Selecting Taxa to Save or Sequence: Desirable Criteria and a Greedy Solution

Lam Si Tung Ho

December 14, 2012

1 Introduction

One important question in conservation biology is determining which collection of evolutionary units (EUs; including species or higher level taxa) should be conserved. The method of selection EUs is expected to have the following properties: *spread*, *stability*, and *applicability*.

The most popular method for this task is maximizing phylogenetic diversity (PD) [2]. The PD of a set of EUs is defined as the total length of the phylogenetic tree that connects these EUs. However, in some cases, this method does not choose a best set of EUs for conservation. For example, we consider the tree with 6 species in figure 1. Assume that we want to choose 3 species to conserve. The maximizing PD method will select 3 species A, B and F. However, this selection seems wrong because A and B are closely related. A more intuition choice is A, D and F. This choice is more intuition because the chosen species are more spread out.

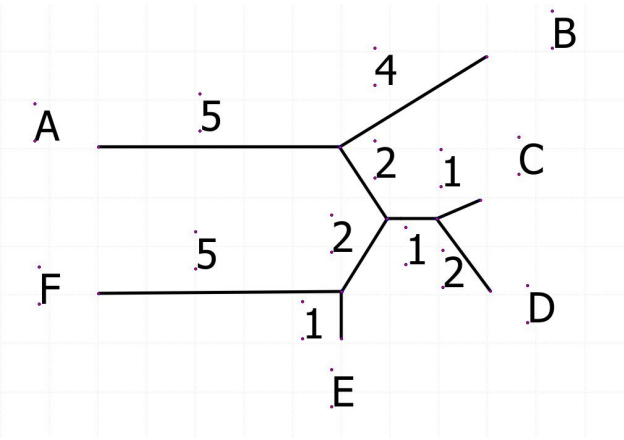

Figure 1: A phylogenetic tree on which maximizing PD method does not choose the best set of 3 species.

In the paper [1], the authors propose a more intuition method which chooses a set of EUs that maximizes the minimum phylogenetic distance between any pair of EUs in the set. The method is called Maximize Minimum Distance (MMD) and only requires to know the distance δ between EUs. Note that MMD method will choose A,D and F as the best 3 species to conserve in the previous example. However, the problem MMD is NP-hard. Hence, the authors suggest the greedy

algorithm GREEDYMMD which is a sequential procedure. At each step, it selects an EU from those who are not already included such that the minimum distance from it to the already chosen ones is maximum. The discussion about the properties: spread, stability, and applicability of GREEDYMMD method compare to maximizing PD method can be found in [1].

2 The performance of GREEDYMMD algorithm

The authors shows that the GREEDYMMD method is a 2-approximation algorithm to the MMD method.

Theorem 1 *Let δ be the distance on X, and suppose that δ satisfies the triangle inequality. Let k be an integer greater than one and let S_k be the set returned by GREEDYMMD(δ,k). Then MD(S_k) is a 2-approximation to MD(Y_{opt}), where Y_{opt} is an optimal solution of size k to MMD.*

To prove theorem 1, we need the following lemma

Lemma 2 *For any element $x \in X \setminus S_{k-1}$, we have $MD(S_{k-1} \cup \{x\}) = \delta(x, s)$ for some $s \in S_{k-1}$.*

Proof of theorem 1. Denote $S_i = \{s_1, s_2, \cdots, s_i\}$ for $i = 1, 2, \cdots, k$. Let $y \in Y_{opt} \setminus S_{k-1}$ such that

$$MD(S_{k-1} \cup \{y\}) = \max_{y' \in Y_{opt} \setminus S_{k-1}} [MD(S_{k-1} \cup \{y'\})].$$

Assign each element of Y_{opt} to the element in S_{k-1} that it is closet to under δ. Note that $|Y_{opt}| > |S_{k-1}|$ and if $|Y_{opt} \setminus S_{k-1}| = 1$, then by lemma 2 we have $MD(Y_{opt}) = MD(S_k)$. Therefore, there exists two distinct elements $y_u, y_v \in Y_{opt} \setminus S_{k-1}$ are assigned to the same element $s \in S_{k-1}$. By lemma 2, we have

$$\begin{aligned} MD(Y_{opt}) &\leq \delta(y_u, y_v) \leq \delta(y_u, s) + \delta(y_v, s) \\ &= MD(S_{k-1} \cup \{y_u\}) + MD(S_{k-1} \cup \{y_v\}) \\ &\leq 2MD(S_{k-1} \cup \{y\}) \leq 2MD(S_k). \end{aligned}$$

Proof of lemma 2. Assume there exists $i < j < k$ such that $MD(S_{k-1} \cup \{x\}) = \delta(s_i, s_j) < \delta(x, s)$. If there is more than one pair s_i, s_j, choose a pair with minimal j. So, $MD(S_{j-1}) > \delta(s_i, s_j)$. Then

$$MD(S_{j-1} \cup \{x\}) = \min[MD(S_{j-1}), \min_{s \in S_{j-1}} \delta(x, s)] > \delta(s_i, s_j) \geq MD(S_j)$$

which is a contradiction.

References

[1] M. Bordewich, A.G. Rodrigo, and C. Semple. Selecting taxa to save or sequence: desirable criteria and a greedy solution. *Systematic biology*, 57(6):825–834, 2008.

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