ON APPROXIMATION OF OPTIMIZING PHYLOGENETIC DIVERSITY FOR CLUSTER SYSTEMS

Beata Faller, Charles Semple and Mike Steel

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

Report Number: UCDMS2009/1 FEBRUARY 2009

ON APPROXIMATION OF OPTIMIZING PHYLOGENETIC DIVERSITY FOR CLUSTER SYSTEMS

BEÁTA FALLER, CHARLES SEMPLE, AND MIKE STEEL

ABSTRACT. A basic question in conservation biology is how to maximize future biodiversity as species face extinction. One way to approach this question is by measuring the diversity of a set of species in terms of the evolutionary history that those species span in a phylogenetic tree. Maximizing the resulting 'phylogenetic diversity' (PD) is one prominent selection criteria for deciding which species to conserve. The basic PD optimization problem aims to find a k-element subset of a given species set that has maximum PD among all such subsets. In this paper, we consider a generalization of this problem, which arises in situations where we do not know the true tree, or where evolution is not tree-like. We show that a greedy algorithm gives a $(1-e^{-1})$-approximation to the general PD optimization problem, and that there is no polynomial-time algorithm that achieves a better approximation ratio unless P=NP.

1. INTRODUCTION

A central task in conservation biology is measuring, predicting, and maximizing future biodiversity. There are numerous ways to approach this problem. In 1992, Faith proposed to measure the 'phylogenetic diversity' (PD) of species sets and to use this quantitative tool as a selection criteria for preserving biodiversity [5]. PD is based on evolutionary distances in a phylogenetic tree; given a subset of taxa, the phylogenetic diversity of that subset is the sum of the distances (or lengths) of the edges of the evolutionary tree that connects this subset. Here, the distance assigned to an edge may refer to the amount of genetic change on that edge, its temporal duration, or perhaps other features such as morphological diversity.

Given the phylogenetic tree of a species set X with lengths on its edges, in the basic PD optimization problem one selects a k-element subset of X that maximizes PD over all k-element subsets. Since this optimization problem assumes that the evolutionary history of the species in X is known, it cannot be used in situations where we do not know the true tree, or where evolution is not tree-like. In these cases, a more general biodiversity measure needs to be defined and, based on it, a more general optimization problem has to be formulated. Spillner et al. [13]

Date: February 23, 2009.

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

Key words and phrases. Combinatorial problem, approximation algorithm, phylogenetic diversity, cluster system.

We thank the New Zealand Marsden Fund and the Allan Wilson Centre for Molecular Ecology and Evolution for supporting this work.

introduced the measure 'phylogenetic diversity for split systems', which can be used when considering species whose evolution is better represented by an unrooted network rather than a tree. In this paper, we give a different generalization of PD: 'phylogenetic diversity for cluster systems' ($\mathrm{PD}_{\mathscr{C}}$). This measure is useful when the evolutionary history is best described by a rooted network [2], or when we do not know the true history, but we have a set of rooted trees (perhaps with different probabilities) and we want to maximize the expected PD. We consider the problem of finding a k-element subset of a given species set, that maximizes $\mathrm{PD}_{\mathscr{C}}$ over all such subsets. A related optimization problem, based on phylogenetic diversity for split systems, is defined and studied in [13].

The paper is organized as follows. In the next section, we state some necessary preliminaries, including the formal definition of $\mathrm{PD}_{\mathscr{C}}$ and an example of its applications, followed by the main result of the paper. Section 3 contains the proof of this result.

2. Main Result

A *rooted phylogenetic X-tree $\mathcal{T}$* is a rooted tree in which the root has degree at least 2 and all other interior vertices have degree at least 3, and whose leaf set is X. Such a tree is commonly used to represent the evolutionary history of a set of present-day species (the leaves) from their hypothetical common ancestor (the root). Let λ be a non-negative real-valued weighting on the edges of $\mathcal{T}$. For a subset Y of X, the *(rooted) phylogenetic diversity* of Y, denoted by $\mathrm{PD}_{\mathcal{T}}(Y)$ or more briefly $\mathrm{PD}(Y)$, is the sum of the edge lengths of the minimal subtree of $\mathcal{T}$ that connects the elements in Y and the root of $\mathcal{T}$ [5, 6].

The following definition generalizes the above described notion of PD. Let X be a finite set, and let $\mathscr{C}$ be a collection of subsets of X. Furthermore, let w be a weighting function on $\mathscr{C}$ that assigns a non-negative real-valued weight to each member of $\mathscr{C}$. For a subset A of X, we define the *phylogenetic diversity of A relative to $\mathscr{C}$*, denoted by $\mathrm{PD}_{\mathscr{C}}(A)$, as the sum of the weights of the members of $\mathscr{C}$ whose intersection with A is non-empty. That is, we set

$$\mathrm{PD}_{\mathscr{C}}(A) = \sum_{C \in \mathscr{C}:\ C \cap A \neq \emptyset} w(C). \tag{1}$$

To see that $\mathrm{PD}_{\mathscr{C}}$ generalizes PD, consider the special case when X is the leaf set of a rooted phylogenetic X-tree $\mathcal{T}$. A subset C of X is a *cluster of $\mathcal{T}$* if there is an edge that has precisely C as its set of descendant leaves. Suppose that the edges of $\mathcal{T}$ have non-negative real-valued weights and let $\mathscr{C}$ be the set of all clusters of $\mathcal{T}$. For a cluster $C \in \mathscr{C}$, let $w(C)$ be the weight of the (unique) edge of $\mathcal{T}$ whose associated cluster is C. It is easy to see that in this setting, the phylogenetic diversity of a subset A of X equals the phylogenetic diversity of A relative to $\mathscr{C}$. That is, for any $A \subseteq X$, we have $\mathrm{PD}_{\mathcal{T}}(A) = \mathrm{PD}_{\mathscr{C}}(A)$.

In this paper, we consider the general case when $\mathscr{C}$ is an arbitrary collection of subsets of a finite set. In the following, we define and feature an optimization problem that is based on $\mathrm{PD}_{\mathscr{C}}$.

Problem: OPTIMIZING PD FOR CLUSTER SYSTEMS
Instance: A finite set X, a collection $\mathscr{C}$ of subsets of X, a non-negative (real-valued) weighting w on $\mathscr{C}$, and a positive integer k.
Goal: Find a subset Y of X of size k that maximizes $\mathrm{PD}_{\mathscr{C}}$ amongst all such subsets.
Measure: The $\mathrm{PD}_{\mathscr{C}}$ score of Y.

In the case when $\mathscr{C}$ is the collection of clusters of a rooted phylogenetic X-tree, OPTIMIZING PD FOR CLUSTER SYSTEMS is just the basic PD optimization problem and is solvable in polynomial time using a greedy algorithm [10, 14]. However, as we will soon see, OPTIMIZING PD FOR CLUSTER SYSTEMS is NP-hard in general.

One of the reasons why we are interested in solving the above problem is highlighted in the following example.

Example. Let $X = \{a, b, c, d\}$ and consider the edge-weighted rooted phylogenetic X-trees shown in Fig. 1. Assume that we do not know the true evolutionary history of the species in X, but we know that either $\mathcal{T}_1$ or $\mathcal{T}_2$ represents it each with probability $1/2$, say. Consider the basic PD optimization problem with $k = 2$; that is, the problem of finding a 2-element subset of X that has maximum PD among all 2-element subsets. If $\mathcal{T}_1$ was the true tree, $\{a, c\}$ would be the optimal solution. However, if $\mathcal{T}_2$ was the true tree, the best 2-element subset would be $\{b, d\}$. In such a situation, it may be safer to choose a subset that maximizes the expected future PD; that is, a 2-element subset of X that has maximum expected PD among all such subsets. Let $W \subseteq X$ be of cardinality 2 and let $\mathbb{E}[\mathrm{PD}_{\mathcal{T}_1,\mathcal{T}_2}(W)]$ denote the expected PD of W with respect to the probability distribution on the two trees. Then, we have

$$\mathbb{E}[\mathrm{PD}_{\mathcal{T}_1,\mathcal{T}_2}(W)] = \frac{1}{2}\,\mathrm{PD}_{\mathcal{T}_1}(W) + \frac{1}{2}\,\mathrm{PD}_{\mathcal{T}_2}(W).$$

Consider now the cluster sets of $\mathcal{T}_1$ and $\mathcal{T}_2$. These are

$$\mathscr{C}_1 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{b, c\}, \{a, b, c\}\}$$

and

$$\mathscr{C}_2 = \{\{a\}, \{b\}, \{c\}, \{d\}, \{a, d\}, \{b, c\}\},$$

respectively. For $i \in \{1, 2\}$, let w_i assign to a cluster in $\mathscr{C}_i$ the weight of the edge corresponding to that cluster in $\mathcal{T}_i$. For example, $w_1(\{d\}) = 1$, $w_2(\{d\}) = 4$, and $w_2(\{a, d\}) = 2$. It is easy to see that $\mathbb{E}[\mathrm{PD}_{\mathcal{T}_1,\mathcal{T}_2}(W)]$ can be written as

$$\mathbb{E}[\mathrm{PD}_{\mathcal{T}_1,\mathcal{T}_2}(W)] = \sum_{C \in \mathscr{C}:\ C \cap W \neq \emptyset} w(C), \tag{2}$$

where $\mathscr{C} = \mathscr{C}_1 \cup \mathscr{C}_2$, $w(C) = \frac{1}{2}w_1(C)I_{C \in \mathscr{C}_1} + \frac{1}{2}w_2(C)I_{C \in \mathscr{C}_2}$, and $I_{C \in \mathscr{C}_i}$ takes the value 1 if $C \in \mathscr{C}_i$ and 0 otherwise. Since the right hand side of (2) is the $\mathrm{PD}_{\mathscr{C}}$ score of W under the above specified X, $\mathscr{C}$, and w, the problem of maximizing the expected PD is equivalent to solving OPTIMIZING PD FOR CLUSTER SYSTEMS for our particular instance. A quick check of all 2-element subsets of X shows that the

unique optimal solution is $\{b, c\}$; this subset has the highest expected PD among all 2-element subsets of X or, equivalently, it maximizes $\mathrm{PD}_{\mathscr{C}}$ over all such subsets.

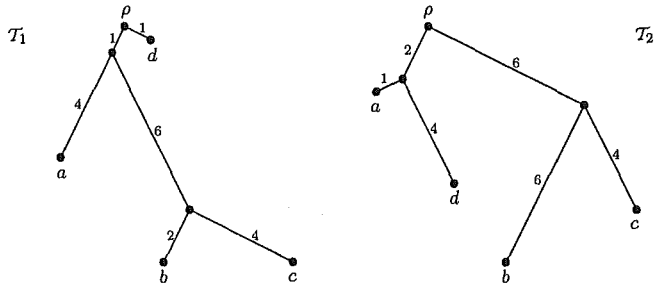

FIGURE 1. Two edge-weighted rooted phylogenetic X-trees $\mathcal{T}_1$ and $\mathcal{T}_2$, both rooted at ρ.

The process described in the example also works in general when we are given a finite set of rooted trees with an arbitrary probability distribution on them; maximizing expected PD always leads to an instance of OPTIMIZING PD FOR CLUSTER SYSTEMS. Furthermore, maximizing expected PD is equivalent to the problem WEIGHTED AVERAGE PD ON t (ROOTED) TREES. For $t = 2$, this problem is solvable in polynomial time [4]. For $t \geq 3$, it is NP-hard [13], and so OPTIMIZING PD FOR CLUSTER SYSTEMS is also NP-hard. However, the main result of this paper, Theorem 2.1, shows that there is a sharp approximation algorithm for it.

The main result of the paper is the following.

Theorem 2.1. OPTIMIZING PD FOR CLUSTER SYSTEMS *is an NP-hard optimization problem. However, it*

(i) *can be approximated by a polynomial-time greedy algorithm with approximation ratio* $1 - e^{-1}$*; and*
(ii) *cannot be approximated in polynomial time with an approximation ratio better than* $1 - e^{-1}$ *unless* $P = NP$.

The proof of (i) uses the fact that $\mathrm{PD}_{\mathscr{C}}$ is a submodular set function. The greedy algorithm that actually gives the above-named approximation ratio is described in [8] in a more general setting. We briefly outline the algorithm here in the language of this paper.

Algorithm: GREEDY$(X, \mathscr{C}, w, k)$
Input: A finite set X, a collection $\mathscr{C}$ of subsets of X, a non-negative (real-valued) weighting w on $\mathscr{C}$, and a positive integer k.
Output: A 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 and return S; otherwise, select an element z of $X - S$ that maximizes $\mathrm{PD}_{\mathscr{C}}(S \cup \{z\}) - \mathrm{PD}_{\mathscr{C}}(S)$ among all elements of $X - S$ (with ties settled arbitrarily).

Step 3 Set $S = S \cup \{z\}$ and $c = c + 1$, and return to Step 2.

GREEDY always produces a solution whose value is at least $1 - (1 - k^{-1})^k$ times the optimal value. This bound can be achieved for each k and has a limiting value of $1 - e^{-1}$ [8].

Remark. It would be interesting for future work to explore extensions to Theorem 2.1 (i) that allow costs to be assigned to the taxa. More precisely, suppose that each taxon has an associated positive real-valued cost associated with its conservation, and there is total budget B available to allocate. Then an extension to OPTIMIZING PD FOR CLUSTER SYSTEMS is to select a subset of taxa to conserve that maximizes the PD score subject to the constraint that the sum of the costs of the taxa conserved does not exceed the budget B (OPTIMIZING PD FOR CLUSTER SYSTEMS corresponds to the special case where all costs take the value 1). Recently, variations on the PD optimization problem on trees that allow taxon costs have allowed pseudo-polynomial time exact algorithms and polynomial-time approximation algorithms [3, 11].

3. PROOF OF THEOREM 2.1

We noted prior to the statement of Theorem 2.1 that OPTIMIZING PD FOR CLUSTER SYSTEMS is NP-hard. Thus, the rest of this section establishes parts (i) and (ii). To prove Theorem 2.1 (i), we first verify the following lemma.

Lemma 3.1. *Let $\mathscr{C}$ be a collection of subsets of a finite set X and let w be a non-negative real-valued weighting on the elements of $\mathscr{C}$. Then, $PD_{\mathscr{C}}$ is a submodular set function. That is, for any subsets A and B of X, we have*

$$PD_{\mathscr{C}}(A \cup B) + PD_{\mathscr{C}}(A \cap B) \leq PD_{\mathscr{C}}(A) + PD_{\mathscr{C}}(B). \tag{3}$$

Proof. Let A and B be arbitrary subsets of X. Apply Eqn. (1) to $A, B, A \cup B$ and $A \cap B$, and partition $\mathscr{C}$ into three sets as follows. For $i \in \{0, 1, 2\}$, let $\mathscr{C}_i$ consist of subsets in $\mathscr{C}$ whose intersection is non-empty with exactly i sets in $\{A, B\}$. Consider now the following cases. For a subset $C \in \mathscr{C}_0$, the weight $w(C)$ affects neither side of (3). For $C \in \mathscr{C}_1$, $w(C)$ appears exactly once on both sides of (3). Finally, for $C \in \mathscr{C}_2$, $w(C)$ appears exactly twice on the right hand side and at most twice on the left hand side. Noting that w is non-negative completes the proof. □

Proof of Theorem 2.1 (i). It is shown in [8] that a greedy heuristic can be used to approximate the following problem with approximation ratio $1 - e^{-1}$. Let S be a finite set and z be a real-valued function defined on the power set of S. Assume that z is submodular and non-decreasing and that $z(\emptyset) = 0$. The problem is to find a subset of S of size at most k that maximizes z amongst all such subsets. We complete the proof by showing that OPTIMIZING PD FOR CLUSTER SYSTEMS is a special case of this problem. Take X as the finite set and $\mathrm{PD}_{\mathscr{C}}$ as the real-valued function on the power set of X. That is, set $S = X$ and $z = \mathrm{PD}_{\mathscr{C}}$. By Lemma 3.1, $\mathrm{PD}_{\mathscr{C}}$ is submodular. It is easy to see that $\mathrm{PD}_{\mathscr{C}}$ is also non-decreasing: for any subset

A of X and for any element a in $X - A$, we have $\mathrm{PD}_{\mathscr{C}}(A \cup \{a\}) - \mathrm{PD}_{\mathscr{C}}(A) \geq 0$. Finally, $\mathrm{PD}_{\mathscr{C}}(\emptyset) = 0$. Theorem 2.1 (i) now follows. □

Before proving Theorem 2.1 (ii), we formally state the problem MAX k-COVER and the definition of a type of approximability preserving reduction, called L-reduction.

Problem: MAX k-COVER
Instance: A finite set $S = \{s_1, \ldots, s_n\}$, a collection $\mathscr{F}$ of subsets of S, and a positive integer k.
Goal: Find a subset $\mathscr{F}' = \{F_1, \ldots, F_k\}$ of $\mathscr{F}$ of size k that maximizes the size of the set $\cup_{i=1}^{k} F_i$.
Measure: The cardinality of $\cup_{i=1}^{k} F_i$.

Feige [7] showed that no polynomial-time approximation algorithm for MAX k-COVER can have an approximation ratio better than $1 - e^{-1}$ unless $P = NP$.

Let Π_1 and Π_2 be two arbitrary optimization problems. An *L-reduction* [1, 9] from Π_1 to Π_2 is a pair of polynomial-time computable functions f and g, and a pair of positive constants α and β that satisfy the following properties:

(I) If I is an instance of Π_1, then $f(I)$ is an instance of Π_2 with
$$\mathrm{opt}(f(I)) \leq \alpha\, \mathrm{opt}(I),$$
where $\mathrm{opt}(I)$ and $\mathrm{opt}(f(I))$ denote the size of an optimal solution to I and $f(I)$, respectively.

(II) If S is a feasible solution to $f(I)$, then $g(S)$ is a feasible solution to I with
$$|\,\mathrm{opt}(I) - c(g(S))| \leq \beta\,|\,\mathrm{opt}(f(I)) - c(S)|,$$
where $c(g(S))$ and $c(S)$ is the size of $g(S)$ and S, respectively.

It follows from the definition that if Π_1 L-reduces to Π_2, and there is a polynomial-time approximation algorithm for Π_2 with approximation ratio ϵ, then there is a polynomial-time approximation algorithm for Π_1 with approximation ratio $\alpha\beta\epsilon$ [9].

Proof of Theorem 2.1 (ii). We prove (ii) by giving an L-reduction with $\alpha = \beta = 1$ from MAX k-COVER to OPTIMIZING PD FOR CLUSTER SYSTEMS. By the previous remarks on MAX k-COVER and L-reduction, this will imply that OPTIMIZING PD FOR CLUSTER SYSTEMS cannot be approximated in polynomial time with an approximation ratio better than $1 - e^{-1}$ unless $P = NP$, as required.

Let I be an instance of MAX k-COVER, and let R be an equivalence relation on S defined as follows. Two elements s_i and s_j of S are *equivalent* if and only if they are elements of precisely the same subsets in $\mathscr{F}$; that is, they satisfy $s_i \in F \Leftrightarrow s_j \in F$, for all F in $\mathscr{F}$. Let $[s_i]$ denote the equivalence class of $s_i \in S$ under R. We now give a function f that constructs from I an instance $f(I)$ of OPTIMIZING PD FOR CLUSTER SYSTEMS; that is, it specifies a set, a collection of subsets of this set, a non-negative real-valued weight assigned to each subset in the collection, and a positive integer. Let $\mathscr{F}$ be the set and let $\mathscr{C}$ be the collection of subsets of $\mathscr{F}$ be

defined as follows. For each equivalence class $[s_i]$ under R, there is a unique member $C_{[s_i]} = \{F \in \mathscr{F} : s_i \in F\}$ of $\mathscr{C}$. Let the weight of $C_{[s_i]} \in \mathscr{C}$ be the cardinality of the equivalence class $[s_i]$. Furthermore, let the positive integer in instance $f(I)$ equal k. Clearly, this construction can be accomplished in polynomial time.

To prove (I), we show that $\text{opt}(I) = \text{opt}(f(I))$, and so $\alpha = 1$. Suppose that $\mathscr{F}' = \{F_1, \ldots, F_k\}$ is an optimal solution to I. Then $\text{opt}(I) = |\cup_{j=1}^{k} F_j|$. Trivially, $\mathscr{F}'$ is a feasible solution to $f(I)$. Moreover, $\mathscr{F}' \cap C_{[s_i]}$ is non-empty precisely if $\cup_{j=1}^{k} F_j$ contains s_i, in which case $[s_i] \subseteq \cup_{j=1}^{k} F_j$. By the choice of weighting, it now follows that $\text{PD}_{\mathscr{C}}(\mathscr{F}') = |\cup_{j=1}^{k} F_j|$, and so $\text{opt}(I) \leq \text{opt}(f(I))$. By choosing an optimal solution to $f(I)$ and reversing this argument, it is also straightforward to show that $\text{opt}(I) \geq \text{opt}(f(I))$, as required.

For (II), let $\mathscr{F}'' = \{F^1, \ldots, F^k\}$ be a feasible solution to $f(I)$. Setting $g(\mathscr{F}'') = \mathscr{F}''$ gives a feasible solution to I with $c(g(\mathscr{F}'')) = c(\mathscr{F}'') = |\cup_{j=1}^{k} F^j|$. This can be seen by arguments similar to those used in the proof of (I). Trivially, g is computable in polynomial time. Thus, (II) is satisfied with $\beta = 1$. This completes the proof of Theorem 2.1 (ii). □

REFERENCES

[1] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi, Complexity and Approximation, Spronger, Berlin (1999).

[2] M. Baroni, C. Semple, and M. Steel, A framework for representing reticulate evolution, Annals of Combinatorics 8 (2004), pp. 391-408.

[3] M. Bordewich and C. Semple, Nature reserve selection problem: a tight approximation algorithm, IEEE/ACM Transactions on Computational Biology and Bioinformatics 5 (2008), pp. 275-280.

[4] M. Bordewich, C. Semple, and A. Spillner, Optimizing phylogenetic diversity across two trees, Applied Mathematics Letters, in press.

[5] D.P. Faith, Conservation evaluation and phylogenetic diversity, Biol. Conserv. 61 (1992), pp. 1-10.

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

[7] U. Feige, A threshold of $\ln n$ for approximating Set Cover, J. ACM 45 (1998), pp. 634-652.

[8] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher, An analysis of approximations for maximizing submodular set functions - I, Mathematical Programming 14 (1978), pp. 265-294.

[9] C.H. Papadimitriou and M. Yannakakis, Optimization, approximation, and complexity classes, J. Comput. System Sci. 43 (1991), pp. 425-440.

[10] F. Pardi and N. Goldmann, Species choice for comparative genomics: Being greedy works, PLoS Genetics 1 (2005), pp. 371.

[11] F. Pardi and N. Goldmann, Resource aware taxon selection for maximizing phylogenetic diversity, Syst. Biol. 56 (2007), pp. 431-444.

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

[13] A. Spillner, B. Nguyen, and V. Moulton, Computing phylogenetic diversity for split systems, IEEE/ACM Computational Biology and Bioinformatics 5 (2008), pp. 235-244.

[14] M. Steel, Phylogenetic diversity and the greedy algorithm, Syst. Biol. 54 (2005), pp. 527-529.

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

BIOMATHEMATICS RESEARCH CENTRE, DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

E-mail address: M.Steel@math.canterbury.ac.nz