Detecting Conflicting Phylogenetic Signals a ‘mixture-model’ approach

Mark Pagel and Andrew Meade
Reading University, England
m.pagel@reading.ac.uk

Some causes of conflicting phylogenetic signals

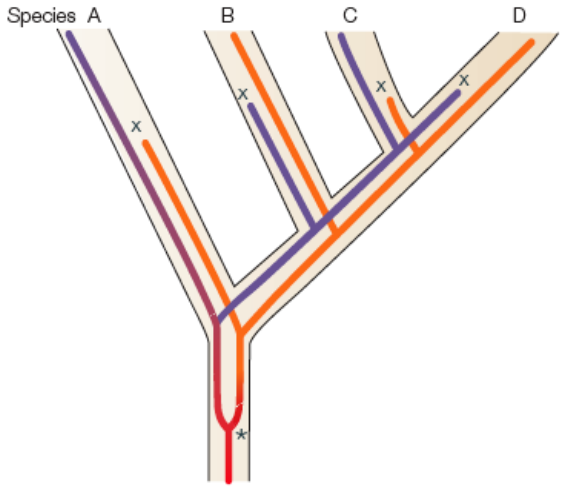
-- gene trees vs species trees (drift, coalescent events)

-- convergent evolution (e.g., lysozyme in cows and monkeys)

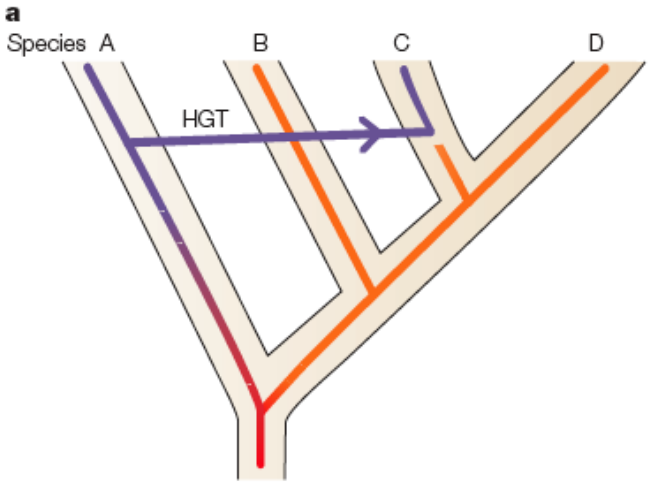
-- gene duplication

-- lateral gene transfer

gene-duplication



lateral gene transfer



Mixture models in phylogenetic inference

- mixture models
- a mixture model in ***T***
- simulation study
- performance at estimating ***T'***s
- application to ecdysozoa/coelomata and prokaryotes

Mixture models in phylogenetic inference

$$L(\mathbf{Q}) \propto P(\mathbf{D}|\mathbf{Q})$$

conventional likelihood model $$P(\mathbf{D}|\mathbf{Q},T) = \prod_i P(\mathbf{D}_i|\mathbf{Q},T)$$

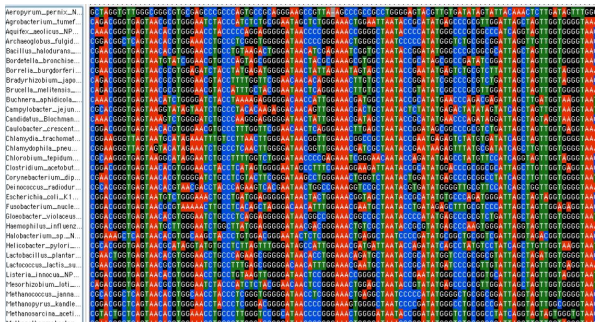
mixture model in **Q** (Pagel and Meade, 2004) $$P(\mathbf{D}|\mathbf{Q}_1,\mathbf{Q}_2,\ldots\mathbf{Q}_J,T) = \prod_i \sum_j w_j P(\mathbf{D}_i|\mathbf{Q}_j,T)$$

mixture model in ***T*** $$P(\mathbf{D}|\mathbf{Q},T_1,T_2\ldots T_J) = \prod_i \sum_j w_j P(\mathbf{D}_i|\mathbf{Q},T_j)$$

How does the multiple-topologies mixture model work?

conventional MCMC inference

alignment



T

iteration

1
2
3
.
.
.
.
.
n

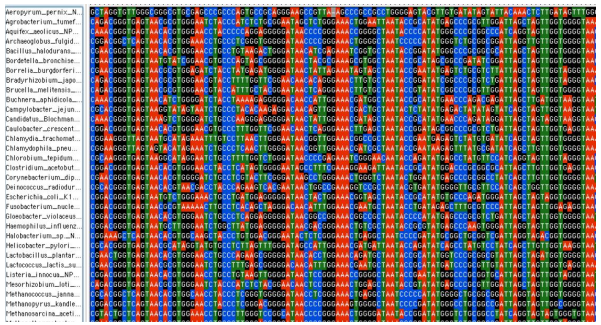
likelihood

$L_{i=1}=P(D|T_1)$
$L_{i=2}=P(D|T_2)$
$L_{i=3}=P(D|T_3)$
.
.
$L_{i=n}=P(D|T_n)$

---

multiple-topologies mixture model

alignment



k distinct Markov chains

$T_1 \quad T_2 \quad .....T_k$

iteration

1
2
3
.
.
.
.
.
n

....

combined likelihood

$L_{i=1}=w_1L_1+w_2L_2+..w_kL_k$
$L_{i=2}=w_1L_1+w_2L_2+..w_kL_k$
$L_{i=3}=w_1L_1+w_2L_2+..w_kL_k$
.
.
$L_{i=n}=w_1L_1+w_2L_2+..w_kL_k$

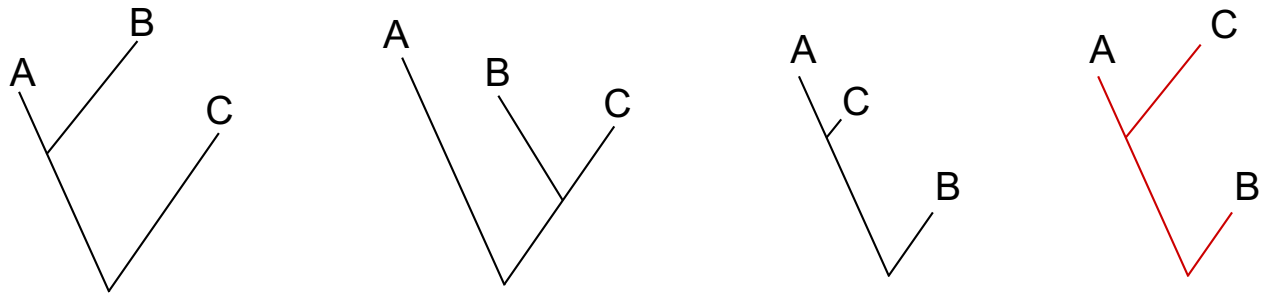
Converged combined likelihood:

$L_{i=n}=w_1L_1+w_2L_2+..w_kL_k$

- **combined likelihood allows tree to diverge if this increases *L***
- **separately calculate posterior probabilities of each sample of trees**

Mixture model in ***T***

mixture model in ***T*** $P(\mathbf{D}|\mathbf{Q}, T_1, T_2 \ldots T_J) = \prod_i \sum_j w_j P(\mathbf{D}_i|\mathbf{Q}, T_j)$



What does the model deliver?

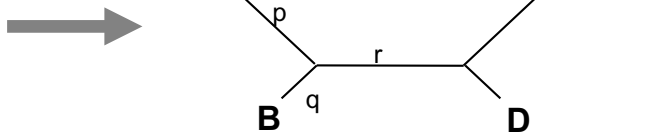
- variation in the topology
- variation in rates across the tree: a non-parametric covarion?
- variation in rates among sites: easy
- variation in patterns (Q): easy

Questions

- does it work?
- can we estimate sets of topologies/branch lengths?
- are there a small number of sets in real data?

Does the multiple-topology model work?

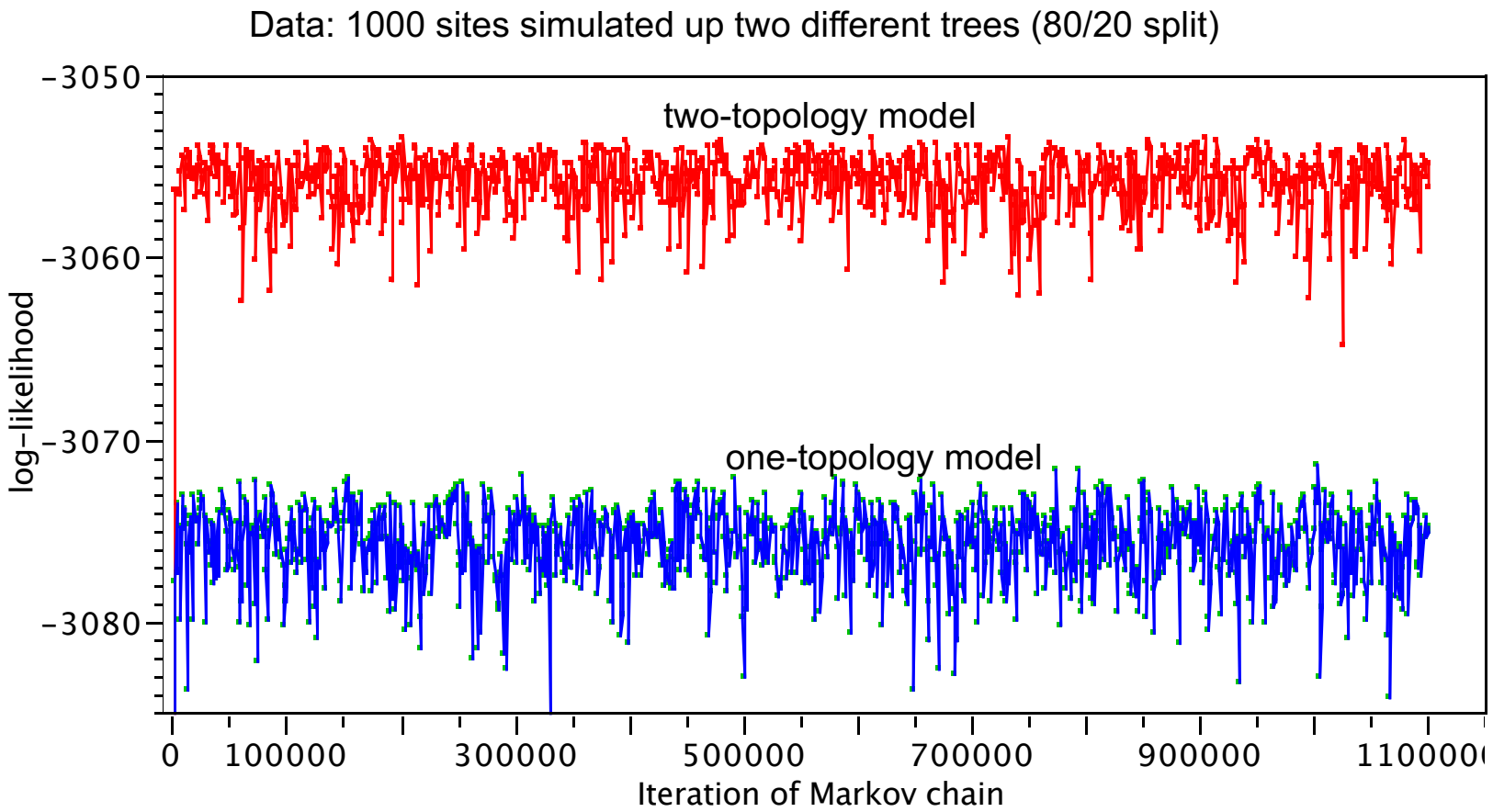
<u>Simulation of two-topologies</u>



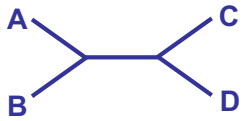
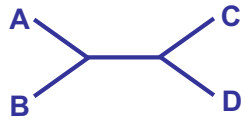
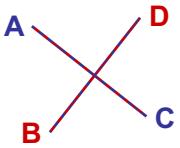
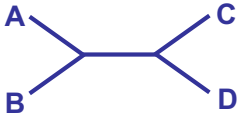
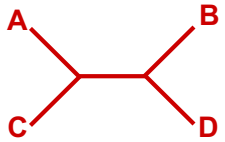
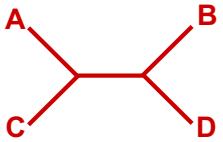
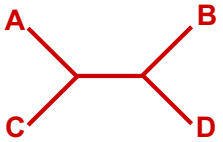
- two four taxon trees
- generate 20,000 random alignments of 1000 sites. Vary the proportion of sites per topology
- analyse with MCMC using: one-tree model (conventional), two-tree model

Can we detect two trees?

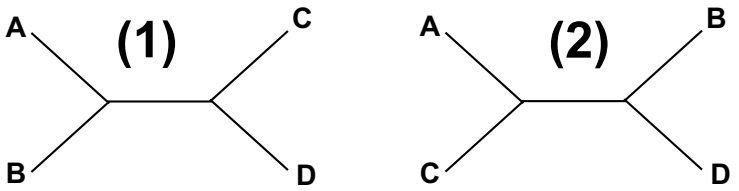
Converged Markov chains for one and two-topology models

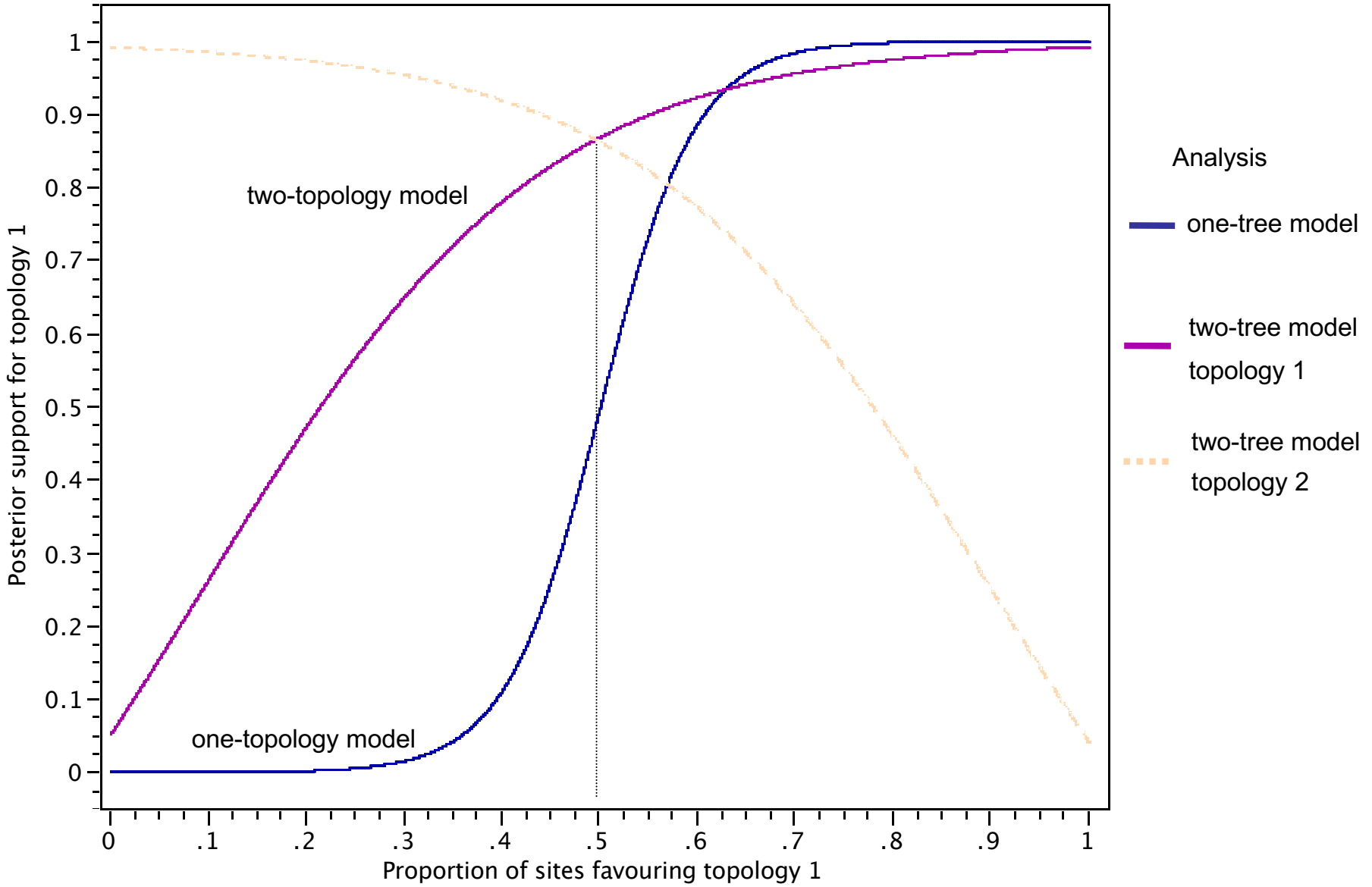


Coarse summary of simulation results

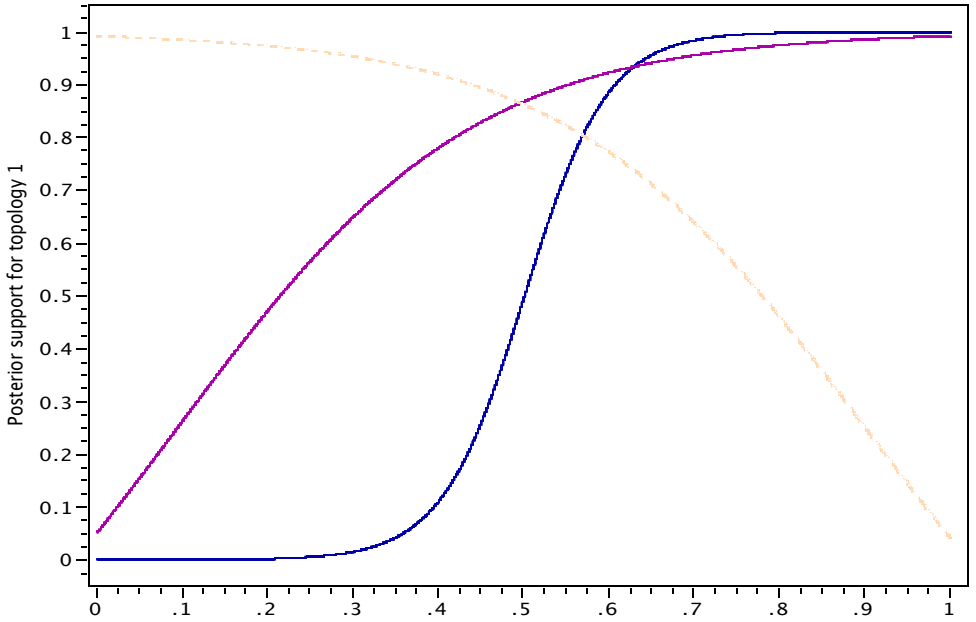
Proportion of sites favouring topology 1	topology 2	Outcome: one-topology model	Outcome: two-topology model
100	0		
50	50		 
0	100		

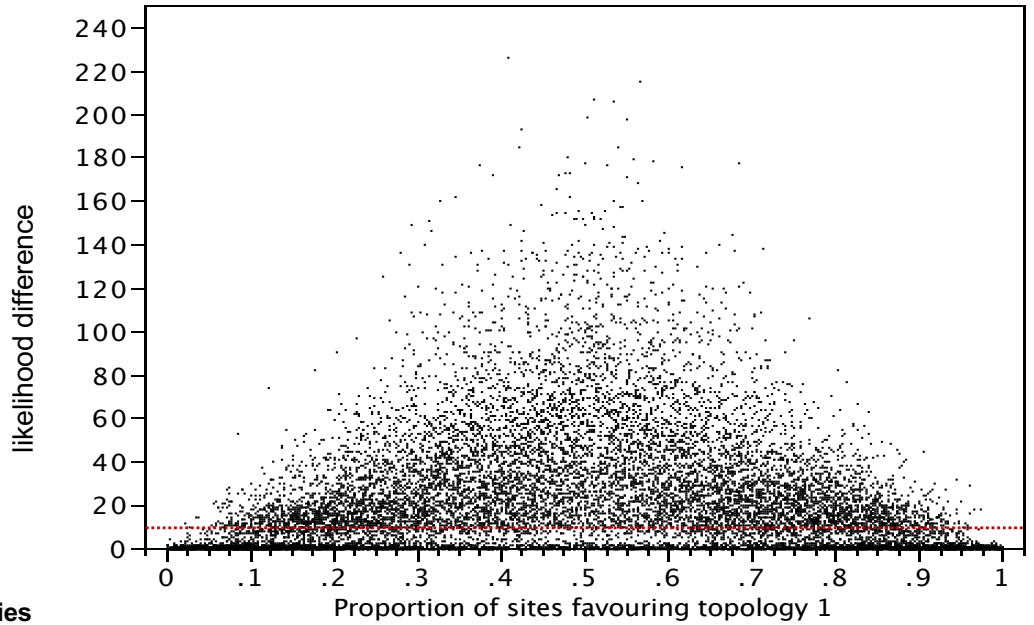
Performance of one and two-topology models given conflicting phylogenetic signal





statistical power: ‘significant’ differences detected with ≤ 10% of sites





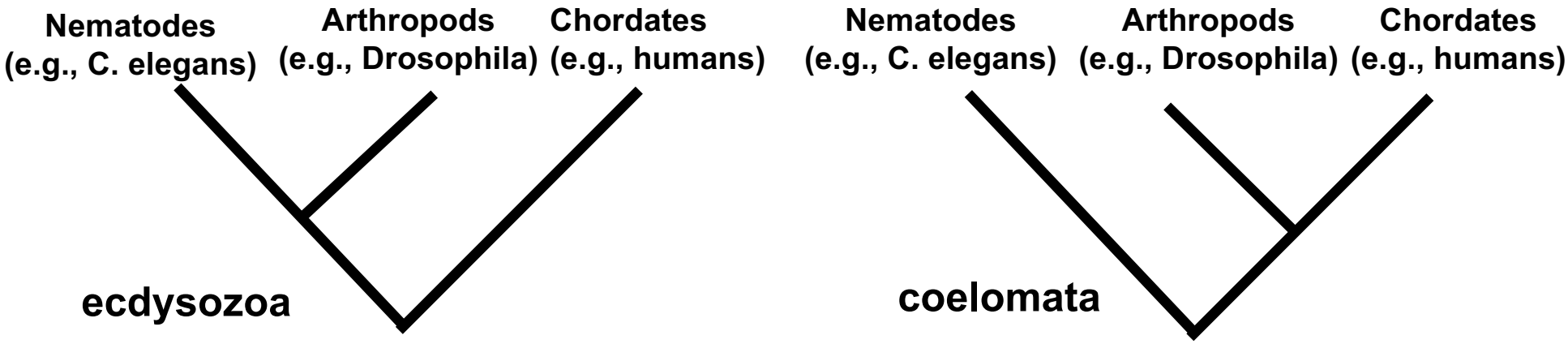
Note: data generated on two topologies

Estimating the number of topologies

Data: 1000 sites simulated up two different trees of 50 tips, 60/40 split

Model	Log-likelihood	s.d	W_1	W_2	W_3
One-topology	-85222.8	7.2			
Two topologies	-64920.4	9.9	0.599294	0.400706	
Three topologies	-64909.9	14.8	0.598558	0.399572	0.001871

Conflicting signal for ecdysozoa/coelomata phylogenies



---

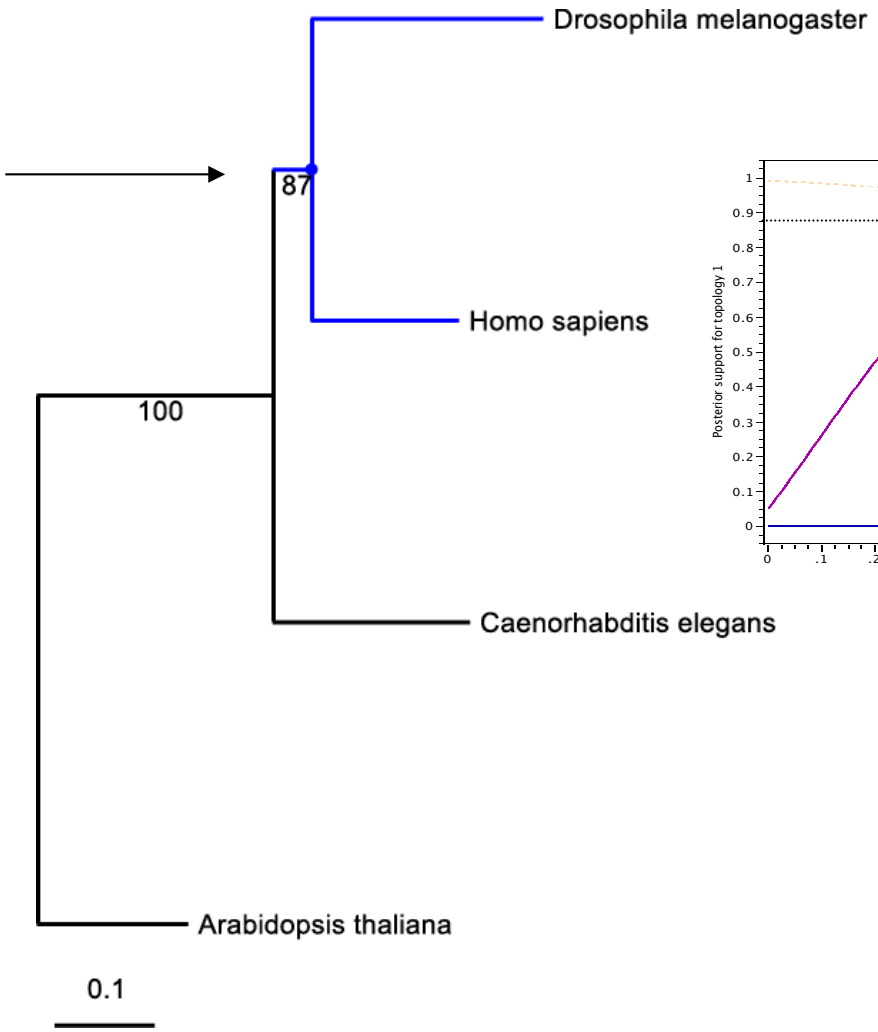
dataset: 13,946 BP alignment of fourteen genes from four eukaryotic species, Homologene (NCBI) database

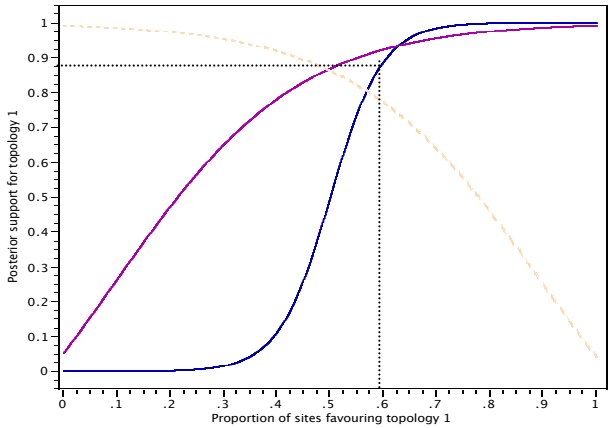
Number	Name	Length
1	Beta tubulin - 2083	536
2	EEF1A1 - 68181	926
3	HSP40-4B-56013	700
4	HSP70-4-1624	1696
5	HSP70-5-3908	1310
6	HSP70-9B-39452	1304
7	HSP70-90	1162
8	HSP90-1A	1414
9	14s	302
10	18s	304
11	40s	522
12	RNA-Bind-Motif-19	1974
13	TUBA6	896
14	TUBB.	900

one-topology model: likelihood = -45435.9±2.58

data from Homologene database (NCBI)

posterior support
for coelomata = 87

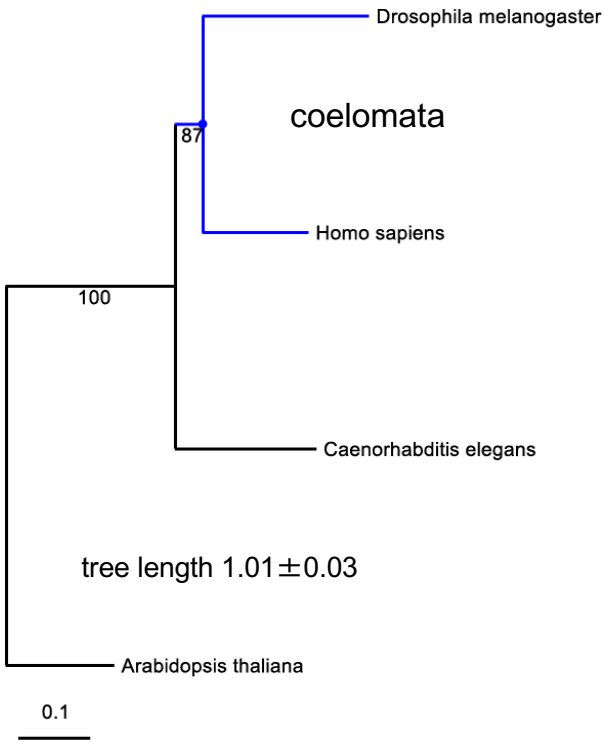




tree length 1.01±0.03

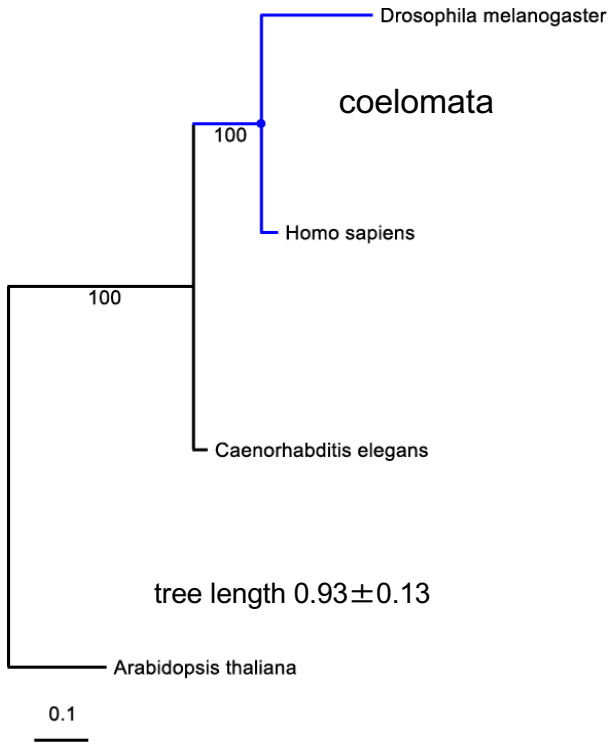
One-topology model versus two-topology results

likelihood = -45435.9±2.58

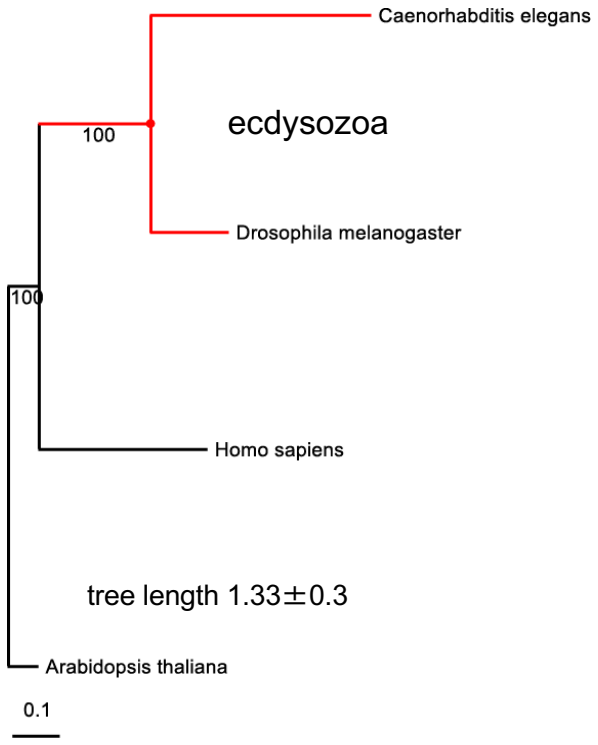


likelihood = -45382.8±3.01

weight = 0.57



weight = 0.43



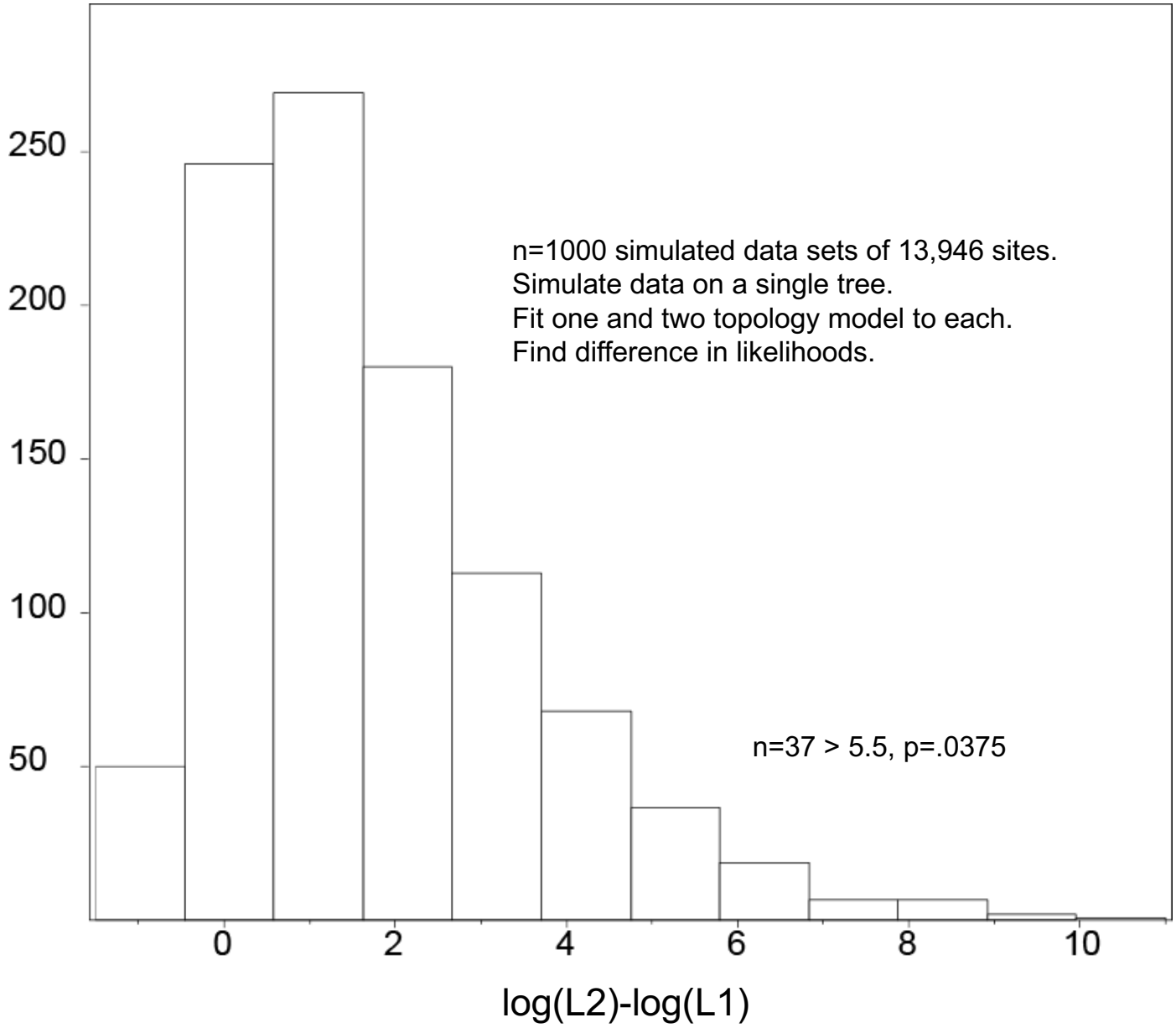
note: two recent papers support ecdysozoa

Δlog-L = 53.1 log units

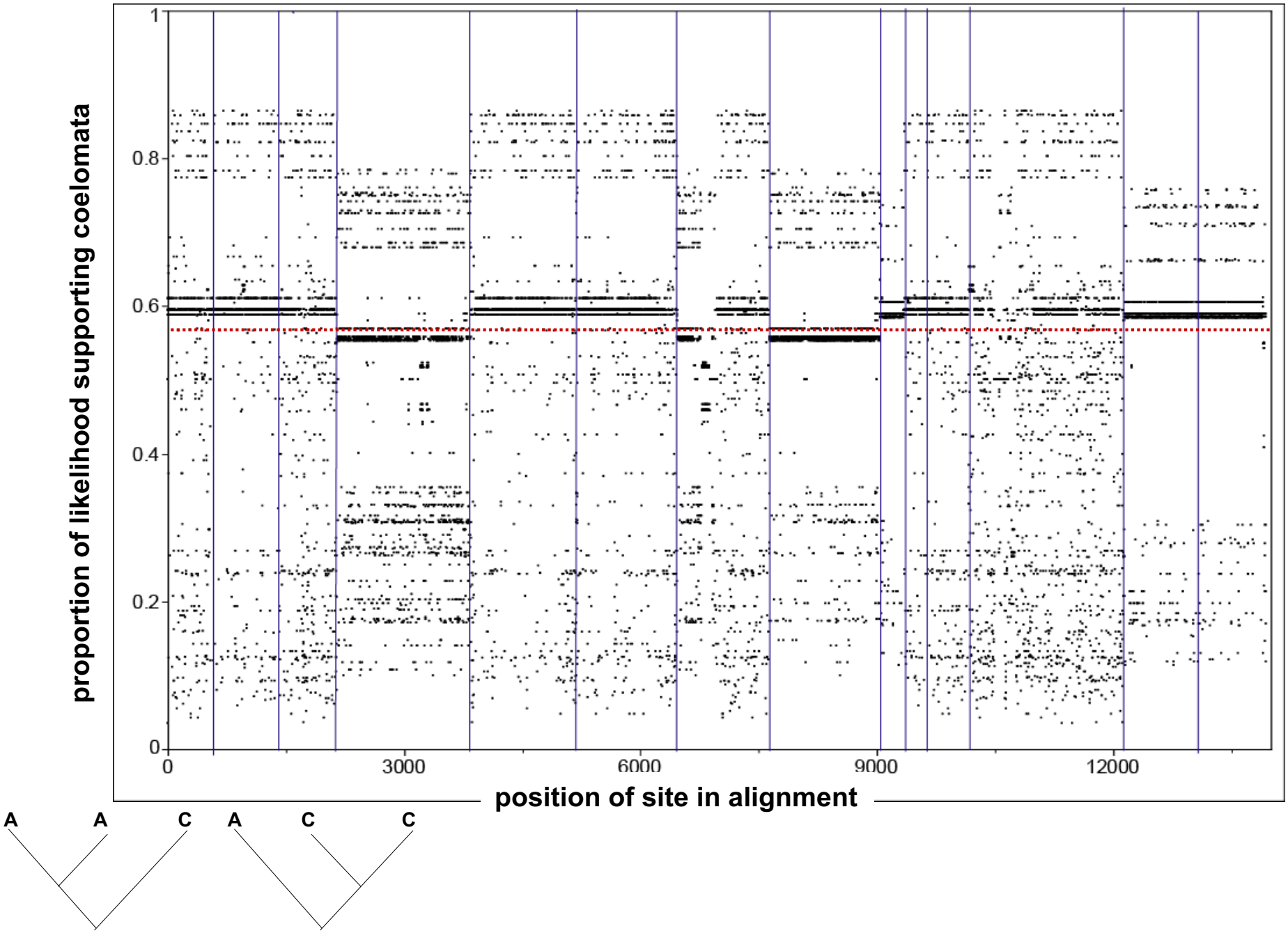
BIC requires $-\log M_2 - \log M_1 > \frac{\Delta p \log(n)}{2} = \frac{(6+2)\log(4)}{2} = 5.55$

note: 6 branch lengths + n-1-p parameters to specify the topology (Charleston conjecture)

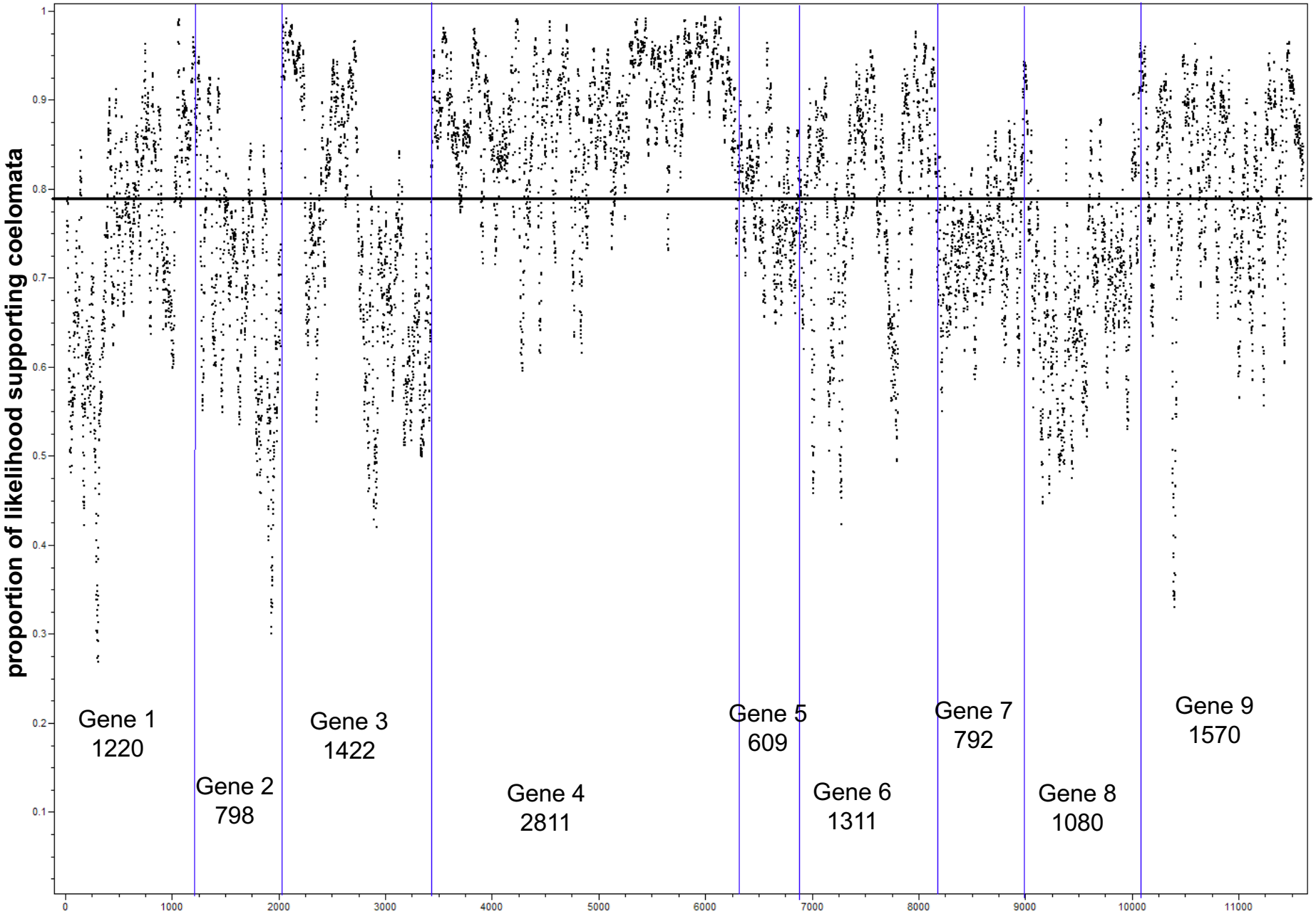
Simulation of null hypothesis distribution of differences in likelihood



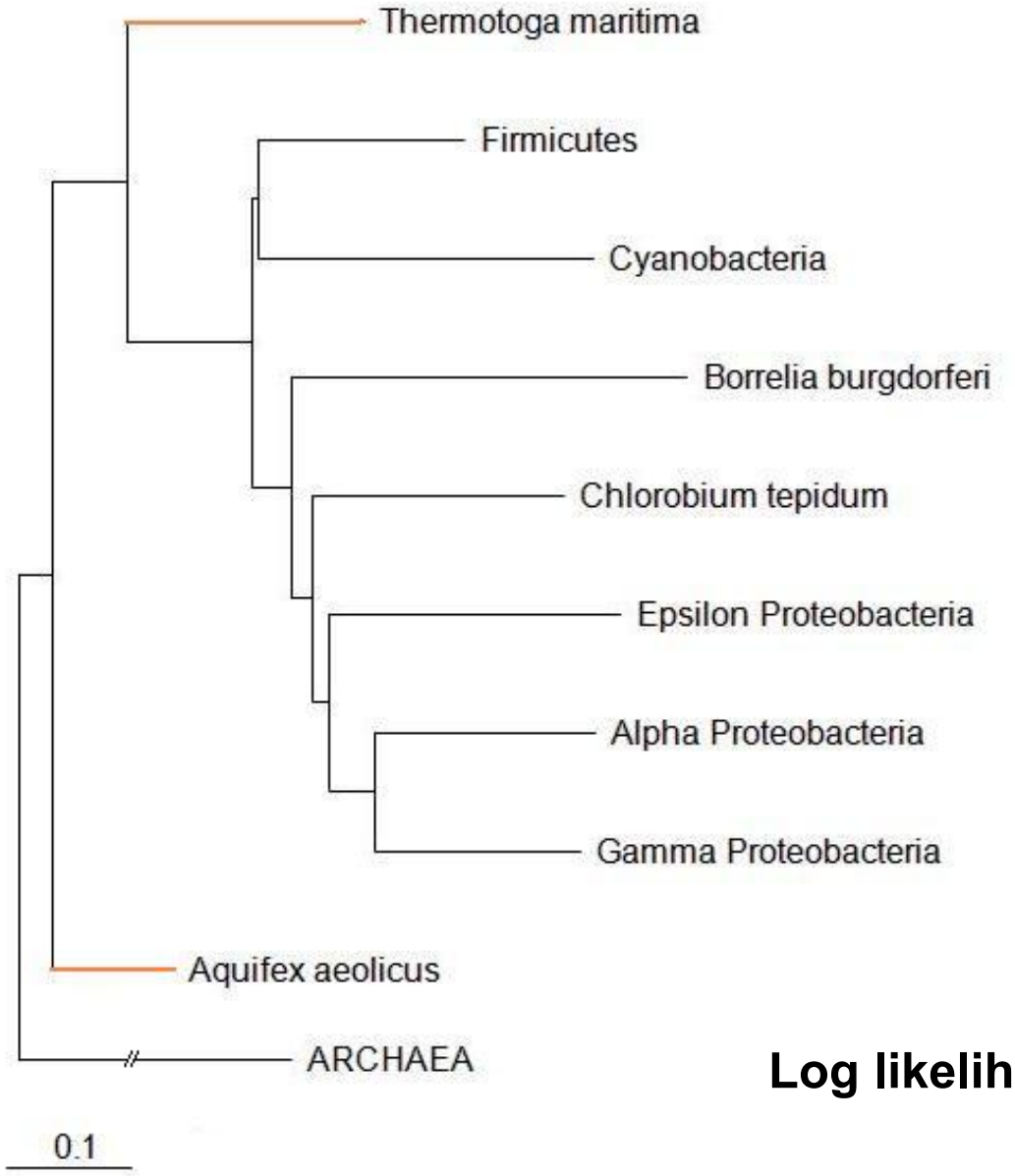
Site likelihoods: showing proportion of likelihood favouring coelomata (4 taxon tree)



Site likelihoods for expanded ecdysozoa/coelomata data

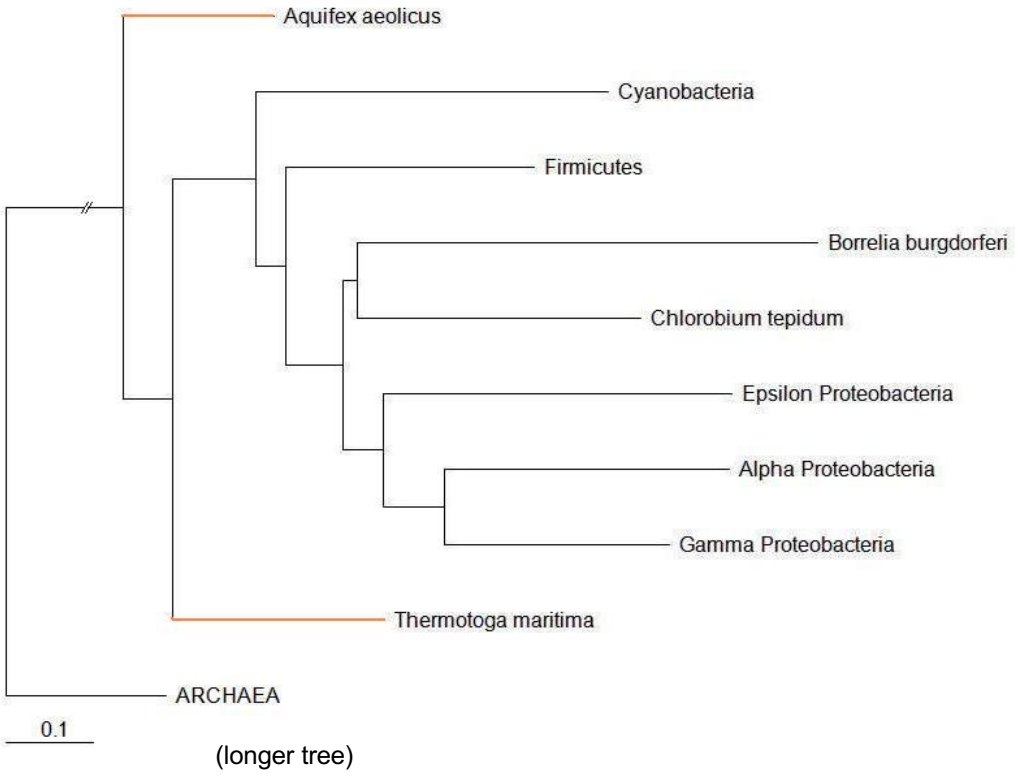


Prokaryote phylogenetics: placement of hyperthermophiles



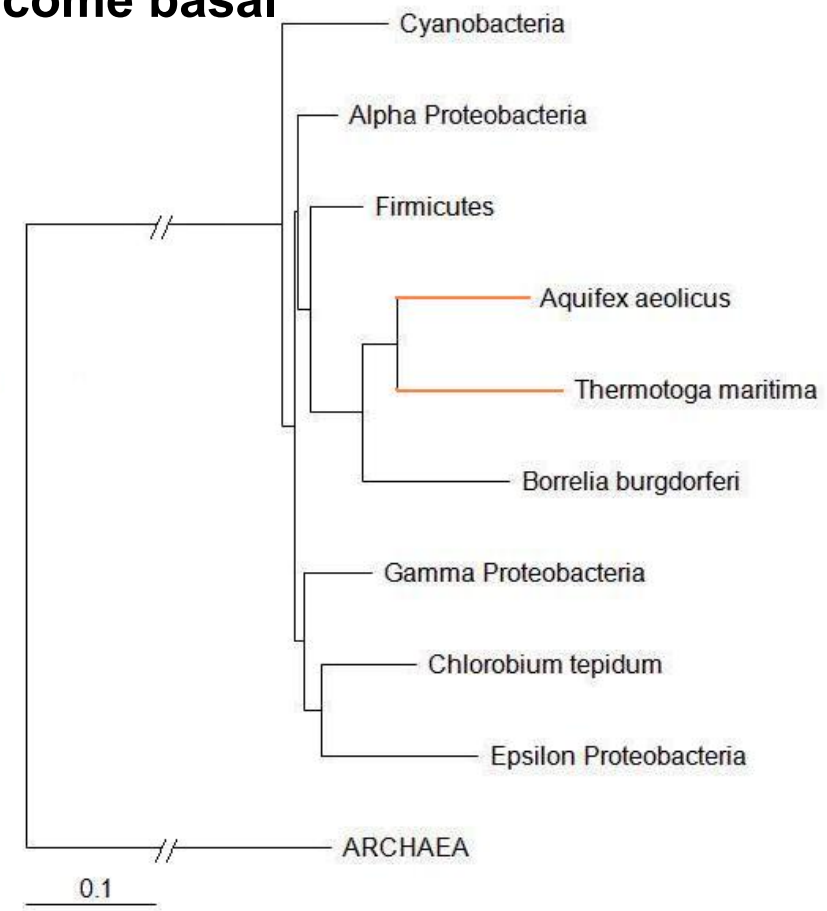
Log likelihood =-199,345

Prokaryotes: two topologies showing that hyperthermophiles become derived and cyanobacteria become basal



(longer tree)

w_1=0.55



w_2=0.45

Log likelihood =-194,022 ΔL = 5323

Summary

Multiple topologies model seems to work

Can estimate more than one topology and identify conflicting signal directly

Some questions

How to test?

How many distinct trees in real data?

Research supported by

Biotechnology and Biological Sciences Research Council (UK)