Combinatorial approaches in phylogenetics

Mike Steel

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

ALLAN WILSON CENTRE

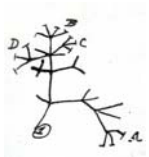
Hamburg, July, 2006

"Unreasonable effectiveness of mathematics" in physics (1960).

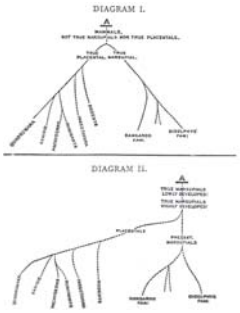
– Eugene Paul Wigner

"The lack of real contact between mathematics and biology is either a tragedy, a scandal or a challenge, it is hard to decide which."

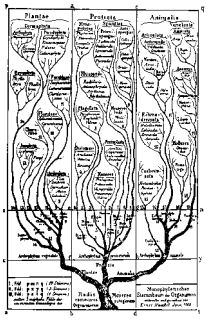
– Gian-Carlo Rota, (1986, in *Discrete thoughts*)



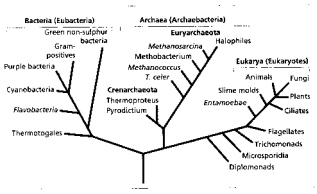
First Notebook on Transmutation of Species, 1837.



Letter from Darwin to Lyell, 1860.



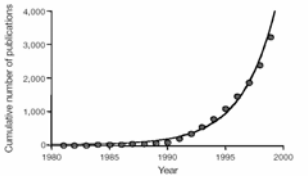
Ernst Haeckel (1866)

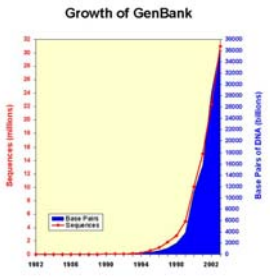


Olsen and Woese (1993)

The genetics era

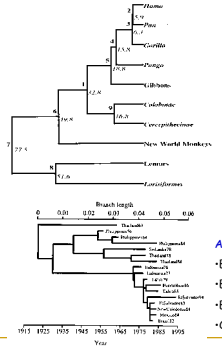
- 1967 Walter Fitch and Emil Margoliash constructed phylogenetic trees from cytochrome c sequences from vertebrates that agreed well with the vertebrate fossil record.





Publications with "molecular" and "phylogenetic" in abstract

Phylogenetic trees

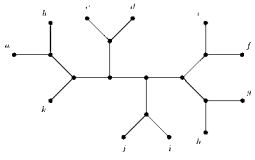


Applications:

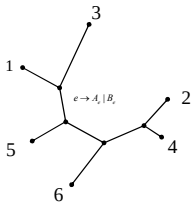
- Evolutionary Biology
- Ecology
- Epidemiology
- Others (language, stemmatology etc)

Phylogenetic trees

- [Definition] A **phylogenetic X-tree** is a tree $T=(V,E)$ with a set X of labelled leaves, and all other vertices unlabelled and of degree ≥ 3.
- If all non-leaf vertices have degree 3 then T is **binary**



Trees and splits



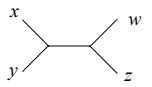
$$\Sigma(T) = \{A_e \mid B_e : e \in E\}$$

Partial order: $(P_X, \leq)$

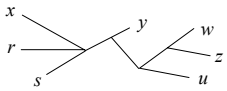
$$T \leq T' \Leftrightarrow \Sigma(T) \subseteq \Sigma(T')$$

Quartet trees

• A **quartet tree** is a binary phylogenetic tree on 4 leaves (say, *x,y,w,z*) written *xy|wz*.



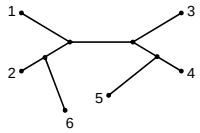
• A phylogenetic X-tree **displays** *xy|wz* if there is an edge in *T* whose deletion separates {*x,y*} from {*w,z*}



Compatibility

A set *Q* of quartets is **compatible** if there is a phylogenetic *X*-tree T that displays each quartet of *Q*

- Example: *Q*={12|34, 13|45, 14|26}



Complexity?

Defining sets

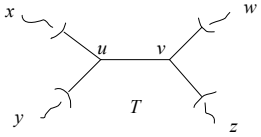
If *T* is the only phylogenetic *X*-tree that displays *Q* (and *X*= L(*Q*)) then we say *Q* **defines** *T* .

- Let *Q*(*T*) be the set of **all** quartets displayed by (any) *T*.

If *T* is binary, then *Q*(*T*) defines *T*.

A necessary condition for *Q* to define *T*

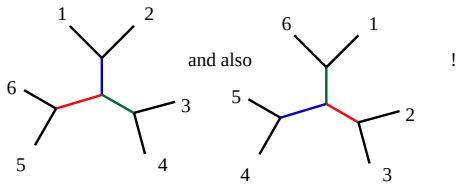
- **Definition:** For a binary phylogenetic tree *T*, a collection *Q* of induced quartet trees *distinguishes* an interior edge {*u,v*} of *T* if there exists a quartet *xy|wz* in *Q* that looks like this:



Observation: If *Q* defines *T* then *T* is binary and *Q* distinguishes every interior edge of *T* (so $|Q| \geq n-3$).

Warning:

Q ={12|45, 56|23, 34|16} distinguishes each interior edge of the tree:



and also

!

Sufficient condition for Q to define T:

- Suppose Q is compatible and distinguishes every interior edge of a binary phylogenetic X-tree T.

Proposition: If there is an element of X that is a leaf of every tree in Q then Q defines T.

Corollary:
There are subsets of $Q(T)$ of size |X|-3 that define T.

Reconstructing trees from characters

Types of "characters"

- Morphology (eg. Wings vs no-Wings)
- DNA sequences (….ACG….)
- Genomic data (gene order, SINEs, RCGs)

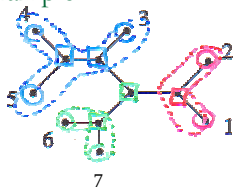
A *character* on X $f : X \to S$

Definitions:

- [Character] A character is any function
  $f : X \to S$
- [Convexity] Given a character $f : X \to S$

and a phylogenetic X-tree $T=(V,E)$, we say f is **convex on *T***
if f extends to $f' : V \to S$
so that $f' | X = f$
and $\{v \in V : f'(v) = s\}$ is connected for all s in S.

Convexity: example



x	1	2	3	4	5	6	7
$f(x)$							

17

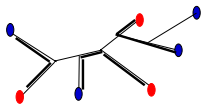
Biological significance of convexity

- Lemma: A character χ is convex on a phylogenetic tree T if and only if χ could have evolved on T (from any root vertex) without any **reversals** or **convergent evolution.**

18

Homoplasy

- $h(f, T)$ = smallest number of reversals/convergent events required to fit f on (rooting of) T.



- $h(f, T)$ easily computed = $l(f, T) - (|f(X)| - 1)$
- For $|f(X)|=2$, h= -1+ max edge-disjoint proper path packing (by Menger's theorem) [S90; extention to $|f(X)|>2$ by ES92]

- $h(f, T)$ = min #SPR operations to transform T into a tree on which f is convex. [Bruen and Bryant 2005]. **Applications**

19

Combinatorial aside 1.

- **Theorem** [Bruen and Bryant 2006]
- $h(T, f)$ = min #SPR operations to transform T into a tree on which f is convex.

Let

$$h(f_1, f_2) := \min_T \{h(T, f_1) + h(T, f_2)\}$$

Construct the partition intersection graph.
Theorem [Bruen and Bryant 2006]
$h(f_1, f_2)$ = #edges + #components – r_1 -r_2

20

Relevance to molecular biology

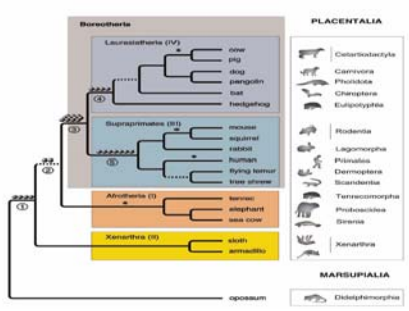
- Large state space
  Example: gene order rearrangements (*n* species, *L* genes, random inversion model)

$$g_1 g_2 \boxed{g_3 g_4 g_5} g_6 g_7 \cdots \Rightarrow g_1 g_2 g_5 g_4 g_3 g_6 g_7 \cdots$$

$$P[h=0] \geq 1 - \frac{2(2n-3)(n-1)}{L(L-1)}$$

- Rare genomic characters (RGC's):
  Examples, Retroposons, SINES, LINES, LTRs, gene content, etc
  (Model, 0->1->?)

Recent example (Kreigs *et al.* PLoS biology, April 2006. Tree of placental mammals)



Character compatibility

- [Compatibility] Characters $f_1, f_2, \ldots, f_k$ are ***compatible*** if there exists a phylogenetic X-tree (a 'perfect phylogeny') on which they are all convex.

- Complexity: NP-hard, but special cases are solvable in polynomial time.

A link to graph theory…

G is **chordal** if every cycle of length four or more has a chord

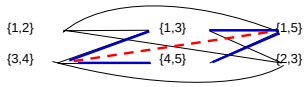
Example

Definition:

- Given $G = (V,E)$ and a partition $V = V_1 \cup V_2 \cup \cdots \cup V_k$ a **restricted chordal completion** of G is any chordal graph $H = (V, E'), E \subseteq E'$
  satisfying $x, y \in V_i \Rightarrow \{x, y\} \notin E' - E$

Characterising compatability

Species	1	2	3	4	5
Characters					
f_1	A	A	B	B	X
f_2	C	E	C	B	B
f_3	U	R	R	S	U

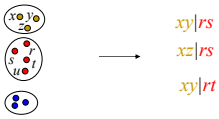


Theorem

- F is compatible if and only if int(F) has a restricted chordal completion.
- If F is compatible, then int(F) has tree-width at most k=|F|.

Equivalence of character and quartet compatibility

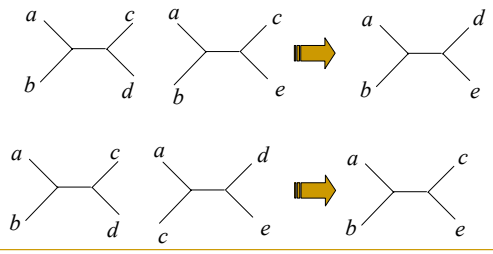
$$C \rightarrow Q(C)$$



Lemma: Each character in C is convex on T if and only if T displays all the quartets in $Q(C)$.

[C is "compatible", C "defines" T iff $Q(C)$ does]

New quartet trees from old ones



Dyadic rules for quartet trees

(Colonius and Schulze; Dekker)

(**Q1**): $\{ab|cd, ab|ce\} \vdash ab|de$

(**Q2**): $\{ab|cd, ac|de\} \vdash ab|ce, ab|de, bc|de.$

Any phylogenetic X–tree that displays the quartet trees on the left of (**Q1**) or (**Q2**) also displays the corresponding quartet tree(s) on the right.

Dyadic quartet closure

$$\mathcal{Q} = \mathcal{Q}_1 \subseteq \mathcal{Q}_2 \subseteq \cdots \subseteq \mathcal{Q}_m = \mathrm{qcl}_\theta(\mathcal{Q})$$

where $\mathcal{Q}_{i+1}$ consists of $\mathcal{Q}_i$ together with all additional quartets that can be obtained from a pair of quartets in $\mathcal{Q}_i$ by applying the rule(s) allowed by θ.

For $\theta \subseteq \{1, 2\}$, let the dyadic quartet closure under rule θ, $\mathrm{qcl}_\theta(\mathcal{Q})$, denote the minimal set of quartet trees that contains $\mathcal{Q}$ and is closed under rule (**Qi**) for each $i \in \theta$.
We denote these closures with: $\mathrm{qcl}_1(\mathcal{Q}), \mathrm{qcl}_2(\mathcal{Q}), \mathrm{qcl}_{1,2}(\mathcal{Q})$.

The closure of a set of quartets

For a compatible set $\mathcal{Q}$ of quartet trees, the *closure* $\mathrm{cl}(\mathcal{Q})$ is defined as

$$\mathrm{cl}(\mathcal{Q}) = \bigcap_{\mathcal{T} \in \mathrm{co}(\mathcal{Q})} \mathcal{Q}(\mathcal{T})$$

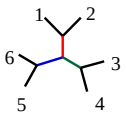
where $\mathrm{co}(\mathcal{Q})$ is the set of phylogenetic trees that display each of the trees in $\mathcal{Q}$.
Thus $\mathrm{cl}(\mathcal{Q})$ consists of precisely those quartet trees that are displayed by every phylogenetic tree that displays $\mathcal{Q}$.

- Rules of order $<p$ (for any fixed p) do not suffice compute $\mathrm{cl}(Q)$.
- There is a set Q that is incompatible but every strict subset Q' is compatible and satisfies $\mathrm{cl}(Q')=Q'$

Example 1: qcl_2

Definition: If Q distinguishes every interior edge of a binary phylogenetic tree T and we can order Q so that each quartet tree in the ordering introduces precisely one new leaf label, we say Q has a **tight ordering** for T.

Example: {12|35, 13|56, 15|34}.



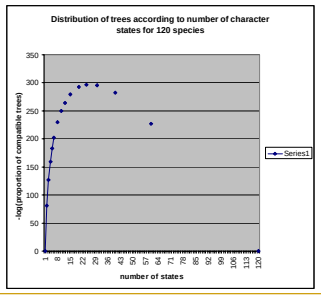
Proposition:
If Q has a tight ordering for T, then $\mathrm{qcl}_2(Q) = Q(T)$
In particular Q defines T.

Application: How many characters are needed to define a binary phylogenetic X-tree?

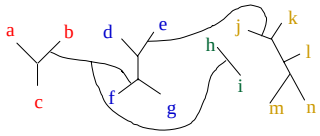
- For binary characters we need n-3 (n=|X|).
- For r-state characters (r fixed) we need at least (n-3)/(r-1)
- What if r is not fixed?

(it is not useful to make r too large!)

I(χ) := - log(Pr[χ is convex on random T])



Where do these numbers come from?

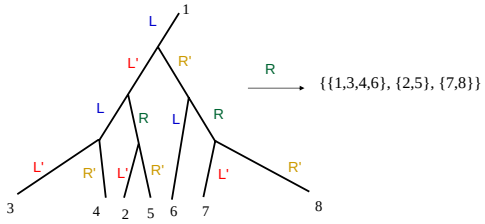


Carter *et al.* (1990); Erdös & Székely (1993).

binary phylogenetic trees with n leaves, $b(n) = 1 \times 3 \times \cdots \times (2n-5)$

of these on which χ is convex = $b(n)\prod_{i=1}^{r} b(a_i+1) \,/\, b(n-r+2)$

Edge-colouring a tree by $Z_2 \times Z_2$



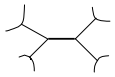
Theorem (Huber, Moulton, S, 2003)

For any tree these ***four*** characters defines T.

Proof: show $Q(C)$ contains a subset with a tight ordering for T.

Application 2: "Short" quartets

- $Q_{short}(T)$



- **Theorem** (Erdös *et al.* 1997)

$Q_{short}(T)$ contains a subset that has a tight ordering for T (and so $qcl_2(Q_{short}(T)) = Q(T)$).

- Application to show that trees can be reconstructed from 'short' sequences evolved under finite-state Markov process

A further application involving qcl_2:

We say Q is **excess-free** if $|L(Q)|-3-|Q| = 0$.
(note: If Q defines a tree, then $exc(Q) \leq 0$).

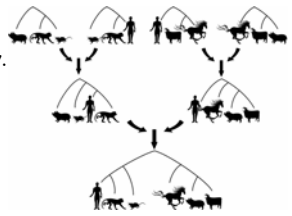
- **Proposition:** Suppose a subset Q of $Q(T)$ contains an excess-free subset Q_0 that defines T. Then $qcl_2(Q)=Q(T)$.

- Why? Let us say a set Q of quartet trees is "good" if (i) Q defines a phylogenetic tree, and (ii) $exc(Q)=0$.

  Theorem [Bocker, Dress 1999] Any good set of ($\geq$2) quartets is the disjoint union of precisely two good sets.

Aside: Phylogenetic patchworks

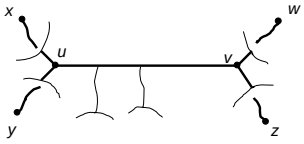
- If Q is 'good' then its good subsets for a 'patchwork', and it contains a maximal hierarchy.



New Theorem (2006; S. Grunewald).
Say Q is 'thin' if $exc(Q')\geq 0$ for all subsets of Q' of Q. Then any thin set is compatible.

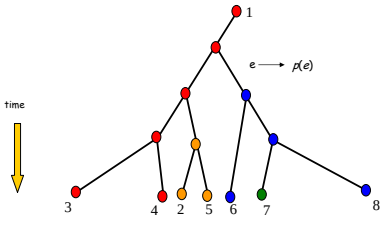
Example 2: qcl_1

- **Definition:** For a binary phylogenetic tree T, a collection Q of displayed quartet trees is a *generous cover* for T if for all pairs u,v of interior vertices of T, we have a quartet $xy|wz$ in Q that looks like this:



Theorem (Dezulian + S, 2003): If Q is a generous cover for T, then $qcl_1(T) = Q(T)$. Thus Q defines T.

Application: How many `evolved' characters are needed to reconstruct a tree?



1 2 3 4 5 6 7 8

Theorem (Mossel +S, 2004)

Elchanan Mossel

[Assume probability of state change on each edge for each character Bounded between (a,b), $0<a<b<1/2$]

The number k of indep. characters required to reconstruct T (correctly with probability>1-ε) is

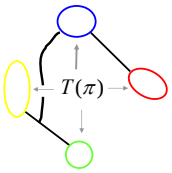
$$k = c \cdot \frac{\log(n)}{p}$$

- n = #species, p =smallest substitution probability, $c=c(\varepsilon,b)$
- The tree reconstruction algorithm is polynomial time (in n,k)

Proof relies on generous cover result.

Combinatorial aside: computing P(σ)

- Recursively
- Mobius-inversion (Evans *et al.* 2004)

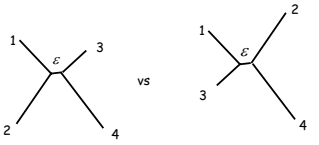


$$\prod_{e \in T(\pi)} (1 - p(e)) = \sum_{\pi \le \sigma} P(\sigma)$$

$$Q(\pi)$$

$$P(\pi) = \sum_{\pi \le \sigma} \mu(\pi, \sigma) Q(\sigma)$$

Topological aside: tree space under Markov models



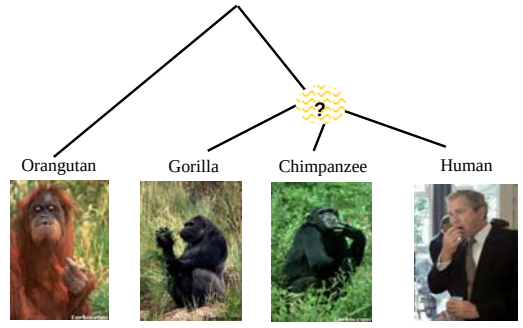
Infinite state Markov process

$$k \propto \frac{1}{\varepsilon}$$

Finite state Markov process

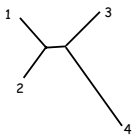
$$k \propto \frac{1}{\varepsilon^2}$$

Example

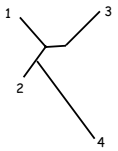


Adapted From the Tree of the Life Website, University of Arizona

Markov models (II)



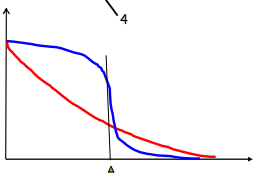
vs



- "site saturation"

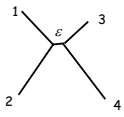
Persi Diaconis

Distance from uniform distribution

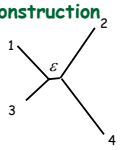


Number of moves

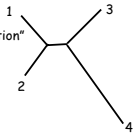
Markov models and tree reconstruction



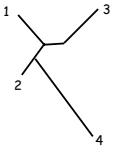
vs



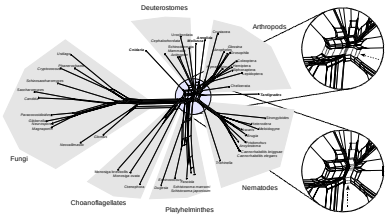
- "site saturation"



vs

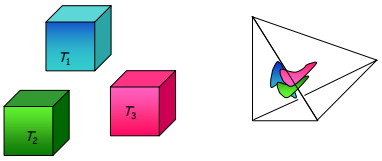


Example: Deep divergence in the Metazoan phylogeny



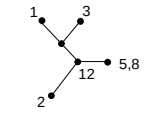
From Huson and Bryant, 2006

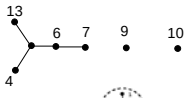
The space of 'phylogenetic oranges'



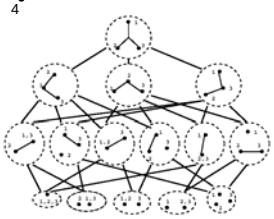
$$TS(X) = \coprod_{T} [0,1]^{e(T)} / \sim$$

Combinatorial aside (II) the Tuffley poset





Chris Tuffley



49

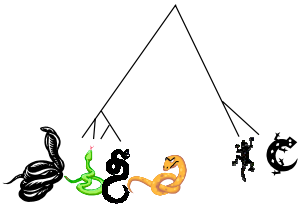
Theorem (2005)

$TS(X)$ is a compact regular cell complex which is

- contractible
- homeomorphic to the geometric realization of the Tuffley poset on X

50

Part 3: phylogenetic diversity and trees from distance data



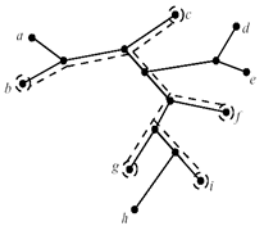
51

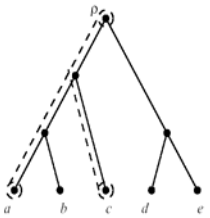
Modifications (I)

- Unrooted trees (with arbitary branch lengths)

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

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





52

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

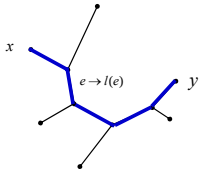
- Proposition: For any two subsets A, B of X with $2 \leq |B| < |A|$ there exists x in A - B so that

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

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

[The sets of maximal PD-score for their cardinality form a (strong) greedoid]

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*]

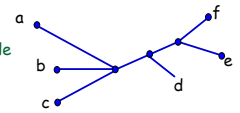
$$l = \sum_{\{x,y\} \subseteq X} \left(\frac{1}{2}\right)^{\Delta(x,y)} d(x,y) \qquad (= \tfrac{1}{16} d(x_1,x_2) + \ldots.)$$

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



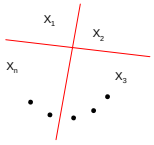
$$l = \frac{1}{3} d(a,b) + \frac{1}{6} d(a,d) + \ldots$$

CHICKEN SCRATCHINGS

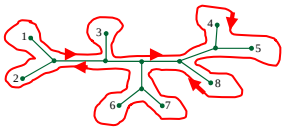
Cyclic Permutation on X

$\pi = (x_1, x_2, ..., x_n)$

$\Sigma^o(\pi) :=$ the splits (bipartitions of X) induced by planar 'cuts'.



[Definition] π is a cyclic ordering for T if $\Sigma(T) \subseteq \Sigma^o(\pi)$



$$l = \frac{1}{2}\sum_{i=1}^{n} d(x_i, x_{i+1})$$

57

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)

58

The end

Further details

•E. Mossel and M. Steel, A phase transition for a random cluster model on phylogenetic trees. *Mathematical Biosciences*, 187 (2004), 189-203.

•C. Semple and M. Steel (2004) Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics* 32(4): 669-680.

• V. Moulton and M. Steel (2004). Peeling phylogenetic 'oranges'. *Advances in Applied Mathematics* 33(4): 710-727.

•K. Huber, V. Moulton and M. Steel (2005). Four characters suffice to convexly define a phylogenetic tree. *SIAM Journal on Discrete Mathematics* 18(4): 835-843.

•M. Steel (2005). Phylogenetic diversity and the greedy algorithm. *Systematic Biology* 54(4): 527--529.

59