What can we reconstruct from the past?

- Joint work with…





Mareike Fischer Laszlo Szekely Elchanan Mossel Olivier Gascuel Elliott Sober

Biomathematics Research Centre

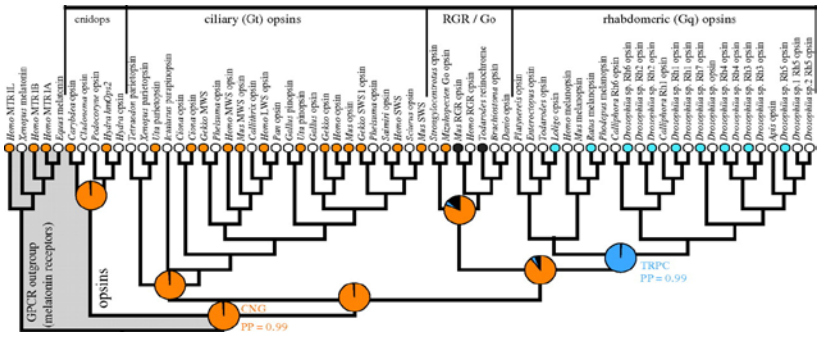
Montpellier, June 2014

ALLAN WILSON CENTRE

Outline of talk

- *Part 1:* **Ancestral state reconstruction**
- *Part 2:* **Information loss on trees**
- *Part 3:* **Reconstructing trees**
  - **20x pushups**
- *Part 4:* **Lateral gene transfer**

Part 1: Ancestral state reconstruction



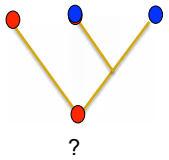
Plachetzki D C et al. Proc. R. Soc. B 2010;277:1963-1969

PROCEEDINGS OF THE ROYAL SOCIETY B

Three methods

Minimum evolution ('parsimony'):

Need tree topology but not branch lengths or model



Majority Rule

Don't even need tree

Maximum likelihood

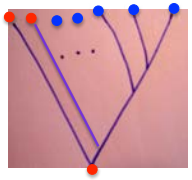
Need tree, branch lengths and model

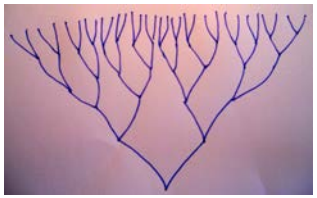
Definition:

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:

$\Pr(M(\text{leaf data}) = \text{state of } v)$

MP vs MR





h

$n = 2^h$

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

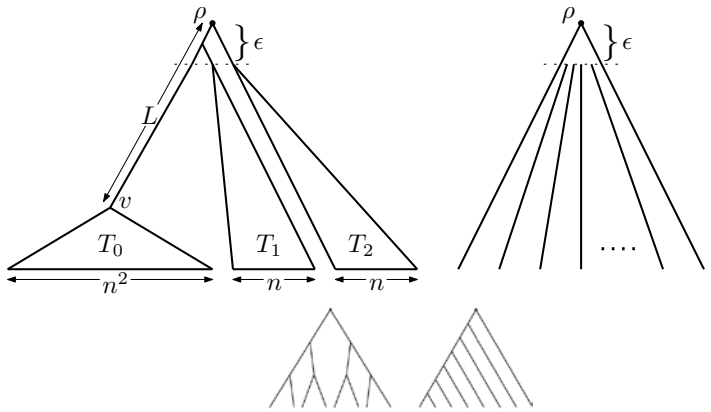
$$f_h = f_{h-1} + f_{h-2}$$

$$f_1 = 2, f_2 = 3$$

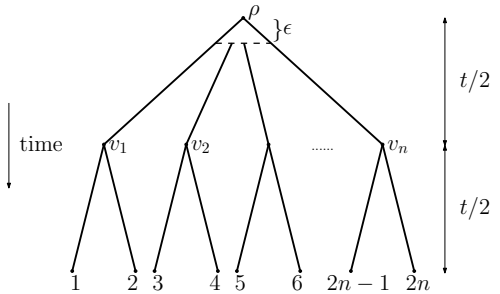
$$f_h \sim c\left(\frac{1+\sqrt{5}}{2}\right)^h$$

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

Q1: Which is more accurate for root state prediction from an 'evolved' character: parsimony or majority?

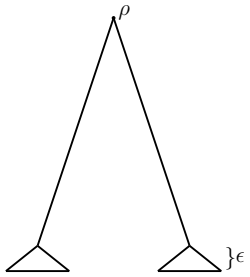


Q2. Is it easier to estimate the ancestral state at the root of the tree, or an interior node?



(a)

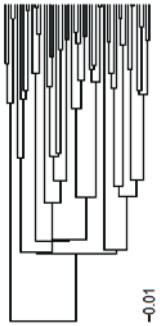
Root state can be estimated with **high** precision but **no** other node can be



(b)

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

Part 3: Random trees



Yule (pure birth) model

Each lineage gives birth independently at some constant rate λ

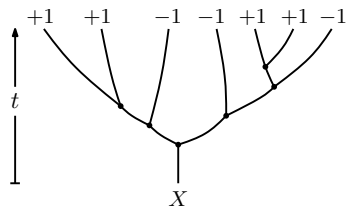
Grow for time t, or till it has n leaves, or condition on both n and t

[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.

So 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

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

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

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

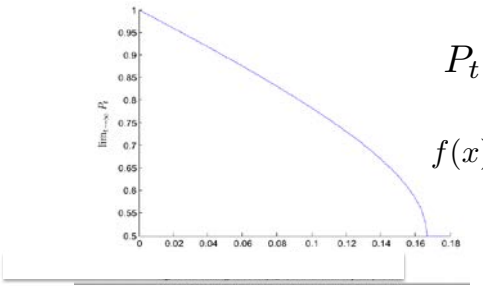
m = mutation rate (of states),
λ = birth rate (of tree)

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

'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

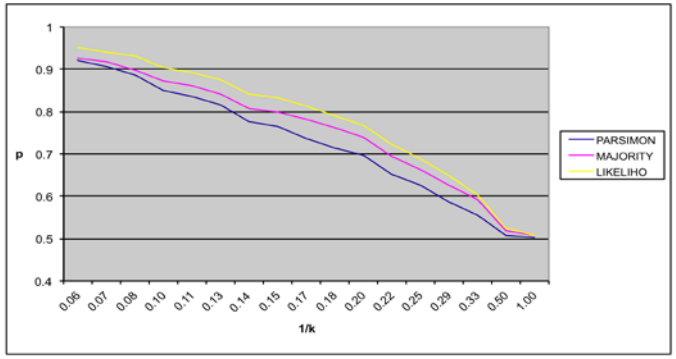


$$P_t \geq \lim_{t\to\infty} P_t = f(x) \text{ where}$$

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

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.

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]

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

All t

What can we say in general (for Yule trees)?

Assume **any** conservative GTR model on any number of states

- **Theorem 1:** The accuracy of any method in predicting the state at the **root node** of the tree vanishes (as *n* or *t* goes to infinity) when the mutation rate passes a threshold (dep. on the speciation rate).
- **Theorem 2:** There is a very simple method that can predict the state of a **randomly selected node** with an accuracy that does not vanish (as *n* or *t* grows) for any fixed mutation rate.

Part 2: Information loss on trees

- **Probability Primer:**
  - Let X and Y be any two discrete random variables.
  - The *mutual information* of (X,Y) is:

$$I(X;Y) = \sum_{x,y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$
$$= D_{\mathrm{KL}}(p_{XY} || p_X p_Y)$$

$$I(X;Y) = 0 \iff X \text{ and } Y \text{ are independent}$$

Fano's lemma

$$I(X;Y) << 1 \Rightarrow Y \text{ cannot accurately predict } X$$

by any method!

Markovian processes that destroy information (exponentially fast)

Theorem: For **any** finite state Markov chain, where a transition from any state to any other is possible in some fixed number of steps with probability $\geq p > 0$ then:

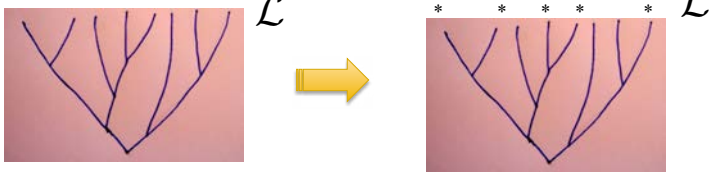
$$I(X_0; X_t) < Ce^{-ct}$$

- Transition rates can vary arbitrarily with time
- Theorem applies to a (1-dimensional) Markov **chain**, not for a Markov process on a (branching) tree (but applies to a path from root to leaf).

Data processing inequality

If $X \to Y \to Z$ (where $X \perp_Y Z$) then

$$I(X;Z) \leq I(X;Y) \text{ and } I(Y;Z)$$



$$X_\rho \to X_{\mathcal{L}} \to X_{\mathcal{L}'}$$

Thus, $I(X_\rho; X_{\mathcal{L}}) \geq I(X_\rho; X_{\mathcal{L}'})$

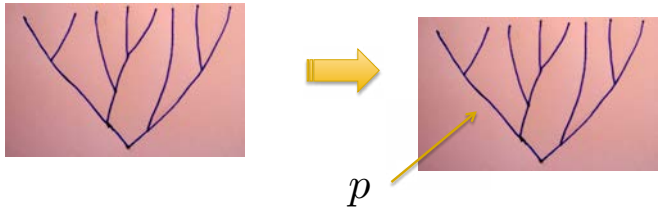
But! – MP can be more accurate on subsets of leaves of (an ultrametric tree) than on all leaves[1]

[1]Fischer, M. and Thatte, B. (2009). *Journal of Theoretical Biology* 260: 290--293.

A caution…

- Throwing away data never increases information (e.g. deleting fast evolving sites for tree estimation) – same for MLE under true model.
  - But when don't know the 'true' model (i.e. always except in simulations!) it can still (sometimes) be a good thing to do (to avoid correcting under an incorrect model)…

Application 2



$$\frac{d}{dp} I(X_\rho; X_{\mathcal{L}}) < 0$$

But! – MP can be more accurate when substitution probabilities increase[1]

[1] S. Roch (Pers. Comm).

Another property of I: subadditivity

If $(Y_1, Y_2, \ldots, Y_k)$ are conditionally independent given X then:

$$I(X; (Y_1, Y_2, \ldots, Y_k)) \leq \sum_{i=1}^{k} I(X; Y_i)$$

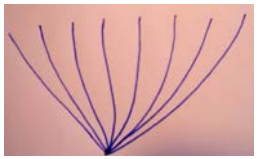
Example: Data = k characters (c_1, c_2,…) generated i.i.d. by an unknown tree topology T, with some prior on branch lengths.

$$I(T; \text{Data}) \leq kI(T; c_1)$$

Application

$X_{\mathcal{L}}$

(unzipping)



$Y_{\mathcal{L}}$

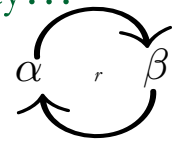
X_ρ $\qquad$ X_ρ

Subadditivity of I implies that $I(X_\rho; Y_{\mathcal{L}}) \leq \sum_{l \in \mathcal{L}} I(X_\rho, Y_l)$

But is $I(X_\rho; X_{\mathcal{L}}) \leq I(X_\rho; Y_{\mathcal{L}})$?

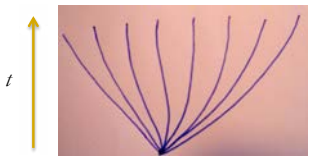
For 2-state symmetric model – yes! Evans et al. (2000)
For r-state symmertic models r>4 – no! Allan Sly, (2011).

Getting our hands (slightly) dirty…



- For the 2-state symmetric model

$$I(X_\rho; Y_l) = \exp(-4rt)$$



$I(X_\rho; X_{\mathcal{L}}) \leq I(X_\rho; Y_{\mathcal{L}})$ $\qquad$ $I(X_\rho; Y_{\mathcal{L}}) \leq \sum_{l \in \mathcal{L}} I(X_\rho, Y_l)$

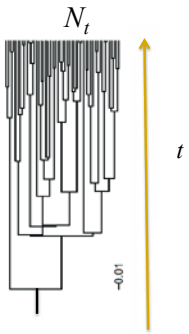
$$I(X_\rho; X_{\mathcal{L}}) \leq n \exp(-4rt)$$

So why was 4 best possible?…
[10 mins ago]

$$I(X_\rho, X_{\mathcal{L}}) \leq n \exp(-4rt)$$

$N_t \sim Geo(e^{-\lambda t})$

$\Rightarrow \mathbb{E}[N_t] = \exp(\lambda t)$

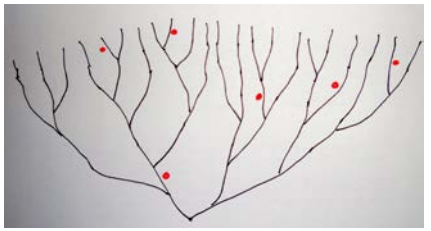


Other useful tricks: "Coupling" and percolation/branching process theory

Reconstructing a complete scenario:

a *curious* combinatorial result

$$D(\mathcal{E}) = \min\{d(e, e') : e, e' \in C\}$$

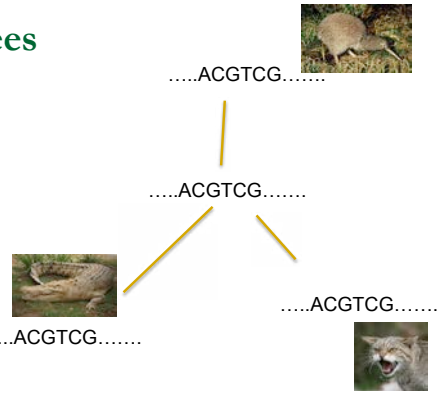


Theorem [S+Penny, 2005]

$D(\mathcal{E}) \geq 3 \Rightarrow$ MP reconstruction from leaf states is correct at every node of $T^{-\rho}$

Part 3: Reconstructing trees

Another combinatorial curiosity...



Theorem [Dress +S, 2005]

Suppose that :

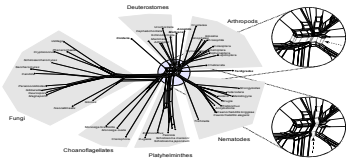
- *n* species evolved on some unknown tree *T*.
- for every three species, the sequence at the median node can be determined.
- the sequences are long enough that each edge of the tree has a transition in at least one site.

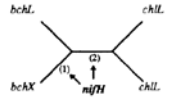
Then the unknown tree *T* can be reconstructed without error.

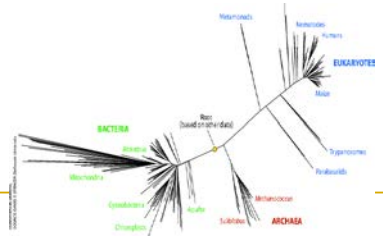
[Easy proof using 'symbolic ultrametric theory' of Boecker and Dress]

Examples of deep and controversial phylogenetic resolutions

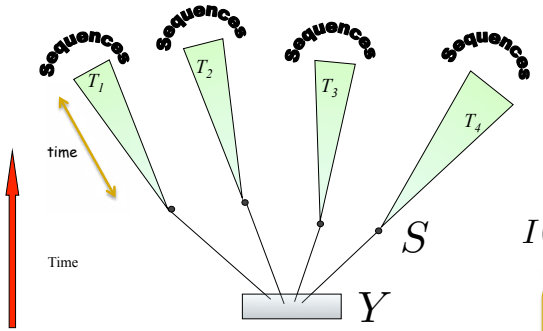
- Origin of metazoa (~550-600 mya)
- Origin of photosynthesis (>2.5 bya)
- Rooting the 'tree' of life (~3.5 bya)







Tree reconstruction: connection with information theory



$X_{\mathcal{L}}$

$Y \to S \to X_{\mathcal{L}}$

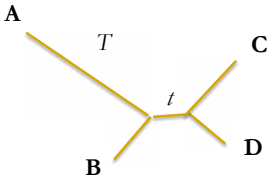
$I(Y; X_{\mathcal{L}}) \leq I(S; X_{\mathcal{L}})$

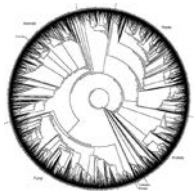
$$I(S; X_{\mathcal{L}}) \leq \sum_{i=1}^{4} I(S; X_{\mathcal{L}_i})$$
$$I(S; X_{\mathcal{L}_i}) \stackrel{k}{=} I(S_i; X_{\mathcal{L}_i})$$
$$I(S_i : X_{\mathcal{L}_i}) \leq \sum_{l \in S_i} \sum_{j=1} I(S_{lj}; X_{lj})$$

$k \geq \exp(4rt)/n$

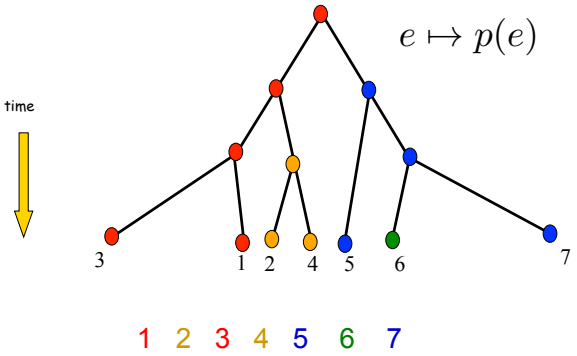
Problems for reconstructing a tree (even when the model is known and nice!)

- **Short interior edges**
- **Long edges**
- **Many taxa (*n*)**





Simplest model



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?

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

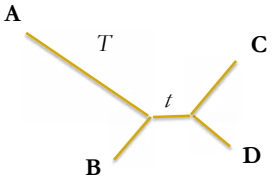
Theorem [Mossel +S, 2004]

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).

What about finite-state models, which will generate homoplasy?

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?

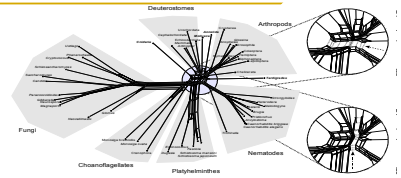
Infinite alleles model as $t \to 0$, k grows at rate $\frac{1}{t}$

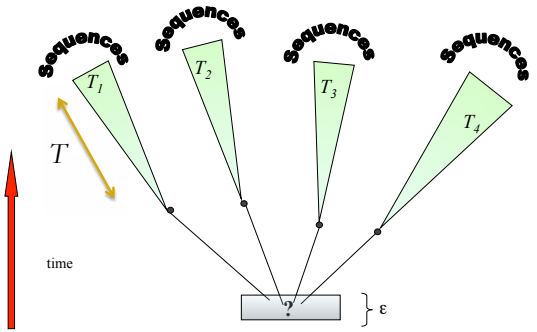
Finite state model

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 $\frac{1}{t}$

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?

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

Recent developments

- **Does 'rates across sites' help?**

- For long edges it can: for certain distributions instead of $k = \Theta(\exp(cT))$ it suffices to have $k = \Theta(T^{\gamma})$

[Martyn and S. (2012). JTB 314: 157-163.]

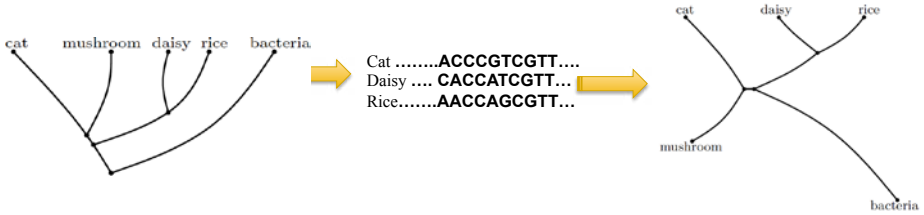
What about additional noise from lineage sorting?

- For short edges, the sequence length can still be kept at

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

Dasarathy, Nowak, Roch (2014). [Data requirement for phylogenetic inference from multiple loci: A new distance method. ArXiv: April 28, 2014:]

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



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

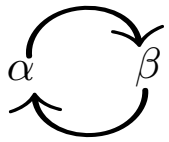
$$= (r^n)^k = r^{nk} \qquad b(n) = 2^{\Omega(n\log(n))}$$

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

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, Szkeley, 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)

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 and '$T = T_1$ or T_2', which is it?

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

Add taxa ➡ build tree ➡ ignore the added taxa

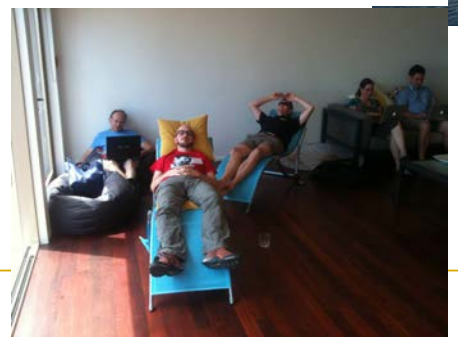
Sequence length required to find the correct tree (on the subset of species) can be reduced logarithmically this way

Intermission….

Portobello 2015

The Interface of Mathematics and Biology
The 19th Annual New Zealand Phylogenomics Meeting

Sunday 1st - Friday 6th February, 2015

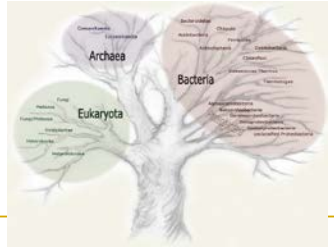


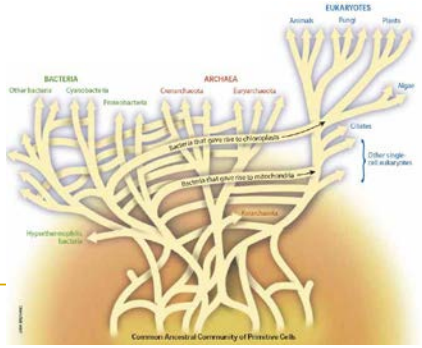
Part 4: Lateral gene transfer

Simone Linz

Daniel Huson

Andreas Pedersen





PNAS

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

- *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

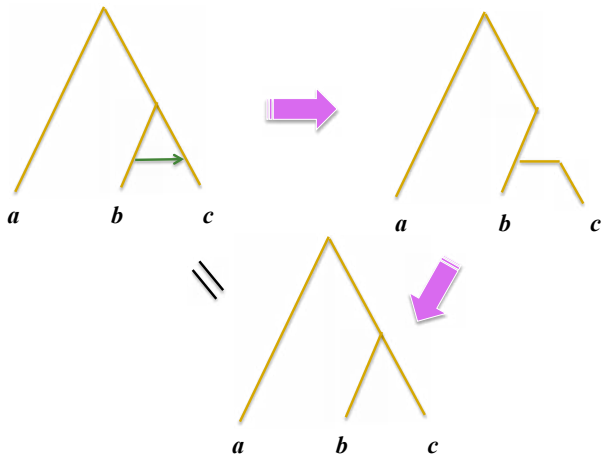
Gergely J. Szöllősi[a,b], Bastien Boussau[a,b,c], Sophie S. Abby[d,e], Eric Tannier[a,b,f], and Vincent Daubin[a,b,1]

PNAS

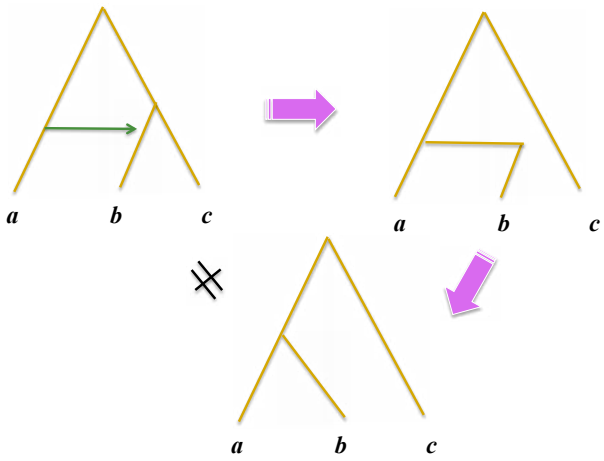
Lateral gene transfer as a support for the tree of life

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

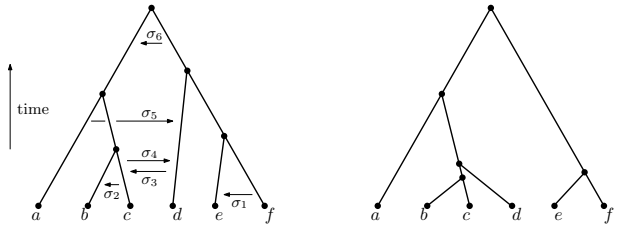
LGT



LGT



LGT combinatorics



(i)

(ii)

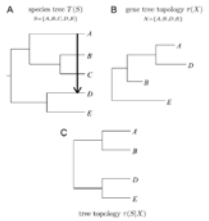
Suppose we have some 'species tree'
(e.g. the tree of bacterial cell divisions)

- *Can random LGT lead to statistical inconsistency in estimating this species tree from gene trees?*
- *When can we infer a species tree from gene trees?*

Possibilities for the LGT rates in the model:

A Likelihood Framework to Measure Horizontal Gene Transfer

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



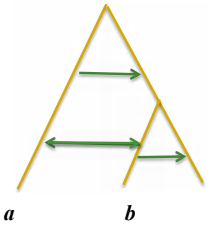
Rate of transfer from x to y is constant

Rate of transfer from x to y depends on the branches

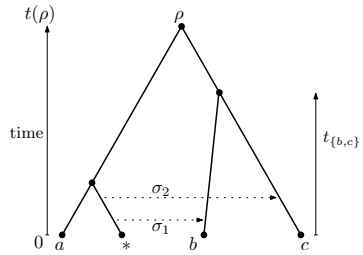
Rate of transfer from x to y depends on d(x,y) and/or time

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

Easy: Under model, 3-taxon tree reconstruction is statistically consistent if no other lineages are involved.

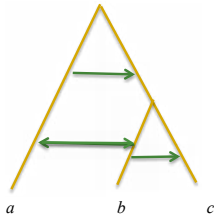


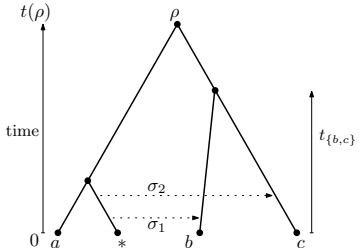
Harder: What about 3-taxon reconstruction if other taxa present?



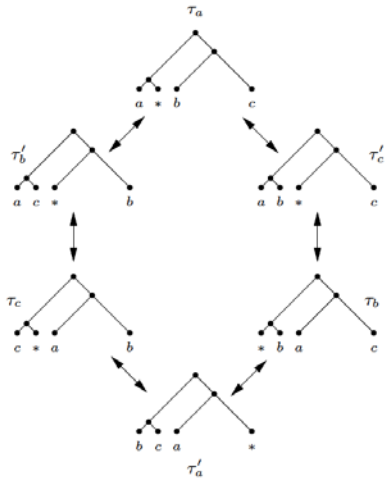
(*cf.* lineage sorting story – there extra lineages don't matter)?

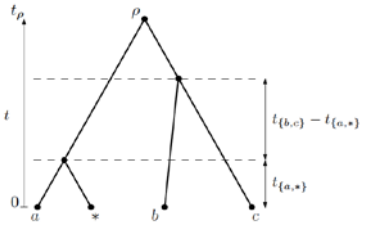
LGT: the gene tree for (a,b,c) ['fixing' and 'moving' transfers]





Walking between trees ("pass the port")

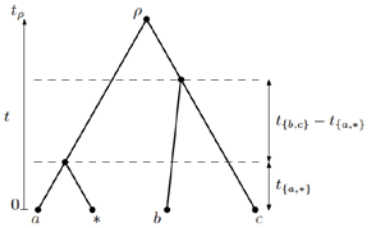




Theorem

$$\mathbb{P}(a|bc) = \frac{1}{3}\left[1 - e^{-7\mu}(1 - e^{-2\mu} - e^{-B}(1 + e^{-2\mu}))\right]$$
$$\mathbb{P}(b|ac) = \mathbb{P}(c|ab)$$
for $\mu = \frac{1}{3}\lambda t_{\{a,*\}}$, *and* $B = 3\lambda(t_{\{b,c\}} - t_{\{a,*\}})$

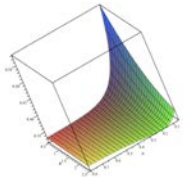
A zone of inconsistency



Corollary

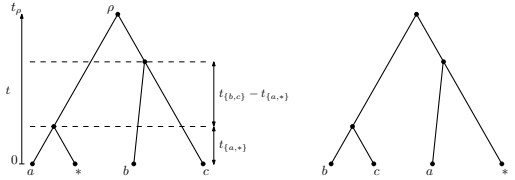
$$\mathbb{P}(a|bc) < \mathbb{P}(b|ac) = \mathbb{P}(c|ab)$$
if and only if $t_{\{b,c\}} - t_{\{a,*\}}$ *is greater than*
$\frac{1}{3\lambda}\ln(\frac{1+e^{-2\mu}}{1-e^{-2\mu}})$ $\quad \mu = \frac{1}{3}\lambda t_{\{a,*\}}$

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



$$B = 3\lambda(t_{bc} - t_{a*})$$

What about the other tree?



(i)

(ii)

More precisely for τ'_a *and* $\mu = \frac{1}{3}\lambda t_{\{b,c\}}$, *and* $B = 3(\lambda t_{\{a,*\}} - t_{\{b,c\}})$, *we have:*

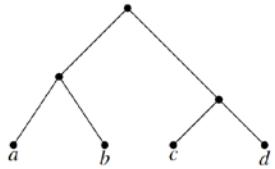
$$\mathbb{P}(a|bc) = \frac{1}{3}\left[1 + e^{-7\mu}(1 + e^{-2\mu} - e^{-B}(1 - e^{-2\mu}))\right]$$

and in this case

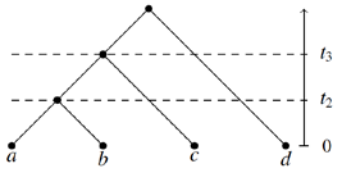
$$\mathbb{P}(a|bc) > \mathbb{P}(b|ac)(= \mathbb{P}(c|ab))$$

for all values of $t_{\{a,*\}}$ *and* $t_{\{b,c\}} - t_{\{a,*\}}$.

The other tree shape

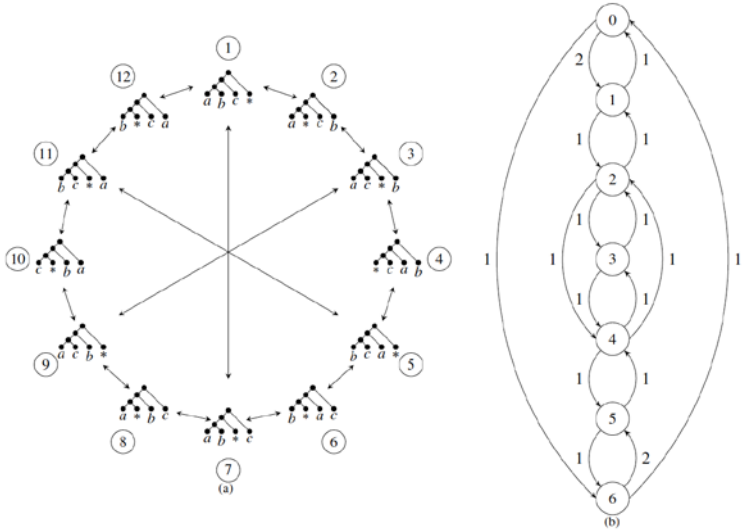


(a) The fork-shaped four-taxon tree topology.



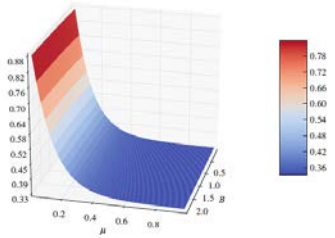
(b) The pectinate four-taxon tree topology $(ab; c; d)$

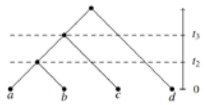
The other tree shape



53

The other tree shape





No statistical inconsistency arises for this tree shape

54

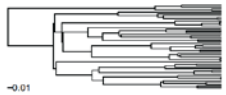
Can we reconstruct a tree under rampant LGT?

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 suffices:

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

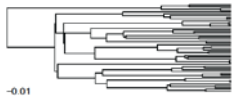


Particular case: [S,Linz, Huson, Sanderson]

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.

55

Can we reconstruct a tree under rampant LGT?



Theorem 1 [Roch and Snir, 2013]

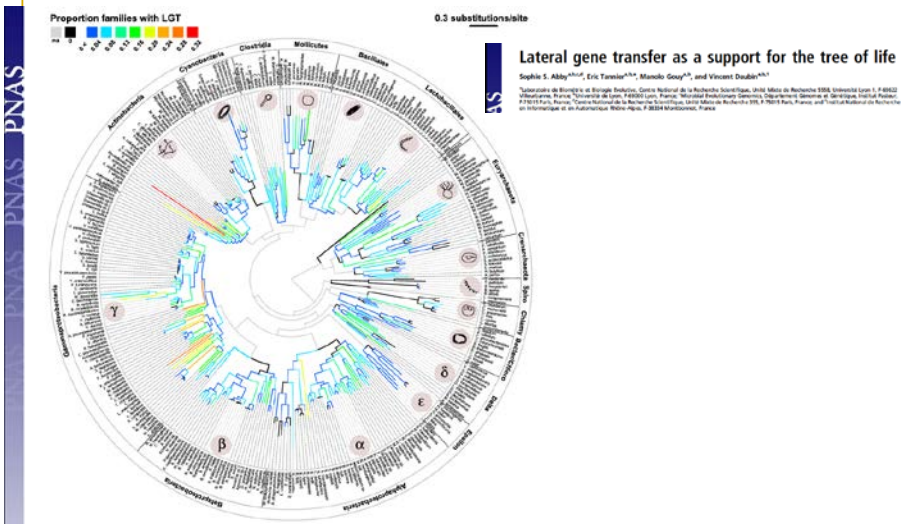
Under the bounded rates (e.g. Yule model), it is possible to reconstruct the topology of a phylogenetic tree for n taxa w.h.p. from $N = \Omega(\log(n))$ gene tree topologies if the expected number of LGT transfers is no more than a constant times $n/\log(n)$.

Theorem 2 [Roch and Snir, 2013]

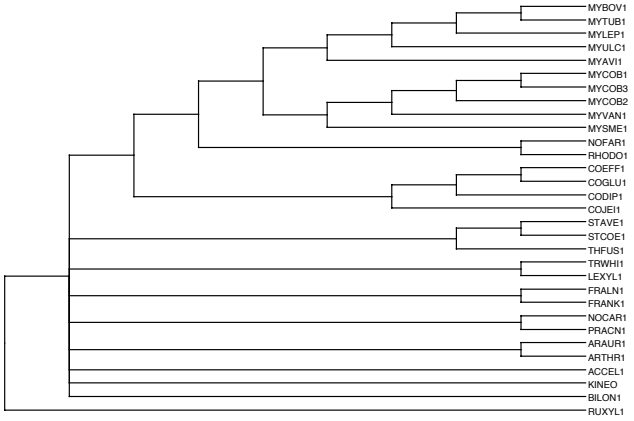
Under the Yule model, it is not possible to reconstruct the topology of a phylogenetic tree w.h.p. from N gene trees if the expected number of LGT events is more than $\Omega(n \log(N))$

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.

56



Example 1



1338 rooted gene trees on variable taxon sets from the Actinobacteria phylum.

"Primordial tree" in Dendroscope

Further details

Gascuel, O. and Steel, M. (2010). Inferring ancestral sequences in taxon-rich phylogenies. *Mathematical Biosciences*, 227: 125-135.

Gascuel, O. and Steel, M. (2014). Predicting the ancestral character changes in a tree is typically easier than predicting the root state. *Systematic Biology*, 63(3): 421-435.

Mossel, E. and Steel, M. (2014). Majority rule has transition ratio 4 on Yule trees under a 2-state symmetric model (ArXiv, April 2014).

Steel, M., Linz, S., Huson, D. and Sanderson, M. (2013) Identifying a species tree subject to random lateral gene transfer. *Journal of Theoretical Biology*, 322: 81--93

Sand, A., and Steel, M. (2013). The standard lateral gene transfer model is statistically consistent for pectinate four-taxon trees. *Journal of Theoretical Biology*, 335: 295-298.

Thanks to:

the Royal Society of New Zealand

Allan Wilson Centre