OPTIMIZING PHYLOGENETIC DIVERSITY WITH ECOLOGICAL CONSTRAINTS

BEÁTA FALLER[1], CHARLES SEMPLE, AND DOMINIC WELSH

ABSTRACT. Given an edge-weighted tree $\mathcal{T}$ with leaf set X, define the weight of a subset S of X as the sum of the edge-weights of the minimal subtree of $\mathcal{T}$ connecting the elements in S. It is known that the problem of selecting subsets of X of a given size to maximize this weight can be solved using a greedy algorithm. This optimization problem arises in conservation biology where the weight is referred to as the phylogenetic diversity of a taxa set S. Here, we consider the extension of this problem whereby we are only interested in selecting subsets of the taxa set that are ecologically 'viable'. Such subsets are specified by an acyclic digraph which represents, for example, a food web. This additional constraint makes the problem computationally hard. In this paper, we analyze the complexity of different variations of the extended problem.

1. INTRODUCTION

In the context of conservation biology, maximizing phylogenetic diversity (PD) is a prominent selection criteria for deciding which species to conserve (e.g. [2, 3, 4, 10, 11, 14, 13, 18]). Intuitively, given a phylogenetic (evolutionary) tree $\mathcal{T}$, the PD of a set of present-day species is the sum of the edges of the minimal subtree of $\mathcal{T}$ that connects the species in the set. In its most direct application to conservation, one selects a k-element subset of species that maximizes PD over all k-element subsets. While PD makes a comparison between species to capture the notion of diversity, the conservation of individual species are considered in isolation. In real ecosystems this can be problematic as species frequently depend on other species for their survival—there is no point conserving a species if all the species it depends on go extinct [6, 21]. In this paper, we consider an extension of the PD selection criteria, where only subsets that are 'viable' are considered for conservation.

A *phylogenetic X-tree* $\mathcal{T}$ is an unrooted tree with no degree-2 vertices and whose leaf set is X. Here, $\mathcal{T}$ represents the evolutionary relationships of the taxa in X. Ignoring the edge weights and dashed lines, a phylogenetic tree with $X = \{a, b, c, d, e, f, g\}$ is shown in Fig. 1(a). Let λ be a non-negative real-valued weighting

Date: 16 September 2008.

1991 *Mathematics Subject Classification.* 05C05; 92D15.

Key words and phrases. Phylogenetic tree, phylogenetic diversity, food web.

The first and second authors thank the New Zealand Marsden Fund, and the third author thanks the New Zealand Institute of Mathematics and its Applications.

[1] Corresponding author.

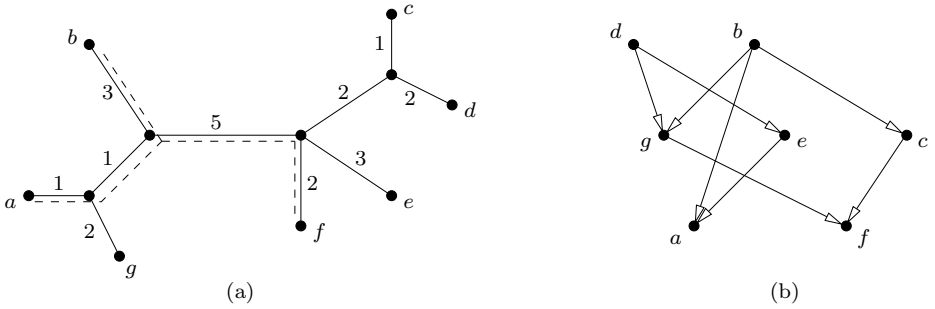

FIGURE 1. (a) A phylogenetic X-tree $\mathcal{T}$ and (b) a food web D on X.

on the edges of $\mathcal{T}$. The *phylogenetic diversity* (PD) of a subset S of X is

$$PD_{\mathcal{T}}(S) = \sum_{e \in E(\mathcal{T}(S))} \lambda(e),$$

where $E(\mathcal{T}(S))$ is the edge set of the minimal subtree of $\mathcal{T}$ connecting the leaves in S. If there is no ambiguity, we frequently denote $PD_{\mathcal{T}}(S)$ by $PD(S)$. Referring to Fig. 1(a), the PD of $\{a, b, f\}$ is the sum of the edge weights of the subtree of $\mathcal{T}$ highlighted with dashed lines. In particular, $PD(\{a, b, f\}) = 12$.

Given a phylogenetic X-tree $\mathcal{T}$ and a fixed integer k, the *PD optimization problem* is to find

$$\max\{PD(S) : S \text{ is a } k\text{-element subset of } X\}.$$

Pardi and Goldmann [11] and Steel [18] independently showed that a solution to this problem can be found in polynomial time using a greedy algorithm.

To allow for ecological dependencies in the conserving of species, we extend the PD optimization problem to additionally include an acyclic digraph $D = (X, A)$. Here, D could be an ecological network, for example a 'food web', where $(u, v) \in A$ precisely if taxon u feeds or preys on taxon v. We say that a subset S of X is *viable* if, for each $s \in S$, there is a directed path in D from s to a vertex with out-degree zero in which every vertex in the path is in S. In Fig. 1(b), $\{a, b, f\}$ is viable. However, $\{a, b, c\}$ is not viable as there is no directed path in D from c to a vertex with out-degree 0 using only vertices in $\{a, b, c\}$. Under the food-web interpretation, a set S is viable if, for each taxon in S that is not at the bottom of the food chain, there is a taxon in S that it feeds or preys on. Formally, the problem we are interested in is the following:

Decision Problem: OPTIMIZING PD WITH DEPENDENCIES
Instance: A phylogenetic X-tree $\mathcal{T}$, a non-negative (real valued) weighting λ on the edges of $\mathcal{T}$, an acyclic digraph $D = (X, A)$, a positive integer k, and a non-negative real number d.
Question: Is there a viable subset S of X of size at most k with $PD(S) \geq d$?

As stated, this problem has been considered by Moulton *et al.* [10] and Spillner *et al.* [17]. The first paper was interested in the problem in the context of greedoids

and greedy algorithms, while the second paper noted without proof that the problem was NP-complete. The purpose of our paper is to establish which variations of Optimizing PD with Dependencies are computationally easy and which variations of it are computationally hard. In addition to the immediate significance of knowing the complexity of the restricted problems, these results increase our knowledge of the essential elements which made the original problem NP-complete.

The organization of the paper is as follows. The next section contains some preliminaries that are used throughout the paper. A *star tree* is a phylogenetic tree with exactly one interior vertex. In Section 3, we show that Optimizing PD with Dependencies is NP-complete even if $\mathcal{T}$ is a star tree. Section 4 considers polynomial-time instances of Optimizing PD with Dependencies when $\mathcal{T}$ is a star tree. Such instances rely on the underlying graph of D containing no (undirected) cycles. The other possibility to consider is when $\mathcal{T}$ is arbitrary, but the underlying graph of D is a rooted tree. However, as we show in Section 5, this particular possibility is also NP-complete. For both intrinsic and practical reasons, greedy algorithms have been frequently considered in the context of phylogenetic diversity. A curious feature of the problem Optimizing PD with Dependencies which gives some additional indication of its hardness is highlighted in Section 6 where we show that it is NP-complete to decide if the feasible solution obtained by the greedy algorithm can be bettered. Throughout most of the paper, we restrict ourselves to unrooted phylogenetic trees. In the last section, we consider the extension of our earlier results to rooted phylogenetic trees, including such trees satisfying the molecular clock hypothesis. The notation and terminology of the paper follows [16].

2. Vertex Cover and the Star Tree Problem

Vertex Cover is a classical NP-complete problem and is frequently used for completeness reductions. As we use Vertex Cover several times in the paper, we give a formal definition of it here. Furthermore, we also describe a problem equivalent to Optimizing PD with Dependencies in case $\mathcal{T}$ is a star tree.

For a graph $G = (V, E)$, a *vertex cover* of G is a subset V' of V such that, for each edge $\{u, v\} \in E$, at least one of u and v belongs to V'. The NP-complete problem Vertex Cover [5] is the following:

Decision Problem: Vertex Cover
Instance: A graph $G = (V, E)$ and a positive integer $m \leq |V|$.
Question: Is there a vertex cover of G of size at most m?

A special instance of Optimizing PD with Dependencies is when $\mathcal{T}$ is a star tree. Because a star tree contains no non-pendant edges, this special instance can be reformulated as a problem on acyclic digraphs. In particular, let w be the weighting on the vertices of D defined by setting $w(v) = \lambda(\{r, v\})$ for all $v \in X$, where r denotes the interior vertex of $\mathcal{T}$. In this setting, for all subsets S of X, set $PD(S) = \sum_{v \in S} w(v)$. It is now easily checked that the following decision problem

is equivalent to Optimizing PD with Dependencies when $\mathcal{T}$ is a star tree and $k \geq 2$.

Decision Problem: Optimizing PD in Vertex-Weighted Food Webs
Instance: An acyclic digraph $D = (X, A)$, a non-negative (real-valued) weighting w on the vertices of D, a positive integer k, and a non-negative real number d.
Question: Is there a viable subset S of X of size at most k with $PD(S) \geq d$?

The above equivalence will be freely used several times in this paper. Although the typical model of evolution is a bifurcating tree, there are instances for which it appears that the underlying model is more star-like than bifurcating (for example, see [20] and the references therein). Thus, restricting Optimizing PD with Dependencies to when $\mathcal{T}$ is a star tree is also of practical importance.

3. NP-Completeness of Optimizing PD with Dependencies

In this section, we show that the decision problem Optimizing PD with Dependencies is NP-complete even if $\mathcal{T}$ is a star tree. In particular, recalling the equivalence in Section 2, we prove the following theorem.

Theorem 3.1. Optimizing PD in Vertex-Weighted Food Webs *is* NP-*complete.*

Proof. Evidently, Optimizing PD in Vertex-Weighted Food Webs is in NP since, given a subset S of X of size at most k, one can easily check in polynomial time if S is viable and $PD(S) \geq d$. To complete the proof of the theorem, we show that there is a polynomial-time reduction from Vertex Cover to Optimizing PD in Vertex-Weighted Food Webs.

Given a graph $G = (V, E)$ and a positive integer m, we construct an instance of Optimizing PD in Vertex-Weighted Food Webs as follows. Let D be the acyclic digraph whose vertex set X is the (disjoint) union of V and E, and whose arc set A is defined to be

$$A = \{(e, v) : e \in E,\, v \in V,\, v \text{ is an end-vertex of } e \text{ in } G\}.$$

Let w be the weight function $w : X \to \mathbb{R}^{\geq 0}$ specified by assigning weight 1 to each vertex in $X \cap E$ and weight 0 to each vertex in $X \cap V$. Clearly, this construction can be accomplished in polynomial time.

We now show that there is a vertex cover of G of size at most m if and only if there is a viable subset S of X of size at most $|E| + m$ with $PD(S) \geq |E|$. First, suppose that $V' \subseteq V$ is a vertex cover for G with $|V'| \leq m$. Then, the construction of D implies that $V' \cup E$ forms a viable subset of X. Since $|V' \cup E| = |V'| + |E| \leq m + |E|$, and $w(V' \cup E) = |E|$, it follows that $V' \cup E$ is a viable subset of X of size at most $|E| + m$ and with weight at least $|E|$. Conversely, suppose that there is a viable subset S of X of size at most $|E| + m$ and with weight at least $|E|$. Since the vertices in $X \cap V$ have weight 0, the subset S must contain all $|E|$ vertices with weight 1; that is, it must contain $X \cap E$. Therefore $S = V' \cup E$ for some $V' \subseteq V$. Since S is viable, there is an arc from each $e \in E$ to some vertex $v \in V'$. In terms

of G, this implies that V' is a vertex cover of G. As $|S| \leq |E| + m$, it follows that $|V'| \leq m$ completing the proof of the theorem. □

Remark. Theorem 3.1 tells us that Optimizing PD with Dependencies is NP-complete even if $\mathcal{T}$ is a star tree. However, the proof of this theorem says the problem remains NP-complete if D is a bipartite digraph with vertex parts V_1 and V_2, where each vertex in V_1 has in-degree 0 and out-degree 2 and each vertex in V_2 has out-degree 0. Moreover, it is also interesting to note that we could have used any restricted version of Vertex Cover for the reduction provided the version is NP-complete. For example, it has been shown that Vertex Cover remains NP-complete if G is cubic and planar [9]. A graph is *cubic* if each vertex has degree three. Thus Optimizing PD with Dependencies remains NP-complete if $\mathcal{T}$ is a star tree and D is a bipartite graph as described above with the additional properties that each vertex in V_2 has in-degree 3 and D is planar. To see planarity, observe that D can be obtained by taking a planar drawing of the planar graph G, subdivide each edge of G and, for each resulting vertex u, direct the incident edges away from u.

4. Star Tree and Food Tree

In contrast to the NP-completeness results of this paper, we have the following theorem.

Theorem 4.1. Optimizing PD in Vertex-Weighted Food Webs *can be solved in polynomial time if D is either*

(i) *a rooted tree with all arcs directed away from the root or*
(ii) *a rooted tree with all arcs directed towards the root.*

Theorem 4.1 is an immediate consequence of what appears to be a well-known dynamic programming algorithm for solving the following problem (for example, see [8]). Let T be a rooted tree with root ρ and let k be a positive integer. Suppose that the vertices of T are assigned real-valued weights. The problem is to find a maximum-weighted subtree of T with root ρ and at most k vertices.

We briefly outline the dynamic programming algorithm here in the language of this paper. For convenience, we view Optimizing PD in Vertex-Weighted Food Webs as an optimization problem. First consider (ii). Let D be a digraph satisfying (ii) in the statement of Theorem 4.1, and let ρ be the root of D. Thus D contains exactly one vertex of out-degree 0, namely ρ. Let v be a vertex of D. We denote the subset of vertices u of D for which (u, v) is an arc in D by $I(v)$. Furthermore, we denote the rooted subtree of D with root v whose vertex set is precisely the subset of vertices x of D for which there is a directed path from x to v by $D(v)$.

For a vertex v of T and a non-negative integer $q \leq k$, let $S(v, q)$ denote the optimal solution of Optimizing PD in Vertex-Weighted Food Webs when D is chosen to be $D(v)$ and the size of the viable subset is at most q. Note that $S(\rho, k)$

denotes the optimal solution for the original problem. Clearly, for any vertex v of T, we have $S(v,0)=0$ and, for each vertex u of in-degree 0, $S(u,q)=w(u)$ for all $1\leq q\leq k$. The dynamic programming algorithm starts at vertices of in-degree 0 and works itself towards ρ using the recursion

$$S(v,q)=w(v)+\max_{\{q_u:\sum_{u\in I(v)}q_u\leq q-1\}}\sum_{u\in I(v)}S(u,q_u)$$

for $1\leq q\leq k$. It is shown in [8] that this approach leads to a quadratic-time algorithm for finding $S(\rho,k)$.

If D is a digraph satisfying (i) in the statement of Theorem 4.1, then we simply modify the above algorithm in the obvious way to find a minimum-weight subtree of D rooted at ρ with at least $n-k$ vertices. The complement of the resulting solution, that is $\sum_{u\in V(D)}w(u)$ minus this solution, gives the desired optimal solution.

Despite the above positive results, we end this section with the following conjecture where no constraints are placed on the direction of the arcs.

Conjecture 4.2. Optimizing PD in Vertex-Weighted Food Webs *when the underlying graph of D is a tree is NP-complete.*

5. Arbitrary Phylogenetic Tree and Food Tree

In this section, we show that Optimizing PD with Dependencies is still NP-complete if $\mathcal{T}$ is an arbitrary phylogenetic tree while D is a rooted tree. In particular, we establish the following theorem.

Theorem 5.1. Optimizing PD with Dependencies *when $\mathcal{T}$ is an arbitrary phylogenetic tree and D is either*

(i) *a rooted tree with all arcs directed away from the root, or*
(ii) *a rooted tree with all arcs directed towards the root*

is NP-complete.

Proof. We prove (i). The proof of (ii) is similar and omitted. Since Optimizing PD with Dependencies is in NP, this particular instance of the problem is also in NP. Like the NP-completeness proof for Theorem 3.1, the reduction is from Vertex Cover. However, for this proof, we use the restricted version of Vertex Cover in which G is cubic and planar. It is shown in [9] that Vertex Cover remains NP-complete under these restrictions.

Let $G=(V,E)$ be a cubic, planar graph. We construct an instance of the restricted version of Optimizing PD with Dependencies described by (i) as follows. Colour the edges of G with three colours $\{1,2,3\}$ such that no two edges incident with the same vertex receive the same colour. Due to a classic construction of Tait [19], this is equivalent to four-colouring the faces of a planar drawing of G which can be done in quadratic time [12]. For each colour $c\in\{1,2,3\}$, let

$$V_c=\{u_c:u\in V\},$$

and let T_c be the tree with leaf set V_c that consists of a (central) vertex z_c of degree $|V|/2$, where the $|V|/2$ neighbours of z_c each have degree 3, and the $|V|$ leaves are arranged so that, for each edge $\{u, v\}$ of G coloured c, the vertices u_c and v_c are adjacent to the same degree-3 vertex. As G is a cubic graph, T_c is well-defined for all c. Let $\mathcal{T}$ be the phylogenetic X-tree that is constructed by starting with components T_1, T_2, and T_3, and two new (isolated) vertices x and y, and then connecting these components with new edges $\{x, y\}$, $\{y, z_1\}$, $\{y, z_2\}$, and $\{y, z_3\}$. Observe that the leaf set of $\mathcal{T}$ is $V_1 \cup V_2 \cup V_3 \cup \{x\}$. We specify the weighting function λ by setting

$$\lambda(e) = \begin{cases} 0 & \text{if } e \text{ is a pendant edge incident with a vertex in } V_1 \text{ or } V_2; \\ N & \text{if } e \text{ is a pendant edge incident with a vertex in } V_3; \\ 0 & \text{if } e = \{x, y\} \text{ or } e = \{y, z_c\} \text{ for some } c \in \{1, 2, 3\}; \\ 1 & \text{otherwise,} \end{cases}$$

where N is sufficiently large, say $N > |E|$. With this construction and weighting, our phylogenetic tree and corresponding edge weighting is complete. Now let D be the associated rooted tree with vertex set $V_1 \cup V_2 \cup V_3 \cup \{x\}$ and arc set

$$\bigcup_{u \in V} \{(x, u_3), (u_3, u_2), (u_2, u_1)\}.$$

Note that x is the root of D. Clearly, both $\mathcal{T}$ and D can be constructed in polynomial time.

We complete the proof by showing that G has a vertex cover of size at most m if and only if there is a viable subset S of X of size at most $3m$ such that $PD(S) \geq |E| + mN$. Suppose first that there is a vertex cover $V' \subseteq V$ for G with $|V'| = m$. By selecting S to be the set $\{u_c : c \in \{1, 2, 3\} \text{ and } u \in V'\}$, we have a viable subset of X of size $3m$. Moreover, observing that there are exactly $|E|$ edges in $\mathcal{T}$ with weight 1 (each corresponding to a distinct edge of G), $PD(S) = |E| + mN$.

Conversely, suppose that there is a viable subset S of X of size at most $3m$ that has PD score at least $|E| + mN$. Since $N > |E|$ and $PD(S) \geq |E| + mN$, it follows that S must contain at least m leaves of T_3 so that the minimal subtree of $\mathcal{T}$ connecting the elements of S includes m edges with weight N. But then, as S is viable, for each such leave u_3 in S, the set S also includes u_1 and u_2. Therefore $|S| = 3m$ and consists of exactly these vertices. For $PD(S) \geq |E| + mN$, it now follows that the minimal subtree of $\mathcal{T}$ connecting the elements in S must contain all $|E|$ edges with weight 1. In turn, this implies that $V' = \{u \in V : u_3 \in S\}$ is a vertex cover of G. As $|V'| = m$, we have completed the proof of the theorem. □

6. Improving Greedy Solutions is Hard

Greedy algorithms have been regularly considered as approaches for solving problems that optimize some measure of diversity (for example, [1, 2, 11, 7, 10, 18]). There are a variety of reasons for this consideration. First, they are fast, simple to use and implement, and, more importantly, solve the original PD problem exactly [11, 18] and provide sharp approximation algorithms for other PD-related problems [1, 2]. Indeed, the fact that the original PD problem can be solved in this way,

motivated Moulton *et al.* [10] to consider PD and the greedy algorithm in detail. Second, in the context of conservation biology, they underlie the desirable property of stability [1]. In particular, one would like the set of species to be targeted for conservation to be stable as budgets vary. For example, if, given some initial budget, one selects a set of species to conserve resulting from a diversity-based method, one would like most of that set to remain if the budget was to be adjusted up or down at a later date and the chosen set of species to conserve was reselected under the new budget.

In this section, we consider the following greedy approach to solving Optimizing PD in Vertex-Weighted Food Webs.

Algorithm: Greedy(D, w, k)
Input: An acyclic digraph $D = (X, A)$, a non-negative (real-valued) weighting w on the vertices of D, and a positive integer k.
Output: A viable subset of X of size k.

Step 1 Let S be the empty set and set counter $c = 0$.
Step 2 If $c = k$, STOP; otherwise, select an element z of $X - S$ so that $S \cup \{z\}$ is viable and maximizes $PD(S \cup \{z\}) - PD(S)$.
Step 3 Set $S = S \cup \{z\}$ and $c = c + 1$, and return to Step 2.

It is not difficult to construct a counterexample to show that Greedy does not necessarily find an optimal solution to Optimizing PD in Vertex-Weighted Food Webs. Of course, since Greedy is trying to solve an NP-hard problem, this is not surprising. However, what is perhaps unexpected is that deciding if there is a feasible solution better than that returned by Greedy is NP-complete as we show next. It would be interesting to know of other situations where improving greedy solutions was a provably hard problem.

Decision Problem: Greedy Optimality
Instance: An acyclic digraph $D = (X, A)$, a non-negative (real-valued) weighting w on the nodes of D, a positive integer k, and the PD score g of the solution returned by Greedy applied to (D, w, k).
Question: Is there a viable subset S of X of size at most k such that $PD(S) > g$?

Theorem 6.1. Greedy Optimality *is* NP*-complete.*

Proof. Greedy Optimality is clearly in NP since, given a subset S of X, one can easily verify in polynomial time whether S is viable and $PD(S) > g$. To complete the proof of the theorem, we show that there is a polynomial-time reduction from Vertex Cover to Greedy Optimality.

Let $G = (V, E)$ and m be a given instance of Vertex Cover. Let D be the acyclic digraph whose vertex set is the union of $V \cup E$ and $U = \{u_1, u_2, \ldots, u_{|E|+m-1}\}$, and whose arc set is the union of

$$\{(e, v) : e \in E,\ v \in V,\ v \text{ is an end-vertex of } e \text{ in } G\}$$

and

$$\{(u_i, u_{i-1}) : i \in \{2, 3, \ldots, |E| + m - 1\}\}.$$

Now let w be any function from the vertex set of D to $\mathbb{R}^{\geq 0}$ that is defined, for all $x \in X$, by setting

$$w(x) = \begin{cases} 0 & \text{if } x \in V; \\ 1 & \text{if } x \in E; \\ \delta & \text{if } x \in \{u_1, u_2, \ldots, u_{|E|+m-2}\}; \\ \alpha & \text{if } x = u_{|E|+m-1}, \end{cases}$$

where $(|E| + m - 2)\delta + \alpha = |E| - \epsilon$ for some $\epsilon > 0$ and $0 < \delta < \frac{1-\epsilon}{|E|+m-2}$. Clearly, such a function exists.

Let $k = m + |E|$ and let g_k be the solution of Greedy applied to (D, w, k). Observe that $g_k = |E| - \epsilon$ and that any set corresponding to this solution must contain all of the elements in U and exactly one element in V. We next show that there is a vertex cover for G of size at most m if and only if there is a viable subset S of X of size at most $m + |E|$ such that $PD(S) > g_k$.

Suppose first that there is a vertex cover V' of G of size at most m. Then, by taking the subset $V' \cup E$ of the vertex set of D, we have a viable subset of size at most $m + |E|$ whose weight is $|E|$. In particular, $PD(V' \cup E) > g_k$.

For the converse, suppose that there is a viable subset S of X of size at most $m + |E|$ such that $PD(S) > g_k$. If $E \subset S$, then we have a vertex cover for G of size at most m by choosing the set $V \cap S$. Therefore we may assume that E is not a subset of S. Furthermore, if $u_{|E|+m-1} \in S$, then, as S is viable, $U \subseteq S$. In this case, as $|S| \leq m + |E|$, we have $S = U$ or $S = U \cup \{v\}$ for some $v \in V$. But then $PD(S) = |E| - \epsilon = g_k$; a contradiction. Thus we may also assume that $u_{|E|+m-1} \notin S$. Since E is not a subset of S, it follows that

$$PD(S) \leq (|E| + m - 2)\delta + |E| - 1. \tag{1}$$

But δ was chosen so that $\delta < \frac{1-\epsilon}{|E|+m-2}$, that is

$$(|E| + m - 2)\delta < 1 - \epsilon.$$

Combining this with (1), we get

$$PD(S) \leq (|E| + m - 2)\delta + |E| - 1 < 1 - \epsilon + |E| - 1 = |E| - \epsilon,$$

contradicting the fact that $PD(S) > g_k = |E| - \epsilon$. It follows that E must be a subset of S and so there is a vertex cover of G of size at most m. Since the reduction can be done in polynomial time, this completes the proof of the theorem. □

As noted in the introduction, Moulton *et al.* considered Optimizing PD with Dependencies in the context of greedy algorithms. They observed that, in the trivial case $\mathcal{T}$ is a star tree in which all edge weights are equal, the problem is solvable via a greedy algorithm. One can extend this observation further by showing that Optimizing PD in Vertex-Weighted Food Webs is solvable via Greedy if D has the property that, whenever P is a directed path in D, then $w(u) \leq w(v)$ for all $(u, v) \in P$. An interesting problem would be to determine precisely when Optimizing PD in Vertex-Weighted Food Webs is solvable via Greedy.

However, Theorem 6.1 shows that any such characterization will not be validated in polynomial time unless P=NP.

7. Rooted Phylogenetic Trees and the Molecular Clock

In practice, one frequently works with rooted phylogenetic trees and therefore the rooted analogue of PD. In this short section, we review the implications of our earlier results in this setting.

A *rooted phylogenetic X-tree* $\mathcal{T}$ is a rooted tree with no degree-2 vertices except perhaps the root and whose leaf set is X. Such a tree commonly describes the evolution of the set X of extant species from their common hypothetical ancestor (the root). Let λ denote a non-negative real-valued weighting on the set of edges of $\mathcal{T}$. For a subset S of X, the *rooted PD* (rPD) of S is the sum of the edge lengths of the minimal subtree of $\mathcal{T}$ that connects the elements in S and the root of $\mathcal{T}$ [3]. The rooted analogue of Optimizing PD with Dependencies, called Optimizing rPD with Dependencies, is the same as that in the unrooted setting but with the rooted phylogenetic tree replacing the (unrooted) phylogenetic tree and using rPD instead of PD. A *rooted star tree* is a rooted phylogenetic tree in which the only interior vertex is the root. As in the unrooted setting, when $\mathcal{T}$ is a rooted star tree, Optimizing rPD with Dependencies is equivalent to Optimizing PD in Vertex-Weighted Food Webs. A minor point to note is that, unlike the unrooted setting where $k \geq 2$ for this equivalence to work, there is no restriction on k in the rooted equivalence. It is now easily seen that Theorems 3.1, 4.1, and 6.1 apply to Optimizing rPD with Dependencies too. Furthermore, the rooted analogue of Theorem 5.1 also holds. This can be easily checked by making minor changes to the proof of Theorem 5.1. In particular, distinguishing the interior vertex y as the root in the constructed tree and using rPD instead of PD in the course of the reduction.

In biology, it is sometimes reasonable to assume that mutations in evolution occur at a constant rate. This assumption is called the *molecular clock* assumption. Mathematically speaking, this assumption implies that, in a rooted phylogenetic tree, the sum of the lengths of the edges from the root to each leaf is the same. The notion of the existence of a molecular clock first appeared in [22] followed by [15]. Now consider Optimizing rPD with Dependencies under the assumption that the edge-weights of $\mathcal{T}$ satisfy the molecular clock. If $\mathcal{T}$ is a star tree, then Optimizing rPD with Dependencies is trivially solvable in polynomial time [10]. However, if $\mathcal{T}$ is arbitrary and D is a food tree, then Optimizing rPD with Dependencies is NP-complete.

Theorem 7.1. Optimizing rPD with Dependencies *when $\mathcal{T}$ is a rooted phylogenetic tree with the molecular clock assumption and D is either*

(i) *a rooted tree with all arcs directed away from the root, or*
(ii) *a rooted tree with all arcs directed towards the root*

is NP-complete.

Proof. We just outline the proof of (i). The proof of (ii) is similar. We use a reduction from the restricted version of VERTEX COVER in which G is cubic and planar. The proof is essentially the same as that of the proof of Theorem 5.1, and so we just highlight the necessary changes.

Distinguish the interior vertex y of the phylogenetic tree constructed in the proof of Theorem 5.1 to obtain a rooted phylogenetic tree $\mathcal{T}_y$ with root y. Using the original weighting function λ , we make $\mathcal{T}_y$ clock-like with the following weighting function λ_y:

$$\lambda_y(e) = \begin{cases} N+1 & \text{if } e = \{x, y\}; \\ N & \text{if } e \in \{\{y, z_1\}, \{y, z_2\}\}; \\ \lambda(e) & \text{otherwise.} \end{cases}$$

Setting $k = 3m+1$ and $d = |E|+(m+3)N+1$ completes the necessary changes. □

ACKNOWLEDGMENTS

We thank Magnus Bordewich, Peter Lockhart, Mike Steel, and Alexander Zelikovsky for useful discussions relating to the work in this paper.

REFERENCES

[1] M. Bordewich, A.G. Rodrigo and C. Semple, Selecting taxa to save or sequence: desirable criteria and a greedy solution, Systematic Biology, in press.
[2] M. Bordewich and C. Semple, Nature reserve selection problem: a tight approximation algorithm, IEEE/ACM Transactions on Computational Biology and Bioinformatics 5 (2008) 275-280.
[3] D.P. Faith, Conservation evaluation and phylogenetic diversity, Biol. Conserv. 61 (1992) 1-10.
[4] D.P. Faith and A.M. Baker, Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges, Evol. Bioinf. Online 2 (2006) 70-77.
[5] R.M. Karp, Reducibility among combinatorial problems, in: Complexity of Computer Computations, R.E. Miller and J.W. Thatcher, Eds., Plenum Press, New York (1972) 85-103.
[6] C.J. van der Heide, C. von den Bergh and E.C. van Ierland, Extending Weitzman's economic ranking of biodiversity protection: Combining ecological and genetic considerations, Ecol. Econ. 55 (2005) 218-223.
[7] B.R. Holland, Evolutionary analyses of large data sets: trees and beyond, PhD thesis, Massey University, New Zealand, 2001.
[8] T.L. Magnanti and L.A. Wolsey, Optimal Trees, in: Network Models, Handbook in Operations Research and Management Science, M.O. Ball et al., Eds., Amsterdam, North-Holland, (1995) 503-615.
[9] B. Mohar, Face covers and the genus problem for apex graphs, J. Combin. Theory Ser. B 82 (2001) 102-117.
[10] V. Moulton, C. Semple and M. Steel, Optimizing phylogenetic diversity under constraints, J. Theoret. Biol. 246 (2007) 186-194.
[11] F. Pardi and N. Goldmann, Species choice for comparative genomics: Being greedy works, PLoS Genetics 1 (2005) e71.
[12] N. Robertson, D. Sanders, P. Seymour and R. Thomas, A new proof of the four-colour theorem, Electron. Res. Announc. Amer. Math. Soc. 2 (1996) 17-25.
[13] A.S.L. Rodrigues, T.M. Brooks and K.J. Gaston, Integrating phylogenetic diversity in the selection of priority areas for conservation: Does it make a difference?, in: Phylogeny and Conservation, A. Purvis at al., Eds., Cambridge University Press, 2005.
[14] A.S.L. Rodrigues and K.J. Gaston, Maximising phylogenetic diversity in the selection of networks of conservation areas, Biological Conservation 105 (2002) 103-111.

[15] V.M. Sarich and A.C. Wilson, Immunological time scale for hominid evolution, Science 158 (1967) 1200-1203.
[16] C. Semple and M. Steel, Phylogenetics, Oxford University Press, 2003.
[17] A. Spillner, B. Nguyen and V. Moulton, Computing phylogenetic diversity for split systems, IEEE/ACM Computational Biology and Bioinformatics 5 (2008) 235-244.
[18] M. Steel, Phylogenetic diversity and the greedy algorithm, Syst. Biol. 54 (2005) 527-529.
[19] P.G. Tait, On the colouring of maps, Proceedings of the Royal Society of Edinburgh, Section A 10 (1980) 501-503.
[20] J.B. Whitfield and P.J. Lockhart, Deciphering ancient rapid radiations, Trends in Ecology and Evolution 22 (2007) 258-265.
[21] L. Witting, J. Tomiuk and V. Loeschcke, Modelling the optimal conservation of interacting species, Ecol. Modell. 125 (2000) 123-143.
[22] E. Zuckerkandl and L.B. Pauling, Molecular disease, evolution, and genetic heterogeneity, in: Horizons in Biochemistry, M. Kasha and B. Pullman, Eds., Academic Press, New York, 1962.

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: B.Faller@math.canterbury.ac.nz

Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: c.semple@math.canterbury.ac.nz

Merton College, University of Oxford, Oxford, United Kingdom

E-mail address: d.welsh@maths.ox.ac.uk