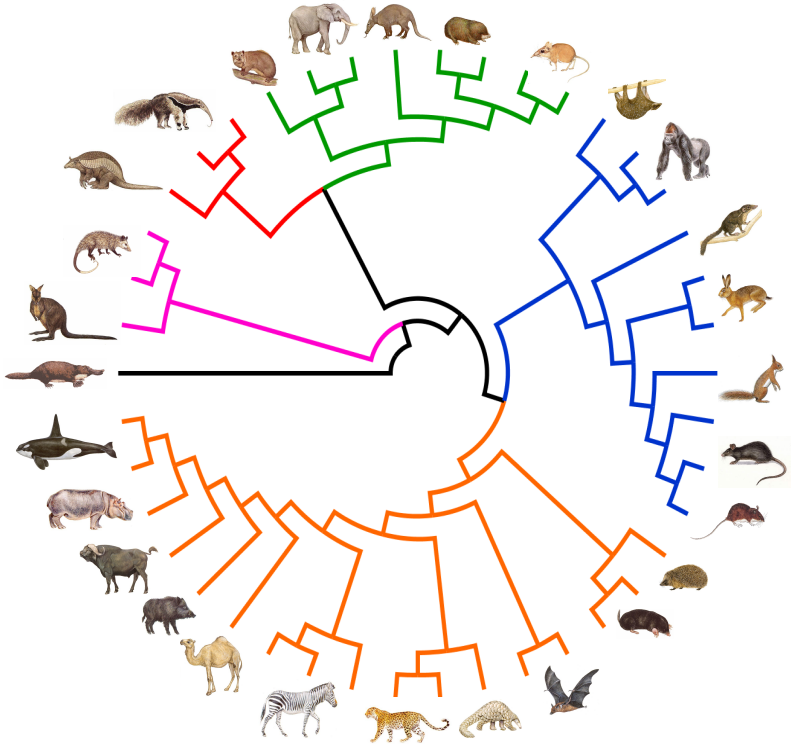
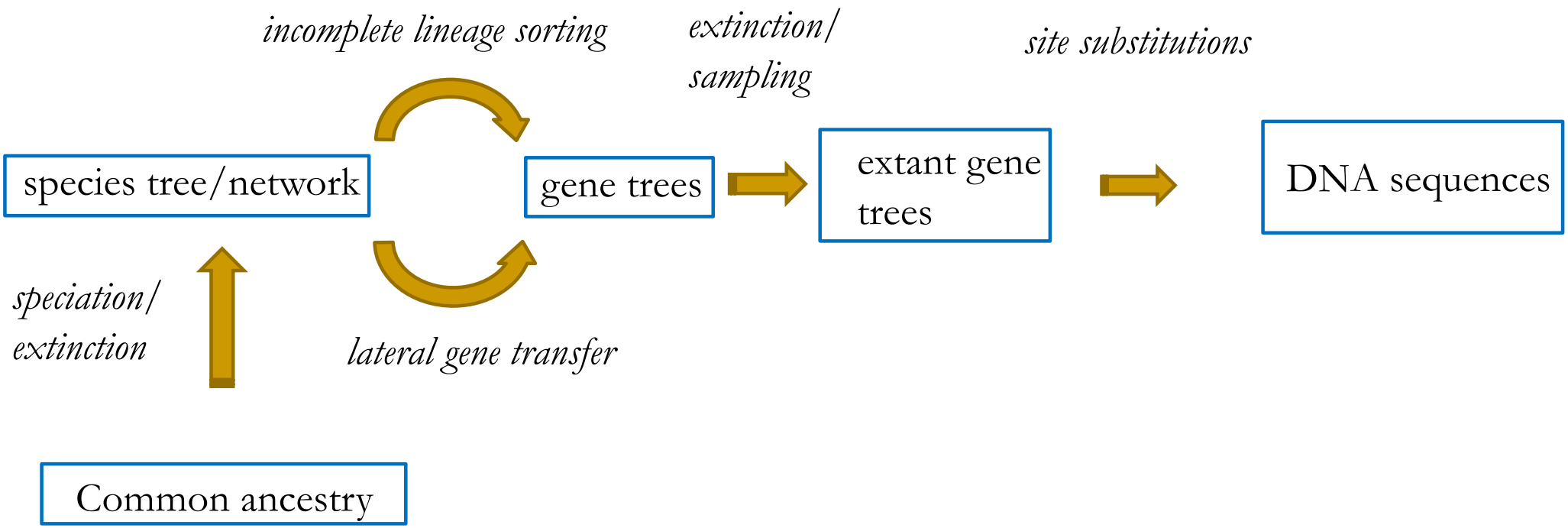
Deciphering a species phylogeny from conflicting gene trees

Mike Steel



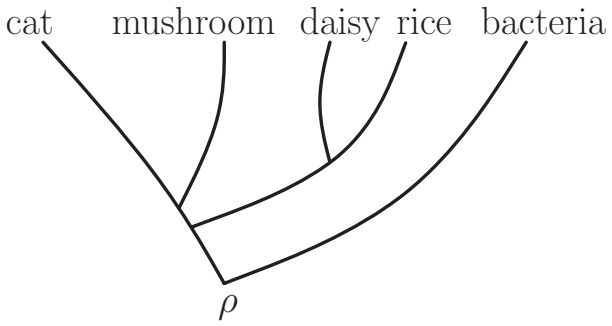
from F. Delsuc and N. Lartillot

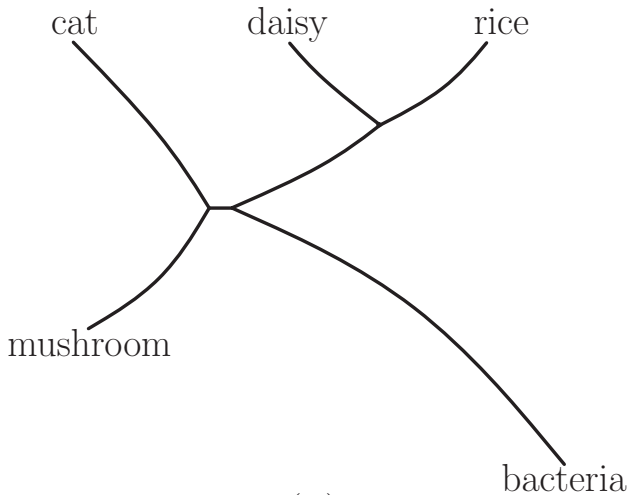
Overview: Stochastic models in phylogenetics



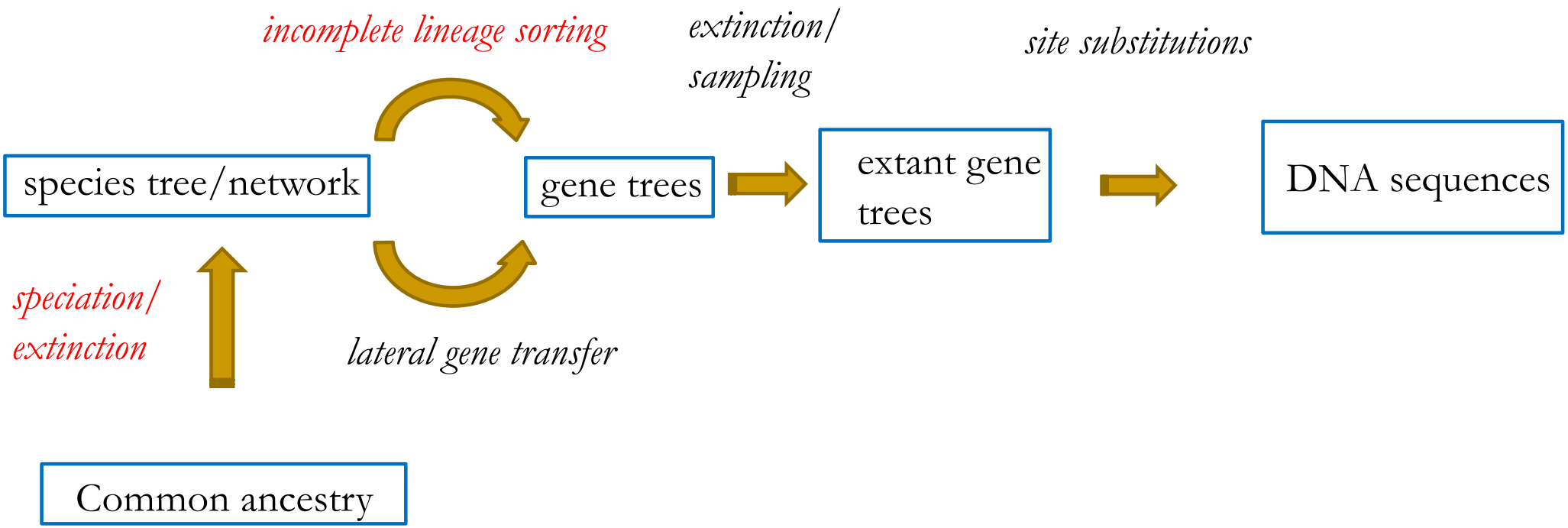
Phylogenetic basics

- rooted vs unrooted
- binary vs non-binary
- tree vs tree-shape
- trees encoded by
  - cluster/splits
  - induced subtree



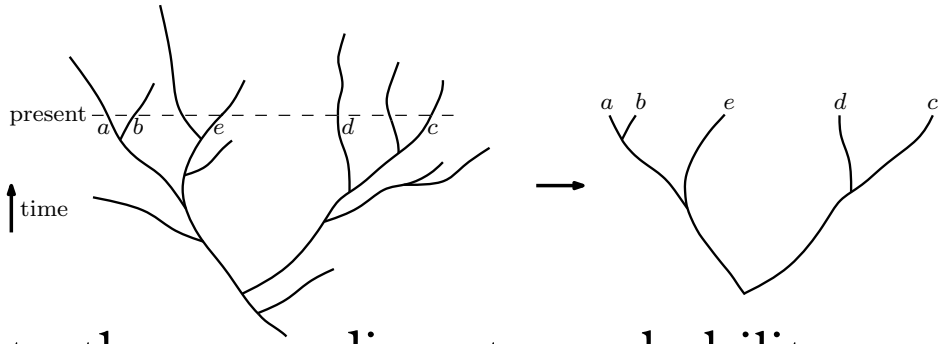


Overview: Stochastic models in phylogenetics



Random trees:

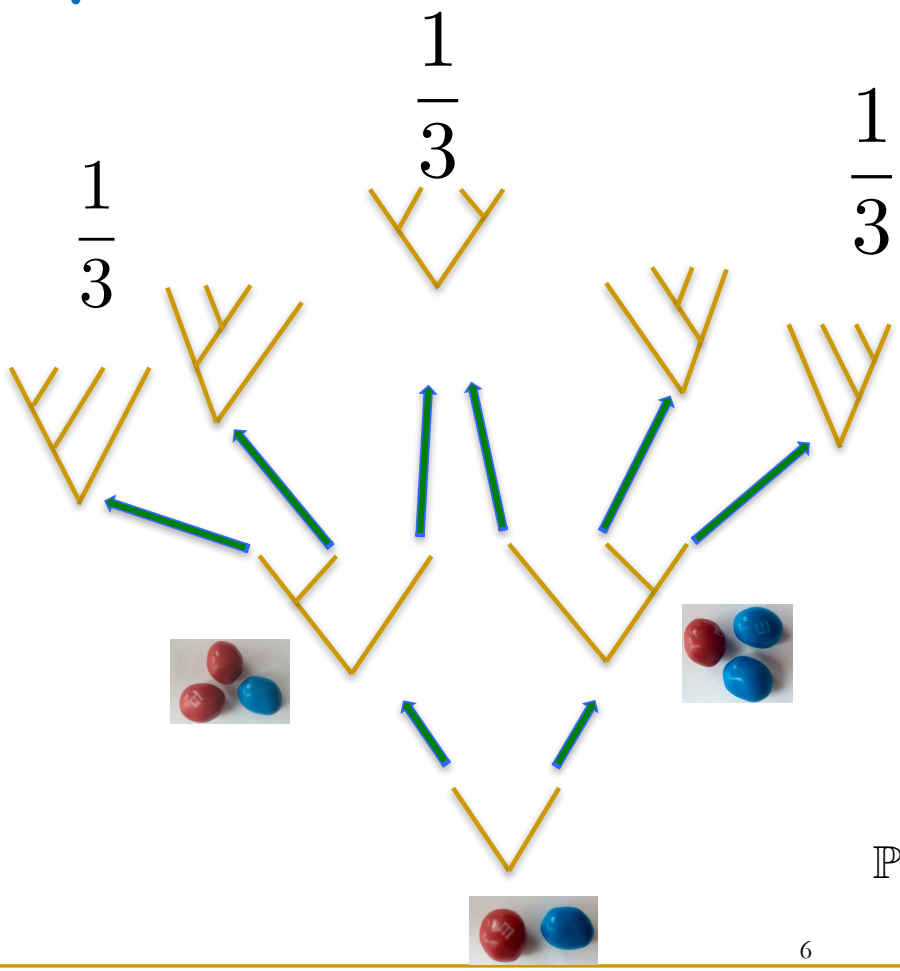
- Kingman coalescent trees – ignore ranking + branch lengths
- Yule pure-birth trees – ignore branch lengths
- Birth-death trees – look at the reconstructed (or 'reduced') trees

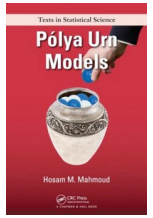


- **Theorem:** All three give rise to the same discrete probability distribution on rooted binary phylogenetic trees.
  - *Yule-Harding* (YH) distribution

[D. Aldous; A. Lambert and T. Stadler]

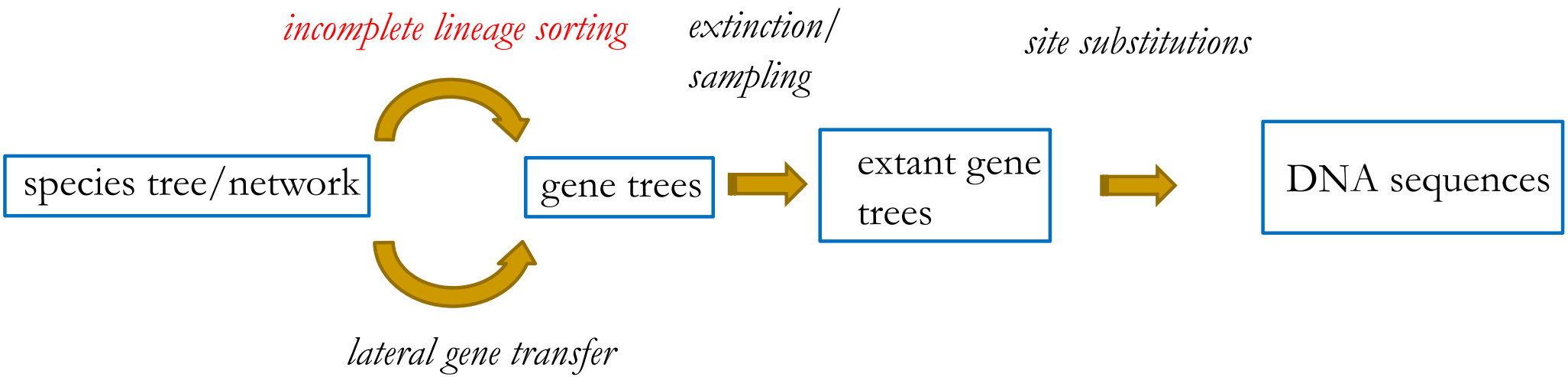
A basic property of this distribution





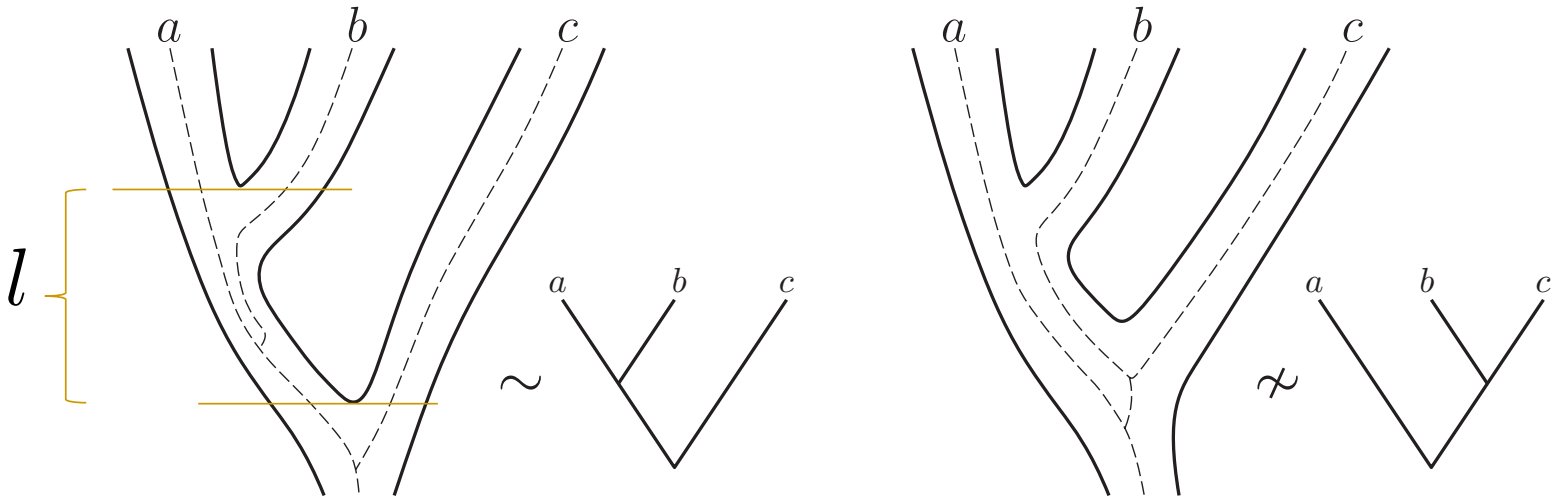
$$\mathbb{P}(Z_n = i) = \frac{1}{n-1}, \text{ for } i = 1, \ldots, n-1$$

Overview: phylogenetic trees/processes/inference



ILS (incomplete lineage sorting)

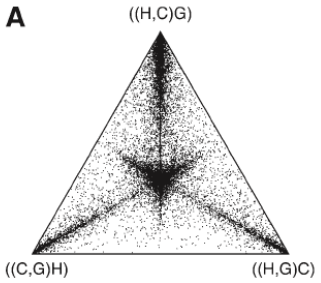
- *multispecies coalescent*



$$\mathbb{P}(\mathscr{T}_g = T_s | T_s, l) = \frac{1}{3}e^{-l} + (1 - e^{-l}) = 1 - \frac{2}{3}e^{-l}$$

For either of the two trees $T \neq T_s$,

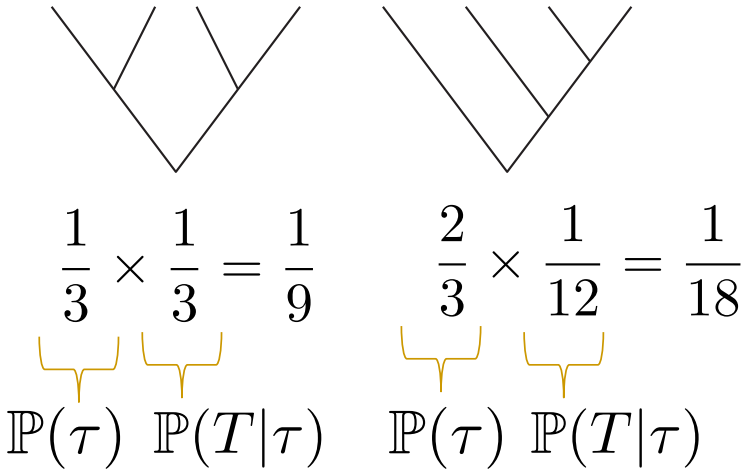
$$\mathbb{P}(\mathscr{T}_g = T_s | T_s, l) = \frac{1}{3}e^{-l}$$



Human/Chimp/Gorilla
(HC)G ~78%
(GC)H ~11% (HG)C ~11%

(Ebersberger et al. MBE 2007)

ILS: gene tree probabilities



- n =4
  - Take interior branch lengths very short

  (so all lineage coalesce above root)

  - So for any one of the 12 unbalanced trees (shape on right) there are branch lengths for which each of the 3 balanced trees has higher probability.
- For n>4:

Is there always a bias towards more 'balanced' trees?

ILS: anomalous gene trees

Gene tree T is an *anomalous gene tree* (AGT) for species tree T_s ($T_s \neq T$) if for some branch lengths ℓ for T_s,

$$\mathbb{P}(\mathcal{T}_g = T | T_s, \ell) > \mathbb{P}(\mathcal{T}_g = T_s | T_s, \ell)$$

- **Theorem** [Rosenberg, Degnan, 2005]
  - For **any** binary species tree T with **five or more** species, there exists an anomolous gene tree.

- Thus, any simple 'voting' strategy using gene trees is a statistically inconsistent estimator of the species tree.

- However, consistent methods exist.

Simple proof that statistically consistency methods exist

- A rooted binary tree T is uniquely determined by its induced rooted 3-leaf trees.
- We saw above, each 3-leaf trees can be recovered in a statistically consistent way
- Thus T can too.

What about clusters?

$$\mu(A) = \text{ expected frequency of gene trees that contain A as a cluster}$$

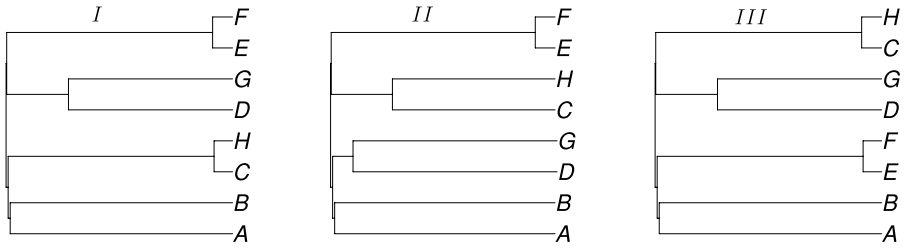
Proposition [Allman *et al.* 2011]

If $\mu(A) > 1/3$ then A is a cluster in the species tree

Allman, Elizabeth S., James H. Degnan and John A. Rhodes. Determining species tree topologies from clade probabilities under the coalescent. *J. Theor. Biol.* 289 (2011): 96-106.

ILS: AGTs and 'wicked forests'

A *wicked forest* is a set W of distinct rooted binary trees having the same leaf set, so that for all ordered pairs T, T' of distinct trees: T (regarded as a gene tree) in W, is an AGT for T' (regarded as a species tree for suitably chosen edge lengths).

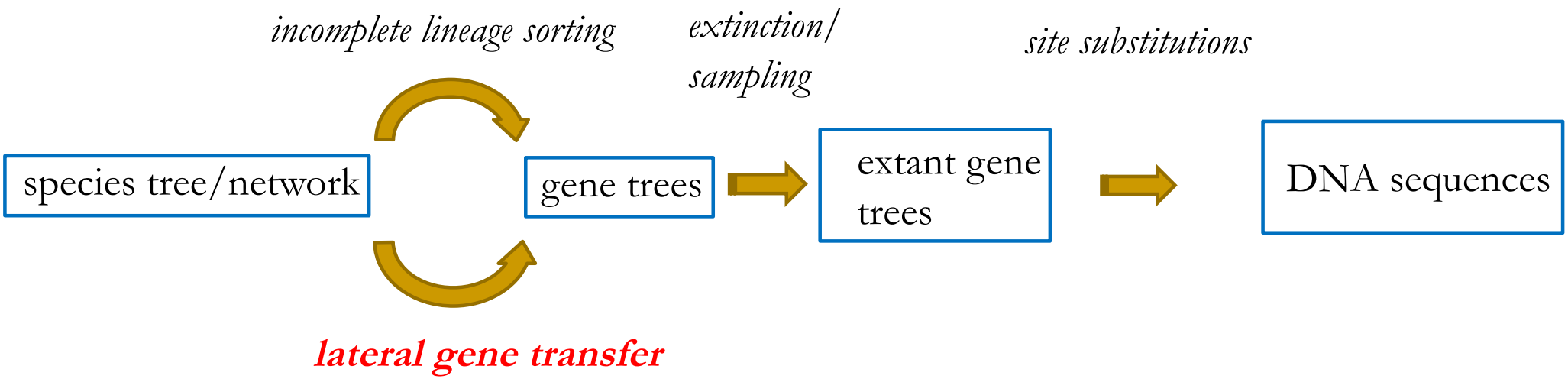


A 'wicked forest' of three trees on eight leaves (from Degan and Rosenberg 2006).

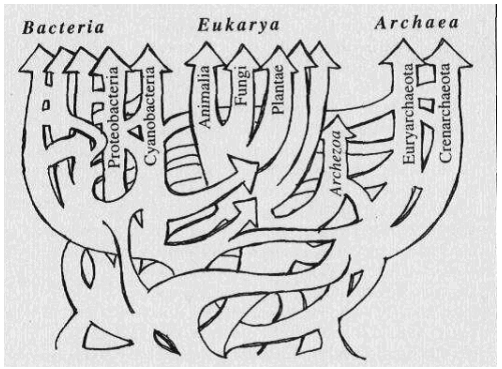
Theorem [Degnan, J.H. and Rhodes, J.A. (T.P.B. 2015)]
A caterpillar tree cannot be an AGT (so there are no caterpillars in a wicked forest).

Question: How large can a wicked forest be for n species?

Overview: Stochastic models in phylogenetics



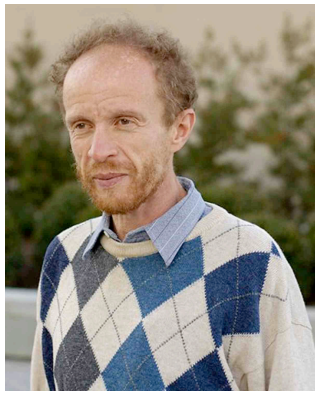
Trees or networks?



"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

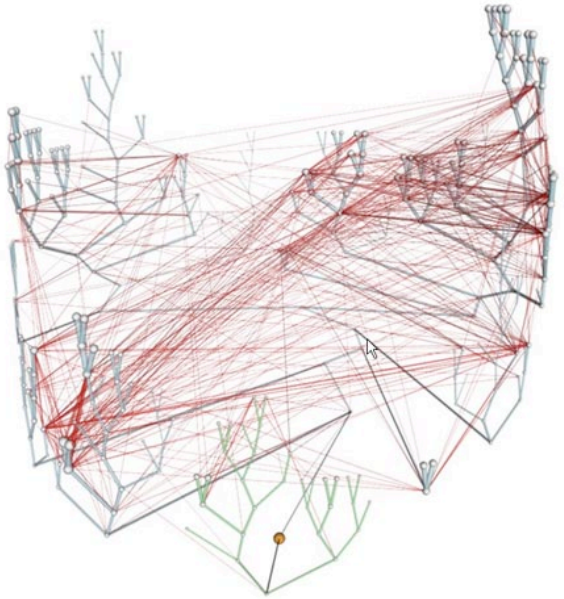
An alternative view



Eugene Koonin, NCBI

- A ‘central tendency’ tree with transfers between its branches

Koonin, E. (2015). The Turbulent network dynamics of microbial evolution and the statistical Tree of Life. *J. Mol. Evol.*, (in press)

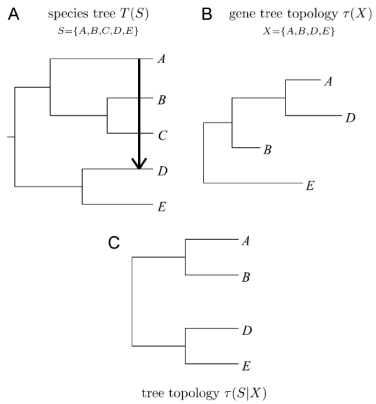


A bird's-eye view of the tree of life, showing the vines in red and the tree's branches in grey [Bacteria] and green [Archaea]. The last universal common ancestor is shown as a yellow sphere.

From: Kunin et al.(2005) The net of life: Reconstructing the microbial phylogenetic net

LGT stochastic models

Independent random transfers from from x to y at rate $\lambda(x, y, t)$ with copy of gene from x replacing copy of gene from y.



A Likelihood Framework to Measure Horizontal Gene Transfer

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

Simple 'vanilla' model: $\lambda(x, y, t) = \lambda$

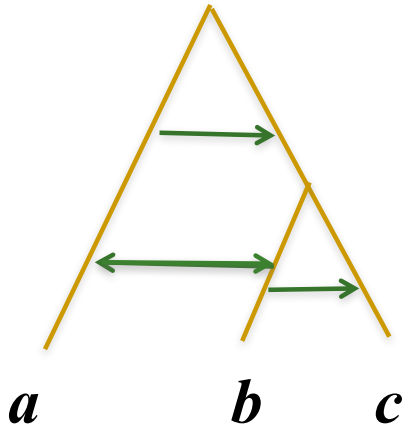
More refined models: $\lambda(x, y, t) = \lambda(b(x), b(y))$

$\lambda(x, y, t) = \lambda(d(x, y))$

$N_T :=$ total number of transfers in T

Vanilla model: $N_T \sim Po(\mu_T), \mu_T = \lambda L_T$

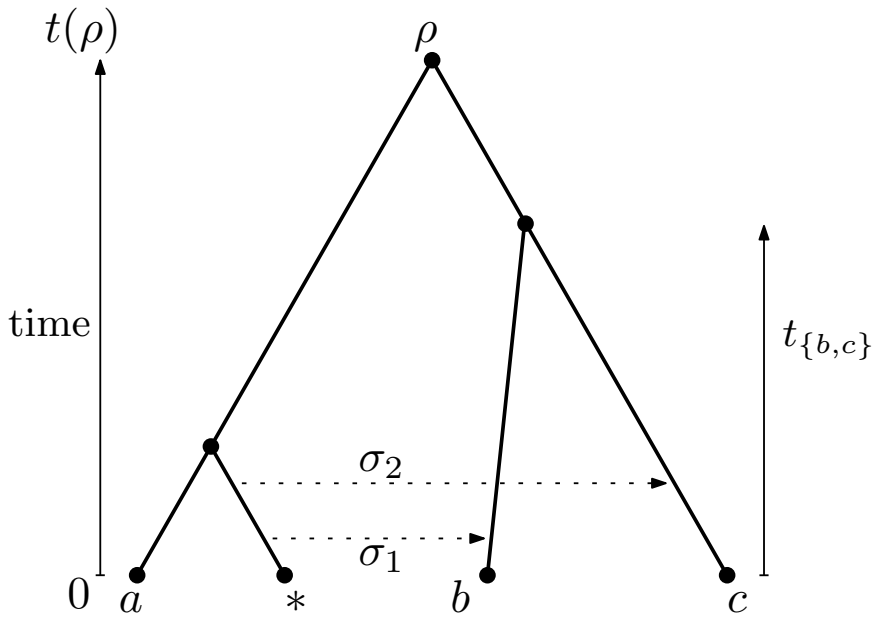
Simple case (n=3)

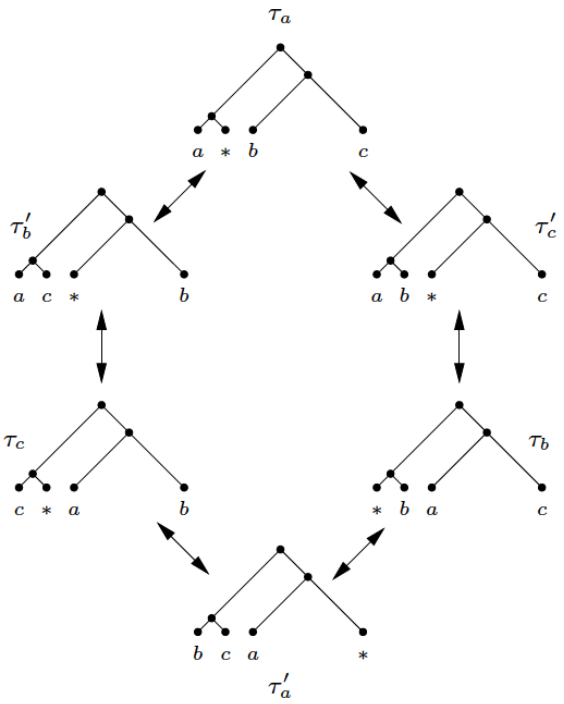


$$\mathbb{P}(\mathcal{T} = a|bc) = \tfrac{1}{3} \cdot p + 1 \cdot (1 - p)$$

$$\mathbb{P}(\mathcal{T} = b|ac) = \mathbb{P}(\mathcal{T} = c|ab) = \tfrac{1}{3} \cdot p$$

Still n=3, but what if there another lineage is present?

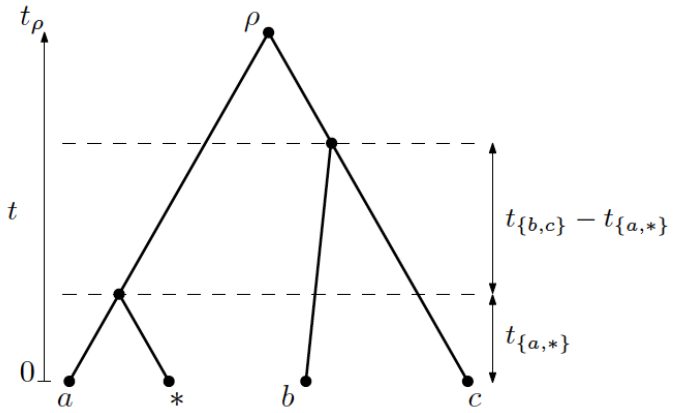




A main difference from ILS: additional lineages can matter!

‘moving’ vs ‘fixing transfers’

Zone of inconsistency



$$\mathbb{P}(b|ac) = \mathbb{P}(c|ab)$$

Theorem

$$\mathbb{P}(a|bc) = \frac{1}{3}\left[1 - e^{-7\mu}(1 - e^{-2\mu} - e^{-B}(1 + e^{-2\mu}))\right]$$

$$\mathbb{P}(a|bc) < \mathbb{P}(b|ac) = \mathbb{P}(c|ab)$$

if and only if $t_{\{b,c\}} - t_{\{a,*\}}$ *is greater than* $\frac{1}{3\lambda}\ln(\frac{1+e^{-2\mu}}{1-e^{-2\mu}})$

for $\mu = \frac{1}{3}\lambda t_{\{a,*\}}$, *and* $B = 3\lambda(t_{\{b,c\}} - t_{\{a,*\}})$

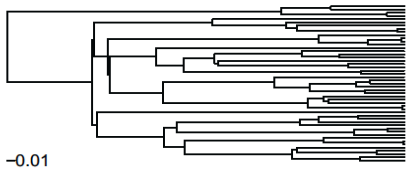
Species tree inference

Theorem [Roch and Snir, 2013; Steel, Linz, Huson, Sanderson 2013]

There 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'.

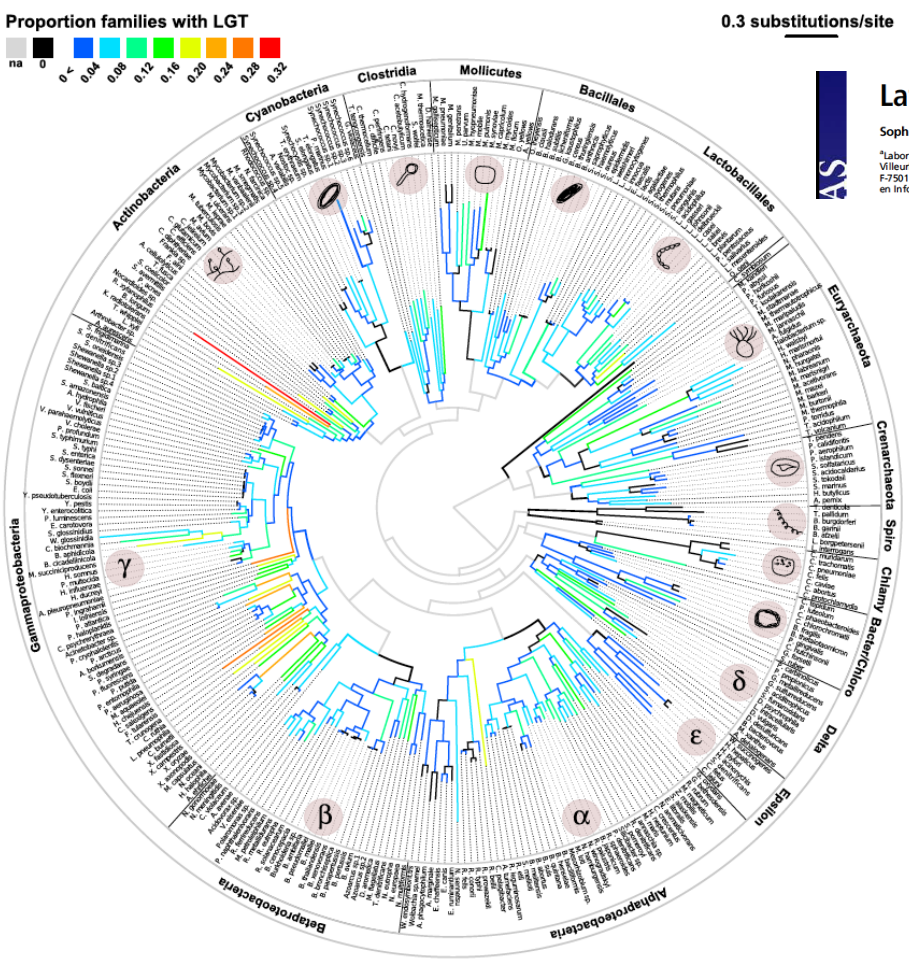
Example: for Yule species trees with n leaves the following suffices:

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



Moreover, $N = \Omega(\ln n)$ gene trees suffice (Roch+Snir, 2013)

Particular case: [Steel, 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.

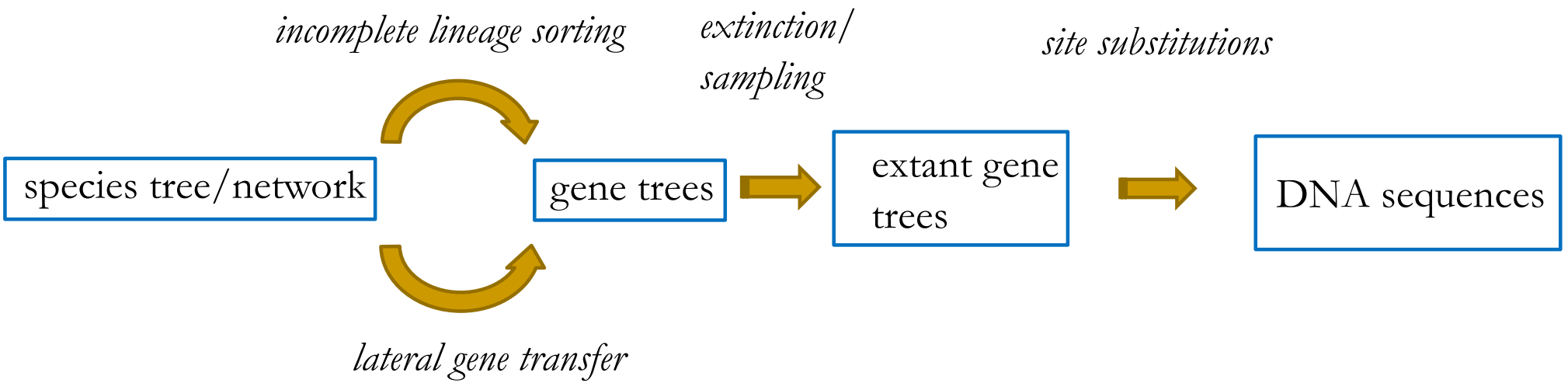


Lateral gene transfer as a support for the tree of life

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

[a]Laboratoire de Biométrie et Biologie Evolutive, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5558, Université Lyon 1, F-69622 Villeurbanne, France; [b]Université de Lyon, F-69000 Lyon, France; [c]Microbial Evolutionary Genomics, Département Génomes et Génétique, Institut Pasteur, F-75015 Paris, France; [d]Centre National de la Recherche Scientifique, Unité Mixte de Recherche 355, F-75015 Paris, France; and [e]Institut National de Recherche en Informatique et en Automatique Rhône-Alpes, F-38334 Montbonnot, France

Overview: Stochastic models in phylogenetics



Taxon coverage matrix (by genus) for cluster set at subtree whose root is 'Cactaceae'

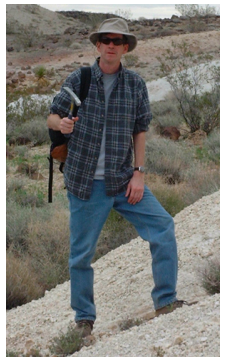
Taxon (108 total)	N_{cl}	0	1	2	14	15	16	17	18	19	20	21	22	23	51	52	53	54	55
Acanthocalycium	3		X			X	X												
Acanthocereus	3		X			X								X					
Acharagma	1	X																	
Ancistrocactus	1	X																	
Aniocarpus	1	X																	
Armatocereus	3		X			X								X					
Arrojadoa	4				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						
Brasiliopuntia	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

Decisiveness

Let T be a phylogenetic tree with leaf set X.

Michelle McMahon



Michael Sanderson

- A collection $\boldsymbol{S} = \{Y_1, \ldots, Y_k\}$ of subsets of X, with union X is **decisive for a tree** T if T is the **only** tree that displays the induced trees $T|Y_1, \ldots, T|Y_k$.

- $\boldsymbol{S}$ is **phylogenetically decisive** if it is decisive for every phylogenetic X-tree.

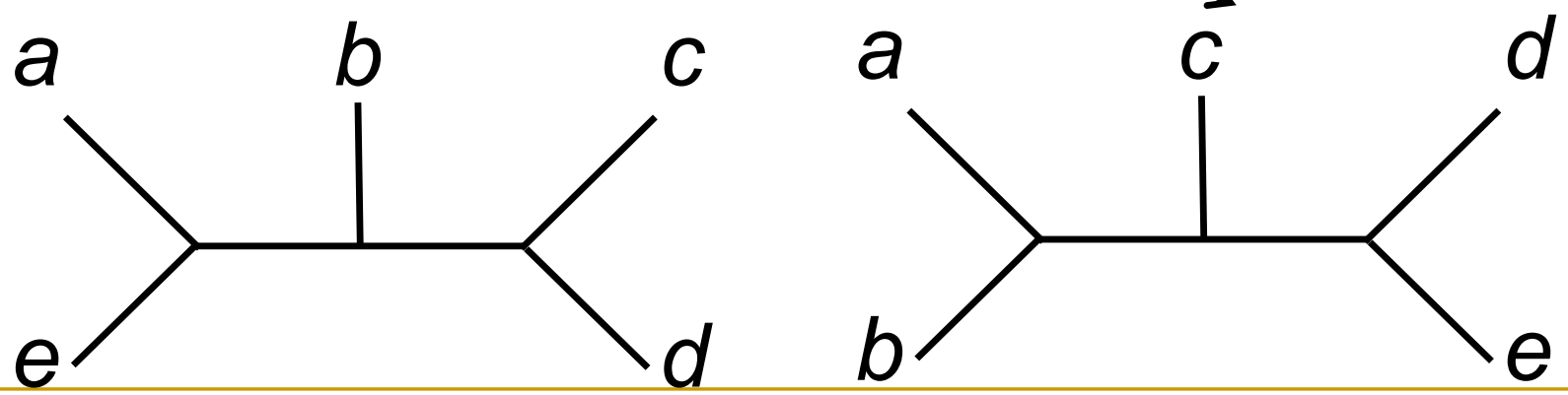
Example: phylogenetically decisive (for some but not all trees)

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



Example: 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

Combinatorial characterization

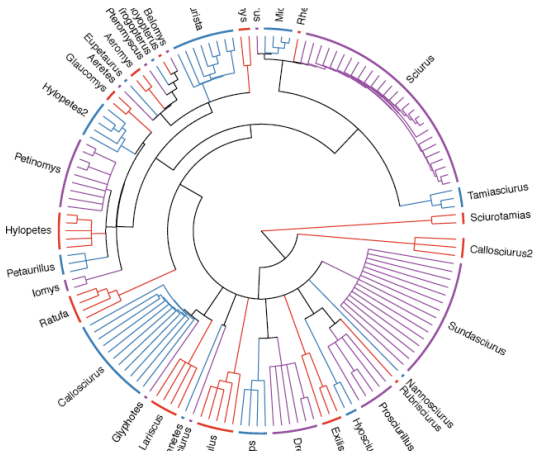
Theorem [S+Sanderson 2010]:

S is phylogenetically decisive

$\Longleftrightarrow$

for all partitions of X into four blocks, there is some $Y_i \in S$ that contains a point from each block.

Modelling random taxon coverage



p_{xj} = probability taxon x is present at loci j.

Choices made independently for k loci to get a pattern $\boldsymbol{S}$ of taxon coverage (n, p_{xj}, k).

Simplest model (**Uniform coverage:** $p_{xj} = p$ for all x, j)

Theorem

For any rooted binary tree *T* with *n* leaves, with coverage probability *p*, the probability that a set **S** of *k* (random) taxon sets is phylogenetically decisive for *T* is at least 1 - ε if

$$k \geq \frac{\ln((n-2)/\epsilon)}{-\log(1-p^3)} \sim \frac{\ln(n/\epsilon)}{p^3}$$

Phylogenetic decisive (for **all** trees) holds if

$$k \geq \frac{\ln(\binom{n}{3}/\epsilon)}{-\log(1-p^3)} \sim \frac{3\ln(n/\epsilon)}{p^3}$$

Fairly close necessary lower bounds on *k* also exist.

Example:

Suppose $n = 100$ and we expect 50% coverage of taxa.

How many loci do we need to be 95% certain that if we reconstruct a (sub)tree for each loci correctly then they will collectively define the underlying tree on all 100 taxa?

Then 57 loci suffice and that at least 49 may be required.

Applications to biological data

- How close to decisiveness in real data-sets?

Taxon	Source	Taxa	Loci	Coverage Density	Min loci	Max loci
Cactaceae	PhyLoTA [a]	488	18	0.14	2932	23891
Papilionoid legumes	McMahon (2006) [b]	1794	72	0.16	2306	16230
Metazoa	Hejnol et al., 2009	94	1487	0.18	1095	7148
Rice	Cranston et al. (2010)	10	9481	0.48	35	91

[b] Their "sparse analysis"
[a] PhyLoTA database

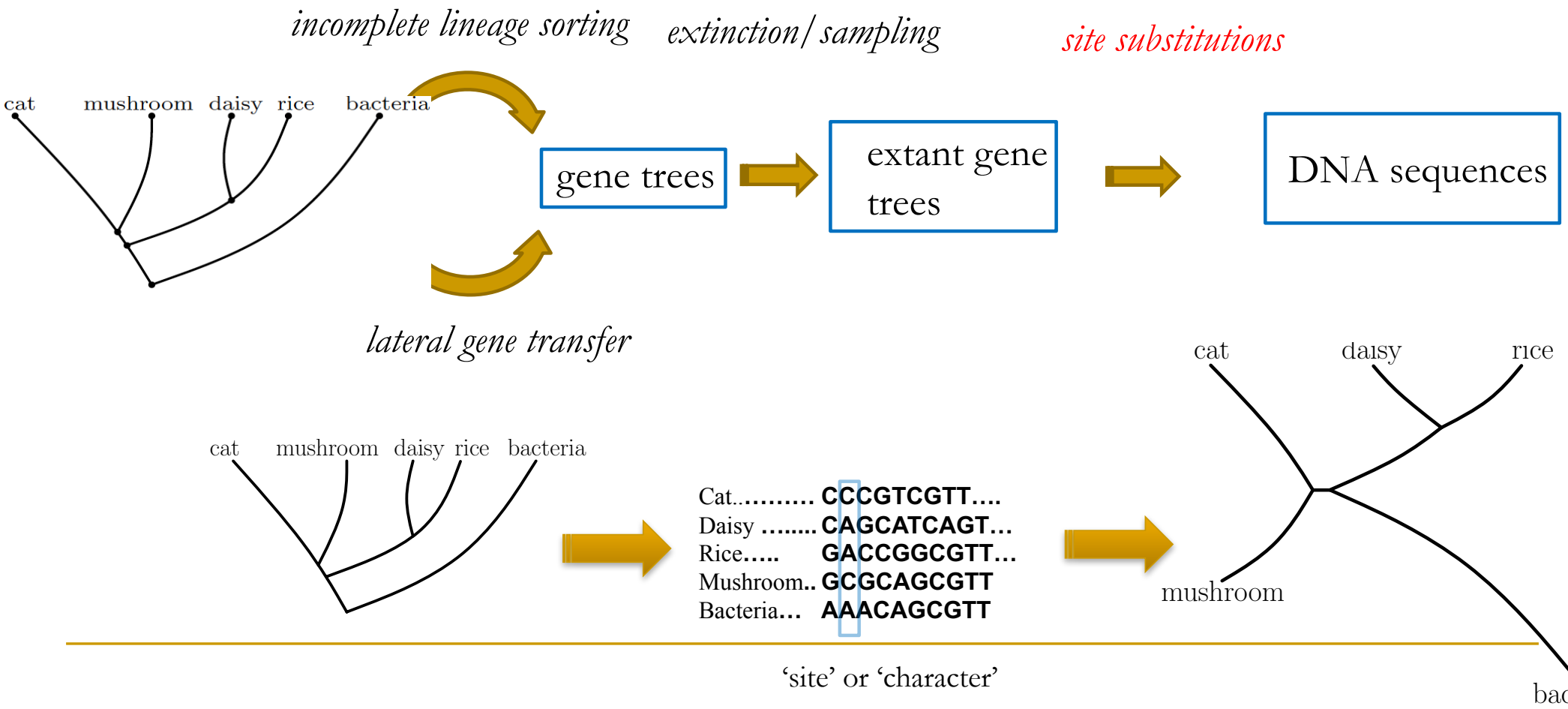
- Patchy taxon coverage also implies the existence of 'terraces' of equally optimal ML trees

Eg. Bouchenak-Khelladi et al. 2008 [Grasses]
298 taxa, 3 loci. **61 million trees**

By removing 12 of 298 original taxa, terrace size reduced from 61 million trees to **one** tree.

Sanderson, McMahon and Steel (2011). Terraces in phylogenetic tree space. *Science* 333: 448-450.

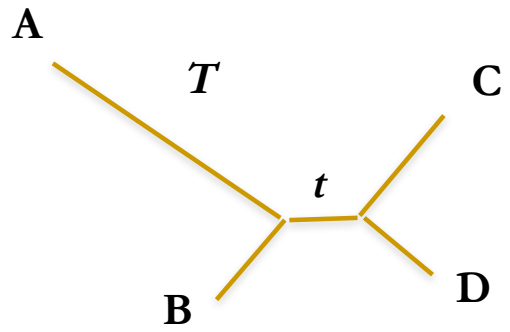
Overview: Stochastic models in phylogenetics



Models and statistical consistency

- Sites evolve i.i.d according to a continuous-time Markov process (rate at site may be sampled according to some distribution).
- Identifiability of tree from sites (from distribution on site patterns) holds under certain mild conditions (so ML consistent).
- With a 'molecular' clock, there is an easy proof of (general) statistical consistency of tree inference.

The impact of 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?

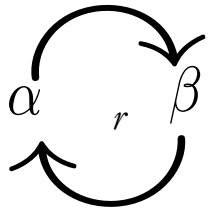
Infinite state model*

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

Finite state model

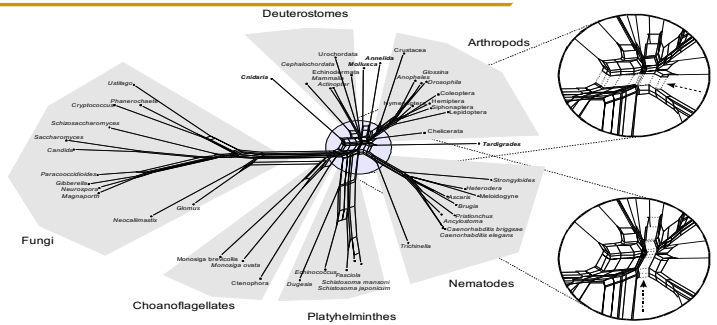
$$\text{as } t \to 0,\ k \text{ grows at rate } \frac{1}{t^2}$$

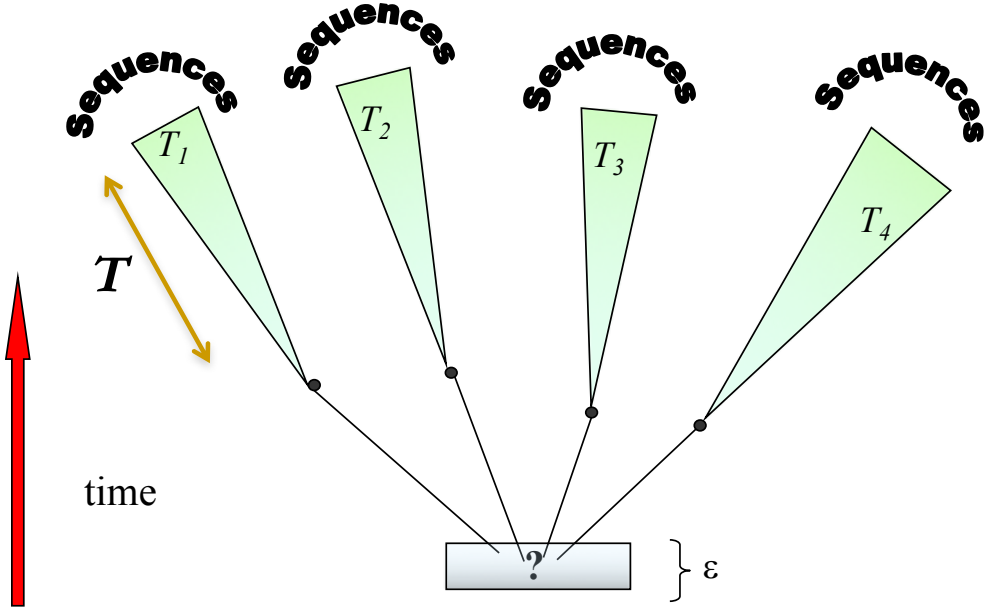
$$\text{but if } T = t \text{ then as } t \to 0,\ k \text{ grows at the rate } \frac{1}{t}$$



*Kimura's infinite alleles model

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?

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

Information-theoretic question:

- How many sites are needed to accurately reconstruct a tree?

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

$$= (r^n)^k = r^{nk}$$

$$b(n) = 2^{\Omega(n \log(n))}$$

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

Theorem (Infinite-alleles model)

If $\max\{p(e)\} = P < \frac{1}{2}, \min\{p(e)\} = p > 0$, then order $\frac{\log(n)}{p}$ characters suffice.

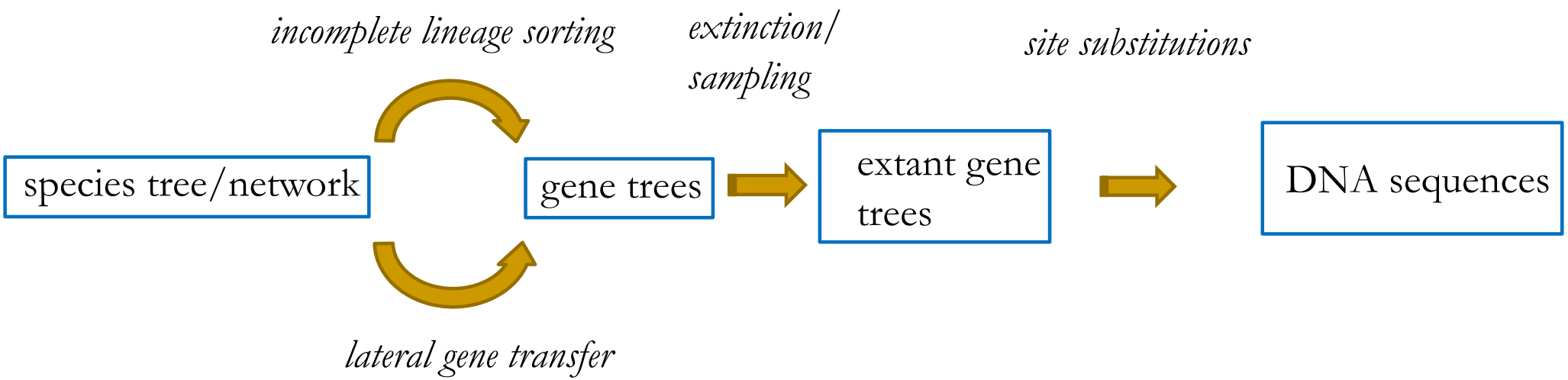
Theorem (2-state model)

If $\max\{p(e)\} = P < \frac{1}{2}\left(1 - \frac{1}{\sqrt{2}}\right), \min\{p(e)\} = p > 0$, then order $\frac{\log(n)}{p^2}$ characters suffice.

E. Mossel and M. Steel, *A phase transition for a random cluster model on phylogenetic trees*, Math. Biosci., 187 (2004), pp. 189–203.

C. Daskalakis, E. Mossel, and S. Roch, *Evolutionary trees and the Ising model on the Bethe lattice: A proof of Steel's conjecture*, Prob. Theor. Relat. Fields, 149 (2011), pp. 149–189.

Putting the pieces together

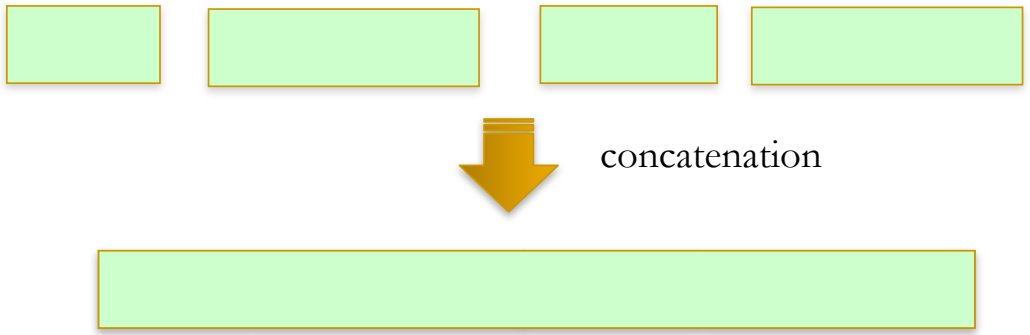


Sometimes controversial

Mirarab, S., Bayzid, M.S., Boussau, B, Warnow, T. (2014), Statistical binning enables an accurate coalescent-based estimation of the avian tree. *Science* 346, 1250463.

Liu L, Edwards S.V. (2015), Comment on "Statistical binning enables an accurate coalescent-based estimation of the avian tree". *Science* 2015; 350 :171.

Combining processes: ILS + site substitution



concatenation

Sequences for each gene (gene trees might be different because of ILS)

Transform (corrected) sequence dissimilarities and use these to build a 'best fit' tree

Apply MLE

Both involve model mis-specification

Statistically consistent estimator of tree topology

Statistically ***in***consistent estimator of tree topology

Dasarathy, Nowak, and Roch (2014)

Roch and Steel (2015) *Theor. Pop. Biol.*

Further details

- Allman, E.S., Long, C. and Rhodes, J. (2018). Species tree inference from genomic sequences using the log-det distance arXiv:1806.04974v1 [q-bio.PE].

- Roch, S., Nute, M. Warnow, T. (2018). Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. arXiv:1803.02800 [q-bio.PE].

- Mossel E. and Roch. S. (2015). Distance-based species tree estimation: information-theoretic trade-off between number of loci and sequence length under the coalescent, arXiv:1504.05289v2 [q-bio.PE].

- Steel, M. (2016). *Phylogeny: Discrete and Random Processes in Evolution*, SIAM