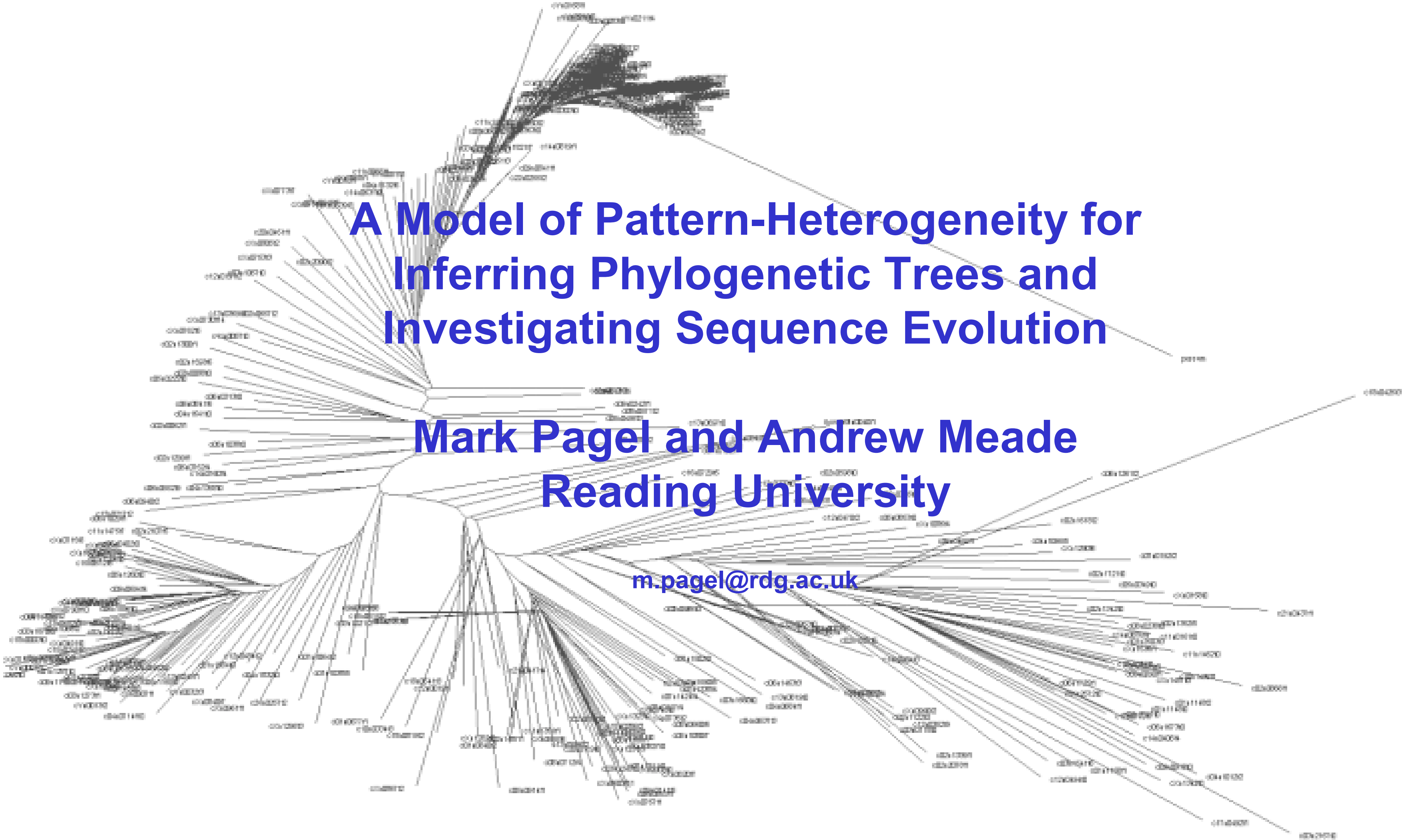
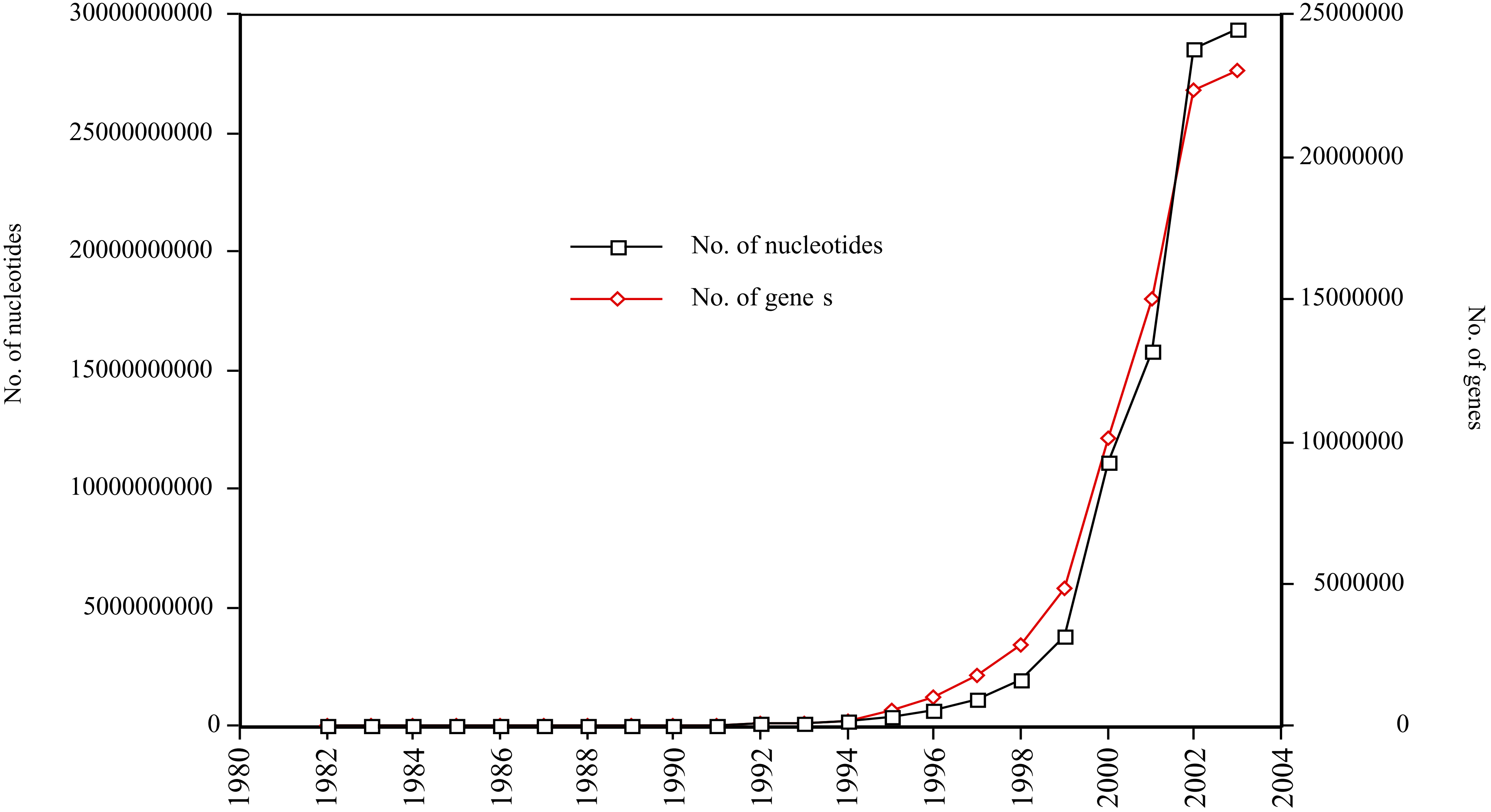


A Model of Pattern-Heterogeneity for Inferring Phylogenetic Trees and Investigating Sequence Evolution

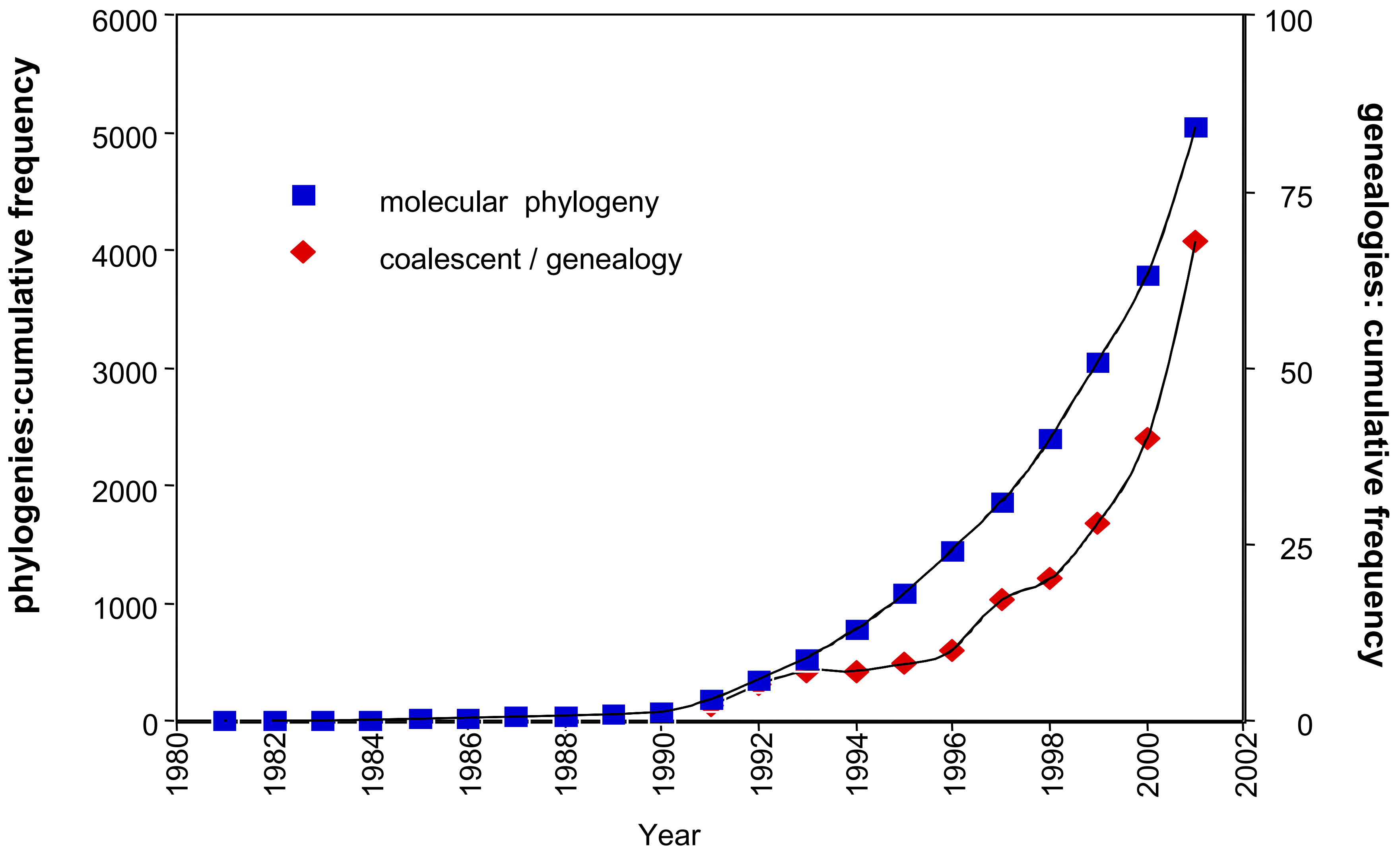
Mark Pagel and Andrew Meade
Reading University

m.pagel@rdg.ac.uk

The growth of GenBank



The use of molecular phylogenies in biological research



Overview: Phylogenetic Inference from gene sequences

1. **Homogeneous Process Model**:

```
Sp 1  AACGTTGTCCCTT
Sp 2  AACGTTCCTTGCT
..
Sp n  CCGGTTGCAAGCT
```

+

	A	**C**	**G**	**T**
A	—	q_{ac}	q_{ag}	q_{at}
C		—	q_{cg}	q_{ct}
G			—	$q_{gt}=1$
T				—

→

2. **Modifications to basic homogeneous model**

a) **rate-heterogeneity** (Yang, 1994): apply homogeneous model but allow rates to vary over sites according to a discrete gamma. Equivalent to fitting *k* rate matrices to each site, where the matrices differ from each other only by a proportional scalar

b) **partitioning data**: apply a different rate matrix to different sites, chosen beforehand by the investigator

Limitations to homogeneous, rate-heterogeneity and partitioning approaches

Homogeneous model: not all sites may evolve according to the same model

Gamma model: rate variation may not be ‘gamma’ or sites may vary in some other way

Partitioning data: presumes investigator knows with certainty which model best applies to each site

Pattern-Heterogeneity Model of Gene-Sequence Evolution

Allow for different sites to evolve in *qualitatively* (or quantitatively) different ways without prior partitioning by the investigator.

Method: fit more than one model of evolution to each site, summing the likelihood over all models. Allow the data to determine the ‘best’ model for each site.

Motivation for model*:* *Pattern-heterogeneity* model will always equal or better the performance of homogeneous, gamma rate heterogeneity and partitioning models. Frequently yields substantial improvements (100’s of log-units)

Applications

Phylogenetic inference

Detecting regions within genes that evolve differently

Detecting differences among genes

An example of pattern-heterogeneity

Sequence data

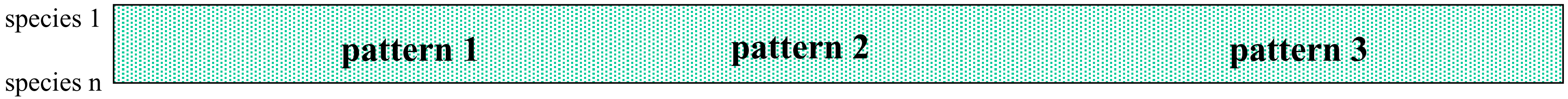
Sp 1   AA**C**GTTG**TCC**CTT
Sp 2   AA**C**GTTC**CTT**GCT
..
Sp n   CC**G**GTTG**CAA**GCT

Two transition rate matrices

	A	**C**	**G**	**T**
A	—	**4.97** **2.11**	**3.41** **3.13**	**0.82** **0.34**
C		—	**0.35** **3.87**	**2.82** **1.49**
G			—	**1** **1**
T				—

Applications of pattern-heterogeneity model

Single gene alignment



species 1

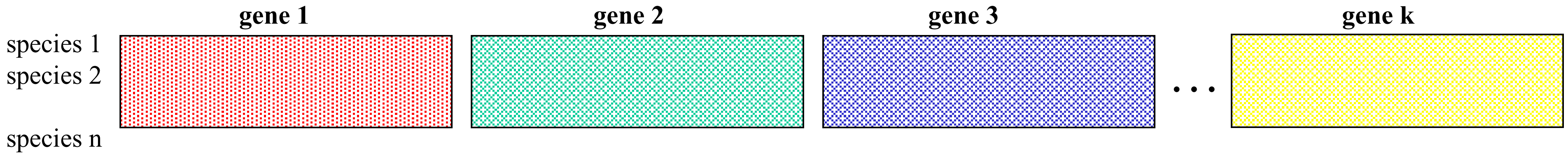
species n

pattern 1

pattern 2

pattern 3

Concatenated gene alignment



gene 1

gene 2

gene 3

gene k

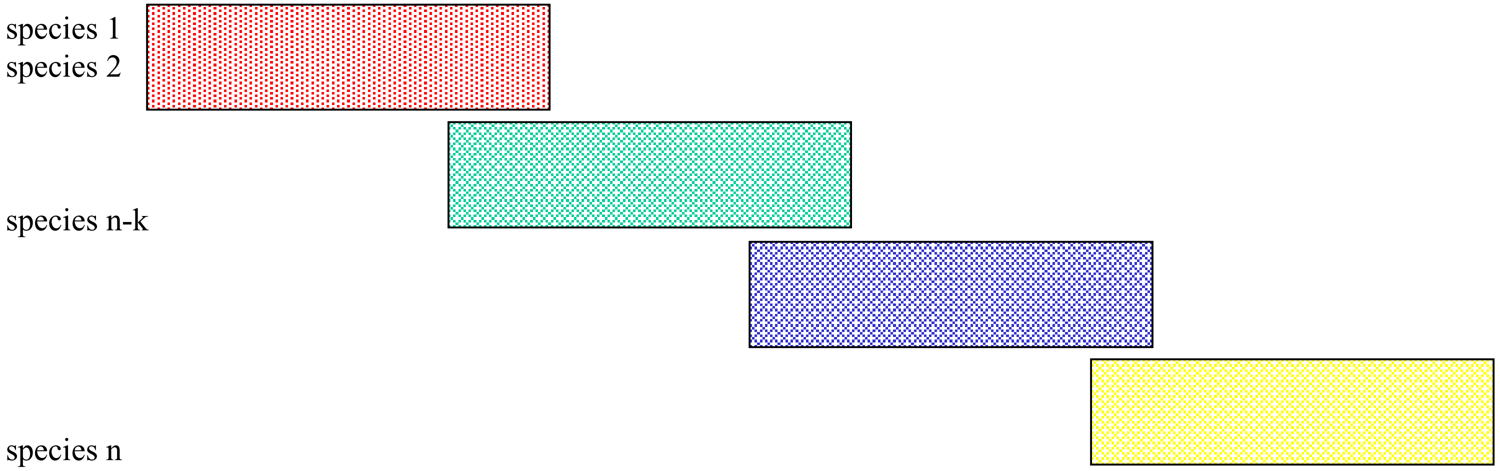
species 1

species 2

species n

. . .

"Supermatrix" alignment



species 1

species 2

species n-k

species n

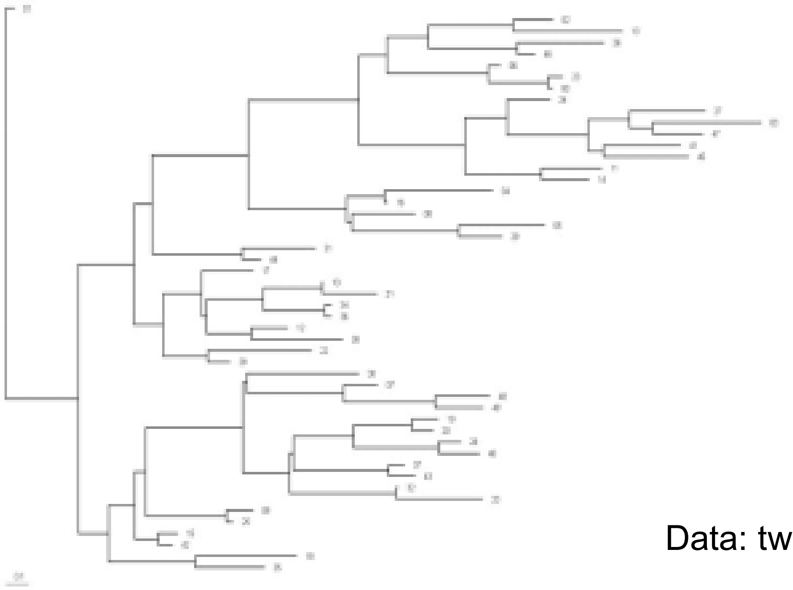
Testing the Pattern-Heterogeneity Model: detecting pattern-heterogeneity in simulated data

Method:

I) generate simulated gene-sequence data on a random tree according to two different models of sequence evolution, creating two ‘genes’ with different patterns of substitutions

2) analyse simulated data using homogeneous, gamma rates and pattern-heterogeneity model. Evaluate trees by their likelihood

Tree used in simulations



rate matrices

	A	C	G	T
A	—	4.97 2.11	3.41 3.13	0.82 0.34
C		—	0.35 3.87	2.82 1.49
G			—	1 1
T				—

Data: two ‘genes’ of length 1200 and 800 sites

Brief digression….

Markov Chain Monte Carlo methods for inferring phylogenies

1 Construct a markov chain whose successive states are possible phylogenetic trees

2. Start at a random tree then guide the chain such that it samples a desirable region of the universe of possible trees (Metropolis-Hastings ratio)

3. Use the sample of trees to estimate parameters of the model of evolution and features of the tree itself

Sampling the Universe of Phylogenetic Trees

Markov-Chain Monte Carlo (MCMC) Methods

- Generate a large number of phylogenetic trees from a Markov Chain
- at equilibrium randomly sample from universe of trees

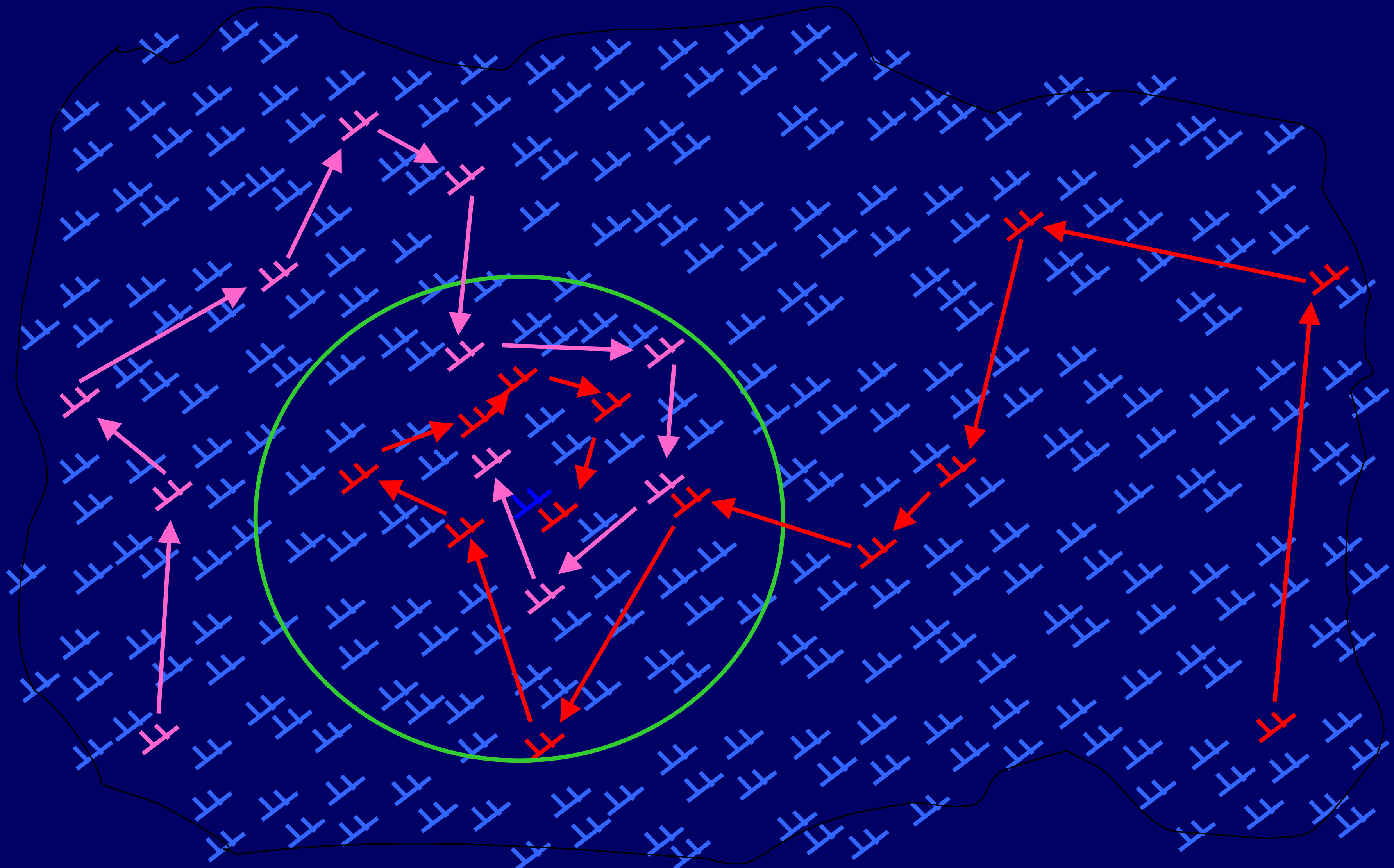
sampling mechanism: The Metropolis-Hastings Algorithm

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

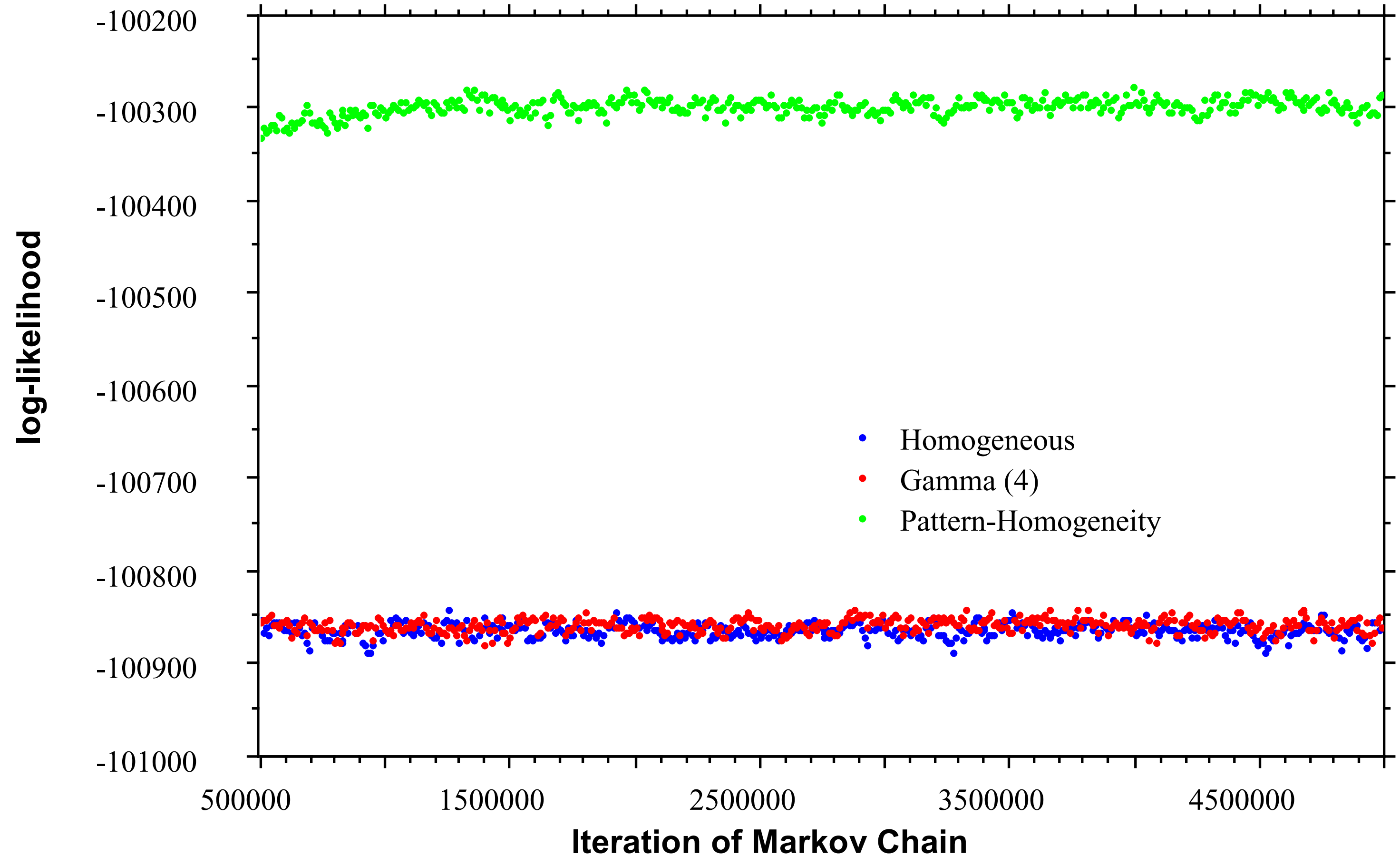
otherwise…

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

Sampling the universe of possible trees: Markov-chain Monte Carlo methods

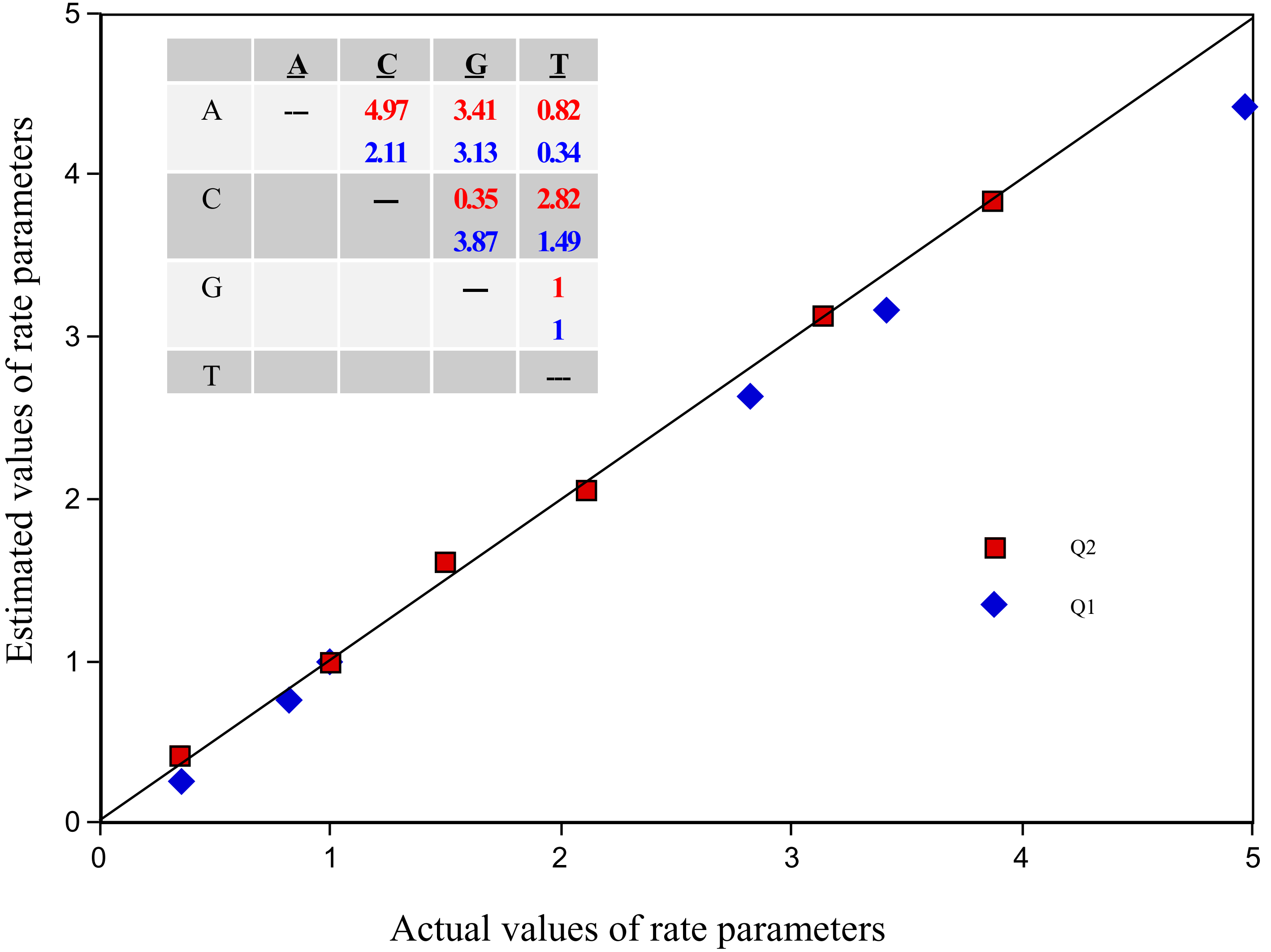


Testing Pattern-Heterogeneity Model: simulations from two random transition rate matrices



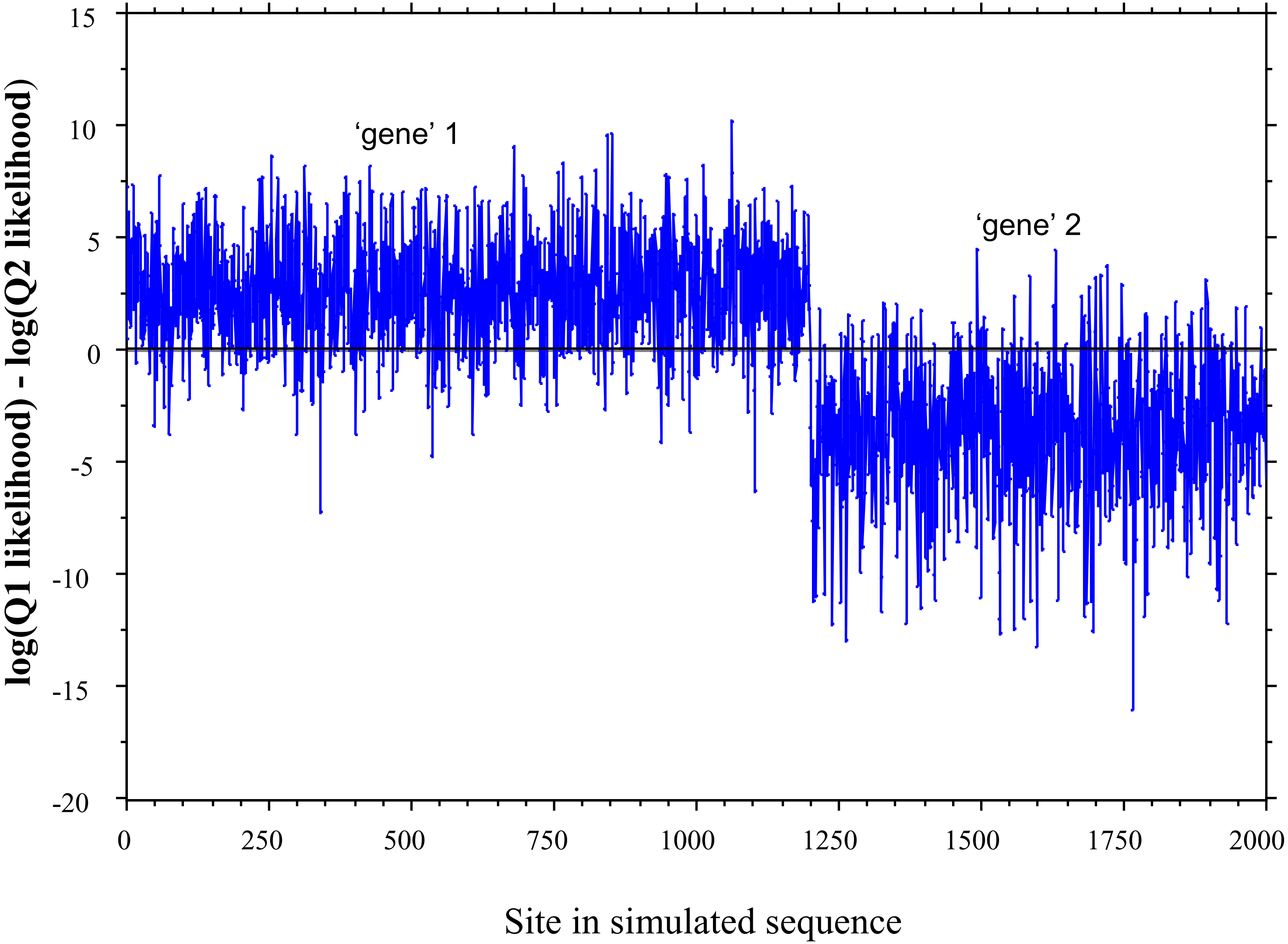
	A <--> C	A <--> G	A <--> T	C <--> G	C <--> T	G <--> T
Q1	4.97	3.41	0.82	0.35	2.82	1
Q2	2.11	3.13	0.34	3.87	1.49	1

Pattern-heterogeneity model: Simulated and estimated values of the rate parameters



	A	C	G	T
A	--	4.97 2.11	3.41 3.13	0.82 0.34
C		—	0.35 3.87	2.82 1.49
G			—	1 1
T				---

Detecting pattern-heterogeneity in simulated data



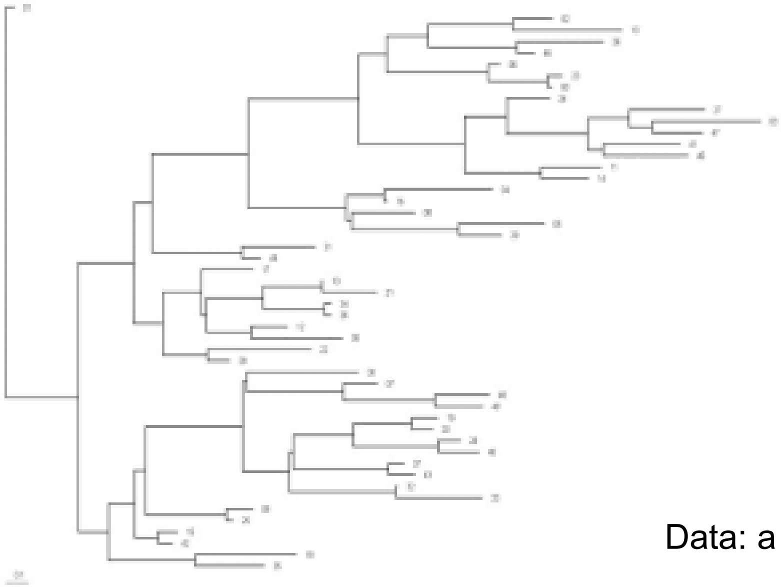
Testing the Pattern-Heterogeneity Model: detecting rate-heterogeneity in simulated data

Method:

I) generate simulated gene-sequence data on a random tree according to a gamma rate heterogeneity model (continuous-gamma, α=1.0)

2) analyse data using homogeneous model, gamma rates (4 categories) and pattern-heterogeneity model (4 rate matrices)

Tree used in simulations

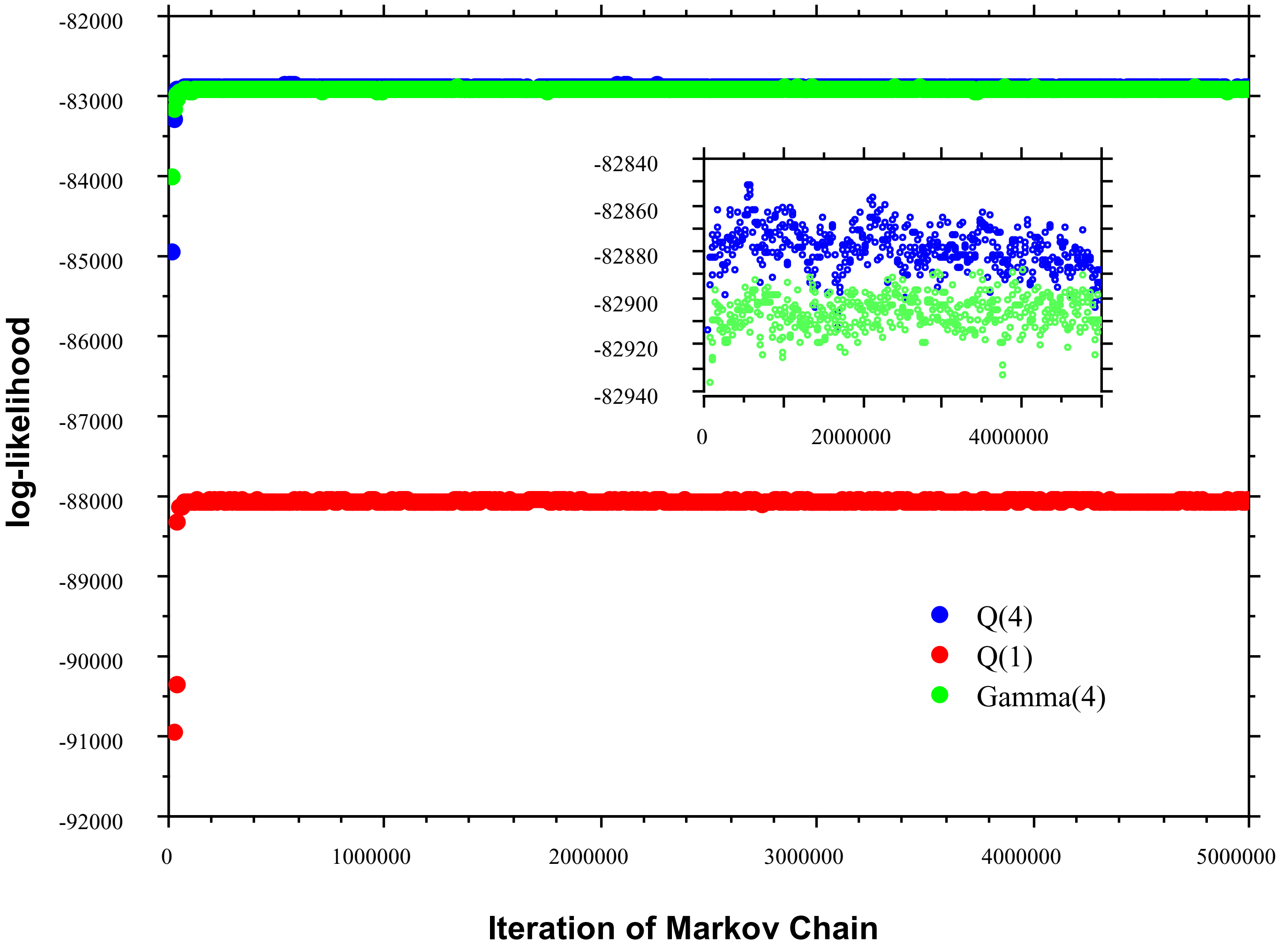


rate matrices

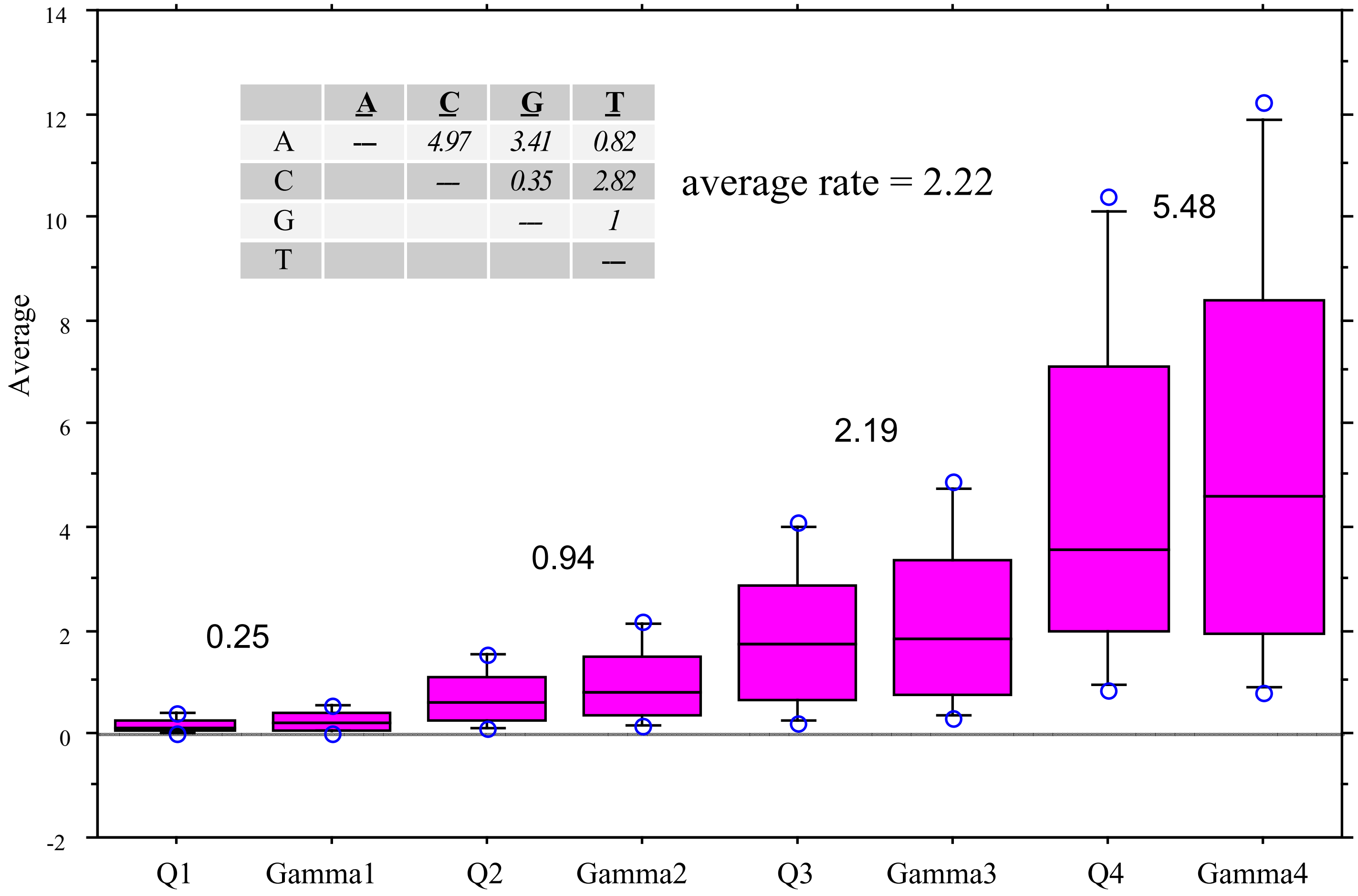
	A	**C**	**G**	**T**
A	—	*4.97*	*3.41*	*0.82*
C		—	*0.35*	*2.82*
G			—	*1*
T				—

Data: a gene of length 2000 sites

Testing the Pattern-Heterogeneity Model: simulated gamma rate variation



Average estimated transition rates from gamma and pattern-heterogeneity models applied to simulated gamma rate heterogeneity data: input values shown



	A	C	G	T
A	—	*4.97*	*3.41*	*0.82*
C		—	*0.35*	*2.82*
G			—	*1*
T				—

Applications of the pattern-heterogeneity model to two real data sets

1. SSU and LSU nrDNA data from Ascomycota fungi: gene differences?

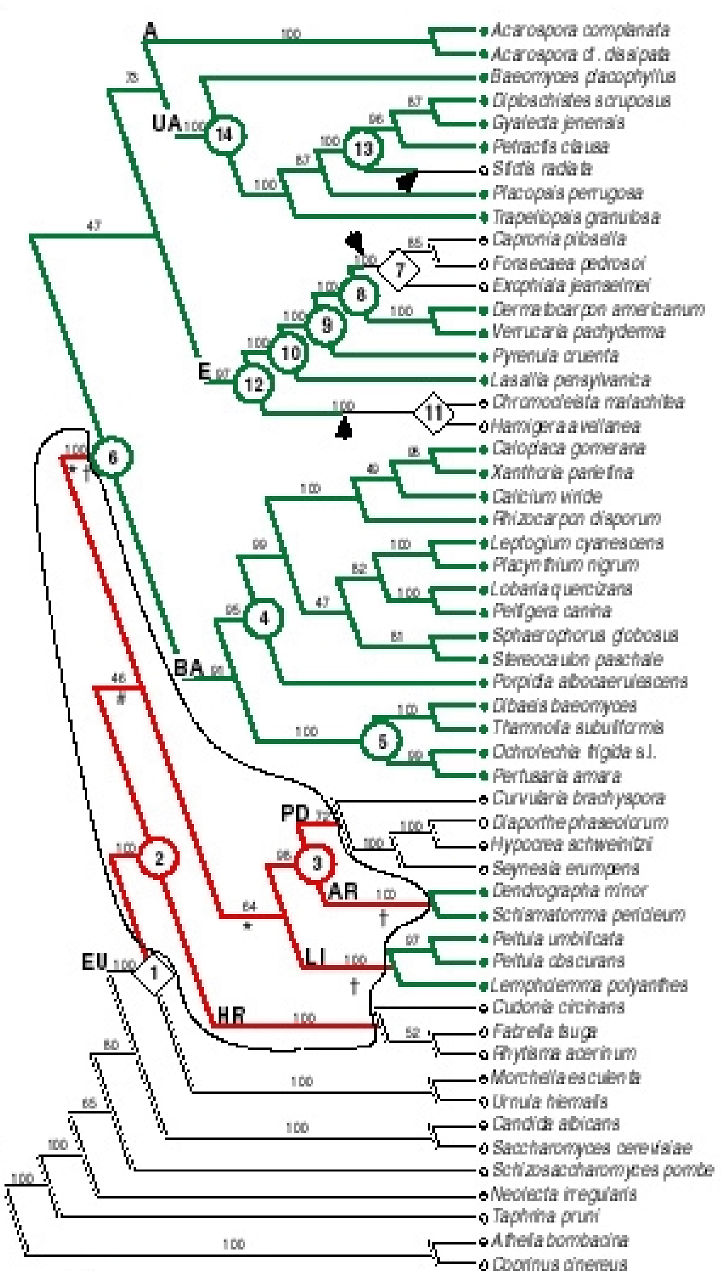
2. Detecting secondary structure in Mitochondrial 12s data from mammals

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

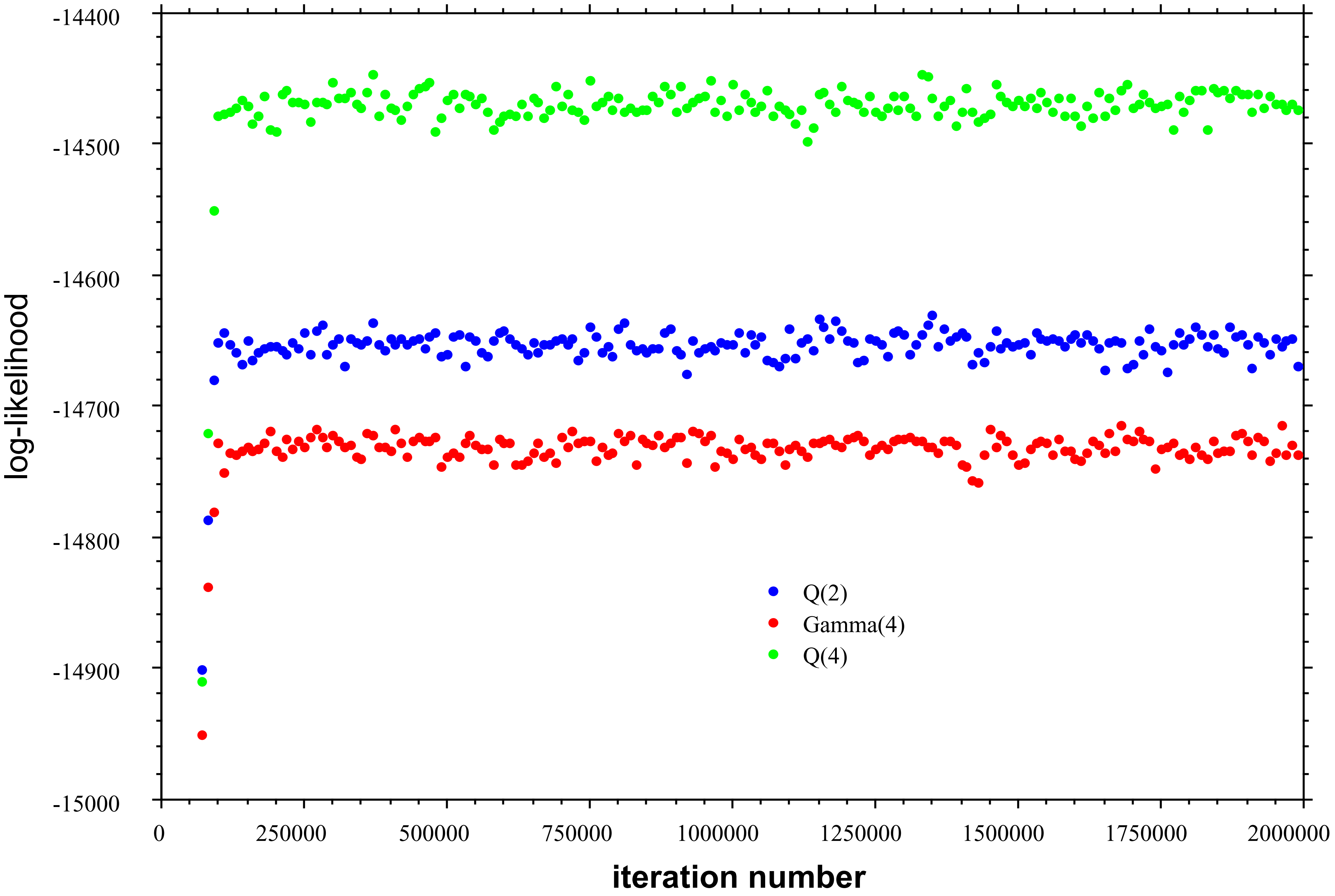
🟩 lichen forming

🟥 ambiguous

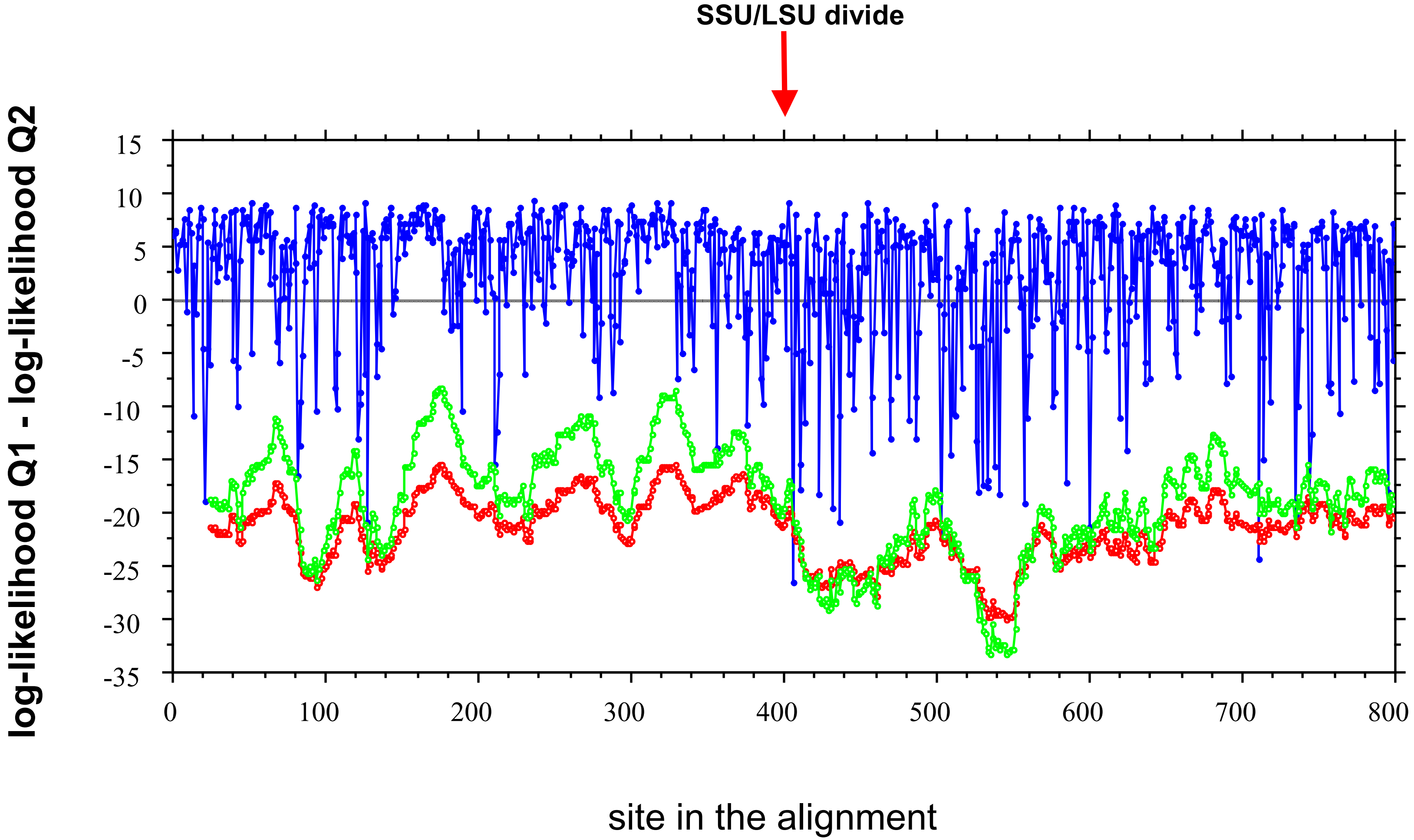
□ not lichen forming



log-likelihoods for combined LSU/SSU nrDNA data set: 54 Ascomycota species n=800 sites

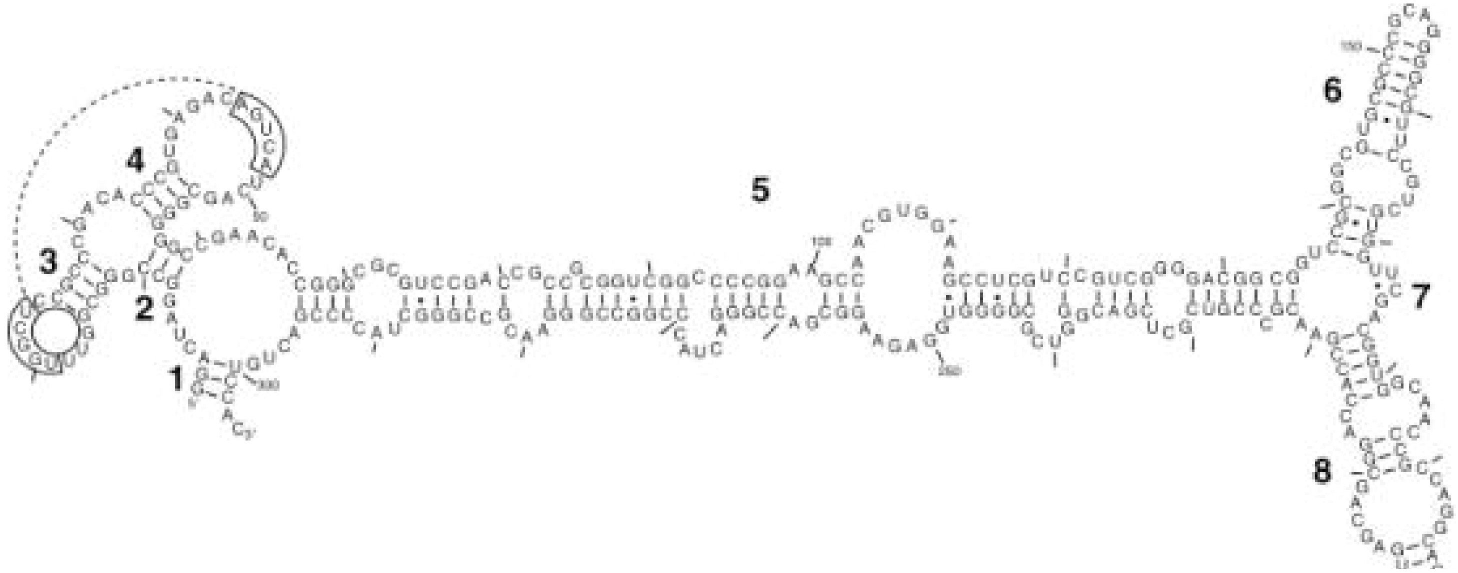


Site by site analysis fitting two independent rate matrices: LSU/SSU nrDNA Ascomycota combined data set



Note: data modelled with two independent rate matrices

Detecting secondary structure using the pattern-heterogeneity model

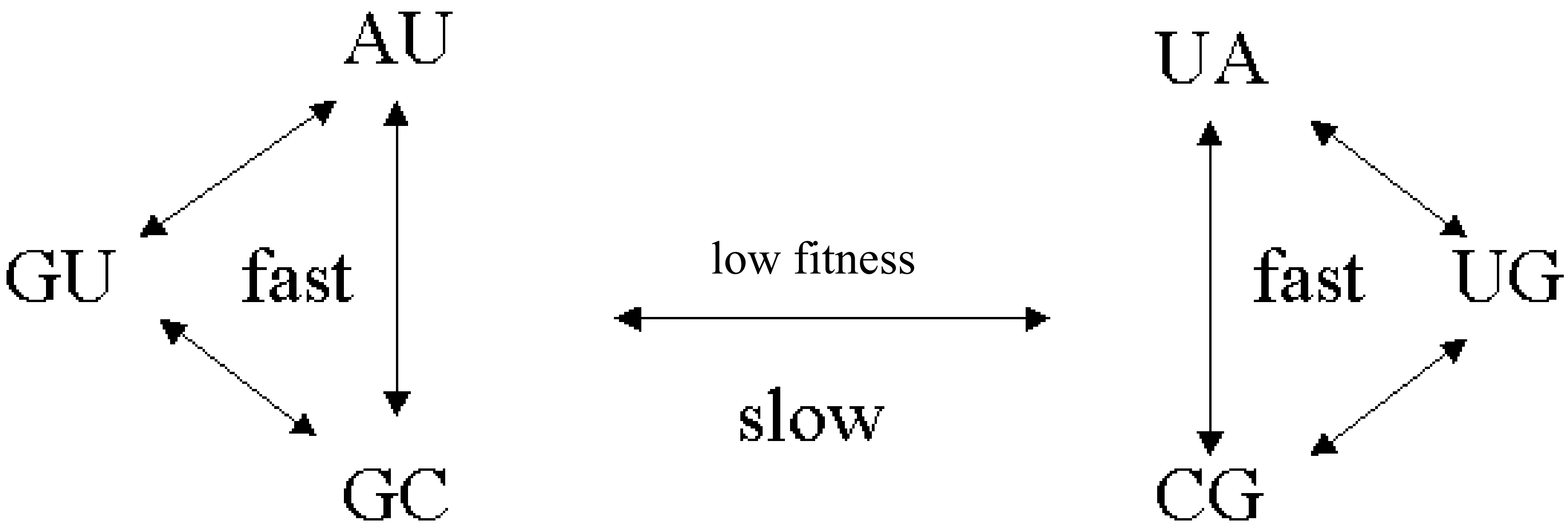


Stem and loop structure leads to predictions about the pattern of nucleotide substitutions

stems: dominated by Watson-Crick base pairings. Therefore expect compensatory substitutions to maintain Watson-Crick pairings. Non-compensatory changes have lower fitness. This predicts that transitional changes will occur at a far higher rate than transversional changes. Often observe Tr/Tv ratio of 10-20 in mitochondrial DNA

loops: no a priori substitutional pattern expected

Compensatory substitutions



AU

GU

fast

GC

low fitness

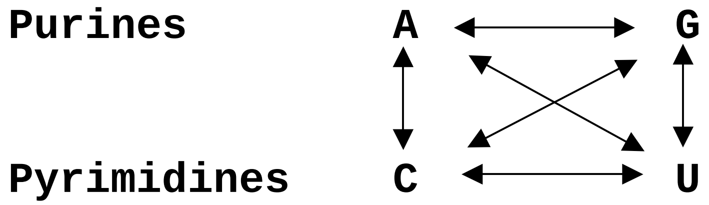
slow

UA

fast

UG

CG



Purines

A

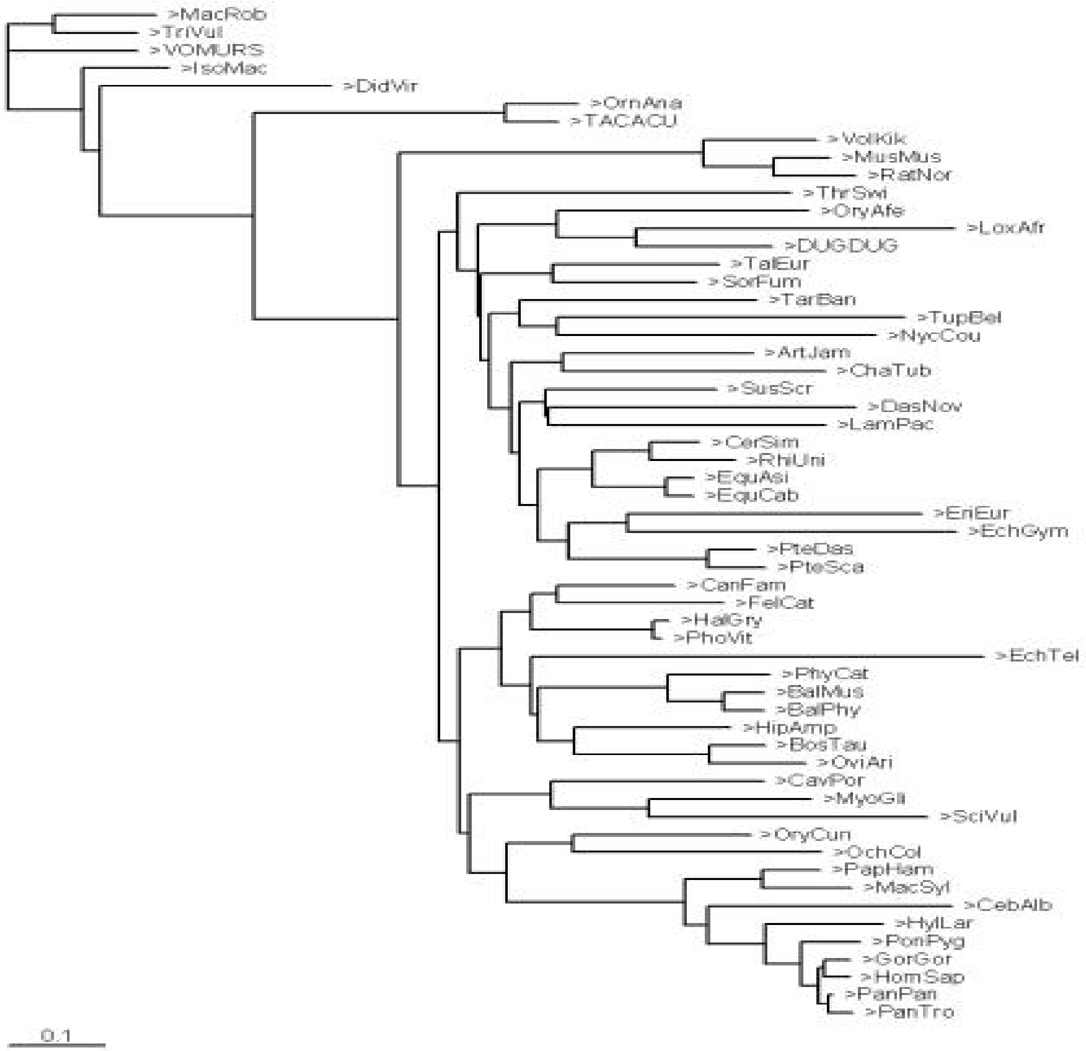
G

Pyrimidines

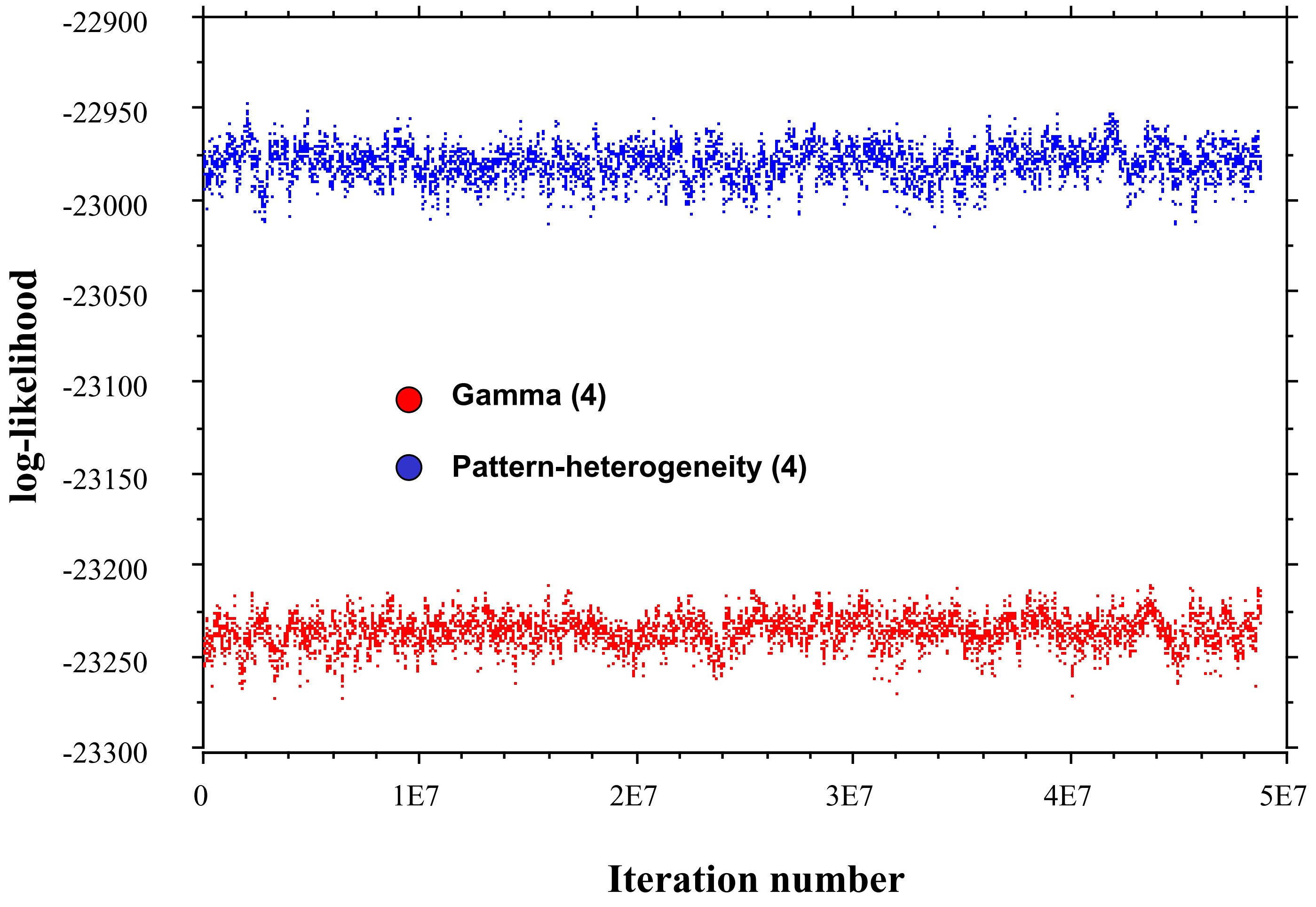
C

U

Detecting secondary structure: application of pattern-heterogeneity model to 12s mitochondrial DNA data on 57 mammals

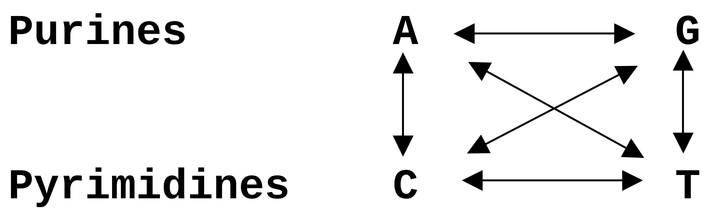


Detecting secondary structure; likelihoods of pattern-heterogeneity and Gamma models



Mammal 12S data: analysis of rate matrices

	A <-> C	**A <-> G**	A <-> T	C <-> G	**C <-> T**	G <-> T	Tr/Tv
Q1	13.75	10.52	7.89	1.69	59.65	3.34	5.26
Q2	44.25	66.37	35.57	10.64	88.46	3.59	3.29
Q3	**1.60**	**45.29**	**1.77**	**1.59**	**23.30**	**1.62**	**20.84**
Q4	0.30	0.80	0.18	0.20	1.62	0.10	6.26



Best fit to loop or stem

	Q1	**Q2**	**Q3**	**Q4**
Stem	57	13	122	272
Loop	102	150	57	236

Conclusions

Pattern-heterogeneity model can detect pattern-heterogeneity in simulated and real data and recover the parameters of the model of sequence evolution

Can lead to large improvements in likelihood over homogeneous process model

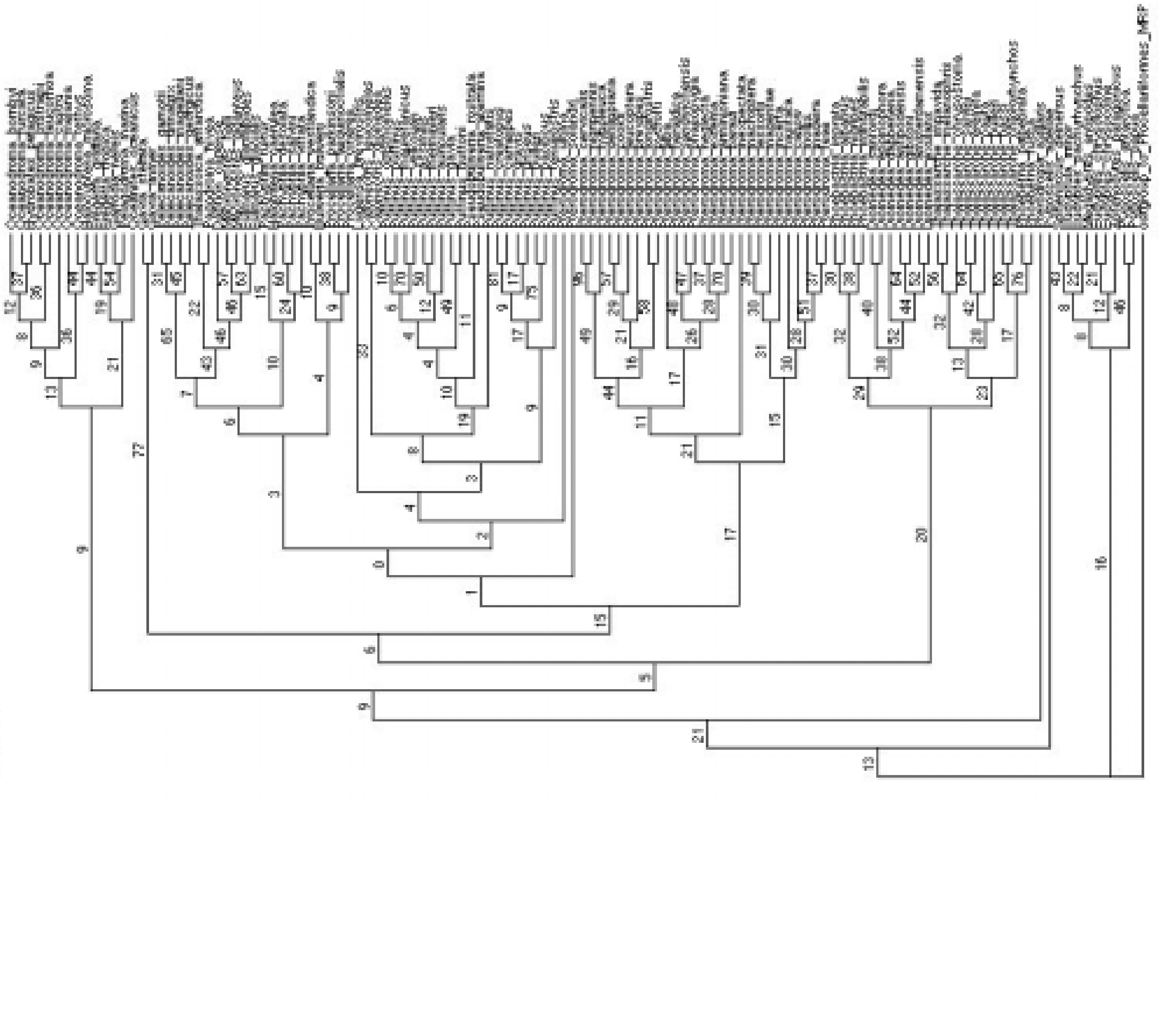
Returns same likelihood as gamma-rates model for data with gamma rate variability and often improves upon gamma model in other situations

Can be used to investigate gene evolution (such as secondary structure) or be applied to concatenated data sets of multiple genes.

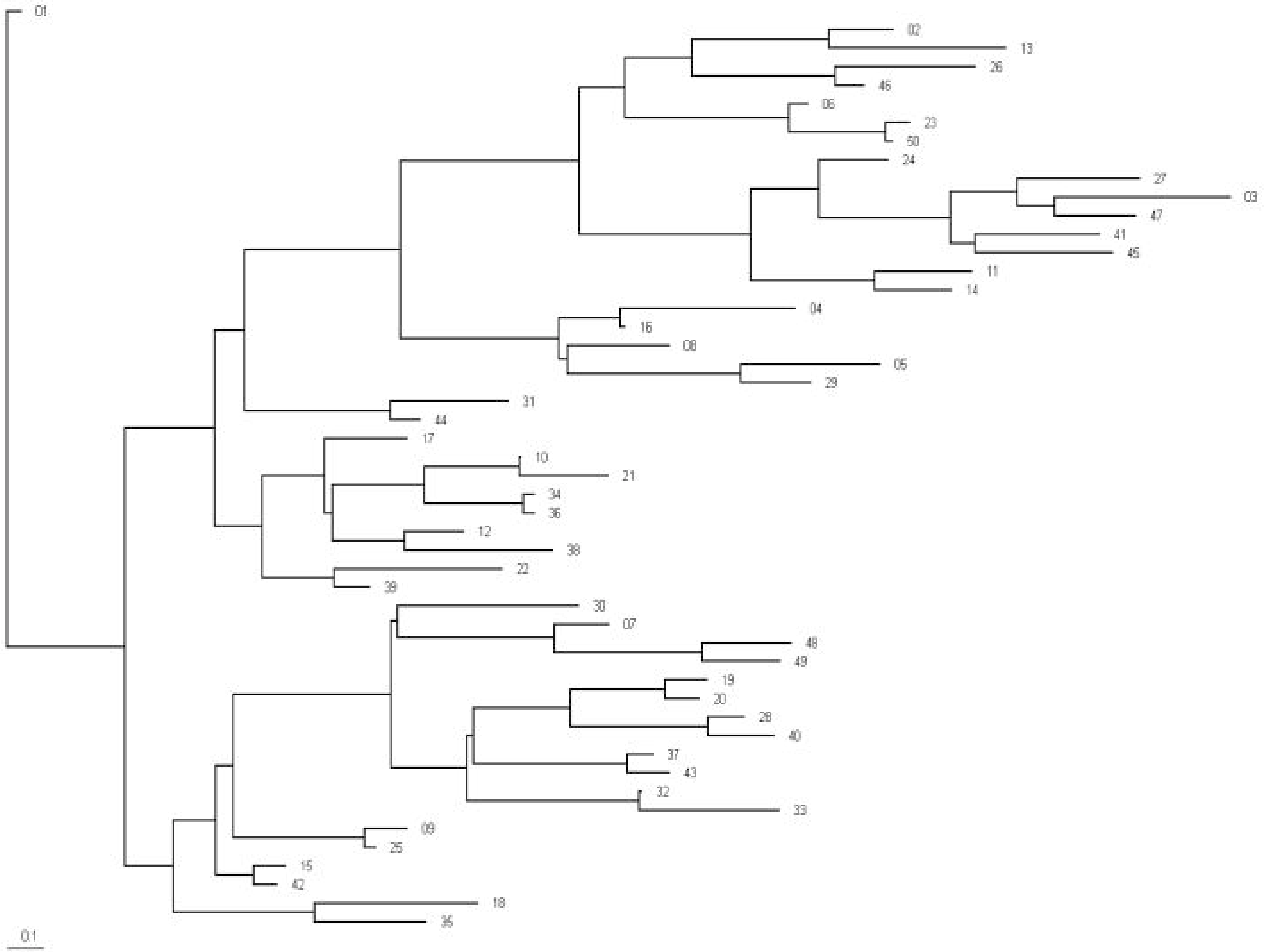
Software available from MP

```
'Oceanodroma_hornbyi'                                                                                                                                                                                    0001000000000000011010
'Gavia_stellata'                                                                                                                                                                                              0000000000000000000110
'Gavia_immer'                                                                                                                                                                                                 0000000000000000000110
'Spheniscus_demersus'                                                                                                                                                                                         1110000000000000000001
'Pygoscelis_adeliae'                                                                                                                                                                            00000000000001
'Eudyptula_minor'                                                                                                                                                                               010000000000011110000000000000000001
'Eudyptes_pachyrhynchus'                                                                                                                                                                        11000000000001
'Megadyptes_antipodes'                                                                                                                                                                          110000000000010110000000000000000001
'Fregetta_grallaria'                                                                                      000000001010000000000000000000000000000000000000000000000000000000000000000000000011010
'Pygoscelis_antarctica'                                                                                   00100000000000000000000000000000000000000000000000000000000000000000000000000000000001
'Pygoscelis_papua'                                                                                        00100000000000000000000000000000000000000000000000000000000000000000000000000000000001              0010000000000000000001
'Eudyptes_chrysolophus'                                                                                   11000000000000000000000000000000000000000000000000000000000000000000000000000000000001
'Eudyptes_chrysocome'                                                                                     11000000000000000000000000000000000000000000000000000000000000000000000000000000000001
'Aptenodytes_patagonicus'                                                                                 01000000000000000000000000000000000000000000000000000000000000000000000000000000000001              0000000000000000000001
'Oceanodroma_melania'                                                                                     00000001000000000000000000000000000000000000000000000000000000000000000000000000000110
'Oceanodroma_tethys'                                                                                      00010001000000000000000000000000000000000000000000000000000000000000000000000000000110
'Halocyptena_microsoma'                                                                                   00010001000000000000000000000000000000000000000000000000000000000000000000000000000110
'Oceanodroma_furcata'                                                                                     00001010000000000000000000000000000000000000000000000000000000000000000000000000000110              0001000000000000011010
'Oceanodroma_tristrami'                                                                                   00000110000000000000000000000000000000000000000000000000000000000000000000000000000110
'Oceanites_oceanicus'                                                                                     00000000000000000000000000000000000000000000000000000000000000000000000000000000011010              0000000000000000011010
'Fregetta_tropica'                                                                                        00000000101000000000000000000000000000000000000000000000000000000000000000000000011010
'Garrodia_nereis'                                                                                         00000000011000000000000000000000000000000000000000000000000000000000000000000000011010
'Pelagodroma_marina'                                                                                      0000000001100000000000000000000000000000000000000000000000000000000000000000000001101000000000000110
'Pelecanoides_garnotii'                                                                                   00000000000000000000000000000000000000000000000000000000000000000000000000000110101010
'Pelecanoides_magellani'                                                                                  00000000000000000000000000000011000000000000000000000000000000000000000000000110101010
'Pelecanoides_georgicus'                                                                                  000000000000000000000000000000110000000000000000000000000000000000000000000001101010100000000011101000000000000000110101010
'Lugensa_brevirostris'                                                                                    00000000000000000000000000000000000000000000000000000000000000000000001010101010101010              0000000001011010101010
'Calonectris_leucomelas'                                                                                  00000000000000000000000000000000000000000000000000000000000000000011011010101010101010
'Puffinus_opisthomelas'                                                                                   00000000000000000000000000000000000000000000000000000000000000001110101110101010101010
'Procellaria_westlandica'                                                                                 0000000000000000000000000000000000000000000000000000010000100000000000011010101010101000000101011010
'Procellaria_parkinsoni'                                                                                  00000000000000000000000000000000000000000000000000001100001000000000000110101010101010
'Procellaria_aequinoctialis'                                                                              00000000000000000000000000000000000000000000000000001100001000000000000110101010101010
'Pachyptila_turtur'                                                                                       0000000000000000000000000000000000000000000000000000000011000000000000011010101010101000010000111010
'Pachyptila_desolata'                                                                                     00000000000000000000000000000000000000000000000000000011110000000000000110101010101010
'Pachyptila_salvini'                                                                                      00000000000000000000000000000000000000000000000000000011110000000000000110101010101010
'Pachyptila_vittata'                                                                                      0000000000000000000000000000000000000000000000000000000111000000000000011010101010101000010000111010000000000111010101010
'Halobaena_caerulea'                                                                                      00000000000000000000000000000000000000000000000000000000010000000000000110101010101010
'Thalassoica_antarctica'                                                                                  00000000000000000000000000000000000000000000000000010000000000000000000001101010101010              0000000000111010101010
'Daption_capense'                                                                                         0000000000000000000000000000000000000000000000000011000000000000000000000110101010101000000000001010
'Macronectes_halli'                                                                                       00000000000000000000000000000000000000000000000101110000000000000000000001101010101010
'Phoebastria_irrorata'                                                                                    00000000000000000001000000000010000000000000000000000000000000000000000000000001101010              0000011000000001101010
'Phoebastria_nigripes'                                                                                    00000000000110000001000000000010000000000000000000000000000000000000000000000001101010              0000001000000001101010
'Diomedea_sanfordi'                                                                                       00000000000000000110000000000010000000000000000000000000000000000000000000000001101010
'Diomedea_dabbenena'                                                                                      00000000000000000101000000000010000000000000000000000000000000000000000000000001101010
'Diomedea_antipodensis'                                                                                   00000000000001011010000000000010000000000000000000000000000000000000000000000001101010
'Diomedea_gibsoni'                                                                                        00000000000001011010000000000010000000000000000000000000000000000000000000000001101010
'Thalassarche_impavida'                                                                                   00000000000000000000100011010100000000000000000000000000000000000000000000000001101010
'Thalassarche_melanophris'                                                                                00000000000000000000100011010100000000000000000000000000000000000000000000000001101010
'Thalassarche_salvini'                                                                                    00000000000000000000011101010100000000000000000000000000000000000000000000000001101010
'Thalassarche_eremita'                                                                                    00000000000000000000011101010100000000000000000000000000000000000000000000000001101010
'Thalassarche_cauta'                                                                                      00000000000000000000001101010100000000000000000000000000000000000000000000000001101010              0000000000000001101010
'Thalassarche_bassi'                                                                                      00000000000000000000000000110100000000000000000000000000000000000000000000000001101010
'Thalassarche_chlororhynchos'                                                                             00000000000000000000000000110100000000000000000000000000000000000000000000000001101010
'Pterodroma_axillaris'                                                        0000000000000000000000000001
'Pterodroma_cervicalis'                                                       1000000000000000000000000001
'Pterodroma_hypoleuca'                                                        000000000000000000000000000100000000000000000000000000000000000000000000000110000000000000000000000000000011010101010              0000000011011010101010
'Pterodroma_defilippiana'                                                     0111000000000000000000001110
'Pterodroma_cookii'                                                           011100000000000000000000111000000000000000000000000000000000010000000000011000000000000000000000000000001101010101000000011011010000000011011010101010
'Pterodroma_leucoptera'                                                       0011000000000000000000001110
'Pterodroma_brevipes'                                                         0001000000000000000000001110
'Pterodroma_longirostris'                                                     00001000000000000000000011100000000000000000000000000000000001000000000001100000000000000000000000000000110101010100
'Pterodroma_pycrofti'                                                         0000100000000000000000001110
'Pterodroma_inexpectata'                                                      0000000000000000000001010110000000000000000000000000000000000001000000011010000000000000000000000000001101010101000000011011010
'Pterodroma_ultima'                                                           0000000000000000000011010110
'Pterodroma_solandri'                                                         0000000000000000000111010110
'Pterodroma_macroptera'                                                       000000000000000000111101011000000000000000000000000000000000000011101101010000000000000000000000000000011010101010
'Pterodroma_magentae'                                                         000000000000000001111101011000000000000000000000000000000000000000010110101000000000000000000000000000011010101010
'Pterodroma_lessonii'                                                         000000000000010111111101011000000000000000000000000000000000000011101101010000000000000000000000000000011010101010
'Pterodroma_incerta'                                                          000000000000110111111101011000000000000000000000000000000000000001101101010000000000000000000000000000011010101010
'Pterodroma_hasitata'                                                         000000000001110111111101011000000000000000000000000000000000000000011101010000000000000000000000000000011010101010              0000000011011010101010
'Pterodroma_cahow'                                                            000000000001110111111101011000000000000000000000000000000000000010001110101000000000000000000000000000011010101010
'Pterodroma_mollis'                                                           000000000000000111111101011000000000000000000000000000000000000000000010101000000000000000000000000000011010101010
'Pterodroma_madeira'                                                          0000000000010001111111010110
'Pterodroma_feae'                                                             000000000001000111111101011000000000000000000000000000000000000010001110101000000000000000000000000000011010101010
'Pterodroma_alba'                                                             0000000000000000000000110110
'Pterodroma_heraldica'                                                        0000000001100000000000110110
'Pterodroma_sandwichensis'                                                    0000010001100000000000110110
'Pterodroma_phaeopygia'                                                       000001000110000000000011011000000000000000000000000000000000011000000011010000000000000000000000000000011010101010
'Pterodroma_neglecta'                                                         000000101010000000000011011000000000000000000000000000000000000000000011010000000000000000000000000000011010101010
'Pterodroma_externa'                                                          000000101010000000000011011000000000000000000000000000000000011000000011010000000000000000000000000000011010101010
'Pterodroma_arminjoniana'                                                     0000000110100000000000110110                                                                                                    0000000011011010101010
'Diomedea_epomophora'                                   001010000000000110                            0000000000000000011000000000001000000000000000000000000000000000000000000000000110101000100000000011000011100000000001101010
'Diomedea_amsterdamensis'                               001010000000000110                            00000000000001110100000000000010000000000000000000000000000000000000000000000001101010
'Phoebastria_immutabilis'                               000110000000000110                            00000000000110000001000000000010000000000000000000000000000000000000000000000001101010              0000111000000001101010
'Phoebastria_albatrus'                                  000110000000000110                            00000000000010000001000000000010000000000000000000000000000000000000000000000001101010
'Phoebetria_palpebrata'                                 100001000000000110                            00000000000000000000000000000110000000000000000000000000000000000000000000000001101010
'Phoebetria_fusca'                                      100001000000000110                            00000000000000000000000000000110000000000000000000000000000000000000000000000001101010
'Thalassarche_chrysostoma'                              010001000000000110                            00000000000000000000000011010100000000000000000000000000000000000000000000000001101010              0000000000000001101010
'Thalassarche_bulleri'                                  010001000000000110                            0000000000000000000000010101010000000000000000000000000000000000000000000000000110101000100000000001100000000000000001101010
'Fulmarus_glacialoides'                                 000000100000011010                            00000000000000000000000000000000000000000000000011110000000000000000000001101010101010
'Hydrobates_pelagicus'                                  000000000000000001                            00001010000000000000000000000000000000000000000000000000000000000000000000000000000110
'Oceanodroma_castro'                                    000000000000000001
'Pterodroma_baraui'                             000000001110                  0000000110100000000000110110
'Pagodroma_nivea'                               000000110110                                              00000000000000000000000000000000000000000000000000000000000000000000000001101010101010
'Procellaria_cinerea'                           000011010110000000000001101010
'Pseudobulweria_rostrata'                       010101010110
'Pseudobulweria_aterrima'                       010101010110
'Pterodroma_nigripennis'                        000000001110                  100000000000000000000000000100000000000000000000000000000000000000000000000000000000000000000000000000011010101010
'Macronectes_giganteus'                         100000110110000000000000011010                            00000000000000000000000000000000000000000000000101110000000000000000000001101010101010              0000000000001010101010
'Calonectris_diomedea'                          001101010110000000000011101010                            00000000000000000000000000000000000000000000000000000000000000000011011010101010101010              0000000100111010101010
'Bulweria_bulwerii'                             000011010110000000000000101010                            00000000000000000000000000000000000000000000000000000000000100000000000001101010101010
'Pelecanoides_urinatrix'                        000000000010                                              00000000000000000000000000000001000000000000000000000000000000000000000000000110101010              0000000000000110101010
'Oceanodroma_leucorhoa'                         000000000001                                              00000110000000000000000000000000000000000000000000000000000000000000000000000000000110              0001000000000000011010
'Diomedea_exulans'                              000000000001                                              00000000000001110100000000000010000000000000000000000000000000000000000000000001101010
'Fulmarus_glacialis'                    00000000100000110110000000100000011010                            00000000000000000000000000000000000000000000000011110000000000000000000001101010101010
'Puffinus_creatopus'                    10000011                                                          00000000000000000000000000000000000000000000000000000000000110000000111010101010101010
'Puffinus_carneipes'                    10000011                                                          00000000000000000000000000000000000000000000000000000000000110000000111010101010101010
'Puffinus_gravis'                       00000011                                                          00000000000000000000000000000000000000000000000000000000000010000000111010101010101010
'Puffinus_griseus'                      00000011                                                          000000000000000000000000000000000000000000000000000000000000100000001110101010101010100000110101101000000001001110101010 10
'Puffinus_tenuirostris'                 00000011
'Puffinus_bulleri'                      01000011                                                          00000000000000000000000000000000000000000000000000000000000001000000111010101010101010
'Puffinus_pacificus'                    01000011001101010110                                              00000000000000000000000000000000000000000000000000000000000001000000111010101010101010
'Puffinus_nativitatis'                  00000101                                                          00000000000000000000000000000000000000000000000000000000000000000101011010101010101010
'Puffinus_mauretanicus'                 00101101            000000011111101010
'Puffinus_yelkouan'                     00101101            000000011111101010
'Puffinus_gavia'                        00011101
'Puffinus_huttoni'                      00011101                                                          0000000000000000000000000000000000000000000000000000000000000000110101101010101010101000001101011010
'Puffinus_assimilis'                    00001101            000000000111101010                            00000000000000000000000000000000000000000000000000000000000000001111010110101010101010
'Puffinus_lherminieri'                  00001101                                                          00000000000000000000000000000000000000000000000000000000000000001111010110101010101010
'Puffinus_auricularis'                  00001101
'Puffinus_puffinus'                     00001101            000000001111101010                            00000000000000000000000000000000000000000000000000000000000000000111010110101010101010
```

Majority rule



Random tree of 60 tips used in all simulations


01
02
13
26
46
06
23
50
24
27
03
47
41
45
11
14
04
16
08
05
29
31
44
17
10
21
34
36
12
38
22
39
30
07
48
49
19
20
28
40
37
43
32
33
09
25
15
42
18
35
0.1