Phylogenetic diversity: from combinatorics to conservation

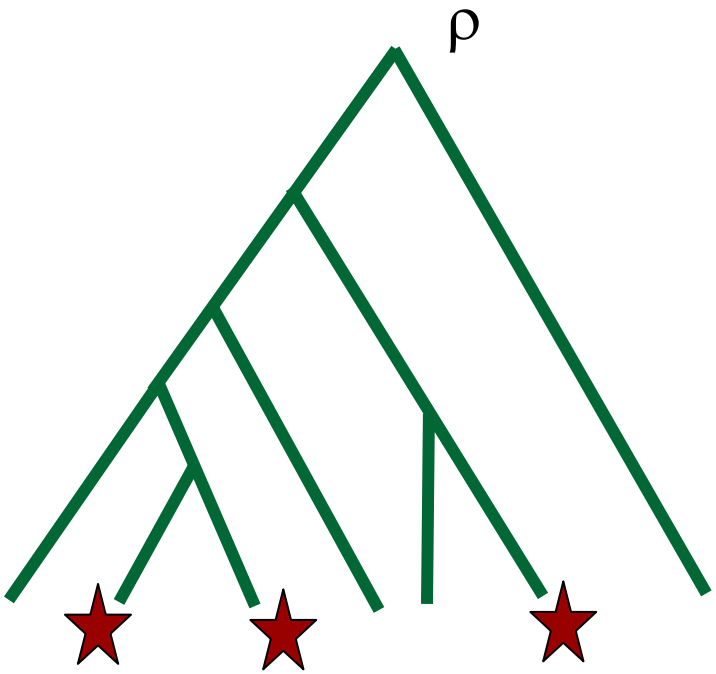
Mike Steel

Allan Wilson Centre for
Molecular Ecology and Evolution
&
Biomathematics Research Centre
University of Canterbury,
Christchurch, New Zealand

ALLAN WILSON CENTRE

Biomathematics Research Centre

PD – what is it?

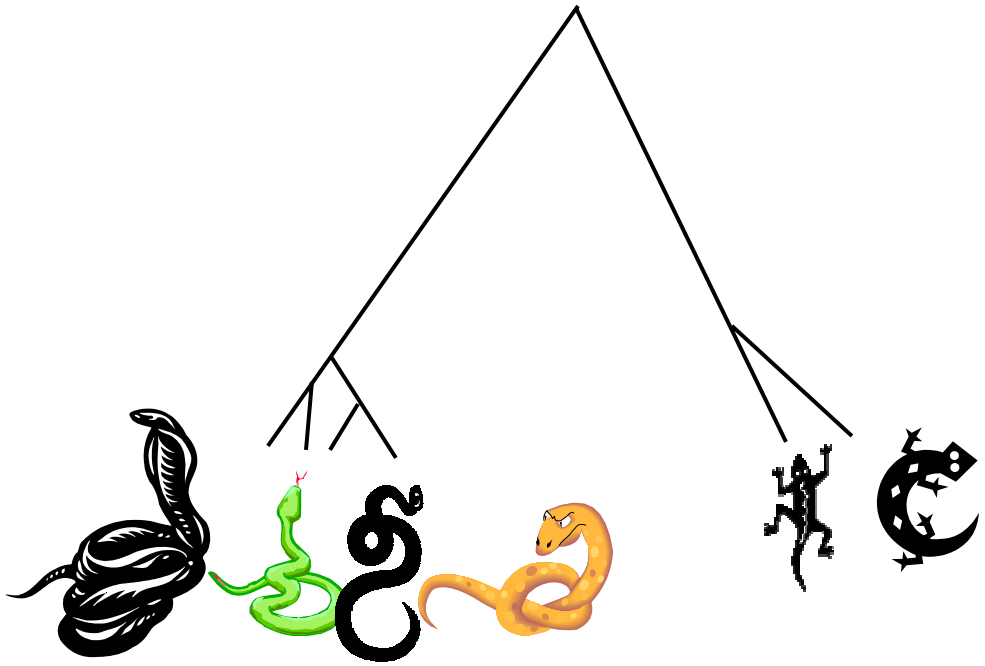


$$PD(W) = \sum_{e \in T(W \cup \rho)} l(e)$$

(D. Faith 1992)

Why do we care?

- Biodiversity conservation
  - Quantifying and comparing biodiversity in different regions/countries etc, and how much is at risk
    [Mooers, Heard, Chrostowski, 2005]
  - Measuring loss of biodiversity (current extinction rate >10—100 times neutral rate) [Prim et al. *Science*, 269: 347—350]).
  - Conservation planning and strategies
- Genomics (L. Pachter *et al.*)
- Phylogeny reconstruction

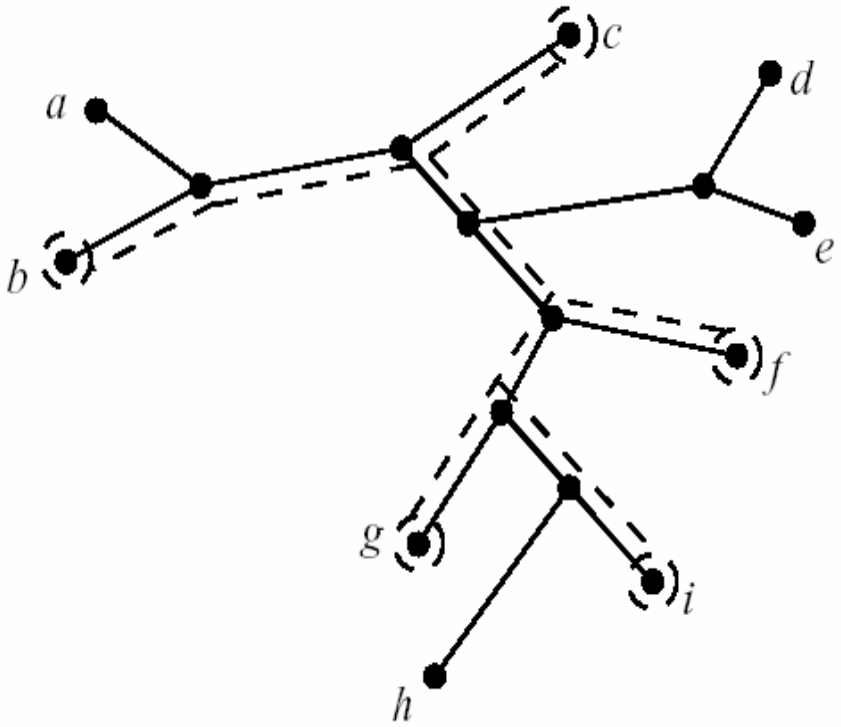


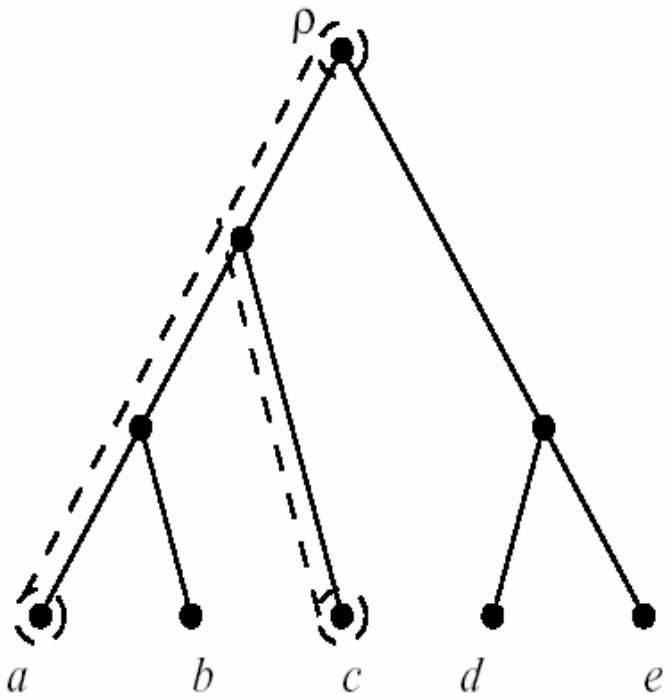
Modifications (I)

- Unrooted trees (with arbitary branch lengths)

$$PD'(W) = \sum_{e \in T(W)} l(e)$$

$$PD(W) = PD'(W \cup \{\rho\})$$





Modifications (II)

Weight function on taxa $f : X \to R$

★ $PD_f(W) = PD(W) + a \cdot \sum_{x \in W} f(x)$

Can regard this as just additional lengthening/shortening of pendant edges

Optimisation problem

- **Problem:** Given a phylogenetic tree T on X with edge weights.

- Find a subset Y_{max} of X given size k to maximise PD.

Nee and May (Science 1997)

For rooted trees with a clock, and standard PD, the greedy algorithm solves this problem

General case?

A combinatorial property

- Proposition: For any two subsets A, B of X with $2 \leq |B| < |A|$ there exists x in A - B so that

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

- Corollary: Y_{max} can always be found by using the 'greedy algorithm'

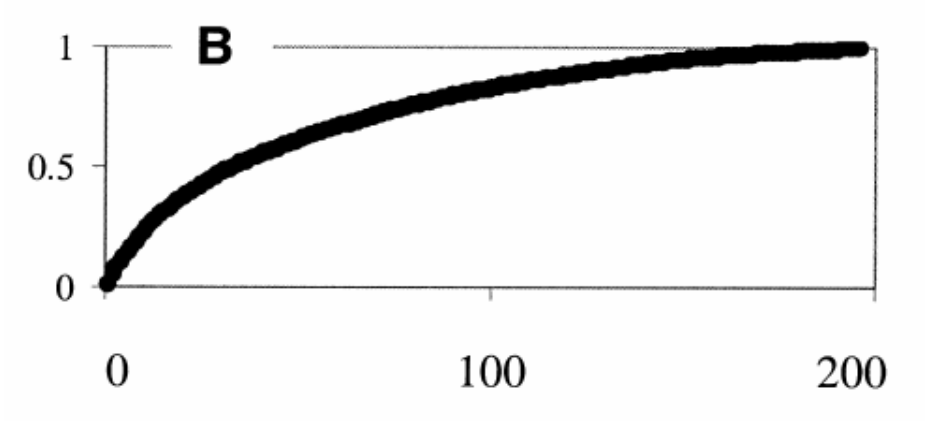
(and in general setting $PD_f(W) = PD(W) + a \cdot \sum_{x \in W} f(x)$)

Applications

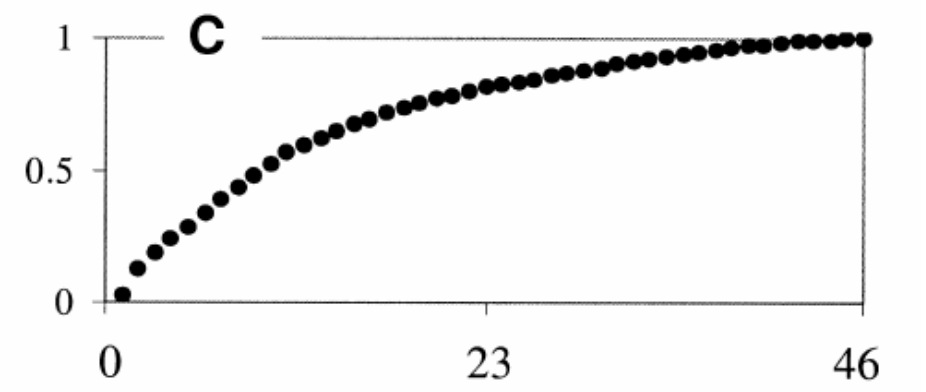
- Ensuring some taxa are in the PD set.
- Which taxa are in ***all*** max. PD sets?
- Which taxa are in ***at least one*** max. PD set?
- Subset Y of X that maximizes

$$PD_f(W) = PD(W) + a \cdot \sum_{x \in W} f(x)$$

$PD(Y_{max})$ vs $|Y_{max}|$



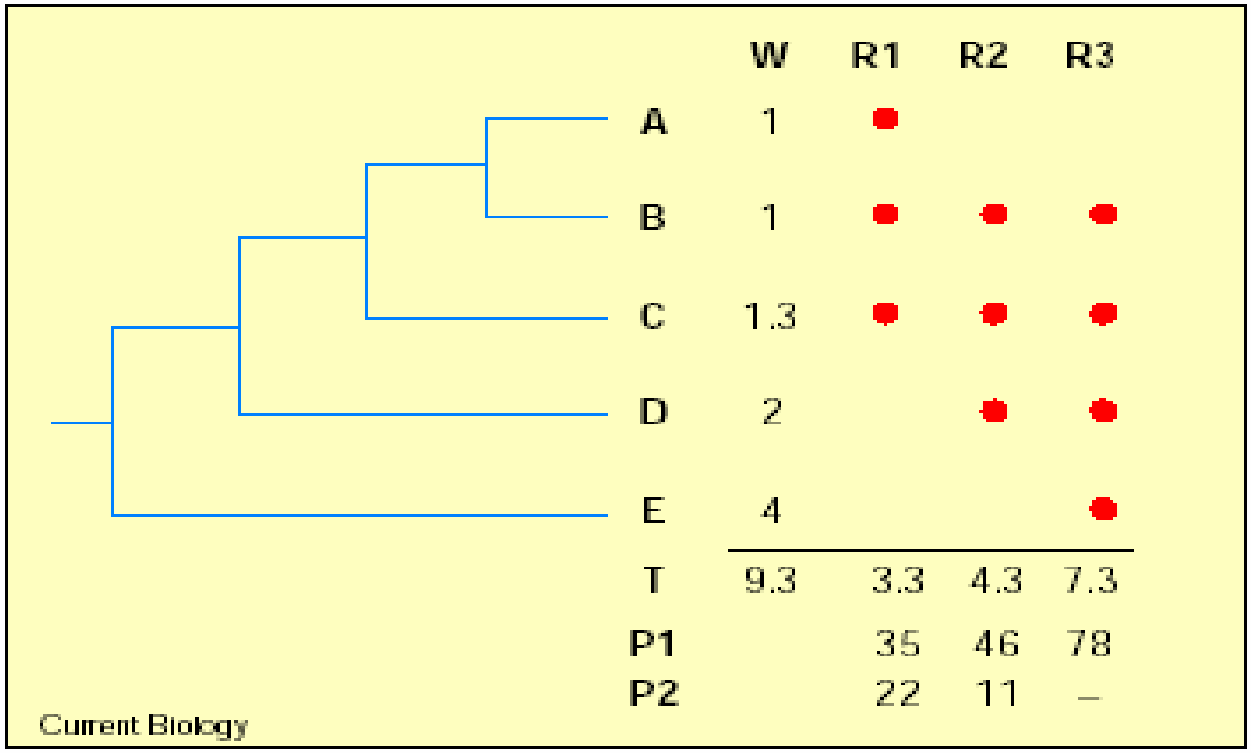
NZ birds



NZ forest birds

From G. Barker, 2002

Related measures

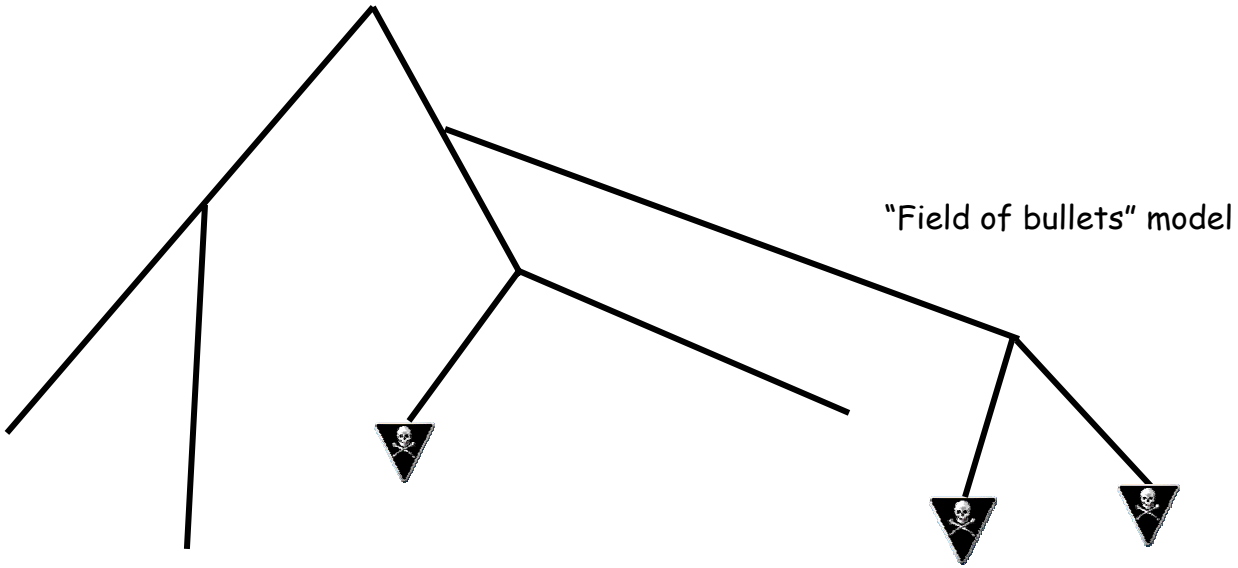


From Vazquez and Gittleman 1998

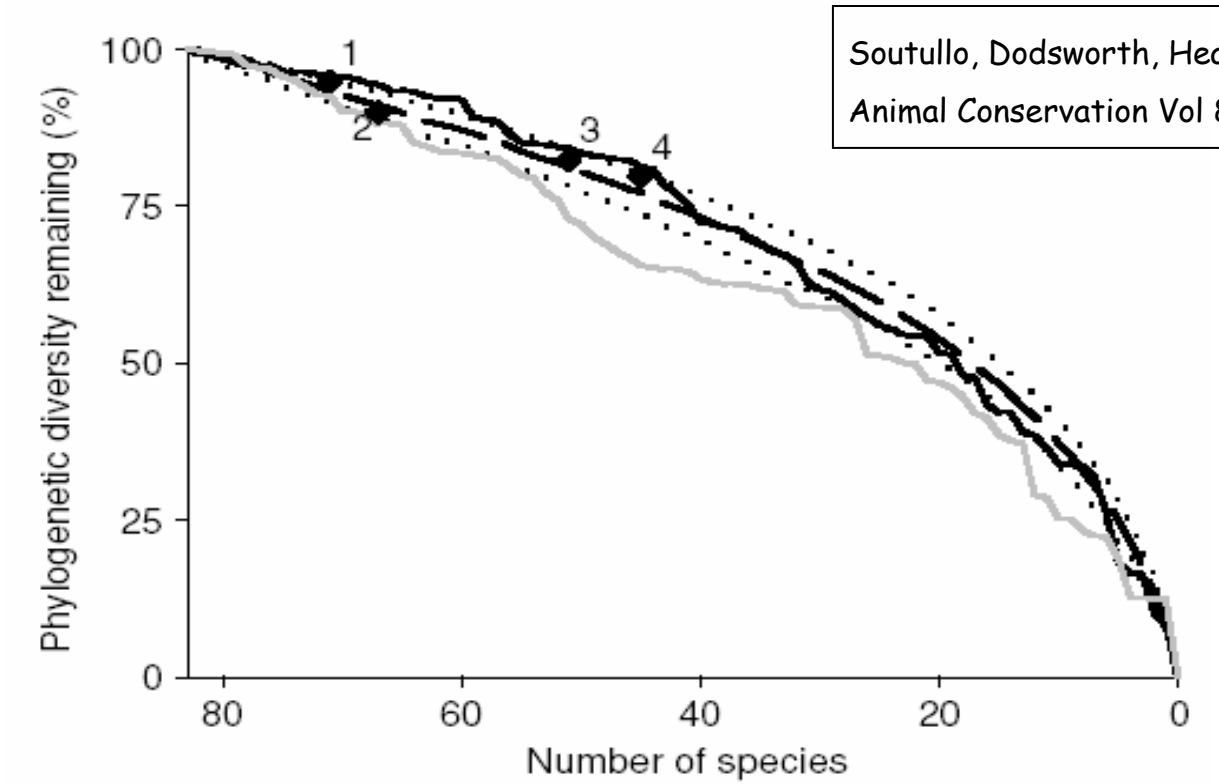
Moulton ,Semple, S ('05) The associated optimization problem is NP-hard

Loss of PD under extinction

(Nee and May 1997 Science)



Example (American carnivore PD)

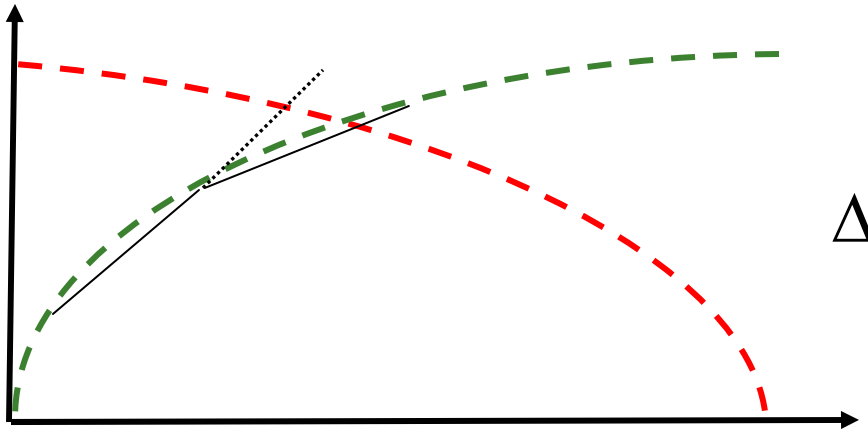


Soutullo, Dodsworth, Heard, Mooers,
Animal Conservation Vol 8 2005

Question: Is concavity of PD loss generic?

Concavity

Expected PD
$\mu_r = E(PD(W_r))$



$$\Delta\mu_r = \mu_r - \mu_{r-1}$$

No. taxa, r No. taxa deleted

Strictly Concave	$\Delta\mu_r - \Delta\mu_{r+1} > 0$
Concave	$\Delta\mu_r - \Delta\mu_{r+1} \geq 0$

Note: μ_r is (strictly) concave if and only if μ_{n-r} is

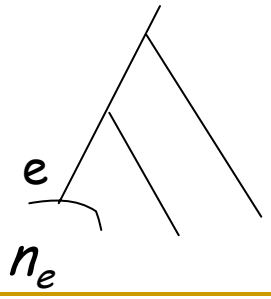
$$\Delta\mu_r - \Delta\mu_{r+1} = 2\mu_r - \mu_{r-1} - \mu_{r+1}$$

Theorem

$$\Delta\mu_r - \Delta\mu_{r+1} = \sum_e l(e)\psi(e,r)$$

$r = 1, 2, \ldots, n-1$, where

$$\psi(e,r) = \frac{n_e(n_e-1)\binom{n-n_e}{r-1}}{r(r+1)\binom{n}{r+1}}$$



- **Summary:** For **any** phylogenetic tree and **any** branch weights E[PD] is a concave function of the # of taxa deleted

Extreme cases

- The following are equivalent

$$\Delta\mu_r = \Delta\mu_{r+1} \quad \text{for all } r$$

T is a 'star' tree

- The following are equivalent

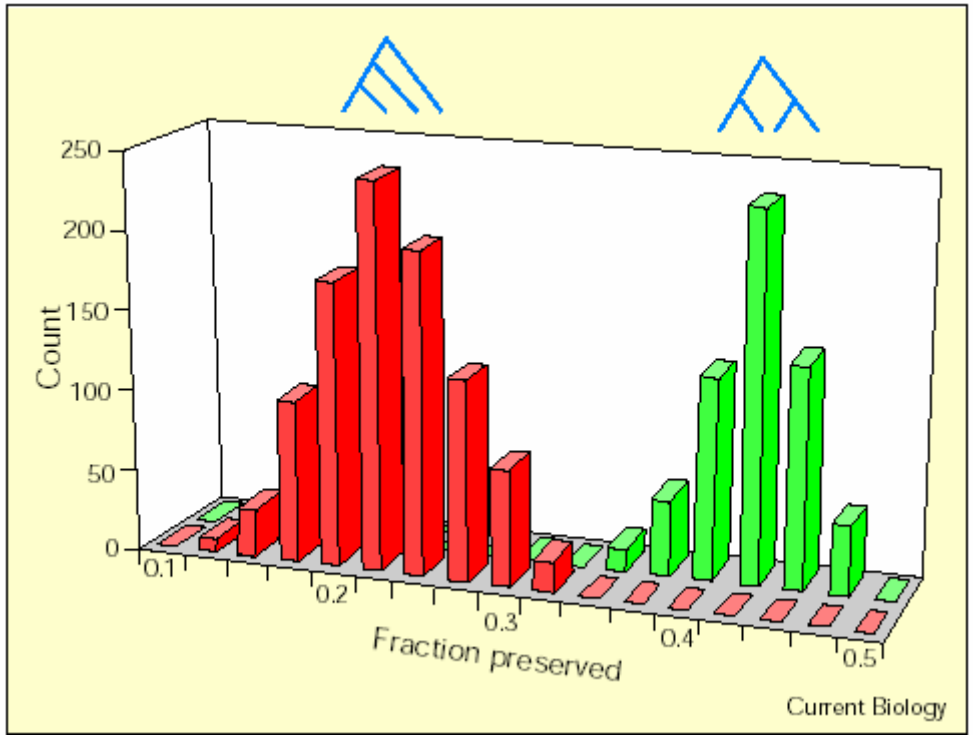
$$\Delta\mu_r > \Delta\mu_{r+1} \quad \text{for all } r$$

T has a cherry

Effect of tree shape

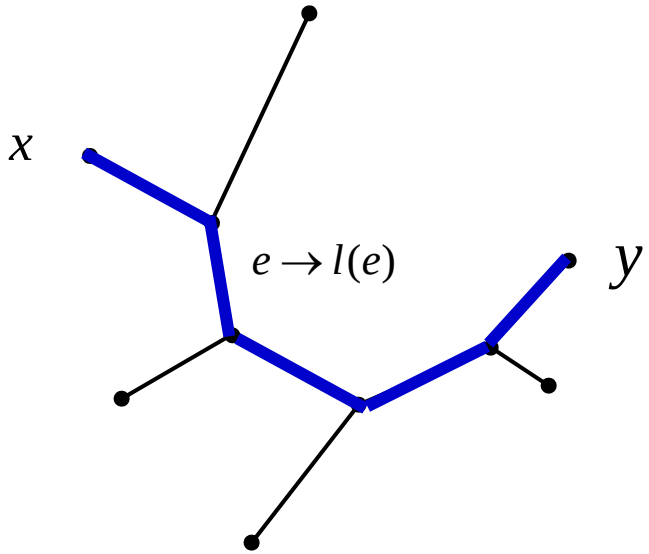
Nee and May 1997

12 species from a 64-leaf tree



From Vazquez and Gittleman 1998

Calculating PD



$$d(x,y) := \sum_{e \in p(T;x,y)} l(e)$$

$$l = l(T,w) := \sum_{e} l(e)$$

Theorem [Yves Pauplin 2000 *Molecular Biology and Evolution*]

$$l = \sum_{\{x,y\} \subseteq X} \left(\tfrac{1}{2}\right)^{\Delta(x,y)} d(x,y)$$

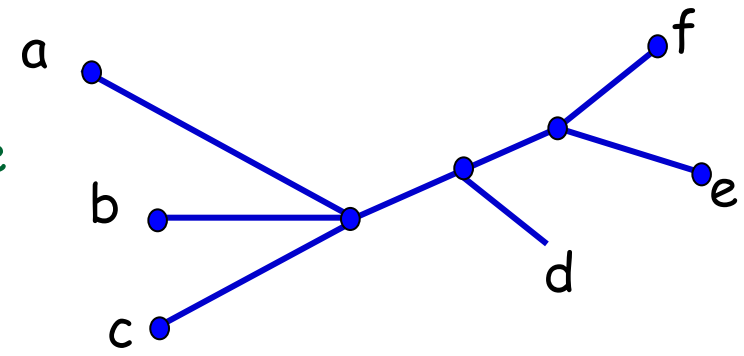
$$(= \tfrac{1}{16} d(x_1,x_2) + \ldots.)$$

Theorem (Semple+S 2004)

For any phylogenetic tree T

$$l = \sum_{\{x,y\}} \frac{1}{\prod_{v \in I(x,y)} (d(v)-1)} d(x,y)$$

Example



$$l = \frac{1}{3} d(a,b) + \frac{1}{6} d(a,d) + \ldots$$

PD for tree reconstruction

$$l = \sum_{\{x,y\}} \frac{1}{\prod_{v \in I(x,y)} (d(v) - 1)} d(x,y)$$

- Can be used as with δ in place of d as a tree reconstruction method (BME)
  - This method is consistent (Desper and Gascuel, 2004)
  - NJ selects the pair of leaves (at each step) to minimize the increase in BME score (Desper and Gascuel, 2004)

PD for tree reconstruction (II)

- (Classic result 1960s) phylogenetic X-tree T can be reconstructed from the $PD_T(Y)$ ($=d_{(T,w)}$) values for all 2-element subsets Y of X.

- Three-way distances (Joly, Le Calve) *J. Classif.* **12** (1995)

- (Pachter and Speyer 04): Any phylogenetic X-tree T can be reconstructed from the $PD_T(Y)$ values of all m-element subsets Y of X if $m \leq |X|+1/2$ (best possible).

Interlude…

Kaikoura06

The tenth Annual New Zealand Phylogenetics Conference
12th --17th February 2006

PD over an Abelian group G

What is a (Abelian) group?

A collection of objects you can add (and subtract) in a "sensible" way.
***Abelian*: $x+y=y+x$**

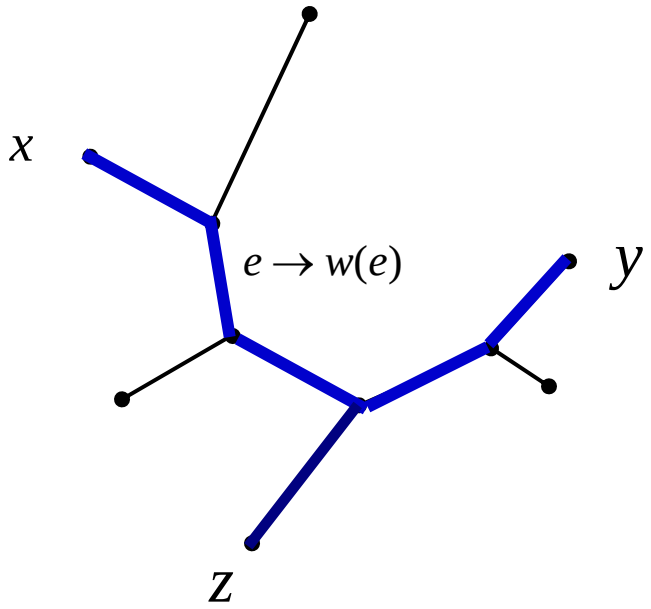
Usual example: The real numbers $\mathbf{R}$

Others: $\mathbf{R}^n$, $\mathbf{Z}$, $\mathbf{Z}^n$

In these groups $x+x=0$ implies $x=0$

(i.e. no "elements of order 2").

PD over G



$$d(x,y) := \sum_{e \in p(T;x,y)} w(e)$$

$$w(e) \in G$$

Warning!

Can have d(x,y)=0!

No NJ

No Pauplin

$$d(x,y,z) := \sum_{e \in T(T;x,y,z)} w(e)$$

Theorem

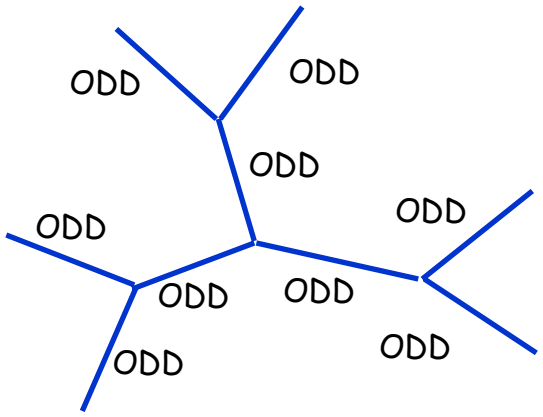
(Bandelt +S, 1994)

Provided

(1) w(e) is not 0_G for any edge e of T
(2) G has no elements of order 2,

any phylogenetic X-tree T can be reconstructed from the $PD_T(Y)$ $(=d_{(T,w)})$ values for all 2-element subsets Y of X.

Assume (1) from now



Not true if G has elements of order 2!

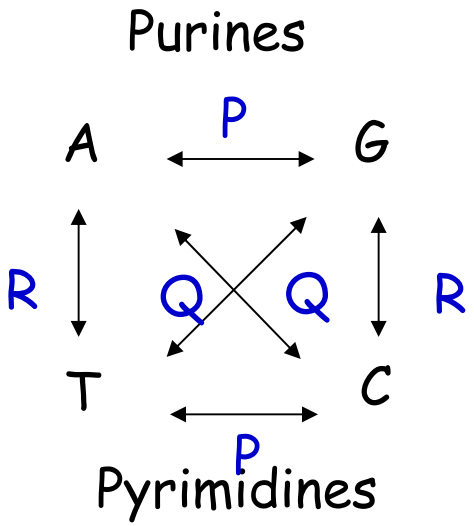
Is abandoning (2) just a technical exercise?

Examples (I)

R ⟷ Y

+	0	1
0	0	1
1	1	0

Z_2

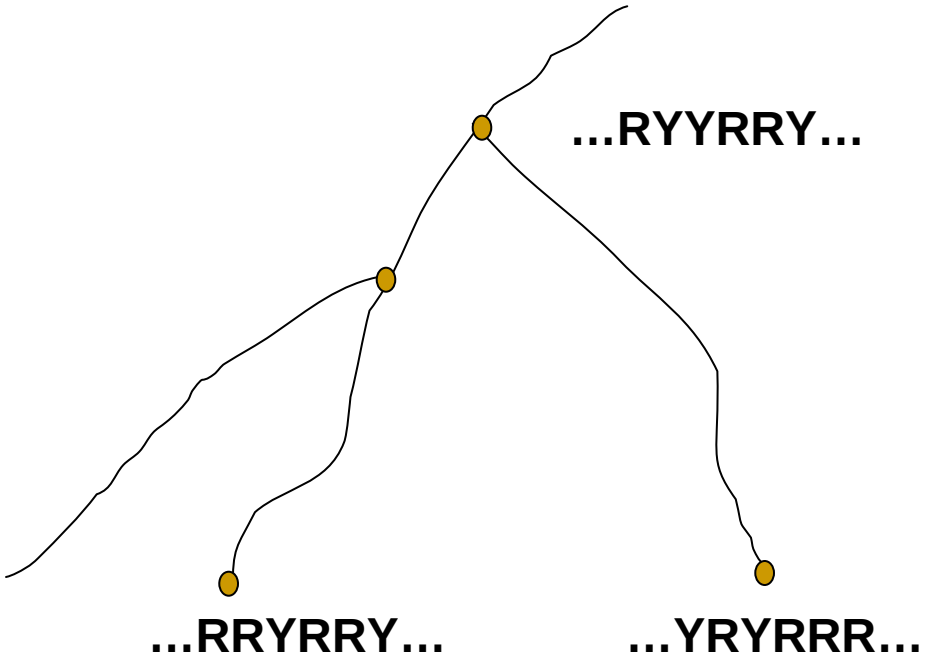


+	0	P	Q	R
0	0	P	Q	R
P	P	0	R	Q
Q	Q	R	0	P
R	R	Q	P	0

$Z_2 x Z_2$

Evans and Speed (1993)

Examples (II)

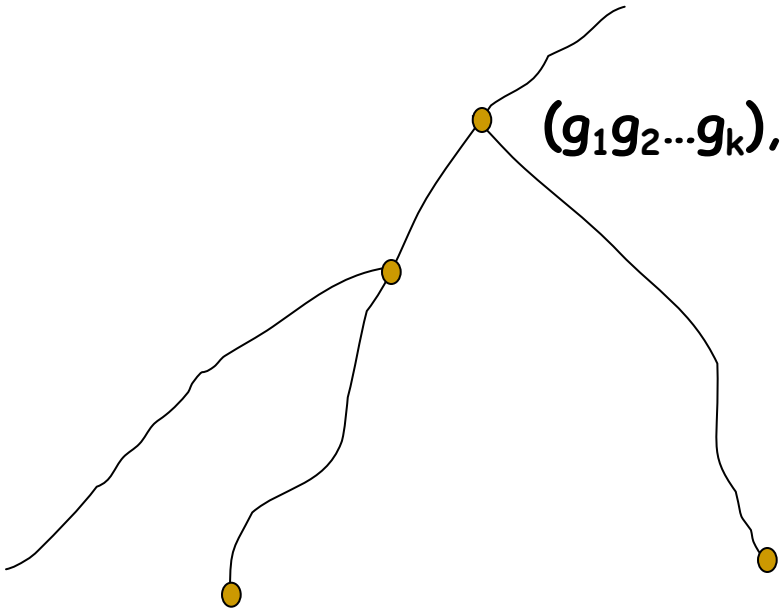


(….0…) + (….0…) = (….0…)
(….0…) + (….1…) = (….1…)
(….1…) + (….1…) = (….0…)

Z_2^k

Examples (III)

- Gene order



$(g_1 g_2 \cdots g_k),\ (g'_1 g'_2 \cdots g'_k),\ \ldots.$ $= (g^1,\ g^2, \ldots, g^k)$

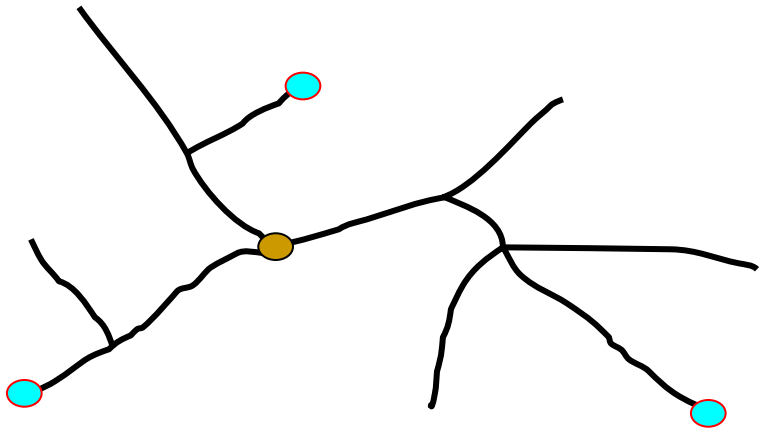
$(g_7 g_5 \cdots g_4),\ (g'_1 g'_3 \cdots g'_7),\ \ldots.$ $(g_2 g_1 \cdots g_k),\ (g'_6 g'_5 \cdots g'_3),\ \ldots.$

CHICKEN SCRATCHINGS!

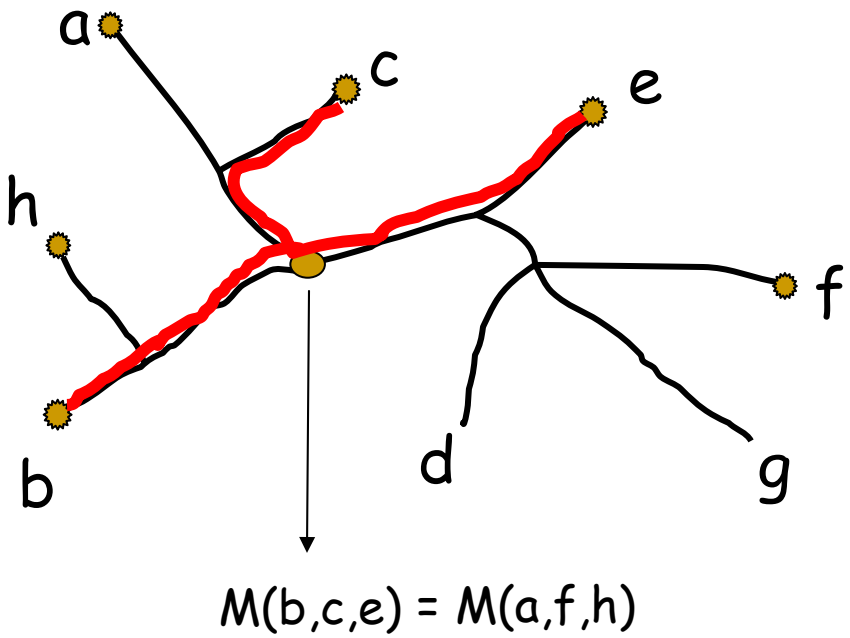
So what can we say?

Proposition (Dress+S, 2005)

For any Abelian group G, T can be reconstructed from the $PD_T(Y)$ values of all subsets Y of X of size ***at most three***.



Idea behind proof



Suppose that for every three taxa i,j,k, one can compute Obj(M(i,j,k)).
Then provided Obj(u) ≠ Obj(v) for all u,v T can be reconstructed

CHICKEN SCRATCHINGS!

Application: Binary sequences

- x = (…RRYRYRRYYYYRRYR…)
- y = (…YRYRRYRYYYRYRRR…)

d(x,y) = (…10001100011010…)

d(x,y) describes at which positions there has been a 'net substitution'.

- d(x,y,z) = ?

The median procedure

- Suppose we take d(x,y,z) to be the median sequence

(…0…), (…1…), (…0…)

When does d(A) = $PD_T(A)$ for all subsets A of X of size 2,3? (d has an "arboreal rep.").

Theorem

- d defined this way from sequences has an arboreal representation for some T if and only if the sites correspond to splits of T (i.e. no 'homoplasy').

Existence

- Proposition (from Bandelt+S, '94): if G has no el's of order 2, then d defined on pairs and triples has an arboreal rep. iff d satisfies a 3- and 4-point condition.

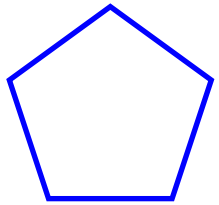
3-point condition: $$d(i,j)+d(j,k)+d(k,i)=2d(i,j,k)$$

4-point condition:

$$d(i,j)+d(k,l)=d(i,k)+d(j,l)=d(i,l)+d(j,k)+2g$$

Corollary: [G has no el's of order 2]. Then d has an arboreal rep. iff d|U does for all subsets U of X of size 4.

Question: Is this true when G has elements of order 2?



No!

Existence

Theorem (Dress+S, 2005)

For any Abelian group, d defined on pairs and triples of X has an arboreal rep. if and only if d satisfies the 3PC, **two** 4PC and one 5PC

Corollary

d has an arboreal rep. iff d|U has one, for all subsets U of X of size 5

Future questions

- Statistical treatment
- The non-abelian case?
  - Relevant to:
    - transition matrices (nucleotide substitution)
    - permutations (gene order)

(G(n), S(n))

- PD for phylogenetic networks?

Further details

• M Steel (2004). Phylogenetic diversity and the greedy algorithm. *Systematic Biology* (in press).

•M. Steel (2005). Tools to construct and study big trees: a mathematical perspective In *Towards the tree of life: taxonomy and systematics of large and species rich taxa* (in press).

•A. Dress and M. Steel (2005). Phylogenetic diversity over an abelian group (in preparation).

•C. Semple and M. Steel, (2004) Cyclic permuations and evolutionary trees, *Advances in Applied Mathematics 32*: 669—680.