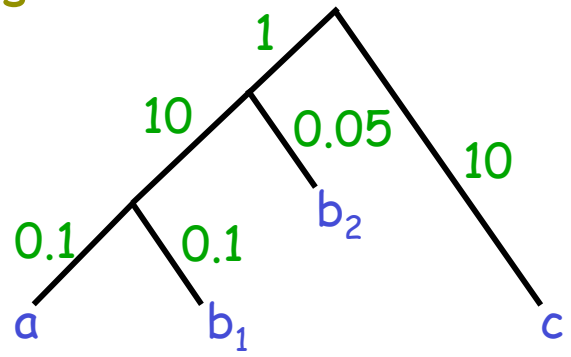
Phylogenetic Diversity with Disappearing Features

Charles Semple
Department of Mathematics and Statistics
University of Canterbury
New Zealand

Joint work with Magnus Bordewich, Allen Rodrigo

Mathematics & Informatics in Evolution & Phylogeny, Hameau de l'Etoile 2008

Conservation biology and comparative genomics



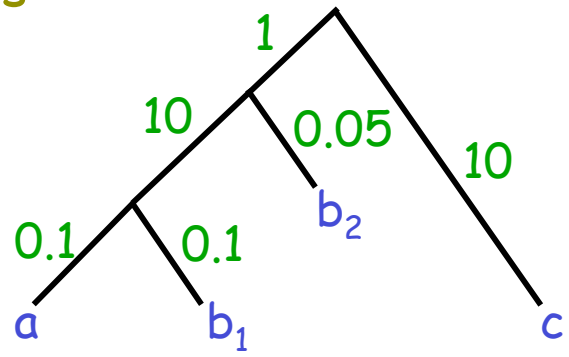
Quantative methods based on biodiversity are used for determining which collection of EUs to save or sequence.

Two criteria:

I. Maximizing Phylogenetic Diversity (PD) For a set S of EUs and a phylogeny T, PD(S) is the sum of the edges of T spanned by S.

- Find a k-element subset of EUs that maximizes PD.

Conservation biology and comparative genomics



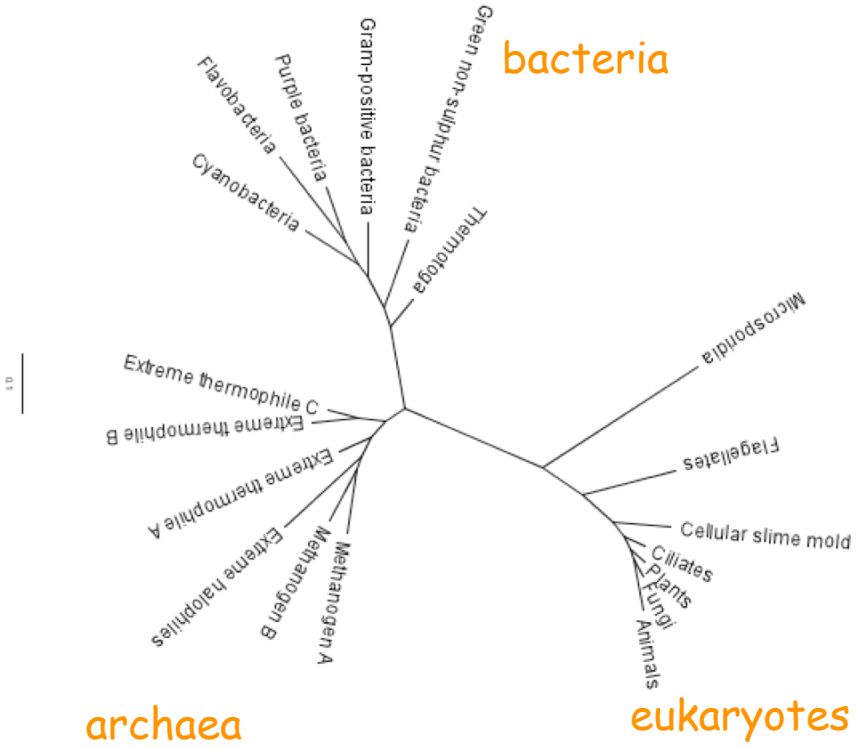
Quantative methods based on biodiversity are used for determining which collection of EUs to save or sequence.

Two criteria:

I. Maximizing Phylogenetic Diversity (PD) For a set S of EUs and a phylogeny T, PD(S) is the sum of the edges of T spanned by S.
   - Find a k-element subset of EUs that maximizes PD.

II. Maximizing Minimum Distance (MD) For a distance d on EUs and a subset S of EUs, MD(S) is the minimum distance between any pair of EUs in S.
   - Find a k-element subset of EUs that maximizes MD(S).

Iconic example: Woese's (1987) small-subunit ribosomal RNA tree

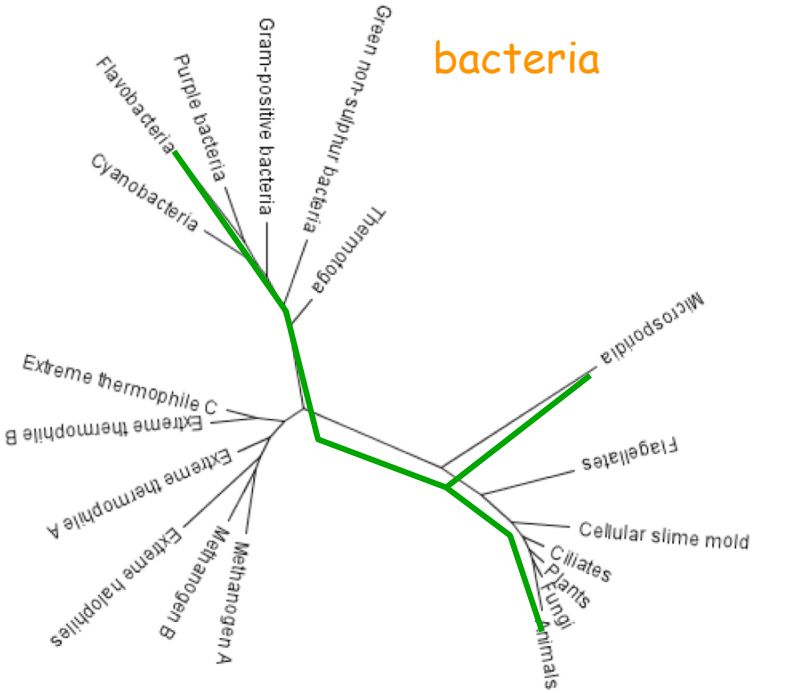


Task: Select 3 EUs for sequencing.

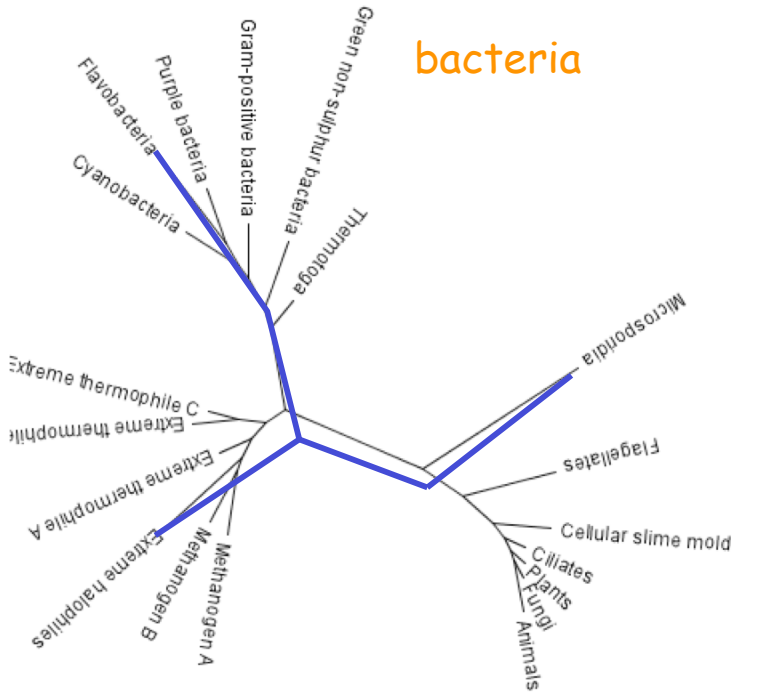
One bacterium, one archaeon, one eukaryote seems an intuitively good selection.

Iconic example: Woese's (1987) small-subunit ribosomal RNA tree

MaxPD



MaxMD



What's going on?

PD measures the expected number of different features shown by the selected EUs.

Assumptions:

I. the length of an edge represents the number of different features arising along that edge;

II. once a feature arises, it persists forever and is present in all descendant EUs.

Why two eukaryotes?

MaxPD chooses an additional eukaryote since an EU connected near the root by a short edge is assumed to contain almost exclusively features shared by every other EU.

What's going on?

Instead, the measure is the expected # of different features shown by the selected EUs under the following model of evolution.

Assumptions:

I. the length of an edge represents the number of different features arising along that edge;

II. once a feature arises, it persists forever and is present in all descendant EUs.

III. features have a constant probability of disappearing on any evolutionary path in which they are present.

It turns out, by choosing a set of EUs that maximize MD, one can obtain a reasonable solution to maximizing this measure.

The model of diversity for which MaxMD is a justifiable heuristic

Assumptions:

I. Features disappear according to an exponential distribution with rate λ independently on any edge.

(Once present, a feature has a constant and memory-less probability $e^{-\lambda}$ of surviving in each time step.)

II. ρ on an infinitely long edge connected to first branching point.

(Full set of features available at the beginning.)

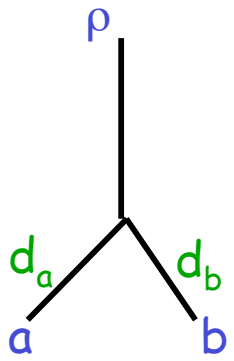
For a subset A of EUs, the # of features present is a random variable F_A.

For a single EU a, $$E(F_{\{a\}}) = \int_0^\infty e^{-\lambda x} dx = \frac{1}{\lambda}$$

(Sum over all points on the path from ρ to a of the probability that the feature arising at that moment is still present at a.)

The model of diversity for which MaxMD is a justifiable heuristic

For two EUs a and b,

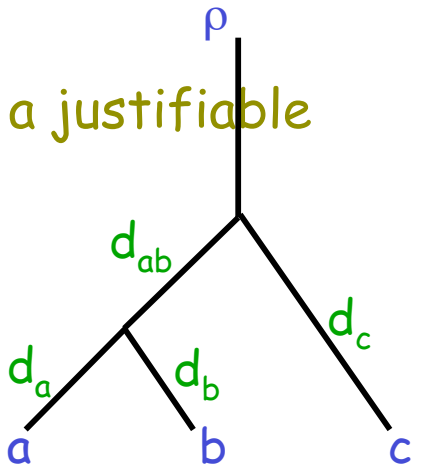


$$E(F_{\{a,b\}}) = \int_0^{d_a} e^{-\lambda x} dx + \int_0^{d_b} e^{-\lambda x} dx + \int_0^{\infty} e^{-\lambda x} (e^{-\lambda d_a} + e^{-\lambda d_b} - e^{-\lambda (d_a + d_b)}) dx$$
$$= \frac{1}{\lambda} (2 - e^{-\lambda (d_a + d_b)})$$

Using the principle of inclusion/exclusion to any size subset of EUs, we can extend the above calculation.

The model of diversity for which MaxMD is a justifiable heuristic

For three EUs a, b, and c,



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

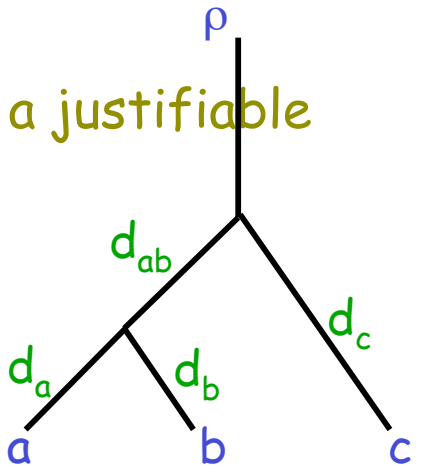
λ very small: $e^{-\lambda m} \approx (1-\lambda m)$ for all $0 \leq m \ll 1/\lambda$. So

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

As $\lambda \rightarrow 0$, $E(F_{\{a,b,c\}}) \rightarrow PD(\{a, b, c\})$.

The model of diversity for which MaxMD is a justifiable heuristic

For three EUs a, b, and c,



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

λ very big: Features die out quickly and $e^{-\lambda m}$ terms become very small.

If λ ``is so large that all features which arise are lost within one unit step, then all species are of equal status (species richness) as there is no predictable redundancy among them, …"

Faith (1994)

The model of diversity for which MaxMD is a justifiable heuristic

Before reaching species richness:

For a k-element subset S of EUs,

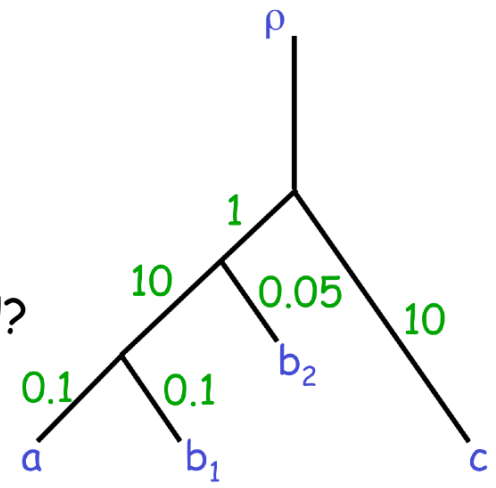
$$\frac{1}{\lambda}\left(k - \sum_{a,b\in S} e^{-\lambda d(a,b)}\right) \le E(F_S) \le \frac{1}{\lambda}\left(k - \sum_{a,b\in S} e^{-\lambda d(a,b)} + \sum_{a,b,c\in S} e^{-\lambda d(a,b,c)}\right)$$

As λ gets big, k/λ and $e^{-\lambda d'}/\lambda$ dominate (d'=distance between closest pair in S).

Thus, if λ big, then to maximize $E(F_S)$ select a set S that optimizes MaxMD.

Example: Selecting a 3-element subset

Selected a & c, do we choose b_1 or b_2 for the third EU?

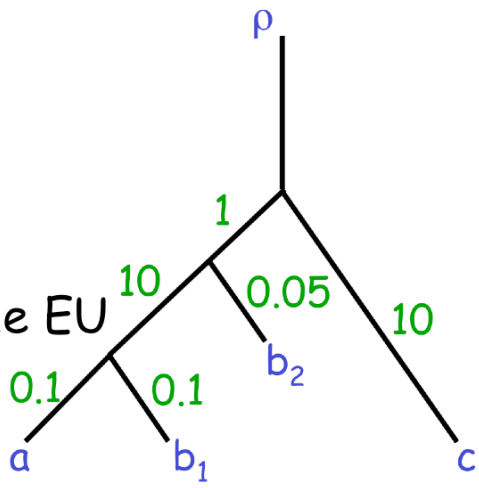


$$E(F_{\{a,b,c\}}) = \frac{1}{\lambda}(3 - e^{-\lambda(d_a + d_b)} - e^{-\lambda(d_a + d_{ab} + d_c)} - e^{-\lambda(d_b + d_{ab} + d_c)} + e^{-\lambda(d_a + d_b + d_{ab} + d_c)})$$

Which is bigger $E(F\{a,c,b_1\})$ or $E(F\{a,c,b_2\})$?
(MaxPD selects b_1, MaxMD selects b_2.)

If λ=0.4, then $E(F\{a,c,b_1\})$=5.19 but $E(F\{a,c,b_2\})$=7.43 (43% gain).

Example: Selecting a 3-element subset



1. How small does λ have to be so that PD will select the EU that maximizes the expected # of features?
To select b_1, $\lambda < 0.00047$.

2. λ large enough, choosing any 3 EUs is good enough.
For S^* an optimal set of 3 EUs and S any set of EUs,

$E(F_{S^*}) - E(F_S)$ within 5%	$\lambda > 9.72$
$E(F_{S^*}) - E(F_S)$ within 1%	$\lambda > 17.6$

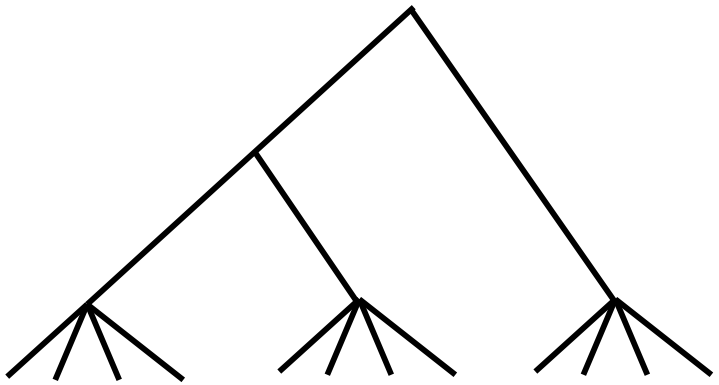
The range for λ in which MaxMD is a better criterion than MaxPD or an arbitrary selection is large---features disappearing between 10 times faster than they arise and 2000 times slower.

Selecting a set under MaxMD.

MaxMD only depends on the closest pair of EUs.

Selecting a set of size 4.

Two possible choices:

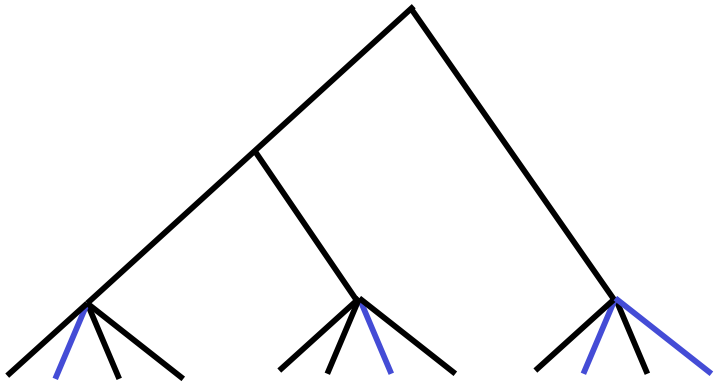


Selecting a set under MaxMD.

MaxMD only depends on the closest pair of EUs.

Selecting a set of size 4.

Two possible choices:



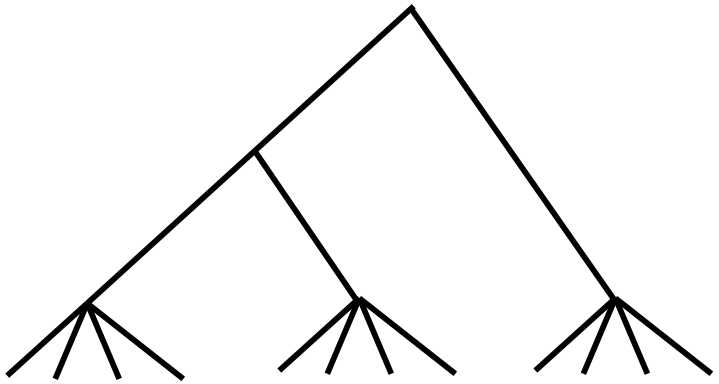
MaxMD is well motivated. It's applicable to an arbitrary distance matrix (no need for a tree).

Selecting a set under MaxMD.

MaxMD only depends on the closest pair of EUs.

Selecting a set of size 4.

Two possible choices:



MaxMD is well motivated. It's applicable to an arbitrary distance matrix (no need for a tree).

GreedyMMD selects EUs that are spread out and it has the property of stability.

GreedyMMD

Selecting a subset of EUs under MaxMD using a greedy approach.

GreedyMMD (d,k):

I. Select the two most distant EUs.

II. Sequentially add EUs that maximize MD until the resulting set is of size k.

If d satisfies the triangle inequality, then GreedyMMD is a 2-approximation to the optimal solution (Tamir, 1991; Ravi et al., 1994).

This approximation is sharp even if d is a tree metric (Bordewich, Rodrigo, S 2008).

GreedyMMD

Selecting a subset of EUs under MaxMD using a greedy approach.

GreedyMMD (d,k):

I. Select the two most distant EUs.

II. Sequentially add EUs that maximize MD until the resulting set is of size k.

If d is an ultrametric, then GreedyMMD returns an optimal set of EUs under MMD and, moreover, this set also maximizes PD.

(Bordewich, Rodrigo, S 2008)