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

Magnus Bordewich[1], Allen G. Rodrigo[2,4], and Charles Semple[3]

[1] *Department of Computer Science, Durham University, Durham DH1 3LE, United Kingdom*

*E-mail: m.j.r.bordewich@durham.ac.uk*

[2] *Bioinformatics Institute and the Allan Wilson Centre for Molecular Ecology and Evolution, University of Auckland, Auckland, New Zealand*

*E-mail: a.rodrigo@auckland.ac.nz*

[3] *Biomathematics Research Centre, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand*

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

[4] *Corresponding author, Telephone +64 09 373 7599, Fax +64 09 367 7136*

*Abstract.*— Three desirable properties for any method of selecting a subset of evolutionary units (EUs) for conservation or for genomic sequencing are discussed. These properties are: *spread, stability,* and *applicability.* We are motivated by a practical case in which the maximization of phylogenetic diversity (PD), which has been suggested as a suitable method, appears to lead to counter-intuitive collections of EUs and does not meet these three criteria. We define a simple greedy algorithm (GreedyMMD) as a close approximation to choosing the subset that maximizes the minimum pairwise distance (MMD) between EUs. GreedyMMD satisfies our three criteria and may be a useful alternative to PD in real world situations. In particular, we show that this method of selection is suitable under a model of biodiversity in which features arise and/or disappear during evolution. We also show that if distances between EUs satisfy the ultrametric condition, then GreedyMMD delivers an *optimal* subset of EUs that maximizes both the minimum pairwise distance and the PD. Finally, since GreedyMMD works with distances and does not require a tree, it is readily applicable to many datasets.

*Key words.* biodiversity conservation, phylogenetic diversity, greedy algorithm

Quantitative methods based on biodiversity have received much attention recently in conservation biology (e.g. Rodrigues and Gaston, 2002; Pardi and Goldman, 2005; Rodrigues et al., 2005; Steel, 2005; Moulton et al., 2007; Spillner et al., 2007; Bordewich and Semple, 2008). These methods are used for determining which collection of evolutionary units (EUs, including species or other higher-level taxa) should be conserved. Maximizing phylogenetic diversity (PD) has emerged as a leading criterion in this regard. (For a set of EUs and a given phylogeny, the PD of the set is defined as the total length of the minimal phylogenetic subtree that connects the EUs in the set.) In its most direct application to conservation (Faith, 1992), we select a $k$-element subset of EUs that maximizes PD over all $k$-element subsets. The use of PD has also been advocated when decisions need to be made about which genomes should be sequenced (Pardi and Goldman, 2005).

However, there are instances when maximizing PD appears to be the wrong criterion to use when selecting a set of EUs for conservation, or deciding which genomes should be sequenced. Perhaps the best example is the iconic small-subunit ribosomal RNA tree of the Bacteria, Archaea, and Eucarya first described in Woese (1987), and reproduced in Figure 1. Suppose that the aim is to select three of these EUs for sequencing. A subset of three EUs chosen to maximize PD will include two eukaryotes and one bacterium, as shown in bold in Figure 1a. Most people would agree that this seems wrong; after all, the second eukaryote may add comparatively little information. In contrast, selecting a bacterium, an archaeon, and an eukaryote (as shown in bold in Fig. 1b) is a choice that few would disagree with; it seems intuitively better. If one were asked to explain why this is more intuitive, one might reply that the chosen EUs are more “spread out” on the phylogenetic tree, and are likely to contain less redundant genetic information (measured, say, by the number of genes in common).Of course, this example is somewhat contrived, but it serves to highlight the fact that maximizing PD is not always the best selection criterion for biodiversity. Since PD does notalways capture what we want from a measure of diversity, we propose an alternative.

A quantitative measure of diversity captures a specific notion of biodiversity under some model of evolution. A reasonable interpretation of PD is that it measures the expected number of different ‘features’ exhibited by the EUs in the selected set, under a model which makes three implicit assumptions about evolution. First, features arise in a homoplasy-free way, second, the length of any branch represents the number of features arising along that branch, and third, once a feature arises in the phylogeny, it persists forever and is present in all descendant EUs. For a good exposition of this interpretation, see Faith (1994). Thus, in the above example, PD chooses an additional eukaryote instead of an archaeon, since a taxon connected near the root of the phylogeny by a short branch is assumed to contain almost exclusively features that are shared with every other taxon in the phylogeny. Being on a short branch means that a very small number of features have arisen since the taxon split off from the common ancestor whose features are shared by all.

Our intuition that an archaeon will add better diversity value simply because it is some distance away in the phylogeny from bacteria and eukaryotes suggests the above assumptions are not always valid. If one assumes that features do not persist forever, but disappear from a lineage at a random time, it transpires that to obtain an intuitively appropriate solution one needs to choose the set of EUs that maximizes the minimum phylogenetic distance between any pair of EUs in the set. Defined formally in the next section, we refer to this selection criterion as MMD (Maximize Minimum-Distance). If MMD is applied to choose three of the EUs in Figure 1, then a desired set of one bacterium, one archaeon, and one eukaryote is obtained. Details of the extended model and corresponding measure of diversity under which MMD is being used as a selection criteria are given later in the paper. As we will show, although MMD is a heuristic for maximizing this measure, it has a strong theoretical justification. We remark here that Faith (1994) described a similar model of diversity based upon features arising and disappearing.

The above discussion prompts us to consider what other properties we want from our chosen method of selecting EUs to conserve or sequence. We have already mentioned the following property:

(1) *Spread.* The selected EUs have close to maximum diversity under some model, achieved by being 'spread out' in some sense on the phylogenetic tree.

However, we would also like the selected set to have the following properties:

(2) *Stability.* The core of the solution set should be stable as the set size fluctuates.

(3) *Applicability.* The method of selection should be able to be applied to as general an input as possible.

We expand onthe latter two properties and why they are desirable below.

### *Stability*

The amount of money available for conservation projects fluctuates with the political climate and with other social and economic factors. The set of species targeted for conservation or set of genomes chosen to sequence needs to be stable as budgets vary. Suppose that you are told that there is sufficient money to conserve $k$ species, or sequence $k$ genomes. Being a good phylogeneticist, you apply a diversity-based method to choose your $k$-element subset of EUs. Your project gets underway, but after a year the budget available is adjusted up or down. Clearly, it is extremely desirable that the set of EUs selected under the new budget is closely related to the original set of EUs chosen. If the number of EUs can be increased, you would like the $k$ EUs previously chosen, and partially worked on, to be guaranteed to be in the larger set of EUs chosen by the algorithm you have used for selection. If the number of EUs is to be reduced, you would like the new solution set to be a subset of the EUs you have already started conserving or sequencing. This notion of stability is captured by a *greedy algorithm*: selecting the pair of EUs that maximizes (pairwise) a measure of diversity, and then iteratively selecting an EU to add to the set that gives the best measure amongst all single EU increments to the existing set, until the desired number of EUs is reached. Therefore, any increase in the size of the chosen set is made by simply adding more EUs to the existing set, whilst any decrease may be made by removing the EUs added last. Thus we see that a greedy selection criteria exactly captures the idea of stability that we are after for our method of selection.

### *Applicability*

We would like to be able to apply the selection method to many varied data sets. It is not always the case that an accurate tree with known branch lengths will be available. However, as a minimum, it seems reasonable to expect that a matrix of pairwise distances between EUs will be available, or at least readily estimated from sequences (e.g. using Hamming distance, see Hamming, 1950). Therefore we would like our method and measure of diversity to apply in this situation, without having to go to the computational expense and loss of information involved in estimating a phylogeny first.

We now consider under what conditions PD and MMD meet these three criteria. Most significant is that PD and MMD both meet the spread criterion, but only under different assumptions. As discussed above, under a model of evolution in which features may both arise and disappear MMD captures the notion of spread that we require, whereas PD is appropriate for a model in which features persist indefinitely. With regard to applicability, since MMD depends only on the pairwise distances between EUs, it can be applied to a distance matrix without reference to any underlying tree. Thus MMD satisfies the applicability criteria. While PD can be computed using a character matrix, a distance matrix could be more problematic for computing PD. If the distances fit a tree metric, then, using the ideas in Faith (1992), it is possible to compute PD from such a matrix. But if the distance matrix does not fit a tree metric, then it is not clear how to compute PD without first estimating a phylogeny.

For stability, the situation is reversed. An optimal set under the PD measure is generated by a greedy algorithm (Pardi and Goldman, 2005; Steel, 2005), while under MMD being greedy does not always return the optimal solution (see Example 3). Hence MMD will not necessarily have the stability property, whereas PD does. Indeed, after an increase of one EU, the revised optimal subset of EUs under MMD could exclude some of the original EUs you had selected (see Example 2). Thus, depending on the model of evolution appropriate to the circumstances, neither MMD nor PD will always satisfy all three criteria. In order to satisfy all three under our model in which features may disappear, we propose a greedy approach to MMD. This method, called GreedyMMD, gains stability and still maintains a good, though not necessarily optimal, solution to MMD, thus retaining spread and applicability.

Provided the pairwise distances satisfy the triangle inequality, it has previously been shown in the context of operations research that GreedyMMD gives a 2-approximation to the optimal MMD solution (Tamir, 1991; Ravi et al., 1994). Observe that this includes the case that the distances fit a tree metric. Noting that it is NP-hard to compute an optimal set under MMD if the distance matrix satisfies the triangle inequality (but is not necessarily a tree metric) (Ravi et al., 1994), we show via a simple example that this approximation is sharp even if the distance matrix is a tree metric. In addition, we show that if the pairwise distances satisfy the ultrametric condition, then GreedyMMD returns an *optimal* set of EUs under MMD and, moreover, this set also maximizes PD. We conclude that the selection method GreedyMMD satisfies the three criteria of spread, stability and applicability, and advocate it as a useful and practical alternative to PD in certain real-world situations.

The paper is organized as follows. The next two sections gives a formal definition of MMD and describe the model under which MMD is an optimal criterion. We then present GreedyMMD and consider how well it performs as an approximation algorithm to MMD if the pairwise distances satisfy the triangle inequality, and if it is applied to a set of pairwise distances that satisfy the ultrametric condition. We end the paper by considering taxonomic distinctness, a measure for diversity with similarities to MMD, and $p$-median diversity.

We end the introduction with a remark on the previous use of MMD in computational biology. The idea of choosing a set of EUs using the criterion of MMD has been considered by Holland (2001) in the context of selecting representative model strains. This criterion was one of several such criteria looked at in her thesis. However, while GreedyMMD was also considered as an efficient way of selecting EUs under MMD, the thesis did not explore any theoretical results or the model for which MMD is an optimal criterion.

## Definitions and Problem Specification

In this section, we detail some definitions that are used throughout the paper and formally describe MMD. An (unrooted) *phylogenetic $X$-tree* is a tree with no degree-two vertices and whose leaf set is $X$. A *rooted phylogenetic $X$-tree* is a rooted tree with no degree-two vertices except the root which may have degree two and whose leaf set is $X$. For the purposes of this paper, we will assume that all the edges of a (rooted or unrooted) phylogenetic tree are assigned non-negative real-valued lengths.

For a set $X$, a *distance* on $X$ is a function $\delta$ that assigns a non-negative real value to each ordered pair in $X \times X$ such that, for all $x, y \in X$, we have $\delta(x, x) = 0$ and $\delta(x, y) = \delta(y, x)$. A distance is said to satisfy the *triangle inequality* if, for all $x, y, z \in X$,

$$\delta(x, z) \leq \delta(x, y) + \delta(y, z).$$

A natural way to obtain a distance on $X$ is from a phylogenetic $X$-tree. In particular, a distance $\delta_{\mathcal{T}}$ on $X$ can be obtained by setting $\delta_{\mathcal{T}}(x, y)$ to be the sum of the edge-lengths on the (unique) path from $x$ to $y$ for all $x, y \in X$. Distances that can be realized via a phylogenetic tree in this way are known as *tree metrics*. Such metrics satisfy the triangle inequality.

A tree metric $\delta_{\mathcal{T}}$ on $X$ is an *ultrametric* if it can be realized by a rooted phylogenetic $X$-tree $\mathcal{T}$ such that, for all $x, y \in X$,

$$\delta_{\mathcal{T}}(\rho, x) = \delta_{\mathcal{T}}(\rho, y), \tag{1}$$

where $\rho$ is the root of $\mathcal{T}$. The equality in (1) means that all leaves are equidistant from the root. Equivalently, for an arbitrary distance $\delta$ on $X$, $\delta$ is an ultrametric precisely if, for every three distinct elements $x, y, z \in X$,

$$\delta(x, y) \leq \max\{\delta(x, z), \delta(y, z)\}.$$

For this equivalence, see Semple and Steel (2003).

Let $\delta$ be a distance on $X$. For a subset $S$ of $X$, we define the *minimum distance* of $S$ to be

$$MD(S) = \min\{\delta(x, y) : x, y \in S\}.$$

For a given positive integer $k$, the problem MMD is to find a subset $S$ of $X$ that maximizes $MD(S)$ amongst all subsets of $X$ of size $k$. Intuitively, such a set $S$ corresponds to selecting a $k$-element subset of $X$ in which each pair of elements is as 'far apart' as possible under $\delta$.

As noted earlier, whereas it is usual to think of phylogenetic diversity as a criterion that requires a phylogenetic tree as a basis for implementation, MMD is more general. In particular, the criterion MMD can be applied to any distance measure.

## The Model of Diversity

In this section, we describe the model for which MMD is an appropriate criterion for optimizing biodiversity. Under PD, one assumes that 'features' arise during evolution at a constant rate—for two points $u$ and $v$ on a rooted phylogenetic $X$-tree $\mathcal{T}$ with $v$ an ancestor of $u$, the distance from $v$ to $u$ represents the number of new features that arose on the evolutionary path from $v$ to $u$. Thus, for every unit of distance, a new feature arises. It is important to note that, under this model, any feature arising at a point $v$ on $\mathcal{T}$ is present at all points descendant from $v$. Now consider the model where, in addition to features arising in this way, features have a constant probability of disappearing on every evolutionary path in $\mathcal{T}$ on which they are present. In mathematical terms, features disappear according to an exponential distribution with rate $\lambda$ independently on any branch. In particular, once a feature is present, it has a constant and memory-less probability $e^{-\lambda}$ of surviving in each time step. So that there is a full set of features available at the beginning, we also assume that the root of $\mathcal{T}$ is on an infinitely long branch connected to its first branching point.

For a subset $A$ of $X$, the number of features present at at least one EU in $A$ is a random variable $F_A$. We next consider the expected value of $F_A$. If $A$ is the singleton $\{a\}$, then the expected value of $F_{\{a\}}$ is the sum over all points on the path from the root to $a$ of the probability that the feature arising at that moment is still present at $a$. In particular,

$$\mathbb{E}(F_{\{a\}}) = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}.$$

If $A$ consists of two EUs $a$ and $b$ whose distances from their most recent common ancestor are $d_a$ and $d_b$, respectively, then, although individually $\{a\}$ and $\{b\}$ each have $1/\lambda$ expected features, the expected value of $F_{\{a,b\}}$ is

$$\begin{aligned}\mathbb{E}(F_{\{a,b\}}) &= \int_0^{d_a} e^{-\lambda x}dx + \int_0^{d_b} e^{-\lambda x}dx + \int_0^{\infty} e^{-\lambda x}(e^{-\lambda d_a} + e^{-\lambda d_b} - e^{-\lambda(d_a+d_b)})dx \\ &= \frac{1}{\lambda}(2 - e^{-\lambda(d_a+d_b)}).\end{aligned}$$

The above calculations can be extended using the principle of inclusion/exclusion to any size subset of $X$. For example, if $A$ consists of three EUs $a$, $b$, and $c$ whose pendent branch lengths are $d_a$, $d_b$, $d_c$, and whose internal branch length from the most recent common ancestor of $a$ and $b$ to the most recent common ancestor of all three EUs is $d_{ab}$, then the expected value of $F_{\{a,b,c\}}$ is the contribution of the three independent EUs, minus the expected number of pairwise common features, plus the expected number of features common to all three EUs:

$$(2) \qquad \begin{aligned}\mathbb{E}(F_{\{a,b,c\}}) = \frac{1}{\lambda}(3 - e^{-\lambda(d_a+d_b)} - e^{-\lambda(d_a+d_{ab}+d_c)} \\ - e^{-\lambda(d_b+d_{ab}+d_c)} + e^{-\lambda(d_a+d_b+d_{ab}+d_c)})\end{aligned}$$

Although the full inclusion/exclusion formula for $k$ EUs has exponentially many terms (in $k$), it is still possible to calculate the expected number of observed features for a subset of $X$ of size $k$ efficiently on a rooted binary phylogenetic tree. This uses the independence of distinct branches and a bottom-up calculation.

Returning to when $A$ consists of three EUs $a$, $b$, and $c$, observe that if $\lambda$ is very small, then we can approximate $e^{-\lambda m}$ with $(1 - \lambda m)$ for all $0 \leq m \ll 1/\lambda$. With this assumption, Equation (2) simplifies to

$$\mathbb{E}(F_{\{a,b,c\}}) \approx \frac{1}{\lambda} + d_a + d_b + d_{ab} + d_c.$$

Recalling that the root is incident with an infinitely long branch, the term $1/\lambda$ is the contribution of the features common to all EUs and may be ignored. It follows that as $\lambda$ tends to zero, the contribution of the selected EUs to the expected number of features tends to $PD(\{a,b,c\})$. This should be expected since zero $\lambda$ corresponds to all features persisting throughout the phylogeny, which is the assumption of PD. This relationship extends to arbitrary sized sets of EUs, and was observed in Faith (1994) under their slightly different model.

On the other hand, if $\lambda$ is very large, so features die out reasonably quickly, then the exponential terms in Equation (2) will be very small. This corresponds to very few features being common to more than one EU. In Faith (1994), it is remarked that if $\lambda$ "is so large that all features which arise are lost within one unit step, then all species are of equal status as there is no predictable redundancy among them," again using a slightly different model. This is sometimes termed 'species richness'. However, before $\lambda$ becomes so large that this happens, we can do better than regarding all exponential terms to be zero. For a subset $S$ of $X$ consisting of $k$ EUs, the inclusion/exclusion formula for the expected number of features of $S$ satisfies

$$\frac{1}{\lambda}\left(k - \sum_{a,b\in S} e^{-\lambda d(a,b)}\right) \leq \mathbb{E}(F_S) \leq \frac{1}{\lambda}\left(k - \sum_{a,b\in S} e^{-\lambda d(a,b)} + \sum_{a,b,c\in S} e^{-\lambda PD(\{a,b,c\})}\right), \tag{3}$$

where $d(a,b)$ is the distance in $\mathcal{T}$ between $a$ and $b$, and $PD(\{a,b,c\})$ is simply a convenient way to write the sum of the edge weights of the minimal subtree of

$\mathcal{T}$ connecting $a, b, c$. In both the upper and lower bound, the largest term is $k/\lambda$ and the second largest term is $e^{-\lambda d'}/\lambda$, where $d'$ is the distance in $\mathcal{T}$ between the closest pair of EUs in $S$. Assuming that in a real-world situation no two pairs of EUs are *exactly* the same distance apart, it follows that, as $\lambda$ becomes large, these two terms will dominate the others, and so the expected value of $F_S$ approaches $\frac{1}{\lambda}(k - e^{-\lambda d'})$. Thus, if $\lambda$ is large, then the most important thing to do in order to maximize the number of observed features in a set of EUs of fixed size is to maximize the minimum distance in $\mathcal{T}$ between the closest pair of selected EUs. In other words, select a subset of EUs that optimizes the MMD criterion. This is illustrated in the following example.

***Example* 1.** Consider the rooted phylogenetic tree shown in Figure 2. Suppose we have already selected EUs $a$ and $c$, and we are now deciding between $b_1$ and $b_2$ for the third EU. In particular, which of the expected values $E(F_{\{a,c,b_1\}})$ and $E(F_{\{a,c,b_2\}})$ given by Equation (2) is bigger? (Note that, under PD, $b_1$ would be chosen, while, under MMD, $b_2$ would be chosen.) If $\lambda = 0.4$, then choosing $b_1$ would give a diversity measure of 5.19, whereas choosing $b_2$ would give 7.43, a gain of over 43%.

As discussed prior to the example, if $\lambda$ is small enough, then PD *will* select the EU that maximizes Equation (2). For this to happen in this example, we would need $\lambda < 0.00047$ for $b_1$ to maximize Equation (2). It was also mentioned above that, for $\lambda$ large enough, you might as well choose any three EUs, since each will contribute a roughly independent set of features. For the difference in Equation (2) of the optimal set of three EUs to a choice of any three EUs to be within 5% and 1%, we would require $\lambda > 9.72$ and $\lambda > 17.6$, respectively. Thus the range of $\lambda$ for which selecting a third EU under MMD would be a significantly more accurate choice than either PD or an arbitrary selection is large—features disappearing between about 10 times faster than they arise and 2000 times slower.

Thus the measure of diversity that we wish to maximize is the expected number of features given by the inclusion/exclusion formula. However, it seems unlikely that one can find a simple and efficient algorithm to select a set of EUs that maximizes this expectation. Thus we must use some form of simplification to proceed. One possible simplification, as taken by PD, is to assume $\lambda$ is zero, or at least extremely small. We do not think this assumption is valid in all circumstances, and so consider a simplification based upon the dominant terms if $\lambda$ is large. One might consider whether it is possible to minimize the sum $\sum_{a,b \in S} e^{-\lambda d(a,b)}$ given in both upper and lower bounds of Equation (3), rather than minimizing the single term $\max_{a,b \in S} e^{-\lambda d(a,b)}$. This also seems very difficult to do in practice. In particular, this is not the same as maximizing $\sum_{a,b \in S} d(a,b)$ (see discussion on taxonomic distinctness). Moreover, in practice, no two pairs of EUs will be at exactly the same distance. In this case, as mentioned above, for large $\lambda$, the single term $\max_{a,b \in S} e^{-\lambda d(a,b)}$ will dominate the sum.

One possible criticism of MMD as a measure of diversity is that it depends only on the closest pair of EUs. Thus one can imagine a situation in which there are $t$ well separated clades, each consisting of a large number of closely related EUs. Any set of $t+1$ EUs will have to contain two EUs from one of the clades. Hence, under MMD, the measure for a set consisting of $t+1$ EUs from a single clade will be similar to the measure for any other set of $t+1$ EUs. This is tantamount to saying that one would ideally prefer to minimize the whole sum of exponentials as discussed above. However, a weaker notion would be to say that amongst sets of EUs that maximize the minimum pairwise distance, we prefer those that also maximize the second smallest pairwise distance, and so on. The greedy approach we propose in the next section will always have a solution for $k$ EUs which contains the solution for $j$ EUs for every $j < k$. This means that not only the minimum distance, but also the distances between all previously chosen EUs will be large. In a situation similar to that outlined here, the GreedyMMD solution for $t + 1$ EUs will contain the solution for $t$ EUs, *i.e.* at least one from every clade.

## A Greedy 2-Approximation Algorithm for MMD

In this section, we analyze the simple greedy approach for selecting a subset of EUs under MMD with distance $\delta$. If $\delta$ is a tree metric, then, as observed in Spillner et al. (2007), one can obtain an optimal solution for MMD in polynomial time using the techniques of Chandrasekaran and Daughety (1981). Nevertheless, because of the desirable property of stability, we are interested in greedy solutions in their own right, and so wish to understand their relative performance. In terms of stability, Example 2 illustrates the potential problem of using the true optimal solution.

***Example* 2.** Consider the phylogenetic tree shown in Figure 3. Under MMD, it is easily checked that $S_5 = \{x_1, x_3, x_5, x_7, x_9\}$ is the unique optimal set of five EUs ($MD(S_5) = 7$), while $S_6 = \{x_1, x_2, x_4, x_6, x_8, x_9\}$ is the unique optimal set of six EUs ($MD(S_6) = 6$). Thus, in this instance, increasing resources to enable an additional EU to be conserved would see three of the currently selected EUs dropped from the optimal set. This example can be extended in the obvious way so that, for an arbitrary $k$, the unique optimal set of size $k+1$ intersects in only two EUs with the unique optimal set of size $k$.

We analyze the following greedy algorithm.

```
GreedyMMD($\delta, k$)

Step 1 Let $S$ be the empty set.
Step 2 Select the two most distant EUs and add to $S$.
       (That is, select two elements $x$ and $y$ that maximize $\delta(x, y)$.)
Step 3 Set counter $c = 2$.
Step 4 If $c = k$, STOP; otherwise, select an EU from those not already included
       in $S$ so that the minimum distance between that EU and those in $S$ is
       maximum amongst all remaining EUs not in $S$.
       (That is, select $z \in X - S$ that maximizes $\min_{y \in S} \delta(z, y)$.)
       Add the selected EU to $S$.
Step 5 Set $c = c + 1$ and return to Step 4.
```

We remark here that the greedy selection criterion in Step 4 of GreedyMMD may be restated simply as: choose $z$ that maximizes $MD(S \cup \{z\})$ amongst all $z \in X - S$.

The next theorem shows that provided $\delta$ satisfies the triangle inequality, then GreedyMMD is a 2-approximation algorithm to MMD. This means that if $S$ is the solution returned by GreedyMMD and $Y_{\text{opt}}$ is an optimal solution, then $2MD(S) \geq MD(Y_{\text{opt}})$, that is, $MD(S)$ is at least half $MD(Y_{\text{opt}})$. It is shown in Ravi et al. (1994) that, assuming $\delta$ is only guaranteed to satisfy the triangle inequality, for any $\epsilon > 0$, no polynomial-time algorithm can return a $(2-\epsilon)$-approximation to MMD unless P=NP. In particular, in this case the problem MMD is NP-hard. Hence, in this setting, GreedyMMD gives the best possible approximation. This theorem has previously been observed in the context of operations research, for example, see Tamir (1991) and Ravi et al. (1994). However, we have included our proof in the appendix as the associated preliminary lemma will be used again in the next section and the proof is written in the language of phylogenetics.

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

It is possible that for a given set of EUs, the $k$-element subset returned by GreedyMMD and that which optimizes MMD will exhibit the same minimum distance or, in fact, be the same subsets. However, we reiterate that the greedy algorithm cannot be guaranteed to work any better than as a 2-approximation algorithm even if $\delta$ is a tree metric, as we show in the following example. We say, therefore, that Theorem 1 is sharp.

***Example* 3.** Consider the phylogenetic $X$-tree $\mathcal{T}$ shown in Figure 4, where the length of the edge incident with $z$ is arbitrarily small. Suppose that GreedyMMD is applied to the distance on $X$ induced by the path lengths in $\mathcal{T}$, where $k=4$. The first two elements selected by GreedyMMD are $x_1$ and $x_2$, the third element is $z$, and the last element is either $y_1$ or $y_2$. If $S$ denotes the resulting set, then

$MD(S) = 3 + \epsilon$. However, it is easily seen that

$$Y_{\text{opt}} = \{x_1, x_2, y_1, y_2\}$$

is an optimal solution and $MD(Y_{\text{opt}}) = 6$. Thus the approximation ratio in this case can be made arbitrarily close to 2 by an appropriate choice of $\epsilon$.

## Ultrametric Distances

In the previous section, we described how well GreedyMMD performed as an approximation to MMD provided the distance measure satisfied the triangle inequality. In this section, we show that there are conditions under which GreedyMMD will always return a subset of EUs that is optimal under MMD. In particular, the next theorem guarantees that if $\delta$ is an ultrametric, then the solution set $S$ returned by GreedyMMD$(\delta, k)$ is an optimal solution of size $k$ to MMD. Moreover, if $\mathcal{T}$ is a rooted phylogenetic tree that realizes $\delta$, then the PD score of $S$, denoted $PD(S)$, is equal to the maximum PD score over all subsets of size $k$. We denote this last score by $PD_k$. Note that, in this setting, $PD(S)$ is the total length of the subtree connecting the elements of $S$ and the root of the phylogeny. A proof of Theorem 2 can be found in the appendix.

**Theorem 2.** *Let $\delta$ be an ultrametric on $X$, and let $k$ be an integer greater than one. Let $S_k$ be the set returned by* GreedyMMD$(\delta, k)$ *and let $Y_{\text{opt}}$ be an optimal solution of size $k$ to* MMD. *Then*

$$MD(S_k) = MD(Y_{\text{opt}}).$$

*Moreover, if $\mathcal{T}$ is a rooted phylogenetic $X$-tree that realizes $\delta$, then $PD(S_k) = PD_k$ on $\mathcal{T}$.*

## Taxonomic Distinctness and $p$-Median Measures

In this section we discuss two existing measures of diversity, namely taxonomic distinctness and the $p$-median. We briefly outline why we do not think these are suitable for our purposes.

### *Taxonomic Distinctness*

A measure closely related to MMD that is widely used in conservation biology, though in slightly different circumstances, is *taxonomic distinctness* (TD), see Clarke and Warwick (1998). Typically, TD is used for comparing the biodiversity of different areas. Each area is visited and the set of taxa observed within the area is recorded. The TD of the area is (effectively) taken to be the average pairwise distance between taxa. This naturally leads to the idea in our setting of selecting a set of EUs that maximizes the *average* distance between EUs (MAD). At first sight, this appears similar to maximizing the minimum pairwise distance as in MMD. However, for many instances in which there is dominant longest path in the tree, *e.g.* to an out-group of EUs, the optimal and greedy sets chosen under MAD unduly try to balance the number of EUs either side of this path, see Example 4. A greedy set under MAD is chosen in the same way as that for MMD except that Step 4 in GreedyMMD is replaced with the following step:

**Step 4′** If $c = k$, STOP; otherwise, select an EU from those not already included in $S$ so that the sum of the pairwise distances between that EU and those in $S$ is maximum amongst all remaining EUs not in $S$.

(That is, select $z \in X - S$ that maximizes $\sum_{y \in S} \delta(z, y)$.)

Add the selected EU to $S$.

***Example* 4.** Consider the ultrametric tree shown in Figure 5. Because of the relatively short distance between any two elements in $\{a_1, a_2, \ldots, a_n\}$, under MMD and PD, a greedy (and therefore optimal) solution set will involve a single element from $\{a_1, a_2, \ldots, a_n\}$; the rest of the set will be made up of a spread of elements from $\{b_1, b_2, \ldots, b_m\}$. Whereas for all $k$, under MAD, both greedy and optimal solution sets of size $k$ will either have $\frac{k}{2}$ or $\frac{k-1}{2}$ of its elements from $\{a_1, a_2, \ldots, a_n\}$. This is straightforward to see by noting that the longest path between any pair of EUs always involves an $a_i$ and a $b_j$.

For completeness, we highlight the computational similarities and differences between MMD and MAD depending on whether $\delta$ is a tree metric or satisfies the triangle inequality. First, if $\delta$ is a tree metric, then, as in the case for MMD, selecting an optimal set of $k$ EUs under MAD can be done in polynomial time (Chandrasekaran and Daughety, 1981). Furthermore, it appears that the exact approximation ratio of the greedy algorithm for MAD is unknown. Ravi et al. (1994) have shown that it is no worse than a 4-approximation; simple examples show that it is no better than a 4/3-approximation. In comparison, the greedy algorithm GreedyMMD returns a 2-approximation to MMD and this is sharp. Second, if $\delta$ satisfies the triangle inequality, Ravi et al. (1994) show that the greedy algorithm for MAD is no worse than a 4-approximation and no better than a 2-approximation algorithm. Again, GreedyMMD returns a 2-approximation to MMD, and this is sharp.

### *The $p$-median*

Faith and Walker (1993) proposed an alternative measure of phylogenetic diversity called "$p$-median diversity" to capture a notion of spread. There are two variants: the continuous and the discrete $p$-median. Both are defined in terms of a *redundancy* which is the complement of diversity. The objective is to minimize the $p$-median redundancy when selecting a set of EUs. As before, let $X$ be a set of EUs and $\mathcal{T}$ be a phylogenetic $X$-tree. For all $p \geq 1$, let $S$ be a $p$-element subset of $X$. The *continuous $p$-median redundancy* of $S$ is the sum, over all points $q$ in $\mathcal{T}$, of the distances $d_{\mathcal{T}}(q, s_q)$, where $s_q$ is the element of $S$ closest to $q$ in $\mathcal{T}$. Here one views each branch of $\mathcal{T}$ as a path consisting of $l$ edges, where $l$ is the length of the branch and, under this viewpoint, a *point* of $\mathcal{T}$ is a vertex of $\mathcal{T}$.

The continuous $p$-median arises naturally in a setting in which combinations of features both arise and disappear in the phylogeny, see Faith (1994) for a description of the proposed model. A set which minimizes the continuous $p$-median redundancy, and thus maximizes the continuous $p$-median diversity, will maximize the number of observed combinations of features (Faith, 1994). However, as pointed out in the same paper, one weakness of this model is that features arising near the end of a pendant (terminal) branch are ignored. The model described in this paper does not ignore such features.

The *discrete* $p$-median redundancy is defined similarly to the continuous, except the sum is restricted to points $q$ in $X$. It is remarked in Faith (1994) that the discrete

$p$-median has some undesirable properties. In particular, they give an example in which the discrete $p$-median does not correctly select a set that maximizes the observed number of combinations of features according to the proposed model. Therefore we focus here only on the continuous $p$-median, although our discussion, below, applies to both variants.

An unusual property of $p$-median is that it depends not only on the set of EUs selected, but also on all those not selected. This leads to several issues. For example, two groups of researchers considering the same subset of EUs will arrive at different views of the diversity value of the set, depending on what other EUs they do or do not include in the analysis. Depending on ones objectives, it may be that the type of representative subset given by the $p$-median is useful. In Horn et al. (1996), it is observed that such a set gives a best representation of the entire tree, in some sense. However, it is not appropriate for our purposes in which we seek a set that intrinsically captures maximum diversity, as we illustrate in the following example.

***Example* 5.** Consider the phylogeny depicted in Figure 6. All edges have the marked length. Suppose we wish to select a set of three EUs to maximize diversity. Under continuous $p$-median, we would select a set containing exactly one of $a$ and $b$, exactly one of $c$ and $d$, and the EU $e$; for example, $\{a, c, e\}$. This is because for a set such as $\{a, b, c\}$, each long edge contributes $\approx \sum_{i=1}^{100} i$ to the $p$-median redundancy. Whereas, for a set such as $\{a, c, e\}$, the long edges leading to $a$ and $c$ contribute only $\approx 2\sum_{i=1}^{50} i$ to the $p$-median redundancy, which is therefore much smaller.

Under MMD and PD, we would select either $\{a, b, c\}$ or $\{a, b, d\}$. This fits in with our intuition that of the features present at the root that survive down to $a$ or $c$, there will be greater overlap in EU $e$ than in EU $b$. Thus including $b$ with $a$ and $c$ will provide the greater additional number of observed features than including $e$. We confirm this using the full model, and in particular (2), described in the model section. Take $\lambda \approx 1/167$, so that many features are expected to survive down a single evolution of 100 time steps, but most do not to survive down two independent evolutions of 100 time steps. Taking set $\{a, b, c\}$, we expect to observe over 18% more features than taking set $\{a, b, e\}$. We also observe that if we are selecting four EUs, then under MMD and $p$-median we would select $\{a, b, c, e\}$ or $\{a, b, d, e\}$, whereas under PD we would select $\{a, b, c, d\}$. Again this fits our intuition and our model, since $c$ and $d$ will share many more features than $e$ shares with any other EU.

Note that this specific example holds for all $\lambda \geq 0$, but is clearest for the value given above. In other examples and for some lambdas a set chosen using $p$-median may exhibit a greater number of expected features under our model. As discussed previously, our approach is for large values of $\lambda$ relative to the phylogenetic distances. For readers considering other ranges of $\lambda$ or interested in further alternatives to PD, we recommend investigating $p$-median (Faith, 1994).

## Discussion

In this paper, we show that a subset of EUs that maximizes phylogenetic diversity may result in a counter-intuitive collection of species earmarked for conservation or as genomic sequencing targets. We argue that as a plausible alternative to phylogenetic diversity, an appropriate criterion is to choose the subset of EUs where the minimum distance between any pair of EUs is maximum, amongst all possible subsets of the same size. This criterion has the virtue of choosing EUs that are "spread out" across the phylogenetic tree and has a strong theoretical justification as a heuristic for selecting EUs that optimize a natural model of diversity.

However, as others have shown (e.g. Moulton et al., 2007), choosing the subset that maximizes the minimum pairwise distance between EUs cannot be achieved by applying a greedy algorithm whereby one "builds up" a subset of $k$ EUs by successively adding to optimal subsets of $2, 3, \ldots, k-1$ EUs. Nonetheless, we think that there is real value in being greedy. When one has the opportunity to sequence yet another genome or save yet another species, one would want all the genomes that have already been sequenced and all the species that have already been saved to be part of the larger optimal subset. A greedy algorithm guarantees this and, moreover, gives a solution such that not only the minimum pairwise distance, but also the pairwise distances between all previously chosen EUs will be large.

It is of value, then, to consider the extent to which a greedy algorithm approximates an optimal implementation of MMD. At worst, the greedy algorithm we defined (GreedyMMD) will choose EUs that are separated on a tree by distances no shorter than half the shortest distance of an optimal subset of EUs under MMD provided the distance satisfies the triangle inequality. This suggests that GreedyMMD will still choose a subset that is reasonably "spread out".

When distances are ultrametric, Theorem 2 shows that GreedyMMD returns a set which is in fact optimal under both MMD and PD. This is significant as it may be possible that even with approximately ultrametric distances GreedyMMD produces a set that is close to optimal under PD. (Note that, although one could compute PD directly using a distance matrix that satisfies the ultrametric condition, it is not clear what one does for a distance matrix that has no underlying tree.)

Lastly, GreedyMMD can be used on a pairwise distance matrix that does not necessarily induce a tree metric. Thus a large set of sequences could be analyzed using this method based upon a simple distance measure such as Hamming distance (Hamming, 1950), without expending the computational time required to reconstruct an accurate phylogeny. Moreover, if our data satisfies the triangle inequality, we have the performance guarantee of Theorem 1.

## Acknowledgements

We thank Dan Faith and an anonymous referee for their helpful and constructive comments.

The first author was supported by the EPSRC, the second was supported by the Allan Wilson Centre for Molecular Ecology and Evolution, and the third author was supported by the New Zealand Marsden Fund. We thank Geoff Nicholls and Mike Steel for useful discussions. This project began at the Isaac Newton Institute for the Mathematical Sciences, under the auspices of the Phylogenetics Programme.

## References

Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc. Natl. Acad. Sci. 93:9188–9193.

Bordewich, M. and C. Semple. 2008. Nature reserve selection problem: A tight approximation algorithm. IEEE/ACM Transactions on Computational Biology and Bioinformatics in press.

Chandrasekaran, R. and A. Daughety. 1981. Location on tree networks: p-centre and n-dispersion problems. Math. Oper. Res. 6:50–57.

Clarke, K. R. and R. M. Warwick. 1998. A taxonomic distinctness index and its statistical properties. The Journal of Applied Ecology 35:523–531.

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

Faith, D. P. 1994. Phylogenetic pattern and the quantification of organismal biodiversity. Philosophical Transactions: Biological Sciences 345:45–58.

Faith, D. P. and P. A. Walker. 1993. Diversity: A software package for sampling phylogenetic and environmental diversity. Canberra: CSIRO Division of Wildlife and Ecology .

Guindon, S. and O. Gascuel. 2003. A simple, fast , and accurate method to estimate large phylogenies by maximum liklihood. Syst. Biol. 52:696–704.

Hamming, R. W. 1950. Error detecting and error correcting codes. Bell System Technical Journal 26:147–160.

Holland, B. R. 2001. Evolutionary analyses of large data sets: Trees and beyond. Ph.D. thesis Massey University.

Horn, M. E. T., D. P. Faith, and P. A. Walker. 1996. The phylogenetic moment - a new diversity measure, with procedures for measurement and optimisation. Environment and Planning A 28:2139–2154.

Moulton, V., C. Semple, and M. Steel. 2007. Optimizing phylogenetic diversity under constraints. J. Theoret. Biol. 246:186–194.

Pardi, F. and N. Goldman. 2005. Species choice for comparative genomics: being greedy works. PLoS Genetics 1:e71.

Ravi, S. S., D. J. Rosenkrantz, and G. K. Tayi. 1994. Heuristic and special case algorithms for dispersion problems. Operations Research 42:299–310.

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

Rodrigues, A. S. L. and K. J. Gaston. 2002. Maximising phylogenetic diversity in the selection of networks of conservation areas. Biological Conservation 105:103–111.

Sanderson, M. J., M. J. Donoghue, W. Piel, and T. Eriksson. 1994. Treebase: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. Amer. Jour. Bot. 81.

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

Spillner, A., B. Nguyen, and V. Moulton. 2007. Computing phylogenetic diversity for split systems, preprint.

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

Tamir, A. 1991. Obnoxious facility location on graphs. SIAM J. Disc. Math. 4:550–567.

Woese, C. R. 1987. Bacterial evolution. Microbiol. Rev. 41.

## Appendix: Proofs of Theorems 1 and 2

Here we give the proofs of Theorems 1 and 2. Both proofs make use of the following lemma.

**Lemma 3.** *Let $\delta$ be a distance on $X$ and let $k$ be an integer greater than two. Let $S_{k-1}$ be the $(k-1)$-element set that is constructed at the completion of the second-to-last iteration of* GreedyMMD$(\delta, k)$*. Then, for any element $x \in X - S_{k-1}$, we have $MD(S_{k-1} \cup \{x\}) = \delta(x, s)$ for some $s \in S_{k-1}$.*

*Proof.* For $2 \leq i \leq k-1$, let $S_i = \{s_1, s_2, \ldots, s_i\}$ denote the $i$-element subset of $S_{k-1}$ that is sequentially constructed by GreedyMMD$(\delta, k)$. If $MD(S_{k-1} \cup \{x\}) \neq \delta(x, s)$ for any $s \in S_{k-1}$, then, for some distinct $s_i, s_j \in S_{k-1}$ with $i < j$, we have $MD(S_{k-1} \cup \{x\}) = \delta(s_i, s_j) < \delta(x, s)$. If there is more than one pair $s_i, s_j$, choose a pair with minimal $j$, and so $MD(S_{j-1}) > \delta(s_i, s_j)$. But then

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

contradicting the way in which GreedyMMD selects elements. The lemma now follows. □

*Proof of Theorem 1.* For all $2 \leq i \leq k$, let $S_i = \{s_1, s_2, \ldots, s_i\}$ denote the $i$-element subset of $S_k$ that is sequentially constructed by GreedyMMD$(\delta, k)$. By Lemma 3, $MD(S_k) = \delta(s_k, s_i)$ for some $i \in \{1, 2, \ldots, k-1\}$.

Let $Y_{\mathrm{opt}} = \{y_1, y_2, \ldots, y_k\}$ be an optimal solution of size $k$ to MMD. Amongst the elements in $Y_{\mathrm{opt}} - S_{k-1}$, let $y$ be an element such that

$$MD(S_{k-1} \cup \{y\}) = \max\{MD(S_{k-1} \cup \{y'\}) : y' \in Y_{\mathrm{opt}} - S_{k-1}\}.$$

By Lemma 3, $MD(S_{k-1} \cup \{y\}) = \delta(y, s_{i'})$ for some $i' \in \{1, 2, \ldots, k-1\}$. Because of the selection criteria of GreedyMMD, $MD(S_k) \geq MD(S_{k-1} \cup \{y\})$, and so $\delta(s_k, s_i) \geq \delta(y, s_{i'})$.

Now assign each element of $Y_{\text{opt}}$ to the element in $S_{k-1}$ that it is closest to under $\delta$. Since $|Y_{\text{opt}}| > |S_{k-1}|$, it follows by the pigeon-hole principle that two distinct elements $y_r, y_s \in Y_{\text{opt}} - S_{k-1}$ are assigned to the same element, $s$ say, in $S_{k-1}$. By the choice of $y$ above,

$$\delta(y_r, s) \leq \delta(y, s_{i'}) \text{ and } \delta(y_s, s) \leq \delta(y, s_{i'}).$$

Then, as $\delta$ satisfies the triangle inequality and it is symmetric,

$$\begin{aligned}
MD(Y_{\text{opt}}) &\leq \delta(y_r, y_s) \\
&\leq \delta(y_r, s) + \delta(s, y_s) = \delta(y_r, s) + \delta(y_s, s) \\
&\leq 2\delta(y, s_{i'}) \\
&\leq 2\delta(s_k, s_i) \\
&= 2MD(S_k) \\
&\leq 2MD(Y_{\text{opt}}).
\end{aligned}$$

That is,

$$MD(Y_{\text{opt}})/2 \leq MD(S_k) \leq MD(Y_{\text{opt}}).$$

Hence $MD(S_k)$ is a 2-approximation to $MD(Y_{\text{opt}})$. □

*Proof of Theorem 2.* The first part of the proof proceeds in a similar way to that used for proving Theorem 1. Nevertheless, for clarity, we include the proof in full.

For $2 \leq i \leq k$, let $S_i = \{s_1, s_2, \ldots, s_i\}$ denote the $i$-element subset of $S_k$ that is sequentially constructed by GreedyMMD$(\delta, k)$. By Lemma 3, $MD(S_k) = \delta(s_k, s_i)$ for some $i \in \{1, 2, \ldots, k-1\}$. Let $Y_{\text{opt}} = \{y_1, y_2, \ldots, y_k\}$ be an optimal solution of size $k$ to MMD. Amongst all elements in $Y_{\text{opt}} - S_{k-1}$, let $y$ be an element that maximizes $MD(S_{k-1} \cup \{y\})$. By Lemma 3, $MD(S_{k-1} \cup \{y\}) = \delta(y, s_{i'})$ for some $i' \in \{1, 2, \ldots, k-1\}$. Now $\delta(s_k, s_i) \geq \delta(y, s_{i'})$ as $MD(S_k) \geq MD(S_{k-1} \cup \{y\})$. Assign each element of $Y_{\text{opt}}$ to the element of $S_{k-1}$ that it is closest to under $\delta$. Since $|Y_{\text{opt}}| > |S_{k-1}|$, there are two distinct elements $y_r$ and $y_s$ in $Y_{\text{opt}} - S_{k-1}$ assigned to the same element $s$ in $S_{k-1}$. By choice of $y$ above,

$$\delta(y_r, s) \leq \delta(y, s_{i'}) \text{ and } \delta(y_s, s) \leq \delta(y, s_{i'}).$$

Thus, since $\delta$ is an ultrametric, we have

$$\begin{aligned} MD(Y_{\text{opt}}) &\leq \delta(y_r, y_s) \\ &\leq \max\{\delta(y_r, s), \delta(y_s, s)\} \\ &\leq \delta(y, s_{i'}) \\ &\leq \delta(s_k, s_i) \\ &= MD(S_k) \\ &\leq MD(Y_{\text{opt}}). \end{aligned}$$

In other words, equality holds throughout and so $MD(S_k) = MD(Y_{\text{opt}})$. This completes the first part of the theorem.

To show that $PD(S_k) = PD_k$ on $\mathcal{T}$, we use induction on $k$. Clearly, the result holds if $k = 2$. So assume that the result holds whenever the set returned by GreedyMMD has size at most $k - 1$, where $k \geq 3$.

Suppose that $S_k = \{s_1, s_2, \ldots, s_k\}$ is returned by GreedyMMD. By Lemma 3, $MD(S_k) = \delta(s_k, s_i)$ for some $i < k$. Let $u$ be the most recent common ancestor of $s_k$ and $s_i$ in $\mathcal{T}$. Then, as $\delta$ is an ultrametric,

$$MD(S_k) = \delta(s_i, s_k) = \delta(s_i, u) + \delta(u, s_k) = 2\delta(u, s_k).$$

Note that, except for $u$, the path from $s_k$ to $u$ in $\mathcal{T}$ does not intersect the minimal subtree of $\mathcal{T}$ connecting the root and the elements in $\{s_1, s_2, \ldots, s_{k-1}\}$; otherwise, $MD(S_k) < \delta(s_k, s_i)$. Thus, in determining $S_k$, GreedyMMD finds the element $s$ in $X - S_{k-1}$ that maximizes the sum of the edge lengths from $s$ to the minimal subtree of $\mathcal{T}$ that connects the elements in $S_{k-1}$. Indeed, this is exactly the greedy selection criterion used in finding a set of EUs that optimizes phylogenetic diversity. In particular, to get a set of size $k$ that optimizes PD, take a set of size $k - 1$ that optimizes PD and find an element that maximizes the sum of the edge lengths joining it to the minimal subtree of $\mathcal{T}$ connecting the root and the elements in the set of size $k - 1$ (Pardi and Goldman, 2005; Steel, 2005). By the induction assumption, $PD(S_{k-1}) = PD_{k-1}$, and so $PD(S_k) = PD_k$. This completes the proof of the theorem. □

## Figure Captions

**Figure 1.** Reproduction of Woese's (Woese, 1987) small-subunit ribosomal RNA tree showing the subtree subtended by three EUs chosen by (a) minimizing PD and by (b) maximising the minimum distance. We constructed this tree using small-subunit ribosomal RNA sequences from an alignment by Barns et al. (1996) available in TreeBase (Sanderson et al., 1994). Maximum likelihood trees were constructed with PHYML (Guindon and Gascuel, 2003) using a GTR model of evolution. The three groups on our tree are represented by the following taxa. ARCHAEA: *Methanococcus vannielli* (Methanogen A), *Methanobacterium* (Methanogen B), *Thermococcus* (Extreme thermophile A), *Thermoproteus* (Extreme thermophile B), *Desulfurococcus* (Extreme thermophile C), *Haloferax* (Extreme halophiles); BACTERIA: *Thermotoga*, *Flavobacteria* (Flavobacteria), *Gloeobacter* (Cyanobacteria), *Escherichia coli* (Purple bacteria), *Bacillus* (Gram-positive bacteria), *Thermomicrobium* (Green non-sulphur bacteria); EUKARYOTES: *Vairimorphal* (Microsporidia), *Euglena geniculata* (Flagellates), *Dictyostelium* (Cellular slime molds), *Zea mays* (Plants), *Homo sapiens* (Animals) and *Coprinus* (Fungi).

**Figure 2.** A phylogeny on which the optimal set of 3 EUs chosen to maximize the minimum pairwise distance yields a greater expected number of observed features for all $\lambda > 0.00047$ and is significantly different from an arbitrary set of 3 EUs for $\lambda < 9.72$.

**Figure 3.** A phylogeny on which the (unique) optimal set $\{x_1, x_3, x_5, x_7, x_9\}$ of 5 EUs and the (unique) optimal set $\{x_1, x_2, x_4, x_6, x_8, x_9\}$ of 6 EUs selected under MMD intersect in only 2 EUs.

**Figure 4.** A phylogenetic tree demonstrating that GreedyMMD cannot be guaranteed to be better than a 2-approximation to MMD. In selecting a set of size 4, GreedyMMD selects a set $S$ containing $x_1$, $x_2$, $z$, and either $y_1$ or $y_2$, while MMD selects $\{x_1, x_2, y_1, y_2\}$. Here $MD(S) = 3 + \epsilon$, but $MD(\{x_1, x_2, y_1, y_2\}) = 6$.

**Figure 5.** An ultrametric tree on which MAD disagrees with MMD and PD. For all $k \geq 2$, both MMD and PD greedily (and therefore optimally) select a set of size $k$ that contains exactly one element from $\{a_1, a_2, \ldots, a_k\}$ and a spread of elements from $\{b_1, b_2, \ldots, b_m\}$. While, under MAD, both greedy and optimal solution sets of size $k$ will either have $\frac{k}{2}$ or $\frac{k-1}{2}$ of its elements from $\{a_1, a_2, \ldots, a_n\}$ depending on whether or not $k$ is even.

**Figure 6.** A phylogeny on which the continuous $p$-median does not give an optimal set under our model. All edges have the marked length. Suppose we wish to select a set of three EUs to maximize diversity. Under $p$-median, we would select a set containing $e$, one of $a$ and $b$, and one of $c$ and $d$; for example, $\{a, c, e\}$. However, under MMD, we would select either $\{a, b, c\}$ or $\{a, b, d\}$. This fits in with our intuition that of the features present at the root that survive down to $a$ or $c$, there will be greater overlap in EU $e$ than in EU $b$. Thus including $b$ along with $a$ and $c$ will provide the greater additional number of observed features than including $e$. We confirm this using the model described in this paper. Taking $\lambda \approx 1/167$, for set $\{a, b, c\}$ we expect to observe over 18% more features than taking set $\{a, b, e\}$.

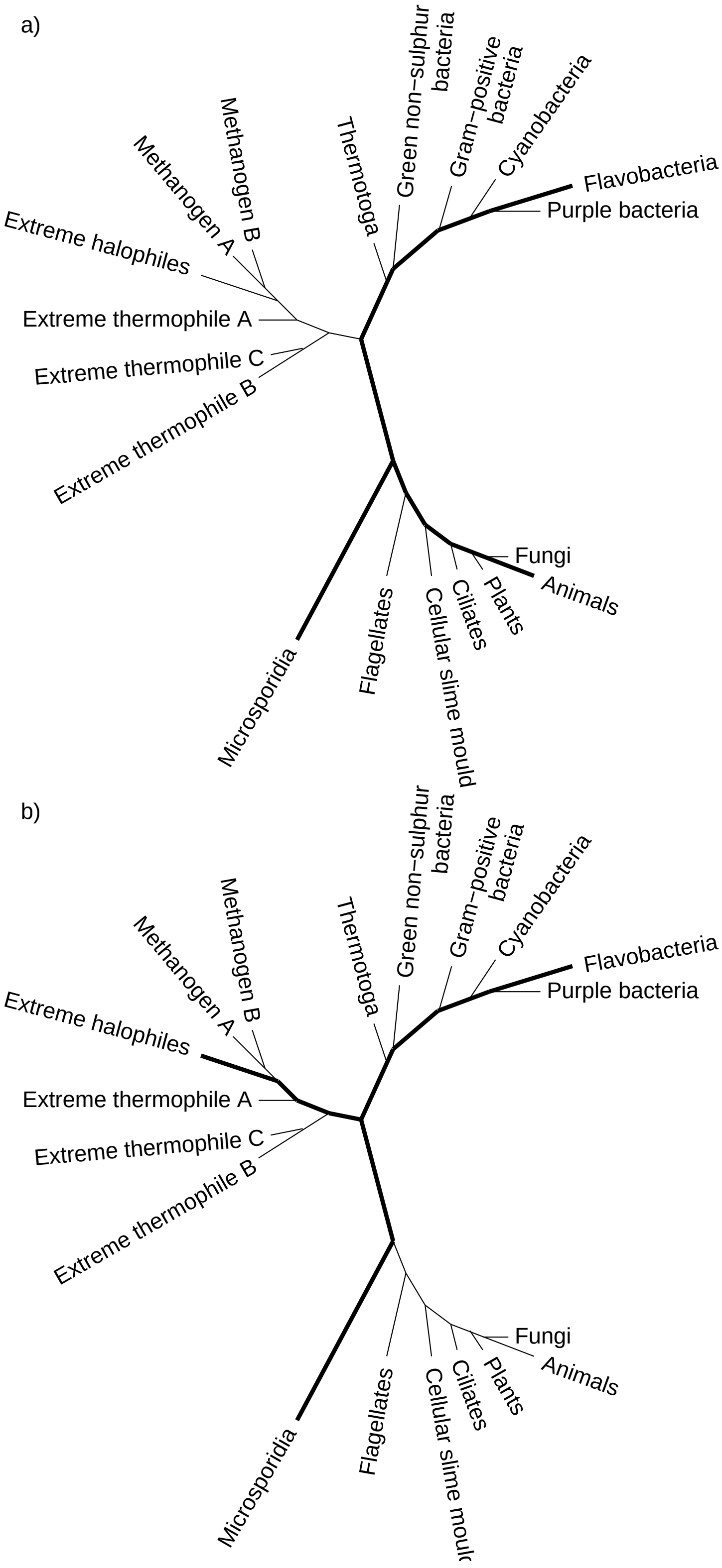
a)
Methanogen B
Methanogen A
Extreme halophiles
Extreme thermophile A
Extreme thermophile C
Extreme thermophile B
Thermotoga
Green non−sulphur bacteria
Gram−positive bacteria
Cyanobacteria
Flavobacteria
Purple bacteria
Fungi
Animals
Plants
Ciliates
Cellular slime mould
Flagellates
Microsporidia
b)
Methanogen B
Methanogen A
Extreme halophiles
Extreme thermophile A
Extreme thermophile C
Extreme thermophile B
Thermotoga
Green non−sulphur bacteria
Gram−positive bacteria
Cyanobacteria
Flavobacteria
Purple bacteria
Fungi
Animals
Plants
Ciliates
Cellular slime moul
Flagellates
Microsporidia

Figure 1

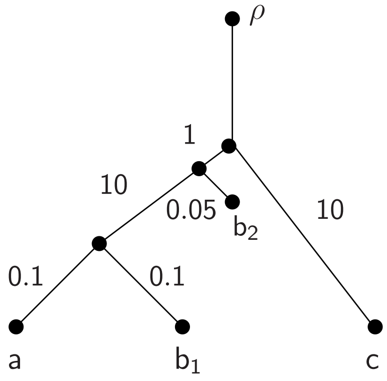

Figure 2

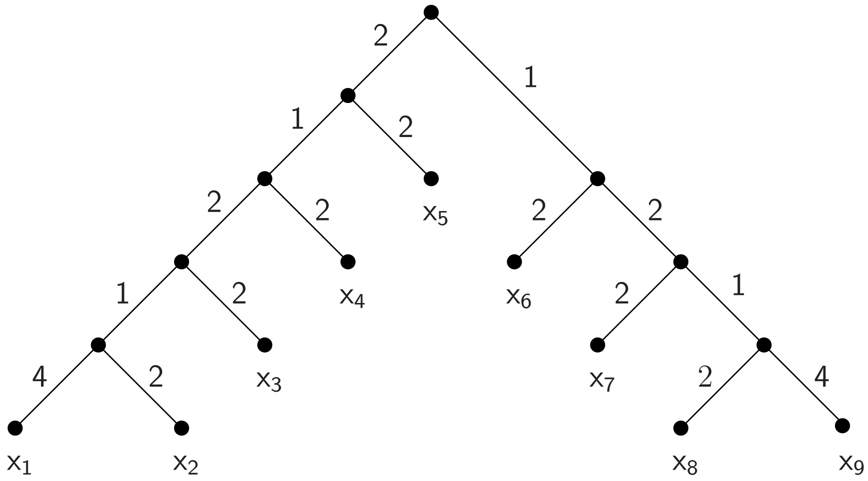

Figure 3

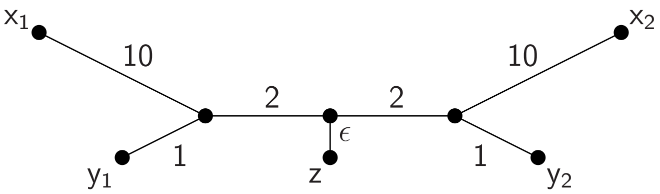

FIGURE 4

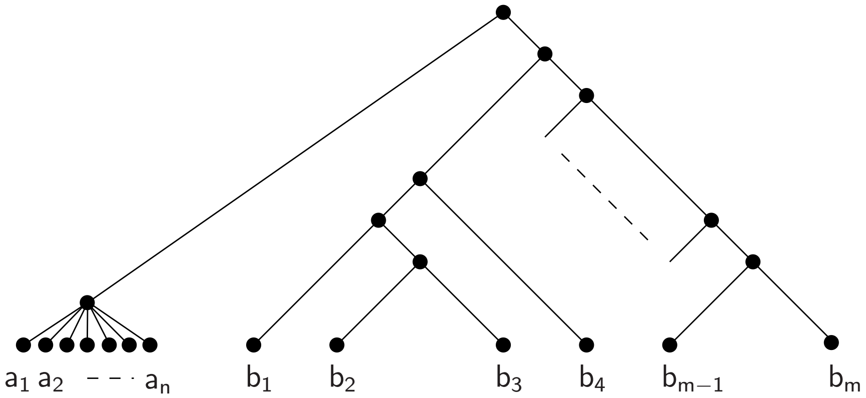

Figure 5

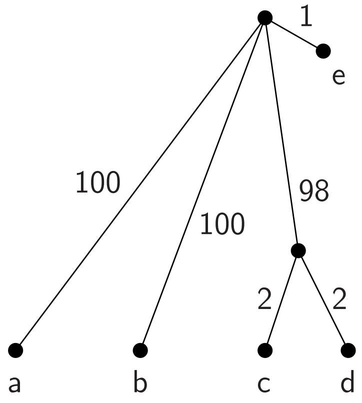

Figure 6