

## CASE 3: PHYLOGENETIC ANALYSIS

Thanks Bret Larget.

Picture of cichlids  
Table of cichlid data  
Some trees

## COMMON METHODS OF TREE RECONSTRUCTION

- Maximum Parsimony.  
(Minimize the number of nucleotide base substitutions.)
  
- Neighbor Joining.  
(Use a clustering algorithm from pair-wise distances.)
  
- Maximum Likelihood.  
(Find the tree for which the observed data is most probable.)

## STANDARD BASE SUBSTITUTION MODELS

- For one site, along one branch of length  $t$ 
  - continuous time Markov chain over 4 states
  - rate matrix  $Q$
  - transition probabilities  $P(t) = \exp(Qt)$
- Conditional independence at branches
- Independence among sites
- Likelihood via *pruning* algorithm and ML by heuristics (Felsenstein, 1981).
- Many variations on  $Q$  etc; see, e.g., see Goldman et al. 1990
  - rate variation across sites; covarion rate changes...
- Extensive code

Markov Random Field perspective

## BOOTSTRAPPING

Felsenstein 1985.

## BOOTSTRAPPING WORKS (THEORETICALLY!)

Newton 1996.

Consider data  $D_n$  from  $n$  iid sites.

Let  $\tau$  be any incorrect tree topology.

Under weak regularity conditions,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P(\hat{\tau}_n = \tau) = c(\tau) < 0$$

and

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log P(\hat{\tau}_n^* = \tau | D_n) = c(\tau) \quad a.s.$$

(plus results on bias...)

## ISSUES

- Computational: repeated optimizations
- Conceptual: if it's Bayesian, why not just do an explicit Bayesian analysis? (e.g. Efron and Tibshirini, 1998)

## BAYESIAN APPROACH

Put prior on state  $S = \{\text{phylogeny, substitution parameters}\}$   
and then compute posterior summaries via MCMC, sampling states  
 $S_1, S_2, \dots$  aimed at target  $P(S|D)$ .

Issues: tree representation, move types, prior...

Rannala and Yang 1996 *JME*

Mau and Newton 1997 *JCGS*

Yang and Rannala 1997 *MBE*

Mau, Newton and Larget 1999 *Biometrics*

Larget and Simon 1999 *MBE*

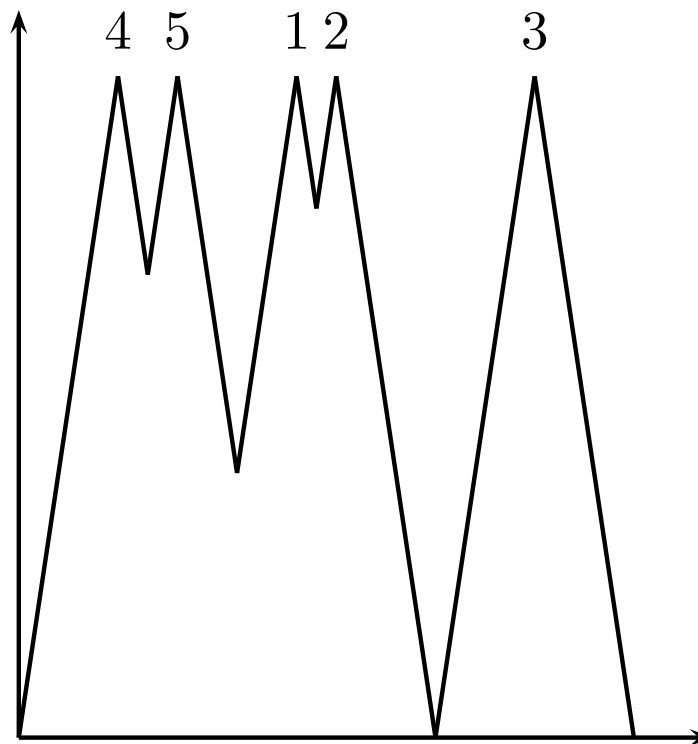
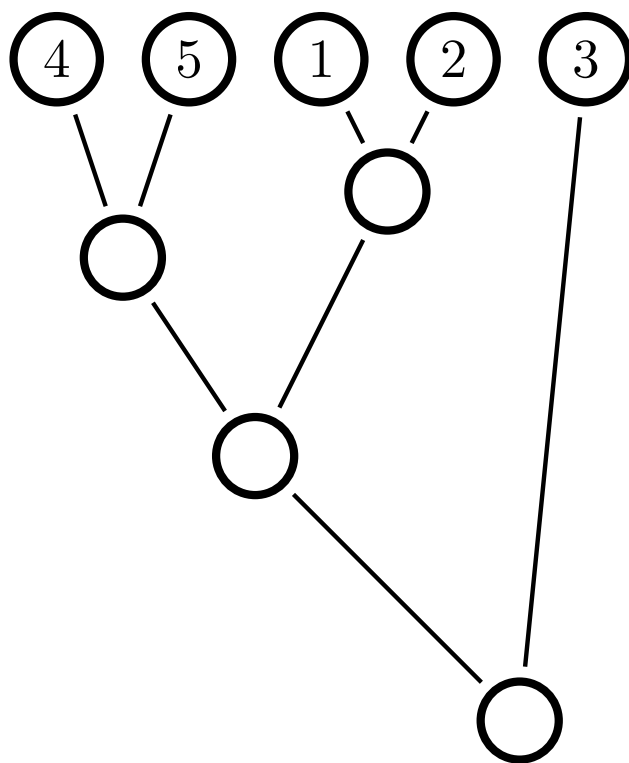
Huelsenbeck et al. 2001 *Science*

Huelsenbeck and Ronquist 2001 *Bioinformatics*

Huelsenbeck et al. 2002 *Sys. Bio.* (good review)

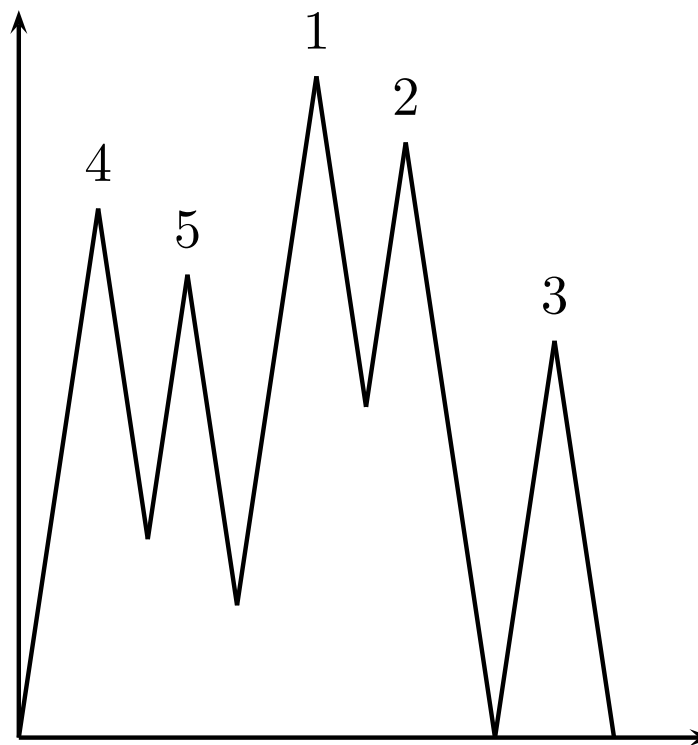
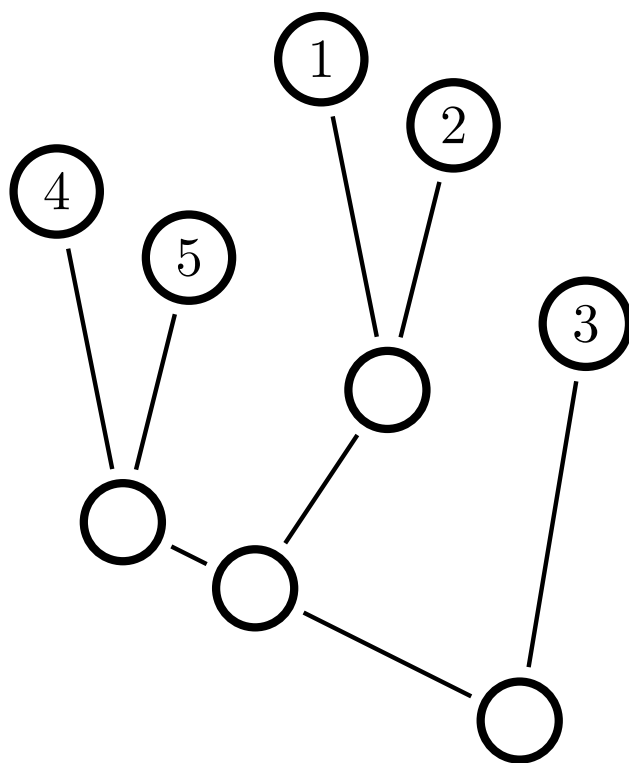
## A TREE REPRESENTATION

Every rooted tree with ordered children may be represented uniquely by a graph of the within-tree distance from the root generated by a depth-first traversal of the tree.



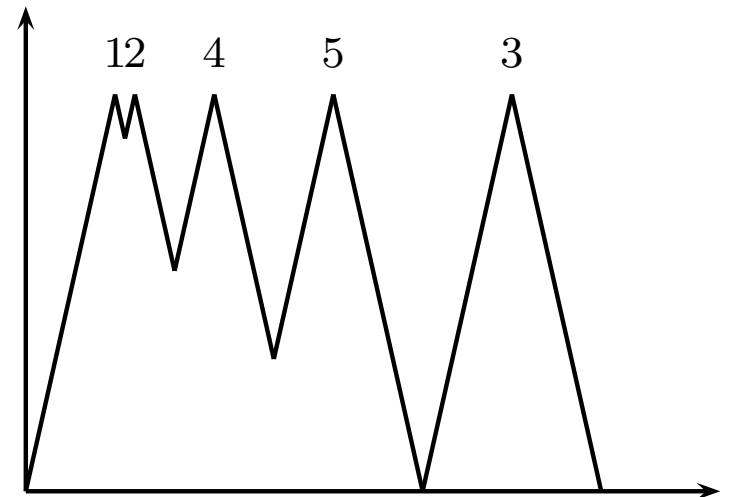
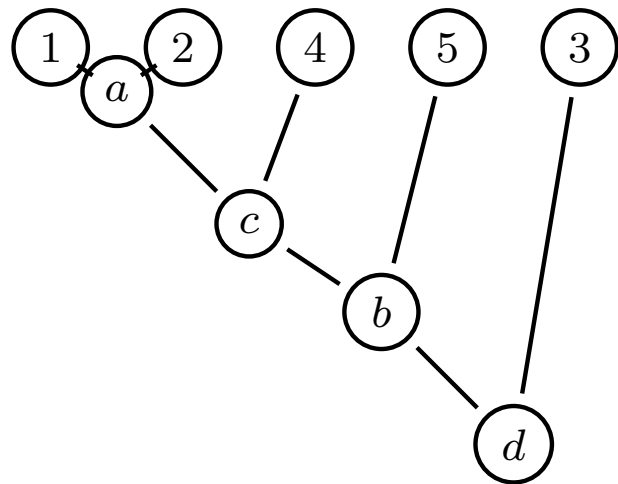
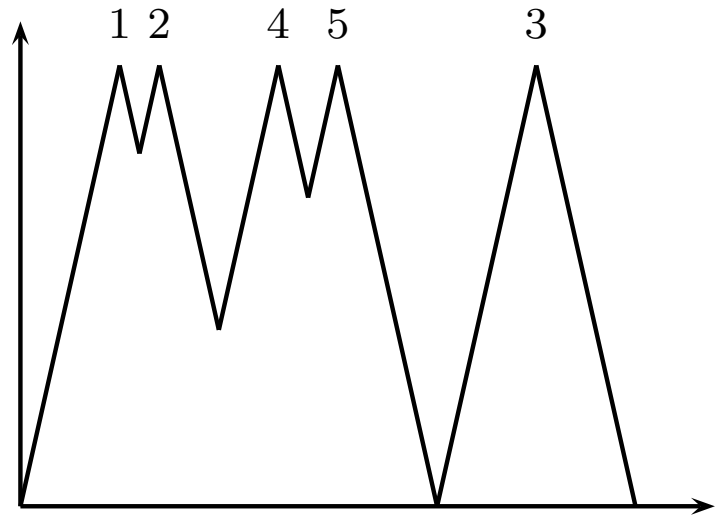
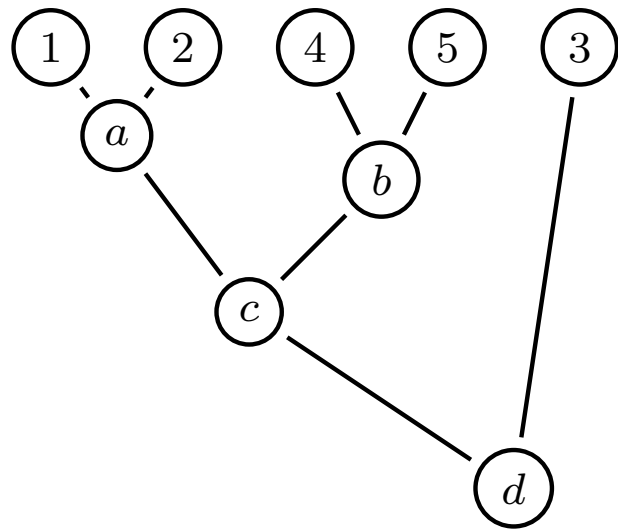
## AN EXAMPLE WITHOUT THE MOLECULAR CLOCK

Each valley in the traversal profile has a left depth and a right depth.

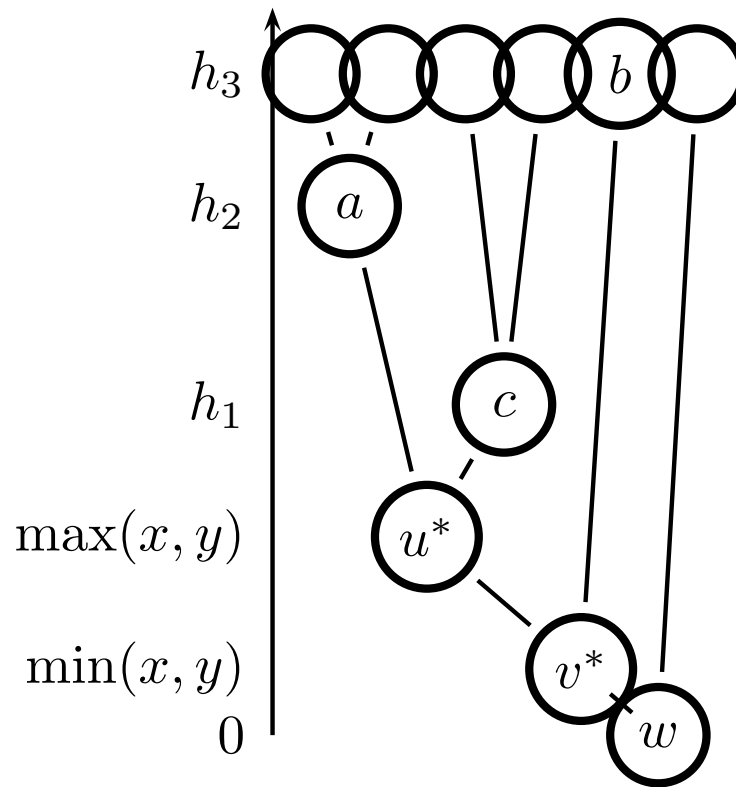
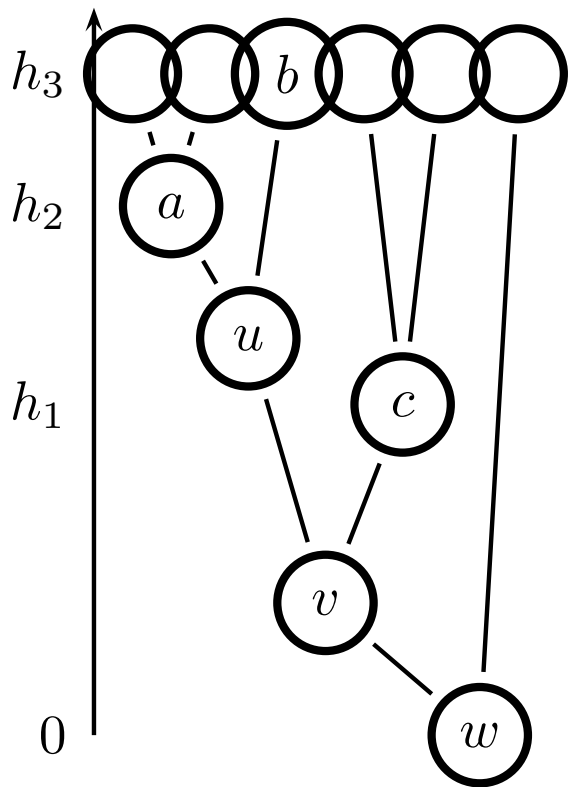


## THE GLOBAL TREE UPDATE

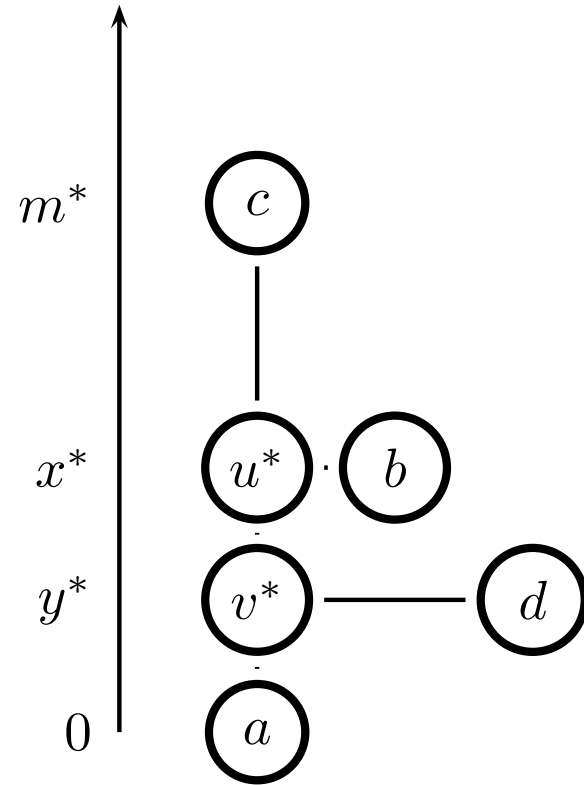
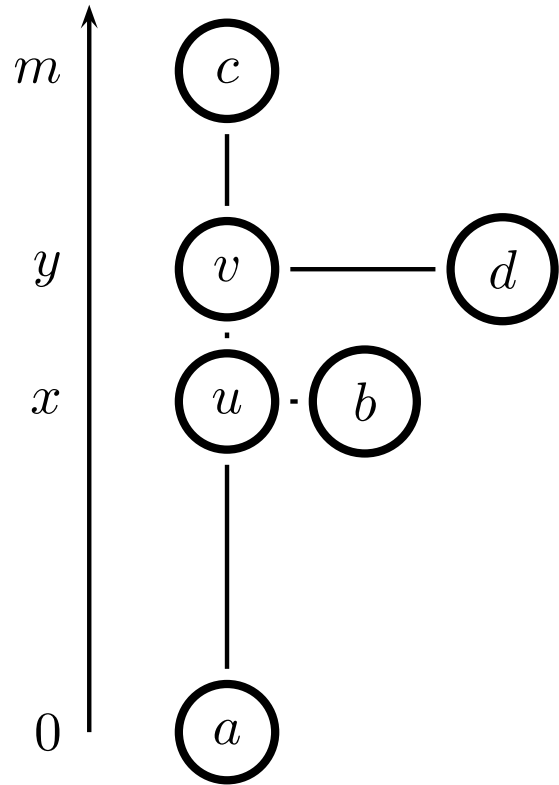
1. Toss a fair coin for each internal node to choose a random left/right orientation. (This chooses one of the  $2^{s-1}$  possible graphical representations of the tree uniformly at random.)
2. Modify each of the “valley depths” of the traversal profile by perturbing at random uniformly within a small fixed window. (There are  $s - 1$ , assuming a molecular clock,  $2(s - 1)$  without assuming a molecular clock).
3. The proposed tree corresponds to the new traversal profile.
4. The proposed tree is either accepted or rejected according to the Metropolis-Hastings algorithm.



## THE LOCAL TREE UPDATE WITH A CLOCK



# THE LOCAL TREE UPDATE WITHOUT A CLOCK



## WHALE PHYLOGENY: LARGET

- Three evolutionary trees have been proposed to describe the evolutionary relationships among whales and dolphins.
- There are three groups considered to be monophyletic:
  1. Dolphins
  2. Toothed whales
  3. Baleen whales
- The traditional taxonomy places dolphins and toothed whales together.
- Several analyses of DNA sequences suggest toothed whales and baleen whales are together.
- A conflicting analysis of different DNA sequences suggests dolphins and baleen whales are together.

## DATA AND MODEL CHARACTERISTICS

- Data: 31 aligned DNA sequences coding cytochrome *b*, 1140 bp, from 2 dolphins, 1 sperm whale, 11 baleen whales, and 17 artiodactyls (hippopotomous, 6 camels, pig, peccary, cow, sheep, goat, black-tailed deer, giraffe, fallow, pronghorn, chevrotain).
- Substitution Model: HKY85, allowing codon-position specific parameters.

## MCMC Run Characteristics

- Four separate runs from random starts.
- 500,000 scans per run after burn-in.
- Each scan uses two move types: (1) uniform window update for substitution parameters, and (2) tree update.
- Burned in with GLOBAL and then used LOCAL.
- Subsampled 1 in 10 scans.

Taxa	Labels
Dolphins	1, 2
Giant Sperm Whale	3
Bowhead Whale	4
Right Whale	5
Minke Whale	6
Antarctic Minke Whale	7
Sei Whale	8
Bryde's Whale	9
Fin Whale	10
Blue Whale	11
Humpback Whale	12
California Whale	13
Pygmy Right Whale	14

\*\*\*\*\* BAMBE Summarize Version 7 \*\*\*\*\*

\*\* Posterior probabilities of clades in most probable tree topology \*\*

Count	Prob.	Clade
...		
200000	1.000	{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
200000	1.000	{1,2,3,4,5,6,7,8,9,10,11,12,13,14}
200000	1.000	{1,2}
198620	0.993	{3,4,5,6,7,8,9,10,11,12,13,14}
200000	1.000	{4,5,6,7,8,9,10,11,12,13,14}
200000	1.000	{4,5}
189343	0.947	{6,7,8,9,10,11,12,13,14}
200000	1.000	{6,7,8,9,10,11,12,13}
200000	1.000	{6,7}
198142	0.991	{8,9,10,11,12,13}
170693	0.853	{8,9,10,11,12}
115978	0.580	{8,9,10}
...		

\*\*\*\*\* BAMBE Summarize Version 7 \*\*\*\*\*

\*\*\*\*\* Named clades \*\*\*\*\*

200000 A {1,2}

200000 A1 (1,2)

200000 B {4,5}

200000 B1 (4,5)

189343 C {6,7,8,9,10,11,12,13,14}

93936 C1 (((6,7),(((8,9),10),(11,12))),13),14)

52282 C2 (((6,7),(((8,9),11),(10,12))),13),14)

6332 C3 (((6,7),(((8,9),10),13),(11,12))),14)

6001 C4 (((6,7),((8,9),10),((11,12),13))),14)

5010 C5 (((6,7),(((8,9),(10,12)),11),13),14)

3577 C6 (((6,7),((8,9),(10,(11,12))),13),14)

3127 C7 (((6,7),(((8,9),10),12),11),13),14)

2737 C8 (((6,7),(((8,9),11),13),(10,12))),14)

### 90% Credible Region of Whales

0.466	0.466	$(A_1, (3, (B_1, ((C_1, (D_1, 13)), 14))))$
0.260	0.726	$(A_1, (3, (B_1, ((C_1, (D_2, 13)), 14))))$
0.031	0.758	$(A_1, (3, (B_1, ((C_1, (((8, 9), 10), 13), (11, 12))), 14))))$
0.030	0.788	$(A_1, (3, (B_1, ((C_1, (((8, 9), 10), ((11, 12), 13))), 14))))$
0.025	0.813	$(A_1, (3, (B_1, ((C_1, (D_3, 13)), 14))))$
0.020	0.833	$(A_1, (3, ((B_1, 14), (C_1, (D_1, 13))))$
0.018	0.851	$(A_1, (3, (B_1, ((C_1, (D_4, 13)), 14))))$
0.015	0.866	$(A_1, (3, (B_1, ((C_1, (D_5, 13)), 14))))$
0.014	0.880	$(A_1, (3, (B_1, ((C_1, (((8, 9), 11), 13), (10, 12))), 14))))$
0.014	0.893	$(A_1, (3, ((B_1, 14), (C_1, (D_2, 13))))$
0.011	0.904	$(A_1, (3, (B_1, ((C_1, (((8, 9), 13), 11), (10, 12))), 14))))$



		F1	F2	F3	F4	F5	F6
-----+							
F1		193542	182	56	25	7	2
F2		177	4044	24	0	0	0
F3		61	19	1117	0	0	0
F4		26	0	0	619	0	0
F5		7	0	0	0	49	0
F6		1	0	0	1	0	40

## A COMPARISON WITH MAXIMUM LIKELIHOOD

- For the full 31 taxa, the most probable tree topology had a posterior probability of 0.413, with an estimated Monte Carlo standard error of 0.013.
- To achieve this level of accuracy with independent draws from the posterior would require about 1400 draws.
- Four separate runs each required about 100 CPU minutes on a 300 Mhz Pentium II PC, for a total of about 7 hours.
- A single analysis using PHYLIP to do maximum likelihood required 180 minutes. Bootstrapping 1400 times to achieve similar statistical accuracy would require about 175 days, or nearly half a year.

“Perhaps the most vexing mystery [in Bayesian analysis of phylogeny] is the observed discrepancy between Bayesian posterior probabilities and nonparametric bootstrap support values.”

– Huelsenbeck *et al.* 2002