Probabilistic models of evolutionary trees

- Joint work with…




Olivier Gascuel [LIRMM, Montpellier]

Arne Mooers [UBC, Vancouver]

Thomas Li [ANU, Canberra]

Tanja Stadler [ETH, Zurich]

Biomathematics Research Centre

ALLAN WILSON CENTRE

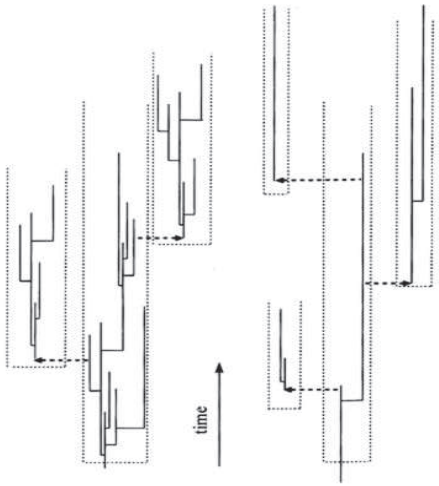
MECB. June 20, 2012

Outline of talk

- ***Part 1***: History, overview
- ***Part 2:*** Discrete models of tree shape
- ***Part 3:*** Continuous trees
- *Part 4:* Applications: phylogenetic diversity, ancestral reconstruction

Yule model

Number of species in genus	Number of genera Observed	Calculated
1	131	130.9
2	35	47.2
3	28	25.2
4	17	16.0
5	16	11.2
6	9	8.3
7	8	6.5
8	8	5.2
9 to 11	13	11.1
12 to 14	3	7.2
15 to 20	7	8.8
21 to 34	14	9.2
35 upwards	4	6.2
Total	293	293.0



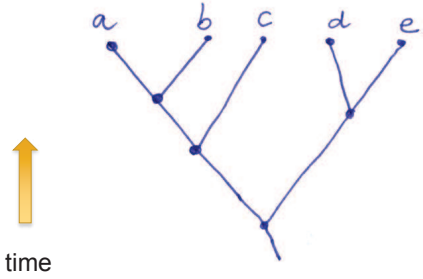
From 'Branching processes in biology' Kimmel and Axelrod

$$\Pr(N=n)=\frac{\mu\Gamma(1+\lambda/\mu)\Gamma(n)}{\lambda\Gamma(n+1+\mu/\lambda)}\approx 1/n^{1+\mu/\lambda}$$

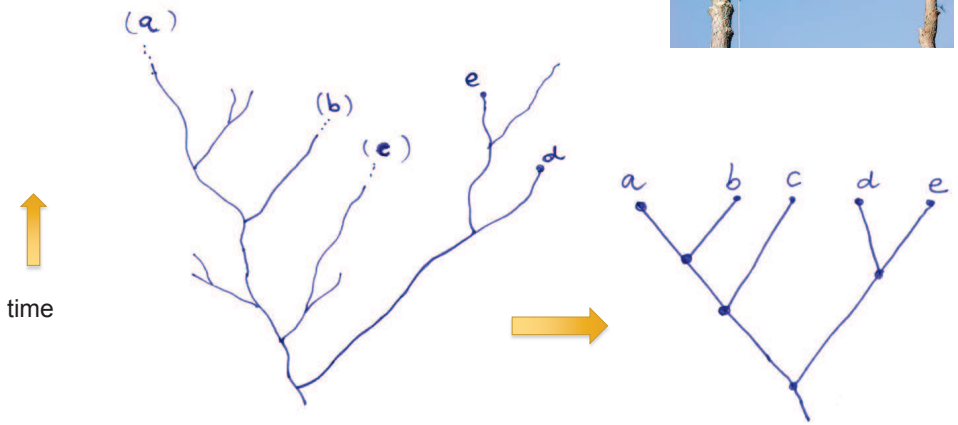
[1]G. U. Yule, A mathematical theory of evolution. Based on the Conclusions of Dr. J.C. Willis, F.R.S. Phil. Trans. Roy. Soc. 213 (1925), 21-87.

Where do evolutionary trees comes from?

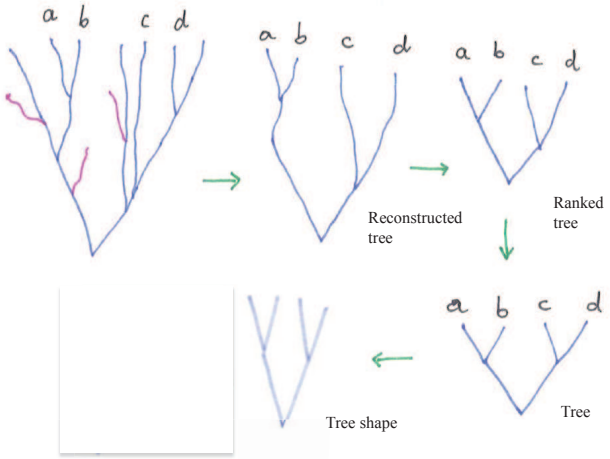
Forestry Unit men tree-felling in Southern Italy



Another viewpoint



The basic picture….



Tree shape: why of interest?

- Speciation/extinction processes make statistical predictions (e.g. about tree 'shape', species distributions etc).
  - So data can be used to test hypotheses about these processes
- Models are used as priors in Bayesian phylogenetics
- Models allow us to estimate, predict quantities of interest (probabilities, expectations, amount of data required etc)

Part 2: Discrete models of tree shape

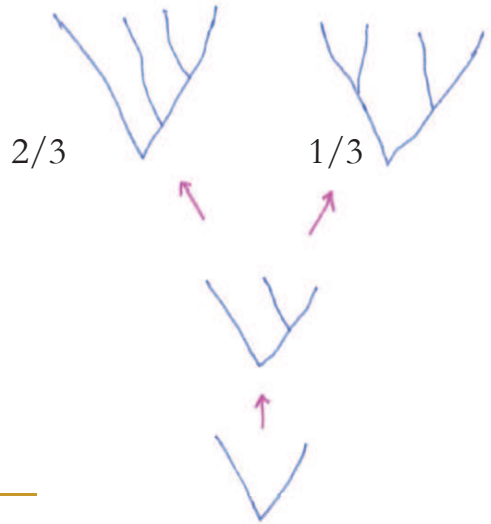
$$\Pr_X(T = t)$$

$$\Pr_X(t), \quad |X| = n$$

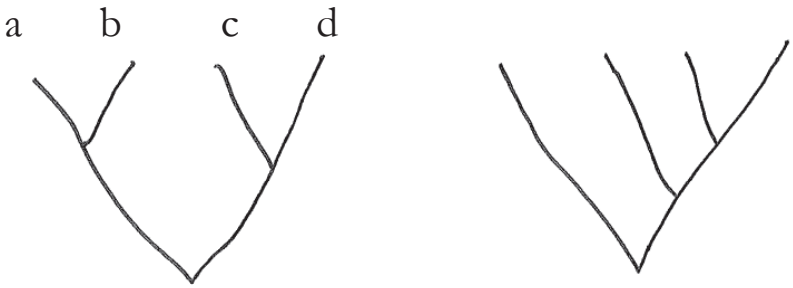
Exchangeability property (EP)

If σ is a permutation of the leaves then $\Pr_X(t^\sigma) = \Pr_X(t)$

The simplest discrete model (Yule-Harding)



Example



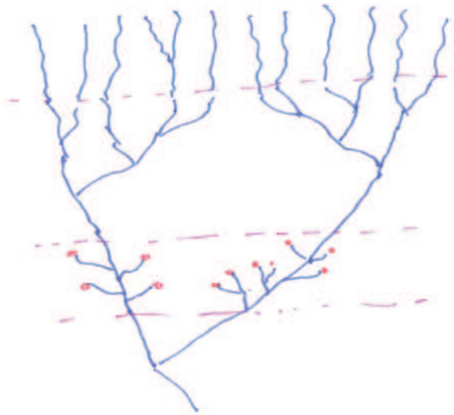
Tree shape probability	1/3	2/3
Tree probability	1/9	1/18

A general process….

[David Aldous, 1995]

- "From time to time there is an "event" which is either an extinction or a speciation, i.e., either some species *B* becomes extinct or some species *A* splits into two species *A* and *A'*.
- The time *t* until the next event, and the chance the next event is an extinction rather than a speciation, may depend on the past in an **arbitrary** way.
- But if the next event is an extinction then each species is equally likely to be the one to become extinct, and if the next event is a speciation then each species is equally likely to be the one to speciate."

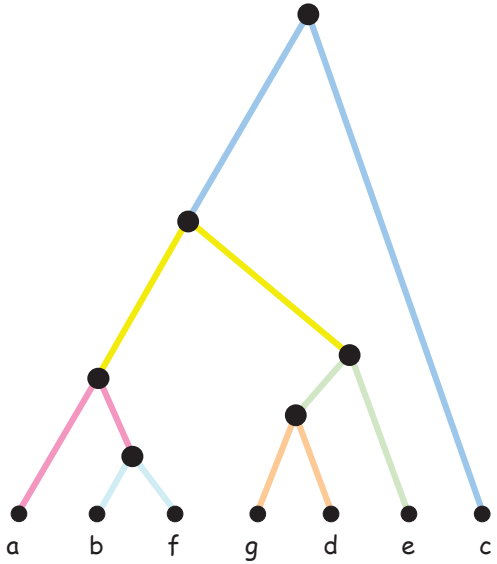
Less can be more….



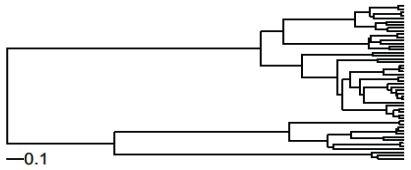
Lemma: All such models lead to the Yule-Harding distribution on discrete trees

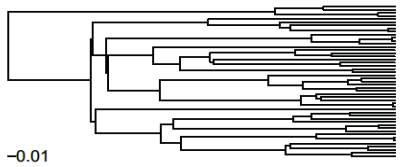
Another further connection…

J.F.C. Kingman



Connection of YH to coalescent?





- **Lemma:**
  - The Yule-Harding and Kingman coalescent lead to identical distributions on discrete trees

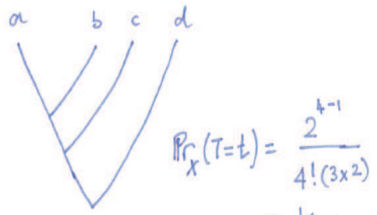
- The ‘YHK’ model

This connection helps!

$$\#\ labelled\ histories = \binom{n}{2} \times \binom{n-1}{2} \times \cdots \times \binom{2}{2}$$

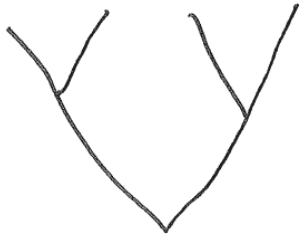
$$\#\ labelled\ rankings\ of\ t = \frac{(n-1)!}{\prod_v (n_v - 1)} \qquad \Pr_X(T = t) = \frac{2^{n-1}}{n! \prod_v (n_v - 1)}$$

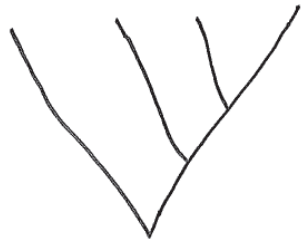
Example



$\Pr_X(T=t) = \frac{2^{4-1}}{4!(3\times 2)} = 1/8$

Uniform on ranked trees is different from uniform on trees (PDA model)





Yule:	1/3	2/3
PDA:	1/5	4/5

PDA – relevant?

- A model?
  - 'Window' of speciation
  - Others
- Random data with maximum parsimony
  - (*cf* YHK~= quartet puzzling [Vinh et al. 2011])

Other discrete models

- Aldous β-splitting

$$-2 \le \beta \le \infty$$

$\beta = -\frac{3}{2}$ PDA

$\beta = 0$ Yule

- Ford α-model

Real 'trees'

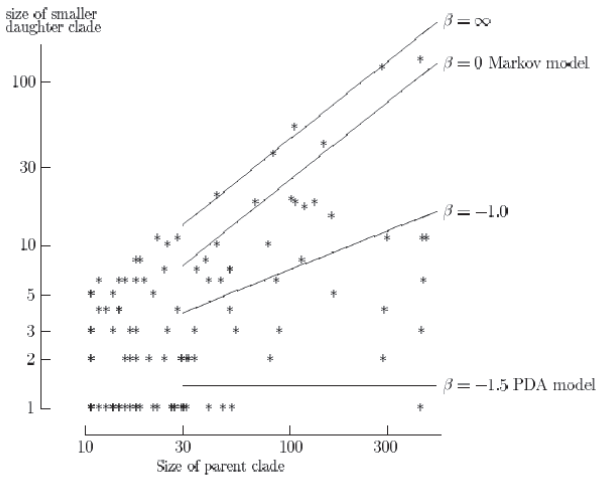
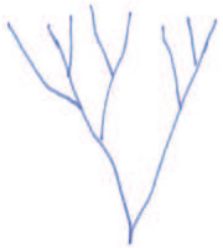
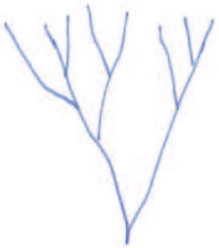


FIG. 3. *Splits in the tree of Chase et al (1993), and approximate median lines for the beta-splitting model. Note the log-log scale.*



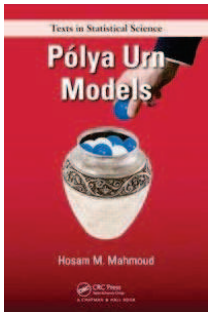
From: Aldous, D. (2001). Stochastic Models and Descriptive Statistics for Phylogenetic Trees, from Yule to Today. *Statistical Science* 16: 23-34

Balance of tree



- Select one of two subtrees incident the root
- Let K = number of leaves in it.
- Under YHK model K is uniform

$$\Pr_X(K = k) = \frac{1}{n-1},\ k = 1, 2, \ldots, n-1$$

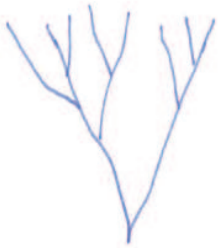


Quiz: Select your favorite taxon x

Generate a YHK tree.

Let K_x =# leaves in the subtree containing x.

Is K_x uniform?

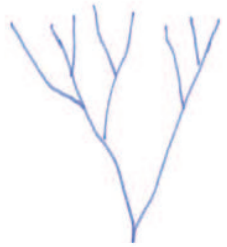


$$\Pr(K_x = k) = \Pr(\#S = k \mid x \in S)$$

$$= \frac{\Pr(x \in S \mid \#S = k) \times \Pr(\#S = k)}{\Pr(x \in S)}$$

$$= \frac{k/n \times 1/n-1}{1/2} = \frac{2k}{n(n-1)}$$

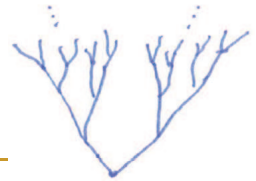
Comparison of K between YHK and PDA



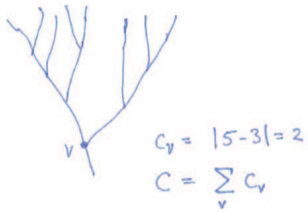
- Select one of two subtrees incident the root
- Let K = number of leaves in it.

- Under PDA $\Pr(Y = k) = \binom{n}{k} \frac{R(k)R(n-k)}{R(n)} \sim \frac{1}{2}, \frac{1}{8}, \frac{1}{16}, \frac{5}{64}, \ldots$

Example: '$Y = n/2$' likelihood ratio $\propto \sqrt{n}$



Measures of balance/depth



- Colless index
- Distance of random leaf to root (or other leaf)
- Sackin index

1. Probability A is a clade

- "Clade"

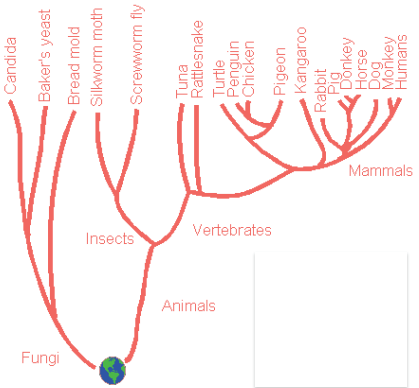
$cl(T)$ = set of clades of T

- YHK (Rosenberg, 2003)

$$\Pr_X(A \in cl(T)) = \frac{2n}{a(a+1)} \binom{n}{a}^{-1}$$

- PDA

$$\Pr_X(A \in cl(T)) = \frac{R(a)R(n-a+1)}{R(n)}$$



Extension to k clades (Zhu, Degnan and Steel 2011).

2. How close is the MRCA of set A of k taxa to the root of the tree?

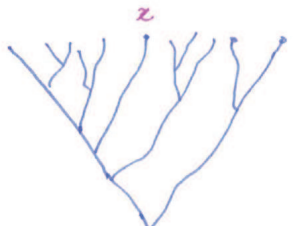
- In YHK need to just sample k=7 taxa to have 50% chance (regardless of n) the MRCA=root.

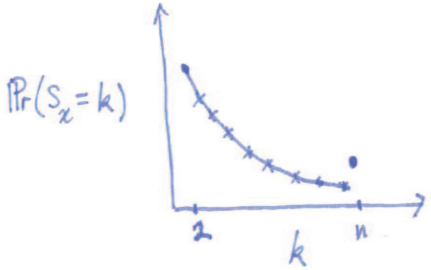
- For PDA you need to have $k > 0.17n$ taxa
  - For YHK, the number of edges from MRCA to root has an (asymptotically) geometric distribution

3. Size of the minimal clade containing x

- [Blum and Francois 2005]

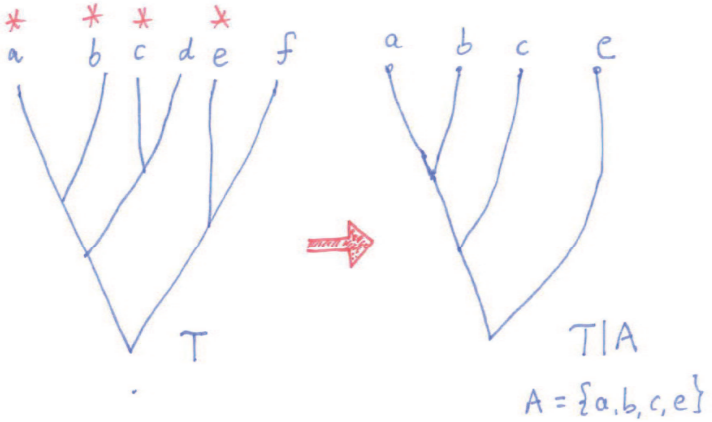
$$\Pr(S_x = k) = \begin{cases} \dfrac{4}{k(k^2-1)}, k = 2,\ldots,n-1 \\ \dfrac{2}{n(n-1)}, k = n \end{cases}$$





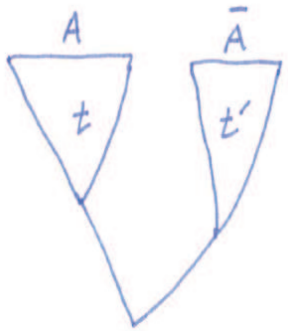
Distribution depends on n only through last term.
Monotone except for last term

Recall induced subtree



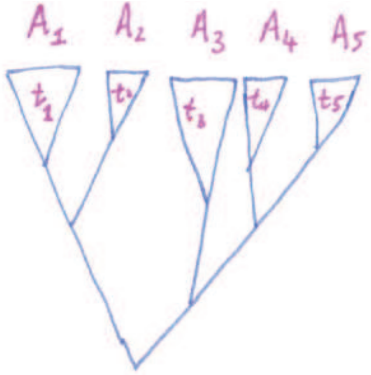
Properties of models

- Markov property (MP)



$$\Pr_X(T \mid A = t \,\&\, T \mid \bar{A} = t' \mid A, \bar{A} \in cl(T)) = \Pr_A(t) \cdot \Pr_{\bar{A}}(t')$$

Properties of models



- (MP) $\Rightarrow$ extended Markov property
  - [application: Sampling YKH trees from an unresolved tree (Bayesian)]
- Marginal Markov property

$$\Pr_X(T \mid A = t \mid A \in cl(T)) = \Pr_A(t)$$

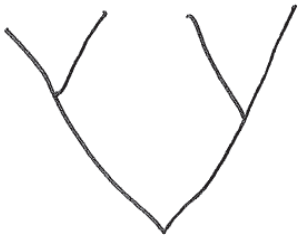
Sampling consistency (SC):

$$For\ any\ \ A \subset X,\ \ \Pr_X(T \mid A = t) = \Pr_A(t)$$

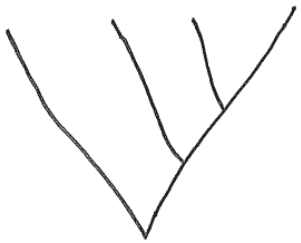
$$\sum_{t' : t' \mid A = t} \Pr_X(t')$$

- "$\Pr_A(t)$ doesn't depend on species you haven't yet sampled."
  - Not implies by the Markov Property
  - Satisfied by YHK, PDA, Comb and *some* values of the beta-splitting model.

Example of a distribution violating (SC):



70%



30%

Why?

$$\left(\frac{3}{5}\right)p_1 + \left(\frac{1}{5}\right)p_2 + \left(\frac{0}{5}\right)p_3 = 0.7$$

Jim Pitman et al (SC+Exchangeability)

Properties: Group elimination

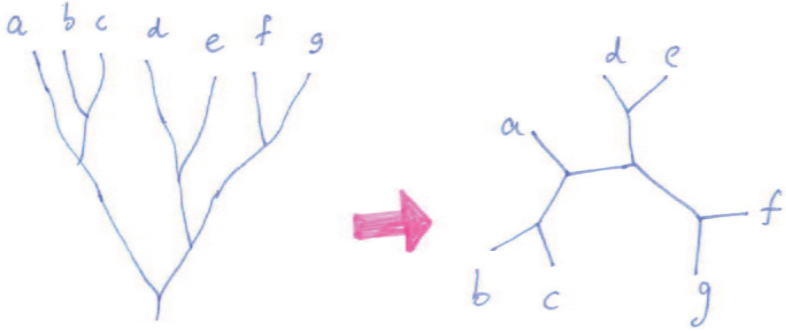
- **If A forms a clade, then the rest of the tree is described by the model**

$$\Pr_X(T \mid \bar{A} = t \mid A \in cl(T)) = \Pr_{\bar{A}}(t)$$

- Satisfied by Yule, PDA, Comb

- *Conjecture* [D. Aldous, 1995]

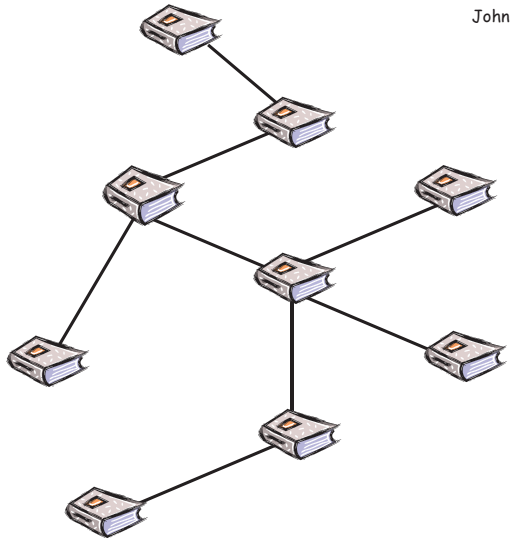
These three are the ONLY distributions on discrete tree (shapes) satisfying GE



Where did it start?

Question: If a tree had 1000 leaves would we have any idea where the root was?

John Haigh 1970

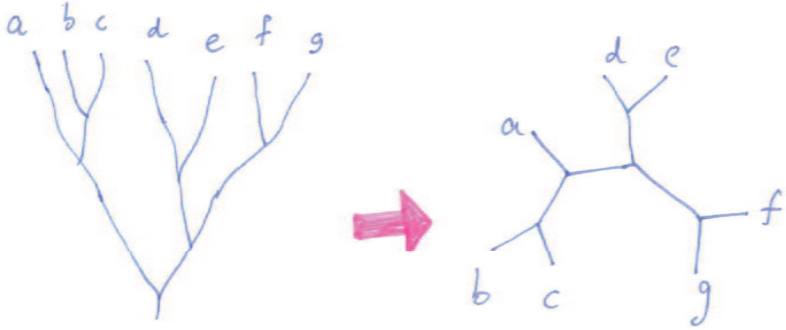


$$1 - \log(2) \sim .307$$

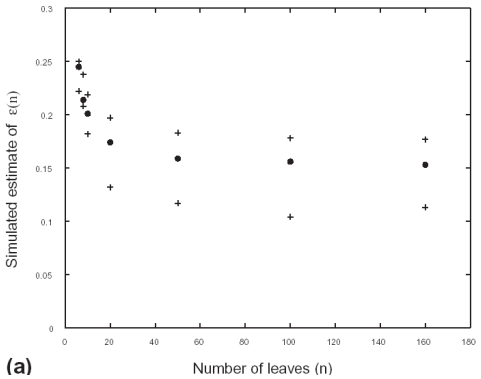
9 vertices: p >99.6%

Probability MLE point is 1,2,3,4th >99.8%

Independent of n



Where did it start?



Theorem (McKenzie+S, 2000)

$$\Pr(e_{ML} = e_0) = 4\log(4/3) - 1 \sim 0.15$$

P(longer of longer < shorter)

A further property (root invariance)

- **"Any rooting of the tree is equally likely"**

Formally: If t' is obtained from t by re-rooting the tree then:

$$\Pr_X(T = t) = \Pr_X(T = t')$$

- Several distributions satisfy (RI) and (EP).
- Several satisfy (EP) and (SC).

- But only one satisfies (RI) and (SC)!

Result:

- **Theorem [S-2012]:**

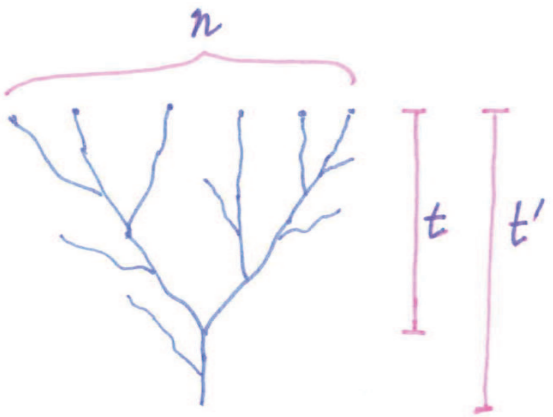
A probability distribution P on rooted phylogenetic trees satisfies (RI) and (SC) if and only if

P is the PDA distribution.

- **Corollary:**

Any non-PDA probability distribution on rooted phylogenetic trees that is sampling consistent must prefer some rooting (of an unrooted tree) over others.

Part 3: Continuous trees



- Conditioning on n and t
- Conditioning just on t
- Conditioning just on n

Pure-birth process

- λ = speciation rate

$$n_t \sim geo(e^{-\lambda t})$$

$$E[n_t] = (2)e^{\lambda t}$$
$$Var[n_t] \approx (2)e^{2\lambda t}$$
$$\lambda_{ML} = \frac{\ln(n/(2))}{t}$$

Birth-death process

- λ = speciation rate
- μ = extinction rate

$$\Pr(n_t = k \mid n_t > 0) \sim geo(p)$$

$$E[n_t] = e^{(\lambda-\mu)t}$$
$$E[n_t \mid n_t > 0] \to \infty$$

What values to take for μ,λ?

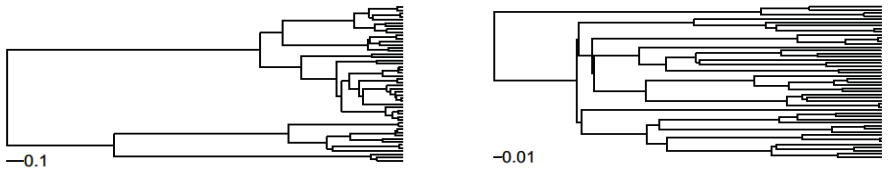
- "Current plant and animal diversity preserves at most 1-2% of the species that have existed over the past 600 my". [Erwim, PNAS 2008].
- Set extinction rate = speciation rate?
- **Problem:** If extinction rate =speciation rate the tree is guaranteed to eventually die out eventually!
- **Solution?**: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

Conditioned critical process (Popovic-Aldous)

Conditioned critical process (Popovic-Aldous)

Set $\lambda=\mu$
Condition on n
Uniform (improper) prior for origin (0, infinity)

Theorem (Stadler):
This leads to expected branch length distribution of the Coalesecent



- **Real** reconstructed trees generally look more like Yule trees with zero extinction rate than birth-death trees with extinction rate = speciation rate (but conditioned on n species today)

[Eg. McPeek (2008) Amer. Natur. 172: E270-284:
Analysed 245 chordate, arthropod, mollusk, and magnoliophyte phylogenies]

Gamma statistic for Yule vs Coalescent trees

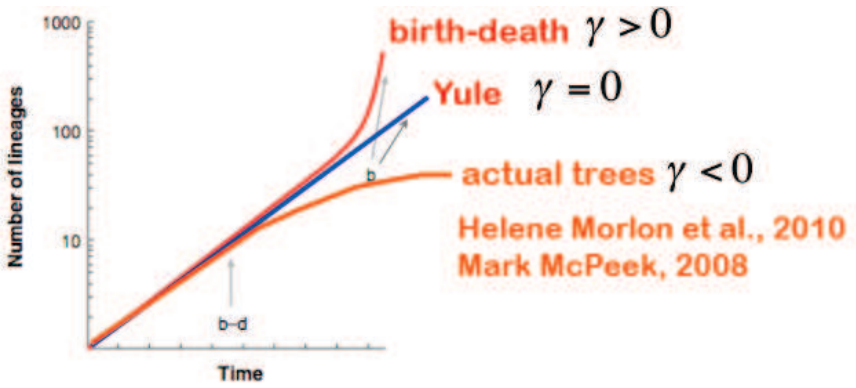
$$\gamma = \frac{\left(\frac{1}{n-2}\sum_{i=2}^{n-1}\left(\sum_{k=2}^{i} k g_k\right)\right) - \frac{TL}{2}}{TL\sqrt{\frac{1}{12(n-2)}}}$$

Oliver Pybus & P. Harvey, 2000

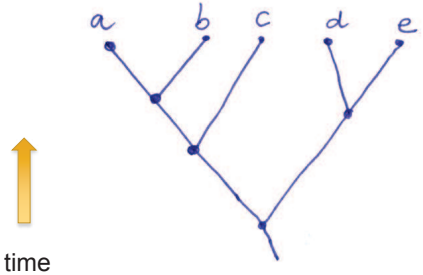
g_k are internode distances

- For Yule pure-birth model $E[\gamma] = 0$
- For Coalescent (or Popovic-Aldous) $\frac{\gamma}{\sqrt{n}} \to \sqrt{3}$

$$E[\gamma] \approx \sqrt{3n}$$

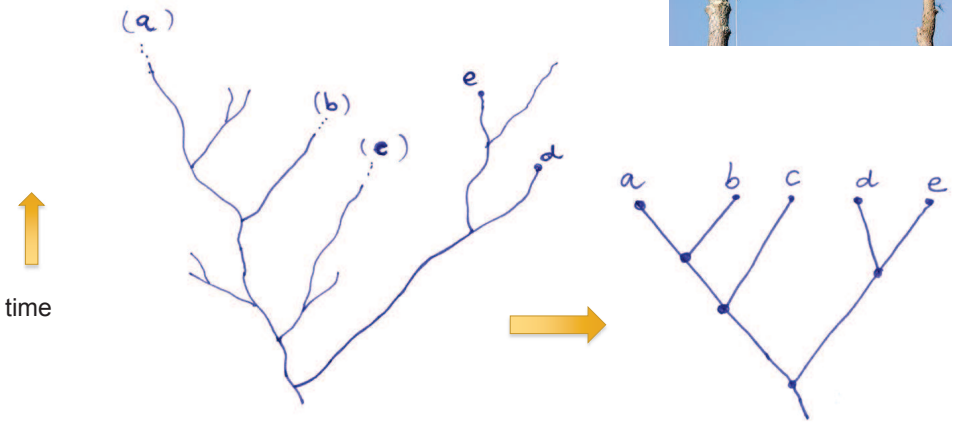


Where do evolutionary trees comes from?



Forestry Unit men tree-felling in Southern Italy

Another viewpoint



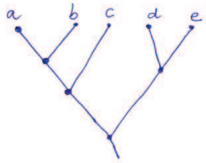
The bus 'paradox'

"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?

The tree puzzle (I):



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

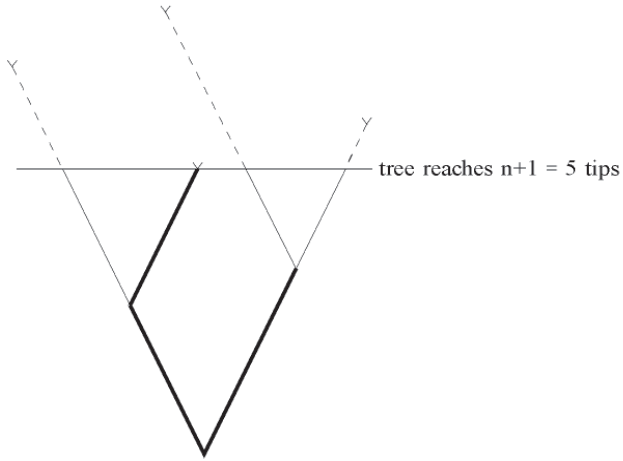
Look at the tree when it has 100 species

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

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:

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

Distribution?

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

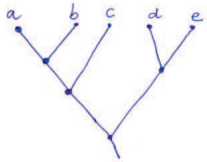
$p(t)$:= expected length of the average pendant edge
$i(t)$: = expected length of the average interior edge

Theorem:

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

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

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? ✓

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.

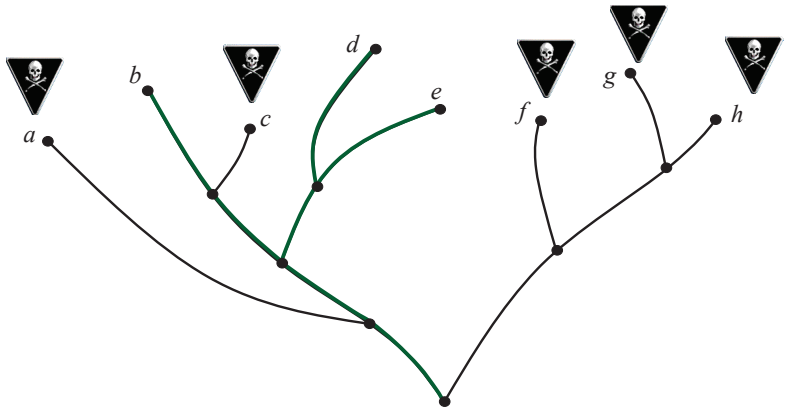
What is the expected length of a randomly selected *extant* lineage?

Answer 1: 1 million years? ✓

Answer 2: 500,000 years? ✗

Relevance?

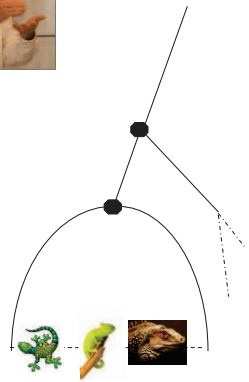
Part 4: Applications:
Application 1: predicting the possible loss of 'evolutionary heritage'



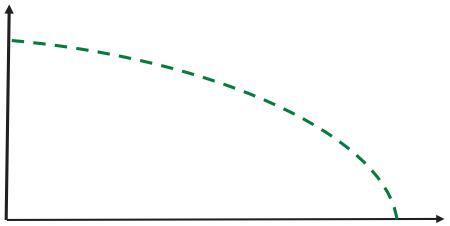
'Field of bullets' models


"…80 percent of the underlying tree can survive even when approximately 95 percent of species are lost."

Nee and May, ***Science***, 1997



Expected 'evolutionary heritage'



No. species that go extinct

However….

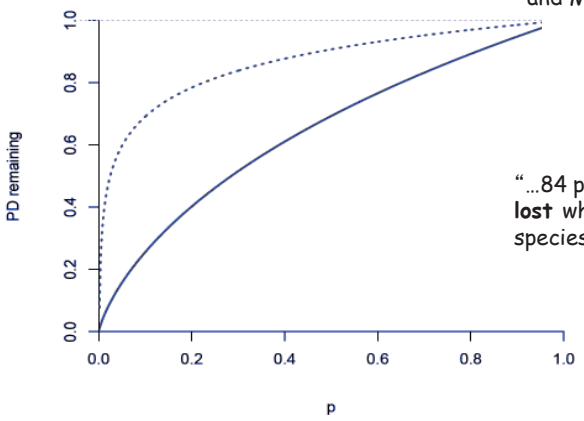
Nee and May's trees are modeled by Coalescent trees.

For Yule model, let $\mu(p)$ = proportion of evolutionary heritage we expect to be preserved in a Yule tree under 'field of bullets' with survival probability p

Theorem:

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

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

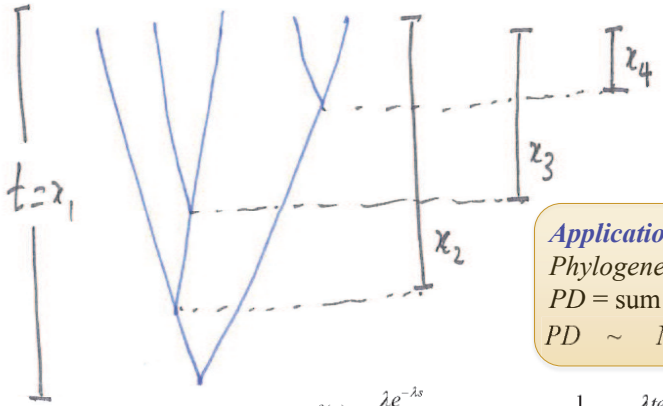


"…80 percent of the underlying tree can survive even when approximately 95 percent of species are lost." Nee and May, *Science*, 1997

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

But that top curve is not 0.8 at p=0.05!

Usefulness of the point process for reconstructed birth-trees (conditioned on n and t)

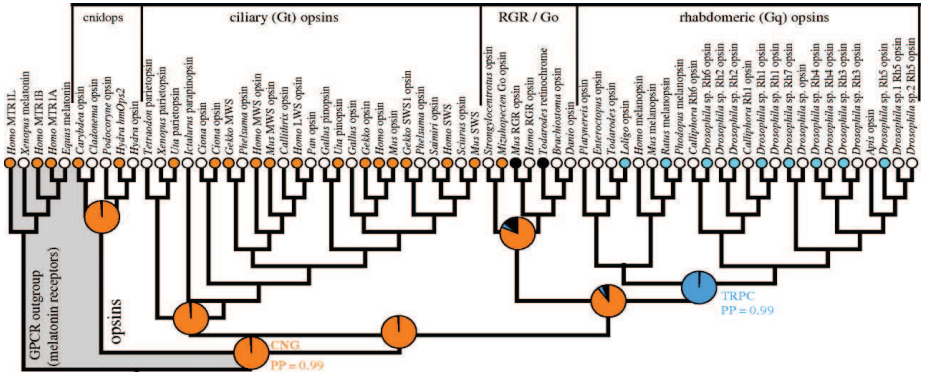


$(x_2,\ldots, x_n)$ are the order statistic or n-2 i.i.d. random variables

Application:
Phylogenetic diversity
PD = sum of branch lengths
$PD \sim N(2t+(n-2)\mu, (n-2)\sigma^2)$

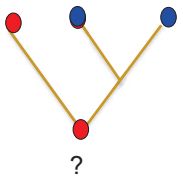
For pure birth, each has density: $f(s)=\frac{\lambda e^{-\lambda s}}{1-e^{-\lambda t}}$ and so $\mu=\frac{1}{\lambda}(1-\frac{\lambda t e^{-\lambda t}}{1-e^{-\lambda t}})$

Application 2: Ancestral state reconstruction



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

Minimum evolution ('parsimony'):



Grow a Yule tree for time t, and evolve binary character on it Let
m = rate of mutation between the two states

Note: we have TWO random processes here.

Estimate root state using minimum evolution.
Let P_t = probability our estimate is correct.

Question: what happens to P_t as t becomes large?

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

Minimum evolution ('parsimony'):

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

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

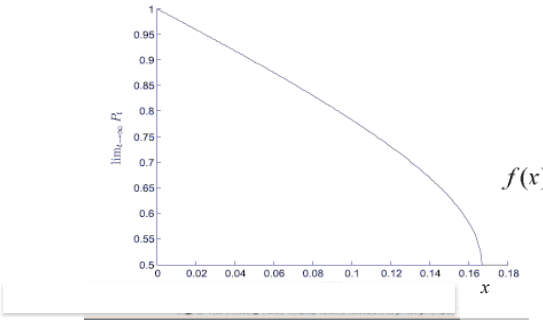
$$\frac{dE_t}{dt} = -\lambda E_t + \lambda(E_t^2 + 2S_DD_t);$$

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

The 'six is (just) enough' theorem:

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

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

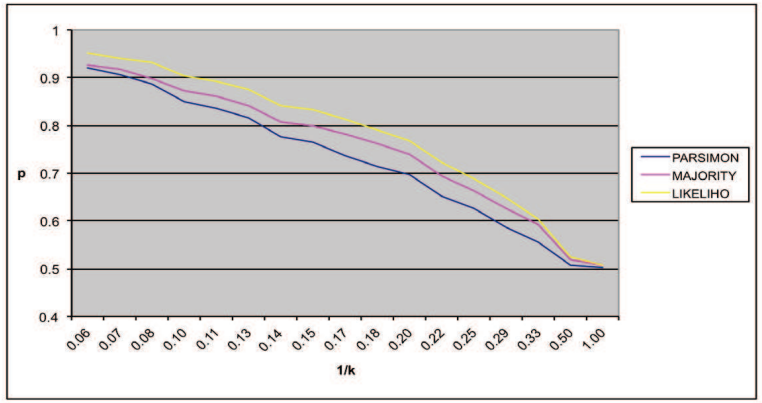


$$f(x) = \frac{1}{2}\left(1 + \sqrt{(1-6x)(1-2x)}\right)$$

Other methods

Majority Rule

Maximum likelihood



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.

Can we do better than six?

If $\frac{\text{speciation rate}}{\text{mutation rate}} < 4$, then we lose *all* information about the ancestral state as t grows **for any method**

If $\frac{\text{mutation rate}}{\text{speciation rate}}$ is between 4 and 6??

x

That's all folks!

Thanks to:

- Royal Society of New Zealand,
- Allan Wilson Centre for Molecular Ecology and Evolution