Phylogenetic diversity: theory and computational challenges

ALLAN WILSON CENTRE

- Joint work with…


Charles Semple


Klaas Hartmann


Fabio Pardi


Arne Mooers


Beata Faller


Vince Moulton

Evolution: June 2007, Christchurch

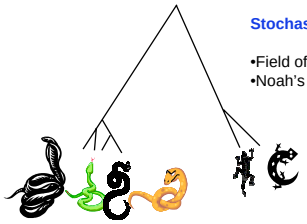
Outline of talk

Combinatorial

- Algorithms
- Ecological complications

Stochastic

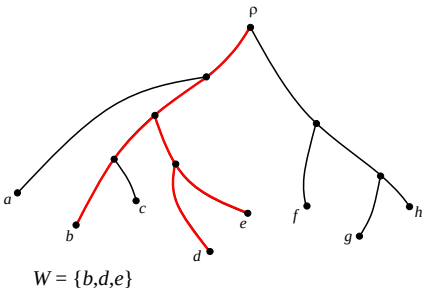
- Field of bullets model
- Noah's Ark problem



Greedy vs Global?

Part 1: Combinatorics

$$uPD(W) = \sum_{e \in T(W)} l(e) \qquad PD(W) = uPD(W \cup \{\rho\})$$

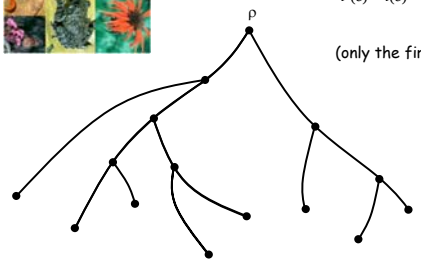


$W = \{b,d,e\}$

Interpretation of branch lengths

- $l(e)$ = Evolutionary time
- $l(e)$ = Genetic divergence
- $l(e)$ = Morphological divergence
- $l'(e) = l(e) + cf(e,x)$.

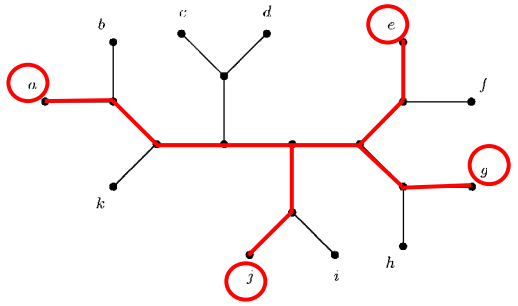
(only the first will be 'clock-like')



Optimisation problem

- **Problem:** Given a phylogenetic tree T on X with branch lengths.
- Find a subset Y_{max} of X given size k to maximise PD.

Greedy strategy



Faith 1992

Greedy works….

Theorem [S., 2005; Pardi & Goldman, 2005]

The greedy algorithm will always choose a maximum PD subset of X with size k.

Notes

• Nee and May showed this result holds for rooted trees with a clock [*Science* 1997]

• Minh *et al.* give $O(n\log k)$ and $O(n+(n-k)\log(n-k))$ algorithms, $n=|X|$ [*Syst. Biol.*, 55, 769-773 2006; Phylogenetic Diversity within Seconds]

… but why does greedy work?

Its….a greedoid!

- The subsets of X of given size that maximize PD form a 'strong greedoid'
- Proposition: For any two subsets A, B of X with $B|<|A|$ there exists x in A - B so that

$$PD(A-\{x\})+PD(B\cup\{x\})\geq PD(A)+PD(B)$$

Extensions (I)

- Ensuring some taxa are in the PD set.
- Which taxa are in ***all*** max. PD sets?
- Which taxa are in ***at least one*** max. PD set?

Extensions (II)

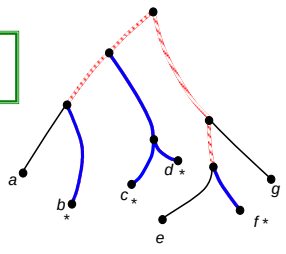
- Suppose each species has a 'cost' and a total budget B. How should B be spent to maximize PD?
- If all costs equal this is just the previous problem
- [F. Pardi and N. Goldman, *Syst. Biol.* 56(3):431–444, 2007]
  A dynamical programming approach works ('pseudo-poly' time by discretizing B).

A related measure

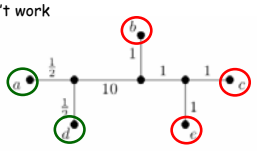
Exclusive molecular phylodiversity (ED)

Lewis, L.A. and Lewis, P. O. (2005). *Systematic Biology*, 54(6), 936-947.

ED(S) = PD(X) - PD(X-S)

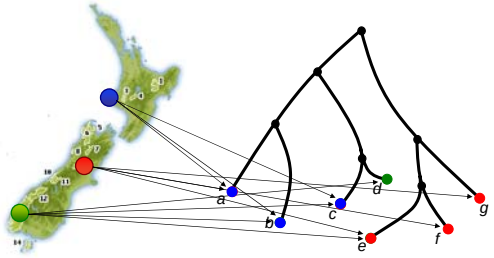


Greedy doesn't work



ED({*a*,*d*}) =15-4=11 (maximal)
ED({*b*,*c*,*e*} = 15-1 = 14 (maximal)

Nature reserve selection problem



- Each region can be conserved at some cost.
- Total budget B available
- ***Problem:*** Which regions should be made reserves (within budget B) so as to maximize the PD of conserved species?

Can we solve the NRSP?

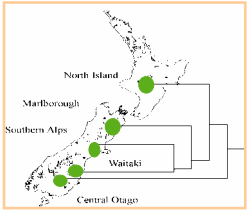
- Greedy doesn't work
- The problem is computationally intractable (NP-hard)
- Can be converted into the integer linear program (MCLP) Rodrigues and Gaston (Biol Conserv. 2002):

Theorem [Bordewich and Semple 2007]
There is a fast algorithm that will find a solution that's within a factor of $1- e^{-1}$ (=63%) of the optimal solution

Getting a better approximation is impossible (unless P=NP).

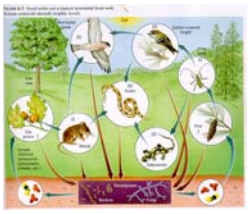
Optimizing diversity with coverage

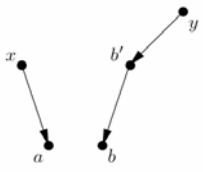
- Each region consists of a monophyletic group of species
- In each region r we want to maintain at least N_r species
- ***Problem:*** Find a set of k species satisfying this constraint and having maximal PD score.



Theorem: This problem can be solved quickly (using a greedy-type approch)

Handling dependencies





The collection of viable sets of size at most 3 is

$$\mathcal{F} = \{\{a\}, \{b\}, \{a, b\}, \{a, x\}, \{b, b'\}, \{a, b, b'\}, \{a, b, x\}, \{b, b', y\}\}.$$

Optimizing diversity with dependencies

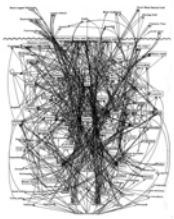
- ***Given:*** Phylogenetic tree, ecological network, k
- ***Problem:*** Find a viable subset of k species of maximal PD

Does greedy work?

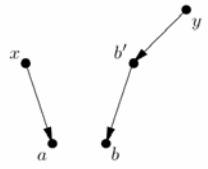
Fact 1: The viable subsets of X form a greedoid.

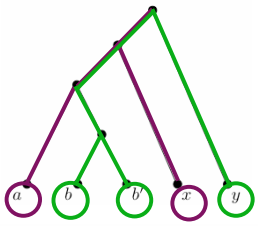
This means the problem can be solved by greedy algorithm in a special case:

(i) the tree is a star and
(ii) the branch lengths are clocklike.



Does greedy work?





The collection of viable sets of size at most 3 is

$\mathcal{F} = \{\{a\}, \{b\}, \{a, b\}, \{a, x\}, \{b, b'\}, \{a, b, b'\}, \{a, b, x\}, \{b, b', y\}\}$

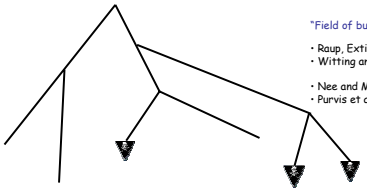
So…

Mol. Clock is not enough (nor is star by itself!)

In general greedy works if the tree and ecological network are `compatible'.

Part 2: Stochastic

Loss of PD under extinction



"Field of bullets" model

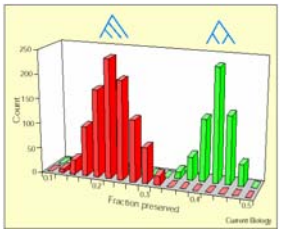
- Raup, Extinction 1993
- Witting and Loeschcke, *Biol. Conserv.* 1995
- Nee and May, *Science* 1997
- Purvis et al., *Science* 2000

ρ_i = rate of extinction of species i

Simple FOB model: $\rho_i = \rho$ (so $p_i = p$). [Nee and May, 1997]

p_i = probability species i is present after time T: $p_i = \exp(-\rho_i T)$

Distribution of PD under FOB



Simulations

Nee and May 1997

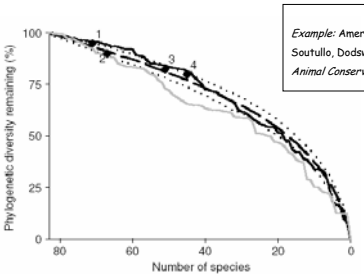
12 species from a 64-leaf tree

(Fig. Vazquez and Gittleman 1998)

- Effect of tree shape
- Normal curve

Theorem [Faller and S, 2007] PD has (asymptotically) a **normal** distribution on a large tree, under a FOB model (even when p_i's vary), under mild assumptions

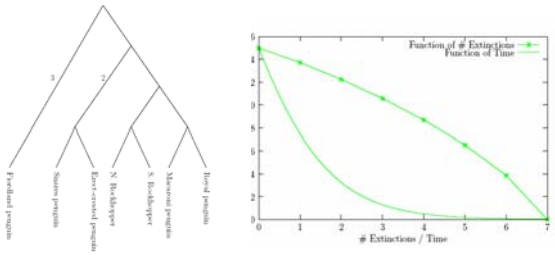
Loss of expected PD



Example: American carnivore PD
Soutullo, Dodsworth, Heard, Mooers, *Animal Conservation* 2005

Theorem: This concave relationship holds for any tree with any branch lengths under the field of bullets model (it is linear only for a 'star' tree).

PD loss vs extinctions or vs time?



Conservation implications

The Noah's Ark Problem

Econometrica, Vol. 66, No. 6 (November, 1998), 1279–1298

THE NOAH'S ARK PROBLEM

BY MARTIN L. WEITZMAN

Given: Phylogenetic tree with branch lengths, budget B, survivial profile for each species.

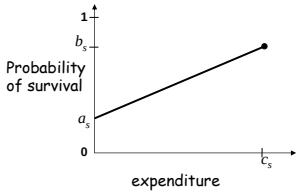
Question: Find an optimal allocation $x = (x_1,\ldots,x_n)$ of expenditure within budget B so as to maximize expected future PD

"Expected future PD"

$= \Sigma_S$ PD(S) x Prob(S is the set of species that survive given allocation x)

$= \Sigma_e$ $l(e)$ x Prob(at least one species below e survives given allocation x).

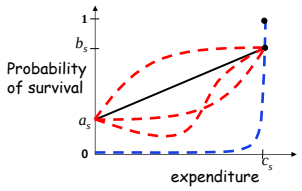
The Noah's Ark Problem



$a_s \xrightarrow{c_s} b_s$ f - NAP

Theorem [Weitzman 1998] For the f-NAP problem there is an optimal allocation which is 'extreme' – i.e. x_s is either 0 or c_s for all (but possibly one) species.

The Noah's Ark Problem



$a_s \xrightarrow{c_s} b_s$ f - NAP

$a_s \xrightarrow{c_s} b_s$ g - NAP

$a_s \xrightarrow{c_s} b_s$ nf - NAP

Note: The $0 \xrightarrow{1} 1$ nf - NAP is just the PD-optimization problem

$a_s \xrightarrow{c_s} b_s$ NAP can be transformed to $0 \xrightarrow{c_s} b_s$ NAP

Noah's Ark Problem [Results]

Theorem [Hartmann+S, *Syst. Biol.* 2006] The NAP can be exactly solved by the greedy algorithm is some special cases.

$a_s \xrightarrow{c} 1$ nf - NAP

$a_s \xrightarrow{c_s} 1$ nf - NAP if branch-lengths are clocklike

Theorem [Pardi+Goldman 2007, *Syst. Biol.* 56(3):431–444, 2007]

$a_s \xrightarrow{c_s} 1$ nf - NAP

Can be solved by a dynamic programming algorithm in (pseudo-)polynomial time (in budget).

On-going work and questions

- Performance on data sets (contact Klass Hartmann Google-funded SoC project)
- Can we solve more general versions of the NAP, for example to handle ecological constraints/dependencies etc?
- Is maximizing expected PD the 'right' objective

  (or eg. Prob(PD > x))?

References

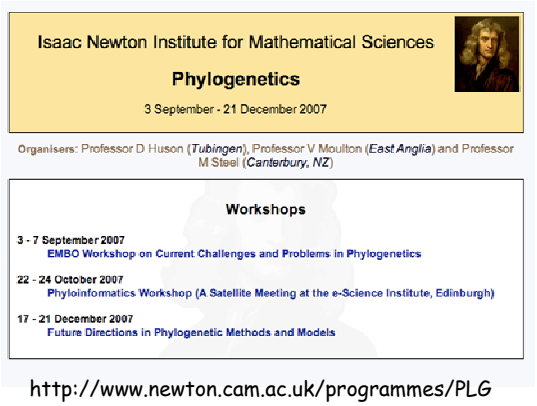
Moulton, V., Semple, C., Steel, M. Optimizing phylogenetic diversity under constraints. *Journal of Theoretical Biology* 246, 186--194 (2007).

K. Hartmann and M. Steel. (2006). Maximimizing phylogenetic diversity in biodiverstity conservation: greedy solutions to the Noah's Ark problem. *Systematic Biology* 55(4), 644-651.

F. Pardi and N. Goldman (2007), Resource-Aware Taxon Selection for Maximizing Phylogenetic Diversity, *Systematic Biology* 56(3):431–444.

K. Hartmann and M. Steel, Phylogenetic diversity: From combinatorics to ecology. Book chapter for: *Reconstructing evolution: New mathematical and computational approaches* (eds. O. Gascuel and M. Steel) Oxford University Press

Bordewich, M. and Semple, C., 2007. Nature reserve selection problem: A tight approximation algorithm, submitted manuscript.



http://www.newton.cam.ac.uk/programmes/PLG