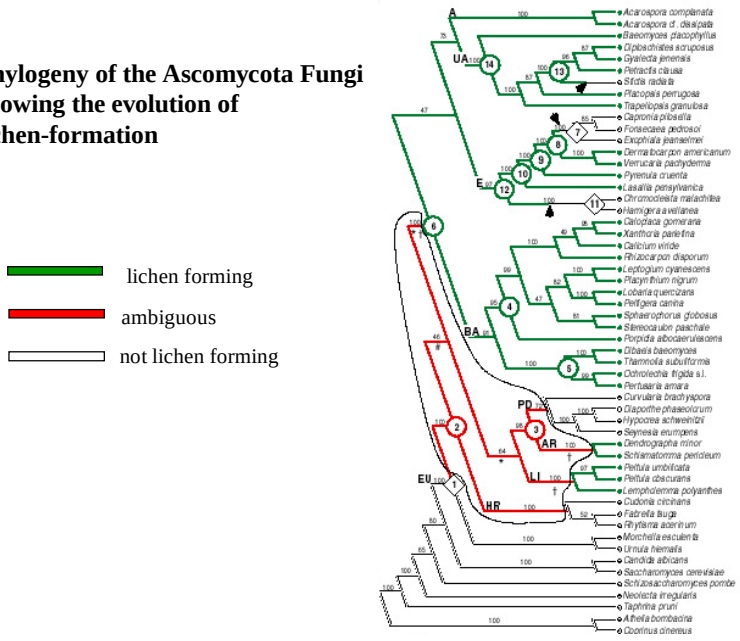
Accounting for Phylogenetic Uncertainty in Comparative Studies: MCMC and MCMCMC Approaches

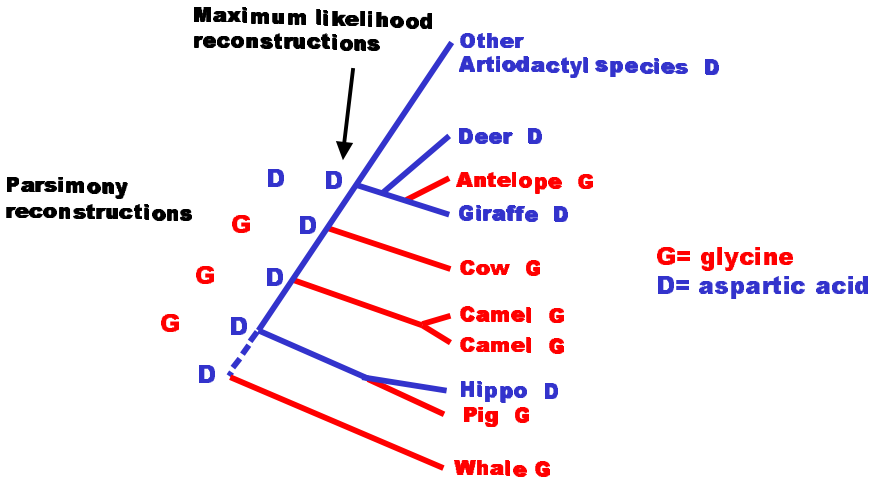
Mark Pagel
Reading University

m.pagel@rdg.ac.uk

Phylogeny of the Ascomycota Fungi showing the evolution of lichen-formation



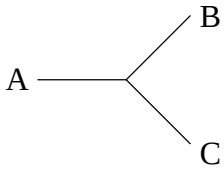
Reconstructing Ancestral Ribonuclease Proteins



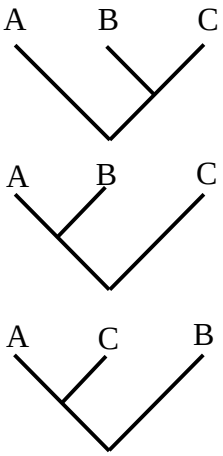
Phylogenetic Uncertainty

Three Species

Unrooted Tree

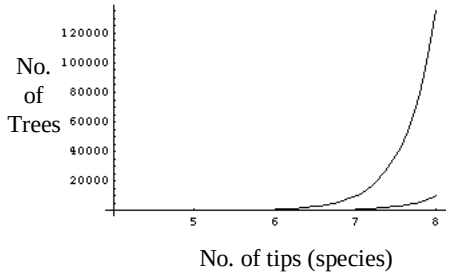


Rooted Trees



Number of Possible Phylogenetic Trees

Species	Unrooted	Rooted
3	**1**	**3**
4	**3**	**15**
5	**15**	**105**
6	**105**	**945**
10	**2,027,025**	**34,459,425**
50	**2.83×10^{74}**	**2.75×10^{76}**



N=50 **No. rooted = 275292135328356515452597297515244306393009730358161960983265537721525878906 25**
No. unrooted = 283806325080779912837729172696128150920628587998105114415737667754150390625

Accounting for Phylogenetic Uncertainty

Markov-Chain Monte Carlo (MCMC) Methods
- generate a long chain of phylogenetic trees (tree proposal and acceptance mechanisms)
- randomly sample from the converged chain
- calculate event or evolutionary process in each

Tree acceptance mechanism: The Metropolis-Hastings Algorithm

Accept new tree with p=1.0 if $L(T_{n+1}) > L(T_n)$

otherwise…

accept with probability $\propto L(T_{n+1})/ L(T_n)$

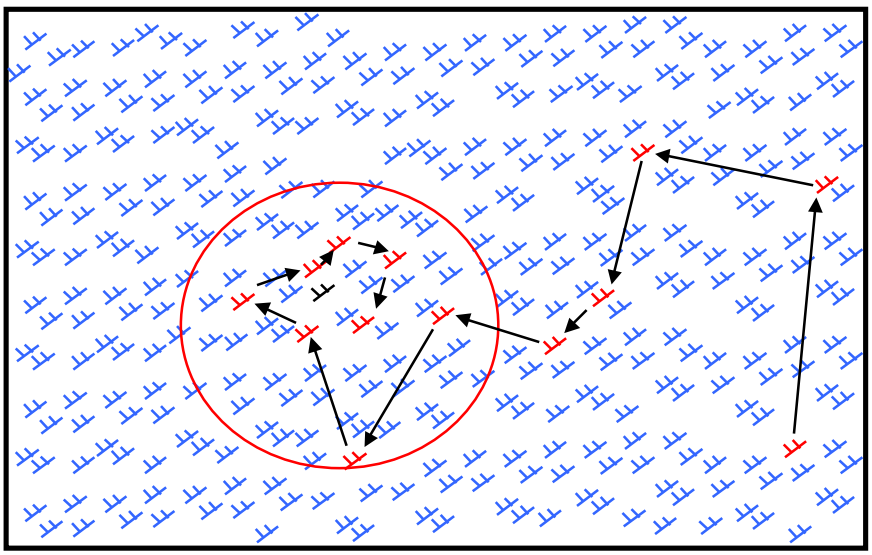
Metropolis-Hastings Algorithm: Accept new tree according to:

$$R = \min\left\{1, \frac{f(X|T')}{f(X|T)} x \frac{f(T')}{f(T)} x \frac{f(T|T')}{f(T'|T)}\right\}$$

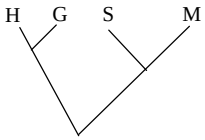
likelihood ratio prior ratio proposal ratio

X=data (e.g., gene sequences) T=tree (topology, branches, parameters)

MCMC Sampling



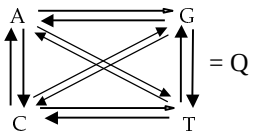
Primer of finding the likelihood of a phylogenetic tree



1. aligned gene- sequence data

```
Sheep      ATGGTGAAAA GCCACATAGG CAGTTGGATC CTGGTTCTCT TTGTGGCCAT
Human      ATGGCGAA-- ----CCTTGG CTGCTGGATG CTGGTTCTCT TTGTGGCCAC
Gorilla    ATGGCGAA-- ----CCTTGG CTGCTGGATG CTGGTTCTCT TTGTGGCCAC
Mink       ATGGTGAAAA GCCACATAGG CAGCTGGCTC CTGGTTCTCT TTGTGGCCAC
```

2. model of sequence evolution



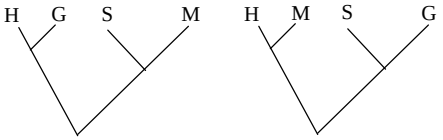
3. the probability of sequence substitutions in a given branch of the phylogeny

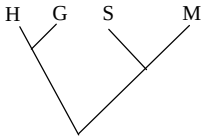
P(t) = Exp[Qt]

4. the likelihood of a given phylogenetic tree

$$L = \prod^{\text{branches}} \mathbf{P(t)} = \prod \mathbf{Exp[Qt]}$$

5. search alternative topologies






```
Sheep      ATGGTGAAAA GCCACATAGG CAGTTGGATC CTGGTTCTCT TTGTGGCCAT
Human      ATGGCGAA-- ----CCTTGG CTGCTGGATG CTGGTTCTCT TTGTGGCCAC
Gorilla    ATGGCGAA-- ----CCTTGG CTGCTGGATG CTGGTTCTCT TTGTGGCCAC
Mink       ATGGTGAAAA GCCACATAGG CAGCTGGCTC CTGGTTCTCT TTGTGGCCAC
```

$$p_i(x \mid T, v, Q) = \sum_{n}^{4^{s-1}} \left[w_{root(i)} \left(\prod_{k=1}^{s} p_{n_k, x_{ki}}(v_k, Q) \right) \left(\prod_{k=1}^{s-2} p_{n'_k, n_k}(v_k, Q) \right) \right]$$

probability of observing ith nucleotide

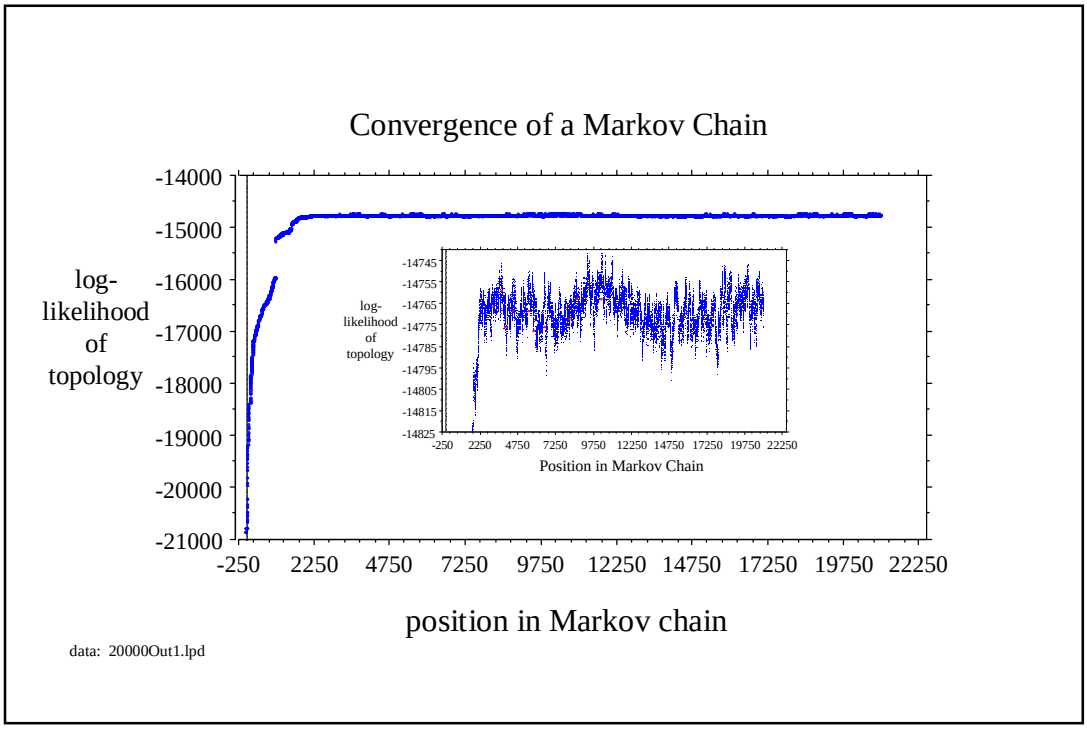
possible assignments of ancestral nodes (e.g., 64)

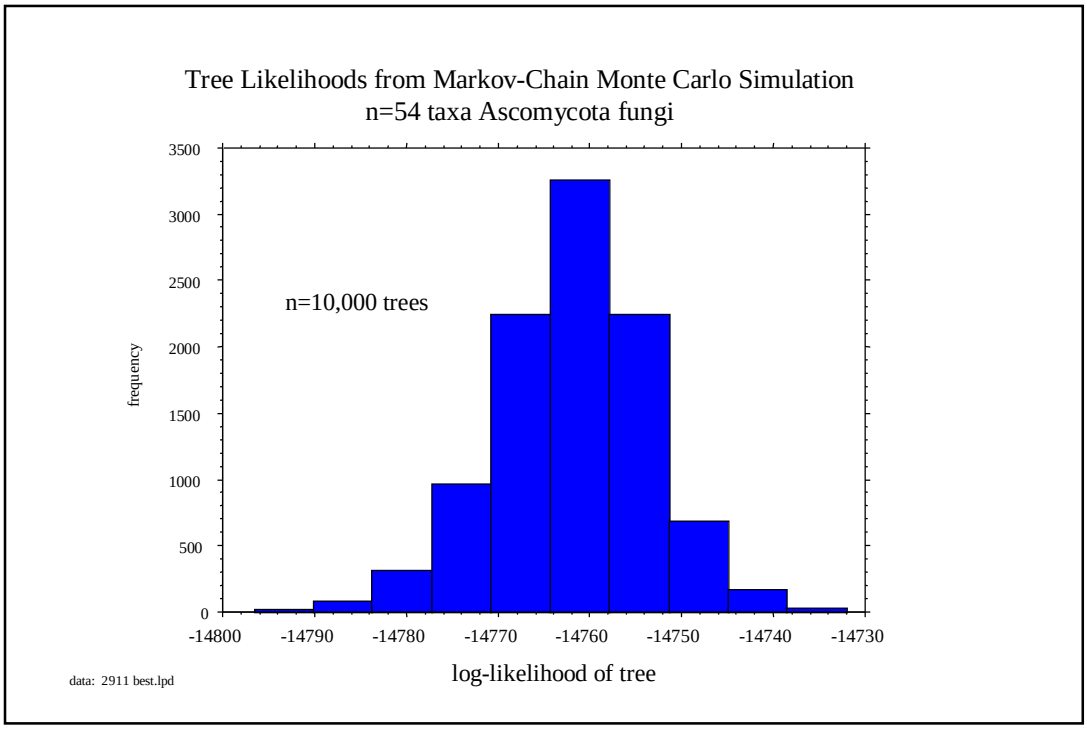
prior weight ith site

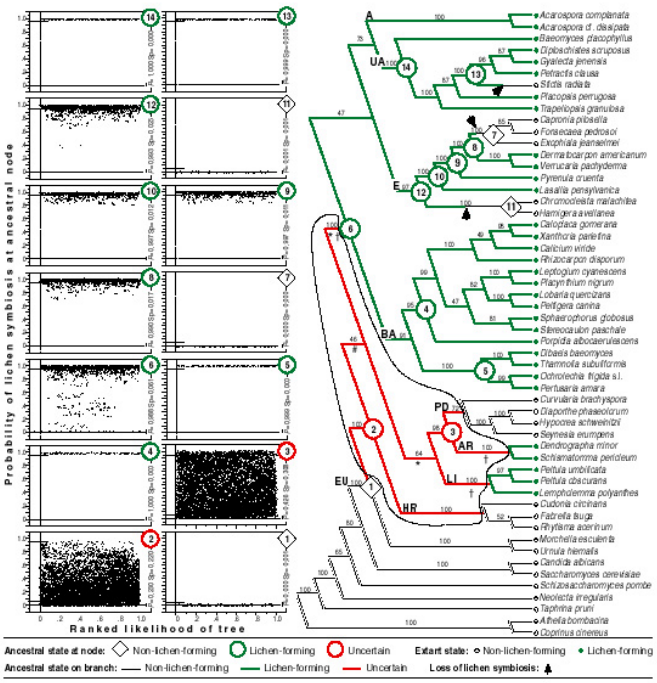
product over s branches leading to species

product over s-2 internal branches

$$L(x \mid T, v, Q) = \prod_i p_i(x \mid T, v, Q)$$ ← product over all i nucleotides


Convergence of a Markov Chain
log-likelihood of topology
-14000
-15000
-16000
-17000
-18000
-19000
-20000
-21000
-250 2250 4750 7250 9750 12250 14750 17250 19750 22250
position in Markov chain
log-likelihood of topology
-14745
-14755
-14765
-14775
-14785
-14795
-14805
-14815
-14825
-250 2250 4750 7250 9750 12250 14750 17250 19750 22250
Position in Markov Chain
data: 20000Out1.lpd

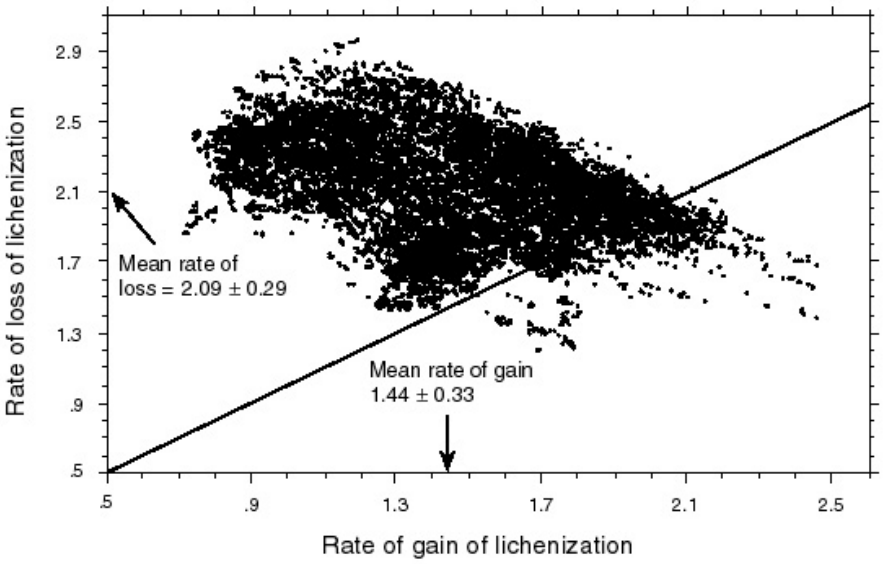

Tree Likelihoods from Markov-Chain Monte Carlo Simulation
n=54 taxa Ascomycota fungi
n=10,000 trees
frequency
3500
3000
2500
2000
1500
1000
500
0
-14800 -14790 -14780 -14770 -14760 -14750 -14740 -14730
log-likelihood of tree
data: 2911 best.lpd



Character Transition Rates
"gains" and "losses" of lichenization

	Single Tree*	MCMC 'Integration'**
gains (q_{01}) =	1.04 (0.05-4.5)	1.47±0.32
losses (q_{10}) =	2.41(0.7-5.6)	2.12 ±0.31

*consensus tree **n=20,000 trees



MCMC Some Issues

Lack of convergence (poor 'mixing)

Tree and parameter proposal mechanisms

Tree and parameter updating schedules

Detecting convergence

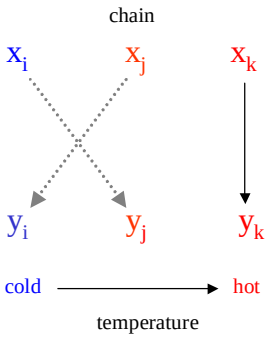
Metropolis-Coupled Markov Chain Monte Carlo (MCMCMC)

Given *m* simultaneous Markov chains, update chains, then swap states among a randomly chosen pair i and j each iteration according to:

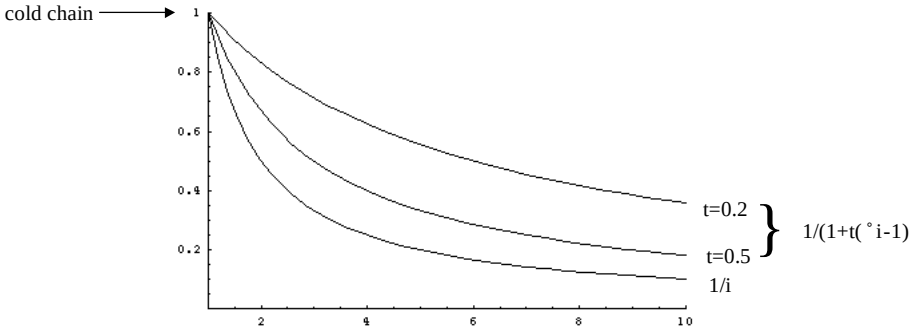
$$R = \min\left\{1, \frac{f_i(y_i)\, f_j(y_j)}{f_i(x_i)\, f_j(x_j)}\right\}$$

{likelihood ratio chain i * likelihood ratio chain j}

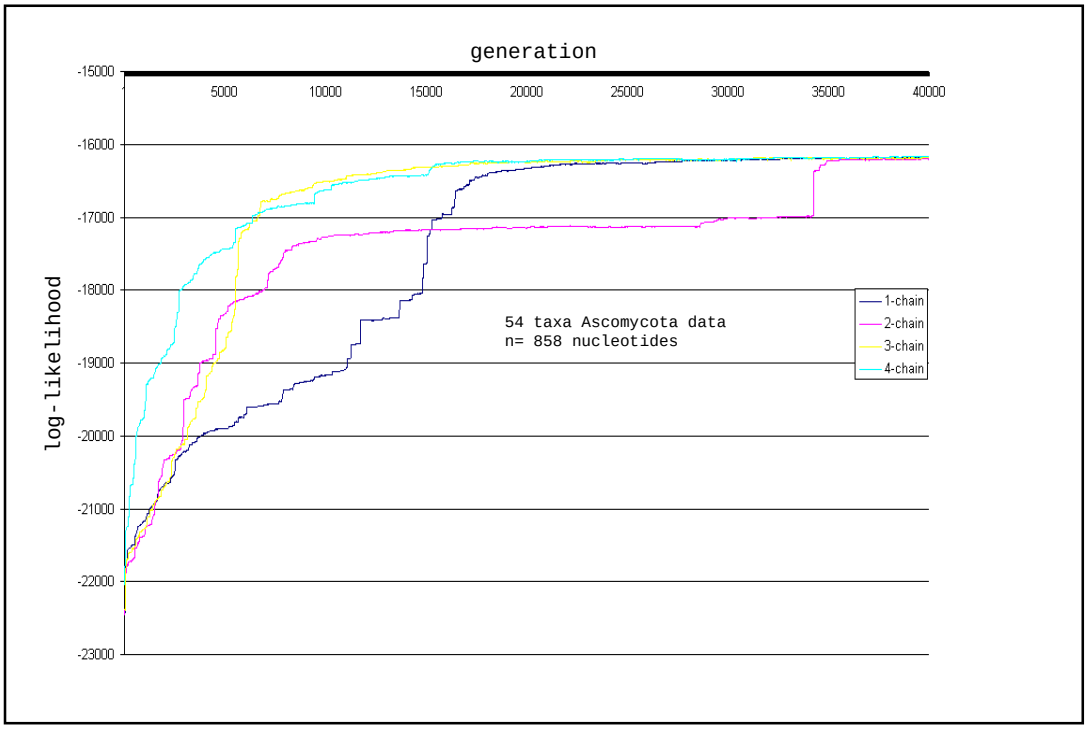
probability of swapping with chain i = R * 2/m

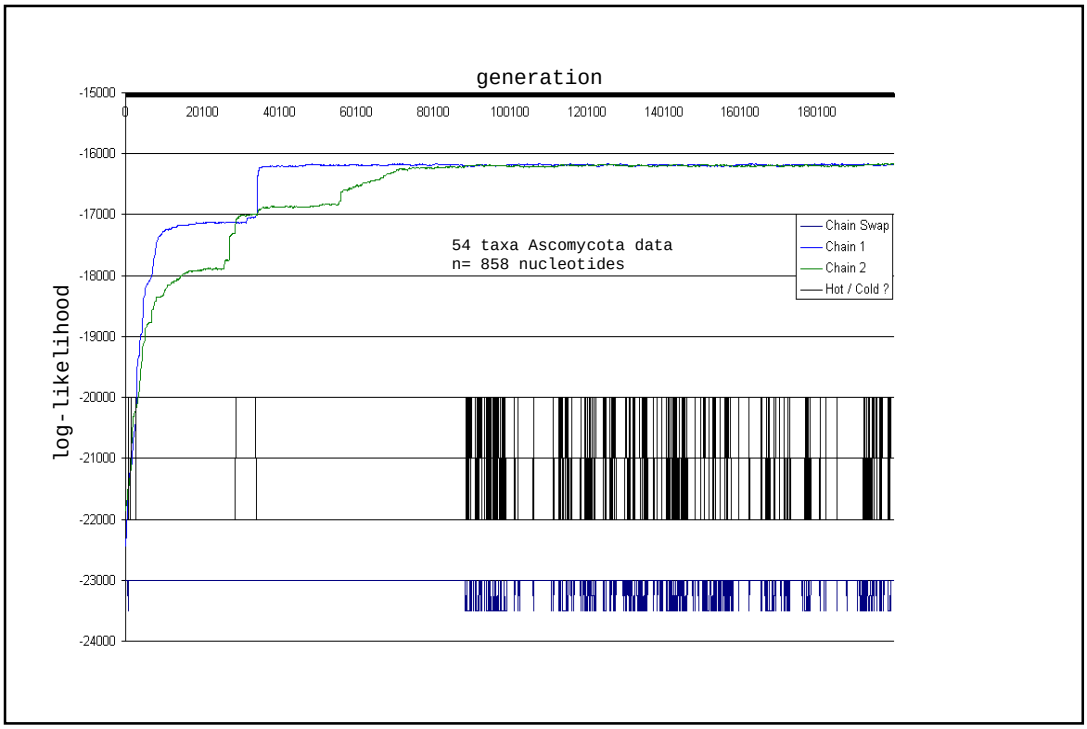


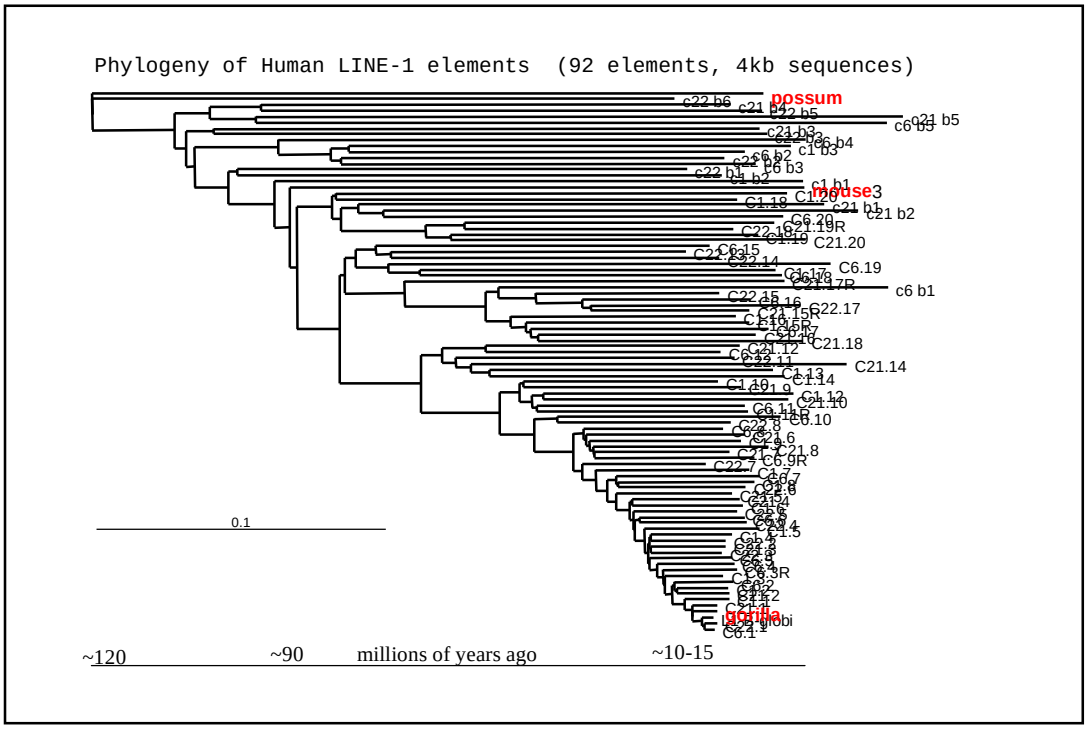
‘Temperatures’ of heated chains

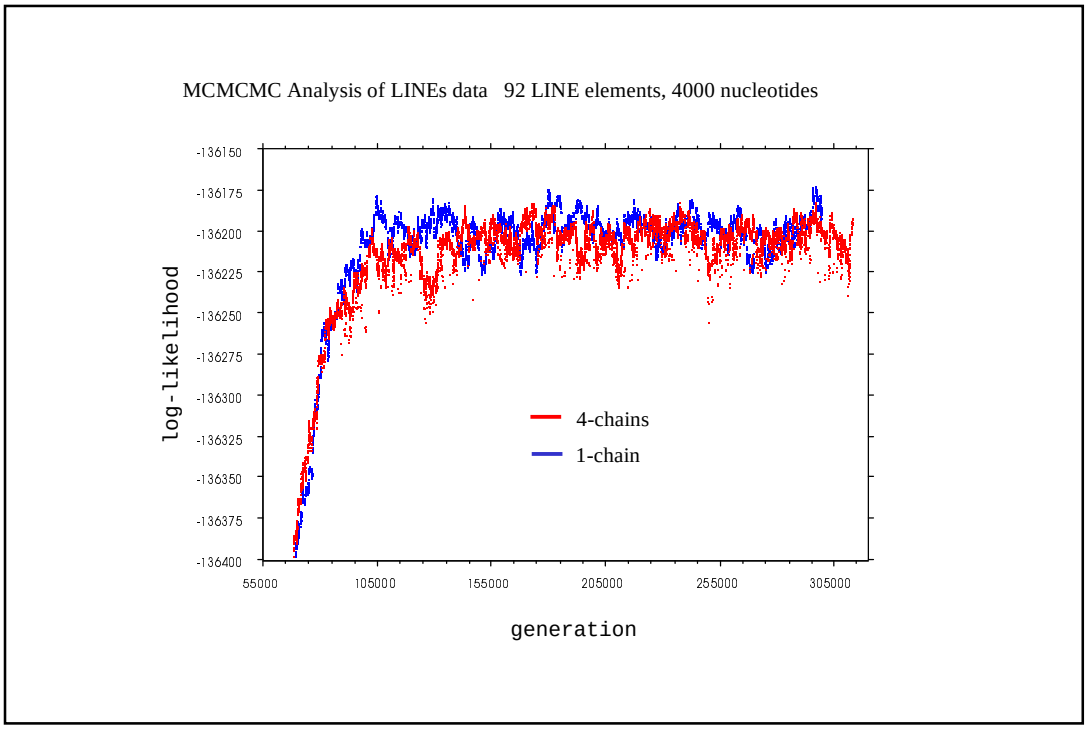


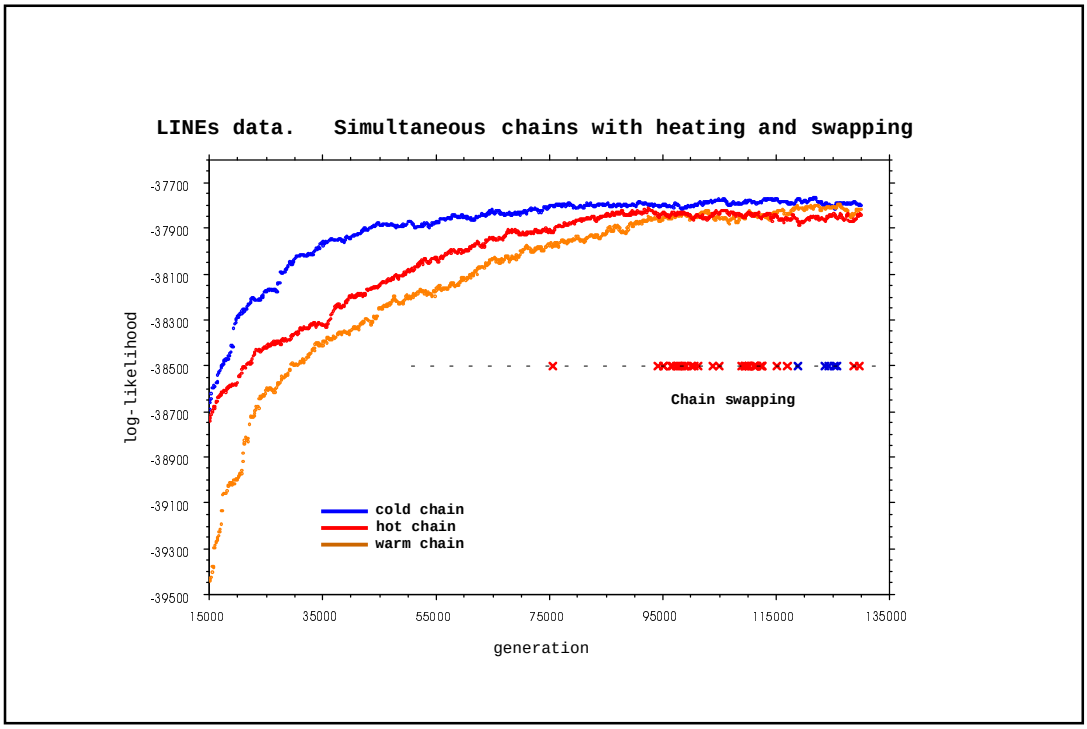
number of chains, i

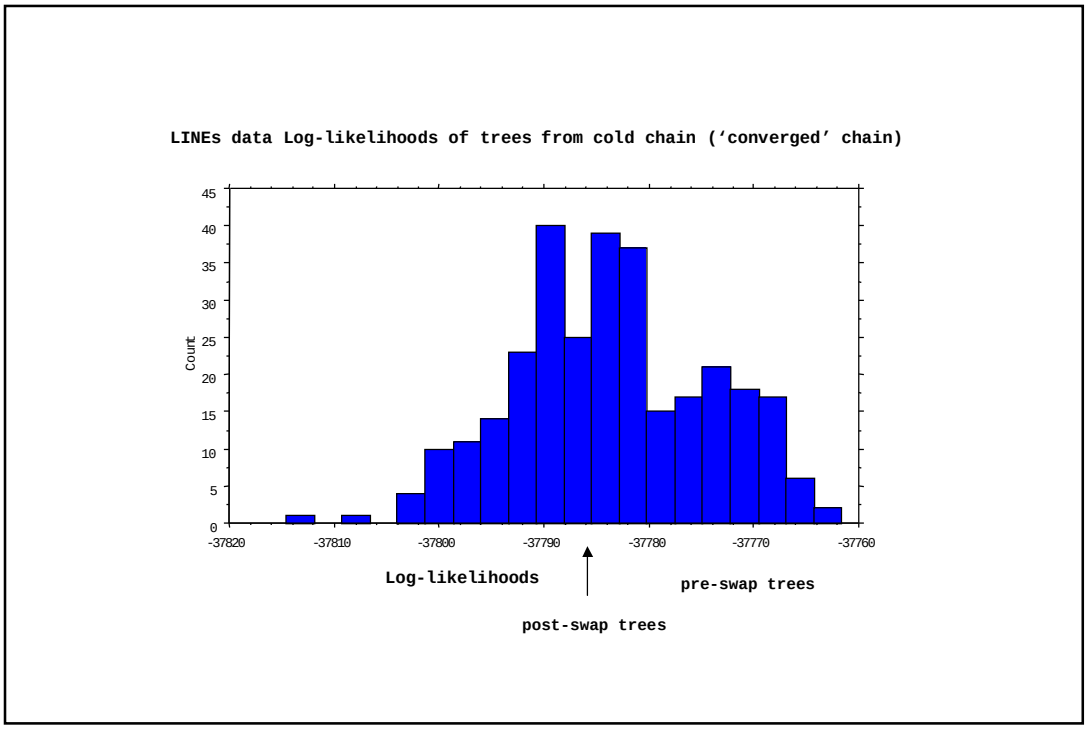

generation
log-likelihood
-15000
-16000
-17000
-18000
-19000
-20000
-21000
-22000
-23000
5000
10000
15000
20000
25000
30000
35000
40000
54 taxa Ascomycota data
n= 858 nucleotides
1-chain
2-chain
3-chain
4-chain

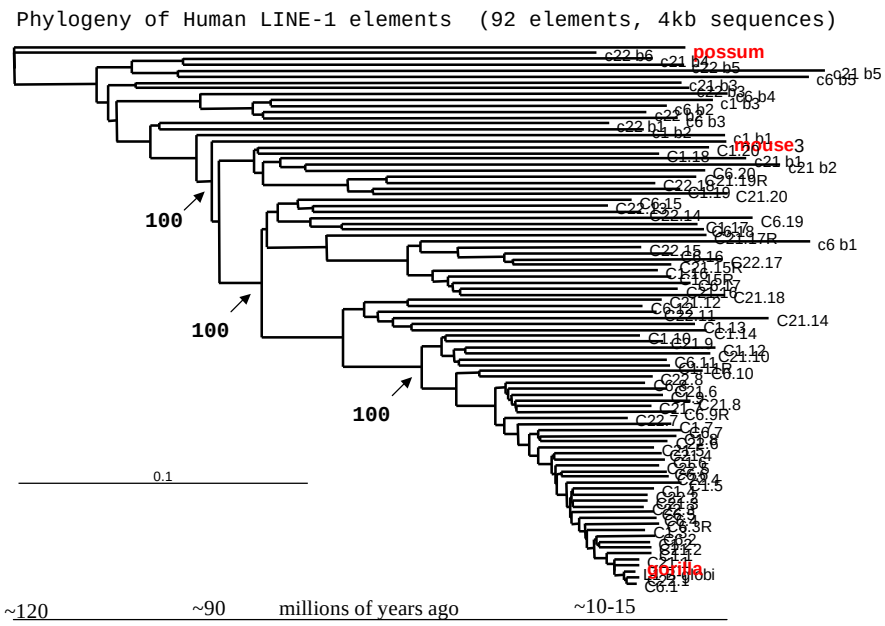

generation
log-likelihood
-15000
-16000
-17000
-18000
-19000
-20000
-21000
-22000
-23000
-24000
0
20100
40100
60100
80100
100100
120100
140100
160100
180100
54 taxa Ascomycota data
n= 858 nucleotides
Chain Swap
Chain 1
Chain 2
Hot / Cold ?


Phylogeny of Human LINE-1 elements (92 elements, 4kb sequences)
possum
mouse
gorilla
0.1
~120
~90
millions of years ago
~10-15


MCMCMC Analysis of LINEs data 92 LINE elements, 4000 nucleotides
log-likelihood
generation
4-chains
1-chain
-136150
-136175
-136200
-136225
-136250
-136275
-136300
-136325
-136350
-136375
-136400
55000
105000
155000
205000
255000
305000


LINEs data. Simultaneous chains with heating and swapping
log-likelihood
-37700
-37900
-38100
-38300
-38500
-38700
-38900
-39100
-39300
-39500
15000
35000
55000
75000
95000
115000
135000
generation
Chain swapping
cold chain
hot chain
warm chain


LINEs data Log-likelihoods of trees from cold chain ('converged' chain)
Count
45
40
35
30
25
20
15
10
5
0
-37820
-37810
-37800
-37790
-37780
-37770
-37760
Log-likelihoods
pre-swap trees
post-swap trees



Some topics and issues

MCMC (getting stuck and slow progress)

Tree proposal algorithms (random, deep, shallow, ???)

Alternation among suites of parameters (topology, branch lengths, model parameters) – what schedule to use?

Optimisation alternating with bouts of M-H selection?

MCMCMC (highly inefficient)

Tree swapping: encourage early tree swapping?

Once converged, cool heated chains and use in inference?