Deep divergences: Can we win the 'war on error'?

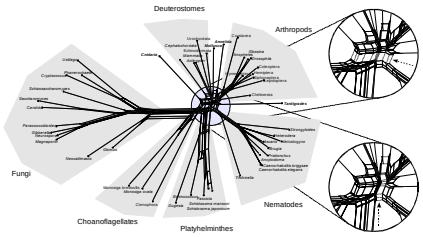
Mike Steel
Allan Wilson Centre for
Molecular Ecology and Evolution
Biomathematics Research Centre
University of Canterbury,
Christchurch, New Zealand

Joint work with **David Penny**

Evolution, Stony Brook, June 2006

What is it?

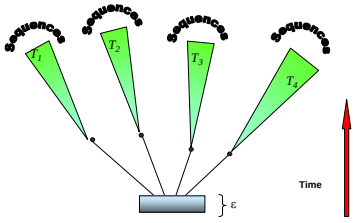
Example: Deep divergence in the Metazoan phylogeny



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

Coelomata vs Ecdysozoa

To a mathematician



Does it matter?

The 'war on error'

- the errorists are everywhere…

(site saturation, lineage sorting*, model violation, HGT, homoplasy, mis-alignment, systematic error…)

- weapons of math instruction (WMIs)

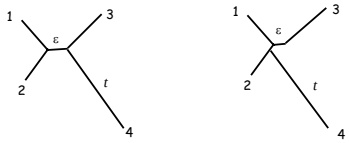
probability theory, combinatorics, algebra

- the "coalition of the D.I.M.B.O.s"

(dynamic interaction: mathematicians and biologists)

*Degnan and Rosenberg, PLoS Genetics 2006

Problem with using DNA sequence models ---



Weak signal (short interior edge) gets swamped by the noise of 'site saturation' of long pendant edges.

How many sites k are needed to resolve such a deep divergence ?

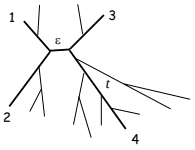
WMI 1: Probability theory

Theorem 1

$$k \propto \frac{1}{\varepsilon^2} \times e^{rt}$$

Follows from Mossel+S 2005

Bound applies for **any** method



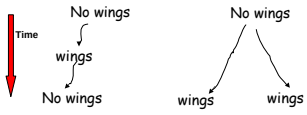
Increased taxon sampling?

Other problems with using site substitution models for resolving deep divergences:

model mis-specification or overparameterization, variation of the process across sites or tree (heterotachy), alignment error etc.

An old idea for new types of data: low-homoplasy characters.

- What is homoplasy-free character evolution?
  - Each state arises just once in the tree.
    or, equivalently,
  - No **reversal** or **convergent evolution**



h = homoplasy (underlying) of data on a tree;

Example 1: Large state space

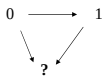
- Eg. gene order rearrangements (n species, L genes, random inversion model)

$$g_1 g_2 \boxed{g_3 g_4 g_5} g_6 g_7 \cdots \Rightarrow g_1 g_2 g_5 g_4 g_3 g_6 g_7 \cdots$$

$$P[h=0] \geq 1 - \frac{2(2n-3)(n-1)}{L(L-1)}$$

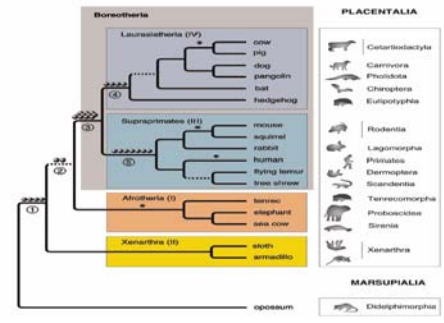
Example 2: Rare genomic changes

- Example: retroposons, SINEs, LINEs, LTRs etc
- Model



Recent example:

Kreigs *et al.* PLoS biology, April 2006. Tree of placental mammals



[Consider a large state space, with probability of state change on each edge for each character is bounded between (a,b), $0<a<b<1/2$]

Theorem [Mossel +S, 2005]

The number k of indep. characters required to reconstruct T (correctly with probability> 1 - δ) is

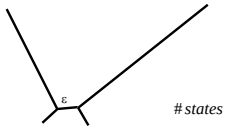
$$k = c \cdot \frac{\log(n)}{\varepsilon}$$

- n = #species, ε =smallest substitution probability, $c = c(\delta,b)$
- The tree reconstruction algorithm is polynomial time (in n, k)
- The model can vary from character to character (NCM)

What the model shows about this type of data

- Can allow different processes across characters
- Instead of $1/\varepsilon^2$ dependence it's $1/\varepsilon$
- Multi-state characters more informative and more robust to saturation

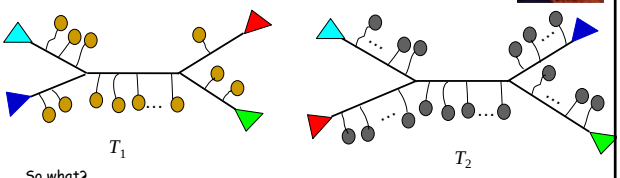
Can we avoid LBA?



$$\#states \geq \frac{c}{\varepsilon}$$

WMI 2: combinatorics

- Theorem [Szekely+S 05] Given ***any*** two different trees T_1, T_2, we can always represent them like this:



So what?

- The number of characters required to accurately 'test' if a given tree is the true tree or not is independent of the # of taxa.

WMI 3: algebra

- For multistate model can we use distances-based methods (NJ)?

Let $d_C(i,j)$ = # characters in C on which i and j differ

If C is homoplasy-free is d_C tree-like?
C binary – yes.
C non-binary – no; and it doesn't escape LBA!

Theorem [Huson and S, 2004]:

For any two trees T_1, T_2 there is a set of multi-state characters C such that

- C *is* homoplasy-free only on T_1 yet
- d_C is tree-like (and satisfies a mol. clock!), but only on T_2.

Moral: for multistate: "distances bad, characters good"

The end

Thanks to:

Allan Wilson Centre
NZ Marsden Fund

Further details:

E. Mossel and M. Steel, A phase transition for a random cluster model on phylogenetic trees. Mathematical Biosciences, 187 (2004), 189-203.

D. H. Huson and M. Steel, Distances that perfectly mislead, Systematic Biology 53(2): 327--332.

M. Steel and L. Szekely, Teasing apart two trees. Submitted to Combinatorics, Probability and Computing.

D. Penny and M. Steel, Deep divergences: can we win the `war on error' (in prep.).