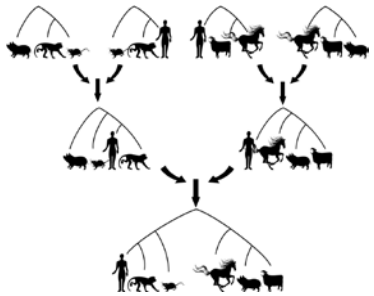
Lecture 5: Subtree-based tree reconstruction



"The analysis of large data sets could proceed by division into overlapping subsets which are classified separately and then recombined to provide a single classification"

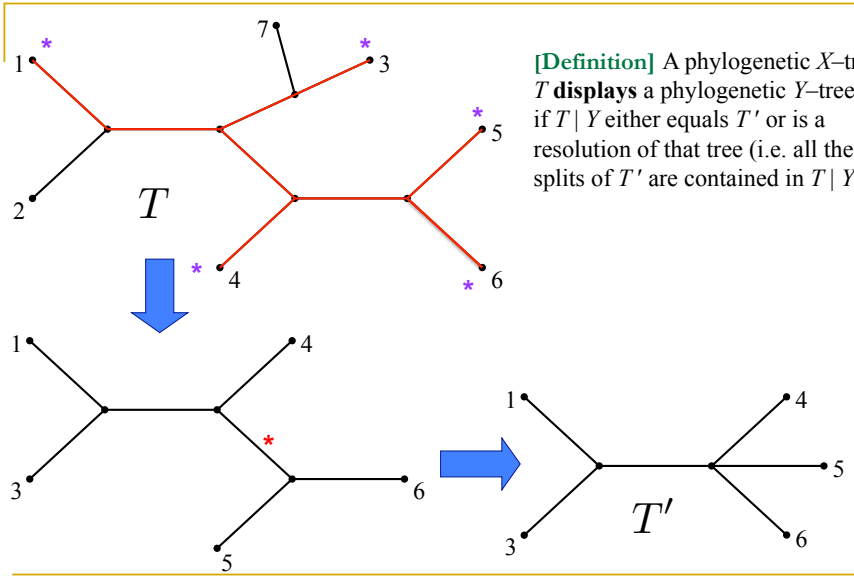
A.D. Gordon, (J. Classif. 1986)

Mike Steel

ALLAN WILSON CENTRE

Outline

- *Part 1:* **Subtrees and supertrees**
- *Part 2:* **Compatibility**
- *Part 3:* **Defining sets**
  - **20x** *Head, shoulders, knees and toes*
- *Part 4:* **Specialist topic: "decisiveness"**



[Definition] A phylogenetic X–tree T **displays** a phylogenetic Y–tree, T' if $T \mid Y$ either equals T' or is a resolution of that tree (i.e. all the splits of T' are contained in $T \mid Y$).

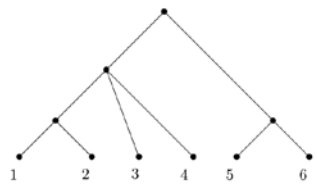
Counting trees II

Quiz:

Suppose T is a binary phylogenetic-tree on leaf set Y (subset of X). How many binary phylogenetic X-trees display Y?

$$b(n)/b(k) \qquad n = |X|, k = |Y|$$

Rooted trees



A rooted phylogenetic tree $\mathcal{T}$ that displays 12|3 and 13|6 but not 13|4 nor 15|4

Display via quartet encodings

Given $T \in U(X)$ and $T' \in U(Y)$ (where $Y \subseteq X$)

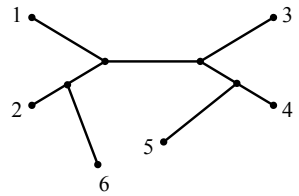
$$T \text{ displays } T' \iff Q(T') \subseteq Q(T).$$

Similarly for rooted trees

Compatibility

A set P of trees is **compatible** if there is a phylogenetic X-tree T that **displays** each tree

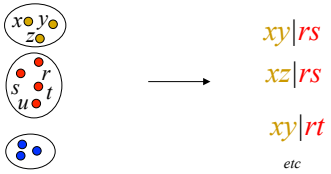
- *Example:* $P = \{12|34,\ 13|45,\ 14|26\}$



If T is the only tree that displays each tree in P we say that P **defines** T.

Equivalence of character compatibility and (quartet) tree compatibility

$$\mathcal{C} \mapsto Q(\mathcal{C})$$



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

How hard is it to tell if a set of trees is compatible?

In general it's (NP)-hard, even for quartet trees (so character compatibility is too, by last slide!)

But it's easy in some special cases…..

Special case 1: Trees have same leaf sets

$$\mathcal{P} = \{T_1, \ldots, T_k\} \subseteq U(X)$$

$$\mathcal{P} \text{ is compatible} \iff \Sigma = \bigcup_{i=1}^{k} \Sigma(T_i) \text{ is p.c.}$$

Special case 2: Two trees

Given T_1 and T_2 on leaf sets X_1, X_2, let $Y = X_1 \cap X_2$

$$\{T_1, T_2\} \text{ is compatible} \iff \{T_1|Y, T_2|Y\} \text{ is}$$

More generally *k* trees, with *k* fixed (FPT)

Special case 3: Q quartets with $|Q| = \binom{n}{4}$

•*Recall from* Part 2: [Colonius and Schultze 1981]

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

$$ab|cd \in Q \Rightarrow ac|bd, ad|bc \notin Q$$
$$ab|cd \in Q \Rightarrow ab|ce \in Q \text{ or } ae|cd \in Q.$$

Corollary:

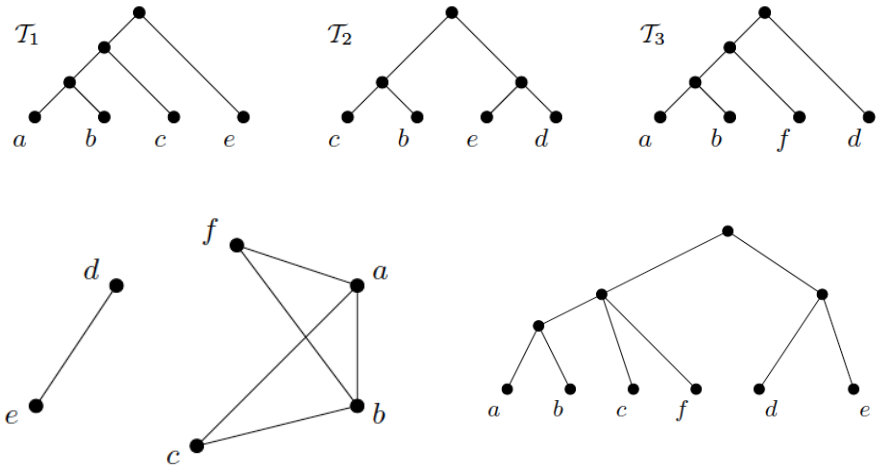
$$\text{If } |Q| = \binom{n}{4} \text{ then } Q \text{ is compatible}$$
$$\Leftrightarrow \text{ every subset of } Q \text{ of size 3 is}$$

Special case 4: Rooted trees

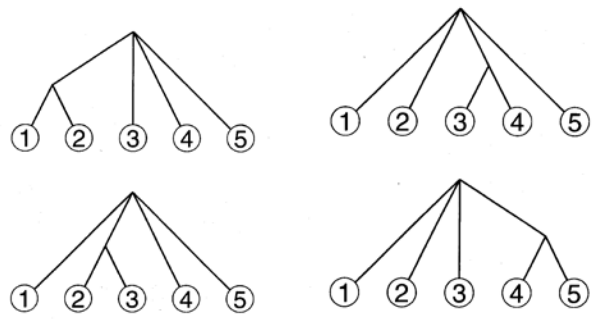
For a set *R* of rooted trees, there is a fast algorithm which determines whether or not *R* is compatible (and if so constructs a canonical tree A_R) that displays each tree in R.

Same applies for a set of unrooted trees that each contain a fixed leaf *x*

Aho *et al* tree ($\mathcal{A}_R$) [1981]



Our example from lecture 2:



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

Properties of the Aho tree

$\mathcal{A}_R$ is a minimal tree that displays R
(but there can be exp. many such trees!)

$$\mathcal{R}_1 = \{ab|c, ac|d\}$$
$$\mathcal{R}_2 = \{ab|g_1, ab|g_2, \ldots, ab|g_n\}$$
$$\mathcal{R} = \mathcal{R}_1 \cup \mathcal{R}_2$$

$\mathcal{A}_R$ is a binary tree if and only if R defines a (that!) binary tree

Proposition [Bryant]

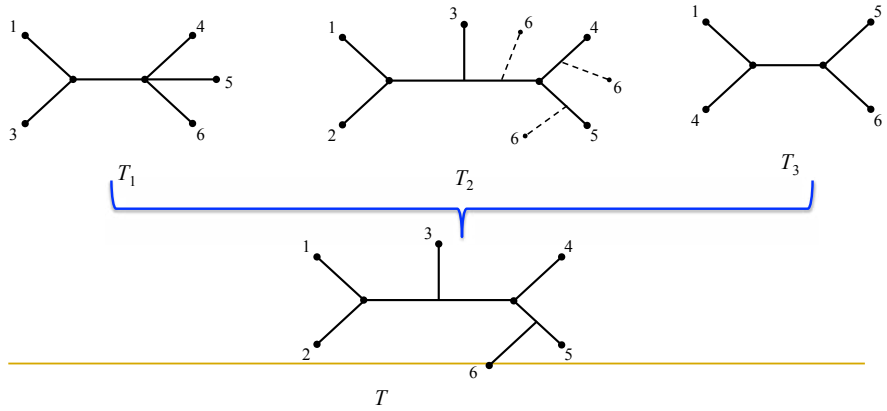
$\mathcal{A}_R$ is the Adams consensus tree of all the rooted X-trees that display R

[Recall definition above]

A collection of phylogenetic trees $T_1, \ldots, T_k$ **defines** a phylogenetic X–tree T if

X is the union of the leaf sets of the trees $T_1, \ldots, T_k$ and

there exists one, and only one phylogenetic X–tree that displays these trees, and this tree is T.



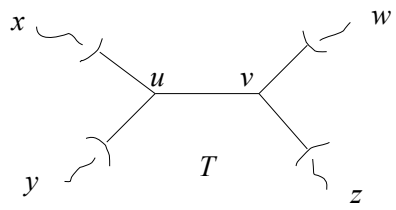
The nice story: Rooted trees

- R defines T if and only if every interior edge of T is 'distinguished' by some rooted triplet $ab|c$ from R

PIC

The unrooted case (more interesting…)

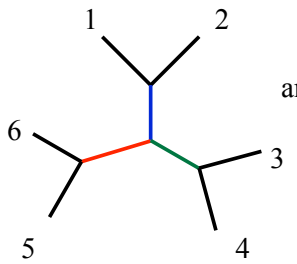
- **Definition:** For a binary phylogenetic tree T, a quartet tree xy|wz *distinguishes* an interior edge e={u,v} of T if T displays xy|wz and and e is the only edge shared by the the paths from x to w and y to z



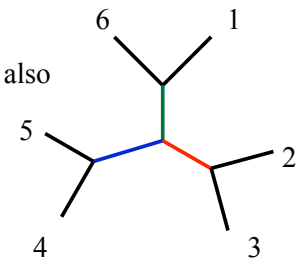
Observation: If Q defines T then T is binary and every interior edge of T is distinguished by at least one quaretfrom Q. So $|Q| \geq n-3$

Warning:

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



and also



!

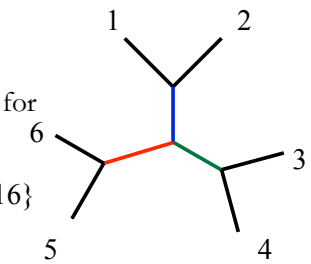
'Islands' in NNI (rooted) tree space

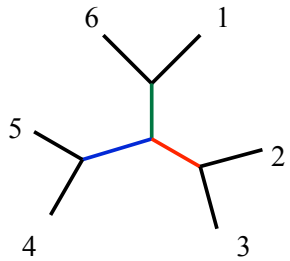
Theorem [Magnus Bordewich PhD thesis (2003)]

The set of rooted phylogenetic trees that display a set of rooted trees is connected under (rooted) NNI operations.

This does *not* hold for unrooted trees!

Q ={12|45, 56|23, 34|16}





Sufficient condition for Q to define T:

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

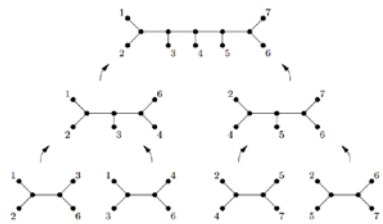
Proposition: If there is an element of X that is a leaf of every tree in Q then Q defines T. [why?]

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

The Böcker-Dress theorem:

Recall if Q defines a tree then $\underbrace{L(Q) - 3 - |Q|}_{\mathrm{exc}(Q)} \leq 0$

Definition: A set Q of quartet trees is "good" if
(i) Q defines a phylogenetic tree, and
(ii) exc(Q) = 0



Theorem
[Böcker, Dress 1999; Grünewald 2012]
Any good set of ($\geq$2) quartets is the disjoint union of precisely two good sets

Observations

Definition: A set Q of quartet trees is "good" if
(i) Q defines a phylogenetic tree, and
(ii) exc(Q) = 0

- Determining if Q defines a phylogenetic tree is NP-hard[1]
- Determining if Q is 'good' is easy.
- Determining if Q contains within it an (unknown) 'good' subset is too!
- Examples of 'good' sets include the 'linked quartet systems' (E. Price and J. Rusinko, 2014).

[1]Maria Luisa Bonet, Simone Linz, and Katherine St. John (2012),

Key idea(s) in the proof of the B+D theorem:

Slim sets; 'patchworks' 

- Grünewald's proof relies on a strong (and suprising) sufficient condition for a set P of phylogenetic trees to be definitive:

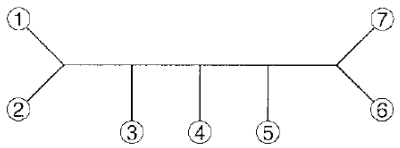
$$\mathrm{exc}(\mathcal{P}) = |L(\mathcal{P})| - 3 - \sum_{T \in \mathcal{P}} |E_{\mathrm{int}}(T)|$$

$\mathcal{P}$ is *slim* if $exc(\mathcal{P}') \geq 0$ for every non-empty subset $\mathcal{P}' \subseteq \mathcal{P}$

Theorem:
Every slim set of binary phylogenetic trees is compatible.

Gruenewald, S. (2012) Slim sets of binary trees Original Research Article
Journal of Combinatorial Theory, Series A, Volume 119(2): 323-330

Quesiton: If Q defines a phylogenetic tree, T, does it always contain an excess-free subset that defines T?



$$Q = \{12|35, 24|57, 13|47, 34|56, 15|67\}$$

A **minimum** defining set of quartets has size *n*-3.

But how big can **minimal** defining set be?

But how big can **minimal** defining set be?

Theorem: [Dietrich, M., McCartin, C., and Semple, C. (2012)]

The largest minimal defining set of quartet trees on n leaves has size between:

$$\tfrac{1}{4}(n^2 - 4n + 3) \text{ and } n^3$$

Conjecture: Quadratic is the actual order!

Supertree methods

- Given different (usually incompatible) phylogenetic trees on overlapping sets of species we want to combine them into a tree that classifies all the species.
- Several methods. The main one in use is MRP ('matrix recoding with parsimony').
- Any supertree method can be used as a consensus method:
  - Bryant's result (lecture 3) implies that any MRP tree refines the strict consensus tree

Quiz time….

- Is there a supertree method for rooted trees with this property:
  - If every tree displays $ab|c$ then the supertree does too.
- Is there a supertree method for unrooted trees with this property:
  - If every tree displays $ab|cd$ then the supertree does too.

Special Topic: Decisiveness

Taxon (108 total)	N_{el}	0	1	2	14	15	16	17	18	19	20	21	22	23	51	52	53	54	55
Acanthocalycium	3		X			X	X												
Acanthocereus	3		X			X								X					
Acharagma	1	X																	
Ancistrocactus	1	X																	
Ariocarpus	1	X																	
Armatocereus	3		X			X								X					
Arrojadoa	4				X	X	X	X											
Astrophytum	2	X	X																
Austrocactus	3		X			X		X											
Austrocylindropuntia	9	X	X	X	X						X	X	X		X	X			
Aztekium	2	X	X																
Bergerocactus	3	X				X								X					
Blossfeldia	7		X	X		X				X	X	X	X						
Brasilopuntia	5		X	X							X	X	X						
Browningia	6		X		X	X	X	X	X										
Calymmanthium	9	X	X	X	X	X					X	X	X	X					
Carnegiea	6					X				X				X			X	X	X
Castellanosia	2		X			X													
Cephalocereus	3					X				X				X					
Cereus	9		X	X	X	X		X		X	X	X	X						
Cintia	3					X	X	X											
Cipocereus	3					X	X	X											
Cleistocactus	3					X	X	X											
Coleocephalocereus	5		X			X	X	X		X									
Copiapoa	7		X		X	X	X	X							X	X			
Corryocactus	3	X	X			X													
Coryphantha	2	X		X															
Dendrocereus	2					X								X					
Denmoza	3					X	X	X											
Discocactus	3					X	X	X											
Disocactus	3		X			X								X					

Group	Taxa	Loci	% Missing	Citation
Metazoa	77	150	55	Dunn et al. 2008
Papilionoid legumes	2228	39	96	McMahon and Sanderson 2006
Asterales	4954	5	91	Smith et al. 2009
Eukaryotes	73060	13	92	Goloboff et al. 2009

Taxon coverage pattern

	Gene1	Gene2
a	x	x
b	x	x
c	x	x
d	x	
e		x

Two taxon sets: {***a,b,c,d***} and {***a,b,c,e***}

S = {{***a,b,c,d***},{***a,b,c,e***}}

Definitions ("decisiveness")

- *Definition:* Decisiveness (for *T*):

For a collection $\boldsymbol{S} = \{Y_1, \ldots, Y_k\}$ of subsets of X, with union X, $\boldsymbol{S}$ is **decisive for a tree** T provided that $T \mid Y_1, \ldots, T \mid Y_k$ defines T.

i.e. T is the only tree that displays $T \mid Y_1, \ldots, T \mid Y_k$

- *Definition:* (Global decisiveness)

A collection $S = \{Y_1, \ldots, Y_k\}$ of subsets of X, is **phylogenetically decisive** if it is decisive for every phylogenetic X-tree.

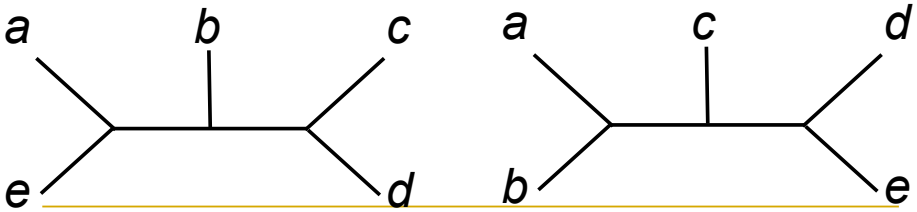
Not phylogenetically decisive (for all trees)

Of the 15 possible binary unrooted trees for this data set...

a	x	x
b	x	x
c	x	x
d	x	
e		x

There are 6, like that below, where the taxon coverage is decisive

...and 9, like that below, where it is not decisive



Phylogenetically decisive (for all trees)

a	x	x	x	x
b	x	x	x	
c	x	x		x
d	x		x	x
e		x	x	x

Necessary condition:

$$\binom{X}{3} \subseteq \bigcup_{j=1}^{k} \binom{Y_j}{3}.$$

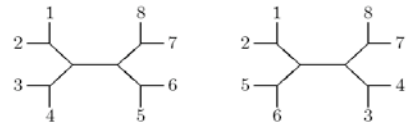
Why?

..but not sufficient!

Example [from Peter Humphries, 2008]

- 8 taxa: 1,2,3,…,8
- All 4-element subsets that contain {1,2}, or {3,4}, or {5,6} or {7,8}.
- Each column has 50% coverage.

1	**x**	**x**	**x**	…	x		…			…	x
2	**x**	**x**	**x**	…			…		x	…	
3				…	**x**	**x**	…	x	x	…	x
4	x		x	…	**x**	**x**	…			…	
5		x		…			…	**x**	**x**	…	
6	x	x		…		x	…	**x**	**x**	…	
7				…	x	x	…			…	**x**
8			x	…			…	x		…	**x**

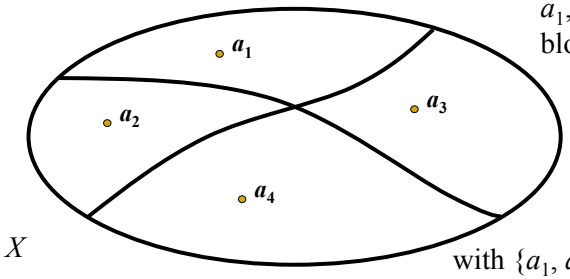


Theorem **[S+Sanderson 2010]:**

S is phylogenetically decisive ⇔ *S* satisfies the *4-way partition property** for *X*.

*For all partitions of *X* into four parts:

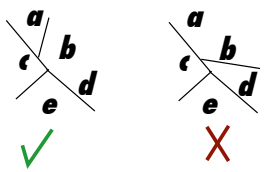
there exists representative a_1, a_2, a_3, a_4 from each block:



with $\{a_1, a_2, a_3, a_4\} \subseteq Y_i$ for some $Y_i \in S$

Examples

a	x	x
b	x	x
c	x	x
d	x	
e		x



1	**x**	**x**	**x**	…	x		…			…	x
2	**x**	**x**	**x**	…			…		x	…	
3				…	**x**	**x**	…	x	x	…	x
4	x		x	…	**x**	**x**	…			…	
5		x		…			…	**x**	**x**	…	
6	x	x		…		x	…	**x**	**x**	…	
7				…	x	x	…			…	**x**
8			x	…			…	x		…	**x**

Complexity of determining decisiveness?

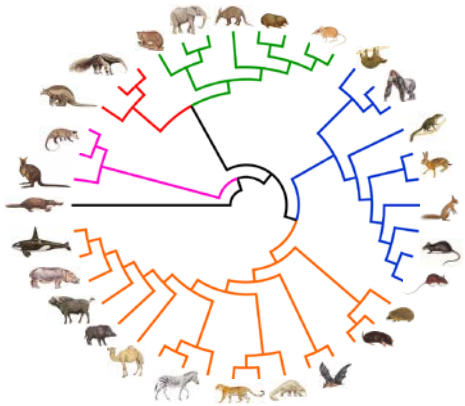
- [*cf.* Manuel Bodirsky's `No rainbow colouring problem' for 3-partitions]

- **Theorem** [June 15, 2012, Mareike Fischer]

There is an $O(n^{16}k)$ algorithm for determining phylogenetic decisiveness!

- THE END

Lecture 6: Stochastic models I



from F. Delsuc and N. Lartillot

Mike Steel

ALLAN WILSON CENTRE

Biomathematics Research Centre

Winthrop lectures, 2014

ALLAN WILSON CENTRE

Outline

- *Part 1:* **Why models?**
- *Part 2:* **Markov processes on trees**
- *Part 3:* **Statistical methods for inference**
  - **20x pushups**
- *Part 4:* **Specialist topic: Ancestral state reconstruction**

Why models?

Genetic characters 'evolve' on a (gene) tree under some random process.

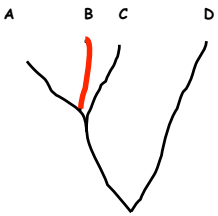
The gene tree is also random (even conditional on the species tree), due to 'lineage sorting' (or LGT).

Some questions:

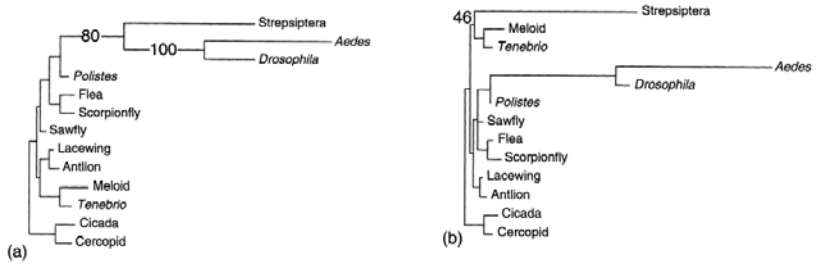
- Will existing methods (parsimony etc) recover the correct tree?
- If not, can can approaches do so (e.g. corrections, ML, Bayesian methods)?
- How much data do we need (to find a tree, or branch length, or resolve a polytomy or estimate an ancestral states) accurately?

Example: The "Felsenstein Zone" (1978)

Joseph Felsenstein

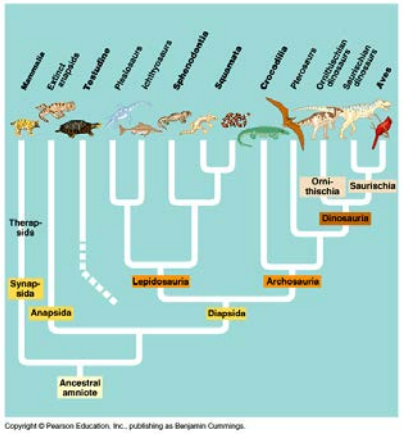


Does it happen?



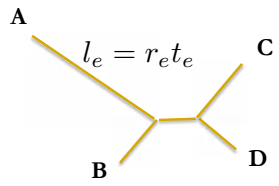
J. Huelsenbeck 1998: Is the Felsenstein Zone a fly trap?

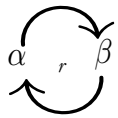
Example 2 (process changes across a tree)



18S rRNA

Simplest model: 2-state symmetric model





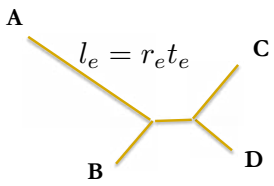
$p =$ probability initial and final state are different

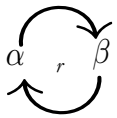
$$p = \frac{1}{2}(1 - \exp(-2rt))$$

$$p_e = \frac{1}{2}(1 - \exp(-2l_e))$$

$$l_e = -\frac{1}{2}\log(1 - 2p_e)$$

Simplest model: 2-state symmetric model





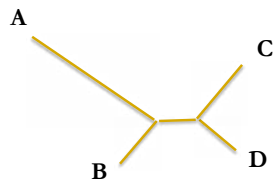
$$p_e = \frac{1}{2}(1 - \exp(-2l_e))$$

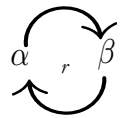
$$l_e = -\frac{1}{2}\log(1 - 2p_e)$$

Path P connecting two vertices of x, y of T:

$$\mathbb{P}(f(x) \neq f(y)) = \frac{1}{2}(1 - \prod_{e \in P}(1 - 2p_e))$$

Remarks





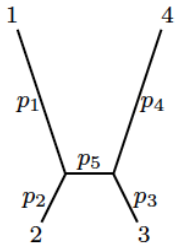
Reversibility

Change events on edges are independent

(more generally, change events on paths are independent if the paths are edge-disjoint).

The $2p$ version of the model

Alternative ways to view the 2-state model

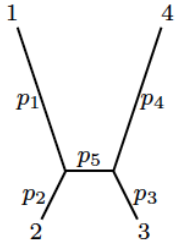


$$p_A = \mathbb{P}(\{x : f(x) \neq f(n)\} = A)$$

$$p_\emptyset = (1-p_1)(1-p_2)(1-p_3)(1-p_4)(1-p_5) + p_1p_2p_3p_4(1-p_5)$$

$$+p_1p_2p_5(1-p_3)(1-p_4) + p_3p_4p_5(1-p_1)(1-p_2).$$

Discrete fourier analysis for the 2-state model



$$p_\emptyset = (1-p_1)(1-p_2)(1-p_3)(1-p_4)(1-p_5) + p_1p_2p_3p_4(1-p_5)$$

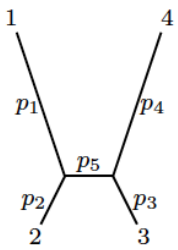
$$+p_1p_2p_5(1-p_3)(1-p_4) + p_3p_4p_5(1-p_1)(1-p_2).$$

$$p_A = \frac{1}{2^{n-1}} \sum_{\substack{B \subseteq [n] \\ |B| \equiv 0 \bmod 2}} (-1)^{|A \cap B|} \prod_{e \in P(T,B)} (1 - 2p_e)$$

$$p_\emptyset = \frac{1}{8}(1 + x_1x_2 + x_3x_4 + x_1x_3x_5 + x_2x_3x_5 + x_1x_4x_5 + x_2x_4x_5 + x_1x_2x_3x_4)$$

$$p_{12} = \frac{1}{8}(1 + x_1x_2 + x_3x_4 - x_1x_3x_5 - x_2x_3x_5 - x_1x_4x_5 - x_2x_4x_5 + x_1x_2x_3x_4)$$

Application 1: Felsenstein zone



$$\mathbb{E}[\Delta] = p_{23} - p_{12}$$

$$p_{23} = \frac{1}{8}(1 - x_1x_2 - x_3x_4 - x_1x_3x_5 + x_2x_3x_5 + x_1x_4x_5 - x_2x_4x_5 + x_1x_2x_3x_4)$$

$$p_{12} = \frac{1}{8}(1 + x_1x_2 + x_3x_4 - x_1x_3x_5 - x_2x_3x_5 - x_1x_4x_5 - x_2x_4x_5 + x_1x_2x_3x_4)$$

$$p_1 = p_4 = P, p_2 = p_3 = p_5 = Q$$

$\mathbb{E}[\Delta] > 0$ precisely if $P^2 > Q(1-Q)$

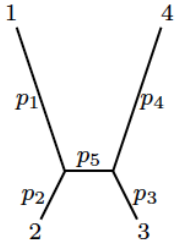
Exercise: Solve the general case!

Some observations and further results

- MP is inconsistent when n=4. Lengths of edges can be arbitrarily small.
- But if the edge lengths are clock-like it **is** consistent for n=4
- For n=5 this inconsistency occurs even at with clock-like branch lengths.
- It's worse when n=6!
- For n large enough MP can even be inconsistent when all edges have the same length (not clock-like).

Conjecture: For some l >0, MP is consistent on **all** binary phylogenetic trees provided all edges have equal length of l (or less).

Application 2: Phylogenetic invariants



$$p_A = \frac{1}{2^{n-1}} \boxed{\sum_{\substack{B \subseteq [n] \\ |B| \equiv 0 \bmod 2}} (-1)^{|A \cap B|}} \prod_{e \in P(T,B)} (1 - 2p_e)$$

A Hadamard matrix of rank 2^{n-1}

$$x_1x_2x_3x_4 = (x_1x_2)(x_3x_4)$$

$$(x_1x_3x_5)(x_2x_4x_5) = (x_1x_4x_5)(x_2x_3x_5)$$

These correspond to two quadratic equations in the p_A values.

Application 3: Homoplasy-rich characters are always unlikely…

$$\mathbb{P}(f) \leq 2^{-\mathrm{ps}(f,T)}$$

Why?

This is best possible $\sup_{0<l_*<\infty} \mathbb{P}(f) = 2^{-\mathrm{ps}(f,T)}$

For any binary character data, the maximum likelihood tree(s) under the 2-state model, in with edge lengths chosen freely for each character are precisely the maximum parsimony tree(s).

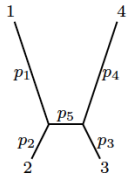
Similar for the r-state symmetric model (but Menger's argument no longer works!)

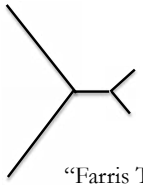
Statistically consistent methods for inferring a tree

- Corrected distances
- ML (maximum likehood)
  - RAxML, PhyML, etc
  - Usual version is 'average ML'
- Bayesian methods
  - MrBayes, 
  - BEAST 
  - Can compare support for hypotheses by averaging over all trees

Statistically consistent methods

- Is ML more accurate on all trees than MP?





"Farris Tree"

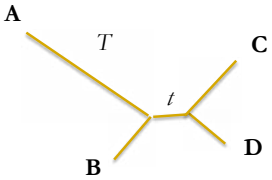
$$\lim_{L\to\infty}\lim_{k\to\infty}\mathbb{P}(ML \text{ returns correct tree}) = 1$$

$$\lim_{k\to\infty}\lim_{L\to\infty}\mathbb{P}(ML \text{ returns correct tree}) \le \frac{1}{2}$$

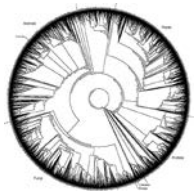
- The dangers of doing simulations….

Problems for reconstructing a tree (even when the model is known and nice!)

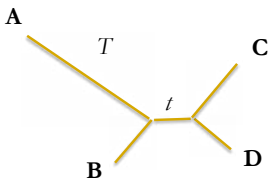
- **Short interior edges**
- **Long edges**



- **Many taxa (n)**



Finite state models: short and long edges

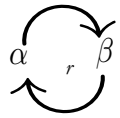


k = sequence length needed to accurately reconstruct this tree

as T grows, k grows at rate $\exp(cT)$

What about is t shrinks?

Finite state model

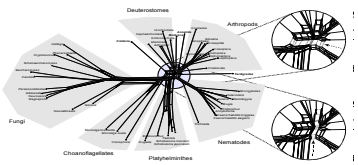


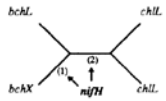
as $t \to 0$, k grows at rate $\frac{1}{t^2}$

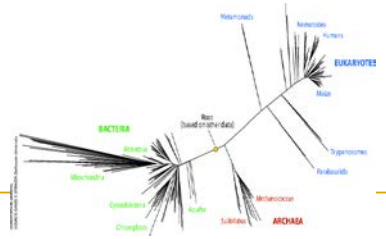
but if $T = t$ then as $t \to 0$, k grows at the rate $\frac{1}{t}$

Examples of deep and controversial phylogenetic resolutions

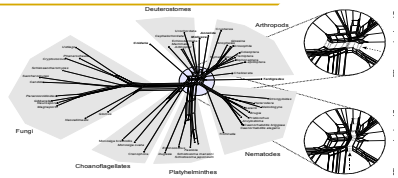
- Origin of metazoa (~550-600 mya)
- Origin of photosynthesis (>2.5 bya)
- Rooting the 'tree' of life (~3.5 bya)

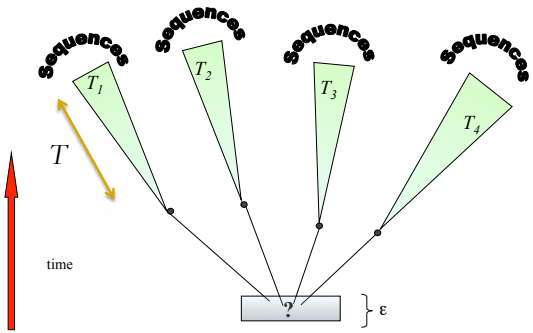






Deep divergences





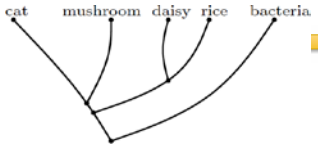
$$k = \Theta\left(\frac{1}{\epsilon^2}\right)$$

$$k = \Theta\left(\frac{\exp(cT)}{n}\right)$$

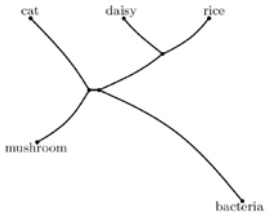
Question: How do these two factors (short, long) interact?

$$k = \Theta\left(\exp(cT) \times \frac{1}{\epsilon^2}\right)$$

How does the required sequence length (for tree reconstruction) depend on n (=# taxa)?



Cat .....**ACCCGTCGTT....**
Daisy .... **CACCATCGTT...**
Rice.....**AACCAGCGTT...**



#data-sets of k characters for n species, over an r-letter alphabet

$$= (r^n)^k = r^{nk} \qquad b(n) = 2^{\Omega(n\log(n))}$$

$$\Rightarrow k \geq c \cdot \log(n)$$

Fine, but what about 'evolved' data

Suppose we evolve k characters independently on a tree under a 2-state symmetric model with

$$p(e) \in [p, P] \text{ for every edge } e$$

Theorem 1 [Erdos, PL, Szekeley, S, Warnow (1999)]

For some ('stringy') trees accurate tree reconstruction is possible with $k = \Theta(\log(n))$

But for other ('bushy') trees our approach required $k = \Theta(n^t)$

However, for almost all trees it suffices to have: $k = \Theta(\log(n)^s)$

Conjecture: Provided that $P < \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right)$ accurate tree reconstruction can be achieved for ALL trees with $k = \Theta(\log(n))$

Theorem 2 [Daskalakis, Mossel, Roch (2011)]

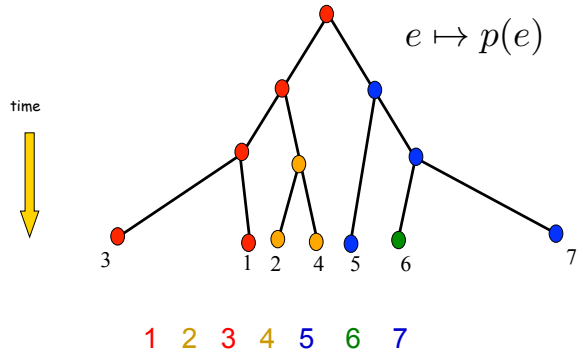
This conjecture holds (and is tight)

Can adding more taxa help (even if you don't care about them)?

Add taxa ➡ build tree ➡ ignore the added taxa

Sequence length required to find the correct tree (on the subset of species) can be reduced logarithmically this way

Specialist topic: a model that generates homoplasy-free data



Kimura and Crow's "infinite alleles" model.
The probability of any partition can be computed via Mobius inversion (Evans et al. 2004)

How many such 'evolved' characters are needed?

$$P = \max\{p(e)\}, p = \min\{p(e) : e \text{ is interior}\}$$

Theorem [Mossel +S, 2004]

For $P < \frac{1}{2}$, the number of characters k needed to corrected reconstruct T (w.p. $> 1 - \epsilon$) is: $k = c \cdot \frac{\log(n)}{p}$

- Proof relies on combinatorial arguments, and basic property of branching processes.
- $P > ½$, k changes to poly(n).

Does finding a tree need more data than to 'test' if a given one is correct?


- *Reconstructing:*
  - Given k characters generated by (unknown) tree T:
    - We need log(n) sites for finite-state and infinite state models to reconstruct T.
- *Testing:*
  - Given data, and candidate tree, T_c, is $T = T_c$?

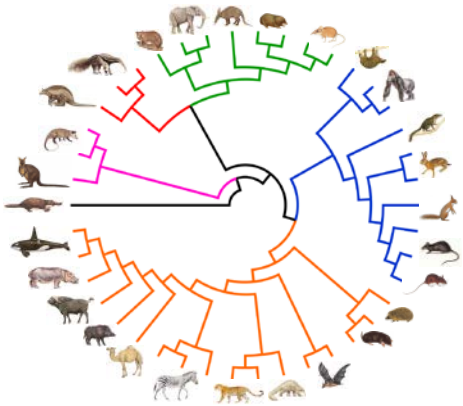
Theorem

- For finite-state data we still need log(n) sites to test
- But for infinite-state data a constant(!) number of sites suffices

- *Teasing:*
  Given data, and that '$T = T_1$ or T_2', which is tree is it?

THE END

Lecture 7-8: Stochastic models II



from F. Delsuc and N. Lartillot

Mike Steel

ALLAN WILSON CENTRE

Biomathematics Research Centre

Winthrop lectures, 2014

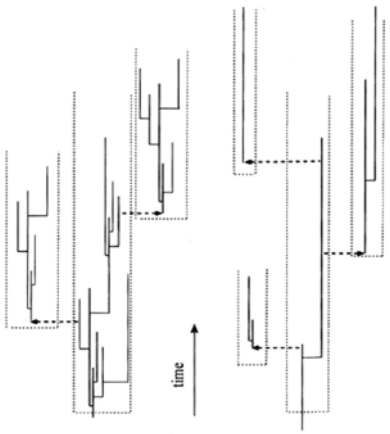
ALLAN WILSON CENTRE

Outline

- *Part 1:* **Speciation/extinction models**
- *Part 2:* **Shapes of trees**
- *Part 3:* **Predicting future PD**
  - **20 x deep breaths**
- *Part 4:* **Specialist topic: Predicting the past**

Yule model

Number of species in genus	Number of genera Observed	Calculated
1	131	
2	35	
3	28	
4	17	
5	16	
6	9	
7	8	
8	8	
9 to 11	13	
12 to 14	3	
15 to 20	7	
21 to 34	14	
35 upwards	4	
Total	293	293.0



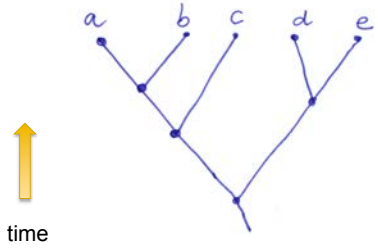
From 'Branching processes in biology' Kimmel and Axelrod

$$\mathbb{P}(N = n) \approx n^{-1-g/\lambda}$$

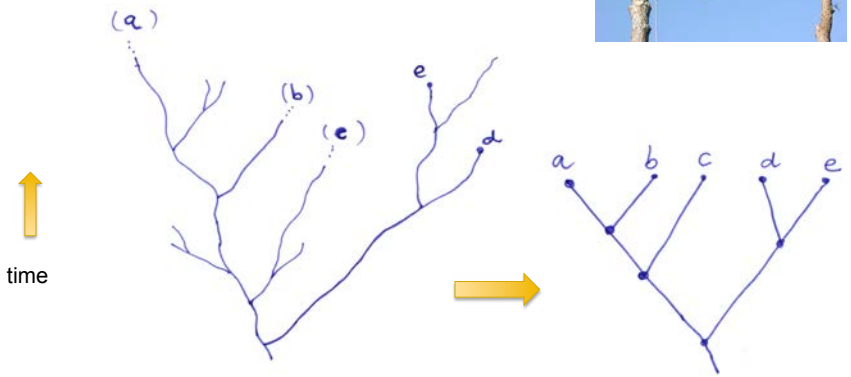
[1]G. U. Yule, A mathematical theory of evolution. Based on the Conclusions of Dr. J.C. Willis, F.R.S. Phil. Trans. Roy. Soc. 213 (1925), 21-87.

Where do evolutionary trees comes from?

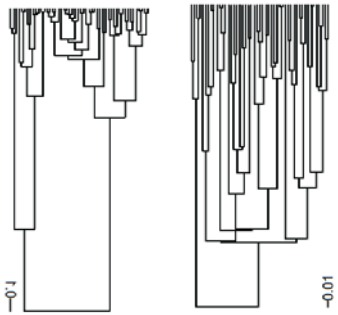
Forestry Unit men tree-felling in Southern Italy



Another viewpoint



Simplest speciation model: pure birth (constant rate)



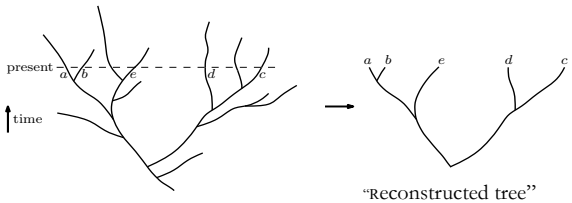
Kingman coalescent trees

Yule (pure birth) model

Each lineage gives birth independently at some constant rate λ

Grow for time *t*, or till it has *n* leaves, or condition on both *n* and *t*.

Birth-death models: simplest case (constant birth-death rates) $\lambda = \text{constant}, \mu = \text{constant}$



"Reconstructed tree"

Sean Nee

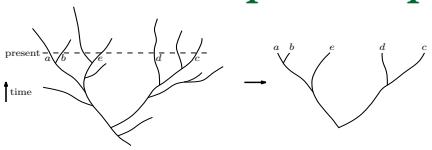
$$\mathbb{E}(N_t) = e^{(\lambda-\mu)t}$$

The 'reconstructed' tree can be conditioned on

- *n, or t or*
- *n* and *t*
- *t* and the event that *n*>0

The 'pull of the present' and 'push of the past'

"Pull of the present/push of the past"

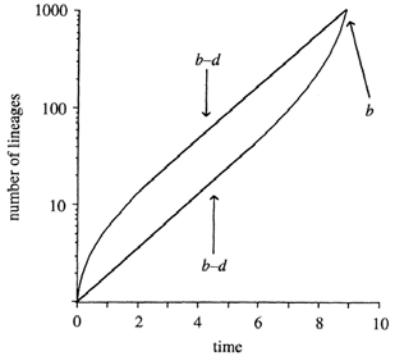


"Reconstructed tree"

$\lambda = \text{constant}, \mu = \text{constant}$

$$\mathbb{E}(N_t) = e^{(\lambda-\mu)t}$$

Sean Nee



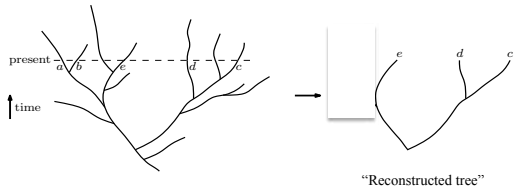

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Extinction Rates can be Estimated from Molecular Phylogenies

Sean Nee, Edward C. Holmes, Robert M. May and Paul H. Harvey

Phil. Trans. R. Soc. Lond. B 1994 **344**, doi: 10.1098/rstb.1994.0054, published 29 April 1994

A nice (but also annoying) property of constant b-d models



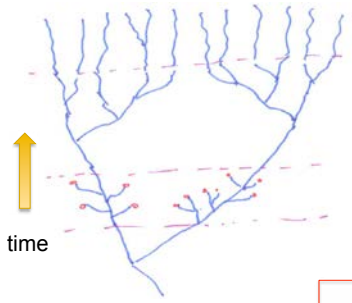
$$f = \text{fraction sampled at present}$$

$$\lambda \geq \mu \geq \lambda(1-f)$$

Conditioning on *n* (or *n* and *t*) the reconstructed tree has the same distribution as complete sampling with adjusted birth-death rates

$$\hat{\lambda} = f\lambda \qquad \hat{\mu} = \mu - \lambda(1-f)$$

Two extensions where *it's just so lovely….*



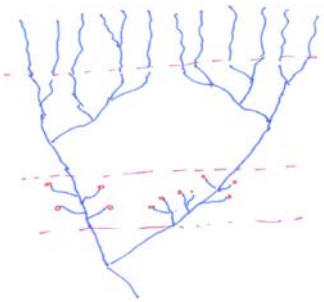
Amaury Lambert

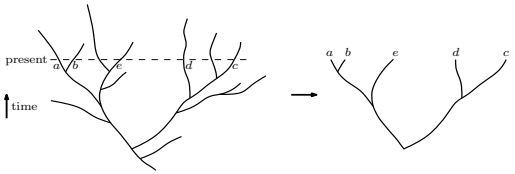
Tanja Stadler

$$\lambda = \lambda(t), \mu = \mu(t,a)$$

$$\boxed{\lambda = \lambda(t,N), \mu = \mu(t,N,a)}$$

Less is more…





Evolutionary tree

Reconstructed tree

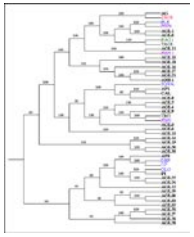
$$\lambda = \lambda(t,N), \mu = \mu(t,N,a)$$

Proposition: [Aldous; Lambert and Stadler]

All such models (as well as **Kingman's coalescent model!**) lead to same distribution on the reconstructed tree (**ignoring branch lengths**) – namely the Yule-Harding distribution (lecture 1)

Real trees

$$\lambda = \lambda(t,N), \mu = \mu(t,N,a)$$



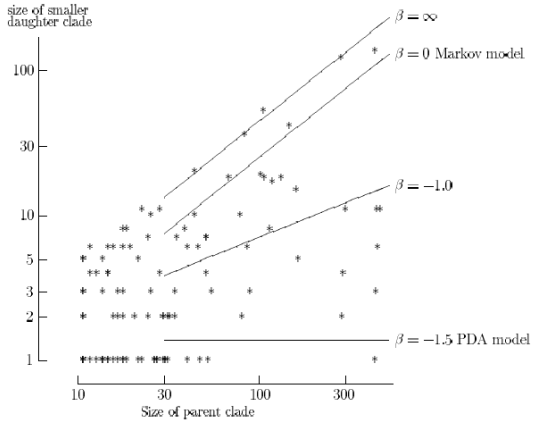
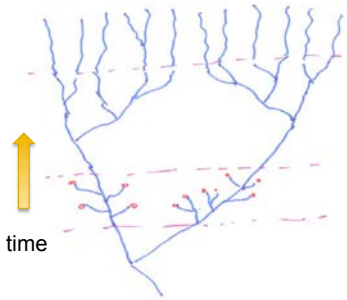


FIG. 3. *Splits in the tree of Chase et al (1993), and approximate median lines for the beta-splitting model. Note the log-log scale.*

From: Aldous, D. (2001). Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science* 16: 23-34

Life gets even better if we are slightly less general

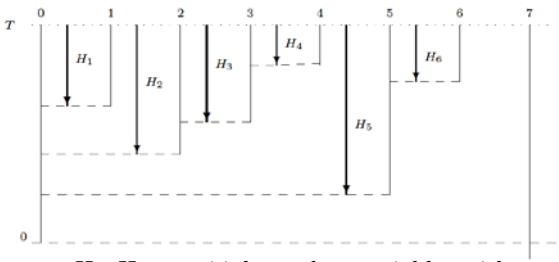


$$\lambda = \lambda(t), \mu = \mu(t, a)$$

$$\lambda = \lambda(t, N), \mu = \mu(t, N, a)$$

Models where the reconstructed tree can be described by a 'coalescent point process'

$$\lambda = \lambda(t), \mu = \mu(t, a)$$

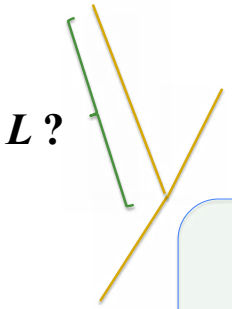


Allows conditioning on *n, t or n* and *t*

$H_1, H_2, \ldots,$ i.i.d. random variables with some distribution F

Example: A pure-birth process $1 - F(t) = \mathbb{P}(H > t) = e^{-\lambda t}$

How long are the branches?



Speciation rate = 1/million years

so the expected value of L equals 1 million years

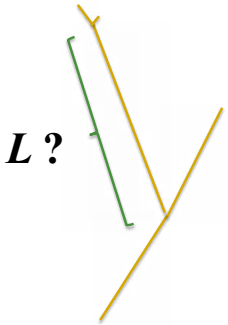
The bus 'paradox'

"It wouldn't hurt to wait around for a little while."

You turn up at a bus stop, with no idea when the next bus will arrive.

- If buses arrive regularly every 20 mins what is your expected waiting time?
- If buses arrive randomly every 20 mins what is your expected waiting time?

Length of a randomly selected branch

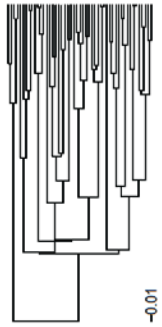


"It wouldn't hurt to wait around for a little while."

Expected value of L is 1 million years

Quiz

A pure-birth tree evolves with each lineage randomly generating a new lineage on average once every **1 million years** (no extinction).



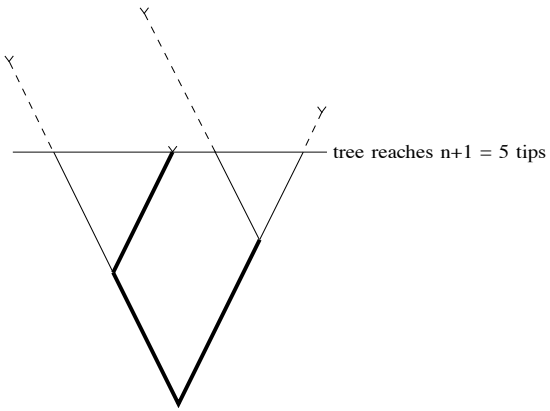
Look at the tree when it has 100 species

What is the expected length of a randomly selected *extant* branch?

Answer 1: 1 million years? ✗

Answer 2: 500,000 years? ✓

The tree puzzle (I):



What about ancestral lineages?

Solution 1: Conditioning on n:

Grow tree till it has n+1 leaves (then go back 1 second!)

p_n = average length of the n pendant edges

i_n = average length of the n-1 internal edges

Theorem:

$$\mathbb{E}[p_n] = \mathbb{E}[i_n] = \frac{1}{2\lambda}$$

same for both!

The tree puzzle (II):

A tree evolves with each lineage randomly generating a new lineage on average once every **1 million years** (no extinction).

Look at the tree **after 500 million years**

What is the expected length of a randomly selected (*extant or ancestral*) lineage?

Answer 1: **1 million years?**

Answer 2: **500,000 years?** ✓

Solution 2: Conditioning on *t*:

In a binary Yule tree, grown for time *t*, let

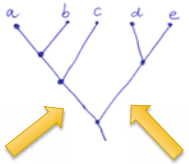
p(t) = expected length of the average pendant edge
i(t) = expected length of the average interior edge

Theorem:

$$\mathbb{E}[p(t)] = \frac{1}{2\lambda} + O(e^{-t})$$
$$\mathbb{E}[i(t)] = \frac{1}{2\lambda} + O(e^{-t})$$

$$\frac{dL}{dt} = \mathbb{E}[N_t] = 2e^{\lambda t}$$
$$\frac{dI}{dt} = \lambda P$$
$$L = I + P$$

What about a 'specific' edge (e.g. a 'root edge')?



A tree evolves with each lineage randomly generating a new lineage on average once every **1 million years** (no extinction).

Look at the tree when it first has 100 species

What is the expected length of a randomly selected *root* lineage?

Answer 1: **1 million years?** ✓

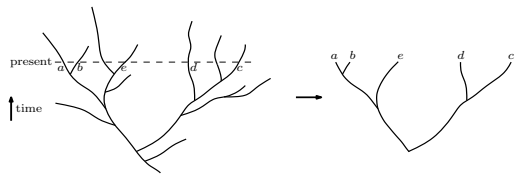
$$E[L \mid n] = \frac{1}{\lambda}\left(1 - \frac{1}{n}\right)$$

Answer 2: **500,000 years?**

Answer 3: **990,000 years** ✓

The tree puzzle (III):

Now **suppose extinction occurs** at the same rate as speciation (one per one million years). Suppose we observe a tree today that has 100 species.



What is the expected length of a randomly selected *extant* lineage?

Answer 1: **1 million years?** ✓

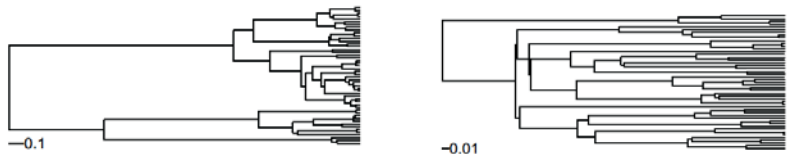
Answer 2: **500,000 years?** ✗

What do 'real' trees look like?

- Current plant and animal diversity preserves at most 1-2% of the species that have existed over the past 600 my". [Erwim, PNAS 2008].
- Set extinction rate = speciation rate?
- **Problem:** If extinction rate =speciation rate the tree is guaranteed to eventually die out eventually!
- **Solution?**: Condition on the tree not dying out (or having n species today)

Less 'realistic models' can fit the data better:

- **Real** reconstructed trees generally look more like Yule trees with zero extinction rate than birth-death trees with extinction rate = speciation rate (conditioned on n species today)
  - [McPeek (2008) Amer. Natur. 172: E270-284:
    Analysed 245 chordate, arthropod, mollusk, and magnoliophyte trees]



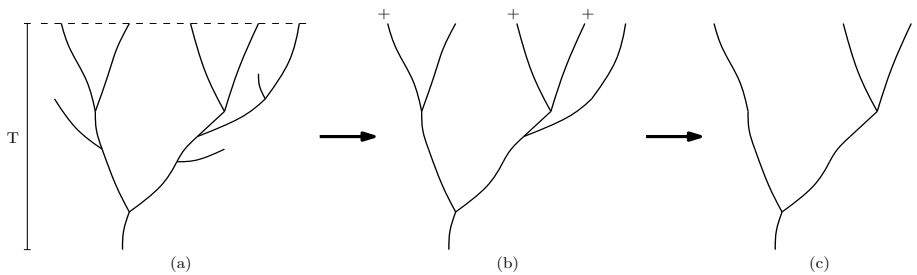
Predicting future phylogenetic diversity loss

Question:

If a random 10% of species from some clade were to disappear in the next 100 years due to current high rates of extinction, how much evolutionary heritage would be lost?

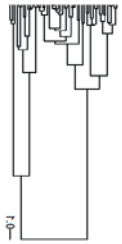
Prediction is very difficult, especially about the future. Niels Bohr, Danish physicist (1885-1962)

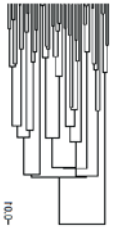
PD (again)

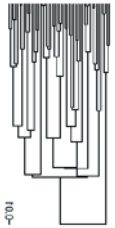


Predict the proportion of diversity that remains if each leaf survives with independently with probability p.

"…80 percent of the underlying tree can survive even when approximately 95 percent of species are lost." Nee and May, ***Science***, 1997







For Yule model, let $\pi_t(p)$ be the expected phylogenetic diversity in a Yule tree, grown for time t, under a 'field of bullets' model with taxon survival probability *p*.

[note 2 random processes]

$$\pi(p) := \lim_{t\to\infty} \frac{\pi_t(p)}{\pi_t(1)} \qquad \mu(p) = \frac{\text{Expected future diversity}}{\text{Expected present diversity}}$$

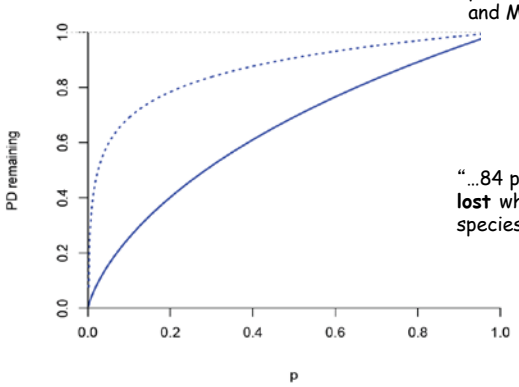
Theorem:

$$\pi(p) = \frac{-p\log(p)}{1-p}$$

$$\pi_t(p) = \frac{2p}{(1-p)\lambda} e^{\lambda t}\left[-\log(p+(1-p)e^{-\lambda t}\right]$$

$$\mu(p) = \frac{-p\log(p)}{1-p}$$

"…80 percent of the underlying tree can survive even when approximately 95 percent of species are lost." Nee and May, *Science*, 1997



"…84 percent of the underlying tree **is lost** when approximately 95 percent of species are lost."

A more recent result (2013):

- Instead of ratio of expected values, what about expected value of 'biodiversity ratio'?

$$\mu(p) = \frac{\text{Expected future diversity}}{\text{Expected present diversity}}$$

$$E\left[\frac{\text{future diversity}}{\text{present diversity}}\right]$$

- What about actual distribution of the biodiversity ratio? And at finite times?

$$\frac{\text{future diversity}}{\text{present diversity}}$$

- What about more general speciation-extinction models?

Theorem [birth rate = $\lambda(t)$, extinction rate = $\mu(t,a)$]

As the number n of species in a random tree of height T grows, the biodiversity ratio converges converges almost surely to a constant $\pi_T(p)$.

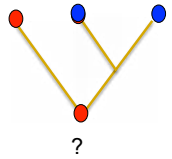
$$\pi_T(p) = p\frac{\int_0^T \frac{1-F_T(t)}{1-(1-p)F_t(t)}dt}{\int_0^T (1-F_T(t))dt}$$

$$\sqrt{np}\left(\frac{\text{future diversity}}{\text{present diversity}} - \pi_T(p)\right) \xrightarrow{D} N(0,\sigma^2)$$

Specialist topic: Ancestral state reconstruction

Minimum evolution ('parsimony'):

Need tree topology but not branch lengths or model



Majority Rule

Don't even need tree

Maximum likelihood

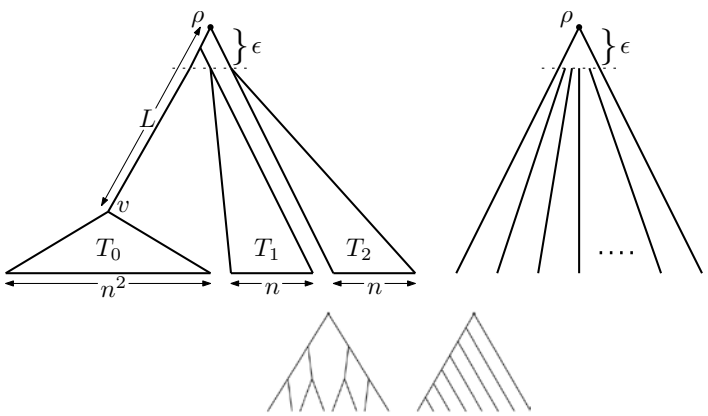
Need tree, branch lengths and model

Definition:

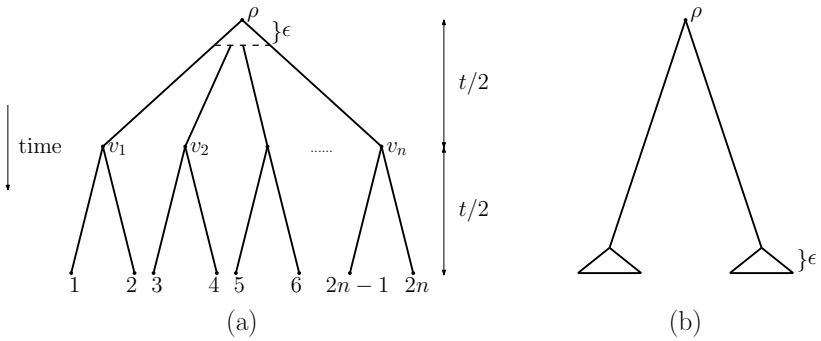
For a method M that estimates the ancestral state at a node v of a tree from leaf data, and a model of character state change, the *Accuracy* of M at v is:

Pr(M(leaf data)= state of v]

Which is more accurate for root state prediction from an 'evolved' character: parsimony or majority?



Q2. Is it easier to estimate the ancestral state at the root of the tree, or an interior node?



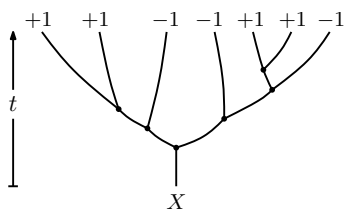
Root state can be estimated with **high** precision but **no** other node can be

Root state can be estimated with **low** precision but **all** other interior nodes can be

What happens on a 'typical' tree?

Grow a Yule (pure-birth) tree at speciation rate λ for time t

Evolve a binary state from the root to the tips binary character (mutation rate m)



Estimate the root state from the tip states using maximum parsimony.

Let P_t = probability our estimate is correct $P_t = S_t + \frac{1}{2}E_t$

Question: what happens to P_t as t becomes large?

Dynamical system

$$\frac{dS_t}{dt} = -(\lambda + m)S_t + mD_t + \lambda(S_t^2 + 2S_tE_t);$$

$$\frac{dD_t}{dt} = -(\lambda + m)D_t + mS_t + \lambda(D_t^2 + 2D_tE_t);$$

$$\frac{dE_t}{dt} = -\lambda E_t + \lambda(E_t^2 + 2S_DD_t);$$

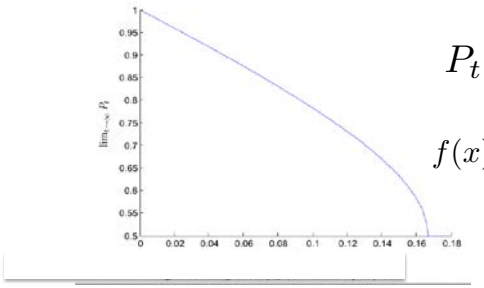
m = mutation rate (of states),
λ = birth rate (of tree)

$$P_t = S_t + \frac{1}{2}E_t$$

'six is (just) enough':

If $\frac{\text{speciation rate}}{\text{mutation rate}} < 6$, then we lose *all* information about the ancestral state as t grows (min evolution).

If $\frac{\text{speciation rate}}{\text{mutation rate}} > 6$, then we don't

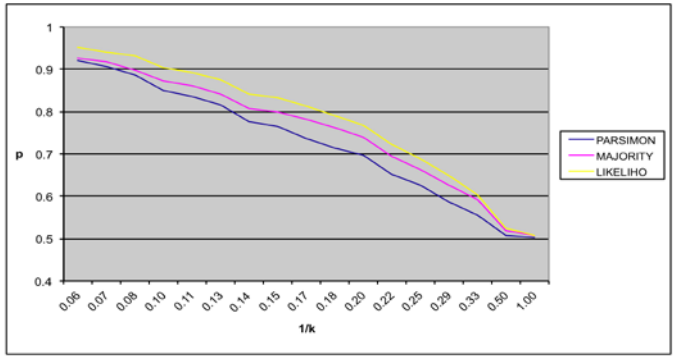


$$P_t \geq \lim_{t\to\infty} P_t = f(x) \text{ where}$$

$$f(x) = \frac{1}{2}\left(1 + \sqrt{(1-6x)(1-2x)}\right)$$

x = mutation rate/speciation rate

Comparisons (simulations)



cf. Hanson-Smith, V., Kolaczkowski, B. and Thornton, J.W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. Mol. Biol. Evol. 27: 1988–99.

Specialist topic: Ancestral state reconstruction

What about majority rule?

If $\frac{\text{speciation rate}}{\text{mutation rate}} < 4$, then ***any*** method loses *all* information about the ancestral state as t grows (we'll see why in 10 mins!).

Theorem [Mossel +S, 2014]

$$\Pr(\text{MR correct}) > \frac{1}{2} + \frac{1}{2}\left(1 - \frac{4m}{\lambda}\right)$$

THE END

Specialist topic 2: Modelling lateral gene transfer (LGT)

PNAS

Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution

Tal Dagan* and William Martin

- *In prokaryotes, if nearly all genes have been transferred between lineages many times is it meaningless to talk about a species 'tree'?*

Biology Direct

BioMed Central

Review

Open Access

Prokaryotic evolution and the tree of life are two different things

Eric Bapteste*[1], Maureen A O'Malley[2], Robert G Beiko[3], Marc Ereshefsky[4], J Peter Gogarten[5], Laura Franklin-Hall[6], François-Joseph Lapointe[7], John Dupré[2], Tal Dagan[8], Yan Boucher[9] and William Martin[8]

Opinion

The tree of one percent

Tal Dagan and William Martin

Address: Institute of Botany, University of Düsseldorf, D-40225 Düsseldorf, Germany.

Correspondence: Tal Dagan. Email: tal.dagan@uni-duesseldorf.de

Published: 1 November 2006

Genome Biology 2006, 7:118 (doi:10.1186/gb-2006-7-10-118)

PNAS

Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations

Gergely J. Szöllosi[a,b], Bastien Boussau[a,b,c], Sophie S. Abby[d,e], Eric Tannier[a,b,f], and Vincent Daubin[a,b,1]

Lateral gene transfer as a support for the tree of life

Sophie S. Abby[a,b,c,d], Eric Tannier[a,b,e], Manolo Gouy[a,b], and Vincent Daubin[a,b,1]

Question:

Suppose we have some 'species tree' (e.g. the tree of bacterial cell divisions). Under a model of independent random LGT events when can we recover this tree from the associated gene trees.

Possibilities for the LGT rates in the model:

Rate of transfer from x to y is constant

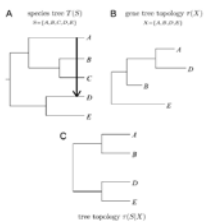
Rate of transfer from x to y depends on the branches

Rate of transfer from x to y depends on d(x,y) and/or time

In all cases, the number of LGT events in the tree has a **Poisson** distribution

A Likelihood Framework to Measure Horizontal Gene Transfer

Simone Linz, Achim Radtke,* and Arndt von Haeseler*†‡§‖*



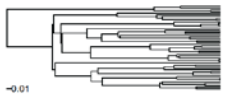
Can we reconstruct a tree under rampant LGT?

Theorem [*c.f.* also Roch and Snir 2013]

Triplet-based (R*) tree reconstruction is a statistically consistent estimator of the species tree under the random LGT model if the expected number G of LGTs per gene is 'not too high'.

Example: for Yule trees with n leaves the following

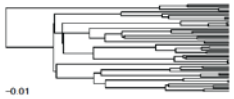
$$G \leq \frac{n-2}{3\ln(n/2)}$$



Particular case: [S,Linz, Huson, Sanderson]

Take n=200 (Yule-shape tree), and suppose each gene is transferred on average 10 times. Then the species tree is identifiable from sufficiently many gene trees.

Can we reconstruct a tree under rampant LGT?



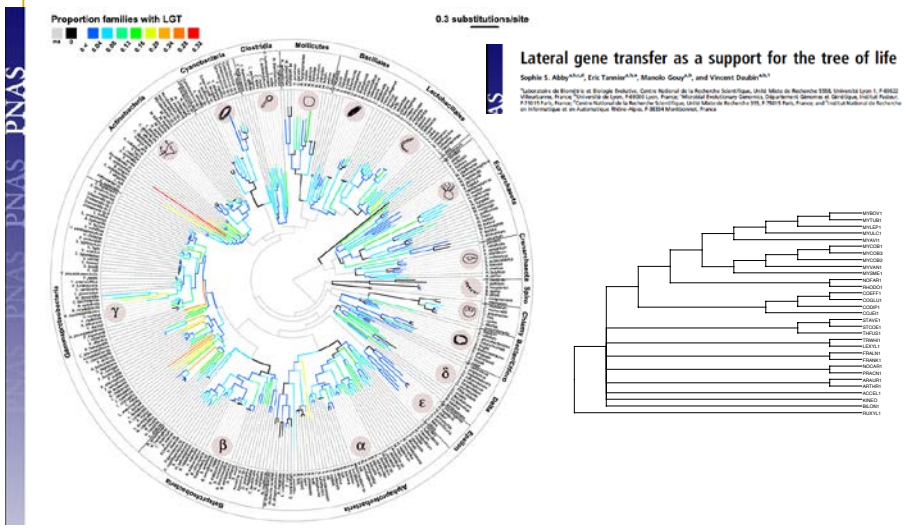
Theorem 1 [Roch and Snir, 2013]

Under the bounded rates (e.g. Yule model), it is possible to reconstruct the topology of a phylogenetic tree for n taxa w.h.p. from $N = \Omega(\log(n))$ gene tree topologies if the expected number of LGT transfers is no more than a constant times $n/\log(n)$.

Theorem 2

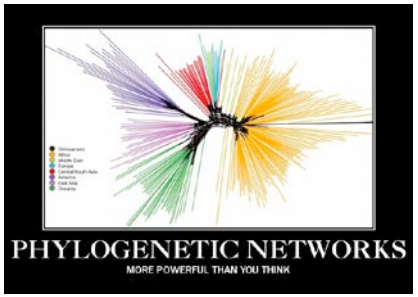
Under the Yule model, it **is not possible** to reconstruct the topology of a phylogenetic tree w.h.p. from N gene trees if the expected number of LGT events is more than $\Omega(n \log(N))$

Roch, S., Snir, S., 2013. Recovering the tree-like trend of evolution despite extensive lateral genetic transfer: a probabilistic analysis. J. Comput. Biol. 20 (2), 93–112.

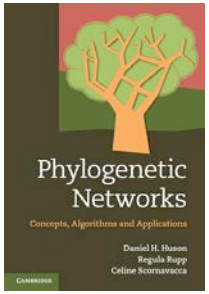


THE END

Lectures 9-10: Phylogenetic Networks I



David Morrison



ALLAN WILSON CENTRE

Mike Steel

The World of PHYLOGENETIC NETWORKS GENEALOGICAL

Biomathematics Research Centre

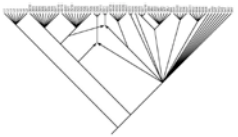
from F. Delsuc and N. Lartillot

Winthrop lectures, 2014

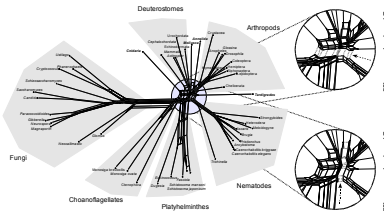
ALLAN WILSON CENTRE

Why networks?

- *Explicit networks:*
  - Species evolution is sometimes reticulate due to:
    - Hybrid species
    - Genetic exchange (eg. Lateral gene transfer)
    - Endosymbiosis
  - Usually represented by rooted networks



- *Implicit networks:*
  - shows conflicting signals in the data (even if evolution is tree-like)
    - SplitsGraphs
    - Neighbor-Net (very widely used)
    - Endosymbiosis
  - Usually represented by unrooted networks



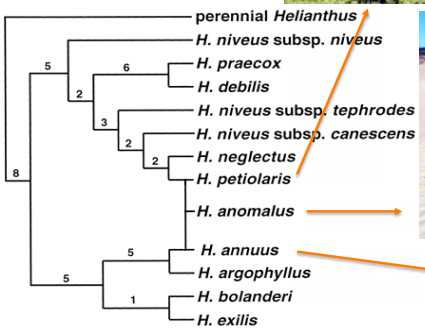
http://phylonetworks.blogspot.co.nz/2012/06/rooted-networks-for-exploratory-data.html

Metazoan phylogeny: From Huson and Bryant (2006). Applications of phylogenetic networks in evolutionary studies, *Mol. Biol. Evol.*

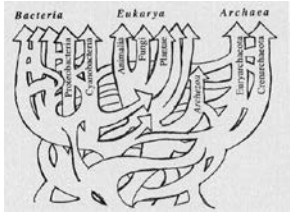
Reticulate evolution

However, sometimes inheritance is from multiple ancestors, because of reticulate events, e.g:

1) **Hybrid speciation**
2) Lateral gene transfer
3) Recombination



Trees or networks?

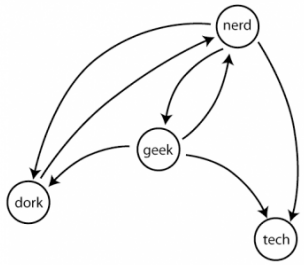


"molecular phylogeneticists will have failed to find the 'true tree' not because their methods are inadequate or because they have chosen the wrong genes, but because the history of life cannot properly be represented as a tree."

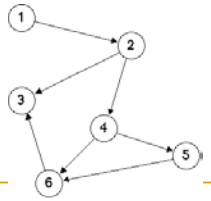
W. F. Doolittle, 1999

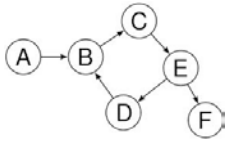
Directed graphs: Basics

In any directed graph $D = (V, A)$
sum of out-degrees =
sum of in-degrees = $|A|$.



Definition: $D = (V,A)$ is *acyclic* if it has no directed cycles ("D.A.G")

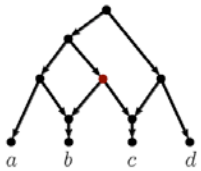




Phylogenetic network:

- A **phylogenetic network on** X is an acyclic network with a single (root) vertex of in-degree 0, X = set of vertices of out-degree 0, and no vertices with in-degree= out-degree=1.
- Unlike phylogenetic trees there are an infinite number of phylogenetic networks on X.
- *Example:* Cluster networks

Three types of network:

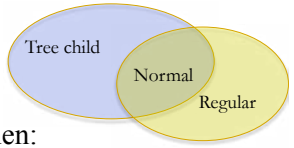


A network that is not tree-child

Tree child: each non-leaf vertex has at least one non-reticulate child ($\Longrightarrow$ Tree path property)

Regular: isomorphic to the Hasse diagram of its clusters

Normal: tree child, no vertices of out-degree 1, no redundant arcs



Theorem 1: Every normal network is regular.

Theorem 2: If $N = (V, A)$ is a **normal** network, then:

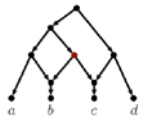
(i) the number r of reticulate vertices is at most $n - 1$

(ii) $|V| \leq (n^2 - n + 1)/2$

c.f. regular

Willson, S.J. 2010. Properties of normal phylogenetic networks, Bulletin of Mathematical Biology 72: 340-358.

Binary phylogenetic networks



- Root has out-degree 2
- A vertex with out-degree 2 has in-degree 1 (and the set of vertices of out-degree 0 is X)
- All other vertices either have in-degree 1 and out-degree 2 or in-degree 2 and out-degree 1 (*reticulate vertices*)

$n = |X|, r = \#$ reticulate vertices, $t = \#$ tree vertices

$$|V| = n + t + r + 1$$
$$t = n + r - 2$$

$$|V| = 2t + 3$$
$$|A| = 3r + 2n - 2$$
$$|A| - |V| + 1 = r$$

Why?

$$r + 2t + 2 = |A| = 2r + t + n$$

OUT IN

Binary phylogenetic networks

Recall (Willson):

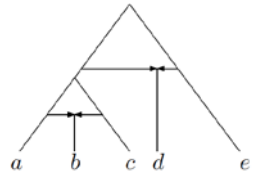
If $N = (V, A)$ is a normal network, then:
(i) the number r of reticulate vertices is at most $n - 1$
(ii) $|V| \leq (n^2 - n + 1)/2$

Theorem:

If $N = (V, A)$ is a **tree-child** binary network, then:
(i) the number r of reticulate vertices is at most $n - 1$
(ii) $|V| \leq 4n - 3$

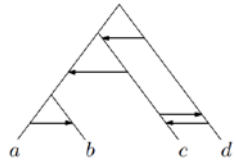
McDiarmid, Semple, Welsh (2014). Phylogenetic networks that display a tree twice. *Bull. Math. Biol.* (in press).

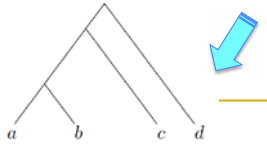
Special classes of [binary] phylogenetic networks



A **reticulation network** is a binary phylogenetic network whose arc set A is the disjoint union of a set of *reticulation arcs*, and a set A_T of *tree arcs*, and such that:

- Each reticulation arc ends at a reticulation vertex;
- Each reticulation vertex has at least one incoming reticulation arc;
- Every interior vertex has at least one outgoing tree arc.



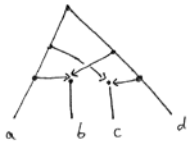


Additional bells and whistles

- A reticulation network is *time-consistent* if there is a 'time-stamp' function

  $t : V(\mathcal{N}) \to \mathbb{R}^{\geq 0}$ such that for each arc (u, v)

  $t(u) = t(v)$ if (u, v) is a reticulation arc and $t(u) < t(v)$ otherwise

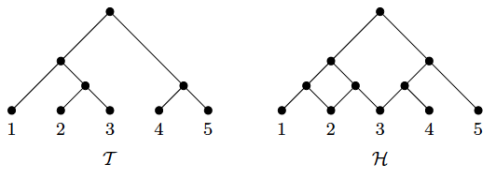


- "Level k" (if N is binary it is level k if k is the maximum number of reticulations in any biconnected component of N)

Mathematical questions about phylogenetic networks

- How many trees do they contain (display)?
- Do these trees allow us to reconstruct the network?
- Given two trees what is the simplest network that contains them?
- What about parsimony?
- How many networks are there?

Tree 'displayed' by a network



Quiz: Is it easy or hard to determine if a given tree is displayed by a given network?

Theorem: [van Iersel et al. 2010]

It is NP-hard, **even for regular networks**.

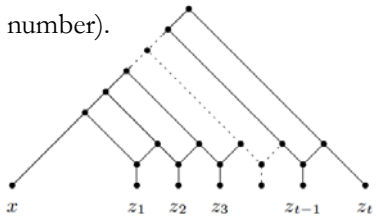
There is is a poly-time algorithm for **tree-child binary networks** and **normal networks** (also level-k networks).

The set of all trees displayed by a network: $Tr(N)$

Observations 

If N has r reticulation vertices, then N displays at most 2^r trees

N can have much fewer than 2^r displayed trees (so one tree is displayed several times). For example, this network displays F_t trees (F = Fibonacci number).



Linz, S., St John, K., and Semple, C. (2013). Counting trees in a phylogenetic network is #P-complete *SIAM Journal on Computing*, 42, 1768-1776.

The set of all trees displayed by a network: $Tr(N)$

Theorem 1*

If N is normal and binary then N displays exactly 2^r trees.

Theorem 2**

Let N be a binary tree-child phylogenetic network on X.
There is an $O(n^2)$ algorithm ($n = |X|$) to decide whether or not N displays a rooted phylogenetic tree with leaf set X twice.

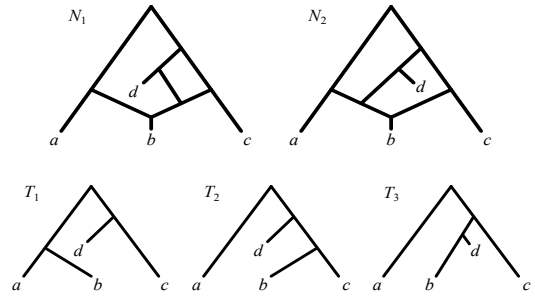
*Special case of Corollary 3.4 of Tree-average distances on certain phylogenetic networks have their weights uniquely determined. Algorithms for Molecular Biology (2012) 7:13

**Phylogenetic networks that display a tree twice Paul Cordue Simone Linz Charles Semple (submitted)

Does $Tr(N)$ determine N?

Not in general? Some networks display the same set of trees

Example:



When does $Tr(N)$=Tr(M) imply N=M?

Theorem [Willson, 2011]

If N is **regular** (or normal) then Tr(N) determines N.

Moreover, there is a poly-time algorithm for reconstructing N from Tr(N).

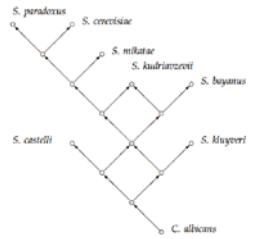
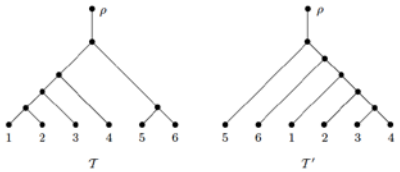


Fig. 5. The normal network which results from utilizing Maximal Child with input the three most common gene trees for the data of [20].

Regular networks are determined by their trees, IEEE/ACM Transactions on Computational Biology and Bioinformatics 8 (2011) 785-796.

Hybridization number of two (binary) trees



$$h((V,A)) = |A| - |V| + 1$$

Given two binary phylogenetic X-trees $\mathcal{T}, \mathcal{T}'$ let:

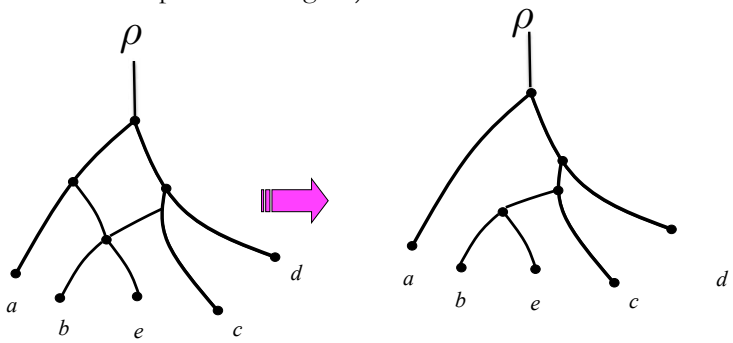
$$h(\mathcal{T}, \mathcal{T}') = \min\{h(\mathcal{N}) : \mathcal{N} \text{ displays } \mathcal{T}, \mathcal{T}'\}$$

$$h(\mathcal{T}, \mathcal{T}') \leq n - 2$$

Quiz: Is computing h(T,T') easy or hard?

Relationship to tree-rearrangement operations

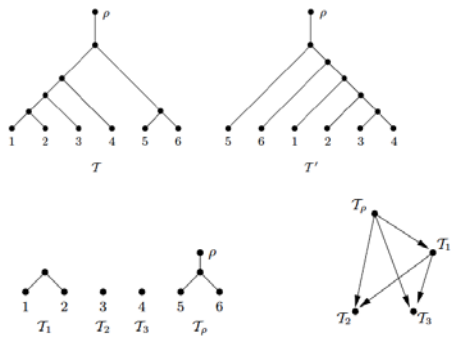
- rSPR (rooted subtree prune and regraft)



$$d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = 1 \Leftrightarrow h(\mathcal{T},\mathcal{T}') = 1$$

How does this generalize?

Hybridization number of two (binary) trees



Maximum *acyclic* agreement forest for *T* and *T'*

Theorem [Baroni, Gruenewald, Moulton, Semple 2005]

- $h(\mathcal{T},\mathcal{T}') = m_a(\mathcal{T},\mathcal{T}')$
- $d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = m(\mathcal{T},\mathcal{T}')$

$h = 4$

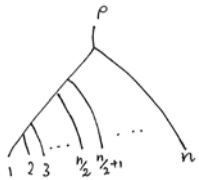
$$d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') \leq h(\mathcal{T},\mathcal{T}')$$

+ [Bordewich and Semple, 2004]

$$h(\mathcal{T},\mathcal{T}') \leq n-2$$

$$d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') \leq h(\mathcal{T},\mathcal{T}')$$

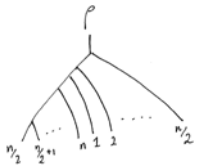
How much less?



Theorem 1

For all $n > 3$, even there exist two binary phylogenetic X-trees with:

$$d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = 2, \text{ and } h(\mathcal{T},\mathcal{T}') = n/2$$



Theorem 2

For all $n > 3$, there exist two binary phylogenetic X-trees with:

$$h(\mathcal{T},\mathcal{T}') - d_{\mathrm{rSPR}}(\mathcal{T},\mathcal{T}') = n - \lceil 2\sqrt{n} \rceil$$

Moreover, this is sharp

Baroni, M., Grunewald, S., Moulton, V., and Semple, C. (2005). Bounding the number of hybridisation events for a consistent evolutionary history. Journal of Mathematical Biology, 51, 171-182.
[Humphries, P.J. and Semple, C.]

Back to our question:

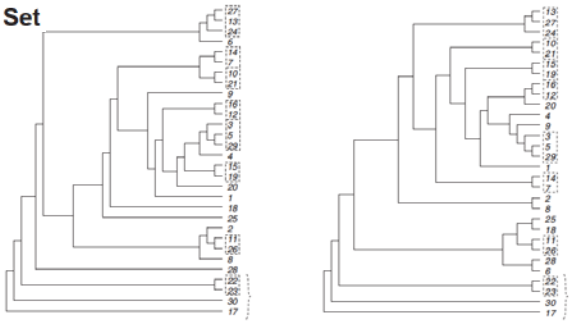
Quiz: Is computing h(T,T') easy or hard?

*NP-hard (but there are algorithms based on max. agreement acyclic forest)

Grass (*Poaceae*) Data Set

n=30

h=8
time =19s



*Bordewich, M. and Semple, C. 2007. Computing the minimum number of hybridisation events for a consistent evolutionary history. *Discrete Applied Mathematics*, 155:914–928.

Counting networks

Recall (lecture 1!):

$$rb(n) \sim \frac{1}{\sqrt{2}}\left(\frac{2}{e}\right)^n n^{n-1} = 2^{n\log_2 n + O(n)}$$

Theorem [Mcdiarmid, Semple, Welsh 2014]

The number of tree-child (or normal) binary networks on n leaves is

$$2^{2n\log_2 n + O(n)}$$

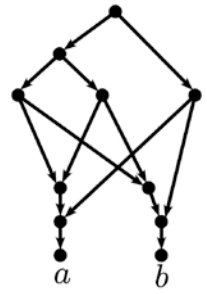
Almost all tree child (or normal) networks with n leaves have $(1+o(1))n$ reticulate vertices and $(4+o(1))n$ vertices in total.

McDiarmid, C., Semple, C. and Welsh, D. (2014). Counting phylogenetic networks. *Annals of Combinatorics* (in press).

Is a network just something you get by adding edges to a phylogenetic tree?

Yes – for tree child networks

No – for some others – e.g. at right (not tree-sibling)



Challenges questions: (phylogenetic networks)

From
Leo van Iersel
and Steven Kelk

ALLAN WILSON CENTRE

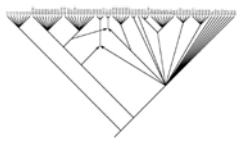
Problem 1 Is the Hybridization Number problem fixed-parameter tractable (FPT)?

Problem 2 Does there exist a polynomial-time 2-approximation algorithm for MAF on two binary trees?

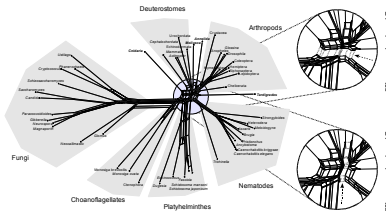
Problem 3 Is there an FPT algorithm for finding a level-k phylogenetic network consistent with a given dense set of rooted triplets, if k is the parameter?

Why networks?

- *Explicit networks:*
  - Species evolution is sometimes reticulate due to:
    - Hybrid species
    - Genetic exchange (eg. Lateral gene transfer)
    - Endosymbiosis
  - Usually represented by rooted networks



- *Implicit networks:*
  - shows conflicting signals in the data (even if evolution is tree-like)
    - SplitsGraphs
    - Neighbor-Net (very widely used)
    - Endosymbiosis
  - Usually represented by unrooted networks



http://phylonetworks.blogspot.co.nz/2012/06/rooted-networks-for-exploratory-data.html

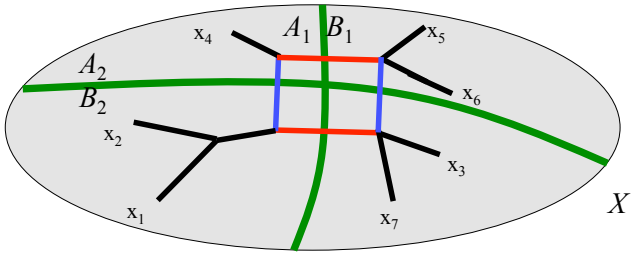
Metazoan phylogeny: From Huson and Bryant (2006). Applications of phylogenetic networks in evolutionary studies, *Mol. Biol. Evol.*

Implicit networks

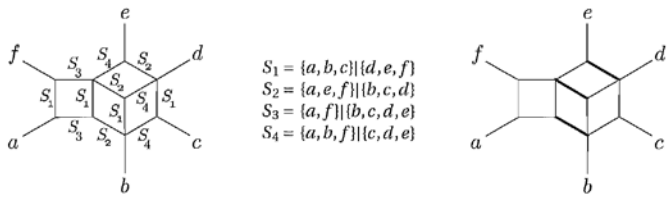
- Two splits $A_1 | B_1$ and $A_2 | B_2$ of X are *compatible*, if one of the following intersections is empty:

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

Two incompatible splits:

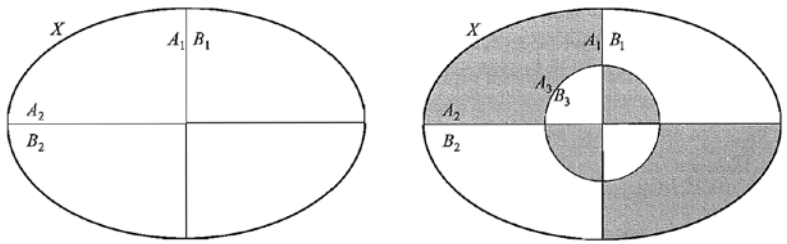


Split Network



(a) Split network N (b) Non-trivial splits (c) Split $S_2 = \frac{\{a,e,f\}}{\{b,c,d\}}$

Weakly compatible splits



The 3 splits are **weakly compatible** if at least one of the white regions and at least one of the grey regions is empty

Weakly compatible: Example and properties

- If Σ is weakly compatible then Σ has size $O(n^2)$.
- Σ is weakly compatible iff $Q(\Sigma)$ has at most two of the three possible resolutions of each quartet
- Connection to 'weak hierarchies':

$$A \cap B \cap C \in \{A \cap B, A \cap C, B \cap C\} \forall A, B, C \in \mathcal{W}$$

Split Decomposition [Bandelt and Dress]

- Notice that a tree metric d can be written as

$$d = \sum_{\sigma \in \Sigma(T)} w_\sigma d_\sigma$$

$$d_\sigma(x, y) = 1 \text{ iff } \sigma \text{ separates } x \text{ and } y$$
$$\text{else } 0$$

- Moreover, if $|X| = 4$ then for any d

$$d = \sum_{\sigma \in W} c_\sigma d_\sigma$$

Split Decomposition [Bandelt and Dress]

- **Theorem:** [Bandelt and Dress ~late 1980s]

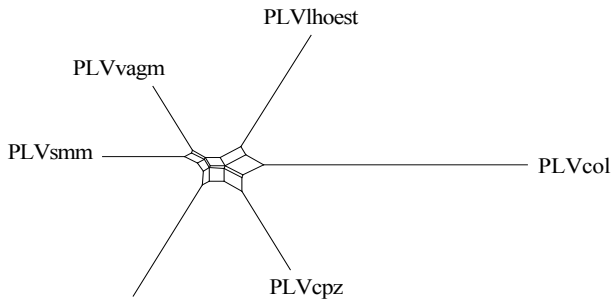
Every distance function on a set X has a unique representation of the form:

$$d = \sum_{\sigma \in W} c_\sigma d_\sigma + \delta$$

where W is a weakly compatible set of X-splits
$c_\sigma > 0$ for all $\sigma \in W$ and δ is 'split prime'

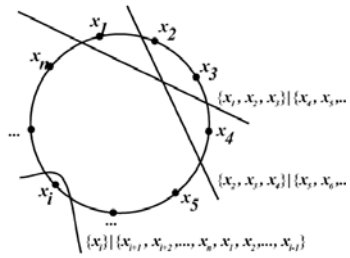
Example

- Split network for primate lentiviruses from whole-genome-based distances using split decomposition:



(Salemi et al, 2003)

Circular split system



Definition:

Σ is circular if there is a circular ordering of X so that each split in Σ is of the form $\{x_p, x_{p+1}, \ldots, x_q\} | X - \{x_p, x_{p+1}, \ldots, x_q\}$

How hard is it to determine if Σ is circular?

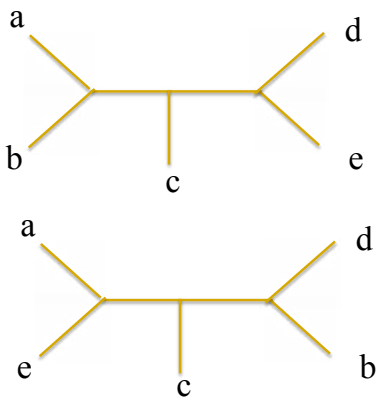
Circular split system implies weakly compatible (but not conversely!)

- **Example:**

$T_1, T_2 \in U(X).$

$$\Sigma = \Sigma(T_1) \cup \Sigma(T_2)$$

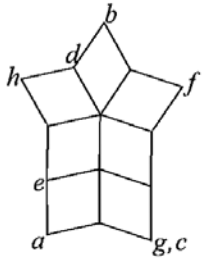
- Σ is always weakly compatible
- But not necessarily cyclic!

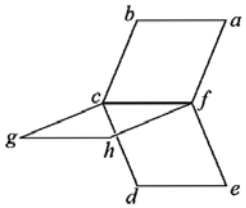


"Outer-labeled planar" networks

- Example

$\{a,b,d,e,h\} \mid \{c,f,g\}$
$\{a,c,d,e,g,h\} \mid \{b,f\}$
$\{a,c,e,g\} \mid \{b,d,f,h\}$
$\{a,c,g\} \mid \{b,d,e,f,h\}$
$\{a,c,e,f,g\} \mid \{b,d,h\}$
$\{a,e,h\} \mid \{b,c,d,f,g\}$



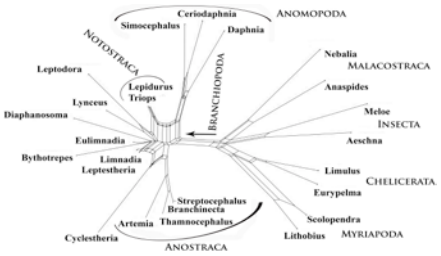


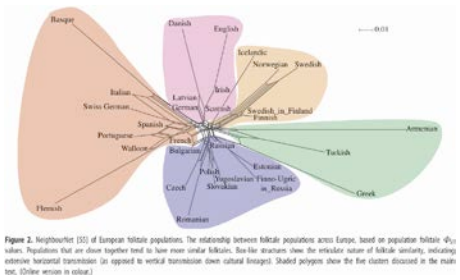
$\{a,b\} \mid \{c,d,e,f,g,h\}$
$\{a,b,c,d,e,f\} \mid \{g,h\}$
$\{a,b,c,f,g,h\} \mid \{d,e\}$
$\{a,e,f,h\} \mid \{b,c,d,g\}$

Cyclic split systems correspond to outer-labelled planar networks

- **Theorem** A set of splits on X is cyclic if and only if it can be represented by an outer-labeled planar network

"NeighborNet"





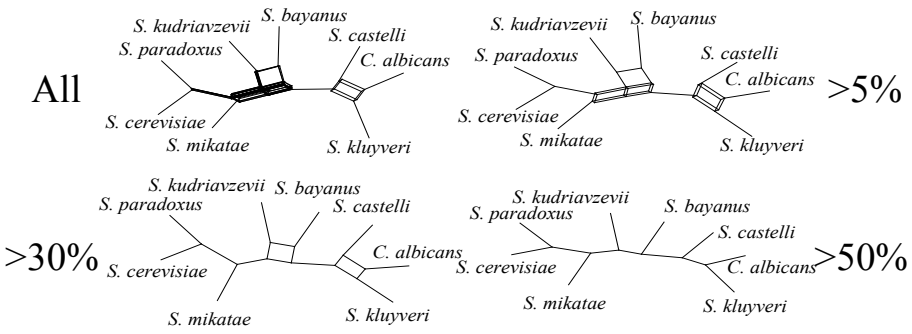
(Waegele and Meyer, 2007). 18S rRNA

Split Networks from Trees

- Consensus splits (Holland et al, 2004)
  - Input: Trees on identical taxon sets
  - Determine splits in more than X% of trees
  - For >50%, result is compatible
- Consensus super splits
  (Huson et al, 2004, Whitfield et al 2008)
  - Input: Trees on *overlapping* taxon sets
  - Use Z-closure to complete partial splits
  - Use "distortion filter" to implement consensus methods

Split Networks from Trees

- Split network for consensus splits on 106 gene trees for yeast:

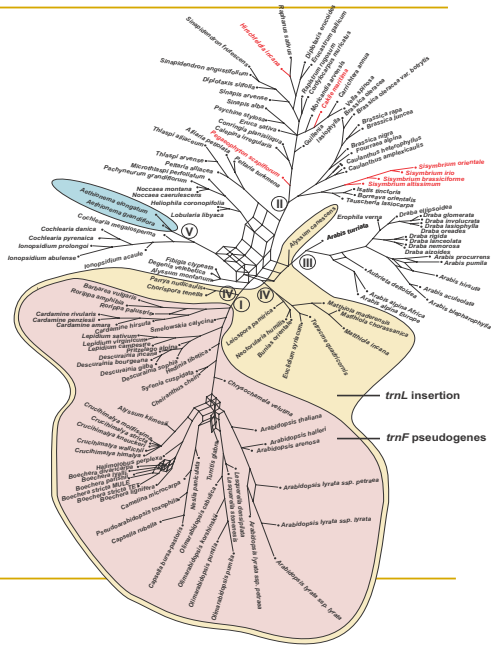


[Rokas et al, 2003, Holland et al, 2004]

Split Networks from Trees

Example:

- Super split network obtained from 5 genes on a total of 71 plant taxa



[Koch et al, MBE 2007]

Useful online resources

Online resources:

The Genealogical World of Phylogenetic Networks
phyloseminar.org
phylobabble
PhyloWiki