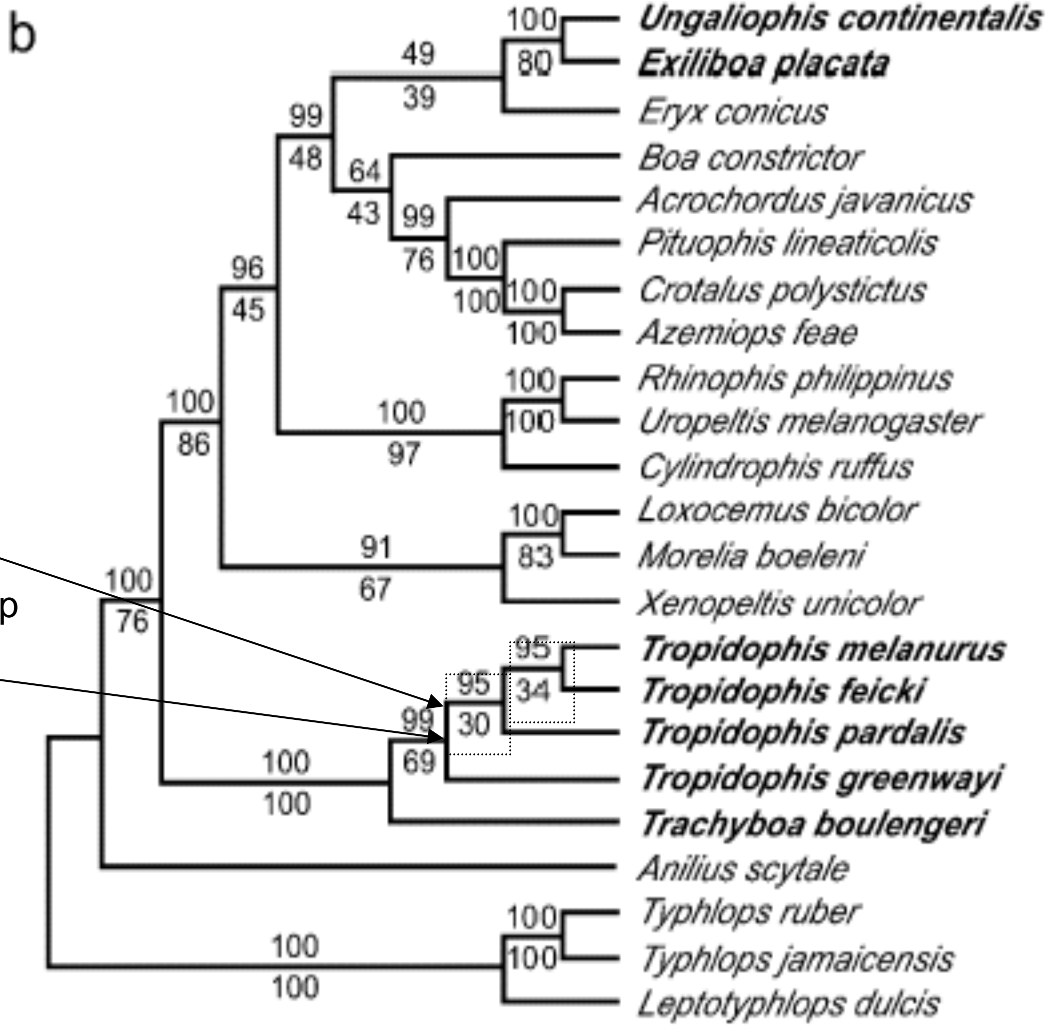
Sometimes ML bootstrap and Bayesian posterior probabilities on clades can be strikingly different...



Bayesian posterior probability

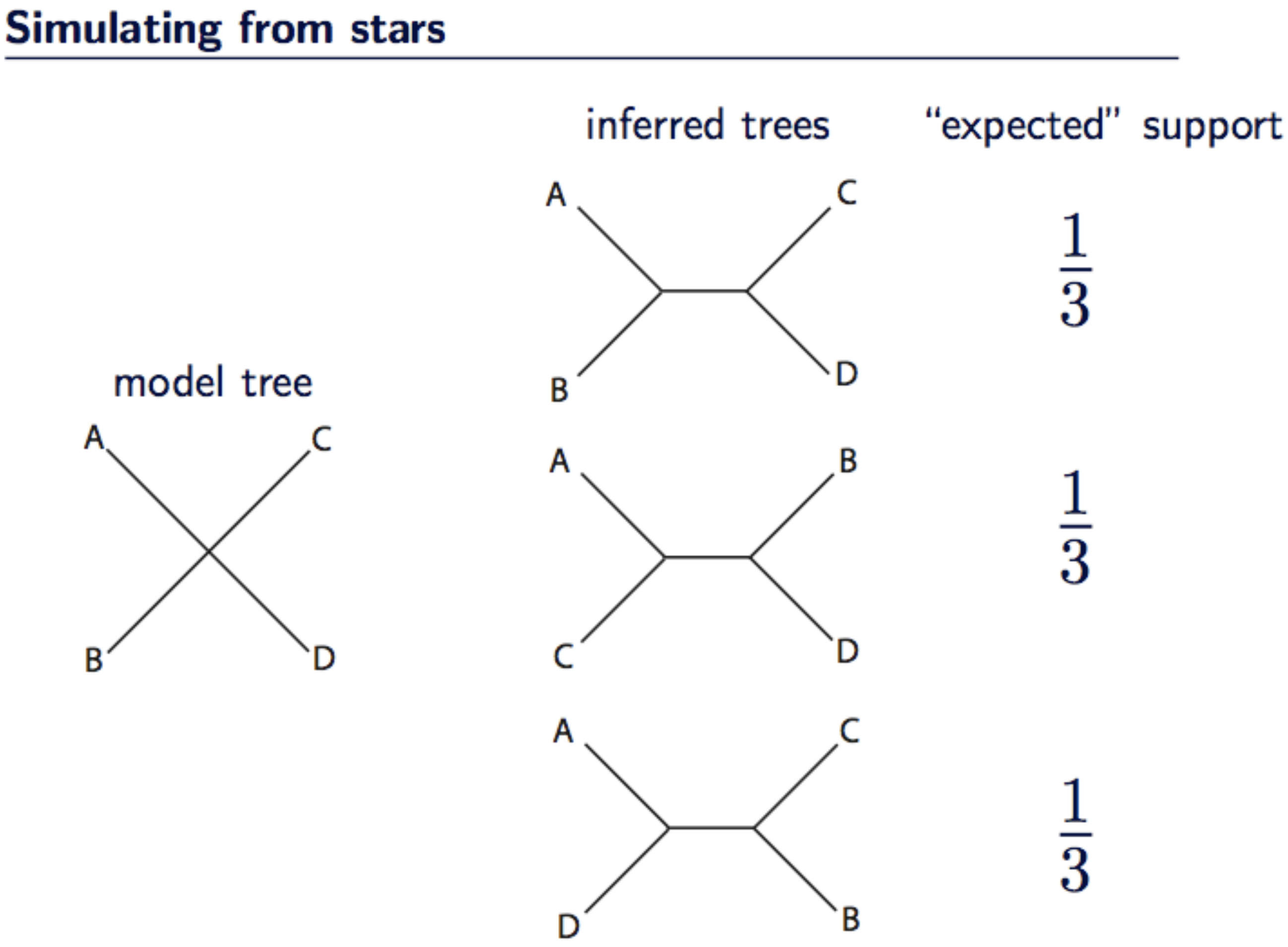
Likelihood bootstrap support

Why are bootstrap values and posterior probabilities inconsistent?

ML bootstrapping and Bayesian phylogenetic inference are not the same thing. While we might expect clades with high bootstrap support to also have high posterior probabilities (and likewise for low values), there is no guarantee that this will be the case.

Another cause of discrepancies is when one or more branches are very short (effectively polytomies)...

The polytomy problem

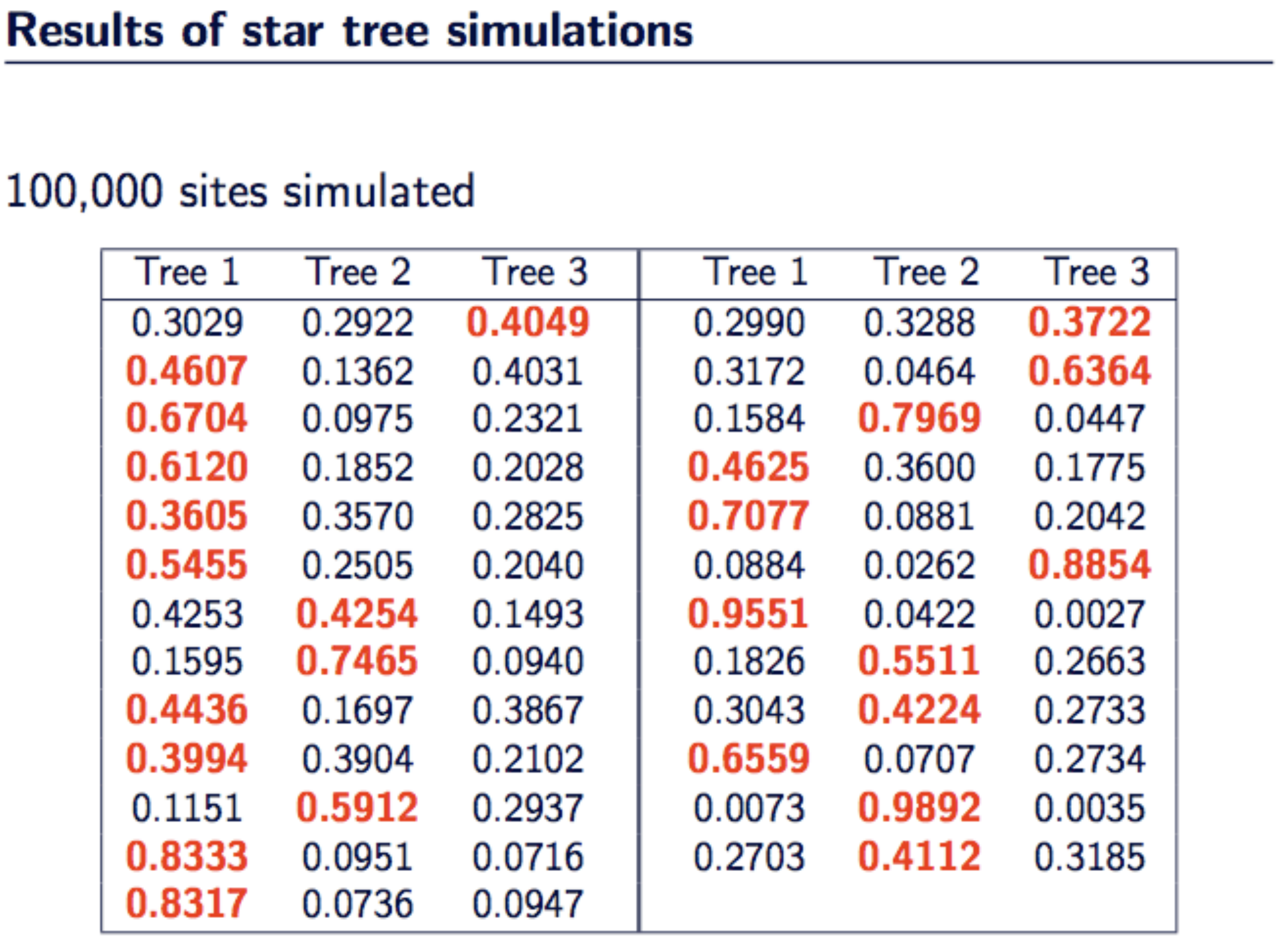


Simulating from stars

model tree	inferred trees	“expected” support
	((A,B),(C,D))	$\frac{1}{3}$
	((A,C),(B,D))	$\frac{1}{3}$
	((A,D),(C,B))	$\frac{1}{3}$

Slide from Mark Holder (Woods Hole 2008)

The polytomy problem



Results of star tree simulations

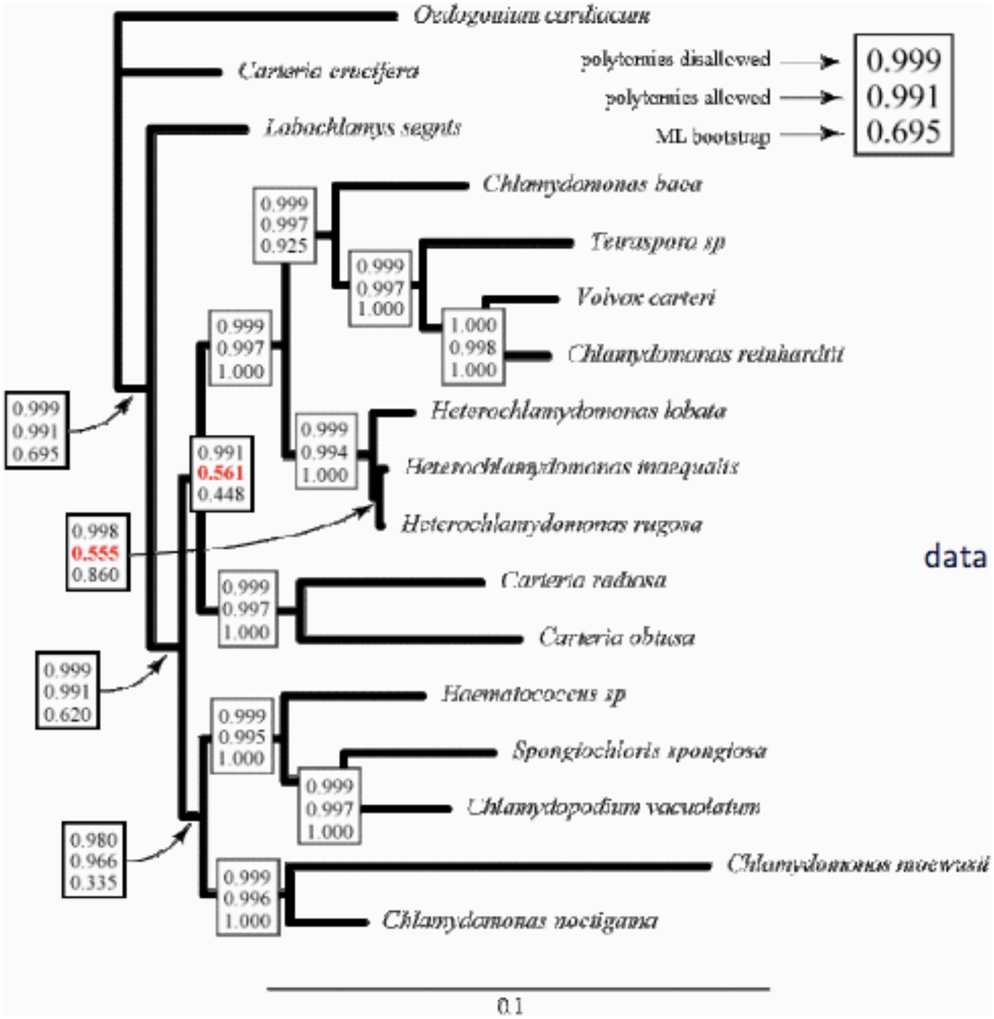
100,000 sites simulated

Tree 1	Tree 2	Tree 3	Tree 1	Tree 2	Tree 3
0.3029	0.2922	**0.4049**	0.2990	0.3288	**0.3722**
0.4607	0.1362	0.4031	0.3172	0.0464	**0.6364**
0.6704	0.0975	0.2321	0.1584	**0.7969**	0.0447
0.6120	0.1852	0.2028	**0.4625**	0.3600	0.1775
0.3605	0.3570	0.2825	**0.7077**	0.0881	0.2042
0.5455	0.2505	0.2040	0.0884	0.0262	**0.8854**
0.4253	**0.4254**	0.1493	**0.9551**	0.0422	0.0027
0.1595	**0.7465**	0.0940	0.1826	**0.5511**	0.2663
0.4436	0.1697	0.3867	0.3043	**0.4224**	0.2733
0.3994	0.3904	0.2102	**0.6559**	0.0707	0.2734
0.1151	**0.5912**	0.2937	0.0073	**0.9892**	0.0035
0.8333	0.0951	0.0716	0.2703	**0.4112**	0.3185
0.8317	0.0736	0.0947			

Slide from Mark Holder (Woods Hole 2008)

The polytomy problem

Effects of allowing for polytomies



data from Shoup and Lewis (2003)

Slide from Mark Holder (Woods Hole 2008)

Why are bootstrap values and posterior probabilities inconsistent?

But another reason is that (especially prior to incorporation of a stopping criterion in MrBayes), chain length has typically been treated as a constant rather than being adjusted to a value appropriate for the size of the data set.

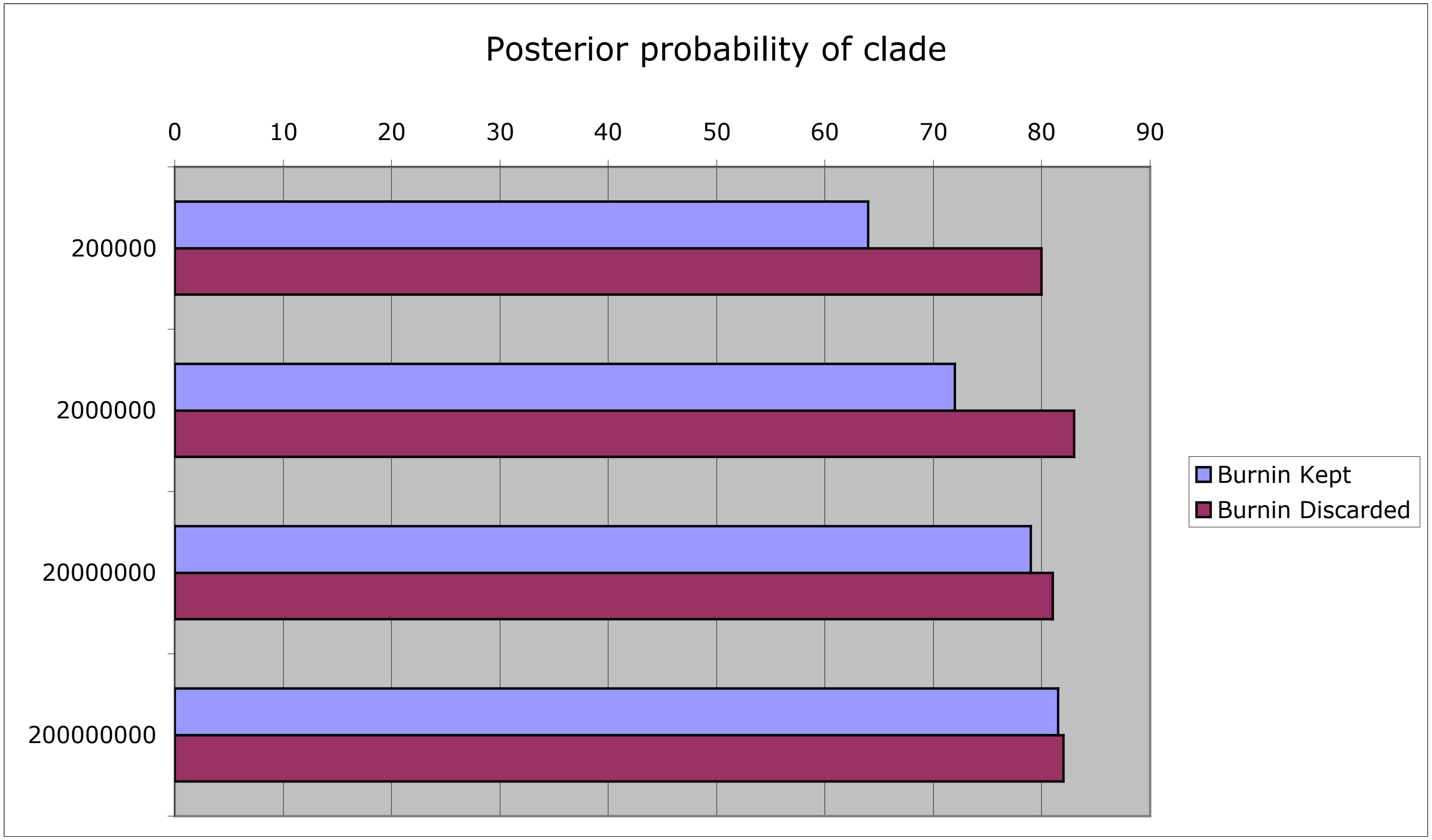
Typically, runs were done using the default number of iterations (generations) or some multiple of multiple of that default.

Is it reasonable to treat chain length as a constant?

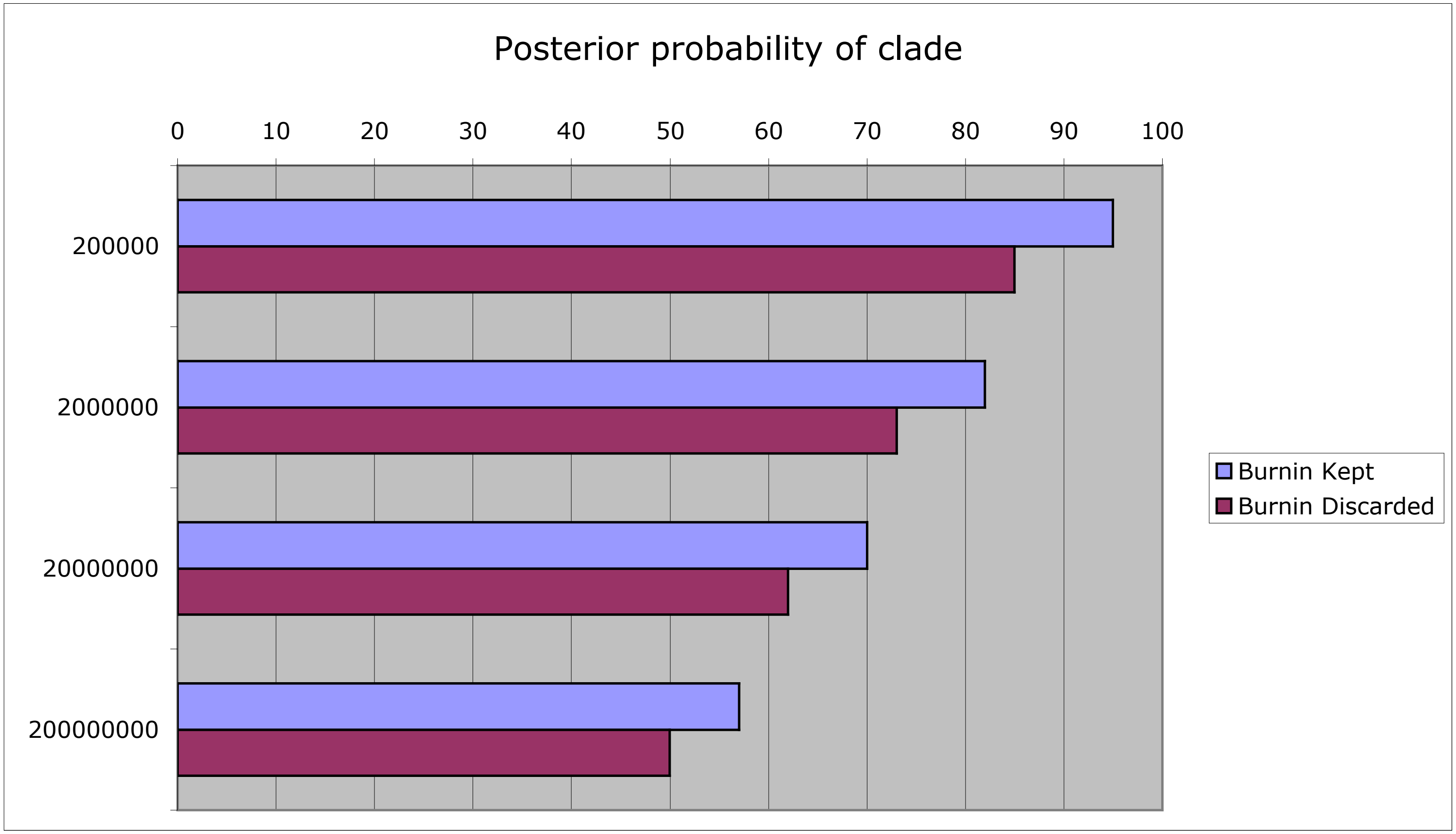
- It seems likely that chain length needed for valid inferences should increase with increasing number of taxa due to exponentially growing state space of possible topologies
- E.g., Brooks and Roberts (1999, Bayesian Statistics 6:415-440):

> Many current applications of MCMC are not just bigger (in terms of the number of parameters) but also more complex structurally. Many of these models exhibit phase transition behavior, which can be interpreted statistically as a high dimensional, and very extreme, multimodality. When phase transition is present, the number of iterations needed to converge typically increases exponentially with the dimension of the problem.

Ideal case: Chain length has little effect on estimated posterior probability of a clade



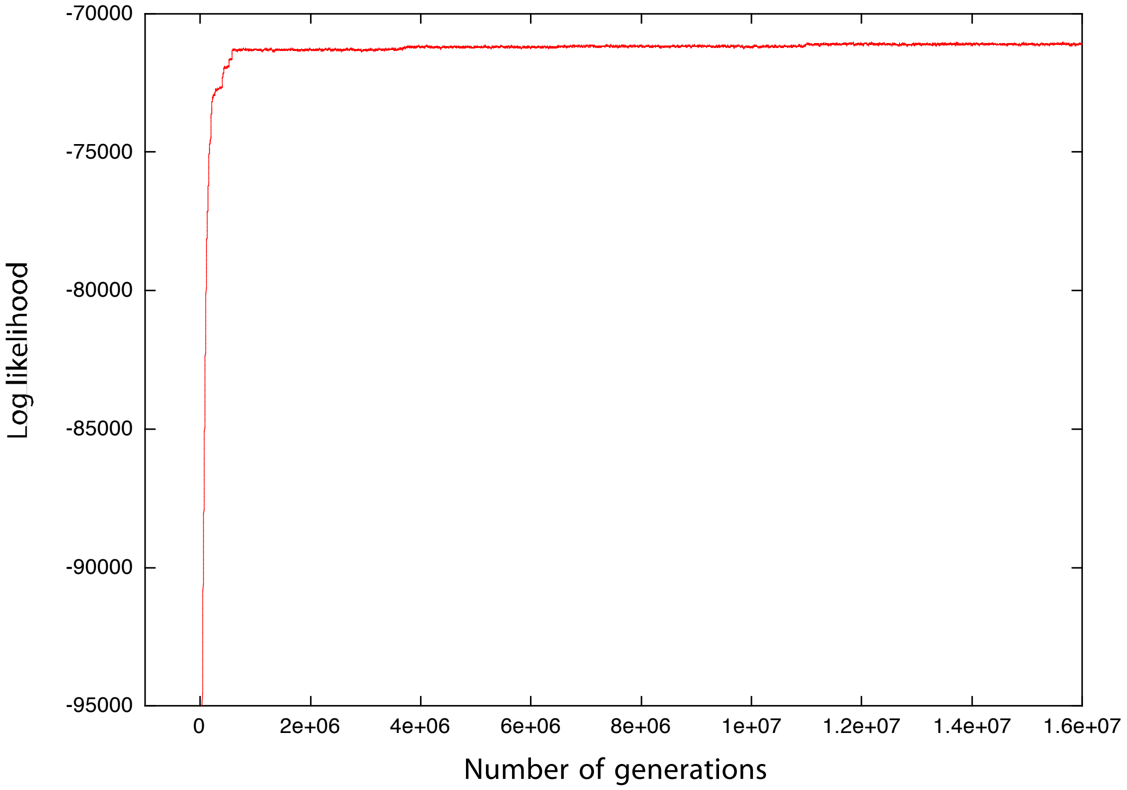
Problem if posterior probability of a clade varies widely when chain is run for different lengths



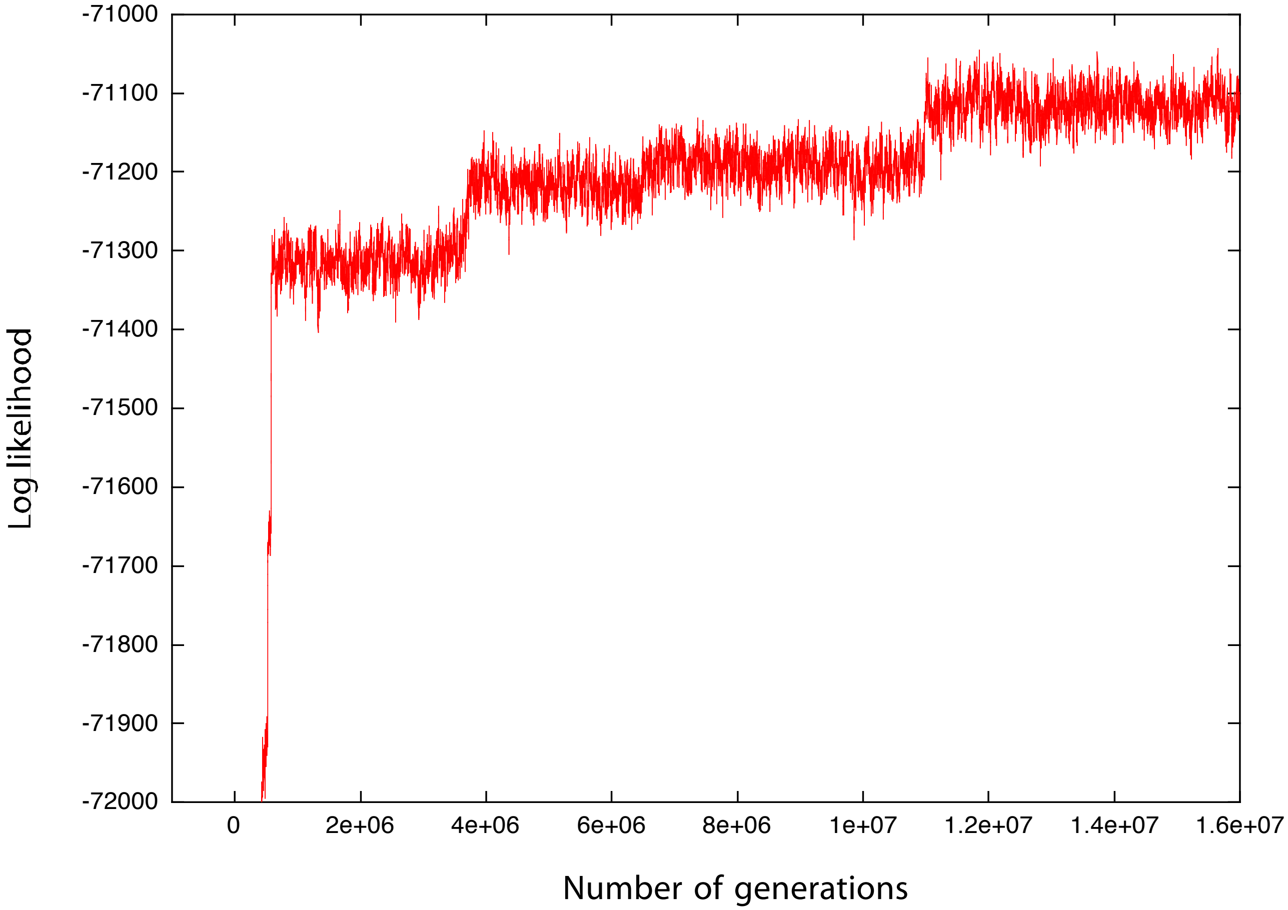
Convergence diagnostics

- Brooks and Roberts (Statistics and Computing 8:319-335): “...because of the fact that theoretical convergence times vary widely (especially in high dimensional problems), and in the absence of any general techniques for a priori prediction of run lengths, it is necessary to carry out some form of statistical analysis in order to assess convergence.”

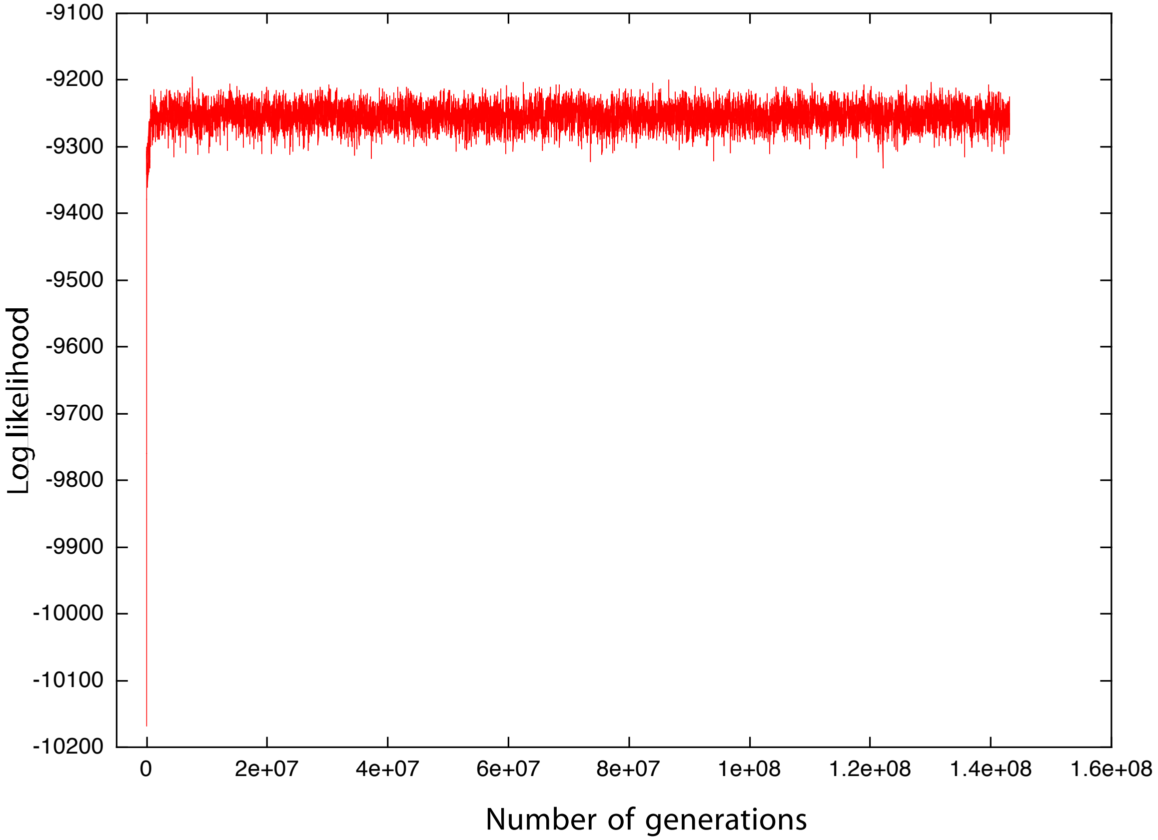
One informal convergence diagnostic: the likelihood trace



Be careful about axis scaling...

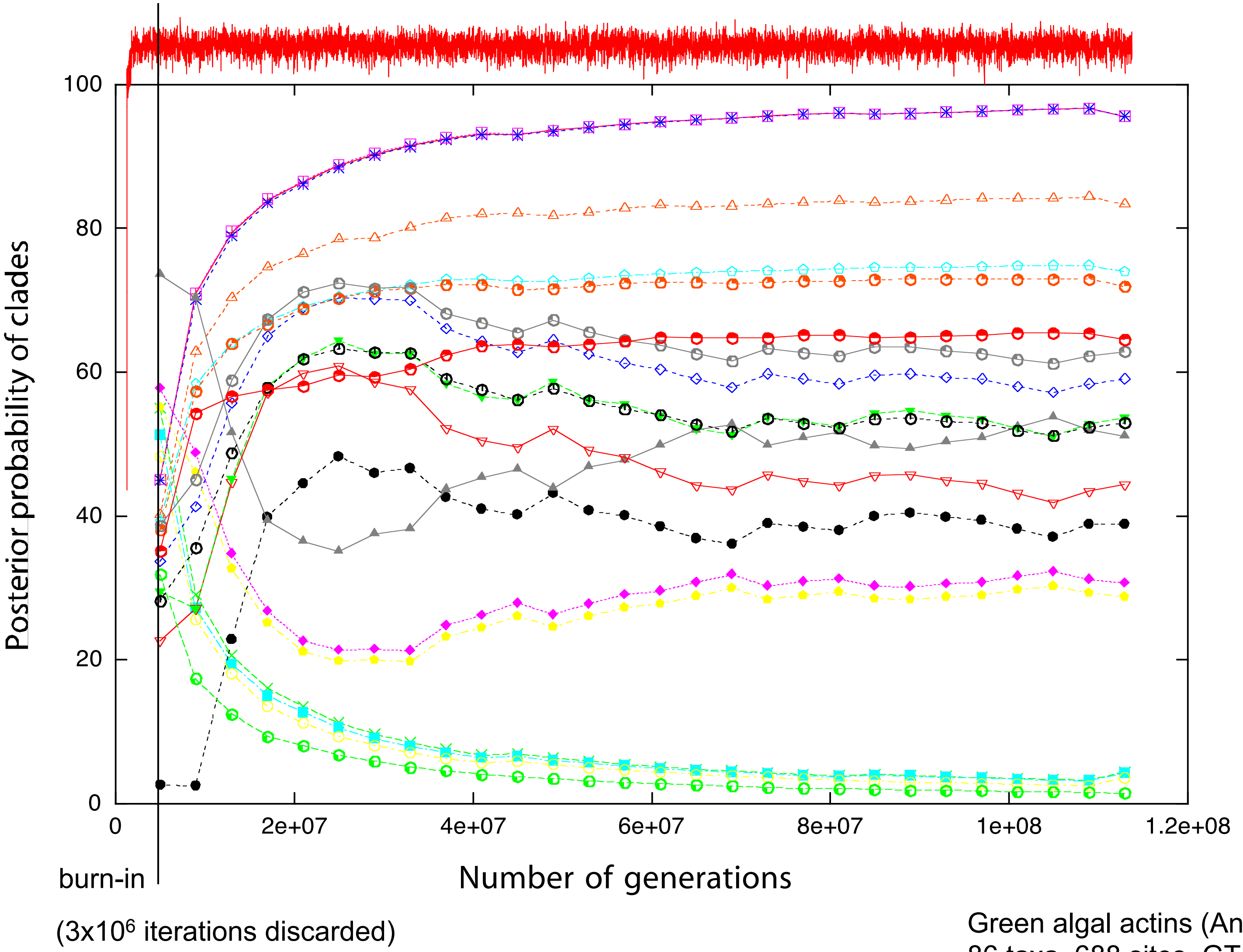


Example—green algal actins



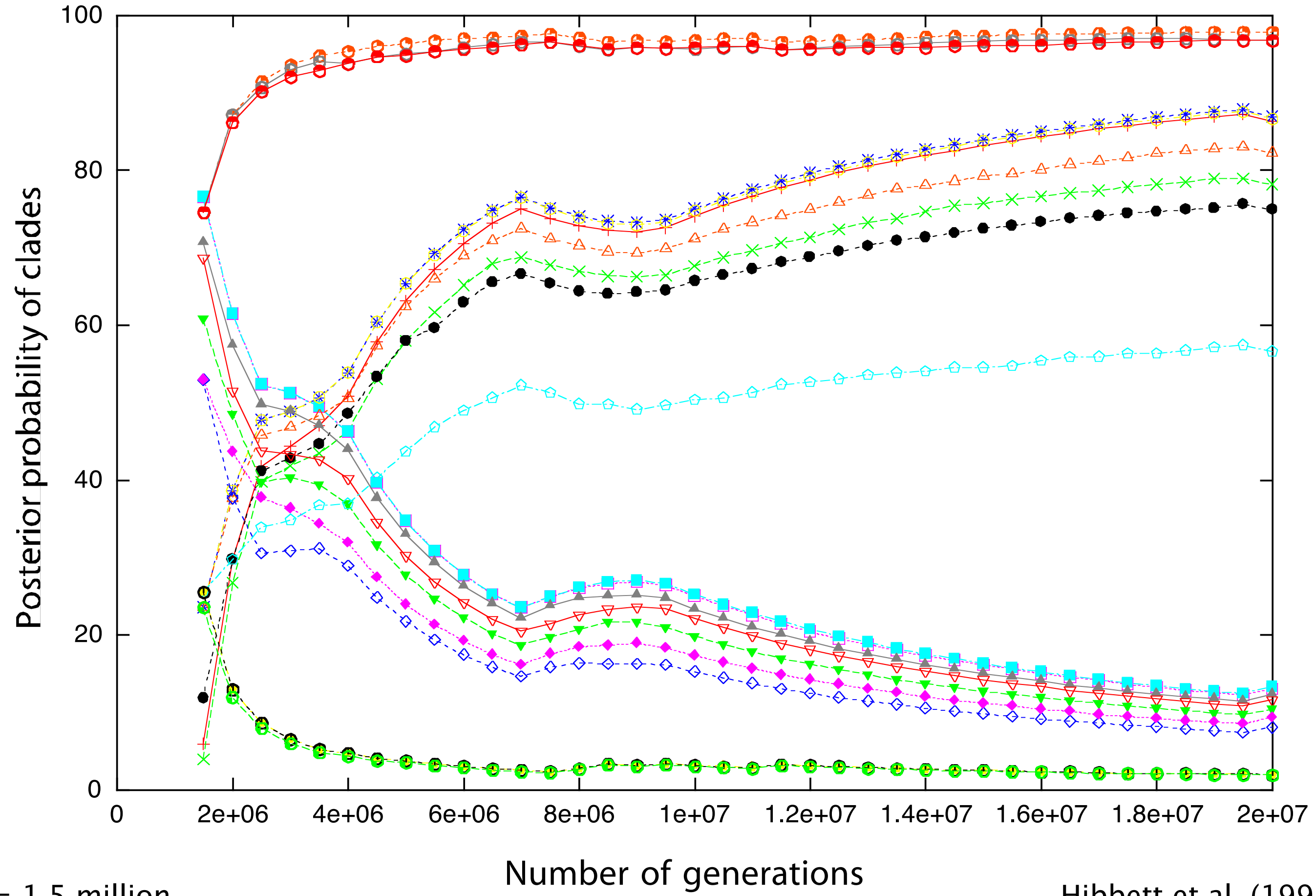
Green algal actins (An et al., 1999)
86 taxa, 688 sites, GTR+gamma

Convergence problems:
Posterior clade probability as a function of chain length



Green algal actins (An et al., 1999)
86 taxa, 688 sites, GTR+gamma

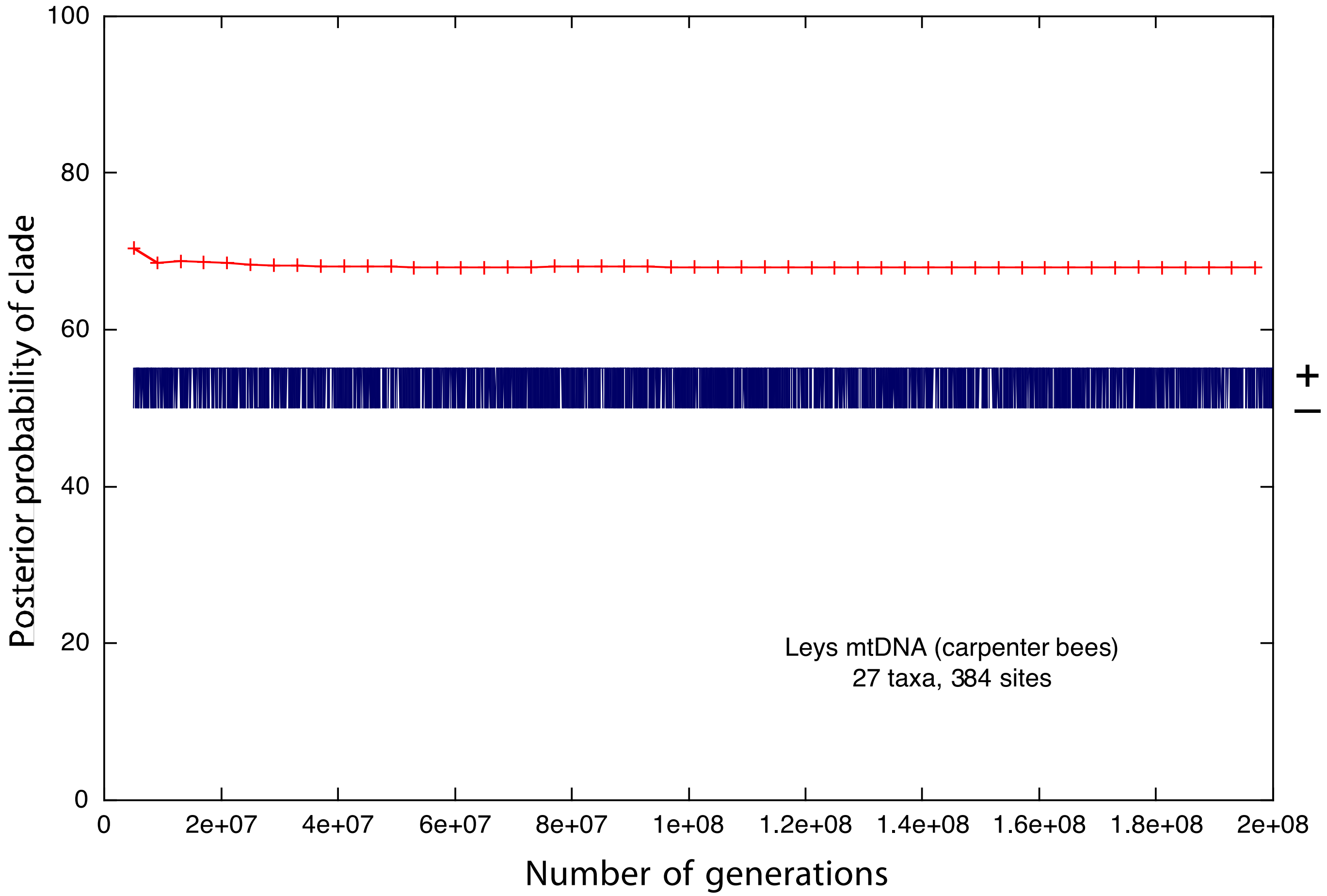
Another example (fungal rDNA)



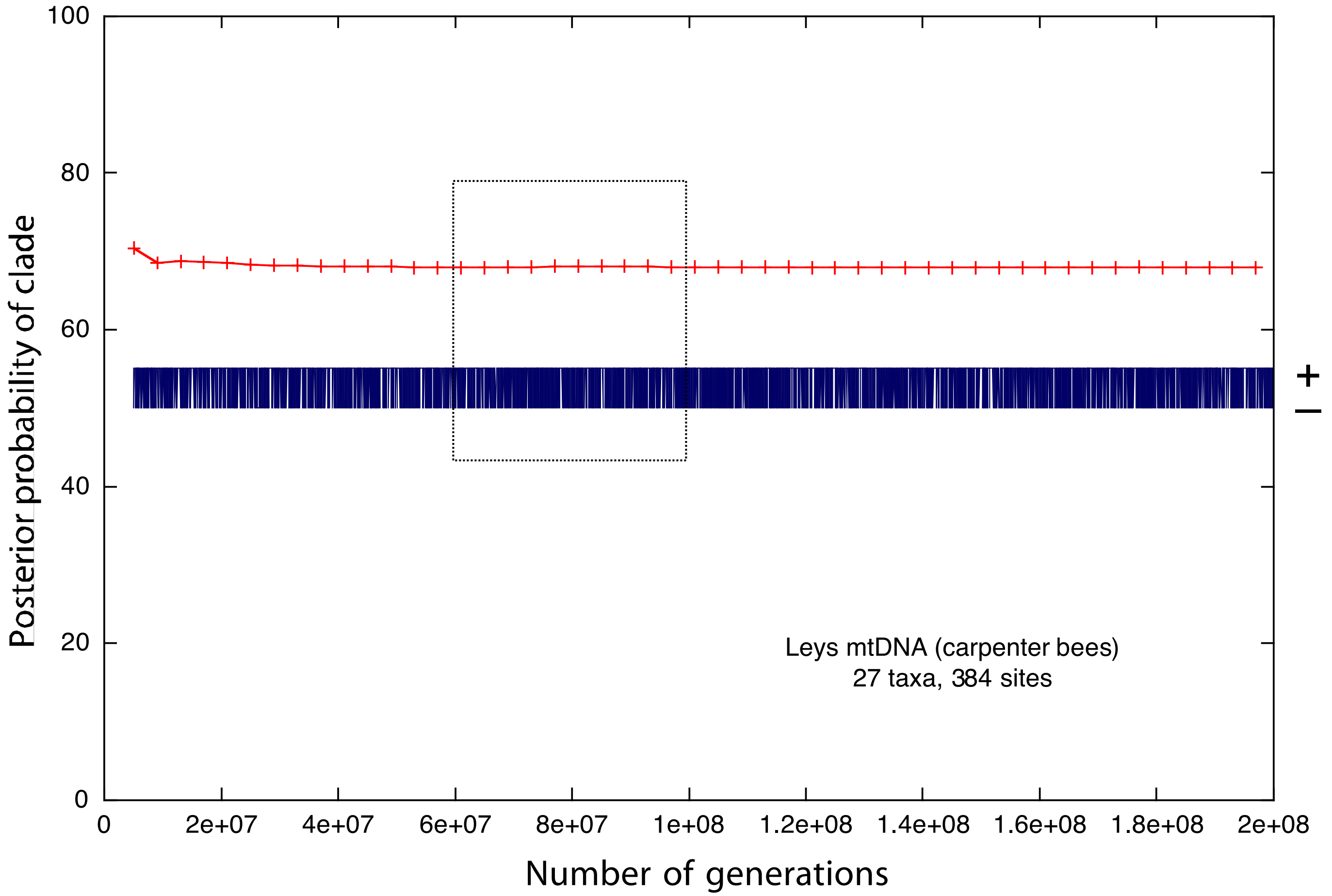
burn-in = 1.5 million

Hibbett et al. (1997)
Mushroom and puffball rDNA
85 taxa, 3487 sites
HKY+I model

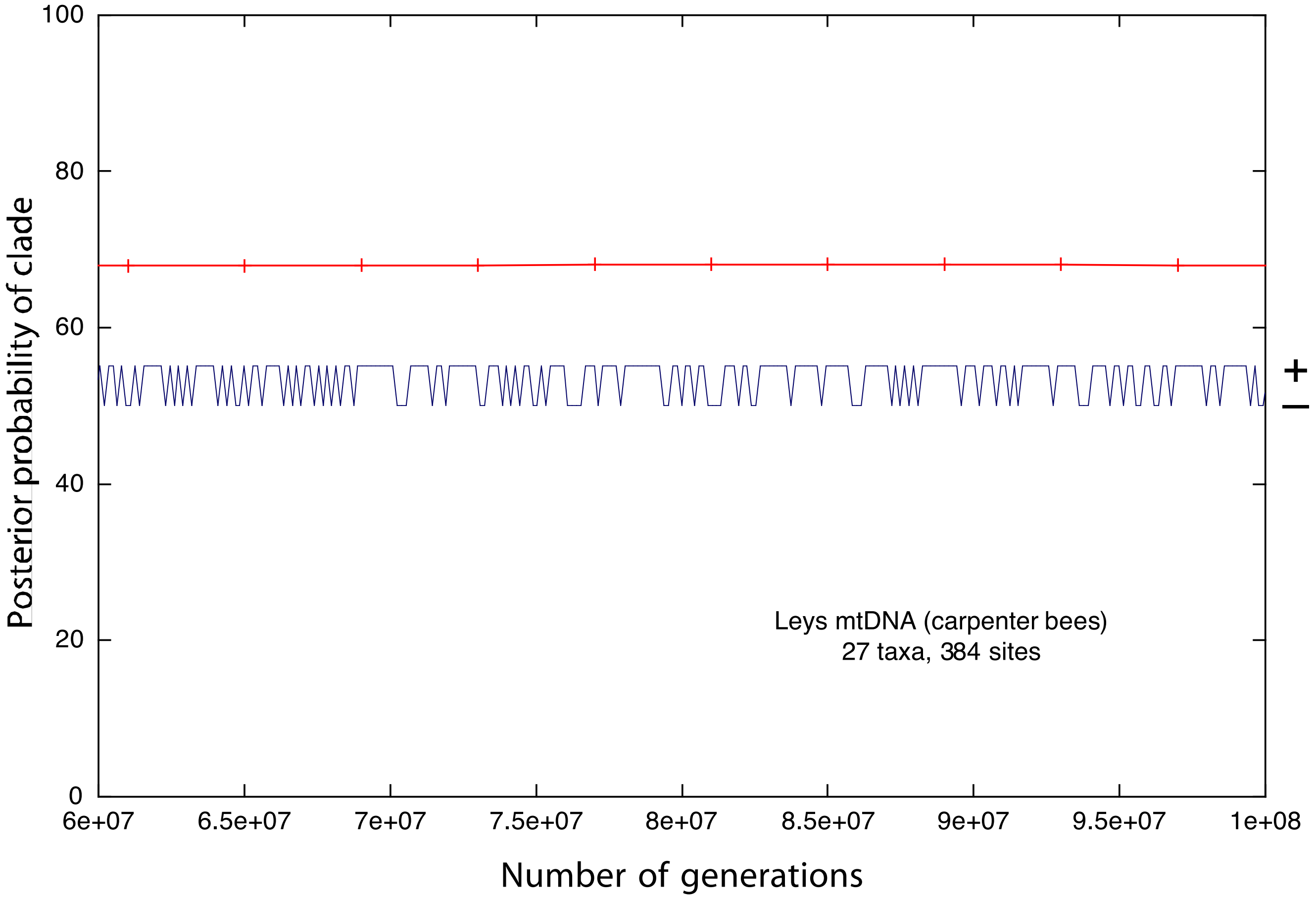
A (possibly) better diagnostic:
Trace plots of clade presence/absence



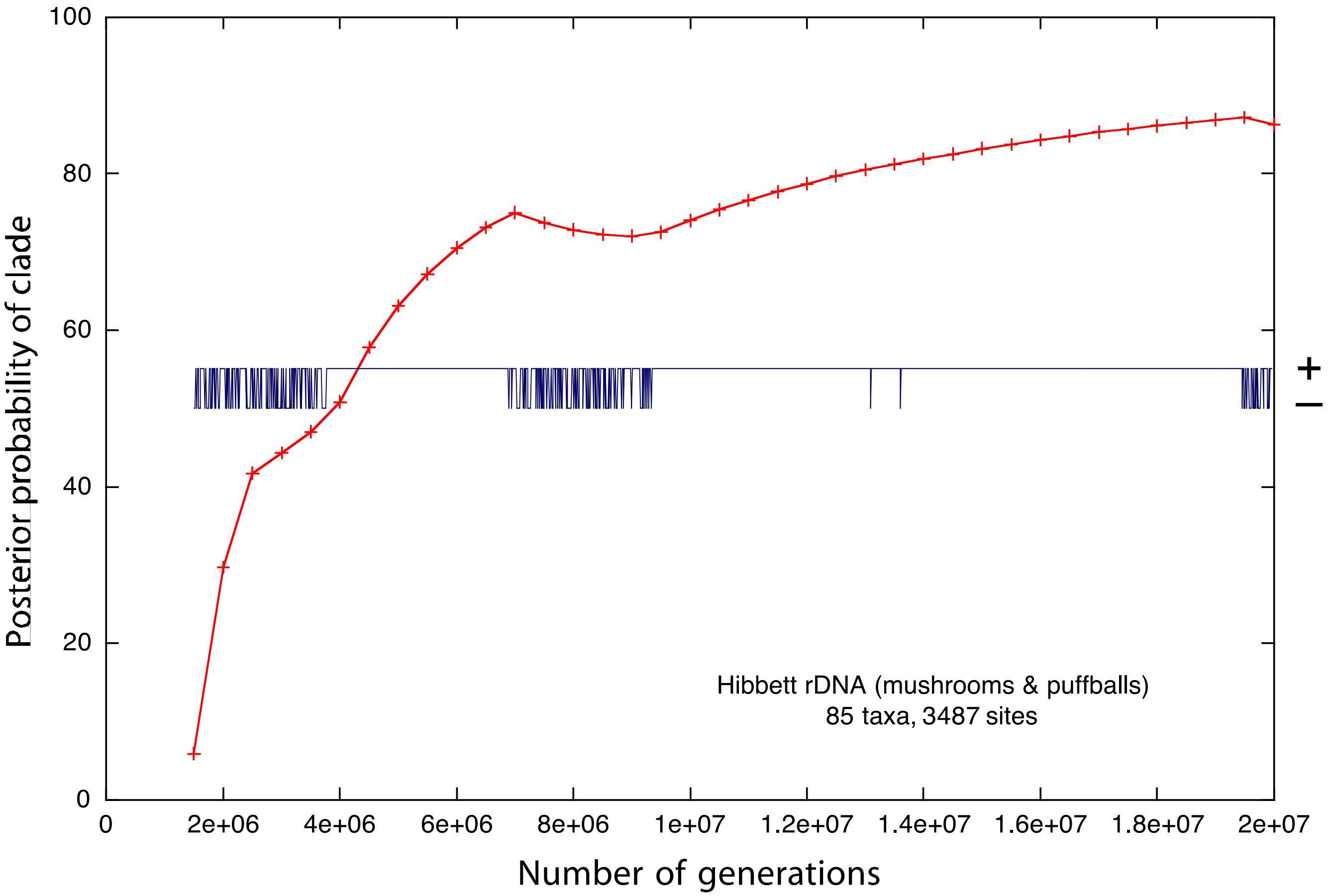
A (possibly) better diagnostic:
Trace plots of clade presence/absence



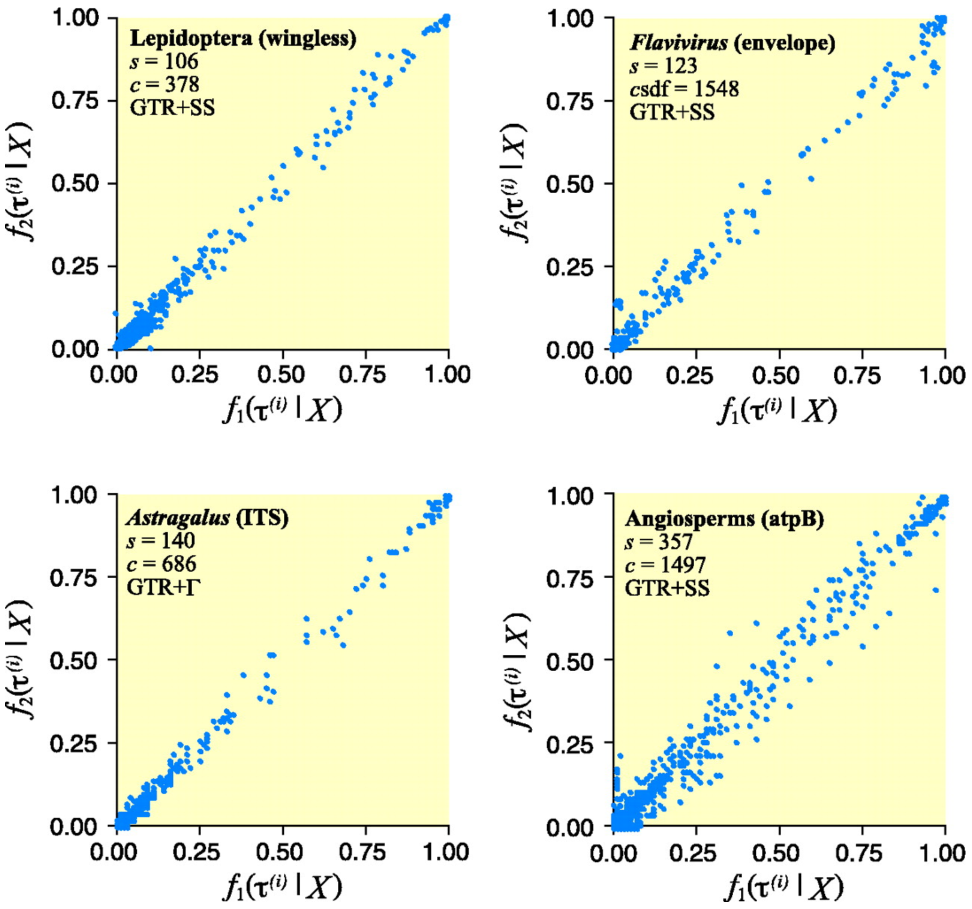
A (possibly) better diagnostic: Trace plots of clade presence/absence



A (possibly) better diagnostic: Trace plots of clade presence/absence



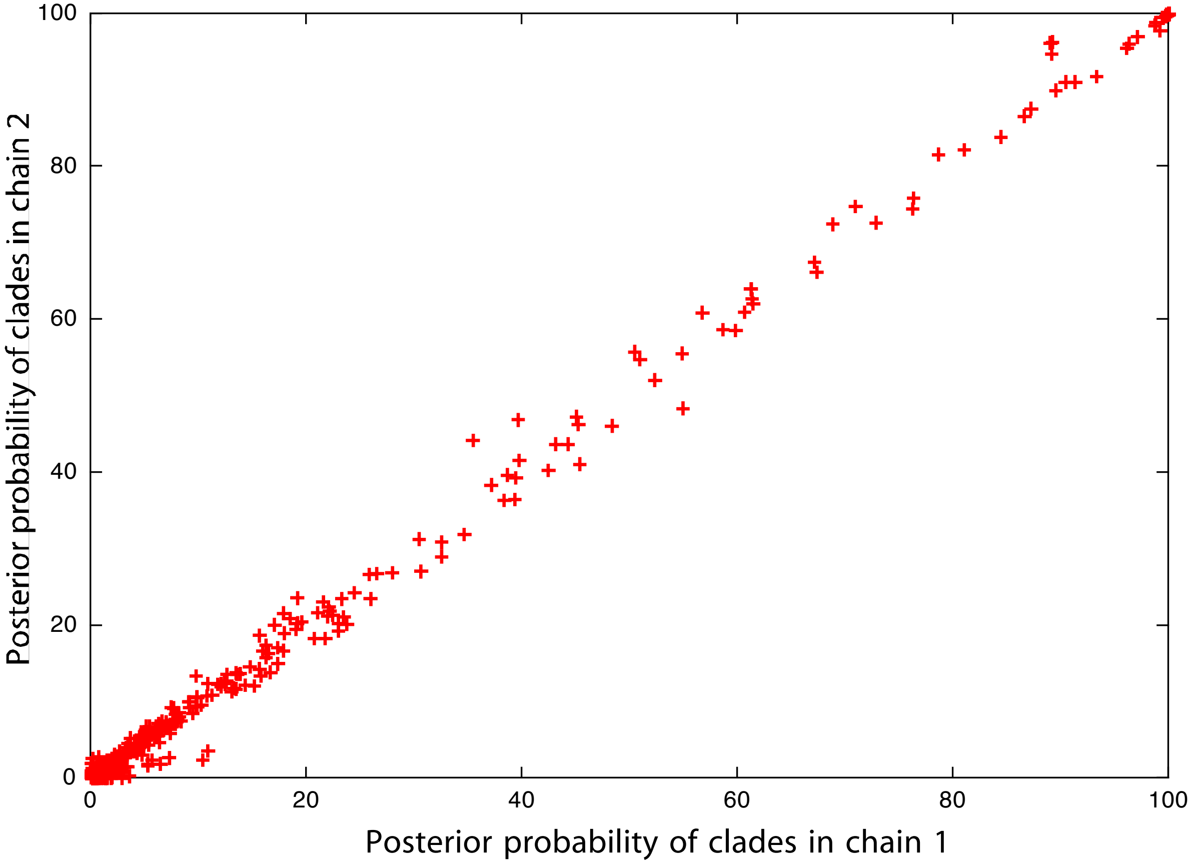
Huelsenbeck et al. (2001, Science 294:2310-2314)
“Two-chain clade probability” diagnostic



Run two chains from different starting points; plot scattergram of the probabilities of each clade estimated by the two chains

“Two-chain clade probability” diagnostic

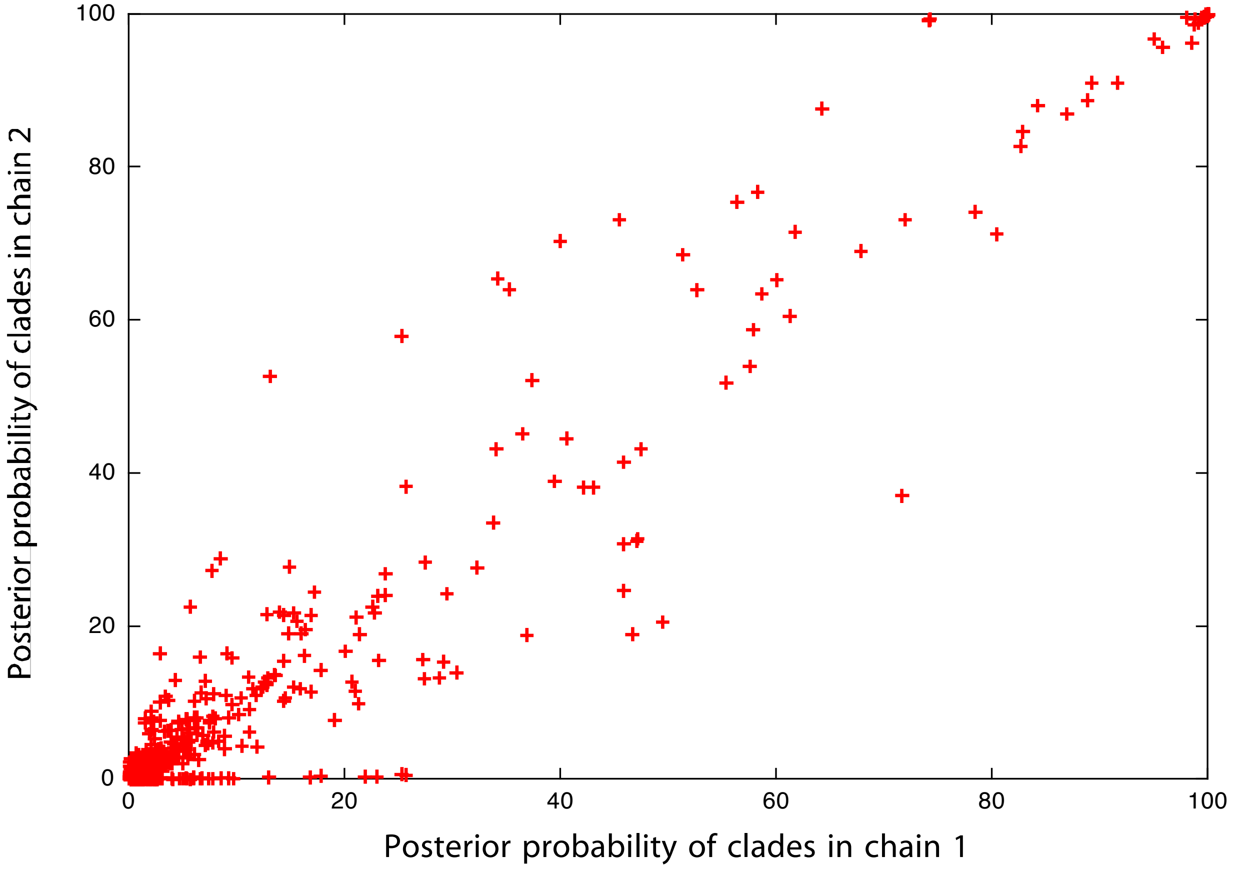
one of the 45 pairwise chain comparisons for 10 chains...



Green algal actins (An et al., 1999)
86 taxa, 688 site, GTR+gamma
20×10^6 generations in each chain

“Two-chain clade probability” diagnostic

but a different pair of chains indicates strong non-convergence...



Green algal actins (An et al., 1999)
86 taxa, 688 site, GTR+gamma
20×10^6 generations in each chain

Main conclusions:

- There is a real possibility of convergence failure with chain lengths many orders of magnitude longer than those typically run, raising the possibility of strongly biased estimates.
- As with other phylogenetic methods (e.g., heuristic searches with maximum likelihood), naive acceptance of default parameter and option settings may lead to trouble.

Caveats:

- We biased our search for datasets by choosing ones that had numerous “islands” in parsimony searches (multimodal). It may turn out that the phenomena we report here are not that common, however, a user who does not attempt to assess whether the chain has converged will never know when it hasn’t. As with other phylogenetic methods (e.g., heuristic searches with maximum likelihood), naive acceptance of default parameter and option settings may lead to trouble.
- MrBayes 3.2 has a better repertoire of topology moves.

Are We There Yet? (AWTY)

A Tool for Graphically Exploring Convergence of MCMC Chains in Bayesian Inference Using MCMC

http://ceb.sc.fsu.edu/awty

Initial version (“Converge”)

Dan Warren
David Swofford

Current version

Jim Wilgenbusch
Johan Nylander
Dan Warren
David Swofford

Nylander, J.A.A., J.C. Wilgenbusch, D. L. Warren, and D. L. Swofford. 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24(4): 581-583.