1

Phylogenetic Tree Shape

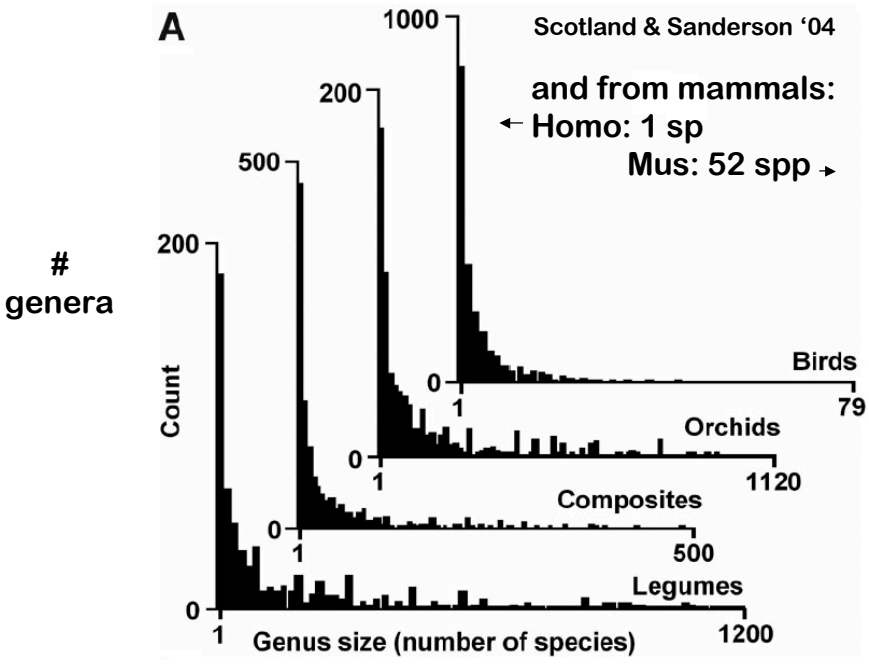
Arne Mooers
Simon Fraser University
Vancouver, Canada

Discussions with generous colleagues & students:

Mark Pagel, Sean Nee (Tree Statistics)
Mike Steel (Tree shapes)
Steve Heard, Andy Purvis (both)
Rutger Vos (Redundancy)
Dave Redding (Equal splits)
Aki Mimoto (Shapley Value)

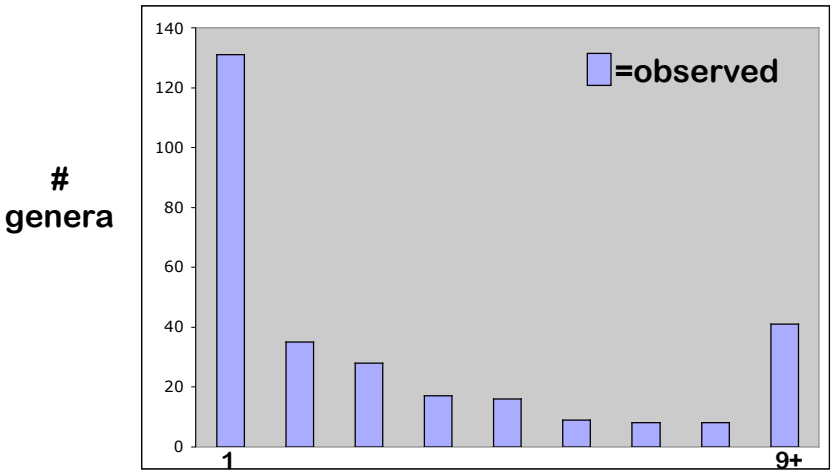
MEP2005, June 20, 2005

2



3

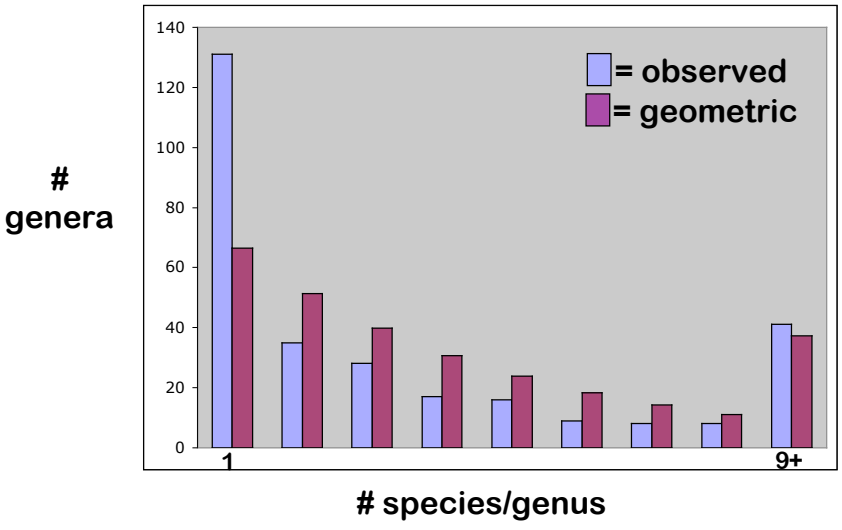
Yule's 1924 data (from Willis) for snakes



Yule, Aldous

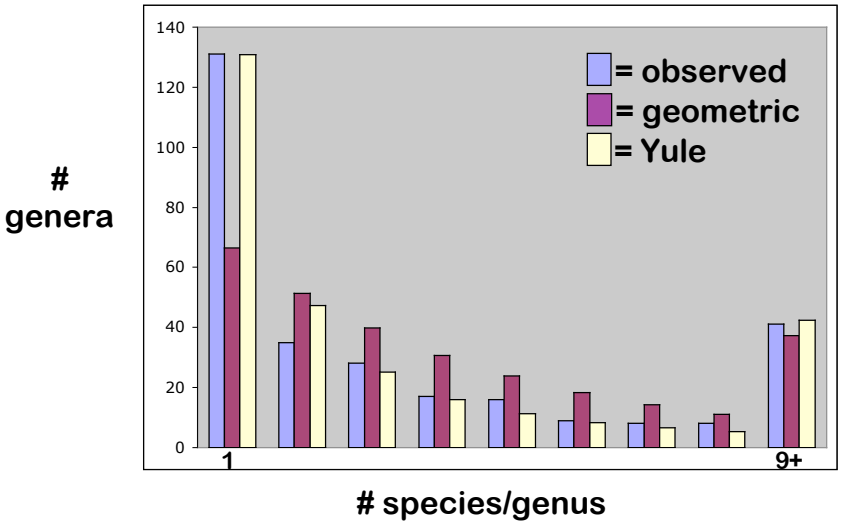
4

Yule's 1924 data for snakes



5

Yule's 1924 data for snakes



6

Yule's 1924 model had two parts:

one: Each genus starts with 1 spp, and each species can give birth to new species with some instantaneous probability λ

This leads to a geometric distribution of genus sizes with mean size $e^{\lambda t}$

ie.

$$P(N(\lambda,t)=n)=e^{-\lambda t}(1-e^{-\lambda t})^{n-1}$$

7

Yule's 1924 model had two parts:

two: Each extant genus gave rise to new genera at some instantaneous rate μ

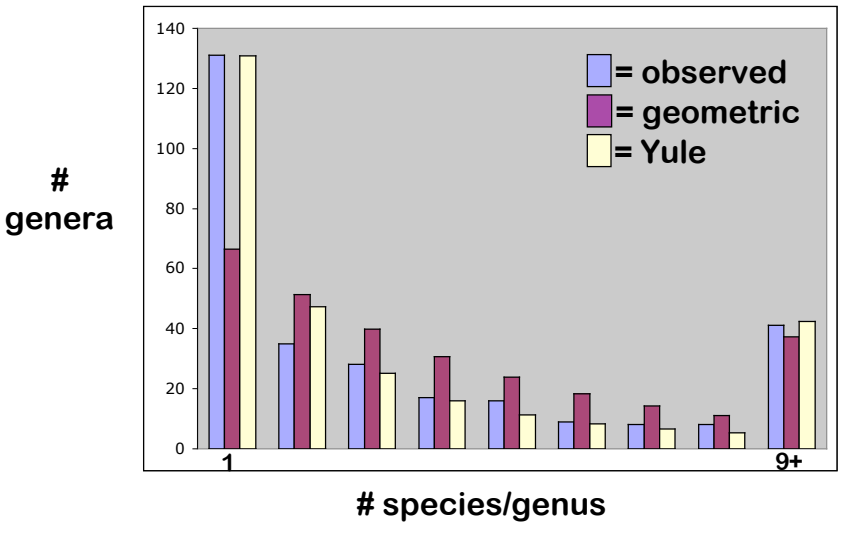
The genera ages (*t*) were not the same, though have same expected age

$$E(t) = \frac{1}{\mu}$$

$$P(N(\lambda,t) = n) = e^{-\lambda t}(1 - e^{-\lambda t})^{n-1}$$

8

Extra variation in genus ages allows for the good fit



Such a compound model gives rise to a 'power law' pattern - or straight lines on log-log plots of frequency(X) on X

Frequency [log(#genera)]

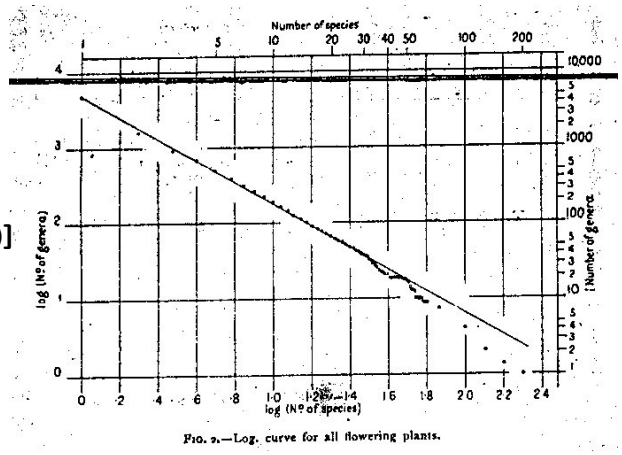


FIG. 2.—Log. curve for all flowering plants.

log(#spp/genus)

Willis & Yule `22

But we know that species (and higher taxa) both grow on trees...

11

But what sort of tree do species grow on?

"Yule tree" or "Equal Rates Markov" (ERM) tree uses part one of the 1924 Yule model:

parent lineages give rise to daughter lineages at some rate (probability) λ, and then daughter and parent lineages are instantly equivalent.

simple & intuitive. All labelled histories are equiprobable (c.f. Prof. Felsenstein)

12

What does a Yule tree look like?

--There are two "dimensions":

1. Expected topology
2. Expected waiting times between splits

13

Yule Topology

Given all possible histories are equally likely,
the probability of a split of size (*k, n-k*) at a node is just

$$\frac{2}{n-1},\ k \neq n-k$$
$$\frac{1}{n-1},\ k = n-k$$

Farris, `76
Slowinski& Guyer, `89

14

Under Yule

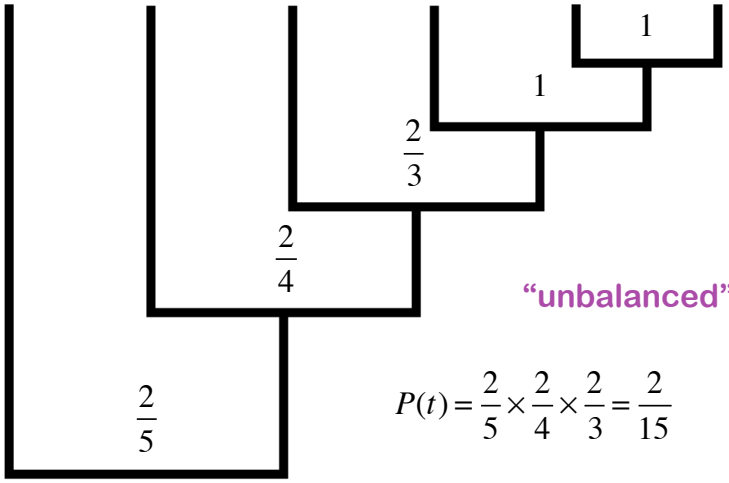
median size of the smaller clade is $\frac{1}{4}n$

Aldous `01

15

Expected topologies under Yule: $\frac{2}{n-1},\ k \neq n-k$

$\frac{1}{n-1},\ k = n-k$



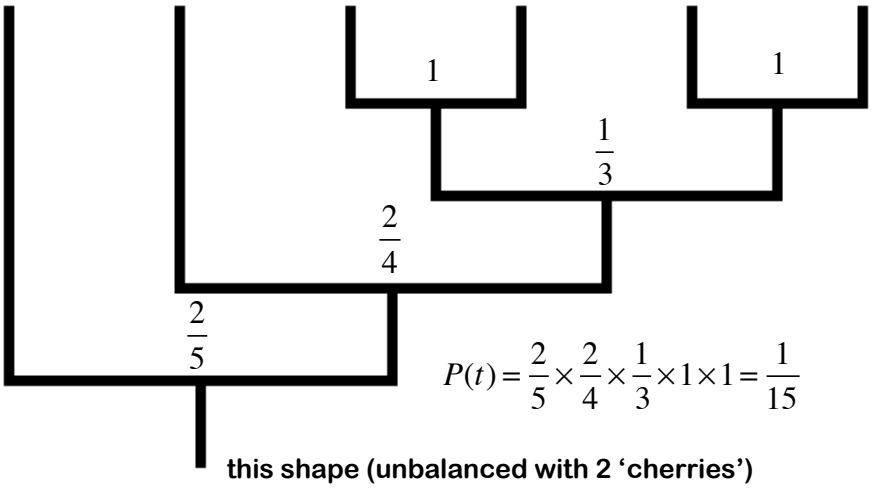
"unbalanced" tree

$$P(t) = \frac{2}{5} \times \frac{2}{4} \times \frac{2}{3} = \frac{2}{15}$$

16

Expected topologies under Yule: $\frac{2}{n-1},\ k \neq n-k$

$\frac{1}{n-1},\ k = n-k$



$$P(t) = \frac{2}{5} \times \frac{2}{4} \times \frac{1}{3} \times 1 \times 1 = \frac{1}{15}$$

this shape (unbalanced with 2 'cherries') is always least likely under Yule (Stone & Repka `98)

But we want to know - how probable is my treeshape relative to an average tree shape?

$$P(\text{average treeshape}) = \frac{1}{\#\text{treeshapes}}$$

Is there is a non-recursive equation for the number of treeshapes [*W(n)*] for n taxa under the Yule model?

$$W(n) = \frac{\sum_{i=1}^{n-1} T_i[T_{(n-i)}]}{2} + E_n$$

T_i = the number of shapes for tree of size i

E_n=0 (n odd), $\frac{W(n/2)}{2}$ (n even), W(1)=1

(Wedderburn `22 in Stone & Repka `98)

The series is:

n	shapes
1	1
2	1
3	1
4	2
5	3
6	6
7	11
8	23
9	46
10	98
11	207
12	451

Stone & Repka `98

19

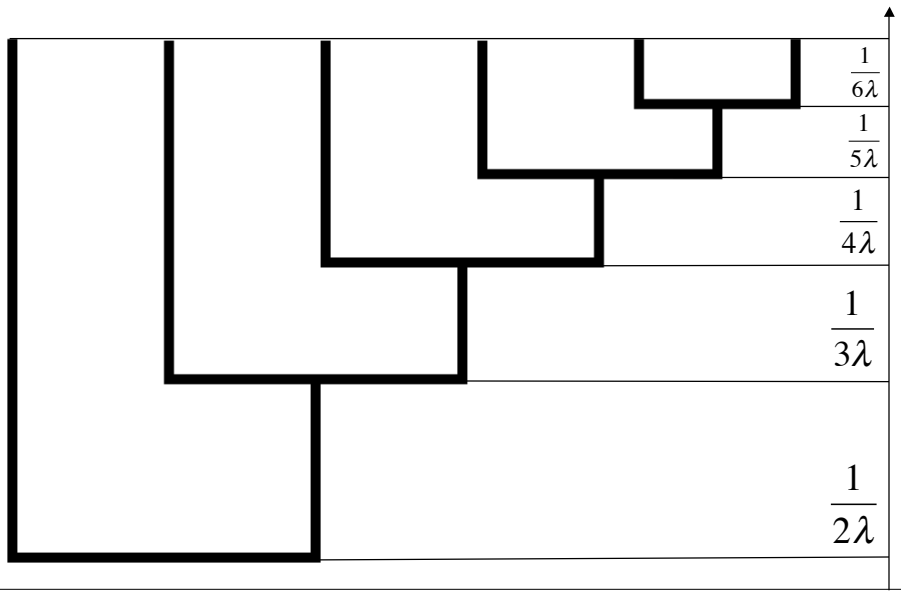
What does a Yule tree look like?

--There are two "dimensions":

1. Expected topology
2. Expected waiting times between splits

20

Yule expected waiting times ~ $\frac{1}{n\lambda}$



21

However, there is second common generating model (Hey 1992)

Species split randomly from other species at some rate λ. A third randomly chosen species goes extinct.

Total number of species remains constant through time.

Formally equivalent to Kingmans' coalescent process for the genealogy of neutral alleles in constant population (Prof. Hey is a population geneticist)

22

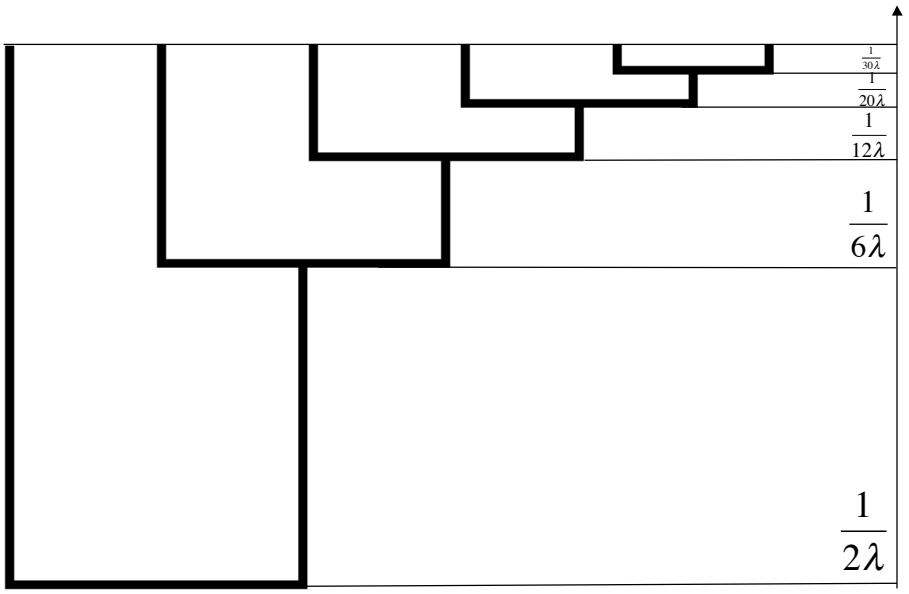
There is second common generating model (Hey 1992)

1. Expected topologies are the same as Yule!
2. Expected waiting times are very different

(Hey, Nee, Aldous)

23

Hey expected waiting times ~ $\frac{1}{n(n-1)\lambda}$

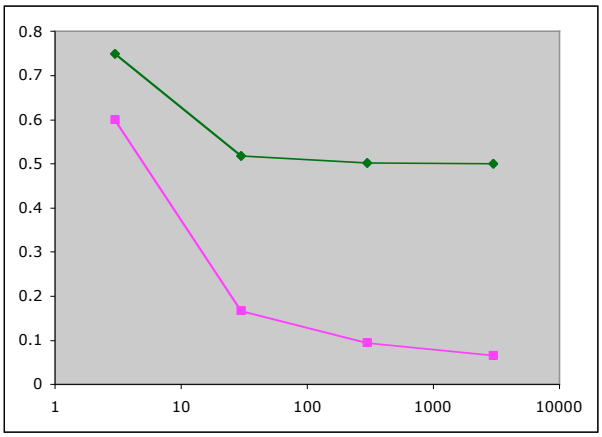


24

First split Hey model: $\geq \frac{1}{2}$ of total depth

First split Yule model: $\leq \frac{3}{5}$ of total depth

$\frac{a}{a+b}$



tree size

25

This property of long root branches on the coalescent is well-known (Prof. Felsenstein's presentation), but is relevant to the discussion of phylogenetic redundancy

26

Hollow curves redux

Scotland & Sanderson (Science, 2004):
-evolved discrete characters down Yule trees
-used character distributions to define 'genera'

This produced extreme hollow curves:
many character changes on terminal branches = many monotypic 'genera' (only one species).

	% monotypes	
taxon	Taxonomy	Simulation
Legumes	0.29	0.50
Composites	0.33	0.51
Birds	0.42	0.54

27

Samples of trees have shapes at odds with Yule/Hey in the opposite direction (too many small clades)

--how established?

1. measure the shape of all trees in sample
2. compare with distribution of shapes of Yule/Hey trees

There are >10 different published measures of shape

Michaël Blum's talk tomorrow erases the red that was to follow...

28

Measures of shape

$$I_c = \frac{\sum_i |r_i - l_i|}{0.5(n-1)(n-2)}$$

← sum of sister clade difference

← maximum possible for tree of size n

-expectation depends strongly on tree size (n)

$$E(I_c) = \frac{2n}{(n-1)(n-2)} \sum_{j=2}^{\frac{n}{2}} \frac{1}{j}, \text{ n even}$$

$$E(I_c) = \frac{2n}{(n-1)(n-2} \left[\frac{1}{n} + \sum_{j=2}^{\frac{(n-1)}{2}} \frac{1}{j} \right], \text{ n odd}$$

tomorrow's talk supercedes this...

Colless, Heard, Brown

What have these measures told us about tree samples?

study	N	treesizes	measure	outcome
Savage '83	<1000	4 -7	prop	=Yule
Guyer & Slowinski '91	120	5	prop	unbalanced
Heard '92	196	4-14	Ic per N	unbalanced
Guyer & Slowinski '93	30	100-20k*	*	unbalanced
Mooers '95	39	8-14	pIc	incomplete < complete
Mooers et al. '95 Purvis (pers. comm.)	31 ''	8-14 ''	pIc I_w	*f*(tree support) *f*(tree support)
Harcourt-Brown et al. '01	100	8-36	Ic	paleo unbalanced
Purvis & Agapow '02	61	6-334	I_w	higher taxa < species
Stam '02	69	8-67	Ic-E(Ic)	not *f*(tree support)
Rüber & Zardoya '05	14	9-102	B_1	all unbalanced

*considered nodes, not trees

What causes these deviations?

A. Trees are biased (methodological)
B. Trees are interesting (biological)

First need to introduce third generating model for topologies:
PDA (proportional-to-distinguishable arrangements) or Uniform

31

PDA or Uniform:
tree shapes expected in proportion to their frequency across all (labelled) cladograms.

$(2n-3)!!$ labelled cladograms for n taxa

$\frac{n!}{2^{\sigma}}$ unique labellings per cladogram,

σ is the number of nodes where subtrees are identical in shape

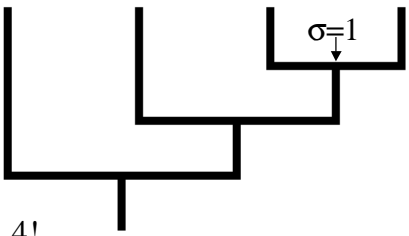
$$P(shape) = \frac{n!}{2^{\sigma}(2n-3)!!}$$

Steel & MacKenzie `01

32

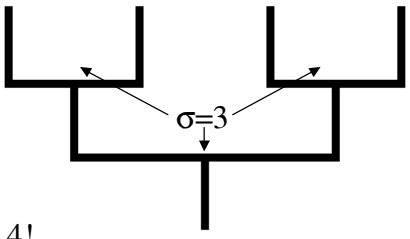
PDA:
treeshapes represented in proportion to their frequency across labelled cladograms.

$(2n-3)!! = 15$ labelled cladograms



$\frac{4!}{2^1}$ unique labellings on this tree

P(tree)=$\frac{12}{15}$



$\frac{4!}{2^3}$ unique labellings on this tree

P(tree)=$\frac{3}{15}$

33

PDA

1. **Not a way to generate trees, so no branchlengths**
2. **Probability of a split (*k, n-k*) at a node is just:**

$$\binom{n}{k}\frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!},\ \text{k<n-k}$$

$$\frac{1}{2}\binom{n}{k}\frac{(2k-3)!!(2(n-k)-3)!!}{(2n-3)!!},\ \text{k=n-k}$$

(and median size of smaller clade is 1.5)

Aldous, `01

34

A. Why are trees unbalanced?

Conjecture 1: "Random data" will produce uniform distribution of all possible labelled trees (ie, PDA)

For MP, this is false (Goloboff, 1991)

--However, some proportion of the randomness will be pure noise, drawing from the PDA (so final result is less balanced)

Mooers, Purvis, Cunningham, Huelsenbeck

35

Do data support a methodological problem?

A. Simulation studies

Huelsenbeck & Kirkpatrick (`96): yes (n_t: ~ 25)

Mooers et al. (`95): yes (n_t: 8)

B. Empirical studies

Mooers et al. (`95): ↑support → ↑balance

Purvis (pers. comm.): ↑ support → ↑balance

Stam (`02): ↑support does not → ↑balance

(larger sample, trees, n_t: 8-67…)

C. (Observational data)

Wilkinson et al. (in press) - Supertrees are too unbalanced, because unbalanced source trees contribute too much data

36

Are our hypotheses biased towards less balanced trees?

Given the sizes of trees built today,
and the rules for sampling from treespace
(e.g NNI, TBR, etc and uniform prior)
Perhaps this needs a second (third) look?
(related to the attributes of the tree landscape?)

37

B. Trees are interesting (biological)

Currently, two classes of model that make $\lambda \neq \text{Constant}$

1. **Heritable variation in** λ
2. $\lambda = f$**(age of lineage)**

(Purvis, Mooers)

38

1. Heritable variation in λ

(i) Processes that produce heritable variation in λ **decrease balance:**

Heard (s `96)
Heard & Mooers (s `02)
Agapow & Purvis (s `02)
Pinelis (a `02)

No data-based studies

39

1. Heritable variation in λ

Given that variation in λ builds up through time, we can make the following prediction:

Older trees should be less balanced

Burlando `90:
Marine taxonomies have steeper log-log slopes
Purvis & Agapow `02:
Higher taxon trees more unbalanced than species trees
(though taxonomy & phylogeny are confounded here.)

(Vazquez, Mooers, Bininda-Emonds)

40

1. Heritable variation in λ

--Processes or situations that increase heritable variation in λ should decrease balance

e.g.
1. Clades with strongly interacting species (radiations on islands)
2. Clades under strongly diversifying selection (biogeographically widespread)
3. Clades with large variation in relevant traits (ecologically distinct)

41

1. Heritable variation in λ

(ii) Tree shape is unaffected by random extinction (Slowinski & Guyer `89); heritable variation in P(extinction) should also decrease balance.

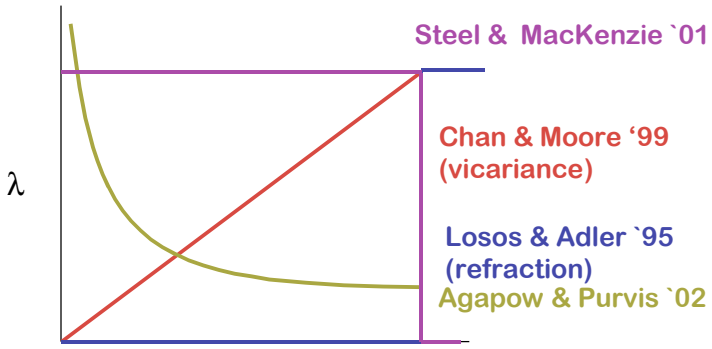
Mooers `95 (s, single extinction event on Yule trees)
Von Euler `00 (d, projected extinction on Bird taxonomy)
Maia `04 (s, continuous extinction on Hey trees)

Should this be investigated further (see last slide)?

42

B. Trees are interesting (biological)

2. Variation in λ related to age of lineage (two classes)

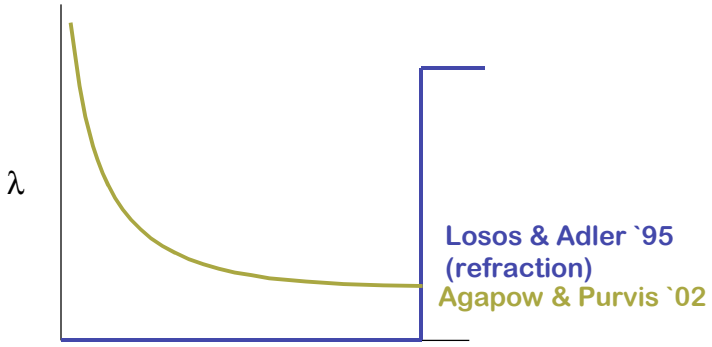


43

B. Trees are interesting (biological)

Refractory periods increase balance

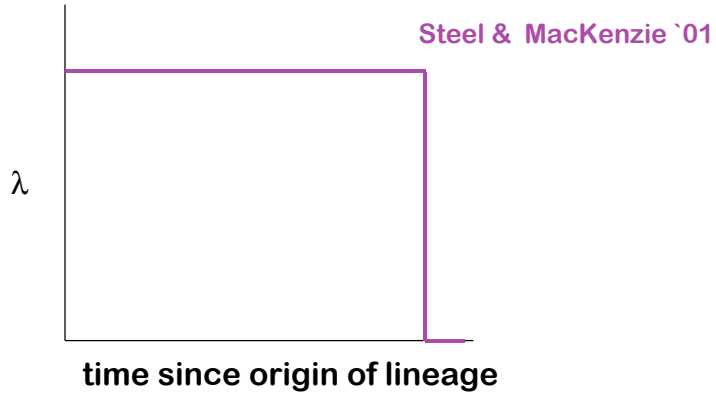
Dying species models decrease balance



44

λ = *f*(age of lineage)
At the limit, for some lineages, λ decreases to 0
(dead species scenario)

Harcourt-Brown et al. (d `01)
Steel and MacKenzie (a `01)
(Pinellis `02)

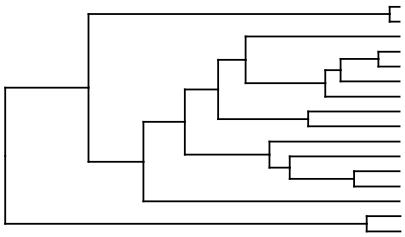


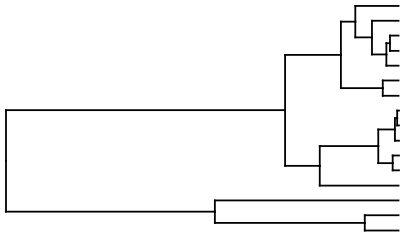
45

For this class of model $\lambda = f$(age of lineage),
I believe there is no expectation that the age of the tree will affect its topology (?)

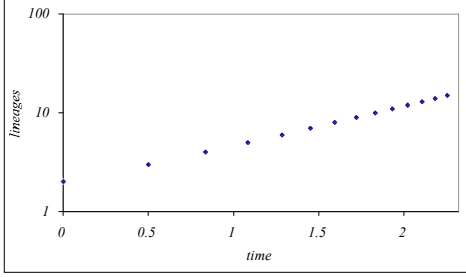
46

Yule **vs.** **Hey**

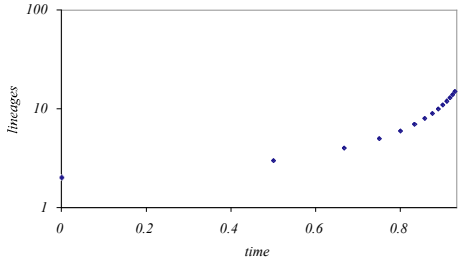




Expected waiting times under Yule model, lineages log transformed



Expected waiting times under Hey model, lineages log transformed



47

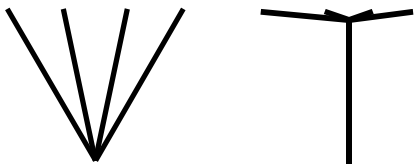
What can we do with a timeline on a tree?

1. Estimate mean and variance in speciation & extinction rates
   (Nee, Bokma, Paradis, Salamin)
2. Correlate these rates with (biological) attributes
   (Paradis `05)
3. Give us a(nother) way to value species
   (Faith, Crozier)

48

Redundancy (R)

measure of distribution of the nodes on a rooted tree from root to leaves:
f(proportion of the tree shared among leaves)



star phylogeny
R~0

umbel phylogeny
R~1

$$R = 1 - \frac{TL - d}{dn - d}$$

TL = total tree length (or PD)

d = depth of tree

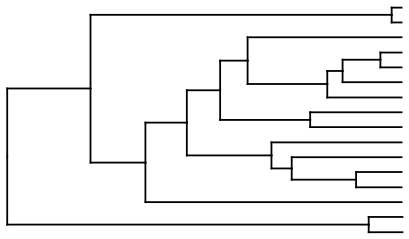
(Mooers & Vos)

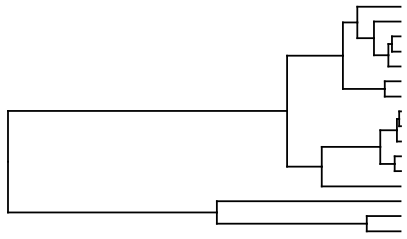
a heuristic measure (c.f. Pybus γ)

49

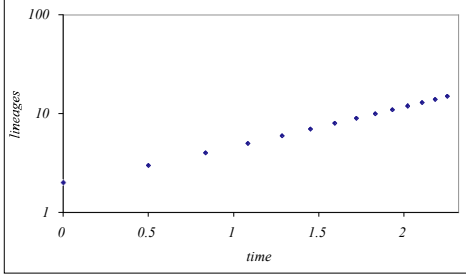
Yule
R= 0.64 (sd = 0.08)

Hey
R= 0.80 (sd = 0.06)

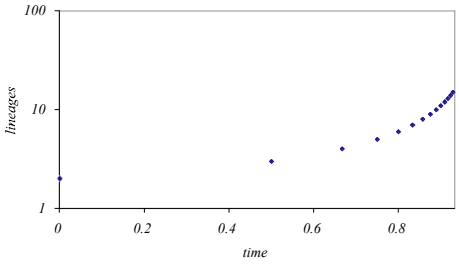




Expected waiting times under Yule model, lineages log transformed

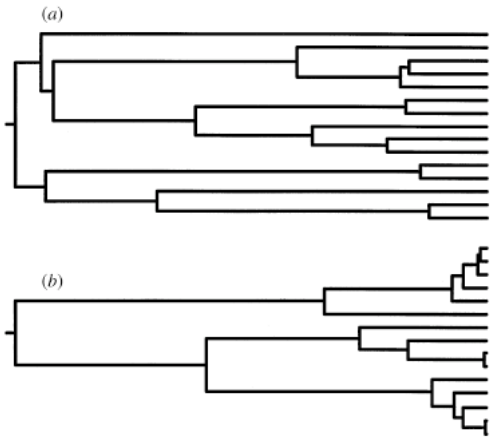


Expected waiting times under Hey model, lineages log transformed



50

Hey looks like Yule with high constant extinction



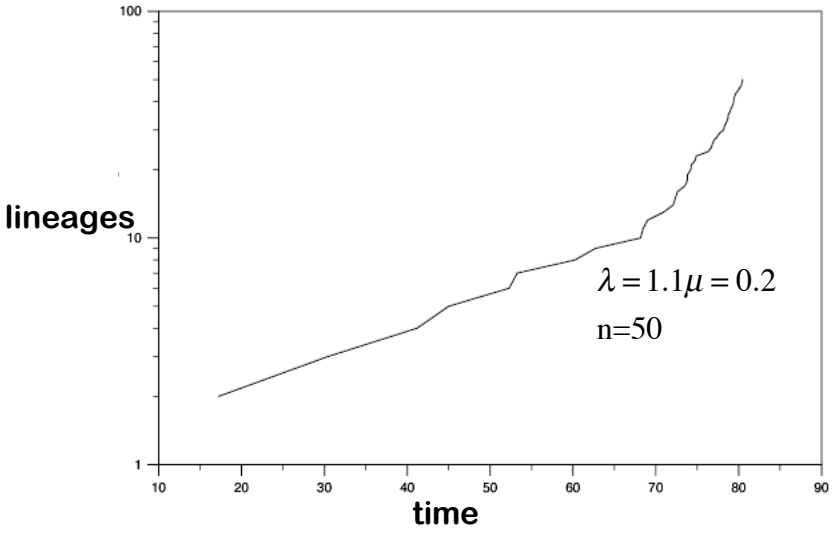
Yule tree, no background extinction

Yule tree, high background extinction

Harvey & Rambaut `98

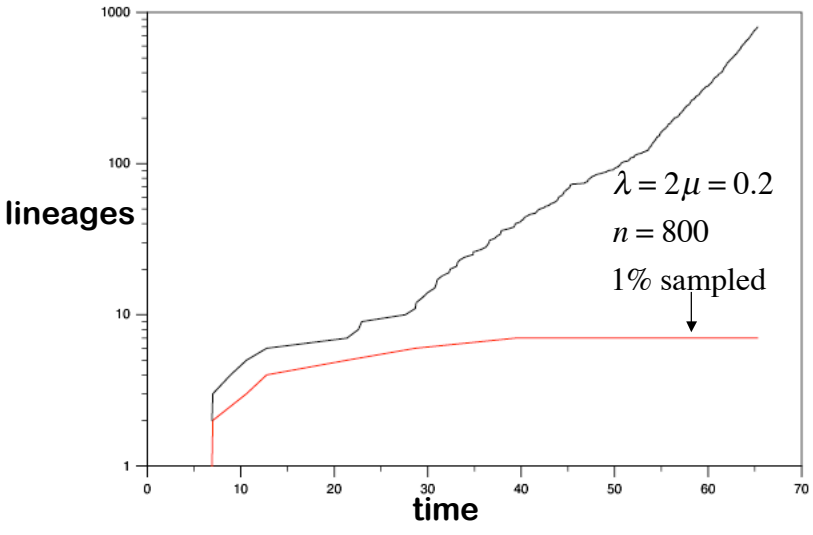
51

Yule with death=0.9*birth



52

A small sample from the present decreases redundancy (R falls) - nodes are nearer root



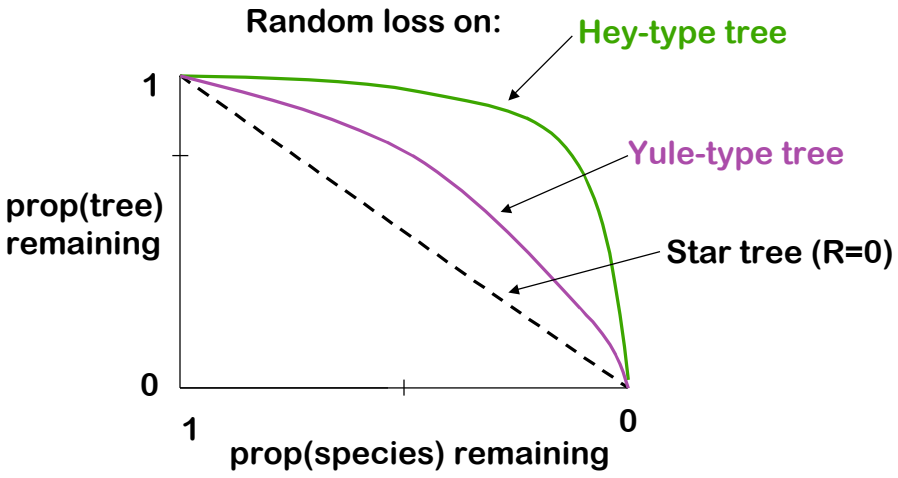
53

Trees with low redundancy have long terminal branches, and so may be harder to reconstruct correctly

Felsenstein `78
Huelsenbeck & Lander `03

54

Under high R, no single species contributes much to tree



55

The Yule and Hey curves are quite different

Save (or sample) 10 species from a clade with 100 extant -

Hey: a random 10% saves ~53%

Yule: a random 10% saves only ~ 25%
(indeed the 10% that maximizes savings captures < 33%)

(Eq. 1-4 in Nee & May `97)

56

If one of the things we would like to conserve is PD, then knowing the redundancy of real clades is important

57

[We don't]

58

Even getting an ultrametric tree is tricky business:
Most samples of DNA sequences are probably rejected by tests for clock-like evolution

59

And rate-smooting (linearizing) procedures may be biased

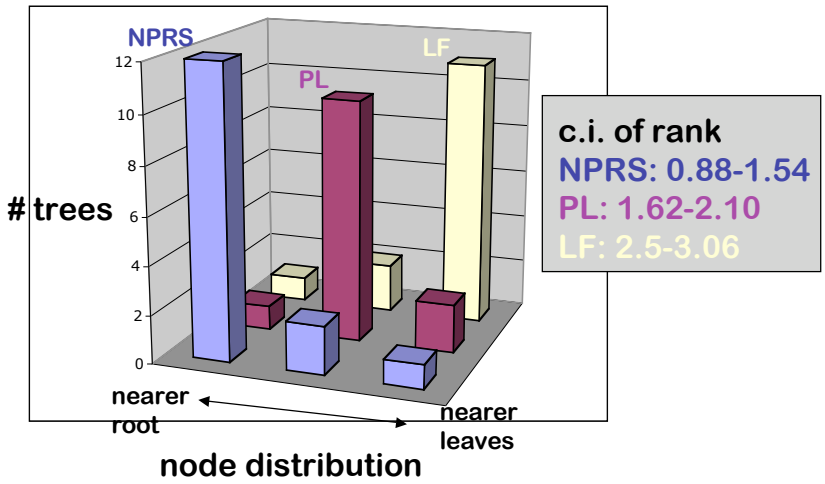
Rüber & Zardoya `05

1. Built trees for 14 clades of marine fish (9-102 spp)
2. Every dataset rejected the clock (cyt b + 12,16s rRNA)
3. Built ultrametric trees anyway (as we do):
   1. Using Langley-Fitch (LF) algorithm (`74)
   2. Using Penalized Likelihood (PL)
   3. Using Non-Parametric Rate Smooting (NPRS) (Sanderson `03)

R&Z were testing for a slowdown in cladogenesis, but...

60

Node distributions differ predictably between methods!



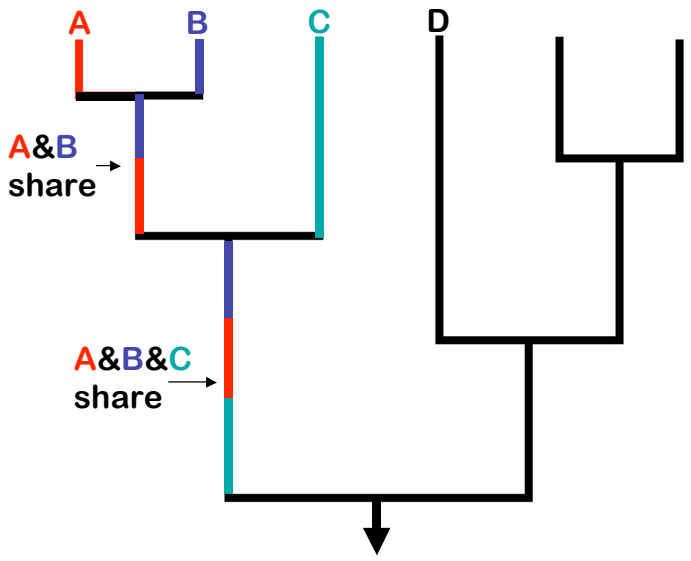
How general a property is this, and why does it happen

61

Let's pretend we do have a clock-like tree

62

1. Can we apportion evolution to the tips?
2. How does trees shape affect this apportioning?



Measures of worth (1)

Pendant edge (PE):

$PE_i = TL_n - TL_{n-i}$

TL_n is treelength of tree of size n

TL_{n-i}=TL of tree minus focal leaf i

Species age:
Altschul & Lippman `90
Faith `94

'Fair proportion (FP)':

$$'FP_i' = \sum_{j=1}^{r} \frac{B_j}{S_{j-1}}$$

j = internal node on direct path from i to root (r)

B_j = edge length from j to $j-1$, B_1 = pendant edge

S_j = size of subtree subtended by j, $S_0 = 1$

(Redding & Mooers)

Measures of worth (1)

'Equal splits (ES)':

$$'ES' = \sum_{j=1}^{r} \frac{B_j}{\prod_{k=1}^{j} (d(k-1)-1)}$$

j = internal node on direct path from i to root (r)

B_j = edge length from internal node j to j - 1

$d(k)$ = degree (3 for bifurcation) at node k

(Redding & Mooers)
(but see also Prof. Steel's presentation)

Measures of worth (1)

Shapley value (*S*)

$$S_i = \frac{1}{n}\sum_{\substack{s \subseteq n \\ i \in s}} \binom{n-1}{s-1}^{-1} (TL_s - TL_{s-i})$$

$$= \frac{1}{n!}\sum_{\substack{s \subseteq n \\ i \in s}} (s-1)!(n-s)!(TL_s - TL_{s-i})$$

i is the focal species

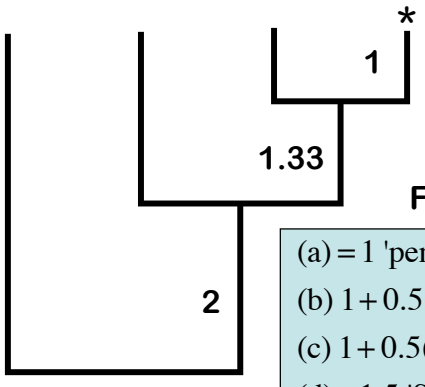
s is the subgroup

TL_s is the length of the tree of s

The contribution of species to i to a randomly chosen subgroup of randomly chosen size

Haake, Kashiwada & Su (Univ. Bielefeld) unpublished

Partitioning a tree among its species



Four measures of w for *:

(a) = 1 'pendant edge'

(b) $1 + 0.5(1.33) + 0.33(2) =$ 2.33 'fair proportion'

(c) $1 + 0.5(1.33) + 0.25(2) = 2.166$ 'equal split'

(d) =1.5 'Shapley value'

67

The measures of worth are correlated to differing degrees:

	pendant	fair	equal	shapley
pendant		0.65	0.71	0.58
fair			0.86	0.98
equal				0.83

correlation coefficients of the log(measures) across 100 16-taxa Hey trees

68

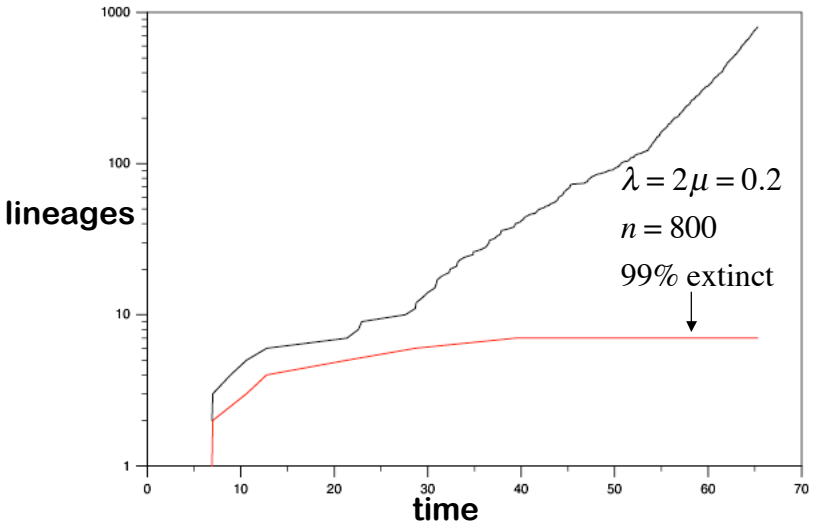
The measures of worth are (differentially) affected by tree shape:

	Partial $F_{1,96}$		
	Ic	R	IC*R
pendant	13	53	ns
fair	63	33	10
equal	84	29	6.8
shapley	84	3.9	8.9

All n=16, N=100, R^2~44-56%, $F_{0.01}$ = 6.9, $F_{0.001}$ = 11.5
response=log(standard deviation of measure)
Ic = Colless' measure of balance
R = Redundancy

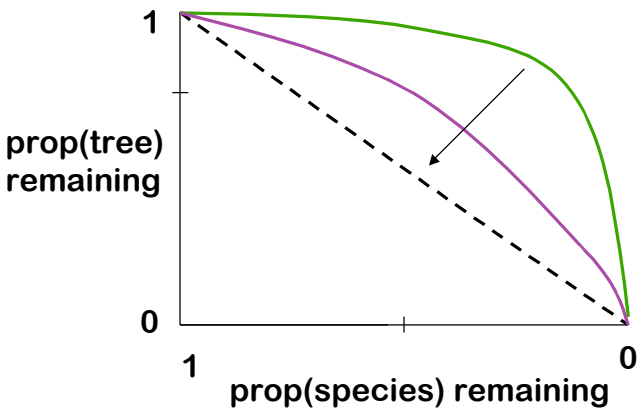
69

Mass extinction in the present decreases redundancy (R falls)



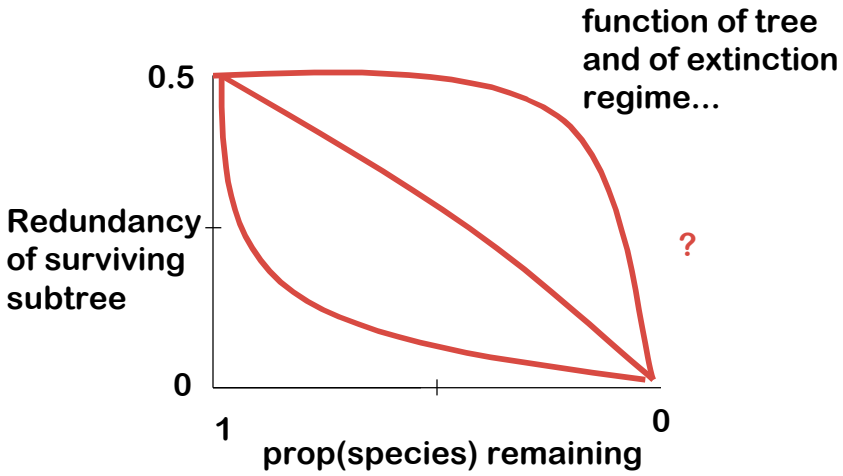
70

And as R decreases, we move toward the line of equality, and individual species become more 'worthwhile'



71

How does R change as species are removed by extinction (non)randomly?



72

Given extinction is nonrandom, are remaining species expected to show more variation in worth, such that some become very valuable ?

picture of cassowary

To recap what I'd love help with:

a: How many shapes are there for n taxa?

d, a, s: Are older trees less balanced?

a: Are our phylogenetic hypotheses biased towards
-less balanced trees?
-decreased redundancy?

a, s, d: At what does extinction increasing variation in species worth?

And most importantly:

d: What is the redundancy of real trees?

Literature cited

Agapow, P. M. and A. Purvis (2002). "Power of eight tree shape statistics to detect nonrandom diversification: A comparison by simulation of two models of cladogenesis." Systematic Biology **51**: 866-872.
Aldous, D. J. (2001). "Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today." Statistical Science **16**: 23-34.
Altschul, S. F. and D. J. Lipman (1990). "Equal animals." Nature, Lond. **348**: 493-494.
Burlando, B. (1990). "The fractal dimension of taxonomic systems." J. theor. Biol. **146**: 99-114.
Chan, K. M. A. and B. R. Moore (1999). "Accounting for mode of speciation increases power and realism of tests of phylogenetic asymmetry." American Naturalist **153**: 332-346.
Cunningham, C. W. (1997). "Is Congruence Between Data Partitions a Reliable Predictor of Phylogenetic Accuracy - Empirically Testing an Iterative Procedure For Choosing Among Phylogenetic Methods." Systematic Biology **46**: 464-478.
Faith, D. H. (1994). Phylogenetic diversity: a general framework for the prediction of feature diversity. Systematics and Conservation Evaluation. P. L. Forey, C. J. Humphries and R. I. Vane-Wright. Oxford, Clarendon Press. Systematics Association special vol. **50:** 251-268.
Farris, J. S. (1976). "Expected asymmetry of phylogenetic trees." Syst. Zool. **25**: 196-198.

Literature cited

Felsenstein, J. (1978). "Cases in which parsimony and compatibility methods will be positively misleading." Systematic Zoology **27**: 401-410.
Haake, C.-J., A. Kashiwade, et al. (2005). "The Shapley value of phylogenetic trees." IMW Working paper **363**.
Harcourt-Brown, K. G., P. N. Pearson, et al. (2001). "The imbalance of paleontological trees." Paleobiology **27**: 188-204.
Harvey, P. H. and A. Rambaut (1998). "Phylogenetic extinction rates and comparative methodology." Proceedings of the Royal Society of London Series B-Biological Sciences **265**: 1691-1696.
Heard, S. B. (1992). "Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees." Evolution **46**: 1818-1826.
Heard, S. B. and A. O. Mooers (2000). "Phylogenetically patterned speciation rates and extinction risks change the loss of evolutionary history during extinctions." Proceedings of the Royal Society of London Series B-Biological Sciences **267**: 613-620.
Hey, J. (1992). "Using Phylogenetic Trees to Study Speciation and Extinction." Evolution **46**: 627-640.
Huelsenbeck, J. P. and M. Kirkpatrick (1996). "Do Phylogenetic Methods Produce Trees With Biased Shapes." Evolution **50**: 1418-1424.

Literature cited

Huelsenbeck, J. P. and K. M. Lander (2003). "Frequent inconsistency of parsimony under a simple model of cladogenesis." Systematic Biology **52**(5): 641-648.
Losos, J. B. and F. R. Adler (1995). "Stumped by trees? a generalized null model for patterns of organismal diversity." American Naturalist **145**: 329-342.
Maia, L. P., A. Colato, et al. (2004). "Effect of selection on the topology of genealogical trees." Journal of Theoretical Biology **226**(3): 315-320.
Mooers, A. Ø., R. D. M. Page, et al. (1995). "Phylogenetic noise leads to unbalanced cladistic tree reconstructions." Syst. Biol. **44**: 332-342.
Nee, S. and R. M. May (1997). "Extinction and the loss of evolutionary history." Science **278**(5338): 692-694.
Paradis, E. (2005). "Statistical analysis of diversification with species traits." Evolution **59**(1): 1-12.
Pinelis, I. (2003). "Evolutionary models of phylogenetic trees." Proceedings of the Royal Society of London Series B-Biological Sciences **270**(1522): 1425-1431.
Purvis, A. and P. M. Agapow (2002). "Phylogeny imbalance: Taxonomic level matters." Systematic Biology **51**(6): 844-854.
Ruber, L. and R. Zardoya (2005). "Rapid cladogenesis in marine fish revisited." Evolution **59**(5): 1119-1127.
Scotland, R. W. and M. J. Sanderson (2004). "The significance of few versus many in the tree of life." Science **303**: 643.

Literature cited

Slowinski, J. B. and C. Guyer (1989). "Testing the stochasticity of patterns of organismal diversity: an improved null model." Amer. Nat. **134**: 907-921.
Stam, E. (2002). "Does imbalance in phylogenies reflect only bias?" Evolution **56**: 1292-1295.
Steel, M. and A. McKenzie (2001). "Properties of phylogenetic trees generated by Yule-type speciation models." Mathematical Biosciences **170**: 91-112.
Stone, J. and J. Repka (1998). "Using a nonrecursive formula to determine cladogram probabilities." Systematic Biology **47**: 617-624.
von Euler, F. (2001). "Selective extinction and rapid loss of evolutionary history in the bird fauna." Proceedings of the Royal Society of London Series B-Biological Sciences **268**: 127-130.
Wilkinson, M., J. A. Cotton, et al. (2005). "The shape of supertrees to come: tree shape related properties of fourteen supertree methods." Syst. Biol.
Willis, J. C. and G. U. Yule (1922). "Some statistics of evolution and geographical distribution in plants and animals, and their significance." Nature **109**: 177-179.
Yule, G. U. (1924). "A mathematical theory of evolution based on the conclusions of Dr J. C. Willis." Philosophical Transactions of the Royal Society (London) Series B **213**: 21-87.