Beyond 'event horizons' in early evolution

Mike Steel

David Penny

Joint work with:

Mareike Fischer

Elchanan Mossel

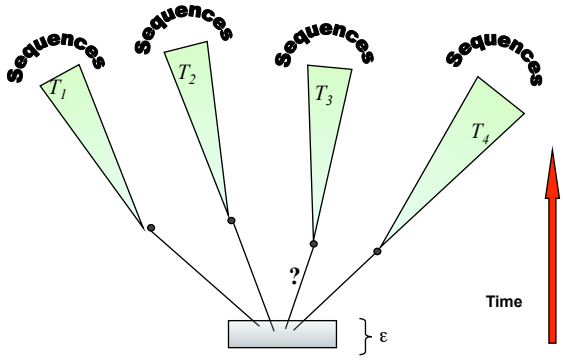
Laszlo Szekely

ALLAN WILSON CENTRE

Biomathematics Research Centre

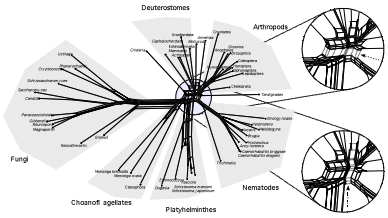
Montpellier, June 10, 2008

Difficult phylogenetic problem

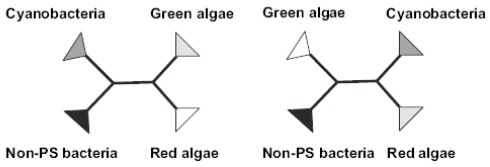


Bushes in the tree of life.
A.Rokas, S.B. Carrol,
Plos Biol. (2006).

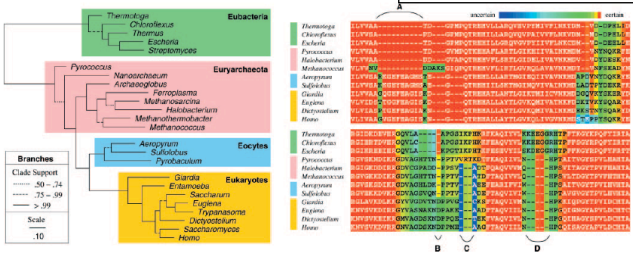
Difficult phylogenetic problem



From Huson and Bryant, Applications of phylogenetic networks in evolutionary studies, MBE. 2006



Lockhart et al. , Heterotachy and tree building, a case with plastids and eubacteria. MBE. 23, 2006



Suchard and Redelings, 2006 (Bioinformatics 22)

3

It's a jungle out there…

- Confounding processes
  - (lineage sorting, alignment error, etc etc etc)
- Model misspecification
- Not enough data
- Non-identifiability

4

Models

Markov (finite-state)

∪

Stationary reversible

$P^e = \exp(Rl_e)$

Mixtures of Markov (finite-state)

$$p_s = \sum_{i=1}^{r} \alpha_i p^i{}_s$$

∪

Arbitrary mixtures (heterotachy)

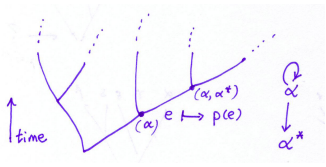
∪

Clocklike mixtures

∪

Rates-across-sites, covarion drift

Random-cluster model

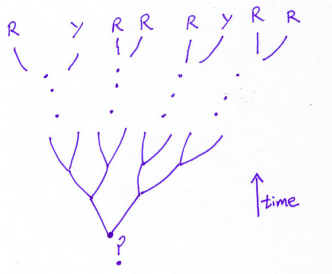


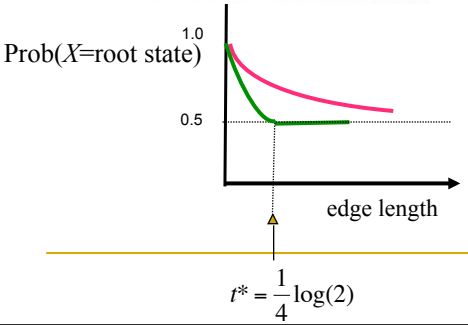
Homoplasy-free data
Mixtures behave similarly

5

Information loss

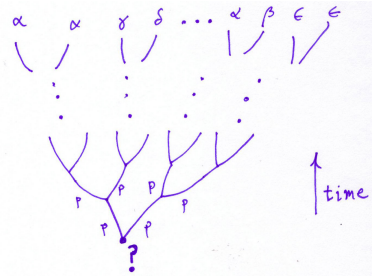
Finite state Markov model

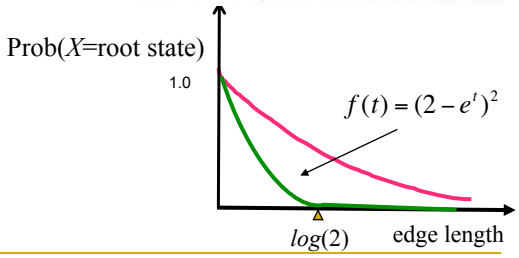




$t^* = \frac{1}{4}\log(2)$

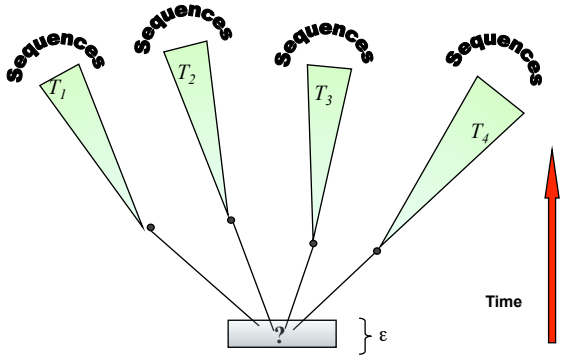
Random cluster model





6

Difficult phylogenetic problem



Let k = sequence length required to resolve the divergence under for i.i.d. sites.

$$k \propto \frac{1}{\varepsilon^2}$$

Finite-state Markov process

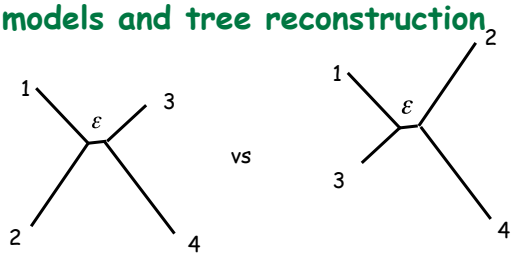
$$k \propto \frac{1}{\varepsilon}$$

Random cluster process

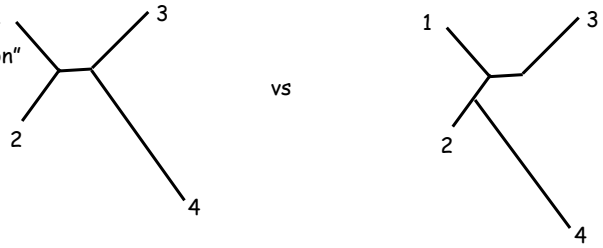
Steel, M., Szekely, L., 2002. SIAM J. Discrete Math 15(4)

Mossel, E., Steel, M., 2004. Math. Biosci. 187, 189-203.

Markov models and tree reconstruction



- "site saturation"



Putting the two together! And for more general models

How many sites required to resolve this basic tree?

Time

- Saitou, N., Nei, M., 1986. J. Mol. Evol. 24, 189-204
  The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence.
- Churchill, G., von Haeseler, A. Navidi, W., 1992. Mol. Biol. Evol. 9(4), 753-769.
  Sample size for a phylogenetic inference.
- Lecointre G, Philippe H, Van Le HL, Le Guyader H., 1994. Mol. Phyl. Evol. 3(4), 292-309.
  How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences.
- Yang, Z., 1998. Syst. Biol. 47(1), 125-133.
  On the best evolutionary rate for phylogenetic analysis.
- Wortley, A.H., Rudall, P.J., Harris, D.J., Scotland, R.W., 2005, How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. Syst. Biol. 54(5), 696—709.
- Townsend, J., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56(2), 222-231.

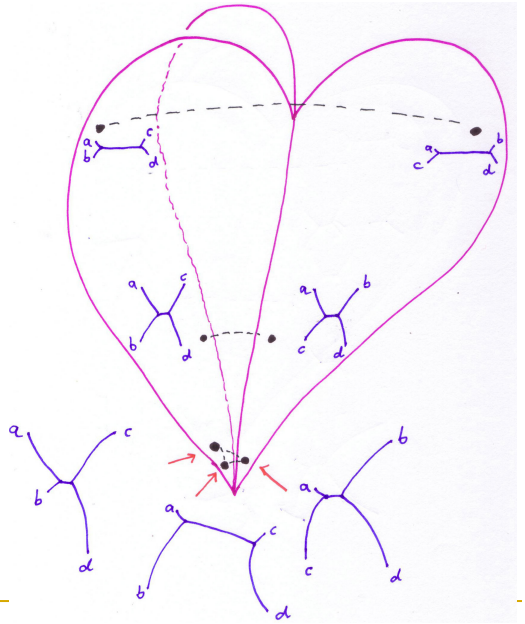
9

(Markov) tree space

- What metric to use?

$$d_H = \sqrt{\sum_{s \in S} (\sqrt{p_s} - \sqrt{q_s})^2}$$

$$d_H^2 = 2\left(1 - \sum_{s \in S} \sqrt{p_s \cdot q_s}\right)$$



10

Fundamental fact:

- To correctly identify (w.p. >1-ε) each of two possible competing hypotheses from k i.i.d. observations of data (of anything, by any method) requires:

$$k \geq \frac{(1-2\varepsilon)^2}{4} \cdot d_H^{-2}$$

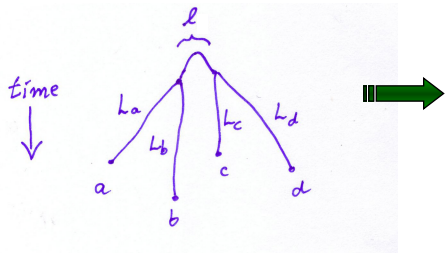
$$H_1 : p_H = \frac{1}{2} + \varepsilon$$
$$H_2 : p_H = \frac{1}{2} - \varepsilon$$

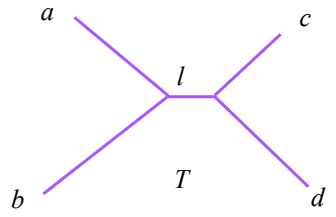
$$H_1 : p_H = \varepsilon$$
$$H_2 : p_H = \varepsilon^2$$

d_H = Hellinger distance between the probability distributions (on a single observation) under the two hypotheses.

11

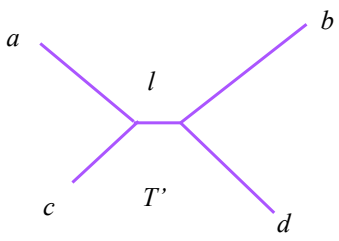
Application (for any Markov process on any state space)





- Proposition [F+S, 08]

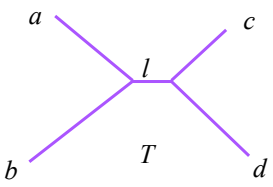
$$d_H^{\,2}(T,T') \leq l^2 \cdot \left(\sum_{s \in S} \frac{D_s^{\,2}}{p_s} \right)$$

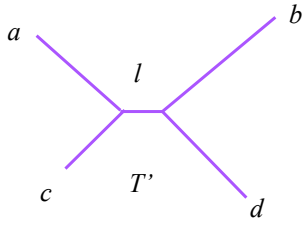


$$p_s = P(s \mid T)$$
$$D_s = P(s \mid T, N > 0) - P(s \mid T', N > 0)$$

12

So…





$$k \geq \frac{(1-2\varepsilon)^2}{4} \cdot d_H(T,T')^{-2} \qquad d_H(T,T')^2 \leq l^2 \cdot \left(\sum_{s \in S} \frac{D_s^2}{p_s} \right)$$

$$k \geq \frac{(1-2\varepsilon)^2/4}{l^2} \cdot \left(\sum_{s \in S} \frac{D_s^2}{p_s} \right)^{-1}$$

13

$$p_s = P(s \mid T)$$
$$D_s = P(s \mid T, N > 0) - P(s \mid T', N > 0)$$

Theorem [F+S, 08]: For 'nice' models*

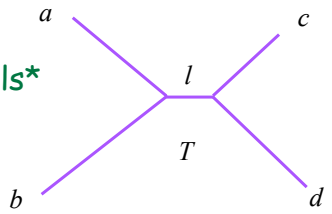
If $L_i \geq L \geq L_0$

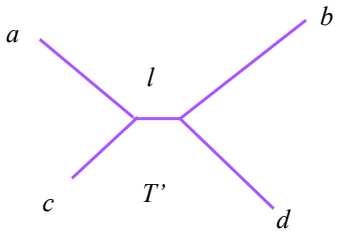
then

$$|D_s| \leq Ae^{-cL}, \; p_s \geq g(L_0)$$

$$k \geq C \cdot \frac{e^{2cL}}{l^2}$$

$$k \geq \frac{(1-2\varepsilon)^2/4}{l^2} \cdot \left(\sum_{s \in S} \frac{D_s^2}{p_s} \right)^{-1}$$





*Finite-state, stationary, time-reversible, irreducible

14

Extension to rates-across-sites models

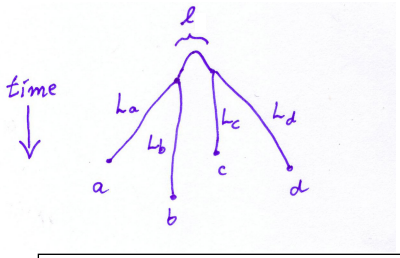
Recall $$d_H(T,T')^2 \le l^2 \cdot \left(\sum_{s\in S} \frac{D_s^2}{p_s}\right)$$

For p=RAS mixture on T, p'= RAS mixture on T' $$d_H(p,p')^2 \le \frac{3}{2} E[l^2 \cdot \left(\sum_{s\in S} \frac{D_s^2}{p_s}\right)]$$

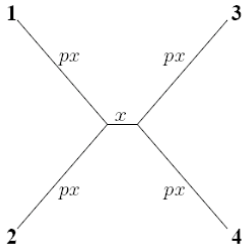
$$k \ge \frac{(1-2\varepsilon)^2}{6} \cdot E\left[\frac{1}{l^2}\left(\sum_{s\in S} \frac{D_s^2}{p_s}\right)\right]^{-1}$$

15

Bounds independent of rates? (fast-genes/slow genes)



$$\frac{L_i}{l} = p$$



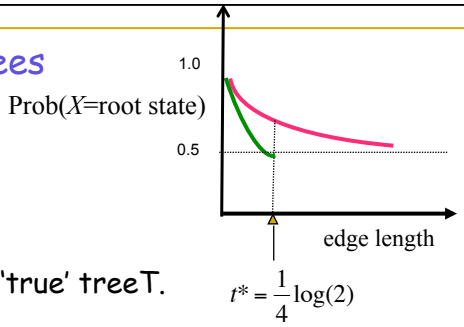
Theorem [F+S, 08]: For 2-state symmetric model

$$k \ge \frac{1}{2}(1-\frac{3}{2}\varepsilon)^2 \cdot p^2$$

Moreover, $k = c \cdot p^2$ can be achieved with MP ($x = 1/4p$)

16

Reconstructing large trees



- **Reconstructing:**
  - Given seq. data find the 'true' treeT.
  - k = c. log(n) can suffice for some models with 'nice' branch lengths (in fixed interval [f,g] independent of n).

If tree evolves under a constant rate Yule speciation process it is likely that sequence length required will grow at rate at least n^2.

17

Is 'testing' a tree, easier than finding it? (stochastic analogue of P=NP)

Reconstructing: Given data find tree
Testing: Given data and tree, did the tree produce data?

[Mossell, Steel, Szekely 2008]

Theorem 1: For finite-state models, testing requires the same order of data (log(n)) for testing as reconstructing.

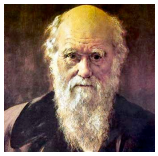
Theorem 2: For the random-cluster model (homoplasy-free) it is possible to test with a fixed (!) number of characters, independent of n (assuming $t_e<\log(2)$).

TEST: Given $c_1, c_2,\ldots,c_k$ and T --- is each character homoplasy-free on T?

If YES, T passes, if NO, T fails. Probability of error?

18

The end (almost)….



***Further information*: Sequence length bounds for resolving a deep phylogenetic divergence. M. Fischer, and M. Steel, 2008 (submitted) available at arXiv:0806.2500**

The 13th Annual NZ Phylogenetics Conference
7-12th Feb. 2009, Kaikoura

Cass Workshop 09

`Wild ideas' in theoretical evolutionary biology
21 Feb-28 Feb, 2009

http://www.math.canterbury.ac.nz/bio/events/

19