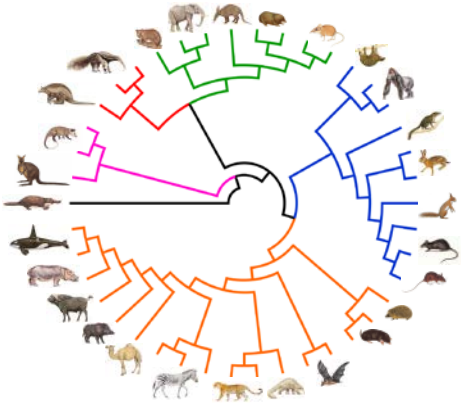
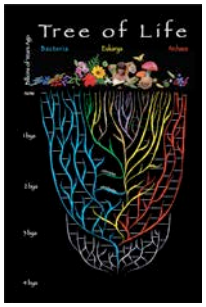
Lecture 1: Introduction to Phylogenetics



from F. Delsuc and N. Lartillot



Mike Steel

Biomathematics Research Centre

Winthrop lectures, 2014

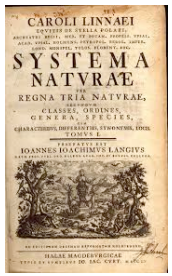
ALLAN WILSON CENTRE

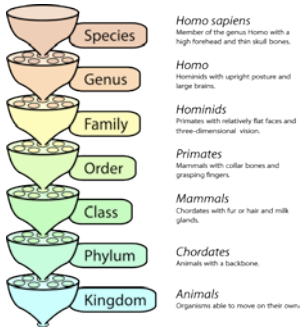
Outline

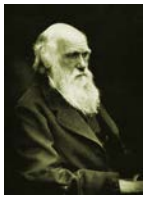
- *Part 1:* **History and background**
- *Part 2:* **Binary phylogenetic trees**
- *Part 3:* **Counting trees, tree shape, rearrangement**
  - **20x pushups**
- *Part 4:* **Specialist topic: models for tree shape**

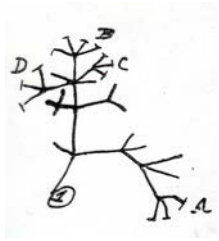
Pre-Darwin

Carl Linnaeus 1701-1778

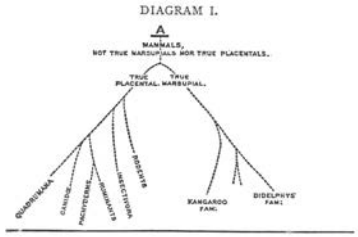




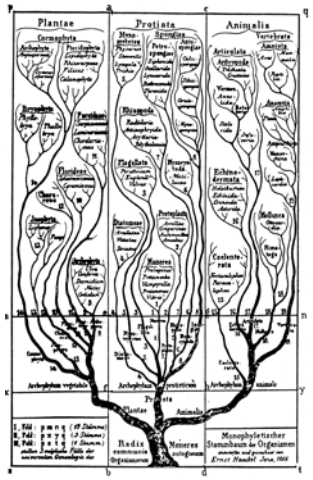




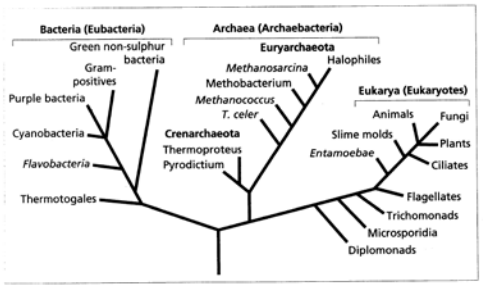
First Notebook on Transmutation of Species, 1837.



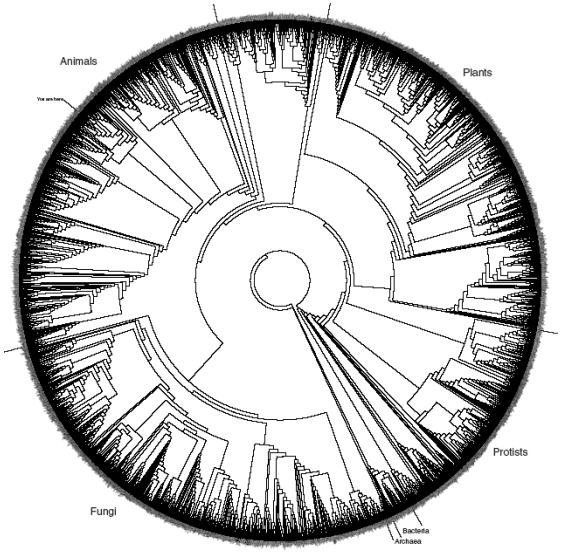
Letter from Darwin to Lyell, 1860.



Ernst Haeckel (1866)



Olsen and Woese (1993)



Hillis lab ~3000 taxa rRNA seq.

The genetic era: early pioneers

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

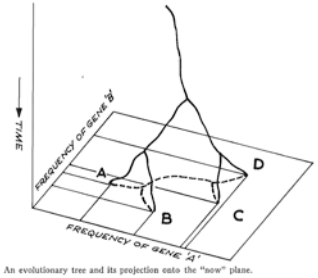


FIG. 1. An evolutionary tree and its projection onto the "now" plane.

L.L. Cavalli-Sforza and Anthony Edwards (late 1960s)

David Sankoff (mathematician)

Early 70s

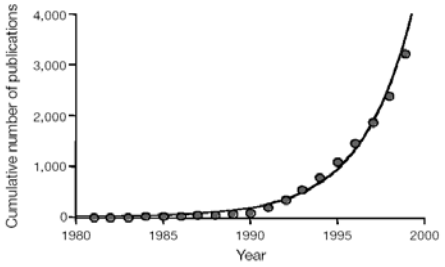
David Sankoff

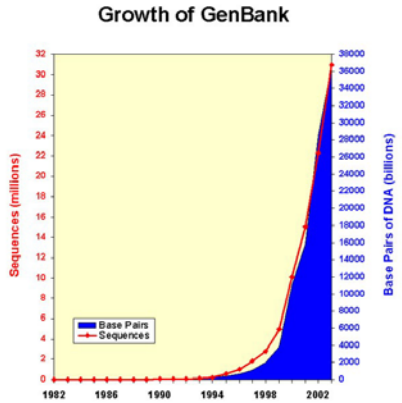
Joseph Felsenstein (statistics)

Peter Buneman;

Fred ('Buck') McMorris + George Estabrook (maths)

The growth of genomics/phylogenetics

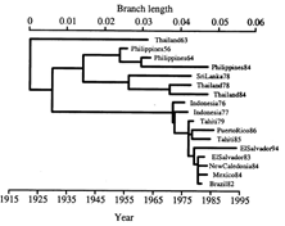


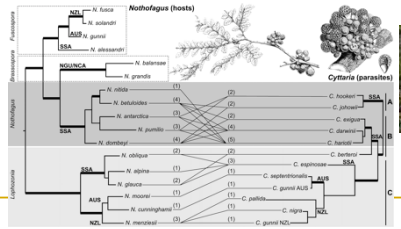


Publications with "molecular" and "phylogenetic" in abstract

Phylogenetic trees

Scientific American. Based on
Cann, Stoneking, Willson 1987 (*Nature*)

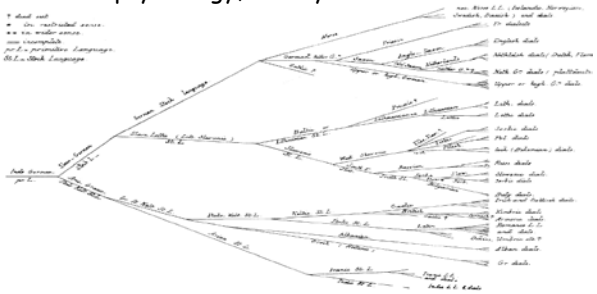




Co-phylogeny
study 2010

Further applications of phylogenetics

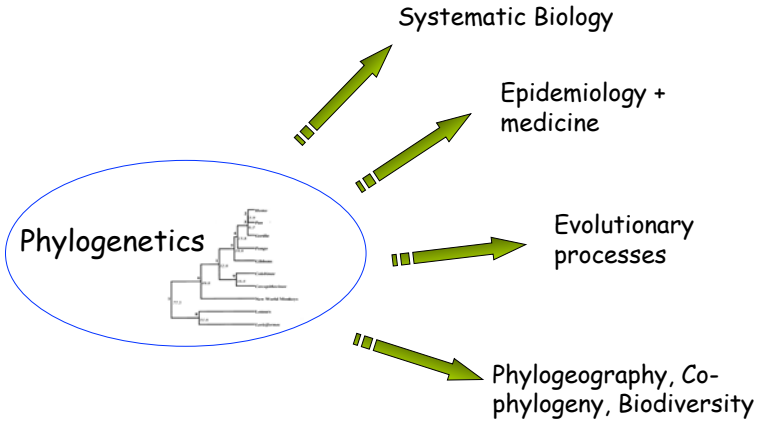
epidemiology, linguistics,
stemmatology,tumour-cell trees,
psychology, whisky

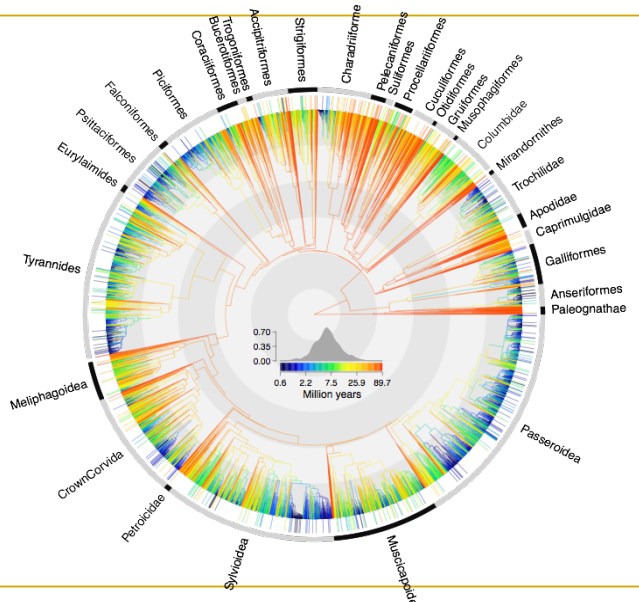


Stammbaum for Indo-Europena. From *Die Darwinshce Theorie und die Sprachwissenschaft* (Schleicher's 1863)

Figure 5: Cladogram for 46 manuscripts of the Wife of Bath's Prologue

1. Why phylogeny?



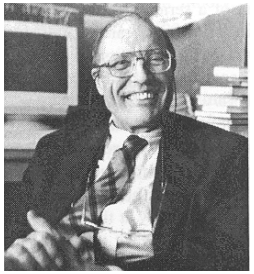


2014 'Tree of birds' ~(10,000 species) with ages colour coded (and with distribution of 575 impelled species at tips (rep. 2.7 billions years of evolution) [Jetz. et al. 2014]

2. Why maths?

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

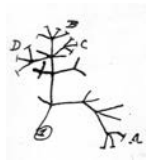
– Eugene Paul Wigner



"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*)

Why maths?



- Analysing existing methods
- Developing better methods
- Help answer questions:
  - Why do some methods lead to different estimated trees?
  - How can we have confidence in a given tree? (or *any* tree?)
  - What can trees tell us about evolutionary processes?
  - How much data do we need to find a tree?

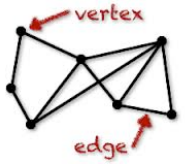
"I hope it arises from your being 10 fathoms deep in the **Mathematics**, & if you are God help you, for so am I, only with this difference**: I stick fast in the mud at the bottom and there I shall remain** "
– C. Darwin to W.D. Fox 29 July, 1828

What sort of math?

- **Discrete mathematics:**
  graph theory, posets, set systems, algorithms, computational complexity.
- **Probability:**
  Markov processes, birth-death processes, coupling, martingale theory, MCMC.

- **Others: algebra, dynamical systems:**
  linear algebra, algebraic geometry, discrete fourier analysis, differential equation modelling

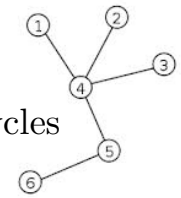
Graphs (and trees)



$$G = (V, E)$$

$$\sum_{v \in V} \deg(v) = 2|E|$$

$$G \text{ connected } \Rightarrow |V| \leq |E| + 1$$

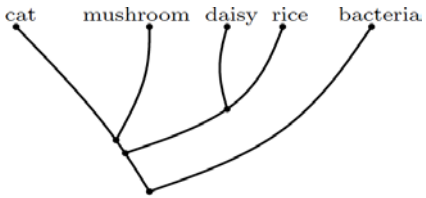


A **tree** is a connected graph with no cycles

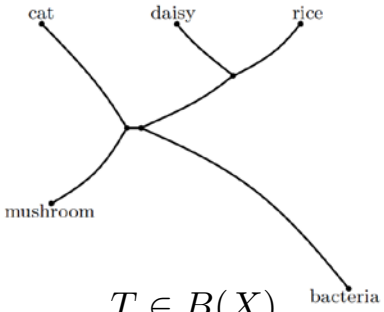
$$G = (V, E) \text{ (connected) is a tree } \Leftrightarrow |V| = |E| + 1$$

"vertex" aka "node"; "edge" aka "branch"

Binary phylogenetic trees (rooted and unrooted)



$T \in RB(X)$



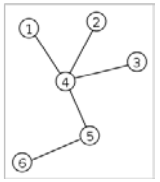
$T \in B(X)$

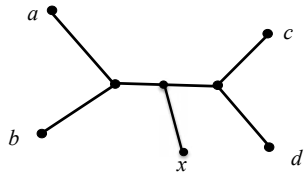
$T \cong T'$ if there is a graph isomorphism
$\varphi : V(T) \mapsto V(T') : \varphi(x) = x, \forall x \in X$

Counting trees

Arthur Cayley, 1889

$$n^{n-2}$$





Erwin Schroeder, 1870

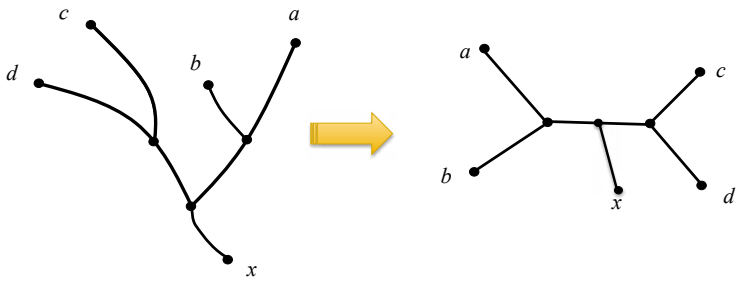
$$B(n) = B(\{1, 2, \ldots, n\}), b(n) = |B(n)|$$

$$T = (V, E) \in B(n) \Rightarrow |E| = 2n - 3$$

$$b(n+1) = b(n) \times (2n-3)$$

$$b(n) = 1 \times 3 \times \cdots \times (2n-5) = (2n-5)!!$$

Counting rooted trees



$$RB(X) \longleftrightarrow B(X \uplus \{x\})$$

$$rb(n) := |RB(X)| = b(n+1) = (2n-3)!!$$

$$\frac{(2n-2)!}{(n-1)!2^{n-1}}$$

An 'algebraic' proof

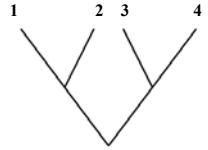
$$\phi(x) = \sum_{n \geq 1} rb(n) \frac{x^n}{n!}$$

$$\phi(x) = x + \frac{1}{2}\phi(x)^2$$

$$\Rightarrow \phi(x) = 1 - \sqrt{1-2x}$$

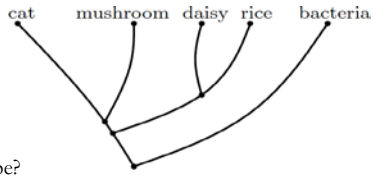
$$rb(n) \sim \frac{1}{\sqrt{2}} \left(\frac{2}{e}\right)^n n^{n-1}$$

Counting trees by shape (via the 'Orbit-Stablizer' theorem)

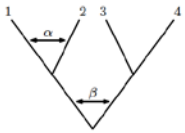


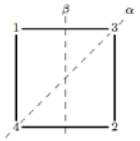
trees in $R(n)$ of this shape?

trees in $R(n)$ of this shape?



$$|O(s)| = \frac{|G|}{|\mathrm{Stab}(s)|} = \frac{n!}{Stab(T)} = n!2^{-s}$$

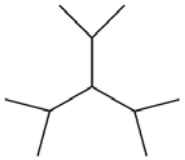




Quiz: how many trees have the same shape as the above?

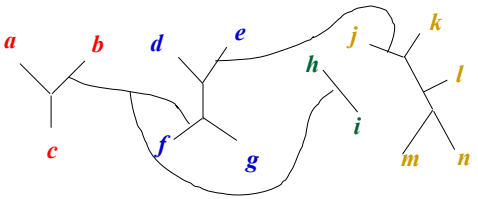
$$\frac{5!}{2^3} = 5 \times 3 = 15$$

Counting trees II



$$n!/|Stab(T)|$$

A more interesting type of counting:



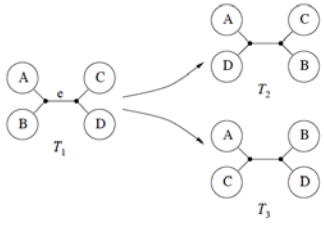
How many binary phylogenetic trees can we construct in this way?

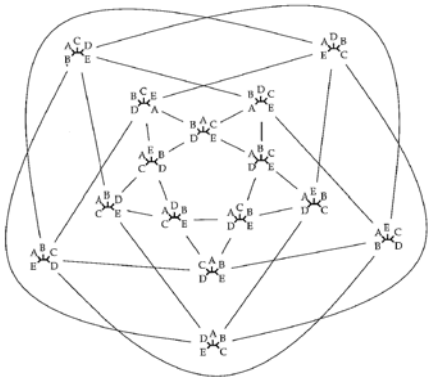
$$\frac{b(n)}{b(n-k+2)} \prod_{i=1}^{k} |E_i| = \frac{b(14)}{b(12)} \times 3 \times 5 \times 1 \times 7$$

$$50,715$$

A first look at tree rearrangment operations (NNI)

What does the space of trees look like look like?

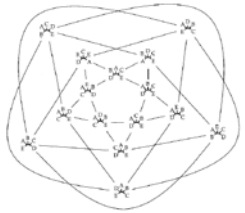




NNI tree space is connected

NNI neighborhood: 2(*n*-3)

Discrete tree space: interesting properties



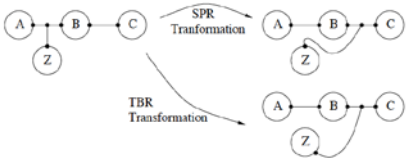
Diameter?

$$\max\{d_{NNI}(T,T')\} = \Theta(n\log(n))$$

Theorem [Gordon, Ford, St John, 2013]
For all *n*, there exists a Hamiltonian path through the *n*-leaf NNI tree-space.

Kevaughn Gordon*, Eric Ford, and Katherine St. John, Hamiltonian Walks of Phylogenetic Treespaces, to appear, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013 Jul-Aug;10(4):1076-9

Other tree rearrangement operations



SPR (**S**ubtree **p**rune and **r**e-graft)

TBR (**T**ree **b**isection and **r**econnection)

Number of neighbors

$$\text{SPR: } 2(n-3)(2n-7)$$

$$\text{TBR: } \Theta(n^2 \log n) - \Theta(n^3)$$

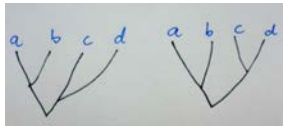
Diameter? $\Theta(n)$

$$[2(n-3)(2n-7)]^d \geq b(n) \Rightarrow d \geq cn$$

Specialist topic: Models for generating discrete random trees

Uniform model – select a tree from RB(n) uniformly at random

Yule-Harding model – select a **ranked** rooted binary phylogenetic tree uniformly at random, then forget the ranking.

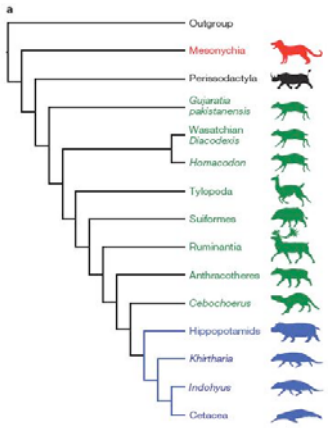


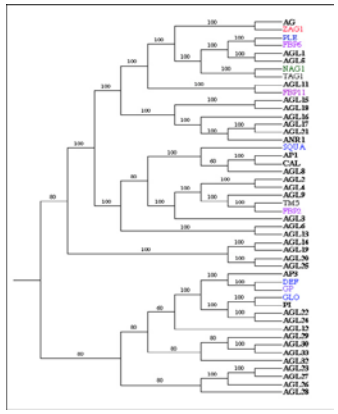
$$\#\text{ ranked trees on } n \text{ leaves } = \prod_{i=2}^{n} \binom{i}{2} = \frac{n!(n-1)!}{2^{n-1}}$$

Quiz: Do these two models produce same probability distribution on RB(n)?

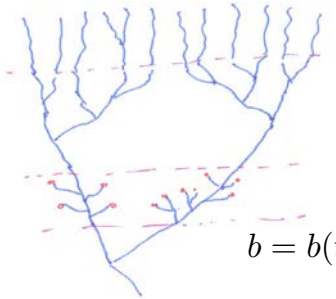
Why of interest?

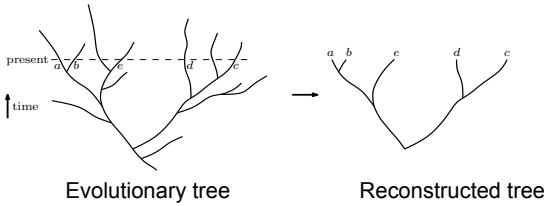
Discrete aspect of tree shape: Balance





All roads lead to ~~Rome~~ Yule-Harding....



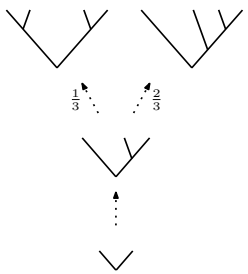


$$b = b(t, N), d = d(t, a, N)$$

Proposition: [Aldous; Lambert and Stadler]

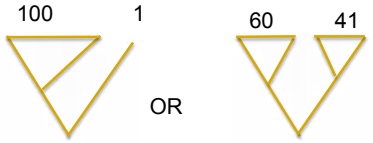
All such models lead to same distribution on the shape of the reconstructed tree (ignoring branch lengths). This is precisely the Yule-Harding distribution.

Evolving 'discrete' Yule-Harding trees



Quiz:

Grow a Yule-Harding tree till it has 101 leaves.
Which is more likely?



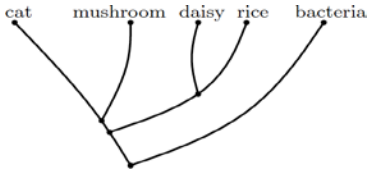
$\frac{1}{2}\times\frac{2}{3}\times\frac{3}{4}\times\ldots\times\frac{99}{100}=\frac{1}{100}$

The exact probability of a tree under YH and U?

$$\mathbb{P}_U(T) = \frac{1}{rb(n)}$$

$$\mathbb{P}_{YH}(T) = \frac{2^{n-1}}{n!\prod_{v\in I(T)}\lambda_v} \qquad \textit{Why?}$$

Example

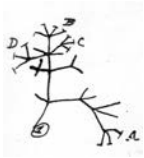


$$\mathbb{P}_U(T) = \frac{1}{rb(5)} = \frac{1}{3\times 5\times 7} = \frac{1}{105}$$

$$\mathbb{P}_{YH}(T) = \frac{2^4}{5!\times 4\times 3\times 1^2} = \frac{1}{90}$$

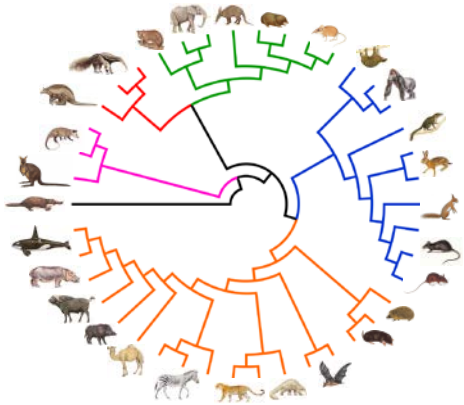
Why maths? (again)…

During the three years which I spent at Cambridge my time was wasted, as far as the academical studies were concerned, as completely as at Edinburgh and at school. I attempted mathematics, and even went during the summer of 1828 with a private tutor (a very dull man) to Barmouth, but I got on very slowly. The work was repugnant to me, chiefly from my not being able to see any meaning in the early steps in algebra. **This impatience was very foolish, and in after years I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics, for men thus endowed seem to have an extra sense.** But I do not believe that I should ever have succeeded beyond a very low grade.



THE END

Lecture 2: Properties of trees



ALLAN WILSON CENTRE

Mike Steel

from F. Delsuc and N. Lartillot

Biomathematics Research Centre

Winthrop lectures, 2014

ALLAN WILSON CENTRE

Outline

- *Part 1:* **Rooted phylogenetic trees, clusters, hierarchies**
- *Part 2:* **Unrooted phyl. trees, splits**
- *Part 3:* **Applications: RF metric, Consensus, Quartet encodings**
  - **break**
- *Part 4:* **Adams consensus**

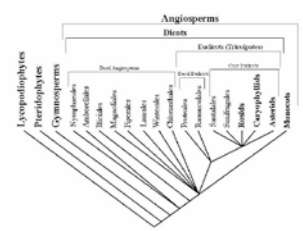
Rooted phylogenetic trees

Definition: A *rooted phylogenetic X-tree* is a rooted tree, with

- X = the set of leaves,
- Every non-root vertex has in-degree 1,
- Every non-leaf vertex has out-degree >1.

$R(X)$ = set of rooted phylogenetic X-trees.

"Polytomy"



$$C(v) = \{x \in X : x \text{ is separated from the root by deleting } v\}$$

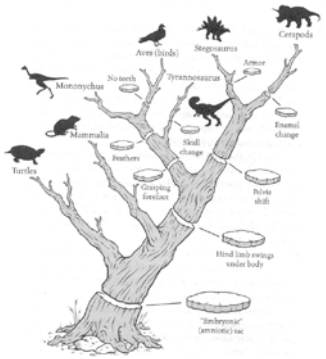
$$C(T) = \{c(v) : c \in V_T\}$$

"Clusters (or clades) of T"
(aka 'momophyletic group')

Hierarchies

A *hierarchy* on H on X is a collection of non-empty subsets of X satisfying:

- $A, B \in H \Rightarrow A \cap B \in \{A, B, \emptyset\}$
- $X \in H$, and $\{x\} \in H, \forall x \in X$



The clusters of any rooted phylogenetic X-tree form a hierarchy on X

Moreover, any hierarchy on X equals C(T) for a unique rooted phylogenetic X-tree T.

Partial order: $T \leq T' \iff C(T) \subseteq C(T')$

What does this order mean?

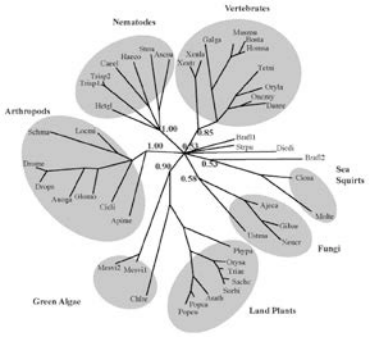
Unrooted phylogenetic X-trees

Definition: A *phylogenetic X-tree* is a tree, with

- X = the set of leaves;
- Every non-leaf vertex has degree at least 3.

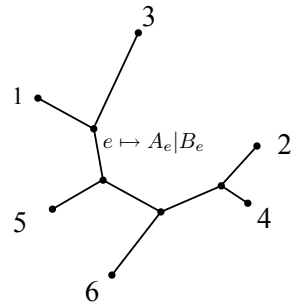
$U(X)$ = set of phylogenetic X-trees.

"Polytomy"
"Isomorphism"
$R(X) \leftrightarrow U(X \uplus \{x\})$



What corresponds to clusters/clades?

Encoding unrooted trees via splits



$$\Sigma(T) = \{A_e|B_e : e \in E\}$$
$$= \{13|2456, \cdots\}$$

$\Sigma(T)$ determines T

$$\Sigma(T) = \Sigma(T') \Leftrightarrow T \cong T'$$

Partial order:

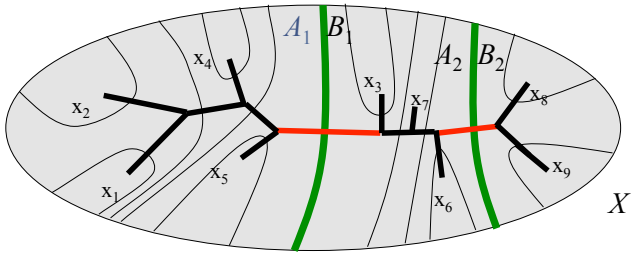
$$T \leq T' \iff \Sigma(T) \subseteq \Sigma(T')$$

When does a set of splits come from a tree?

- Two splits $A_1|B_1$ and $A_2|B_2$ of X are *compatible*, if one of the following intersections is empty:

$$A_1 \cap A_2, A_1 \cap B_2, B_1 \cap A_2, B_1 \cap B_2$$

Two compatible splits:



The set of splits of a phylogenetic tree is pairwise compatible

Conversely….

If Σ is a set of pairwise compatible splits, and

★ $\{x\}|(X - \{x\}) \in \Sigma$ for all $x \in X$

then $\Sigma = \Sigma(T)$ for a (unique!) phylogenetic X-tree

Without ★ the same applies with the phylogenetic replaced by "X-tree"

Simple algorithm for reconstructing T from $\Sigma(T)$ ('tree popping')

The link(s) between pc X-splits vs hierarchies

Obvious one: Select $x_0 \in X$

$A|B \mapsto$ the set (A or B) that does not contain x_0

Σ is pc iff the induced set system is a hierarchy on $X - \{x_0\}$

Example: Σ = {1|234, 2|134, 3|124, 4|123, 12|34}

More subtle…

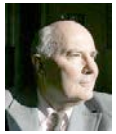
$A|B \mapsto$ smaller of A or B

Σ is pc iff the induced set system is a hierarchy on X

Example: Σ = {1|234, 2|134, 3|124, 4|123, 12|34}

Applications of split encoding I: Tree metrics

Robinson-Foulds metric [1981]



Les Foulds

$$d(T,T') = |\Sigma(T) \nabla \Sigma(T')|$$

"symmetric difference"

Interpretation?

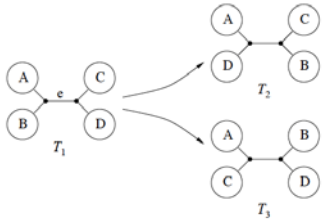
$d(T,T')$ is the minimum number of interior edges we need to collapse in T and in T' (combined) to arive at the same tree T^*

Properties of RF metric on UB(n)

$$d(T,T') = 2|\Sigma(T) - \Sigma(T')| = 2|\Sigma(T') - \Sigma(T)|$$

d is even!

$$\min_{T \neq T'} \{d(T,T')\} = 2$$



$$d_{NNI}(T,T') = \min\{k : T_0 = T, \cdots, T_k = T', d(T_i, T_{i+1}) = 2\}$$

$$\max\{d(T,T')\} = 2n - 6$$

Most big trees share only few (and tiny!) non-trivial splits

$s(T,T') = \#$ non-trivial splits that T and T' share

Given $T \in UB(n)$, and
T' (random) from UB(n)

$$\mathbb{P}(s(T,T') = k) \sim e^{-\lambda_T} \frac{\lambda_T^k}{k!}$$

$$\lambda_T = \frac{\#\text{ cherries in } T}{2n}$$

How many cherries are there in a binary tree?

Tree model	**mean # cherries**
Yule-Harding	$\frac{n}{3}$
Uniform model	$\sim \frac{n}{4}$

$$\lambda_T = \frac{\#\text{ cherries in } T}{2n}$$

Corollary: Two trees chosen uniformly at random share a Poisson number of non-trivial splits with mean (1/8). So 88% share no non-trivial splits.

An interesting combinatorial challenge:

Show that the number of trees in UB(n) that have exactly c cherries is:

$$\frac{n!(n-4)!}{(n-2c)!(c-2)!c!2^{2c-2}}$$

$$\binom{n}{2c} \cdot \frac{(2c)!}{c!2^c} \times \frac{(2c-4)!}{(c-2)!2^{c-2}} \; (= b(c)) \; \times \binom{k+(2c-3)-1}{k} \cdot k!$$

$$k = n - 2c$$

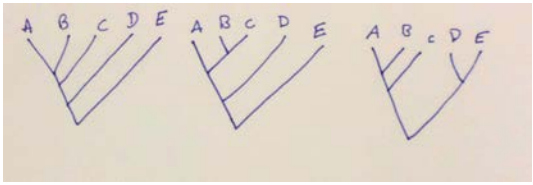
Applications of split encoding II: Consensus

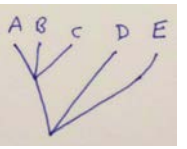
A *consensus method* is a function that assigns to each 'profile' (sequence) of phylogenetic X-trees

$$\mathcal{P} = (T_1, T_2, \ldots, T_k)$$

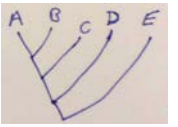
a single phylogenetic X-tree. *Why interesting?*

Example





Strict consensus



Majority-rule consensus

Strict and majority rule consensus

$$\mathcal{P} = (T_1, T_2, \ldots, T_k)$$

Strict consensus:
Let $\Sigma_{100\%}$ be the splits that appear in **all** the trees.

Majority rule consensus:
Let $\Sigma_{>50\%}$ be the splits that appear in **more than half** the trees.

Proposition: $\Sigma_{>50\%}$ is pairwise compatible (and so determines a tree.

Proof: Applies the 'pigeonhole principle'

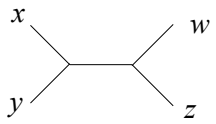
A nice exercise:

Theorem [McMorris]: When n is odd, the Majority Rule tree is the unique phylogenetic X-tree T that minimizes the median RF-distance:

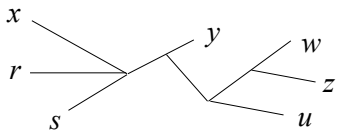
$$\sum_{i=1}^{k} d(T, T_i)$$

Quartet trees

• A **quartet tree** is a binary phylogenetic tree on four 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*}



(Displayed) quartet trees encode any phylogenetic tree

Given a phylogenetic X-tree T, let $Q(T)$ be the set of quartets that T displays.

Then $\forall T, T' \in U(X)$

$$Q(T) = Q(T') \Leftrightarrow T \cong T'$$

Quiz: Why?

Harder question: How many questions do we need to ask of the form 'what is T|q' for a quartet q in order to reconstruct T?

When is Q=Q(T) (for some T)?

•[Colonius and Schultze 1981]

$Q = Q(T)$ for some $T \in U(X)$ iff the following hold

$$ab|cd \in Q \Rightarrow ac|bd, ad|bc \notin Q$$

$$ab|cd \in Q \Rightarrow ab|ce \in Q \text{ or } ae|cd \in Q.$$

Another tree metric – the 'quartet metric'

$$d_Q(T, T') = |Q(T) \nabla Q(T')|$$

Less 'sensitive' than RF

Mean = easy to compute $\frac{1}{3}\binom{n}{4}$

Complexity?

The diameter is a difficult unsolved problem!

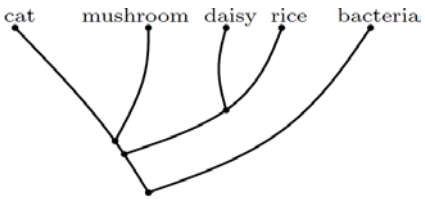
Conjecture:

$$\max\{d_Q(T, T') : T, T' \in UB(n)\} = \left(\frac{1}{3} + o(1)\right)\binom{n}{4}$$

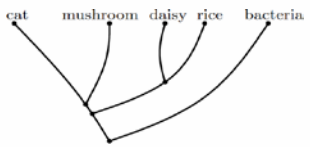
(S. Grunewald seems to have a proof)

Analogous theory for rooted trees:

In place of quartet trees $ab|cd$, one has rooted triples $ab|c$



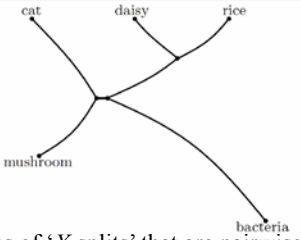
Summary: Encoding trees



Hierarchies on X (a subsets of X that do not 'overlap').

Collection of rooted 3-leaf phylogenetic trees that are compatible

Distance function on X that satisfies a 3-point condition (ultrametrics)



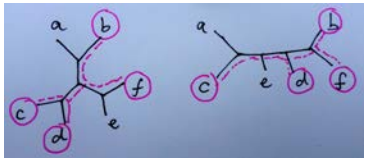
Collections of 'X-splits' that are pairwise compatible.

Collection of unrooted 4-leaf trees that are compatible

Distance function on X that satisfy a 4-point condition

Maximum agreement subtree I:

Definition



$T_1, T_2, \ldots T_k$ phylogenetic X-trees

$\mathrm{MAST}(T_1, T_2, \cdots, T_k) :=$

$\max\{|Y| : Y \subseteq X, T_1|Y = T_2|Y = \cdots T_k|Y\}$

Algorithms? Two trees

Three trees

k-trees, with $\geq$1 tree having no vertex of high degree

Maximum agreement subtree II: mathematical aspects

Randomized question

Given two trees generated 'at random' (uniform or Yule) what can we say about the size of their max. agreement subtree? [see Katherine St John!]

Extremal question

- Any two trees on UB(6) have a max. agrement subtree of size at least 4.
- There are two trees in UB(n) that have max. agreement subtree of size $\log_2(n)$

Conjecture: The max. agreement subtree of any two trees in UB(n) has size at least $c\log(n)$ for some constant c.

Specialist topic: Axiomatic aspects of consensus methods

Three properties we'd expect any consensus method to have:

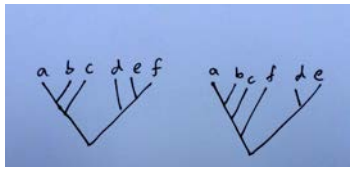
- Unanimity: $\mathcal{P} = (T, T, \ldots, T) \Rightarrow \psi(\mathcal{P}) = T$
- Tree order invariance:

$$\psi(T_1, T_2, \ldots, T_k) = \psi(T_{\sigma(1)}, T_{\sigma(2)}, \ldots, T_{\sigma(k)})$$

- Taxon permutation equivariance:

$$\psi(T_1^\sigma, T_2^\sigma, \cdots, T_k^\sigma) = \psi(T_1, T_2, \cdots, T_k)^\sigma$$

Another consensus method: (Adams consensus) [E.N. Adams III, 1972, 1986]



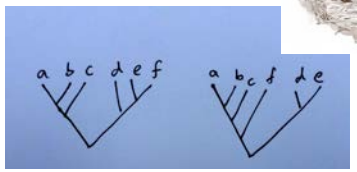
Given partitions $\pi_1, \pi_2, \ldots, \pi_k$ consider the *product partition*

$$\pi_1 \otimes \pi_2 \cdots \otimes \pi_k$$

Adams consensus tree for above two trees

Adams concensus and nestings

"A **nests in** B in T" means that the MRCA of A is a strict descendant of the MRCA of B in T



[Ad1] If A nests in B for each tree in the profile, then A nests in B in the Adams tree

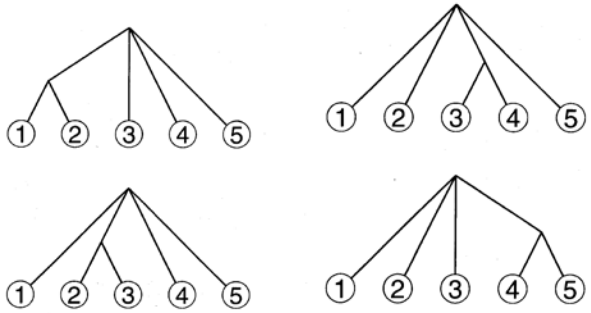
[Ad2] If A, B are clusters of the Adams consensus tree, and A nests in B then A nests in B in every tree in the profile.

[Moreover, Adams consensus is the **only** tree satisfying these properties]

Note that [Ad1] implies: If all input trees display $xy|z$ so too does Adams tree

Can we do better than that?

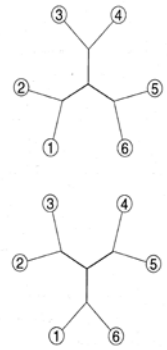
If at least one input tree displays $xy|z$, and no input tree displays $xz|y$ or $yz|x$ then the consensus tree should display $xy|z$



Output tree should display 12|5, 23|5, 34|1 and 45|1 – but there is no tree that does this!

What about Adams for unrooted trees?

If each tree in $\mathcal{P}$ displays $ab|cd$ then $\psi(\mathcal{P})$ does too

$\mathcal{P} =$ 

These trees each display

$\{12|45, 34|16, 56|23\}$

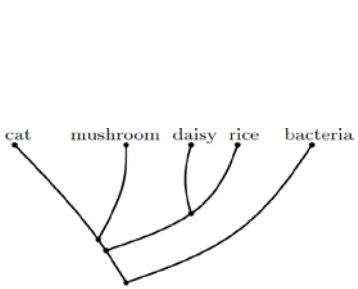
They are the **only** trees that display these quartets.

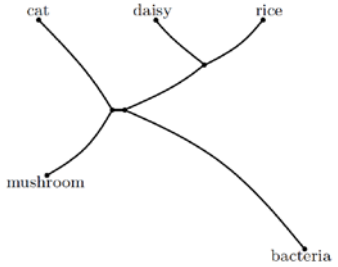
The cyclic permutation

(123456) interchanges the two trees in $\mathcal{P}$

- THE END

Revision



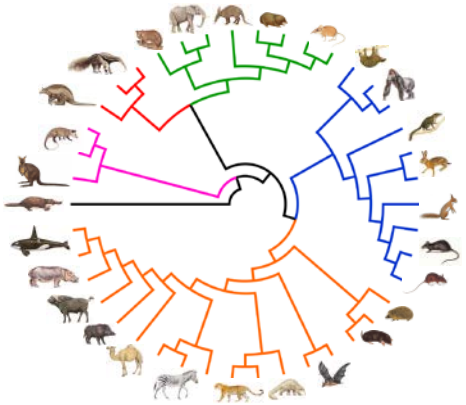


Quiz:

If A is a cluster/clade of a rooted tree T, and we suppress the root of T, is $A|X\text{-}A$ a split of T?

If $A|B$ is a split of an unrooted tree T, and we root T, is A a cluster of T?

Lecture 3: Character data



Mike Steel

ALLAN WILSON CENTRE

from F. Delsuc and N. Lartillot

Biomathematics Research Centre

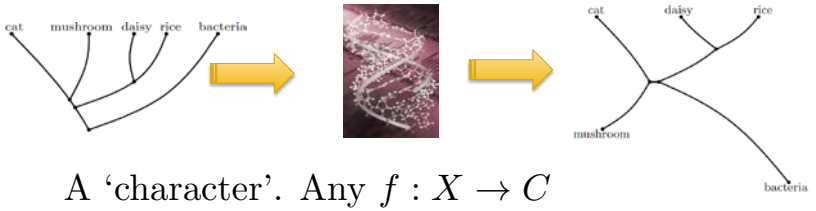
Winthrop lectures, 2014

ALLAN WILSON CENTRE

Outline of talk

- *Part 1:* **Discrete characters and homoplasy**
- *Part 2:* **Perfect phylogeny**
- *Part 3:* **Parsimony**
  - **20x pushups**
- *Part 4:* **Specialist topic: The 'joys of being mean'**

Tree reconstruction



A 'character'. Any $f : X \to C$

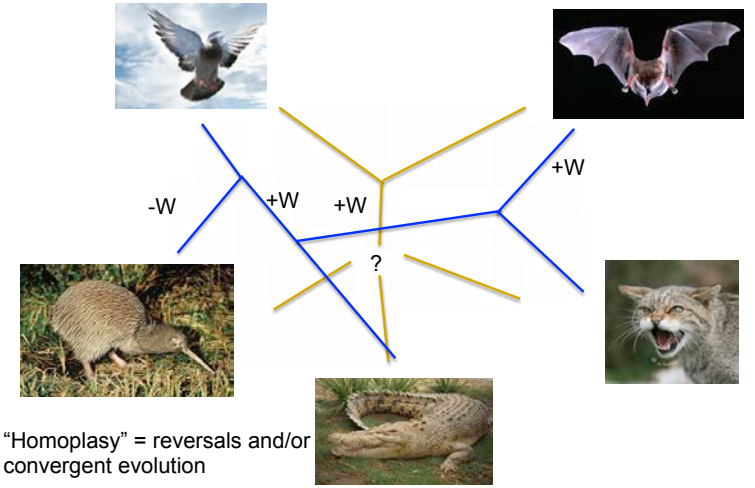
Discrete data: $(f_1, f_2, \ldots, f_k)$

Species	Attribute	1	2	3	4
Kangaroo		T	R	U	E
Chimpanzee		B	R	E	T
Human		B	R	O	E
Gorilla		C	O	E	E
Hippopotamus		C	A	P	O
Whale		C	A	U	P
Lion		D	R	A	O
Tiger		D	R	U	G

Types of "characters"

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

Signal in data (and why it be misleading...)



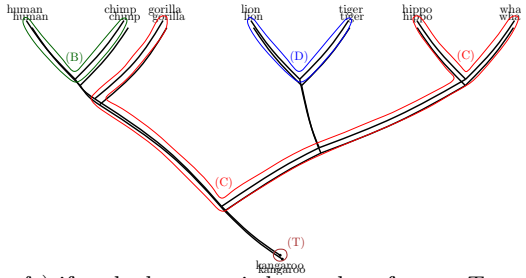
"Homoplasy" = reversals and/or convergent evolution

$h(f, T)$ = minimal number of such events required to fit f to T

Homoplasy-free: $h(f, T) = 0$

$\iff$ the minimal subtrees of T connecting the leaf sets $f^{-1}(c)$ and $f^{-1}(c')$ are vertex-disjoint, for each $c \neq c'$

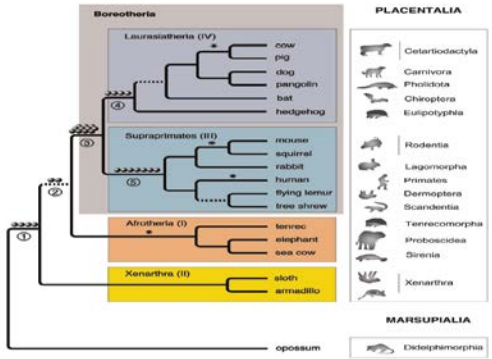
Species	Attribute	1	2	3	4
Kangaroo		T	R	U	E
Chimpanzee		B	R	E	T
Human		B	R	O	E
Gorilla		C	O	E	E
Hippopotamus		C	A	P	O
Whale		C	A	U	P
Lion		D	R	A	O
Tiger		D	R	U	G



T is a *perfect phylogeny* for $(f_1, f_2, \ldots, f_k)$ if each character is homoplasy-free on T

Example of low-homoplasy data I (SINEs)

[Kreigs *et al.* PLoS biology, 2006. Tree of placental mammals]



Example of low homoplasy-data II

- Gene order rearrangements (n species, L genes, random inversion model)

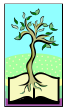
$$g_1 g_4 \boxed{g_3 g_4 g_5} g_6 g_7 \cdots \Rightarrow g_1 g_2 g_5 g_4 g_3 g_6 g_7 \cdots$$

$$\mathbb{P}(h(f,T)=0) \geq 1 - \frac{2(2n-3)(n-1)}{L(L-1)}$$

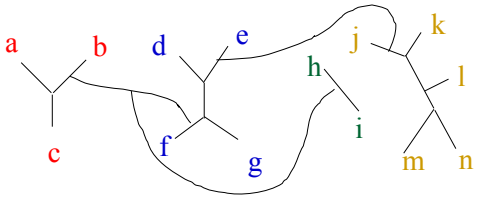
Semple+S, Adv. Appl. Math. 2002

N=10, L=100, Prob>0.97

How many trees have $h(f, T) = 0$?



a b c
d e f g
h i
j k l m n



How many binary phylogenetic trees can we construct in this way? (*c.f.* lecture 1)

$$\frac{b(n)}{b(n-k+2)} \prod_{i=1}^{k} |E_i| = \frac{B(14)}{B(12)} \times 3 \times 5 \times 1 \times 7 = 50,715$$

$$\text{So } \#T : h(T,f) = 0 \text{ is } \frac{b(n)}{b(n-k+2)} \prod_{i=1}^{k} rb(n_i)$$

Application (example)

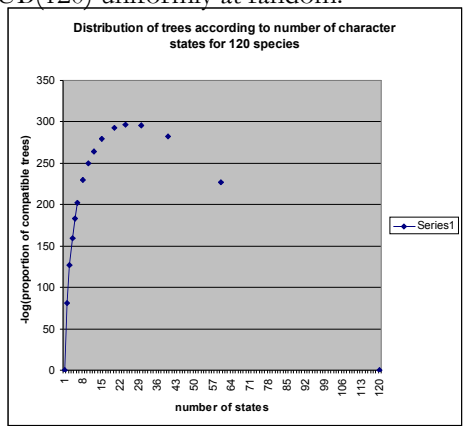
Consider an r-state character f on 120 species, with $120/r$ species in each state. Select a tree T from UB(120) uniformly at random.

Let

$$I(f) := -\log(\mathbb{P}(h(f,T)=0))$$

A good case for mathematics over simulations…

Nice problem for a student: How does r_{max} grow with n?



When does a perfect phylogeny exist?

- **Definition:** Characters $f_1, f_2, \ldots, f_k$ are *compatible* if there exists a perfect phylogeny for them.
- **Special case:** Binary characters are compatible if and only if the associated set of X-splits Σ is pairwise compatible.

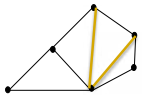
$$\Sigma = \{f_i^{-1}(0) | f_i^{-1}(1); i = 1, \ldots, k\}$$

  - **Corollary:** A set of binary characters are compatible iff each pair is; and there is a unique minimal perfect phylogeny.
  - Both parts of this corollary fail for 3-state characters.

A link to graph theory…

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

Example



Definition:

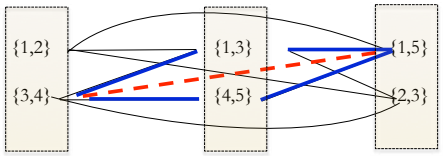
- Given $G = (V,E)$ and a partition $V = V_1 \cup V_2 \cup \cdots \cup V_k$ a **restricted chordal completion** of G is any chordal graph satisfying

$$H = (V, E') : E \subseteq E'$$

$$x, y \in V_i \Rightarrow \{x, y\} \notin E' - E$$

Characterising compatibility

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



$\mathrm{Int}(\{f_1, f_2, f_3\})$

Partition Intersection Graph (PIG)

Theorem

- $C = (f_1, .., f_k)$ is compatible if and only if int(C) has a restricted chordal completion. [why?]
- If $|C| = 2$, then C is compatible if and only if int(C) has no cycles [why?]

How hard is the perfect phylogeny problem?

Given characters $f_1, f_2, ..., f_k$ what is the complexity of deciding whether or not they are compatible?

- Easy for binary characters
- Poly-time for r-state characters (r bounded)
- NP-hard in general (we'll see why in the lecture 5!)

Special 'easy' case:

Characters f, g are *strongly compatible* if $f^{-1}(s) \cup g^{-1}(s') = X$ for some s, s'

Theorem Suppose $\mathcal{C} = (f_1, \ldots, f_k)$ is pairwise strongly compatible. Then $\mathcal{C}$ is compatible, and has a unique minimal perfect phylogeny.

Dress, A.W.M., Moulton and S. 1997.

A curious result….

When $r = 2$ or $r = 3$, a set of r-state characters $\{f_1, f_2, ..., f_k\}$ is compatible if and only if every subset of r characters is compatible[1].

How does this generalize?

Theorem[2] For all $r \geq 2$ there is an incompatible set of $\lfloor \frac{r}{2} \rfloor \cdot \lceil \frac{r}{2} \rceil + 1$ r-state characters with every r-subset compatible

Interesting unsolved problem:

Is (quadratic) behaviour 'as bad as it gets'?

[1] r=3, recent result due to Dan Gusfield

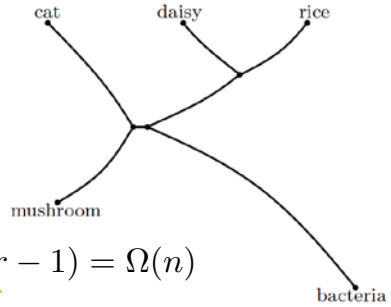
[2] B. Shutters, S. Vakati, D. Fernandez-Baca, Incompatible quartets, triplets, and characters, J Mol. Biol. 8 (2013) 11.

Question: how many characters are needed so that T is the *only* perfect phylogeny for this data?

T must be binary!

'Binary characters': $f : X \to C, |C| = 2$

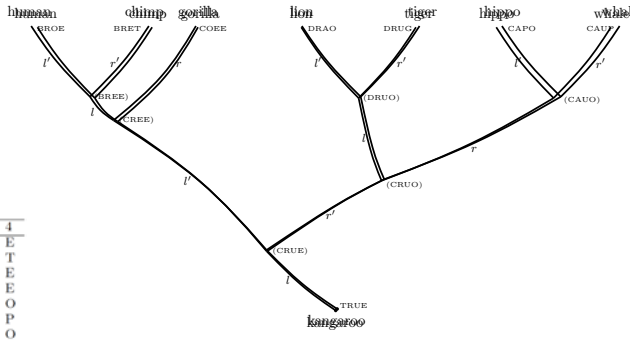
If T is the only perfect phylogeny for $(f_1, ..., f_k)$ then $k \geq n - 3$



r-state characters : $k \geq (n-3)/(r-1) = \Omega(n)$

The 'four is enough' theorem

Every binary phylogenetic tree – on any number of species – is a unique perfect phylogeny for at most *four* characters.



Species	Attribute	1	2	3	4
Kangaroo		T	R	U	E
Chimpanzee		B	R	E	T
Human		B	R	O	E
Gorilla		C	O	E	E
Hippopotamus		C	A	P	O
Whale		C	A	U	P
Lion		D	R	A	O
Tiger		D	R	U	G

Maximum parsimony (minimum evolution)

$\mathrm{ps}(f, T)$ *The "parsimony score" of character f on T*

= the minimum number of edges that need to have different states assigned to their ends in order to extend f to all vertices of T.

PIC

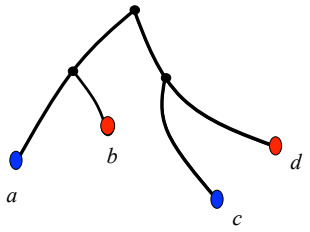
Easy or hard?

Easy – by dynamic programming.

Moreover, the "Fitch-Hartigan algorithm' is linear-time algorithm (in *n* and *r*) due to Walter-Fitch (formalized and mathematically verified by John Hartigan).

Homoplasy (again)

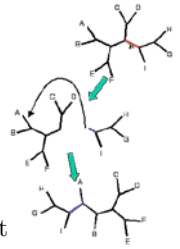
- $h(f, T) =$ smallest number of reversals/convergent events required to 'evolve' f on (any rooting of) T.



- Easily computed: $h(f, T) = \mathrm{ps}(f, T) - [|f(X)| - 1]$

Homoplasy as measure of tree distortion from a perfect fit

SPR (**S**ubtree **p**rune and **r**e-graft operation)



$h(f,T) - h(f,T') \in \{0, \pm 1\}$ if T and T' are one SPR apart

Theorem [Bruen and Bryant 2008]

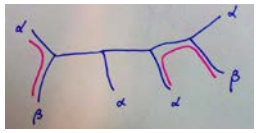
$h(f,T) =$ min #SPR operations to transform T into a tree on which f is homoplasy-free.

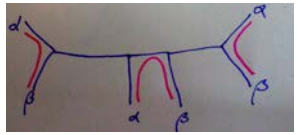
A curious (+useful) property for binary characters…

For **binary** characters:

$\text{ps}(f,T) =$ max number of edge-disjoint paths in T, each of which connects a leaf in one state to a leaf in a different state.

[*Proof:* by Menger's theorem]





Exercise

Show if $T \in UB(2n)$ then $\#f : \text{ps}(f,T) = n$ is 2^n

Generalization due to Peter Erdös and Laszlo Székely.

Maximum parsimony trees

$$\mathcal{C} = (f_1, \ldots, f_k) \quad \text{ps}(\mathcal{C}) = \min_T \sum_{i=1}^{k} \text{ps}(f_i, T)$$

How hard is to compute this?
(and find an optimal 'maximum parsimony tree' T?)

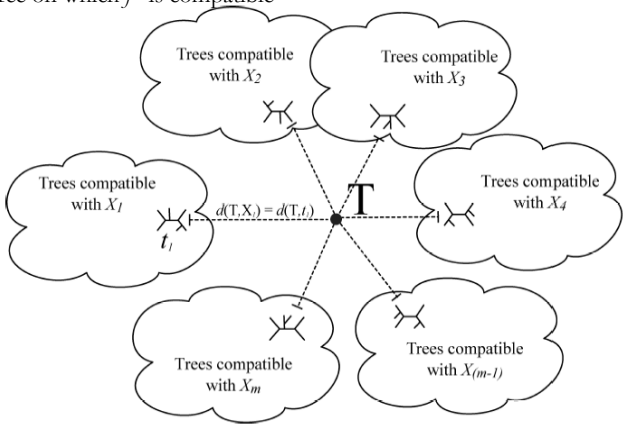
For binary characters it's already NP-hard

But there are some special cases that can be solved exactly;
and also bounds, for example:

Proposition [Foulds] The score of the MP tree for C is at most twice the score of a min cost spanning tree of X under Hamming distance d_C

A way of thinking of the MP tree (from Bryant and Bruen's result)

$h(f,T) =$ min #SPR operations to transform T into a tree on which f is compatible



From Bruen and Bryant 2008

Further mathematical properties of the MP tree I

Proposition[Bruen and Bryant 2008]
For any **two** characters $C = (f_1, f_2)$ the score of the MP tree is determined by int(C).

$$\mathrm{ps}(\mathcal{C}) = \#\text{ edges of } \mathrm{int}(\mathcal{C}) - \#\text{ components of } \mathrm{int}(\mathcal{C}) + 2$$

Proposition[Bryant 2003]

If C consists of just binary characters, and one of them, say f, is compatible with all others, then:

$$f^{-1}(0)|f^{-1}(1)$$

is a split of *every* MP tree for C.

An extension of this: H.J.- Bandelt's result that all MP trees lie in the 'median network' for C

Counting: How many trees in UB(n) have parsimony score k for a binary character f?

$a = |f^{-1}(0)|; b = |f^{-1}(1)|$

$$N(a,b;k) = \#T \in UB(n) : \mathrm{ps}(f,T) = k$$

Example: $N(2,2,k) = ?$

David Penny's remarkable conjecture:

$$N(a,b;k)/b(n) = 2^k \cdot \frac{k(2n-3k)}{(2a-k)(2b-k)} \cdot \frac{(2a-k)!}{(a-k)!} \cdot \frac{(2b-k)!}{(b-k)!} \cdot \frac{(n-k)!}{k!(2n-2k)!}$$

$$a + b = n; 0 \le k \le \min(a,b)$$

Proof uses several ideas above (Menger's theorem; counting trees that can be constructed by joining trees etc) plus some new ideas.
The original proof involved generating functions and a computer-assisted use of the multivariate Lagrange inversion formula

Maximum parsimony trees III

If T is the unique perfect phylogeny on n leaves for k characters then we need k to be at least n-3 (and this suffices for the right choice!)

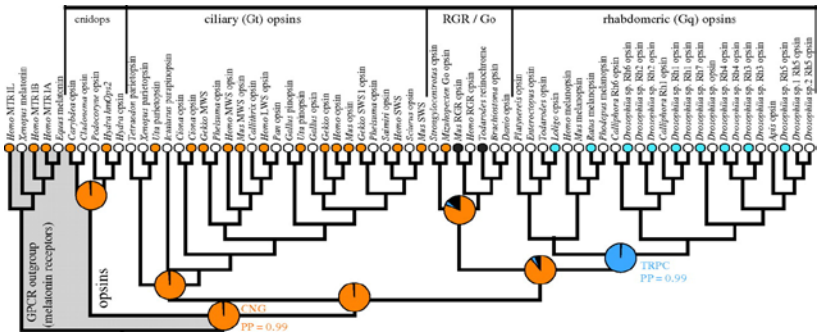
But what if we want T to be the unique MP tree?
We can do this with fewer than n-3? Sublinear?

A primitive counting argument gives a lower bound of log(n).
Remarkably, this can be achieved…

Theorem [Chai and Housworth, 2011]

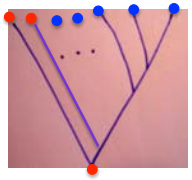
For every $T \in UB(n)$ there is a set of $\Theta(\log n)$ binary characters with T as the unique MP tree.

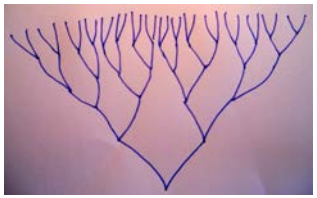
Parsimony: Ancestral state reconstruction



Plachetzki D C et al. Proc. R. Soc. B 2010;277:1963-1969

Maximum Parsimony vs Majority Rule





h

$n = 2^h$

$f_h = \min\{\#\text{ red tips } : MP(\text{root}) = \{\text{red}\}\}$

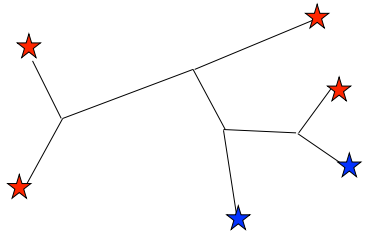
$$f_h = f_{h-1} + f_{h-2}$$

$$f_1 = 2, f_2 = 3$$

$$f_h \sim c\left(\frac{1+\sqrt{5}}{2}\right)^h$$

$$f_h/2^h \to 0$$

Specialist topic: Random thoughts about parsimony



Binary tree with T leaves.

S_T := average value of ps(f, T) over all 2^n binary f.

A curious recursion....

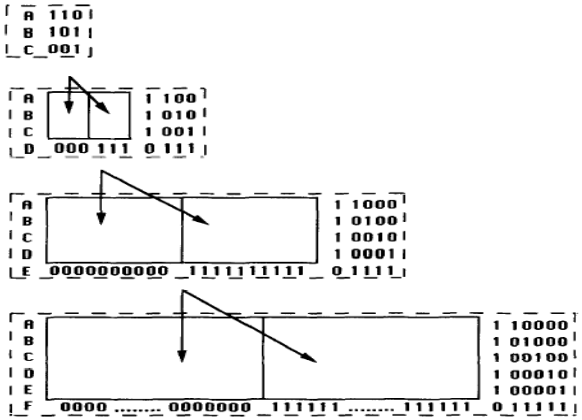
$$S_T = S_{T-1} + S_{T-2} + 2^{T-2} + 21 + 5T + 3\sum_{i=2}^{T-3} i + \sum_{X=0}^{T-8/2} \binom{T-8-X}{T-8-2X} \times (15X+53)$$

$$+ \sum_{X=0}^{T-7/2}\left(\sum_{Y=0}^{T-7-2X} \left(\binom{T+Y}{Y} \times (S_{T-4-2X-Y} + ((2^{T-4-2X-Y} - T + 2 + 2X + Y) \times (X+2))\right)\right)$$

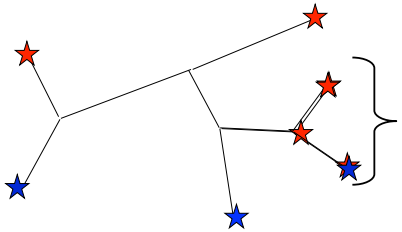
$$+ \sum_{Y=1}^{T-8-2X}\left(\binom{X+Y-1}{Y-1} \times (42 + 10X + ((T-2-2X-Y) \times (X+3))\right.$$

$$\left. + \sum_{Z=2}^{T-4-2X-Y} (X+3+(Z \times (X+4)))\right)\Bigg)$$

where does it comes from?



Random thoughts...



$$\mathrm{PS} = \begin{cases} \mathrm{PS}' & \text{if } x = y; \\ \mathrm{PS}'' + 1 & \text{if } x \neq y. \end{cases}$$

$$\mathbb{E}[\mathrm{PS}] = \frac{1}{2}\mathbb{E}[\mathrm{PS}'] + \frac{1}{2}(\mathbb{E}[\mathrm{PS}''] + 1)$$

$$S_T = \frac{1}{2}S_{T-1} + \frac{1}{2}(S_{T-2} + 1) \qquad S_T = \frac{3T - 2 - \left(\frac{-1}{2}\right)^T}{9}$$

Two solutions

$$S_T = \frac{3T - 2 - (-\frac{1}{2})^T}{9}$$

$$S_T = S_{T-1} + S_{T-2} + 2^{T-2} + 21 + 5T + 3\sum_{i=2}^{T-3} i + \sum_{X=0}^{T-8/2} \binom{T-8-X}{T-8-2X} \times (15X + 53)$$
$$+ \sum_{X=0}^{T-7/2} \left(\sum_{Y=0}^{T-7-2X} \left(\binom{T+Y}{Y} \times (S_{T-4-2X-Y} + ((2^{T-4-2X-Y} - T + 2 + 2X + Y) \times (X+2))) \right) \right)$$
$$+ \sum_{Y=1}^{T-8-2X} \left(\binom{X+Y-1}{Y-1} \times (42 + 10X + ((T-2-2X-Y) \times (X+3)) \right.$$
$$\left. + \sum_{Z=2}^{T-4-2X-Y} (X + 3 + (Z \times (X+4))) \right) \Big)$$

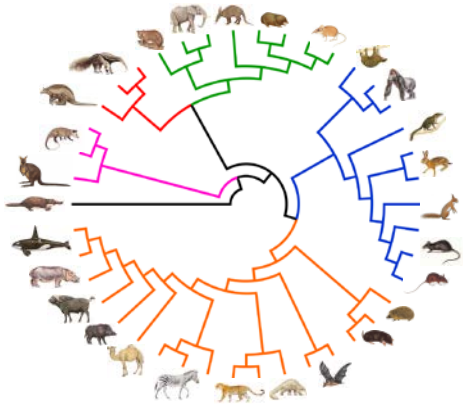
- Actually the recursion gives more….

$$\mathbb{P}(\mathrm{ps} = k) = \frac{(2n-3)}{k} \binom{n-k-1}{k-1} 2^k$$

for any T in UB(n) (independent of shape), k>0, and this is asymptotically normal as n grows.

THE END

Lecture 4: Distance-based tree reconstruction



from F. Delsuc and N. Lartillot

Mike Steel

ALLAN WILSON CENTRE

Biomathematics Research Centre

Winthrop lectures, 2014

ALLAN WILSON CENTRE

Outline

- *Part 1:* **Encoding trees by distances, 4PC, ultrametrics**
- *Part 2:* **Reconstruction methods**
- *Part 3:* **Phylogenetic diversity and BME**
  - **20x pushups**
- *Part 4:* **Specialist topic: Do we need all the distances?**

The unified view

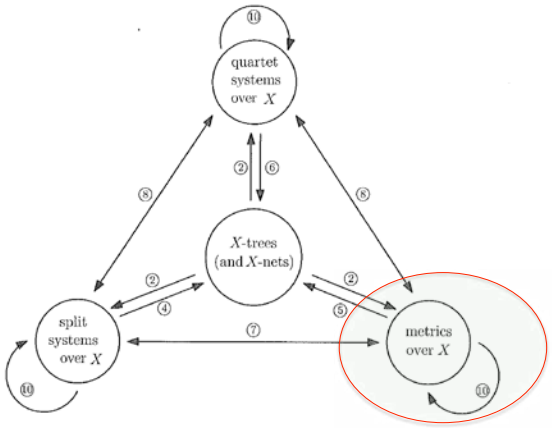
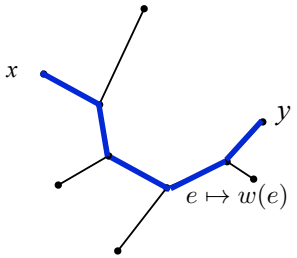


Figure 1 In this figure, we indicate the manifold relationships between various combinatorial objects relevant in phylogenetic analysis that will be studied in this book.

From: *Basic Phylogenetic Combinatorics*

Edge-weighted trees and tree metrics



$$w(e) > 0$$

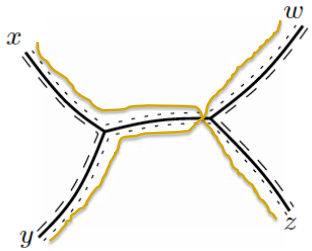
$$d_{(T,w)}(x,y) = \sum_{e \in P(T;x,y)} w(e)$$

If this holds for all pairs of leaves we that d is '**tree metric**' with a '**representation on *T***'

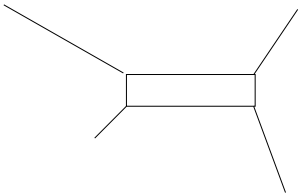
When can a distance (metric) on X be represented on a phylogenetic X-tree?

$n = 3$, always!

$n = 4$?



$$d(x,y) + d(w,z) < d(x,w) + d(y,z)$$
$$d(x,w) + d(y,z) = d(x,z) + d(y,w)$$



What does this tell us?

When about general *n*?

For any four points x, y, w, z let

$$S_1 = d(x,y) + d(w,z)$$
$$S_2 = d(x,w) + d(y,z)$$
$$S_3 = d(x,z) + d(y,w)$$

If d is a tree metric then, for i =1,2,3

$$S_i \leq \max\{S_j, S_k\}$$

This is called the **four point condition (4PC)**

A classic result (1960s/early 70s)

Theorem

d is a tree metric if and only if it satisfies the 4PC

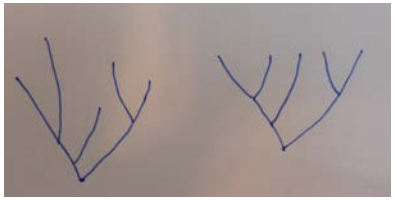
And the choice of T and $w > 0$ to represent d is unique

Proof?

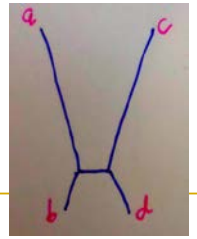
Ultrametrics trees (aka. 'equidistant tree', 'clock-like tree')

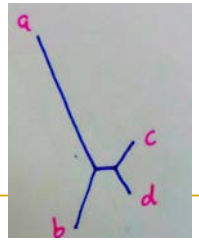
Defintion: A rooted tree with edge weighting (T,w) is an 'ultrametric tree' if the distance from the root to each leaf is the same (rooted).

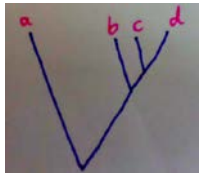
For an unrooted tree – it is ultrametric if it can be rooted (at some point)so this holds).



Quiz: are these unrooted trees ultrametric trees





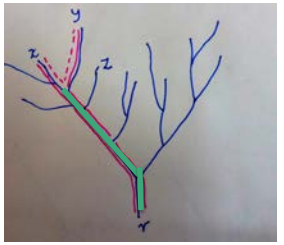


Ultrametrics

Definition: D is an **ultrametric** on X if it satisfies the 3-point condition:

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

The connection: D is an ultrametric on X if and only if there is a tree T in $R(X)$ on which D has ultrametric branch lengths (and then T, w unique)



Transforming an arbitrary tree metric into an ultrametric (Farris/Gromov transform):

$$D(x,y) = \begin{cases} d(x,y) - d(x,r) - d(y,r), & x \neq y; \\ 0, & x = y. \end{cases}$$

Distances vs characters – the 'darndest thing'!

Let $d_C(i,j)$ = # characters in C on which i and j differ (sequence dissimilarity).

QUIZ: If T is a perfect phylogeny for C does d_C a tree metric (on T)?

C binary characters – yes.

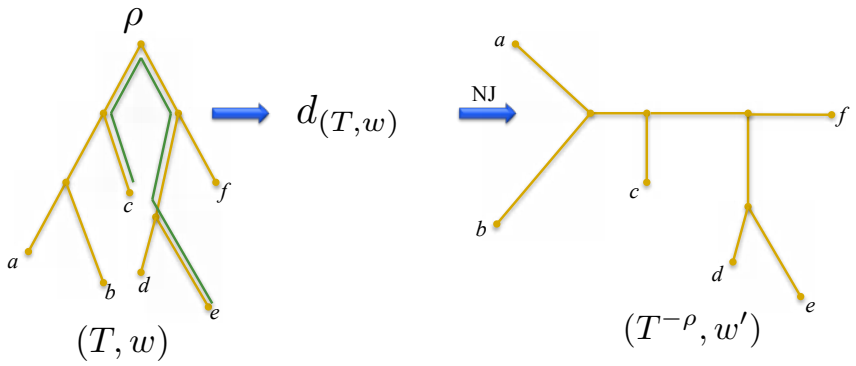
C non-binary characters – not necessarily.

Theorem [Fischer and Bandelt, H.-J. 2008; Huson and S, 2004]:

For **any** two trees T_1, T_2 there is a set of 3-state characters C such that:

- T_1 is the unique perfect phylogeny for, yet
- d_C is a tree metric (ultrametric!) represented only by T_2.

Distance-based tree reconstruction methods



Desirable property: perturbing d slightly leads to same tree

A simple approach

Select $p \in X$

Let x, y maximize $d(x,p) + d(y,p) - d(x,y)$

$$q(x,y)$$

Then x, y form a cherry of T

Why?

Neighbor-Joining

>36,000 citations: The neighbor-joining method: new method for reconstructing phylogenetic trees. N Saitou, M Nei *Molecular biology and evolution* 4 (4), 406-425

Neighor-Joining

$$Q(x,y) = d(x,y) - \frac{1}{n-2}\sum_p d(x,p) - \frac{1}{n-2}\sum_p d(y,p)$$

Select (x,y) to minimize Q

(1) If $d = d_{(T,w)}$ then (x,y) selected by Q is a cherry of T

(2) Q is a linear function of d

(3) If Q selects (x,y) and σ is a permutation of taxa, then Q applied to d^{σ} will select $(\sigma(x), \sigma(y))$.

Theorem [Bryant] If a selection criterion Q^* satisfies (1), (2) and (3) then Q^* makes the same selection as Q

An inconvenient truth

- Biological distances were not created by a mathematician!
  - Our best hope: If δ is 'close' to $d = d_{(T,w)} | \mathcal{L}$ then $NJ(\delta) = T$

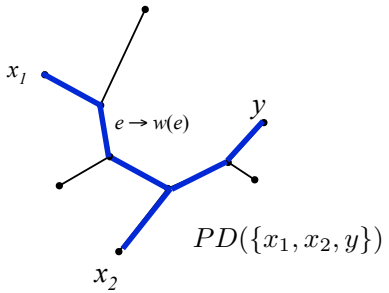
We say that a distance-based method M has *safety radius* r if following holds: $\forall x, y \in X$

$$|\delta(x,y) - d_{(T,w)}(x,y)| < r \cdot w^* \Rightarrow M(\delta) = T$$

$$w^* = \min\{w(e) : e \in E_{\text{int}}(T)\} \quad T \in UB(n)$$

- No method allows $r > ½$. (why?)
- But NJ has safety radius ½ [K. Atteson]
  (also there is an 'edge safety radius' result also)
- NJ exhibits discontinuity (when far from tree metric)

Phylogenetic diversity (PD)



$PD(\{x_1, x_2, y\})$

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

$$PD(Y) = \sum_{e \in T(Y)} w(e)$$

$$L = PD(X)$$

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

$$L = PD(X) = \sum_{\{x,y\} \subseteq X} \left(\frac{1}{2}\right)^{\Delta_T(x,y)} d(x,y)$$

$\Delta_T(x,y) = \#$ int. vertices between x and y in T

$$= \left(\frac{1}{2}\right)^3 d(x_1,x_2) + \cdots$$

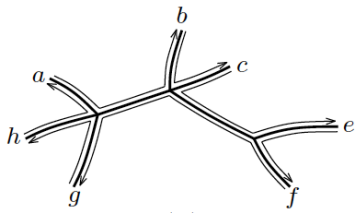
Balanced minimum evolution (BME)

Given d (not necessarily a tree metric) select the phylogenetic tree(s) T to minimizes L according to the Pauplin formula

$$L = PD(X) = \sum_{\{x,y\} \subseteq X} \left(\frac{1}{2}\right)^{\Delta_T(x,y)} d(x,y)$$

BME is 'consistent'
(Desper and Gascuel, 2004)

A way to view Pauplin's formula [original proof by induction]

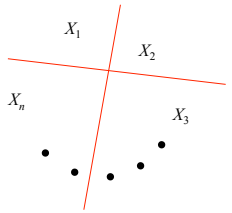


$$L = \frac{1}{2}[d(a,b) + d(b,c) + d(c,e) + d(e,f) + d(f,g) + d(g,h) + d(h,a)]$$

Each 'cyclic ordering' of the leaves of T gives a different way of writing L
Each is an arbitrary choice, so let's average over all of them –
what do we get?

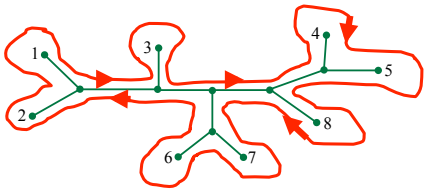
Cyclic Permutation on X

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



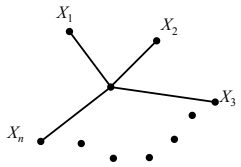
$\Sigma^o(\pi) :=$ the splits (bipartitions of X) induced by planar 'cuts'.

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

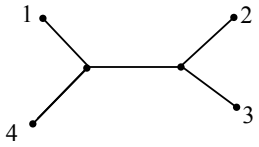


$$L = \frac{1}{2}\sum_i d(x_i, x_{i+1})$$

Question: How many cyclic orderings does T have?

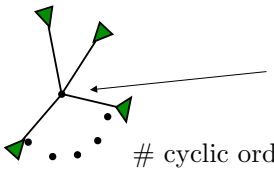


$(n-1)!$

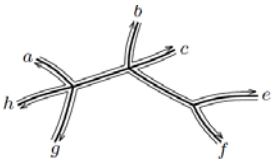


(1234), (1324), (1423), (1432)

General case



($d(v)$-1)!



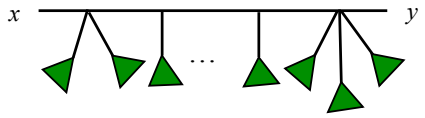
cyclic orderings for T is $\prod_{v \in I(T)} (\deg(v) - 1)!$

For a binary tree this is 2^{n-2}

More counting

How many cyclic orderings for T for with …xy…?

At least one!



Lemma: The proportion of co's of T with …xy… equals

$$\prod_{v \in I(x,y)} (\deg(v) - 1)^{-1}$$

Corollary:

Given a cyclic ordering π the # binary phylogenetic trees for which π is a cyclic ordering is the Catalan number

$$\frac{1}{(n-1)}\binom{2n-4}{n-2}$$

Why?

$$\#(T,o) := b(n)2^{n-2} = (n-1)!Q$$

Back to Pauplin…

$$L = \frac{1}{|o(T)|}\sum_{(x_1,\ldots,x_n)\in o(T)}\left(\frac{1}{2}\sum_i d(x_i,x_{i+1})\right)$$

$$= \frac{1}{2}\sum_{(x,y)}\left(\frac{n_T(x,y)}{|o(T)|}\right)d(x,y)$$

$$= \sum_{\{x,y\}\subseteq X}\lambda_T(x,y)d(x,y)$$

where

$$\lambda_T(x,y) = \prod_{v\in I(x,y)}(\deg(v)-1)^{-1}.$$

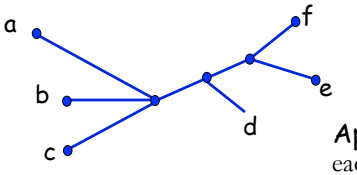
Summary:

For any phylogenetic tree T the average of all the representations of L is:

$$L = PD(X) = \sum_{\{x,y\}\subseteq X}\lambda_T(x,y)d(x,y)$$

$$\lambda_T(x,y) = \prod_{v\in I(x,y)}(\deg(v)-1)^{-1}.$$

Example



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

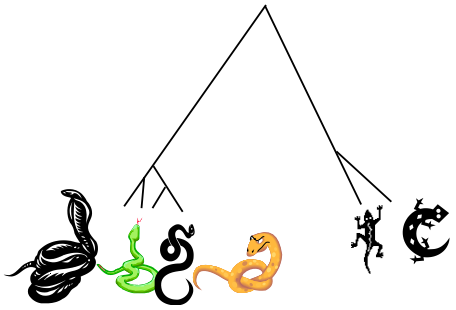
Application: NJ selects the pair of leaves (at each step) to maximizes the reduction in BME score [Desper and Gascuel, 2004; Gascuel and S, 2006]

Phylogenetic diversity (again)

$$PD(Y) = \sum_{e\in T(Y)} w(e)$$

Relevant for:

- Conservation biology
- Ensuring evolutionary 'coverage' in study designs
- Tree reconstruction



A nice combinatorial property of PD

- **Problem:** Find a subset Y_{max} of X given size k to maximise *PD*.
- **Theorem:** Y_{max} can always be found by using the 'greedy algorithm'.

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

- *Why?*

If $1 < |Y_1| < |Y_2|$ there exists $y \in Y_2 - Y_1$:

$PD(Y_1 \cup \{y\}) + PD(Y_2 - \{y\}) \geq PD(Y_1) + PD(Y_2)$.

An alternative measure: "max-min"

Select set S of k leaves to maximise $\min\{d(x,y) : x, y \in S, x \neq y\}$

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) maximizing the minimum distance. We constructed this tree using small-subunit ribosomal RNA sequences

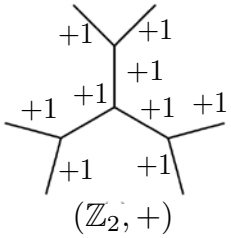
Min-max selection on a tree is easy, but not via greedy Bordewich and Semple

What's the connection?

Theorem [Bordewich and Semple (Syst. Biol. 2012)]

For clock-like branch lengths the optimal max-min selection of k species is **identical** to the optimal PD selection.

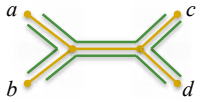
Quiz: What happens to our **distance** and **PD** results if the edges are weighted by an (Abelian) group?



Specialist topic: Do we need all of the distances?

Given $\mathcal{L} \subseteq \binom{X}{2}$ and $d = d_{(T,w)}|\mathcal{L}$

Does d determine T? (and/or w?)

Example: 

$\mathcal{L} = \{ab, cd, ac, bd\}$

$d|\mathcal{L}$ determines T but not w

$\mathcal{L}' = \mathcal{L} \cup \{ad\}$ determines T and w

$\mathcal{L}'' = \{ab, ac, ad, bc, bd\}$ determines *neither*

Draw graph!

How few distances do we need?

Classic result: [Yusmanov, 1984]

For any binary tree T with n leaves, there is a set $\mathcal{L} \subseteq \binom{X}{2}$ of size $2n$-3 so that the $d_{(T,w)}|\mathcal{L}$ determines both T and w.

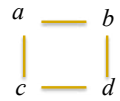
[Dress, Huber, S, (2014)] If we just want to define T, we can reduce the size of $\mathcal{L}$ by 1 (but no more!)

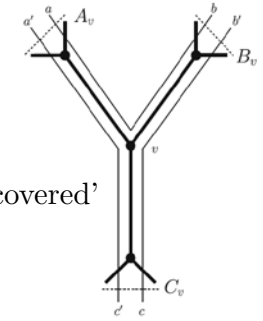
Quiz If $d_{(T,w)}|\mathcal{L}$ determines w on given T, does it also determine T?

Necessary conditions for $d|L$ to determine T and its edge weights

$(X, \mathcal{L})$ must be connected and contain an odd cycle

$\mathcal{L} = \{ab, cd, ac, bd\}$





If T is binary, each interior vertex must be '3-covered'

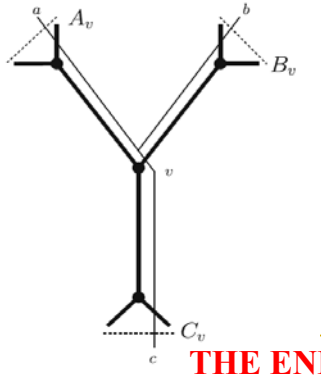
$$\{ab, b'c, c'a'\} \subseteq \mathcal{L}$$

The 'Triplet Cover' conjecture

$\mathcal{L}$ is a *triplet cover* for $T \iff$ for each $v \in I(T)$:

$$\exists a \in A_v, b \in B_v, c \in C_v :$$

$$\{ab, ac, bc\} \subseteq \mathcal{L}$$



Conjecture: If $\mathcal{L}$ contains a triplet cover for T then $d_{(T,w)}|\mathcal{L}$ determines T (and so also w).

THE END