomics (Pardi and Goldmann, 2005).

In the PD optimization problem, we wish to find a subset of X of given size (perhaps also containing a given subset of species) that has maximal PD score amongst all such subsets. Although the concept of PD (and a greedy algorithm for constructing high-PD sets) has been around for 14 years (Faith, 1992), it was only in 2005 (Steel, 2005; Pardi and Goldmann, 2005) that the greedy algorithm was formally shown to solve the PD optimization problem. The implication of this is that it is now realistic to solve this optimization problem exactly for thousands of species (Minh et al., 2006).

In this paper, we apply greedoid theory (a branch of combinatorics related to matroid theory) to study optimization problems based on PD, including three variations that are biologically motivated. We first make explicit the underlying role of greedoids in the original PD optimization problem. We then consider the following variations on the problem that include a geographic component: in the first problem we wish to maximize the PD of the species chosen so that at least certain numbers are conserved from

*Corresponding author. Tel.: +64 3 364 2600.
E-mail addresses: vincent.moulton@cmp.uea.ac.uk (V. Moulton), c.semple@math.canterbury.ac.nz (C. Semple), m.steel@math.canterbury.ac.nz (M. Steel).

0022-5193/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.12.021

each of a set of regions. We show that this problem has a polynomial-time algorithm based on the greedy algorithm.

Next we show that the problem (described in Rodrigues et al., 2005) of selecting a given number of regions to maximize the PD of the species that occur within at least one of the selected regions is NP-hard.

In the third variation, we incorporate an obvious ecological constraint: the extinction of a certain set of species will necessarily lead to the extinction of other species (for example, if that species depends on at least one of the species in the set for its survival)—that is, not all subsets of X are (ecologically) 'viable' (this point has been raised by other authors, such as van der Heide et al., 2005; Witting et al., 2000). Consequently it is desirable to restrict the PD optimization problem just to the 'viable' subsets of X. Again there is an underlying greedoid structure to this problem (though in a quite different sense to the standard PD optimization problem) and we describe precisely when the greedy algorithm solves this restricted PD optimization problem.

The final section of the paper considers three functions related to PD, and provides examples to show that for all three the corresponding optimization problem is not solved by the greedy algorithm. We begin by recalling some fundamental concepts from greedoid theory, in particular the concept of a greedoid and the formal definition and properties of the greedy algorithm.

2. Some facts from greedoid theory

Let X be a finite set and let $\mathscr{F}$ be a collection of subsets of X. The pair $(X,\mathscr{F})$ is a *greedoid* if it satisfies the following two conditions:

(G1) If $F \in \mathscr{F}$ and $F \neq \emptyset$, then there is an $x \in F$ such that $F - \{x\} \in \mathscr{F}$.

(G2) If $F_1, F_2 \in \mathscr{F}$ and $|F_2| = |F_1| + 1$, then there is an $x \in F_2 - F_1$ such that $F_1 \cup \{x\} \in \mathscr{F}$.

Note that a consequence of (G1) is that $\emptyset \in \mathscr{F}$.

For a greedoid $(X,\mathscr{F})$, the members of $\mathscr{F}$ are called *feasible*. Furthermore, the maximal feasible sets, that is, the feasible sets not properly contained in any other feasible set, are called *bases*. Observe that, because of (G2), all bases have the same size. As an explicit example, let G be a connected graph with edge set E. Let $\mathscr{F}$ denote the collection of subsets of E that contain no cycle, that is, induce a forest. Then $\mathscr{F}$ certainly satisfies (G1) and one can show that $\mathscr{F}$ satisfies (G2), and so $(E,\mathscr{F})$ is a greedoid. As G is connected, the bases of this greedoid are the edge sets of spanning trees of G.

Condition (G1) implies that $(X,\mathscr{F})$ is an 'accessible' set system. The implication of this is that if F is a feasible set, then it can be obtained from the empty set by sequentially adding elements of X such that at each stage the set so far constructed is feasible. In particular, there is a sequence of feasible sets $\emptyset = F_0, F_1, F_2, \ldots, F_k = F$ such that, for all i, we have that $F_{i-1} \subseteq F_i$ and $|F_i| = |F_{i-1}| + 1$.

Now consider the following condition, a strengthening of (G2):

(G2′) If $F_1, F_2 \in \mathscr{F}$ and $|F_2| = |F_1| + 1$, then there is an $x \in F_2 - F_1$ such that $F_1 \cup \{x\} \in \mathscr{F}$ and $F_2 - \{x\} \in \mathscr{F}$.

If $(X,\mathscr{F})$ satisfies (G1), (G2′), then $(X,\mathscr{F})$ is called a *strong* greedoid or, equivalently, a *Gauss* greedoid. Bryant and Brooksbank, and Goecke derive a number of properties of this type of greedoid in (Bryant and Brooksbank, 1992) and (Goecke, 1988), respectively.

One natural way to obtain one greedoid from another is stated in the following well-known proposition (for example, see Björner and Ziegler, 1992).

Proposition 2.1. *Let $(X,\mathscr{F})$ be a greedoid (resp. a strong greedoid), k a non-negative integer, and $\mathscr{F}^{(k)}$ denote the subset of $\mathscr{F}$ containing all feasible sets with at most k elements. Then $(X,\mathscr{F}^{(k)})$ is a greedoid (resp. a strong greedoid).*

The original motivation for greedoids was to provide a unified approach to various greedy algorithms that can be successfully applied to optimization problems. Generically, these algorithms work by sequentially selecting objects of maximum weight with no backtracking. Their simplicity is highlighted by the fact that the sole criteria for each selection is the weight of the objects—the available object with the biggest weight is the one that is selected. In this section, we formally describe the greedy algorithm and give one direction of an algorithmic characterization of greedoids (see Theorem 2.2). This characterization in terms of the greedy algorithm justifies the original motivation.

To aid the reader in what follows, consider the explicit example above. Suppose that the edges of G are weighted with (assigned) positive real numbers. One natural problem is to find a spanning tree of G whose sum of edge weights is maximized. In terms of the greedy algorithm below, $\mathscr{F}$ is the collection of subsets of E that contain no cycle and the objective function f is the function on $\mathscr{F}$ that assigns $f(F)$ to be the sum of the weights of the edges in F for all $F \in \mathscr{F}$.

Formally, the greedy algorithm is stated as follows.

Algorithm. GREEDY
Input: A collection $\mathscr{F}$ of subsets of a set X, and an objective function $f : \mathscr{F} \to \mathbb{R}$.
Output: A member of $\mathscr{F}$.
1. Set $F_0 = \emptyset$ and $i = 0$.
2. Given F_i, choose an element x in $X - F_i$ such that
   (i) $F_i \cup \{x\} \in \mathscr{F}$ and
   (ii) $f(F_i \cup \{x\}) \geqslant f(F_i \cup \{y\})$ for all $y \in X - F_i$ with $F_i \cup \{y\} \in \mathscr{F}$.
3. Set $F_{i+1} = F_i \cup \{x\}$.

4. If F_{i+1} is not a maximal member of $\mathscr{F}$, set $i = i+1$ and go to Step 2; otherwise output F_{i+1}.

To state Theorem 2.2, we need one further definition. Suppose that $(X, \mathscr{F})$ is a greedoid and let $f : \mathscr{F} \to \mathbb{R}$ be an objective function on $\mathscr{F}$. We say that f is *compatible* with $(X, \mathscr{F})$ if the following property holds:

- Let $F \subseteq G$ and $x \in X - G$, and assume that

  $F, G, F \cup \{x\}, G \cup \{x\} \in \mathscr{F}$.

  Then, for all $y \in X - F$ with $F \cup \{y\} \in \mathscr{F}$ such that $f(F \cup \{x\}) \geqslant f(F \cup \{y\})$, we have

  $f(G \cup \{x\}) \geqslant f(G \cup \{z\})$

  for all $z \in X - G$ with $G \cup \{z\} \in \mathscr{F}$.

Informally, this property says that if x is the current best choice, then it is also the best choice at any latter stage. An example of an objective function that is compatible with every greedoid $(X, \mathscr{F})$ is the cardinality function, which is the objective function f on $\mathscr{F}$ that is defined by setting $f(F) = |F|$ for all $F \in \mathscr{F}$.

The following theorem gives one direction of the characterization of greedoids in terms of Greedy (see Björner and Ziegler, 1992, Theorem 8.5.2; Korte et al., 1991, Theorem 1.3, p. 155).

Theorem 2.2. *Let $(X, \mathscr{F})$ be a greedoid and let $f : \mathscr{F} \to \mathbb{R}$ be an objective function that is compatible with $(X, \mathscr{F})$. Then* Greedy *is applied to $(X, \mathscr{F})$ and f outputs a basis of $(X, \mathscr{F})$ of maximum weight.*

An important and well-known observation to note is the following. Let $(X, \mathscr{F})$ be a greedoid and let $f : \mathscr{F} \to \mathbb{R}$ be an objective function that is compatible with $\mathscr{F}$. Then, by Theorem 2.2, Greedy applied to $(X, \mathscr{F})$ and f finds a basis, F say, of $(X, \mathscr{F})$ of maximum weight. To find F, the algorithm finds a nested sequence of feasible sets

$$\emptyset = F_0 \subset F_1 \subset F_2 \subset \cdots \subset F_r = F,$$

where $|F_i| = |F_{i-1}| + 1$ for all $i \in \{1, 2, \ldots, r\}$. While $F_r = F$ is a feasible set of size r of maximum weight, it also turns out that, for all i, the set F_i is a maximum weight feasible set of size i. To see this, let $k \in \{0, 1, 2, \ldots, r\}$. Then, by Proposition 2.1, $(X, \mathscr{F}^{(k)})$ is a greedoid. Let $f_k : \mathscr{F}^{(k)} \to \mathbb{R}$ denote the function that is obtained from f by setting $f_k(F) = f(F)$. Since f is compatible with $(X, \mathscr{F})$, f_k is compatible with $(X, \mathscr{F}^{(k)})$ and so Greedy applied to $(X, \mathscr{F}^{(k)})$ and f_k finds a basis of maximum weight. But this basis is also a maximum weight feasible set of size k of $(X, \mathscr{F})$. By considering Greedy is applied to $(X, \mathscr{F})$ and f, the desired outcome follows routinely.

The following result will also be useful in what follows. It provides a slight strengthening of part of Theorem 4 in Bryant and Brooksbank (1992).

Lemma 2.3. *Let $(X, \mathscr{F})$ be a strong greedoid and let $f : \mathscr{F} \to \mathbb{R}$ be a function on $\mathscr{F}$. Suppose that, for all $F_1, F_2 \in \mathscr{F}$ with $|F_2| = |F_1| + 1$ and $x \in F_2 - F_1$ such that $F_1 \cup \{x\}, F_2 - \{x\} \in \mathscr{F}$, we have*

$$f(F_1 \cup \{x\}) + f(F_2 - \{x\}) \geqslant f(F_1) + f(F_2). \tag{1}$$

Then

$$\mathscr{F}^* = \{F \in \mathscr{F} : f(F) = \max\{f(F') : F' \in \mathscr{F}, |F'| = |F|\}\}$$

is the collection of feasible sets of a strong greedoid on X (with respect to f). In particular, $(X, \mathscr{F}^*)$ is a strong greedoid.

Before proving Lemma 2.3, we note that the set $\mathscr{F}^*$ is the subset of $\mathscr{F}$ consisting of all feasible sets of each size of maximum weight.

Proof of Lemma 2.3. To verify condition (G1), let $F \in \mathscr{F}^*$. Since $\mathscr{F}$ satisfies (G1) and $F \in \mathscr{F}$, there is an element $y \in F$ such that $F - \{y\} \in \mathscr{F}$. Consequently, there is some element $F' \in \mathscr{F}^*$ of cardinality $|F| - 1$. As $(X, \mathscr{F})$ is a strong greedoid, there is an element $x \in F - F'$ such that $F' \cup \{x\}, F - \{x\} \in \mathscr{F}$. Applying inequality (1) with $F_1 = F'$ and $F_2 = F$, we have

$$f(F' \cup \{x\}) + f(F - \{x\}) \geqslant f(F') + f(F). \tag{2}$$

Since $F, F' \in \mathscr{F}^*$, $|F' \cup \{x\}| = |F|$, and $|F - \{x\}| = |F'|$, it follows that $f(F' \cup \{x\}) \leqslant f(F)$ and $f(F - \{x\}) \leqslant f(F')$. By considering (2), we deduce that $f(F - \{x\}) = f(F')$, and so there is an element in F, namely x, such that $F - x \in \mathscr{F}^*$. It follows that $\mathscr{F}^*$ satisfies (G1).

To show that $\mathscr{F}^*$ satisfies (G2′), let $F_1, F_2 \in \mathscr{F}^*$ with $|F_2| = |F_1| + 1$. Since $\mathscr{F}$ satisfies (G2′), there exists some element $x \in F_2 - F_1$ such that $F_1 \cup \{x\}, F_2 - \{x\} \in \mathscr{F}$. Furthermore, by hypothesis,

$$f(F_1 \cup \{x\}) + f(F_2 - \{x\}) \geqslant f(F_1) + f(F_2).$$

Since $|F_2 - \{x\}| = |F_1|$ and $|F_1 \cup \{x\}| = |F_2|$ and since $F_1, F_2 \in \mathscr{F}^*$, it follows that $f(F_1 \cup \{x\}) \leqslant f(F_2)$ and $f(F_2 - \{x\}) \leqslant f(F_1)$. Combining these last two inequalities with the previous inequality gives $f(F_1 \cup \{x\}) = f(F_2), f(F_2 - \{x\}) = f(F_1)$, and so $F_1 \cup \{x\}, F_2 - \{x\} \in \mathscr{F}^*$. Hence $\mathscr{F}^*$ satisfies (G2′). We conclude that $(X, \mathscr{F}^*)$ is a strong greedoid. □

Remark. Let $(X, \mathscr{F})$ be a strong greedoid and let $f : \mathscr{F} \to \mathbb{R}$ be an objective function on $\mathscr{F}$ satisfying the property of its namesake in the statement of Lemma 2.3. A consequence of Lemma 2.3 is that if F is a feasible set of size k that maximizes f over all feasible sets of size k, then it is possible for Greedy when applied to $\mathscr{F}$ and f to construct a nested sequence that includes F. To see this, observe that, as $(X, \mathscr{F}^*)$ is a greedoid on X, there is a nested sequence of feasible sets

$$\emptyset = F_0 \subset F_1 \subset F_2 \subset \cdots \subset F_k = F$$

such that, for each i, we have $|F_i| = |F_{i-1}| + 1$ and F_i maximizes f over all subsets of $\mathscr{F}$ of size i. Now consider applying Greedy to $\mathscr{F}$ and f. Beginning with F_0 at Step 1, we can subsequently choose F_1 at the first iteration, F_2 at

the second iteration, and so on. Eventually, GREEDY chooses $F_k = F$ at the kth iteration.

3. Phylogenetic diversity

In this section the notation and terminology follows Steel (2005) (also see Semple and Steel, 2003). A *phylogenetic X-tree* is a tree with no degree-2 vertices and whose leaf set is X. Let $\mathcal{T}$ be a phylogenetic X-tree with edge set E and let $\lambda : E \to \mathbb{R}^{\geqslant 0}$ be an assignment of lengths (weights) to the edges of $\mathcal{T}$. For example, ignoring the dashed edges, the tree shown in Fig. 1 is a phylogenetic tree whose edges are weighted with non-negative real numbers. For a subset S of X, the PD score of S, denoted by $PD_{(\mathcal{T},\lambda)}(S)$, is the sum of the edge lengths of the minimal subtree of $\mathcal{T}$ that connects S. If there is no ambiguity, we frequently denote $PD_{(\mathcal{T},\lambda)}(S)$ by $PD(S)$. Referring to Fig. 1, if $S = \{a, b, f\}$, then $PD(S)$ is equal to the sum of the weights of the minimal subtree (dashed edges) that connects a, b, and f. In particular, $PD(S) = 12$.

Following Pardi and Goldmann (2005), it is also useful to consider an extended version of PD by restricting attention to those subsets of X that contain a fixed non-empty subset W of X. For example, if we take W to be a singleton, $\{z\}$ say, we may regard z as providing a root for the tree $\mathcal{T}$ (in which case, if the edge incident with leaf z is assigned weight 0, the concept of rooted PD coincides with that used in biology). For a fixed subset W of X, let $PD_{W,k}$ denote the maximum PD score over all subsets S of X of size k that contain W. Let $\mathcal{F}_W$ be the collection of all subsets F of $X - W$ that have the property that $PD(W \cup F) = PD_{W,|W|+|F|}$. In other words, $\mathcal{F}_W$ is the collection of subsets of $X - W$ that together with W maximize the PD score for their cardinality under the restriction that they contain W. Within this setting, the standard PD problem can be formally stated as follows.

Problem. OPTIMIZING DIVERSITY.
Instance. A phylogenetic X-tree $\mathcal{T}$, a weighting $\lambda : E \to \mathbb{R}^{\geqslant 0}$ of the edge set of $\mathcal{T}$, a subset W of X, and a positive integer k.
Question. Find a subset X' of X of size k that contains W and maximizes the PD score amongst all such subsets of X.

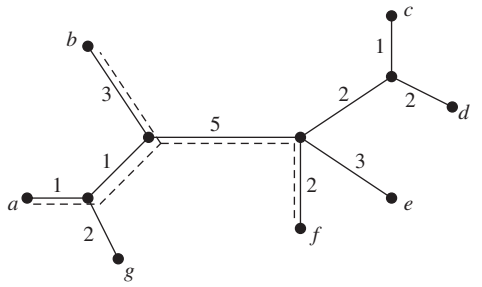

Fig. 1. A phylogenetic X-tree with edge lengths, where $X = \{a, b, c, d, e, f, g\}$.

It is shown in Steel (2005) and Pardi and Goldmann (2005) that the greedy algorithm can be used to solve OPTIMIZING DIVERSITY in polynomial time. Essentially, this is done by applying GREEDY to the collection of all subsets of X and the PD function on this collection. In particular, provided W is non-empty, we begin with W instead of the empty set in Step 1 of GREEDY and add elements sequentially as in Steps 2 and 3. Once the set that contains W has size k, we stop; this set maximizes the PD score over all subsets of X of size k that contain W. Interestingly, if W is empty and $k \geqslant 2$, one proceeds in the same way but begins with an initial subset of size two that maximizes the PD score over all subsets of size two.

In the rest of this section, we put OPTIMIZING DIVERSITY in the setting of greedoids and show that all optimal solutions can be obtained via GREEDY. To achieve this aim, we begin with a lemma, which we will use to show that $(X, \mathcal{F}_W)$ is a (strong) greedoid.

Lemma 3.1. *Let $\mathcal{T}$ be a phylogenetic X-tree and let $\lambda : E \to \mathbb{R}^{\geqslant 0}$ be a weighting of the edge set E of $\mathcal{T}$. Let U and V be subsets of X both containing at least one common element with $1 \leqslant |U| < |V|$. Then there is an element $x \in V - U$ such that*

$$PD(V - \{x\}) + PD(U \cup \{x\}) \geqslant PD(V) + PD(U).$$

Proof. The case when $|U| \geqslant 2$ is established at the beginning of the proof of Theorem 1 in Steel (2005). Therefore assume that $U = \{z\}$ and let x be any element in $V - U$. Then, as $PD(U) = 0$ and both U and V contain z, it is easily seen that

$$PD(V) + PD(U) \leqslant PD(V - \{x\}) + P(U \cup \{x\}).$$

This completes the proof of the lemma. □

Theorem 3.2. *Let $\mathcal{T}$ be a phylogenetic X-tree and let $\lambda : E \to \mathbb{R}^{\geqslant 0}$ be a weighting of the edge set E of $\mathcal{T}$. Let W be a fixed non-empty subset of X. Then $(X, \mathcal{F}_W)$ is a strong greedoid.*

Proof. Define $f : 2^{X-W} \to \mathbb{R}$ by setting $f(F) = PD(W \cup F)$ for all $F \in 2^{X-W}$. Let F_1, F_2 be subsets of 2^{X-W} with $|F_2| = |F_1| + 1$. Since W is non-empty, it follows by Lemma 3.1 that there is an element x in $(F_2 \cup W) - (F_1 \cup W) = F_2 - F_1$ such that

$$f(F_2 - \{x\}) + f(F_1 \cup \{x\}) \geqslant f(F_2) + f(F_1).$$

Observing that the pair $(X, 2^{X-W})$ is trivially a strong greedoid, it now follows by Lemma 2.3 that $(X, \mathcal{F}_W)$ is a strong greedoid. □

The important consequence of Theorem 3.2 for us is that $(X, \mathcal{F}_W)$ is a greedoid. It now follows from the remarks after Lemma 2.3 that, for all W and k, every optimal solution of OPTIMIZING DIVERSITY can be chosen by GREEDY. Indeed, by considering all possible choices at Step 2 of GREEDY, a straightforward modification of this algorithm can produce such a set of solutions.

4. Optimizing Diversity with Coverage

In this section we consider the following problem. Suppose that each species in our set X possesses one of several possible (discrete state) attributes. For example, we may have a collection of geographic regions, and we record for each species the region(s) it is found in (each region may contain several species, and each species can be present in several regions). We may wish to select a subset of species of maximal PD, so that we select at least one (or more generally some positive number) from each region. We will show how this problem can be solved by a polynomial-time approach based on the greedy algorithms, provided that (i) the regions are chosen sufficiently large so that each species is present in just one location and (ii) species of given attribute are compatible with the tree (i.e. they divide the tree up into non-overlapping subtrees). First we formalize the problem.

Consider a phylogenetic X-tree $\mathscr{T}$ and λ a weighting of the edges by non-negative real numbers. Let A be a finite set, and let $f : X \to 2^A$ be a function. For $a \in A$, let

$$X_a = \{x \in X : a \in f(x)\}, \tag{3}$$

that is, X_a is the set of species with attribute a. Now suppose we wish to sample k species from X, so that for each $a \in A$ at least $n_a \geqslant 1$ species are selected from X_a in order to maximize the PD score of the set over all such selections. In other words, we have the following problem.

Problem. Optimizing Diversity with Coverage.
Instance. A pair $(\mathscr{T}, \lambda)$, $f : X \to 2^A$, positive integer k, and a positive integer n_a for each $a \in A$.
Question. Find a subset X' of X of maximum PD score amongst all subsets of X of size k that satisfy the constraint that, for each $a \in A$, at least n_a species in X_a are included in X'.

Let $\mathscr{T}(X_a)$ denote the minimal subtree of $\mathscr{T}$ that connects the leaves in X_a (note that this tree may have vertices of degree 2). Following Semple and Steel (2003), we say that f is *convex on* $\mathscr{T}$ if the collection $\{\mathscr{T}(X_a) : a \in A\}$ of subtrees is vertex disjoint. Intuitively, f is convex on $\mathscr{T}$ if there is a subset of edges whose deletion results in a graph such that, for all $a \in A$, the elements of X having attribute a are in exactly one component of the graph, and no component has elements of X with different attributes. Furthermore, we say that f is *atomic* if $|f(x)| = 1$ for all $x \in X$, that is, x has precisely one attribute.

Lemma 4.1. *Let $\mathbb{N} = \{1, 2, 3, \ldots\}$ and let $m \in \mathbb{N}$. For each $i \in \{1, 2, \ldots, m\}$, let $n_i \in \mathbb{N}$ and let $f_i : \mathbb{N} \to \mathbb{R}$ be an increasing function. Let $k \in \mathbb{N}$ with $k \geqslant \sum_{i=1}^{m} n_i$. Then, with respect to m and k, the following problem can be solved in polynomial time: construct an m-tuple $(x_1, x_2, \ldots, x_m) \in \mathbb{N}^m$ that maximizes $\sum_{i=1}^{m} f_i(x_i)$ subject to the constraints*

(i) $\sum_{i=1}^{m} x_i = k$, *and*
(ii) *for all* $i \in \{1, \ldots, m\}$, $x_i \geqslant n_i$.

Proof. For all i between 1 and m and all j between $\sum_{l=1}^{i} n_l$ and k, let $M(i,j)$ denote a sequence $(m_1, m_2, \ldots, m_i) \in \mathbb{N}^i$ that maximizes $\sum_{l=1}^{i} f_l(m_l)$ subject to the constraints $m_l \geqslant n_l$ for all $l \in \{1, 2, \ldots, i\}$ and $\sum_{l=1}^{i} m_l = j$. Let $m(i,j)$ denote the corresponding value of $\sum_{l=1}^{i} f_l(m_l)$.

The algorithm for solving the desired problem is inductive. For $i = 1$, we have $M(1,j) = (j)$ for each $j \geqslant n_1$. To construct a valid choice for $M(i+1,j)$ for all values of j between $\sum_{l=1}^{i+1} n_l$ and k from the sequences $M(i,j')$ (for j' between $\sum_{l=1}^{i} n_l$ and k) observe that

$$m(i+1,j) = \max\left\{ m(i,r) + f_{i+1}(s) : r + s = j, r \geqslant \sum_{l=1}^{i} n_l, s \geqslant n_{i+1} \right\}. \tag{4}$$

Thus in $O(j)$ steps we can find a pair r and s to maximize the expression on the right-hand side of (4) and we can then extend the sequence $M(i,r)$ to a sequence $M(i+1,j)$ by appending s as the $(i+1)$th coordinate. Continuing in this way constructs a desired sequence $M(m,k)$. □

Theorem 4.2. *If f is atomic and convex on $\mathscr{T}$, then* Optimizing Diversity with Coverage *can be solved in polynomial time by a method based on the greedy algorithm.*

Proof. First note that, as f is atomic, we may assume that $k \geqslant \sum_{a \in A} n_a$. Let

$$E_0 = E(\mathscr{T}) - \bigcup_{a \in A} E(\mathscr{T}(X_a))$$

and let E_1 denote the subset of E_0 containing those edges with at least one end vertex in

$$\bigcup_{a \in A} V(\mathscr{T}(X_a)).$$

Let $\Lambda_f = \sum_{e \in E_0} \lambda(e)$. For each $a \in A$, let $\mathscr{T}_a$ denote the tree that is obtained from $\mathscr{T}(X_a)$ by adjoining a new leaf (via a new edge) to each vertex v of $\mathscr{T}(X_a)$ that is an end vertex of an edge in E_1 in $\mathscr{T}$. Note that v may be a degree-2 vertex of $\mathscr{T}(X_a)$. For each $a \in A$, let W_a denote the resulting set of new leaves and observe that $\mathscr{T}_a$ is a phylogenetic tree with leaf set $X_a \cup W_a$. Now assign each edge incident with a leaf in W_a weight 0, thereby extending the restriction of λ to $\mathscr{T}(X_a)$ to an edge weighting λ_a of $\mathscr{T}_a$. For each positive integer j between n_a and $|W_a| + |X_a|$, let

$$f_a(j) = \max\{PD_{(\mathscr{T}_a, \lambda_a)}(Y) : W_a \subseteq Y \subseteq X_a \cup W_a, |Y| = j\},$$

and let $Y_a(j)$ denote any set Y that realizes this maximum. It follows from Theorems 2.2 and 3.2 and the comments after Theorem 2.2 that the sequence $f_a(j)$ and a set $Y_a(j)$

can be computed by the greedy algorithm for $j = n_a, n_a + 1, \ldots, |W_a| + |X_a|$.

Now, for every subset X' of X that satisfies the condition $|X' \cap X_a| \geqslant 1$ for all a, we have

$$PD_{(\mathscr{T},\lambda)}(X') = \Lambda_f + \sum_{a \in A} PD_{(\mathscr{T}_a,\lambda_a)}((X' \cap X_a) \cup W_a).$$

Consequently, a set X' that maximizes this last quantity and is subject to the constraints $|X' \cap X_a| \geqslant n_a \geqslant 1$ and $|X'| = k$ is the disjoint union

$$\dot{\bigcup_{a \in A}} (X' \cap Y_a(j_a)),$$

where the sequence $(j_a, a \in A)$ is chosen to maximize the expression

$$\sum_{a \in A} f_a(j_a)$$

subject to the constraints $j_a \geqslant n_a + |W_a|$ and $\sum_{a \in A} j_a = k + \sum_{a \in A} |W_a|$. The construction of the sets $Y_a(j_a)$ can now be carried out by applying Lemma 4.1. □

Theorem 4.2 shows that Optimizing Diversity with Coverage can be efficiently solved in the special case where f is both atomic and convex on $\mathscr{T}$. Requiring both these conditions is clearly a strong assumption, and it would be of interest to investigate the complexity of this problem under weaker constraints.

5. Optimizing Diversity via Regions

In this section we consider a variation on the phylogenetic coverage problem, discussed in Rodrigues et al. (2005). The motivation for the problem is as follows. Suppose we have various regions (for example, nature reserves) each of which contains a subset of species. We can conserve each region at some cost, and we wish to select certain regions to conserve so as to (i) keep within the allowed budget and (ii) maximize the PD score of the species that are 'safe' (i.e. present within at least one conserved region).

Here the set-up is similar to Optimizing Diversity with Coverage, but the question is different. Let $\mathscr{T}$ be a phylogenetic X-tree with positive edge weighting λ, let A be a set of regions, each containing some subset of X, and let $f : X \to 2^A$ be the function defined by setting $f(x)$ to be the set of regions that contain x, for each $x \in X$. Given a non-negative integer k, the problem is to find a subset A' of A of size k that maximizes the PD score of those species that are contained in at least one region in A' amongst all such choices of A' of size k. Formally, the problem can be stated as follows (recall the definition of X_a from (3)).

Problem. Optimizing Diversity via Regions.
Instance. A phylogenetic X-tree $\mathscr{T}$, a positive weighting λ on the edges of $\mathscr{T}$, a set of regions, a function $f : X \to 2^A$, and a positive integer k.
Question. Find a subset A' of A that maximizes the PD score of $\bigcup_{a \in A'} X_a$ over all subsets of A of size k.

A more general version of the problem would be to have an additional cost function $c : A \to \mathbb{R}^{>0}$ and a budget B and so the choice of A' is also subject to the constraint $\sum_{a \in A'} c(a) \leqslant B$. However, Optimizing Diversity via Regions is computationally hard.

Theorem 5.1. Optimizing Diversity via Regions *is* NP-*hard*.

Proof. To establish the theorem, we use a polynomial-time reduction from the NP-complete problem SET COVER Garey and Johnson (1979). In this latter problem, one is given a collection C of subsets of a finite set X and a positive integer k'. The question is whether there exists a subset of C of size k' whose union is X. Given an instance of this problem, we construct an instance of Optimizing Diversity via Regions as follows. Take the phylogenetic X-tree $\mathscr{T}$ having exactly one interior vertex and assign weight 1 to each edge of $\mathscr{T}$. Let $f : X \to 2^C$ be the function that is defined by setting $f(x)$ to be the collection of sets in C that contain x. Here C corresponds to the set A in the problem Optimizing Diversity via Regions. For a subset C' of C, the union of the members of C' is a subset of X and, referring to $\mathscr{T}$, its PD score is the cardinality of that set (provided it has size at least two) since the pendant edges of $\mathscr{T}$ connecting these elements of X all have weight 1.

If we could solve Optimizing Diversity via Regions in polynomial time, then we could decide in polynomial time whether or not the maximum PD score amongst all subsets of C of size k' was $|X|$ or not. This is precisely the condition that C contains a subset of size k' whose union is X, and thus we would obtain a solution to the given instance of Set Cover. Since the above reduction is clearly polynomial time in the size of the input, it follows that Optimizing Diversity via Regions is NP-hard. □

Theorem 5.1 implies there is no polynomial-time algorithm for solving Optimizing Diversity via Regions (unless P = NP). However, if some constraints are placed on the function f it may be possible to efficiently solve certain instances of this problem. Of particular relevance in biodiversity conservation would be mild constraints on f that permit an efficient algorithm for the more general problem in which costs are assigned to elements of A (mentioned just before Theorem 5.1).

6. Optimizing Diversity with Dependencies between species

In this section we consider a complication that arises in maximizing PD in real ecosystems. Namely, often species depend on other species for their survival; that is, only certain sets of taxa are 'viable' and selecting sets to

maximize PD should respect this constraint (van der Heide et al., 2005; Witting et al., 2000).

We can model this formally as follows. Suppose that, as well as our phylogenetic X-tree $\mathscr{T}$ with its edge lengths λ, we also have a acyclic digraph $D=(X,A)$; this could represent for example a 'food web' where an arc $(u,v)\in A$ is present precisely if taxon u feeds on taxon v. We say that a subset S of X is *viable* if the following property holds:

- for every $x\in S$, either x has out-degree 0, or there exists some $s\in S$ such that $(x,s)\in A$.

For the food-web interpretation, this translates into the condition that S is viable if every species under consideration that needs to predate at least one of the other species under consideration has such a species available to it within the set S.

Proposition 6.1. *Let $D=(X,A)$ be a digraph and let $\mathscr{F}$ be the collection of subsets F of X that are viable. Then $(X,\mathscr{F})$ is a greedoid. Moreover, $\mathscr{F}$ has the property that if $F_1,F_2\in\mathscr{F}$, then $F_1\cup F_2\in\mathscr{F}$.*

Proof. Let G be a rooted digraph with vertex set V and root vertex r. Let $\mathscr{F}'$ be the collection of subsets F' of $V-\{r\}$ such that $F'\cup\{r\}$ is the set of vertices of a subtree of G directed away from r. Then $(V-\{r\},\mathscr{F}')$ is a greedoid. This greedoid is sometimes referred to as the *vertex search greedoid* of G (Björner and Ziegler, 1992).

To show that $(X,\mathscr{F})$ is a greedoid, consider the rooted digraph that is obtained from D in the following way:

(i) Add a new vertex r that is adjoined to precisely the vertices of D that have out-degree 0. Initially, the direction of the arcs incident with r are directed towards r.
(ii) Now reverse the direction of all the arcs, and let D' denote the resulting digraph.

It is easily seen that the collection $\mathscr{F}'$ of feasible sets of the vertex search digraph of D' is equal to $\mathscr{F}$, and so $(X,\mathscr{F})$ is a greedoid. Furthermore, as $\mathscr{F}'$ is closed under union (Björner and Ziegler, 1992), it follows that $\mathscr{F}$ is closed under union. This completes the proof of the proposition. □

We note in passing that Proposition 6.1 implies that the pair $(X,\mathscr{F})$ has the structure of an 'antimatroid' (see Björner and Ziegler, 1992, Proposition 8.2.7).

An immediate consequence of Theorem 2.2 and Proposition 6.1 is the following.

Corollary 6.2. *Let (X,A) be an acyclic digraph and let $(X,\mathscr{F})$ denote the greedoid described in Proposition* 6.1. *Let $f:\mathscr{F}\to\mathbb{R}$ be an objective function. If f is compatible with $(X,\mathscr{F})$, then* Greedy *applied to $\mathscr{F}$ and f finds, for all k, a feasible subset of size k that maximizes f.*

Consider now the question of optimizing PD while respecting the dependencies specified by (X,A). The biological motivation for this is that there is no point conserving a species if all of the taxa it depends on go extinct. Formally, we have the following problem.

Problem. Optimizing Diversity with Dependencies.
Instances. A phylogenetic X-tree $\mathscr{T}$ with edge set E, a function $\lambda:E\to\mathbb{R}^{\geqslant 0}$, an acyclic digraph (X,A), and a positive integer k.
Question. Find a subset of X of size at most k that is viable and maximizes the PD score over $\mathscr{T}$ amongst all such subsets.

By Corollary 6.2, Optimizing Diversity with Dependencies can also be solved by Greedy if the weighting on the viable subsets of X induced by their PD score over $\mathscr{T}$ is compatible with $(X,\mathscr{F})$, where $\mathscr{F}$ is the collection of viable subsets of X in (X,A).

For example, suppose the edge lengths of $\mathscr{T}$ are *clock-like*—that is, the sum of the lengths of the edges from the root to each leaf is the same. If, in addition, $\mathscr{T}$ is the phylogenetic X-tree consisting of exactly one interior vertex (the 'star tree'), then Greedy solves this special case of Optimizing Diversity with Dependencies (we can use Corollary 6.2 by choosing f to be the cardinality function).

However, for an arbitrary phylogenetic tree, even with clock-like edge lengths, the greedy algorithm does not solve Optimizing Diversity with Dependencies, as the following example shows.

Let $X=\{a,b,b',x,y\}$ and let D be the digraph shown in Fig. 2(a). The collection of viable subsets of X of size at most three is

$$\mathscr{F}=\{\{a\},\{b\},\{a,b\},\{a,x\},\{b,b'\},\{a,b,b'\},\{a,b,x\},\{b,b',y\}\}.$$

Now consider the rooted phylogenetic tree $\mathscr{T}$ shown in Fig. 2(b) with the indicated edge lengths which are clearly clock-like. The unique member of $\mathscr{F}$ that maximizes the PD score on $\mathscr{T}$ is $\{b,b',y\}$. However, this set does not contain the unique member of $\mathscr{F}$ of size two, namely $\{a,x\}$, that maximizes the PD score on $\mathscr{T}$.

Remark. In computing the PD score of a set of taxa on a rooted phylogenetic tree, one may or may not insist that

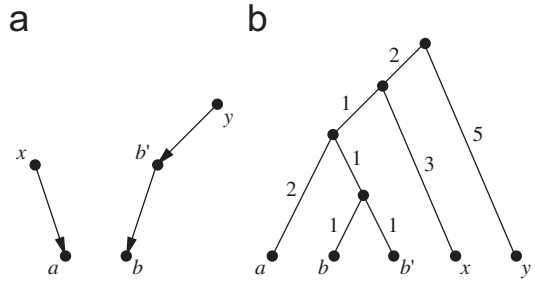

Fig. 2. (a) A digraph D on X and (b) a rooted phylogenetic tree with clock-like edge lengths.

the root be part of this set. To illustrate, in the example shown in Fig. 2, the PD score of $\{a, b, b'\}$ without the root is 5, while its PD score with the root is 8. Nevertheless, the example is valid regardless of which of the two ways we define the PD score for rooted trees.

7. Variations and extensions of PD

We now describe some variations and extensions of PD, and investigate whether the greedy algorithm is guaranteed to find optimal solutions.

First, suppose we have a function $f : 2^X \to \mathbb{R}$. For $k \in \{1, \ldots, |X|\}$ let $m(f, k) = \max\{f(A) : A \subseteq X, |A| = k\}$ and let $M(f, k) = \{A \in X : |A| = k, f(A) = m(f, k)\}$. We say that f satisfies the *nested optimality property* if for every $k \in \{2, \ldots, |X|\}$, there exists a pair $A \in M(f, k), A' \in M(f, k-1)$ with $A' \subset A$. In particular, if GREEDY maximizes f amongst all sets of given cardinality, then f satisfies the nested optimality property, and so to demonstrate that GREEDY fails it suffices to show for some k that the nested optimality property fails. We will use this observation repeatedly in what follows.

Lewis and Lewis (2005) defined a measure of diversity on a phylogenetic X-tree with a non-negative real-valued edge weighting as follows. For a subset S of X, let

$$ED(S) = PD(X) - PD(X - S).$$

Note that selecting a subset S of X of size at most k to maximize ED is equivalent to selecting a subset S' of X of size $|X| - k$ to minimize PD.

However, a simple example shows that ED fails to have the nested optimality property. Consider the phylogenetic tree shown in Fig. 3. It is easily checked that $S = \{b, c, e\}$ is the unique maximum weight subset of X of size three. However, $T = \{a, d\}$ is the unique maximum weight subset of X of size two, and T is not a subset of S.

An alternative way to measure the diversity of a subset S of X has been proposed by Holland (2001) in the context of model strain selection. Here we set

$$M(S) = \min\{PD(\{x, y\}) : x, y \in S\}.$$

Selecting a subset S of X of size k to maximize $M(S)$ corresponds to selecting a subset of k elements of X each pair of which is as 'far apart' as possible in the tree. Note that, as before, sets which maximize M do not necessarily maximize PD and vice versa. As with the previous variation, M does not have the nested optimality property.

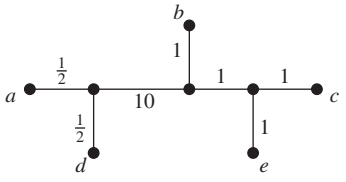

Fig. 3. A phylogenetic tree with edge lengths. Here $\{b, c, e\}$ is the unique subset of size three that maximizes $ED(\{b, c, e\}) = 15 - 1 = 14$, while $\{a, d\}$ is the unique subset of size two that maximizes $ED(\{a, d\}) = 15 - 4 = 11$.

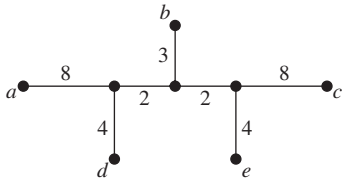

Fig. 4. A phylogenetic tree with edge lengths. Here $\{a, c, d, e\}$ is the unique subset of size four that maximizes $M(\{a, c, d, e\}) = 12$, but it does not contain $\{a, b, c\}$, the unique subset of size three that maximizes $M(\{a, b, c\}) = 13$.

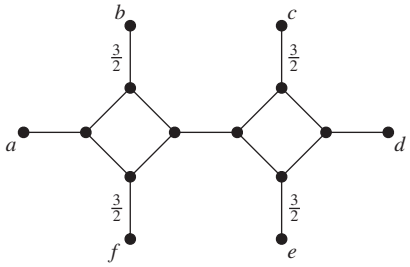

Fig. 5. A split network (Huson and Bryant, 2006) corresponding to a circular split system. Apart from the edges with weight $\frac{3}{2}$, all edges have weight 1. Here $\{b, c, e, f\}$ is a subset of size four that maximizes $PD(\{b, c, e, f\}) = 11$, but it does not contain $\{a, d\}$, the unique set of size two that maximizes $PD(\{a, d\}) = 7$.

Consider the phylogenetic tree shown in Fig. 4 with its edge lengths. The set $\{a, c, d, e\}$ is the unique subset of size four that maximizes M, but it does not contain $\{a, b, c\}$, the unique subset of size three that maximizes M.

Instead of varying PD, it is natural to extend PD on phylogenetic trees to more general structures. For example, we may regard a phylogenetic X-tree as a collection Σ of pairwise compatible X-splits (Buneman, 1971) and we can regard edge lengths as a map $\lambda : \Sigma \to \mathbb{R}^{\geqslant 0}$, in which case the PD score of a subset S of X is

$$\sum \lambda(A|B),$$

where the sum is over all $A|B \in \Sigma$ with $A \cap S \neq \emptyset$ and $B \cap S \neq \emptyset$. Extending this definition of PD on a pairwise compatible collection of λ-weighted splits to an arbitrary collection of splits, we can ask the question of whether the pair (Σ, λ) has the nested property. Here we could impose, for example, that Σ is either circular or weakly compatible (Bandelt and Dress, 1992).

In general, (Σ, λ) does not have the nested property. For example, consider the network shown in Fig. 5. This network is a pictorial way of describing a collection of weighted splits that are circular. Splits are obtaining by deleting 'parallel edges'. Except for the four edges with weight $\frac{3}{2}$, all edges have weight 1. The distance between two labelled vertices u and v is the length of a shortest path joining u and v. It is easily checked that $\{b, c, e, f\}$ is the unique subset of size four that maximizes PD. However, this set does not contain the unique set of size two, namely $\{a, d\}$, that maximizes PD, and so the nested optimality

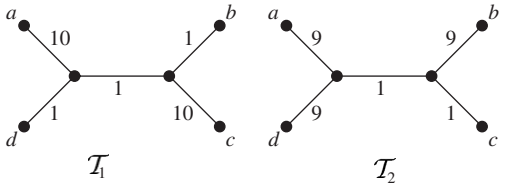

Fig. 6. For the multivariate function $G_g(S) = \max\{x_1, x_2\}$, the set $\{a, b, d\}$ is the unique 3-element set that maximizes G_g, but it does not contain $\{a, c\}$, the unique 2-element set that maximizes G_g.

property fails. Nevertheless, there are collections Σ of non-compatible splits for which (Σ, λ) has the nested optimality property. An interesting (and possibly challenging) problem is to characterize these structures.

Fast algorithms for constructing high-diversity subsets for networks could be useful in case trees are constructed from various genomic regions (Minh et al., 2006). In this context, it is worth mentioning one further extension of PD. Given a collection of phylogenetic X-trees $\mathscr{T}_1, \ldots, \mathscr{T}_n$ and a subset S of X we define

$$G_g(S) = g(PD_1(S), \ldots, PD_n(S)),$$

where $PD_i(S)$ is the PD score of S with respect to tree cT_i for $1 \leqslant i \leqslant n$ and g is some multivariate function. In general, for $n > 1$, the greedy algorithm will again fail to find subsets of X optimizing G_g. For example, suppose that $G_g(S) = \max\{x_1, \ldots, x_n\}$, where $x_i = PD_i(S)$ for all i, and consider the two phylogenetic trees in Fig. 6. Here $\{a, b, d\}$ is the unique 3-element subset of X that maximizes G_g, yet it does not contain $\{a, c\}$, the unique 2-element subset of X that maximizes G_g, and so the nested optimality property fails. Note that $G_g(\{a, b, d\}) = 28$ and $G_g(\{a, c\}) = 21$.

Despite the last example, a subset S of X of size k maximizing $G_g(S) = \max\{x_1, \ldots, x_n\}$ may be found by applying the greedy algorithm to each cT_i to find a set of size k with highest PD_i score, and then taking the highest scoring set amongst all of these sets. It would be interesting to investigate whether other multivariate functions G_g could be optimized using variants of the greedy algorithm.

Acknowledgments

The first author thanks the Department of Mathematics and Statistics, University of Canterbury, New Zealand for hosting him during the preliminary stages of this work, during which time he was supported by a University of Canterbury Erskin Fellowship. The first author was supported in part by an EPSRC Grant (EP/D068800/1), while the second and third authors were supported by the New Zealand Marsden Fund (06UOC02).

References

Bandelt, H.-J., Dress, A., 1992. A canonical decomposition theory for metrics on a finite set. Adv. Appl. Math. 7, 47–105.

Barker, G.M., 2002. Phylogenetic diversity: a quantitative framework for measurement of priority and achievement in biodiversity conservation. Biol. J. Linn. Soc. 76, 165–194.

Björner, A., Ziegler, G.M., 1992. Introduction to greedoids. In: White, N. (Ed.), Matroid Applications. Cambridge University Press, Cambridge.

Bryant, V., Brooksbank, P., 1992. Greedy algorithm compatibility and heavy-set structures. Eur. J. Combin. 13, 81–86.

Buneman, P., 1971. The recovery of trees from measures of dissimilarity. In: Hodson, F.R., Kendall, D.G., Tautu, P. (Eds.), Mathematics in the Archaeological and Historical Sciences. Edinburgh University Press.

Faith, D.P., 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61, 1–10.

Faith, D.P., Baker, A.M., 2006. Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges. Evol. Bioinf. Online 2, 70–77.

Garey, M.R., Johnson, D.S., 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman, San Francisco, CA.

Goecke, O., 1988. A greedy algorithm for hereditary set systems and a generalization of the Rado–Edmons characterization of matroids. Discr. Appl. Math. 20, 39–49.

Hartmann, K., Steel, M., 2007. Phylogenetic diversity: from combinatorics to ecology. In: Gascuel, O., Steel, M. (Eds.), New Mathematical Models in Evolution. Oxford University Press, Oxford, in press.

Holland, B.R., 2001. Evolutionary analyses of large data sets: trees and beyond. Ph.D. Thesis, Massey University, New Zealand.

Huson, H., Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. Mol. Biol. Evol. 23, 254–267.

Korte, B., Lovász, L., Schrader, R., 1991. Greedoids, Algorithms and Combinatorics. Springer, Berlin.

Lewis, L.A., Lewis, P.O., 2005. Unearthing the molecular phylodiversity of desert soil green algae (Chlorphyta). Syst. Biol. 54, 936–947.

Minh, B.Q., Klaere, S., von Haesler, A., 2006. Phylogenetic diversity within seconds. Syst. Biol. 55, 769–773.

Pardi, F., Goldmann, N., 2005. Species choice for comparative genomics: being greedy works. PLoS Genet. 1, e71.

Rodrigues, A.S.L., Brooks, T.M., Gaston, K.J., 2005. Integrating phylogenetic diversity in the selection of priority areas for conservation: does it make a difference? In: Purvis, A., Gittleman, J.L., Brooks, T. (Eds.), Phylogeny and Conservation. Cambridge University Press, Cambridge.

Semple, C., Steel, M., 2003. Phylogenetics. Oxford University Press, Oxford.

Steel, M., 2005. Phylogenetic diversity and the greedy algorithm. Syst. Biol. 54, 527–529.

van der Heide, C.J., von den Bergh, C., van Ierland, E.C., 2005. Extending Weitzman's economic ranking of biodiversity protection: combining ecological and genetic considerations. Ecol. Econ. 55, 218–223.

Witting, L., Tomiuk, J., Loeschcke, V., 2000. Modelling the optimal conservation of interacting species. Ecol. Modell. 125, 123–143.