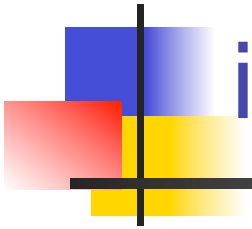
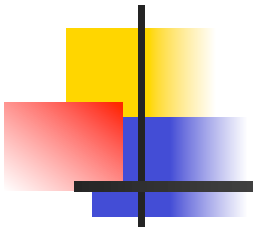


Bayesian Estimation of Species Trees
The BEST way to sort out incongruent gene phylogenies

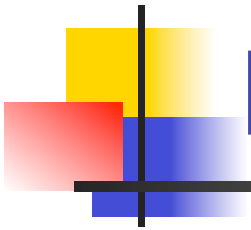
MBI Phylogenetics course

May 18, 2010



Outline

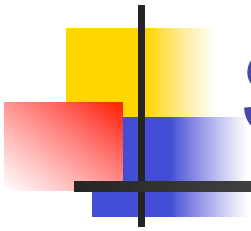
- Estimating Species Tree distributions
  - Probability model
  - The BEST Algorithm
  - Examples



References

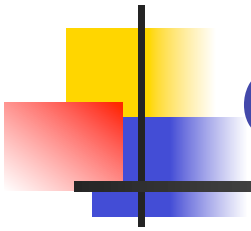
- Liu & Pearl, *Systematic Biology*, 2007
- Edwards, Liu, & Pearl, *PNAS*, 2007
- Liu, Pearl, Brumfield, & Edwards, *Evolution*, 2008

- Rannala & Yang, *Genetics*, 2003
- Degnan & Salter, *Evolution*, 2005
- Kubatko & Degnan, *Systematic Biology*, 2007



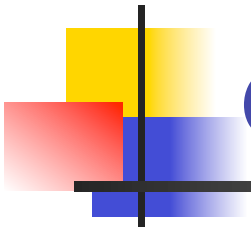
Scales

- *Unrooted gene trees* allow arbitrary mutation rate; branch lengths are in expected number of substitutions per site.
- *Rooted gene trees* assume a molecular clock; branch lengths are in time units (equivalent to units for unrooted tree if assumption holds)
- *Species trees* are rooted; branch lengths are in expected number of substitutions per generation (often rescaled per 2N generations) per site.



Gene Tree – Species Tree differences

- *Estimates* of gene trees can be different from each other when they follow the same true tree.
- Also, actual Gene Trees can be different from each other – and from the underlying Species Tree.
  - Recombination
  - Gene Duplication
  - Horizontal gene transfer
  - Deep Coalescence

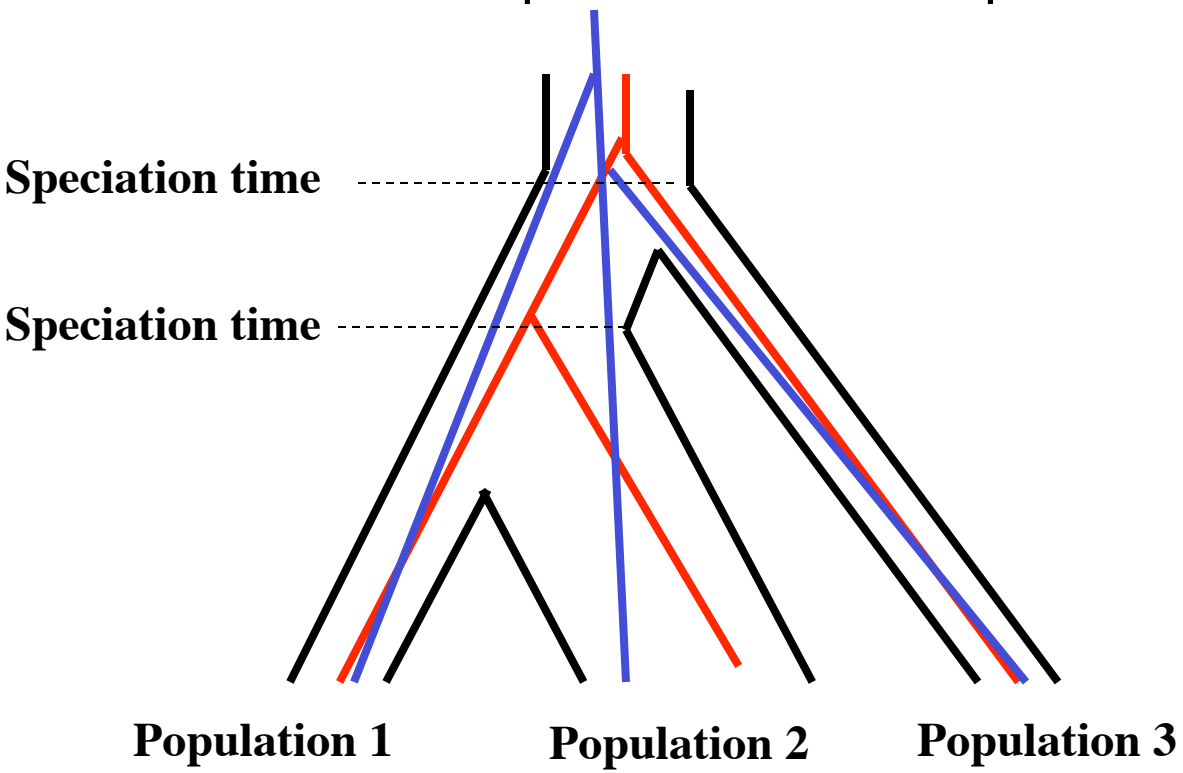


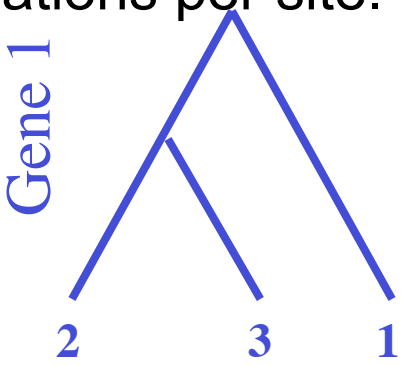
Gene Tree – Species Tree differences

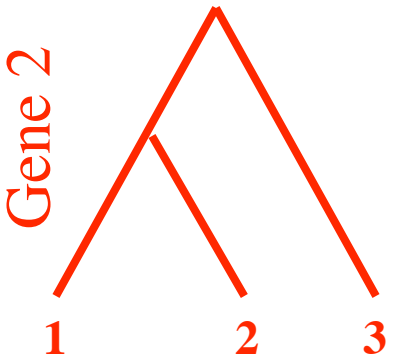
- *Estimates* of gene trees can be different from each other when they follow the same true tree.
- Also, actual Gene Trees can be different from each other – and from the underlying Species Tree.
  - Recombination
  - Gene Duplication
  - Horizontal gene transfer
  - Deep Coalescence

Species Trees

- follows lineages of populations, branch lengths are rescaled to measure expected mutations per 2N generations per site.







Kingman's coalescent process

Coalescent trees of gene copies within species (Kingman, 1982)

Random collision of lineages as go back in time (sans recombination)

Collision is faster the smaller the effective population size

Average time for

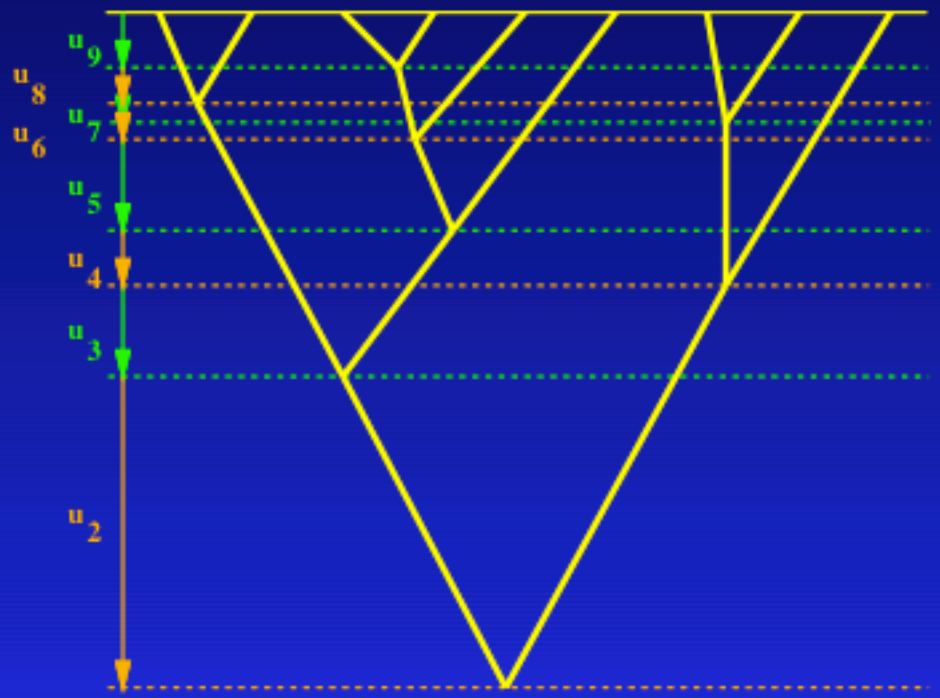
k copies to coalesce to

$$k-1 = \frac{4N}{k(k-1)}$$

Average time for

two copies to coalesce

= 2N generations



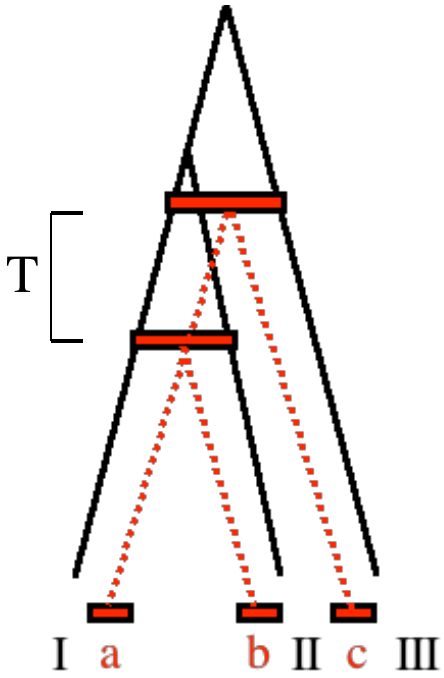
In a diploid population of effective population size N,

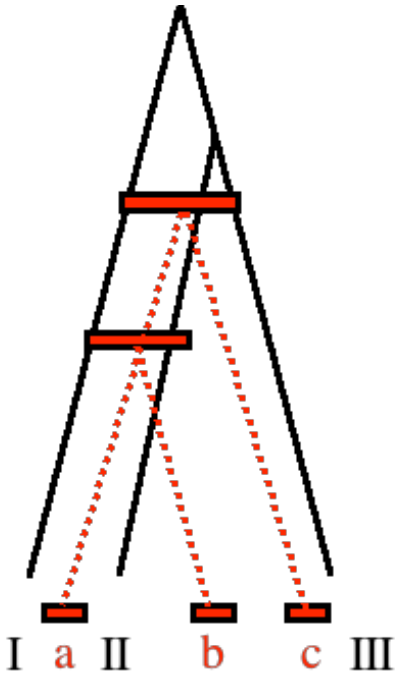
Average time for n copies to coalesce

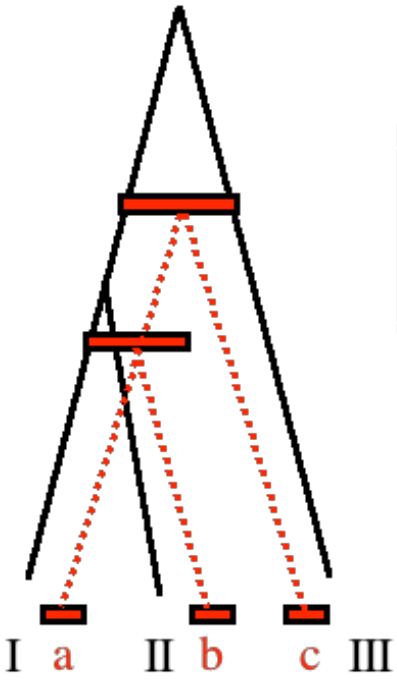
$$= 4N\left(1-\frac{1}{n}\right)$$

generations

(Figure from Felsenstein notes , 2005)







Gene trees compatible with a fixed species tree

When times between speciation events are short - the probability of topology error is high

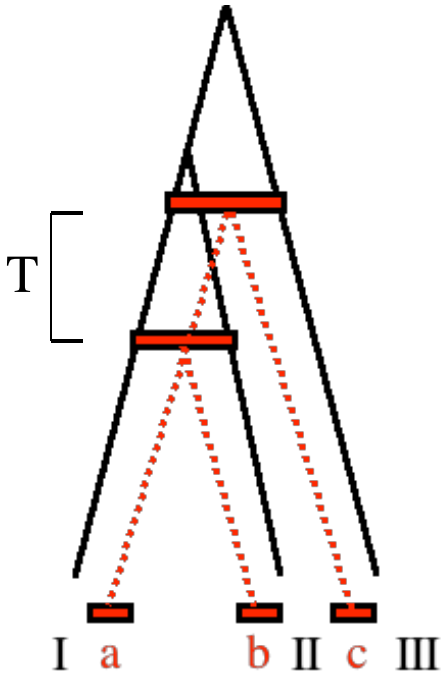
- Coalescent theory: time since two gene copies had a common ancestor is Exponentially distributed with mean 2N
- For the three species case this makes

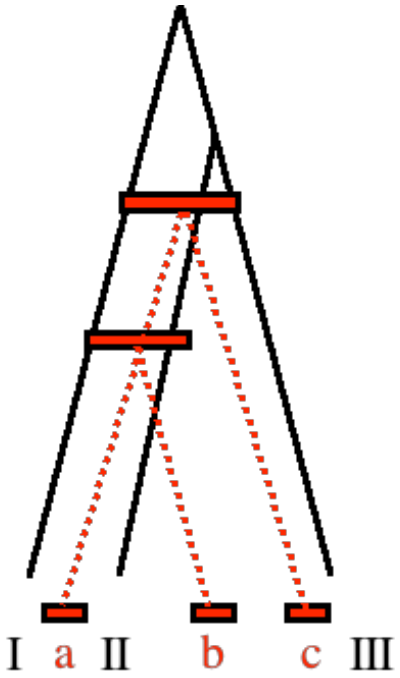
P(gene tree topology ≠ species tree topology) = $\frac{2}{3} e^{-T/2N}$

Where T = time between speciation events

(see next figure)

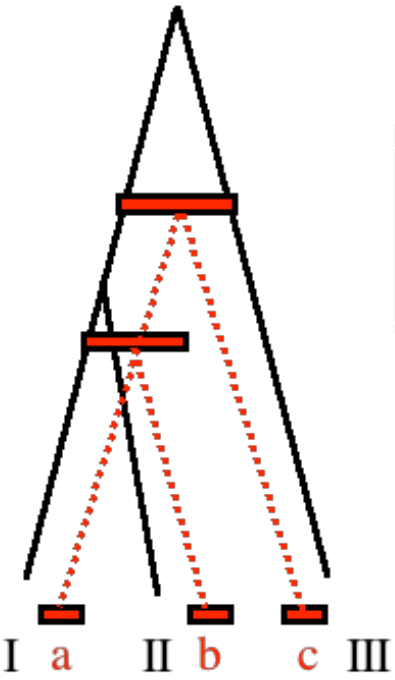
$\frac{1}{3}e^{-T/2N}$





$\frac{2}{3}e^{-T/2N}$

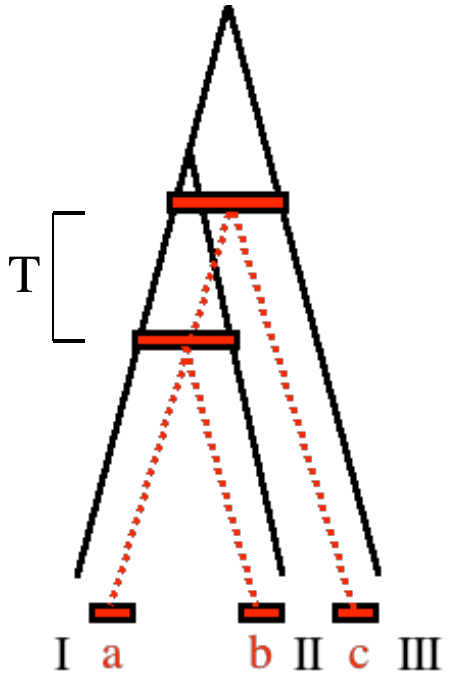
$1-e^{-T/2N}$

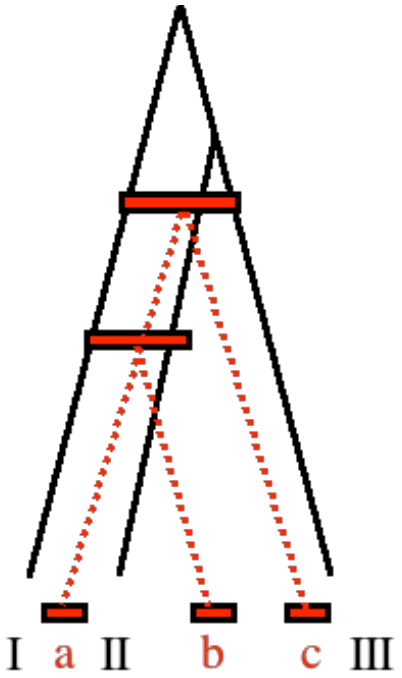


Gene trees compatible with a fixed species tree

Note the problem: with small T, up to 2/3 of the probability lies with the wrong topology.

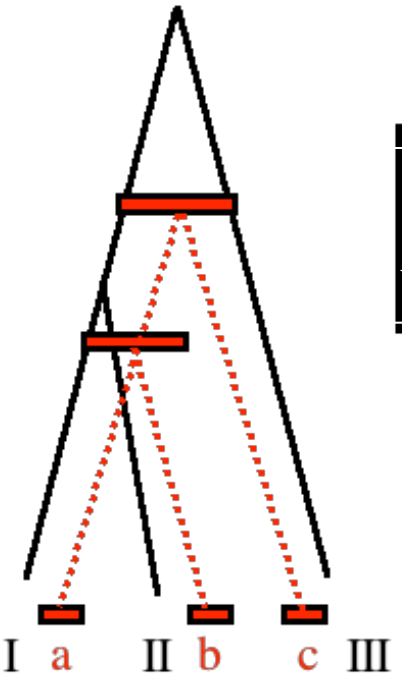
$\frac{1}{3}e^{-T/2N}$





$\frac{2}{3}e^{-T/2N}$

$1-e^{-T/2N}$



Gene trees compatible with a fixed species tree

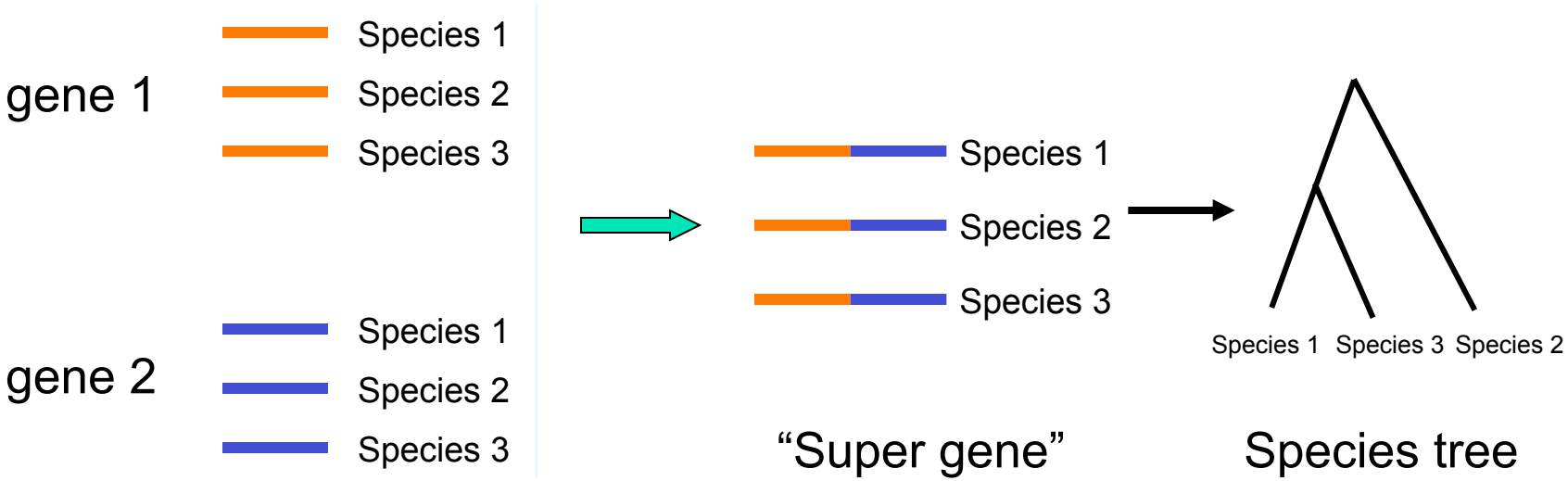
Ref: Nei (1987) *Molecular Evolutionary Genetics*

Generalized by

Degnan & Salter (2005) *Evolution*

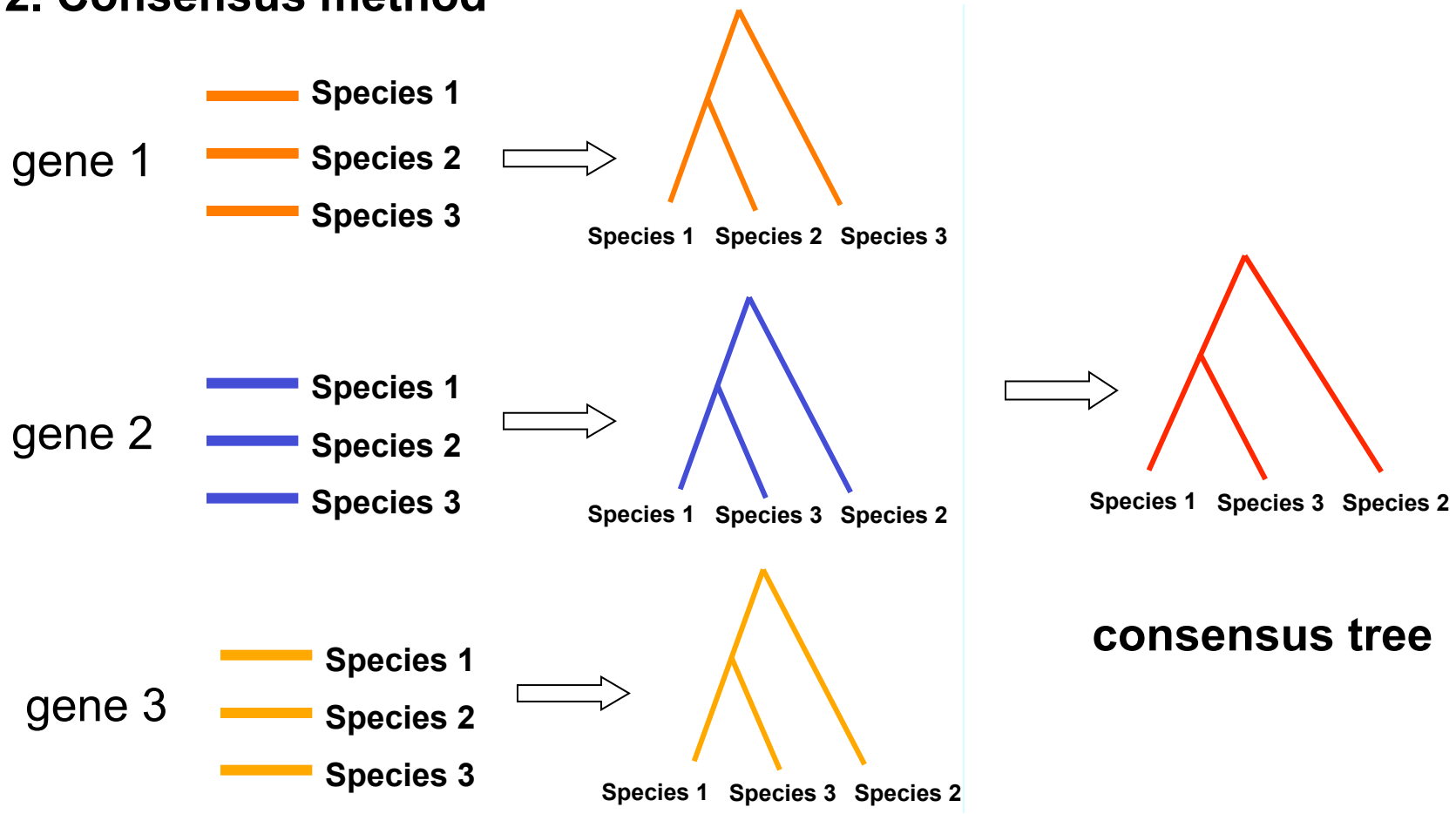
Current methods to estimate species phylogeny

1. Concatenation method



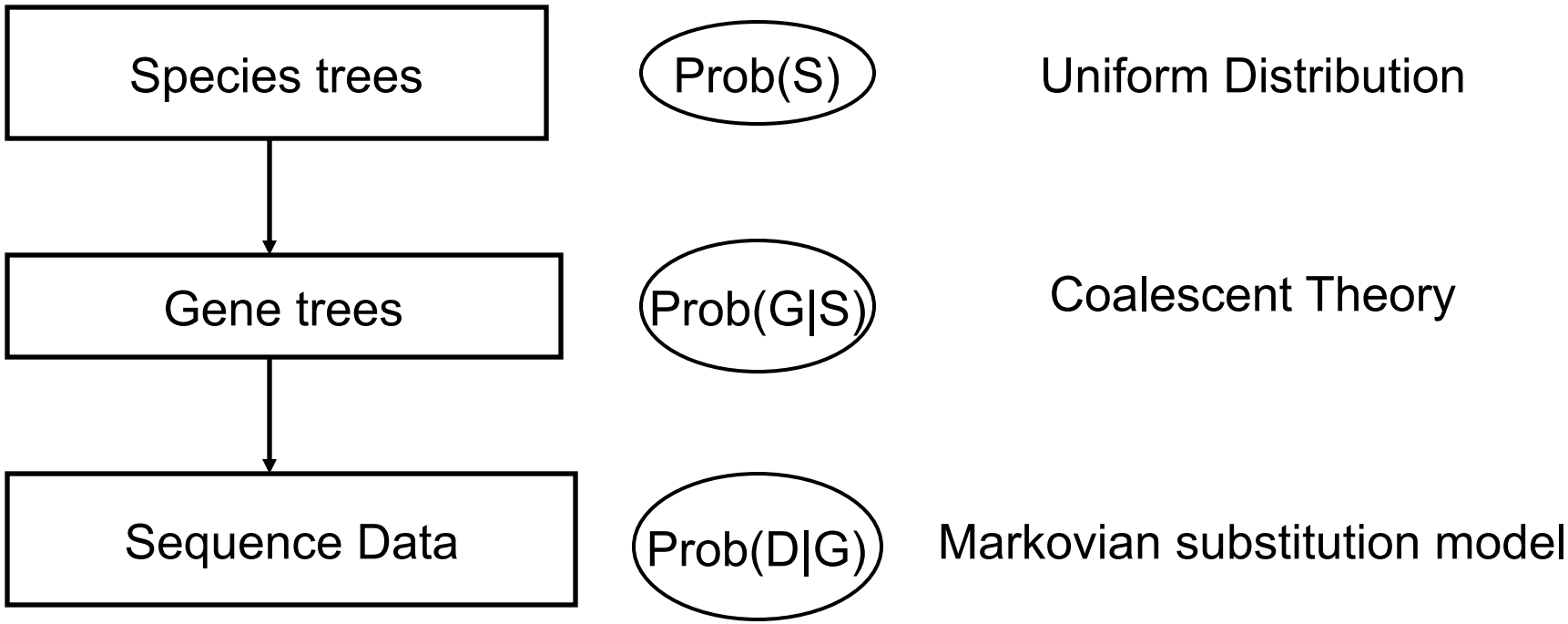
Current methods to estimate species phylogeny

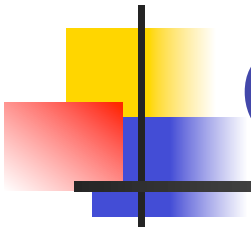
2. Consensus method



Bayesian **E**stimation of **S**pecies **T**rees (BEST)

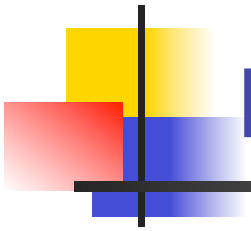
Hierarchical model





Compare

- Under concatenation approaches, the gene trees are completely linked – all assumed to follow the same tree = the species tree.
- Under Consensus approaches, the gene trees are completely unlinked – the species tree is assumed to be the tree of highest probability across genes
- Under the BEST model, the genes are correlated only by their coalescent relationship to the common species tree.



BEST Method Assumptions

- Given the species tree, the gene trees are conditionally independent.
- Given the gene tree, the DNA sequences are conditionally independent of the species tree.
- Random mating in each population.
- No gene flow after species divergence.
- No recombination within a locus.

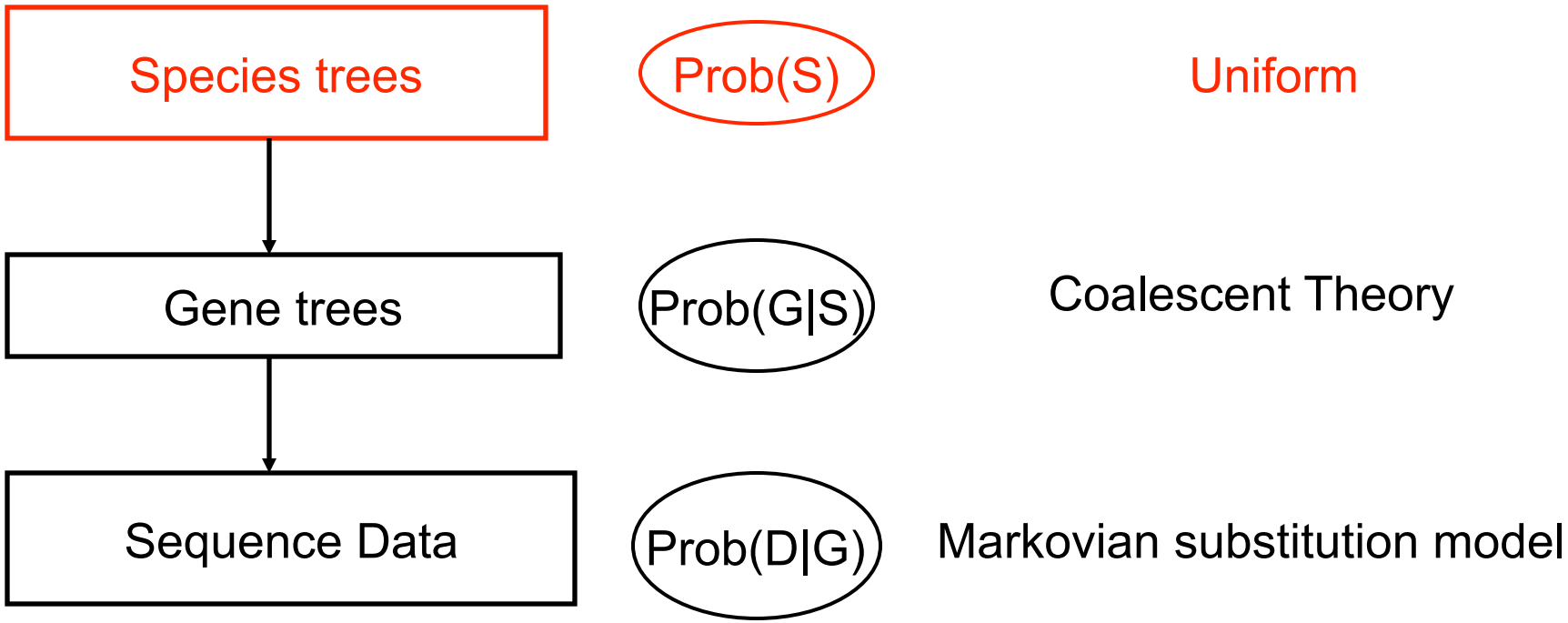
The BEST posterior

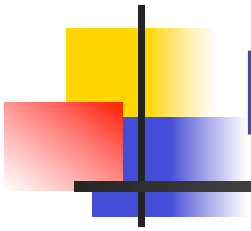
- The BEST method provides the joint posterior distribution of the gene trees and species tree.

$$f(S,\mathbf{G} \mid D) = \frac{f(D \mid \mathbf{G}) f(\mathbf{G} \mid S) f(S)}{f(D)}$$

Bayesian Estimation of Species Trees (BEST)

Hierarchical model



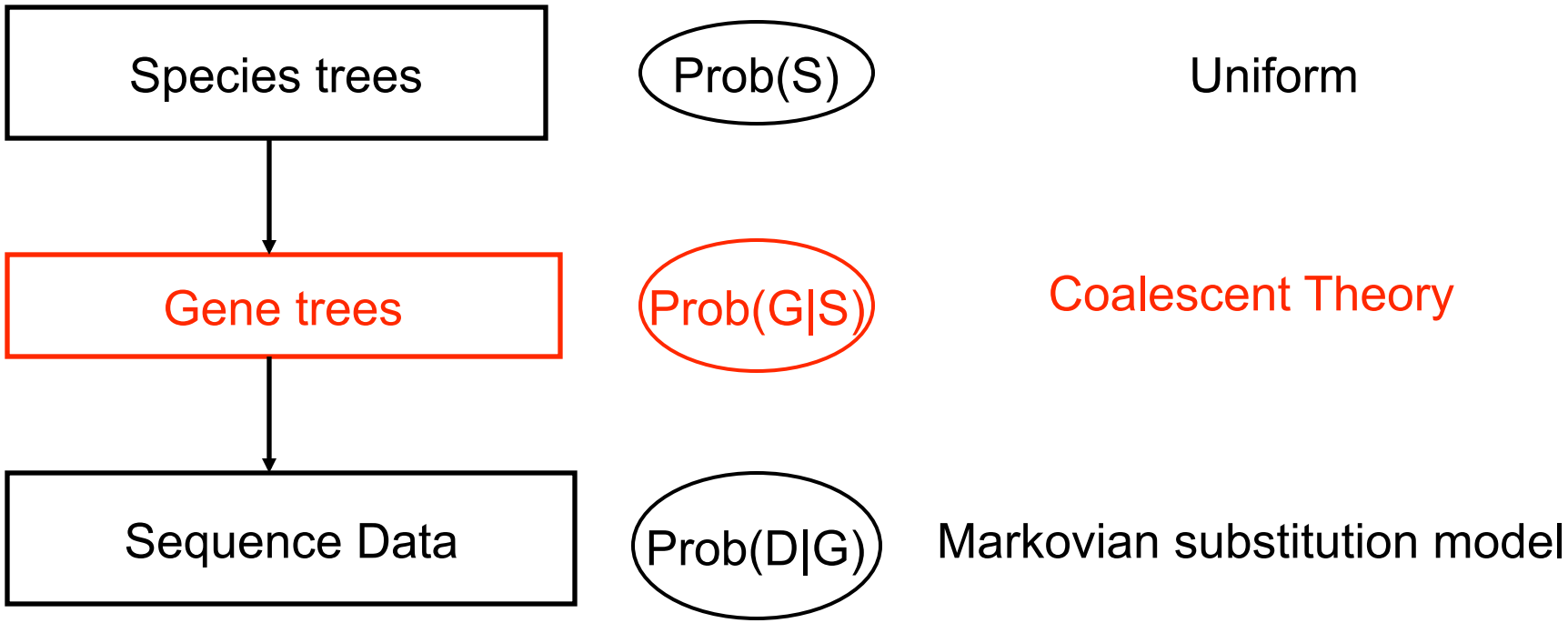


Prob(S) in BEST

- Use uniform distribution for topology of the species tree.
- Independent inverse gamma (α,β) distributions for the population sizes.
  - Prior mean of θ is $\beta/(\alpha-1)$.
  - Prior variance of θ is $\beta^2/((\alpha-1)^2(\alpha-2))$.
  - This is a conjugate prior!

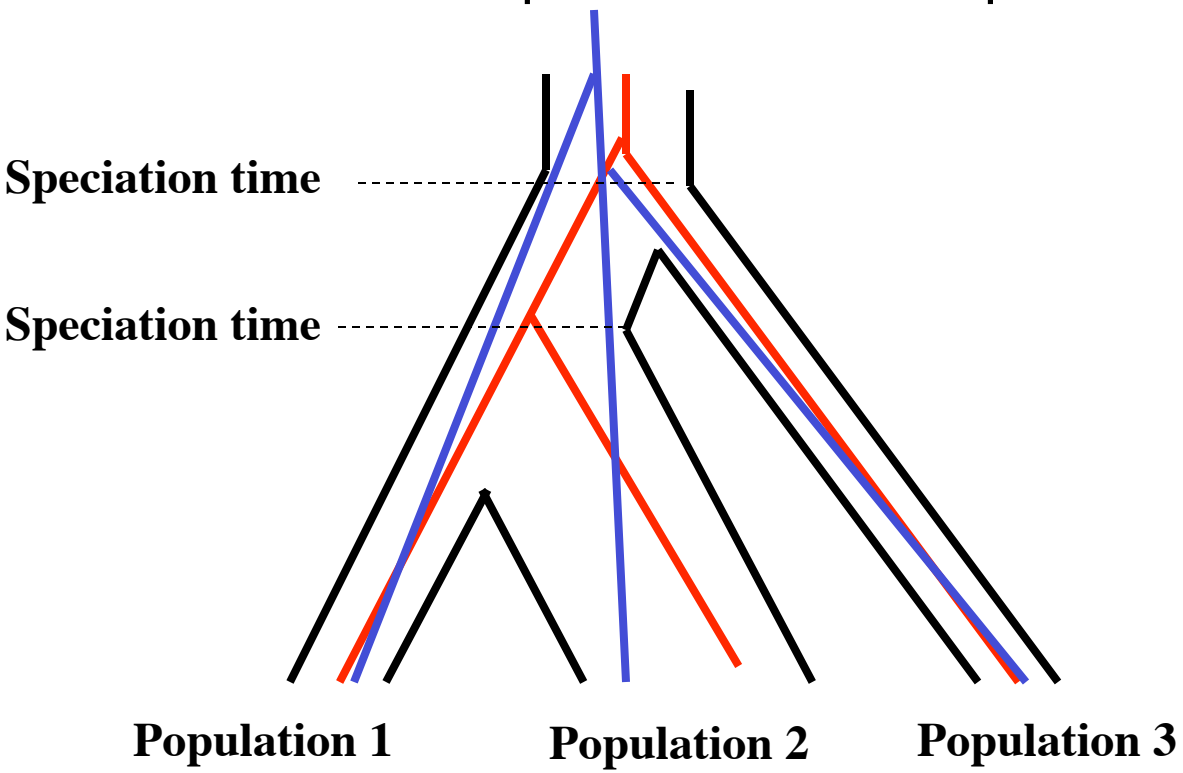
Bayesian Estimation of Species Trees (BEST)

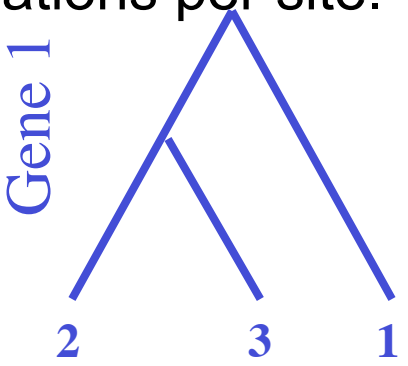
Hierarchical model

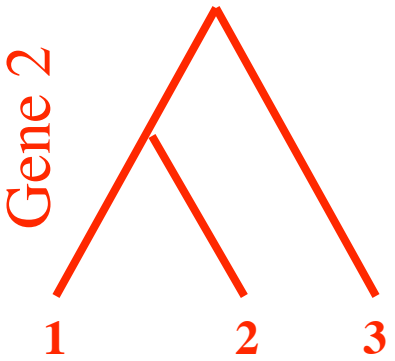


Species Trees

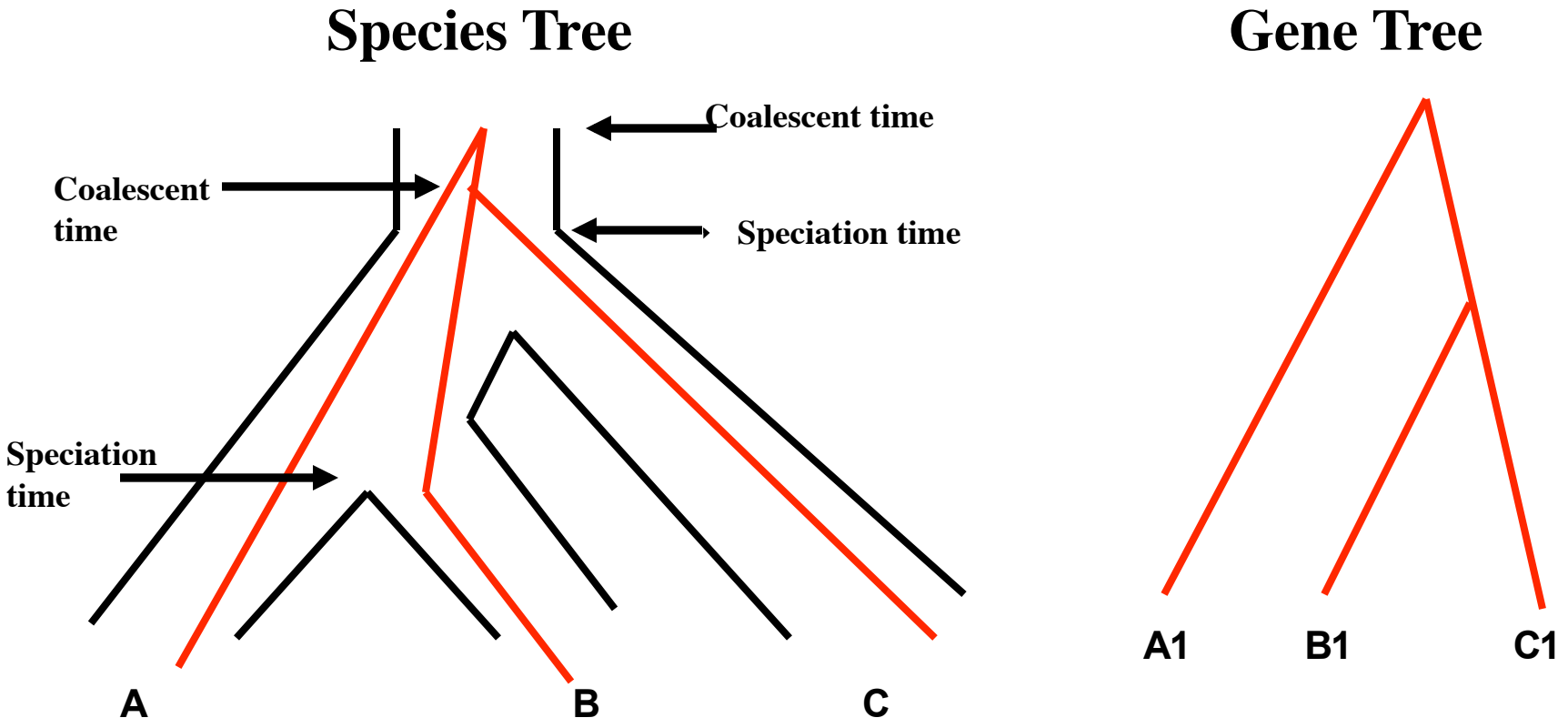
- follows lineages of populations, branch lengths are rescaled to measure expected mutations per 2N generations per site.



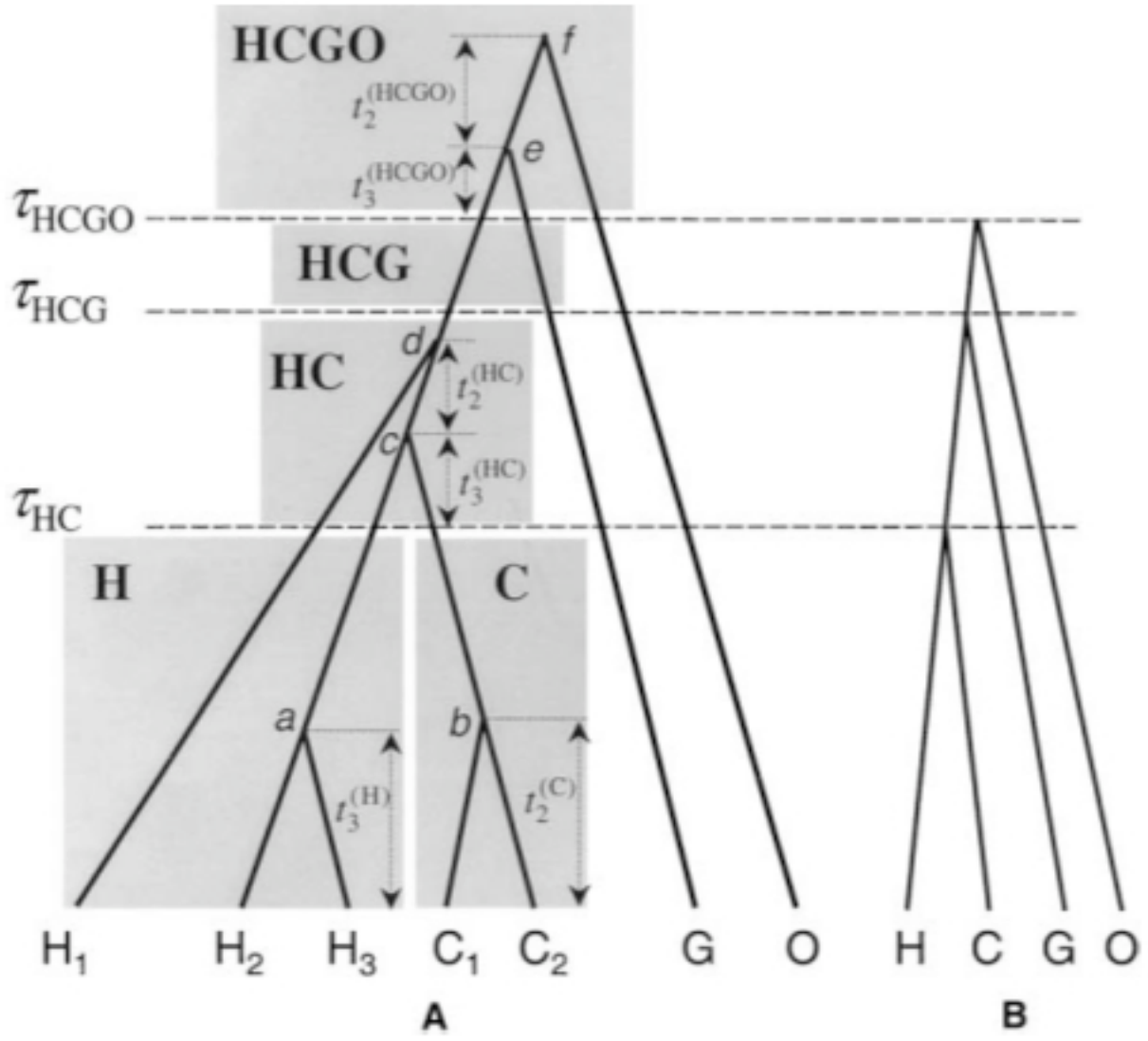




Prob(G|S)



Given the species tree, gene trees are assumed independent. The joint distribution of gene trees given the species tree is just the product of the individual conditional probability distributions [an explicit formula for this distribution is given, for example, in Rannala & Yang (2003)].



Gene Tree (A) compatible with a species tree (B) – more general situation

FIGURE FROM RANNALA & YANG, 2003

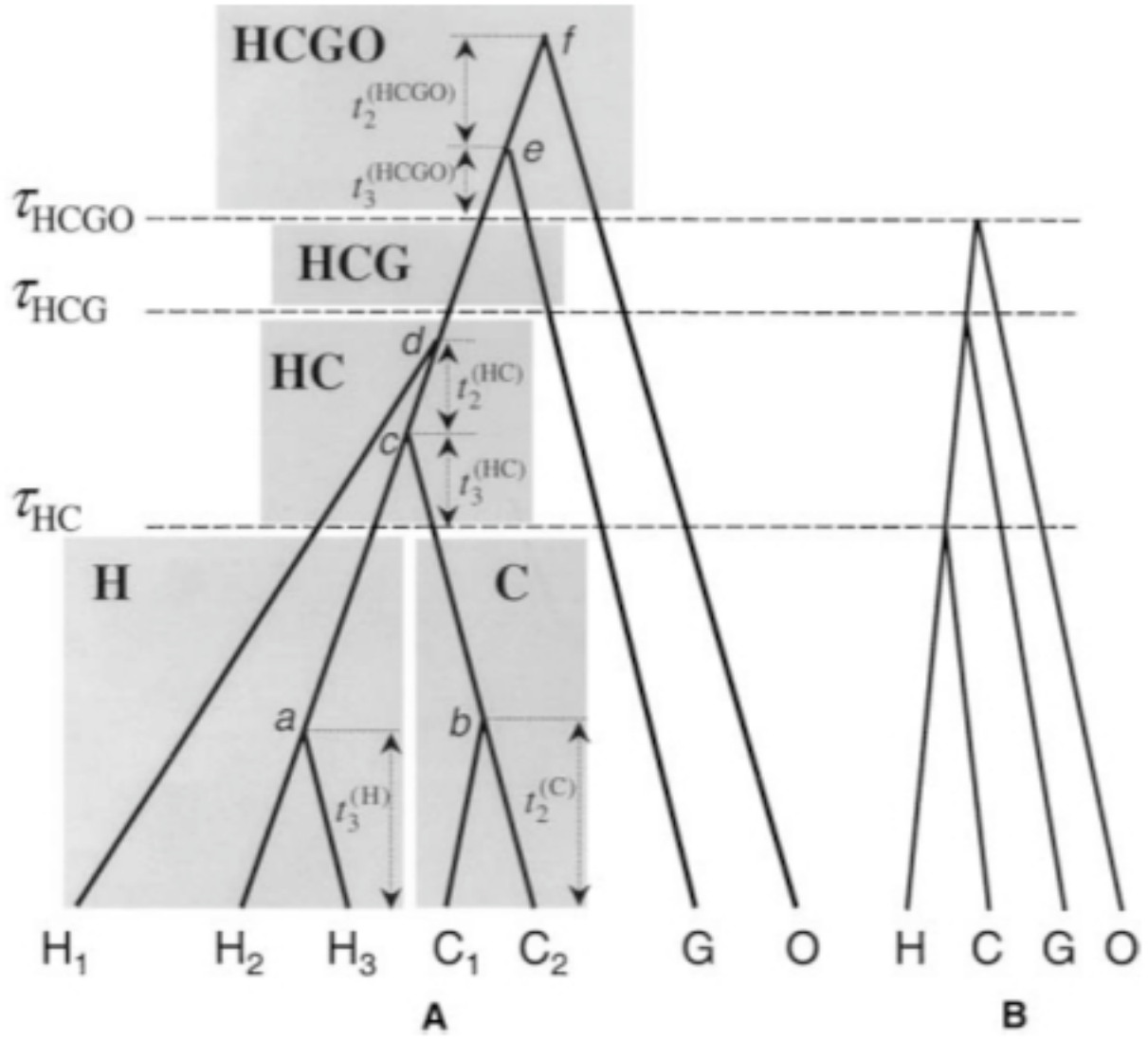
Calculating P(**G**|S)
Rannala & Yang (2003)

- Suppose you have m lineages going into a population and n coming out.

- In each population, the joint probability distribution of the gene tree topology in the population and its coalescent times t_m , t_{m-1} ….t_{n+1} is

$$[\prod_{j=n+1}^{m} \frac{2}{\theta} \exp\{-\frac{j(j-1)}{\theta} t_j\}] \times \exp\{-\frac{n(n-1)}{\theta}(\tau - (t_m + t_{m-1} + ... + t_{n+1}))\}$$

- The probability of the gene tree and coalescent times for the locus is the product of such probabilities across all the populations.



$$\left[\frac{2}{\theta_H}e^{\{-\frac{6t_3^H}{\theta_H}\}}e^{\{\frac{-2(\tau_{HC}-t_3^H)}{\theta_H}\}}\right]\left[\frac{2}{\theta_C}e^{\{\frac{-2t_2^C}{\theta_C}\}}\right]\left[\frac{2}{\theta_{HC}}e^{\{\frac{-6t_3^{HC}}{\theta_{HC}}\}}\frac{2}{\theta_{HC}}e^{\{\frac{-2t_2^{HC}}{\theta_{HC}}\}}\right]\left[e^{\frac{-2(\tau_{HCG}-(\tau_{HC}+t_3^{HC}+t_2^{HC}))}{\theta_{HCG}}}\right]\left[\frac{2}{\theta_{HCGO}}e^{\{\frac{-6t_3^{HCGO}}{\theta_{HCGO}}\}}\frac{2}{\theta_{HCGO}}e^{\{\frac{-2t_2^{HCGO}}{\theta_{HCGO}}\}}\right]$$

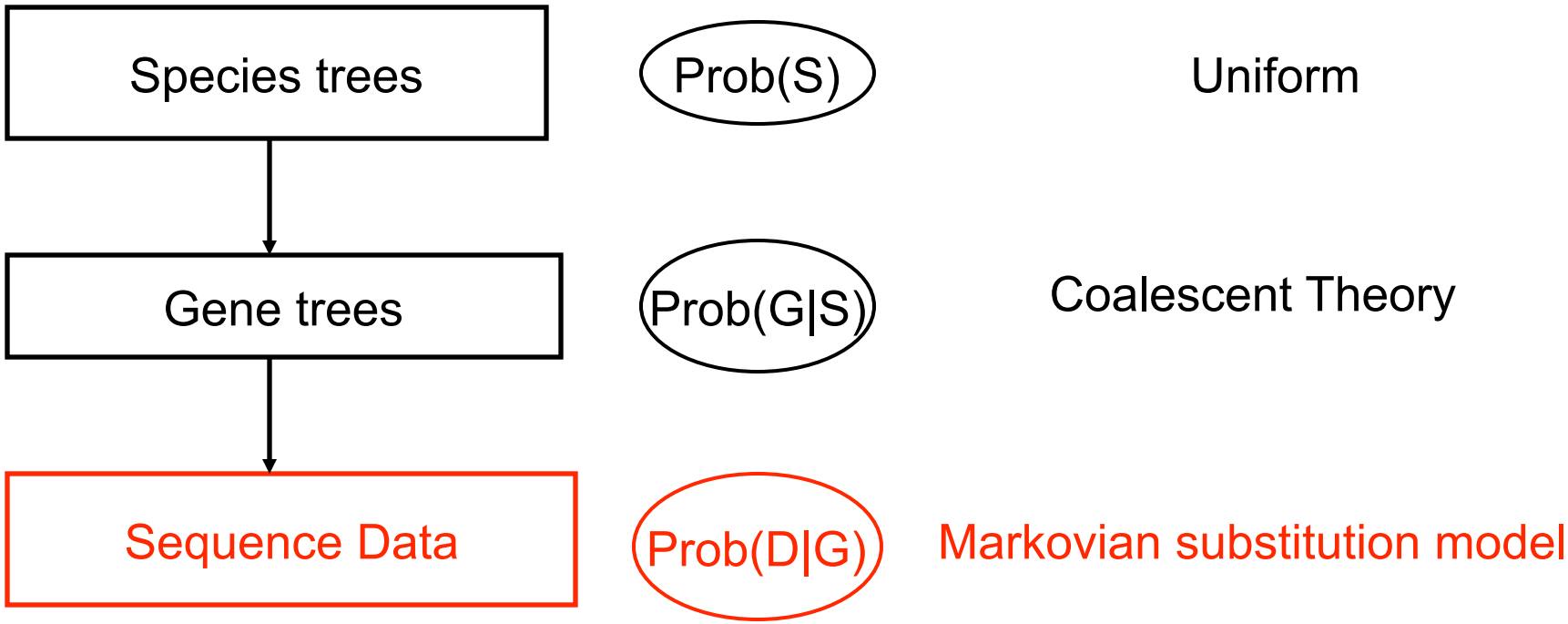
Calculating P(**G**|S)
Rannala & Yang (2003)

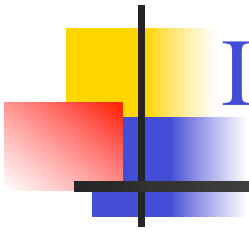
- Given the species tree, gene trees are independent. The joint distribution of gene trees given species tree is the product of probability distribution of each gene tree given species tree.

$$\prod_{loci}\prod_{population}\prod_{j=n+1}^{m}[\frac{2}{\theta}\exp\{-\frac{j(j-1)}{\theta}t_j\}]\times\exp\{-\frac{n(n-1)}{\theta}(\tau-(t_m+t_{m-1}+....+t_{n+1}))\}$$

Bayesian Estimation of Species Trees (BEST)

Hierarchical model





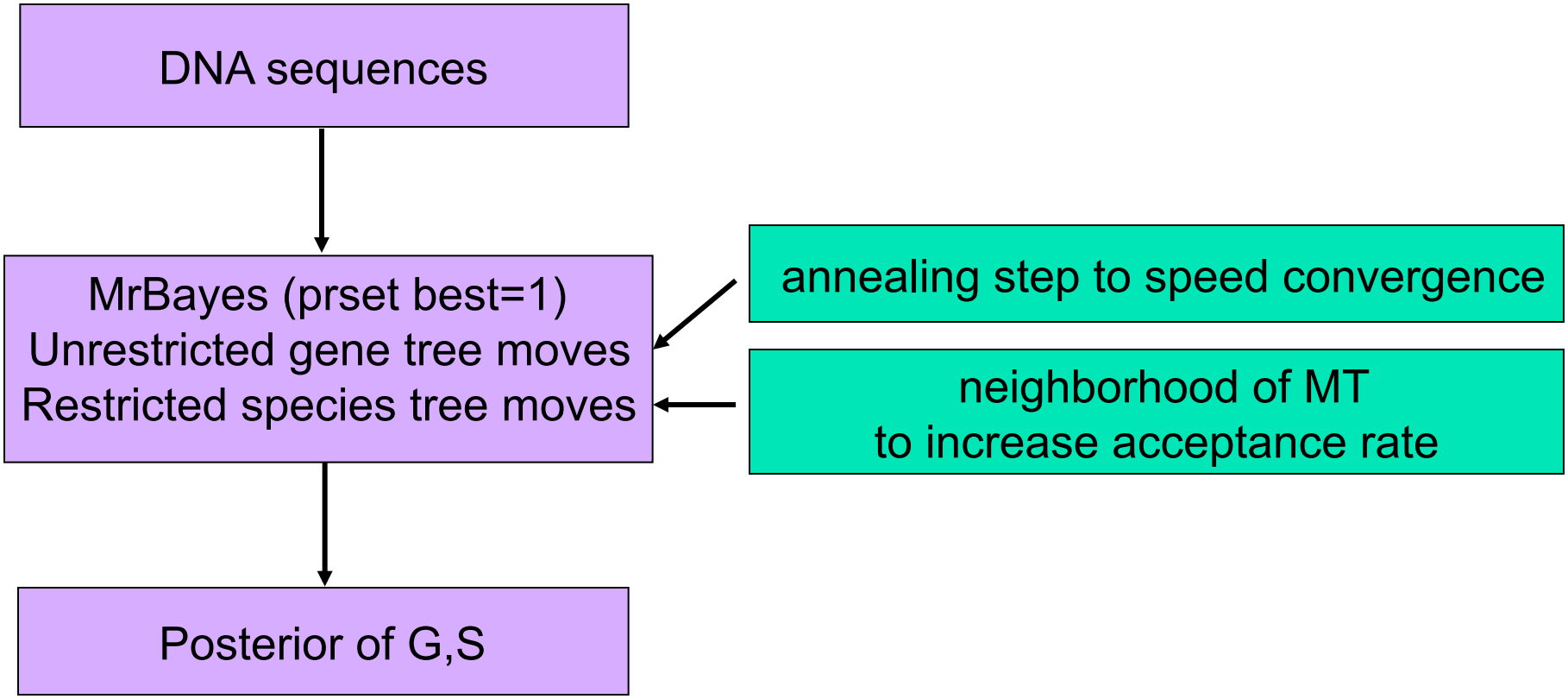
Implementation: MrBayes with BEST

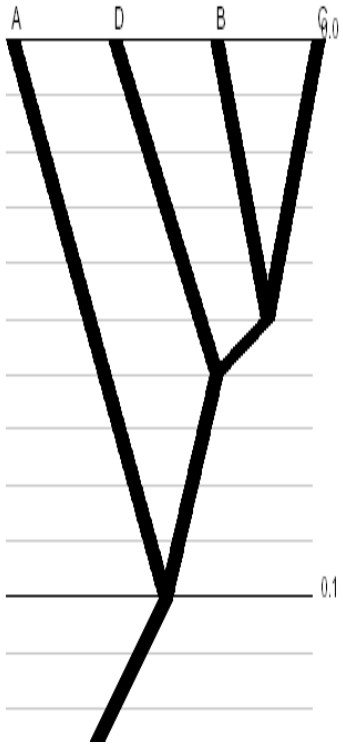
Step 1: Use MrBayes to propose vectors of joint gene trees (unlinked).

Step 2: Given those gene trees, propose a compatible species tree.

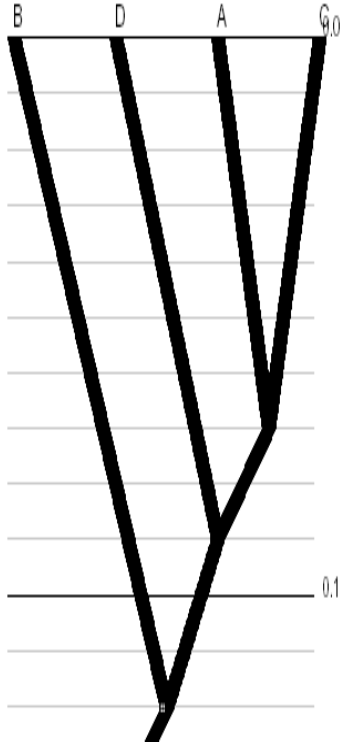
Step 3: Implement the chain fully within MrBayes using the usual properties of the MCMC as proposed by the user.

The Algorithm

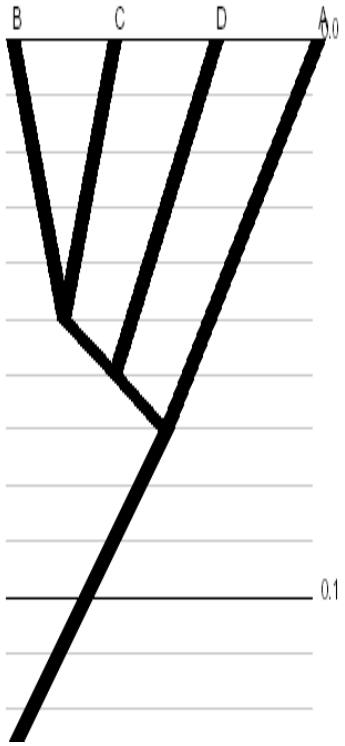




Gene tree 1

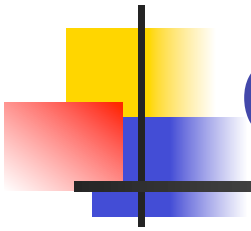


Gene tree 2



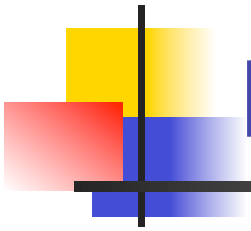
Max tree

Pair of species	Gene split time for gene 1	Gene split time for gene 2	minimum
A,B	0.1	0.12	0.1
A,C	0.1	0.07	0.07
A,D	0.1	0.09	0.09
B,C	0.05	0.12	0.05
B,D	0.06	0.12	0.06
C,D	0.06	0.09	0.06



Calculation of the Maximum Tree

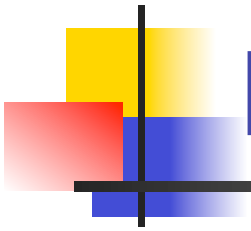
- Step 1: choose the smallest of the n(n-1)/2 constraints as the first constraint (involve S_1 and S_2)
- Step 2: choose the smallest constraint in which one species is S_1 or S_2 and the other is a new species S_3
- Step 3: repeat until all species are added into the set
- Step 4: build an ultrametric tree from the new set of (n-1) constraints.



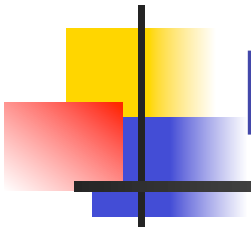
Properties of the Maximum Tree

- Provides the tree with the largest possible speciation times in the space restricted by available gene trees.
- Provides the maximum likelihood estimate of the species tree given a set of gene trees and equal population sizes.
- Provides a consistent estimate of the species tree even with unequal population sizes – as long as gene tree estimates are themselves consistent.

Implemented as an add-on to MrBayes



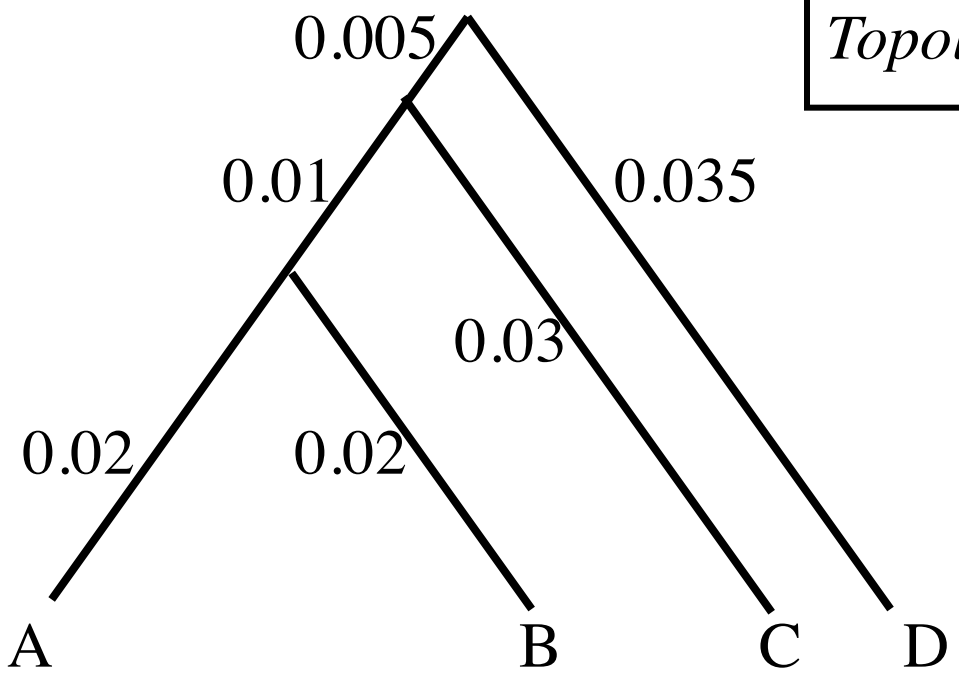
- Use the prset command in MrBayes to specify the parameters of BEST
- Example:
  - prset thetapr=invgamma(3,0.003) genemupr=uniform(0.2,1.8) BEST=1;
  - unlink topology=(all) brlens=(all) genemu=(all);



New output

- The species trees with divergence times and population sizes are saved in the “filename.sptree.t” file.
- The MrBayes “.p” will inclde the Genemu parameter estimates.
- After running sumt in MrBayes you will get
  - a “.con” file with the consensus species tree
  - A “.trprobs” file with the nodal posterior probabilities
  - A “.parts” file with means & st. devs. For times and population sizes.

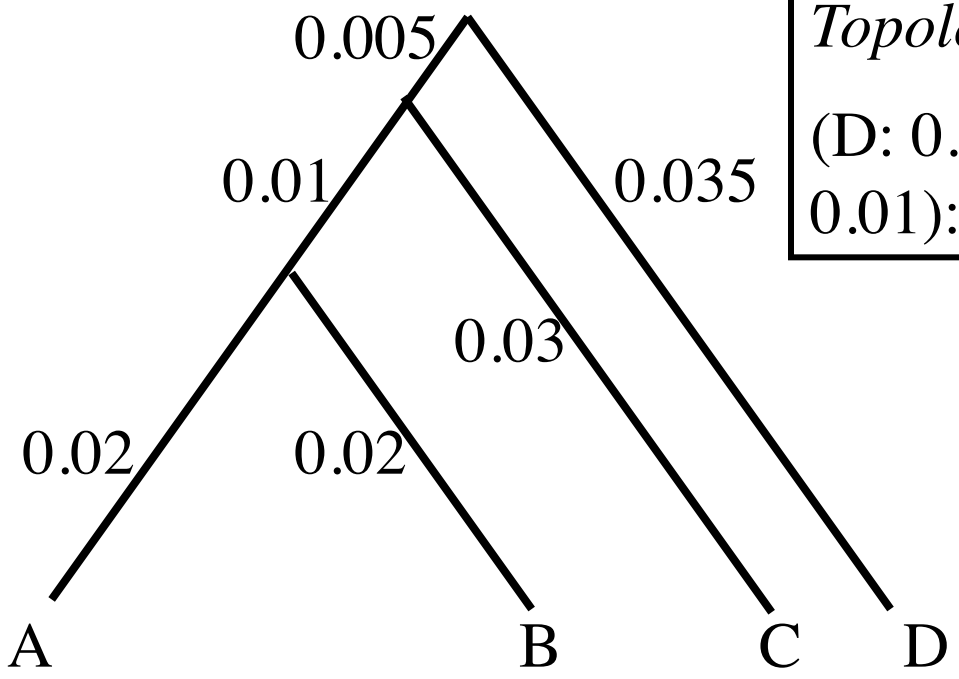
Species Tree Notation



Topology: (D(C(A,B)))

$$\theta_{AB} = 0.3,\ \theta_{ABC} = 0.2,\ \theta_{ABCD} = 0.25$$

Species Tree Notation

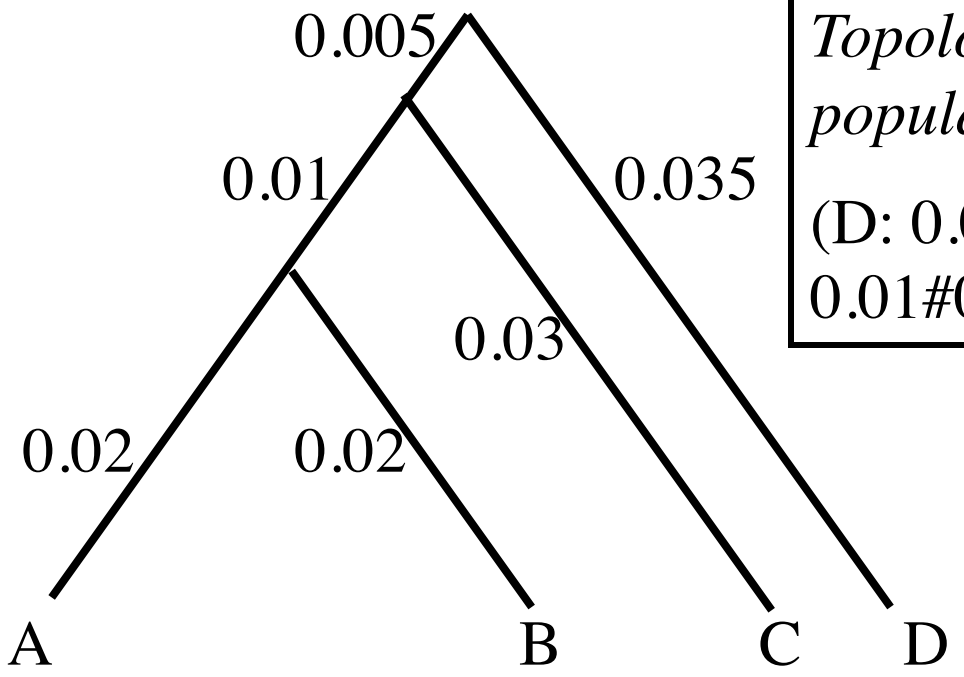


Topology & branch lengths:

(D: 0.035(C:0.03(A:0.02,B:0.02): 0.01):0.005)

$$\theta_{AB} = 0.3,\ \theta_{ABC} = 0.2,\ \theta_{ABCD} = 0.25$$

Species Tree Notation



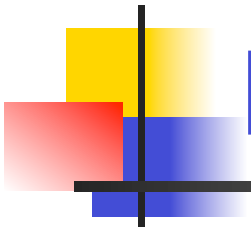
Topology, branch lengths, & population sizes:

(D: 0.035(C:0.03(A:0.02,B:0.02): 0.01#0.3):0.005#0.2)#0.25

$$\theta_{AB} = 0.3,\ \theta_{ABC} = 0.2,\ \theta_{ABCD} = 0.25$$

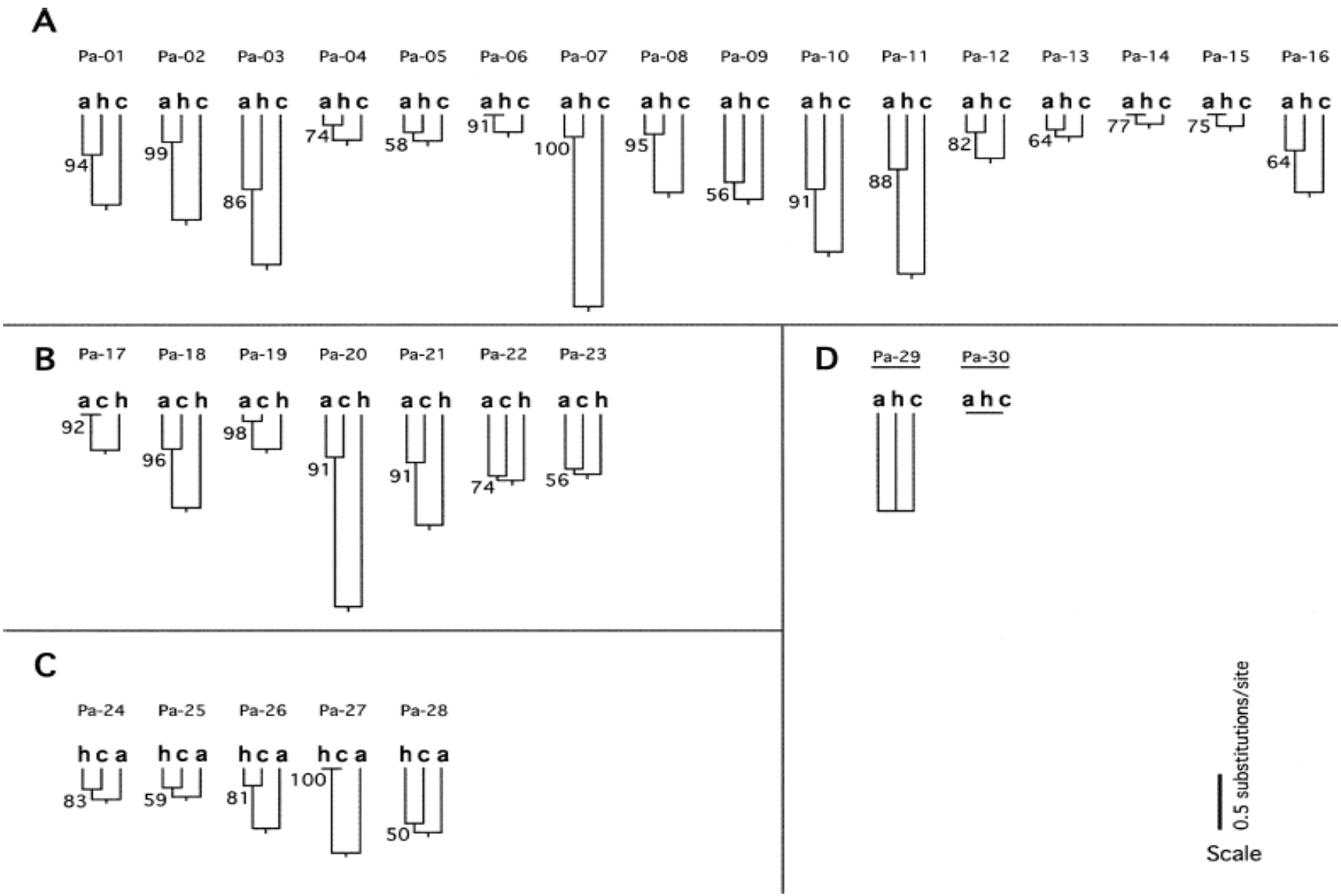
Simulation Results: comparison with concatenation

- Gene trees generated from true species tree: (8,(7,(6,(5,(4,(3,(1,2))))))); then 500bp sequences generated along each tree.
- Concatenation method finds (8,(7,((5,6),(4,(3, (1,2)))))) with probability 0.98
- BEST method finds (8,(7,(6,(5,(4,(3,(1,2))))))) with probability 0.96.
- Applying the coalescent model successfully recovered the true species tree while the concatenation method estimated the wrong tree.

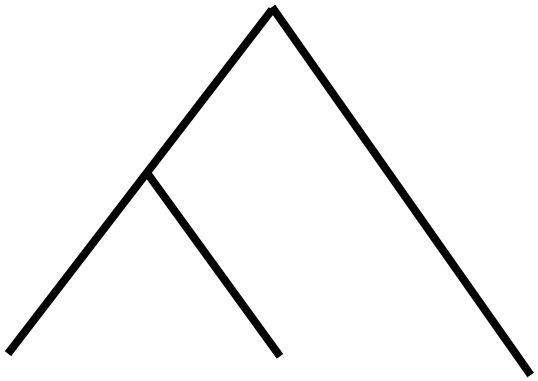


Example 1: Australian Finches

The BEST method was applied to a data set of three closely related Australian grass finches (Poephila) distributed in northern Australia. There are 30 genes and 4 species (one is an outgroup to root the tree) in the data. This data has already been analyzed by Jennings and Edwards, 2005.

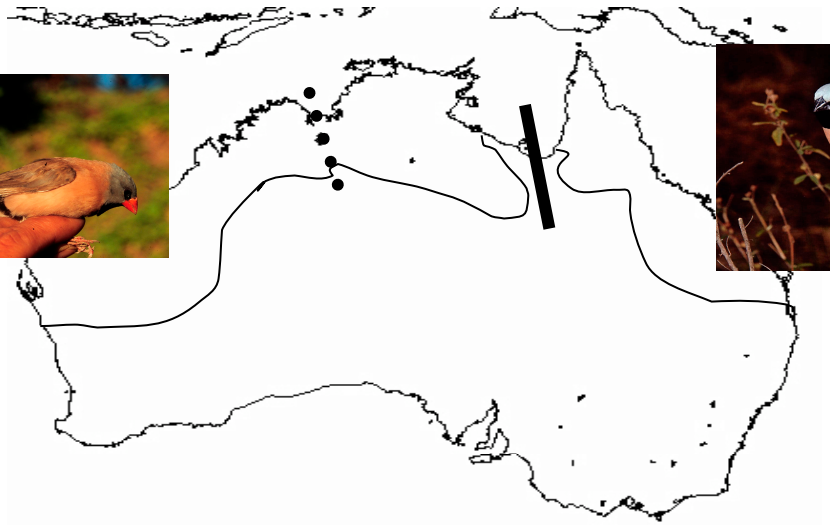

A
Pa-01
Pa-02
Pa-03
Pa-04
Pa-05
Pa-06
Pa-07
Pa-08
Pa-09
Pa-10
Pa-11
Pa-12
Pa-13
Pa-14
Pa-15
Pa-16
a h c
94
99
86
74
58
91
100
95
56
91
88
82
64
77
75
64
B
Pa-17
Pa-18
Pa-19
Pa-20
Pa-21
Pa-22
Pa-23
a c h
92
96
98
91
91
74
56
C
Pa-24
Pa-25
Pa-26
Pa-27
Pa-28
h c a
83
59
81
100
50
D
Pa-29
Pa-30
a h c
0.5 substitutions/site
Scale

Consensus method weakly supports



Assumed phylogeny of *Poephila* finches

Long-tailed Finch

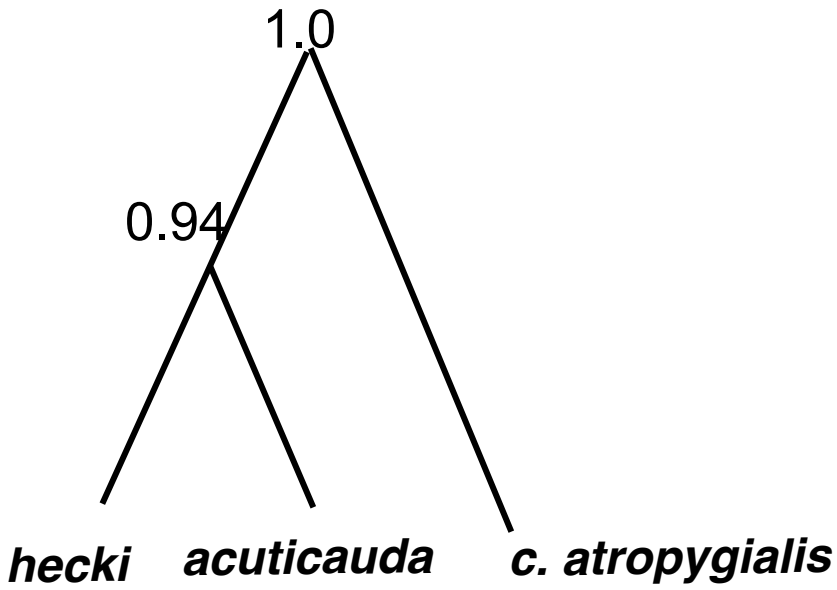


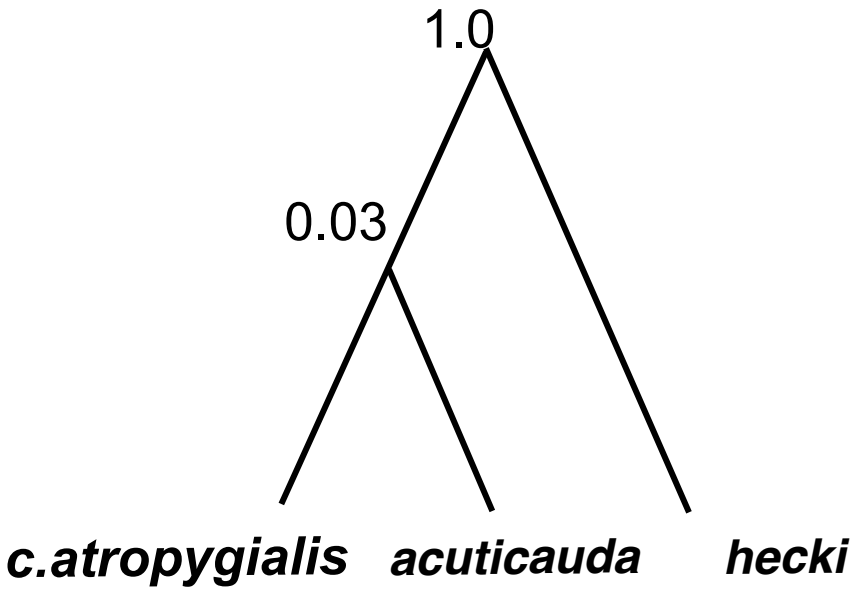
Black-throated Finch

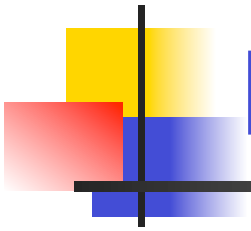
Estimated species tree distribution using BEST

	Independent prior			Joint prior		
	(2,(1,3))	(3,(1,2))	(1,(2,3))	(2,(1,3))	(3,(1,2))	(1,(2,3))
Pa-1	0.184	0.671	0.146	0.171	0.683	0.146
Pa-2	0.337	0.353	0.309	0.299	0.375	0.326
Pa-3	0.062	0.88	0.058	0.056	0.915	0.029
Pa-4	0.331	0.331	0.337	0.221	0.452	0.327
Pa-5	0.319	0.319	0.361	0.264	0.398	0.338
Pa-6	0.012	0.966	0.022	0.047	0.894	0.059
Pa-7	0	1	0	0	1	0
Pa-8	0	1	0	0	1	0
Pa-9	0.042	0.912	0.046	0.026	0.935	0.038
Pa-10	0.222	0.547	0.232	0.117	0.699	0.184
Pa-11	0	1	0	0	1	0
Pa-12	0.319	0.353	0.327	0.293	0.449	0.258
Pa-13	0.493	0.503	0.004	0.257	0.743	0
Pa-14	0.242	0.503	0.255	0.254	0.497	0.249
Pa-15	0.325	0.349	0.325	0.151	0.578	0.271
Pa-16	0.335	0.333	0.331	0.233	0.496	0.271
Pa-17	0.042	0.02	0.938	0.073	0.156	0.772
Pa-18	0	0	1	0	0	1
Pa-19	0	0	1	0	0	1
Pa-20	0	0.002	0.998	0	0	1
Pa-21	0	0	1	0	0	1
Pa-22	0.04	0.076	0.884	0.045	0.085	0.87
Pa-23	0.014	0.064	0.922	0.019	0.046	0.935
Pa-24	0	1	0	0.002	0.998	0
Pa-25	0.311	0.339	0.349	0.232	0.503	0.265
Pa-26	0.782	0.212	0.006	0.482	0.5	0.018
Pa-27	1	0	0	1	0	0
Pa-28	0.389	0.305	0.305	0.298	0.431	0.271
Pa-29	0.01	0.653	0.337	0.001	0.739	0.26
Pa-30	0.333	0.327	0.339	0.164	0.68	0.156
Average	0.205	0.434	0.361	0.157	0.508	0.335
concatenation	0	1	0			
Joint prior (1,139)	0.08	0.88	0.04			
Joint prior (1,1389)	0.03	0.95	0.02			
Joint prior (1,10)	0.08	0.89	0.03			
Joint prior (1,1)	0.01	0.94	0.05			

Estimated species tree distribution using BEST







Example 2: Yeast Data

Here the BEST method was applied to 106 protein coding regions sequence3d from eight species of yeast. This data has already been analyzed by several authors including Rokas, *et. al.*, 2003.

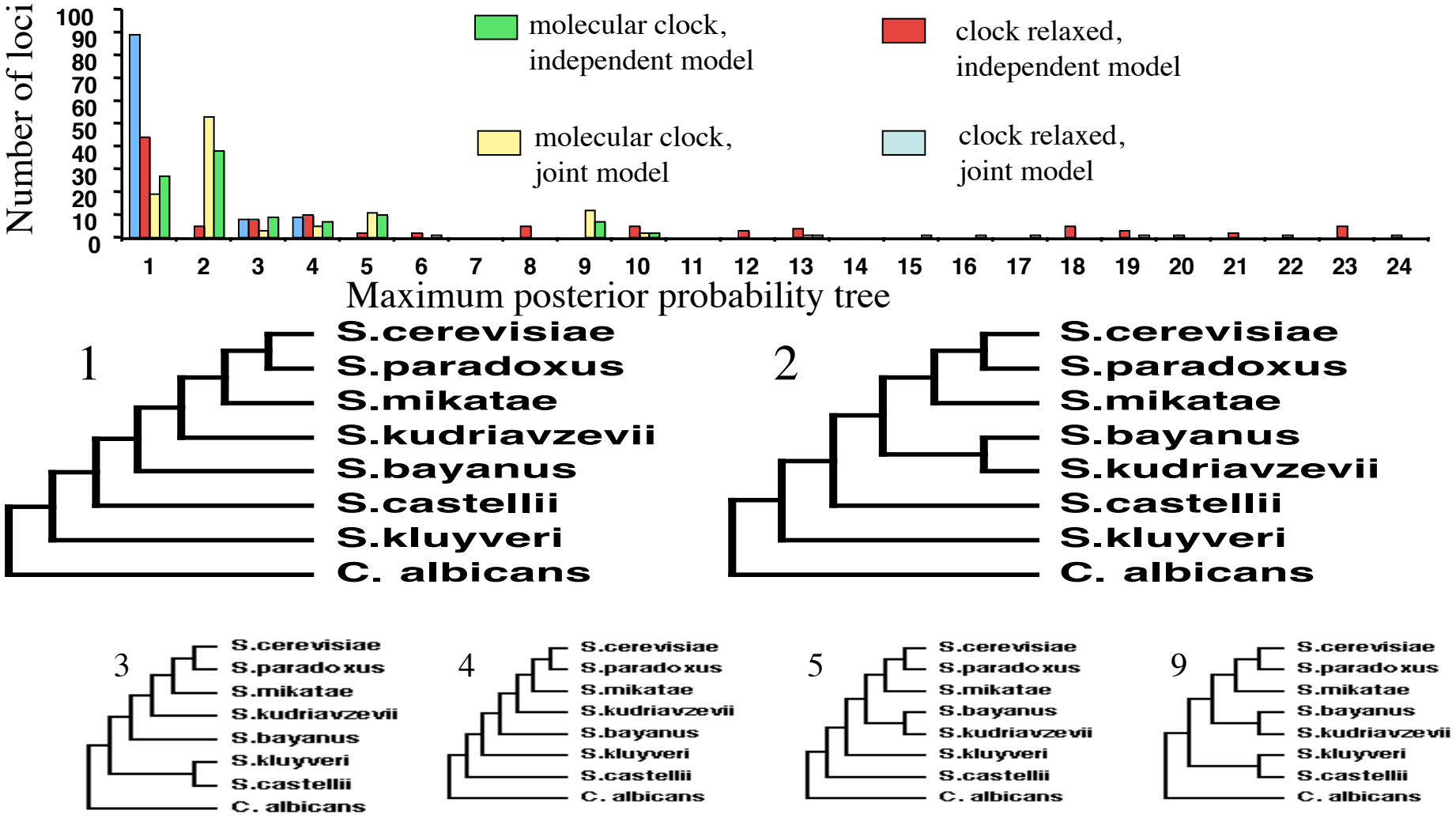

Number of loci
100
90
80
70
60
50
40
30
20
10
0
molecular clock, independent model
clock relaxed, independent model
molecular clock, joint model
clock relaxed, joint model
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
Maximum posterior probability tree
1
S.cerevisiae
S.paradoxus
S.mikatae
S.kudriavzevii
S.bayanus
S.castellii
S.kluyveri
C. albicans
2
S.cerevisiae
S.paradoxus
S.mikatae
S.bayanus
S.kudriavzevii
S.castellii
S.kluyveri
C. albicans
3
S.cerevisiae
S.paradoxus
S.mikatae
S.kudriavzevii
S.bayanus
S.kluyveri
S.castellii
C. albicans
4
S.cerevisiae
S.paradoxus
S.mikatae
S.kudriavzevii
S.bayanus
S.kluyveri
S.castellii
C. albicans
5
S.cerevisiae
S.paradoxus
S.mikatae
S.bayanus
S.kudriavzevii
S.kluyveri
S.castellii
C. albicans
9
S.cerevisiae
S.paradoxus
S.mikatae
S.bayanus
S.kudriavzevii
S.kluyveri
S.castellii
C. albicans

Fig. 2

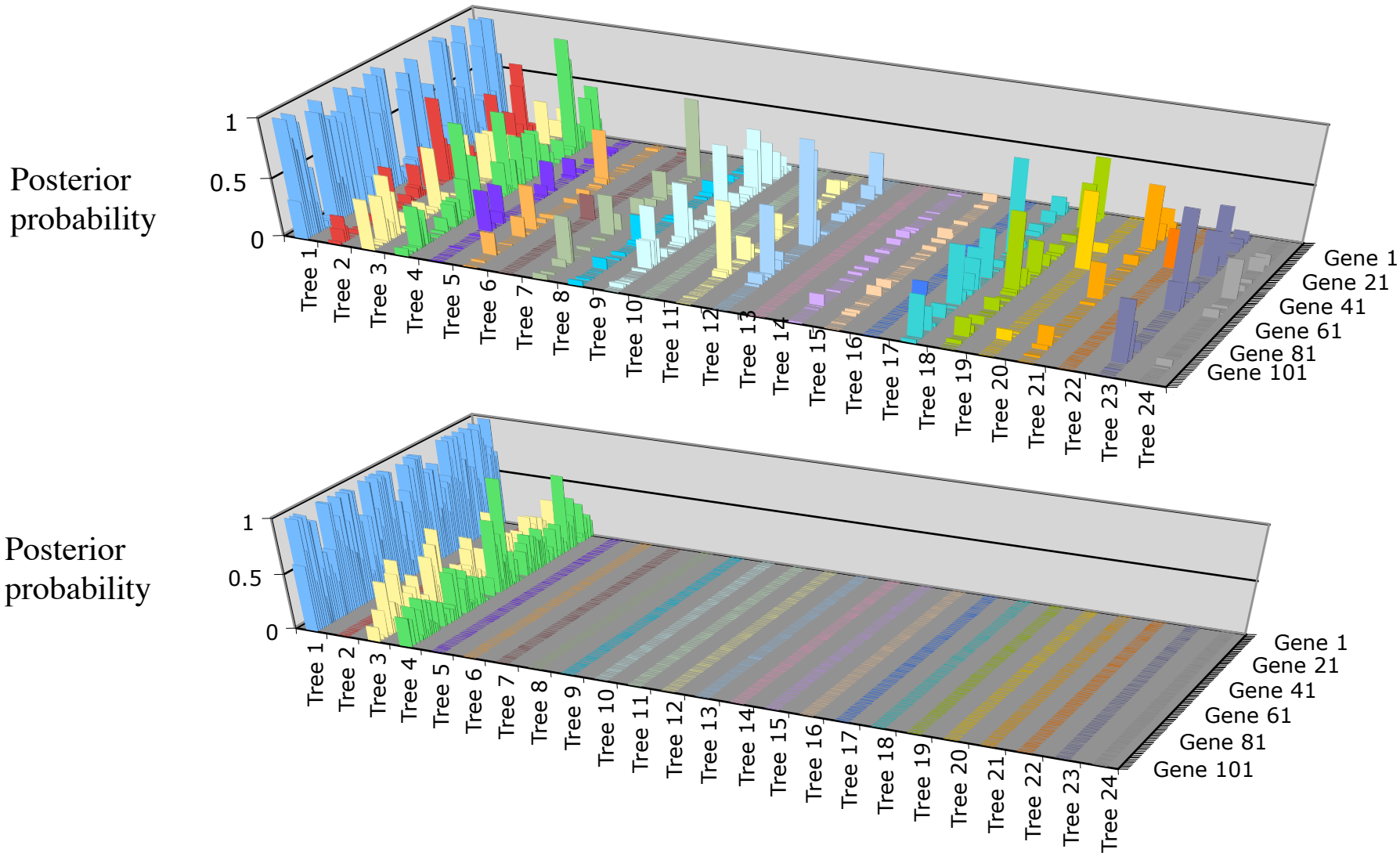

Posterior probability
1
0.5
0
Tree 1
Tree 2
Tree 3
Tree 4
Tree 5
Tree 6
Tree 7
Tree 8
Tree 9
Tree 10
Tree 11
Tree 12
Tree 13
Tree 14
Tree 15
Tree 16
Tree 17
Tree 18
Tree 19
Tree 20
Tree 21
Tree 22
Tree 23
Tree 24
Gene 1
Gene 21
Gene 41
Gene 61
Gene 81
Gene 101
Posterior probability
1
0.5
0
Tree 1
Tree 2
Tree 3
Tree 4
Tree 5
Tree 6
Tree 7
Tree 8
Tree 9
Tree 10
Tree 11
Tree 12
Tree 13
Tree 14
Tree 15
Tree 16
Tree 17
Tree 18
Tree 19
Tree 20
Tree 21
Tree 22
Tree 23
Tree 24
Gene 1
Gene 21
Gene 41
Gene 61
Gene 81
Gene 101

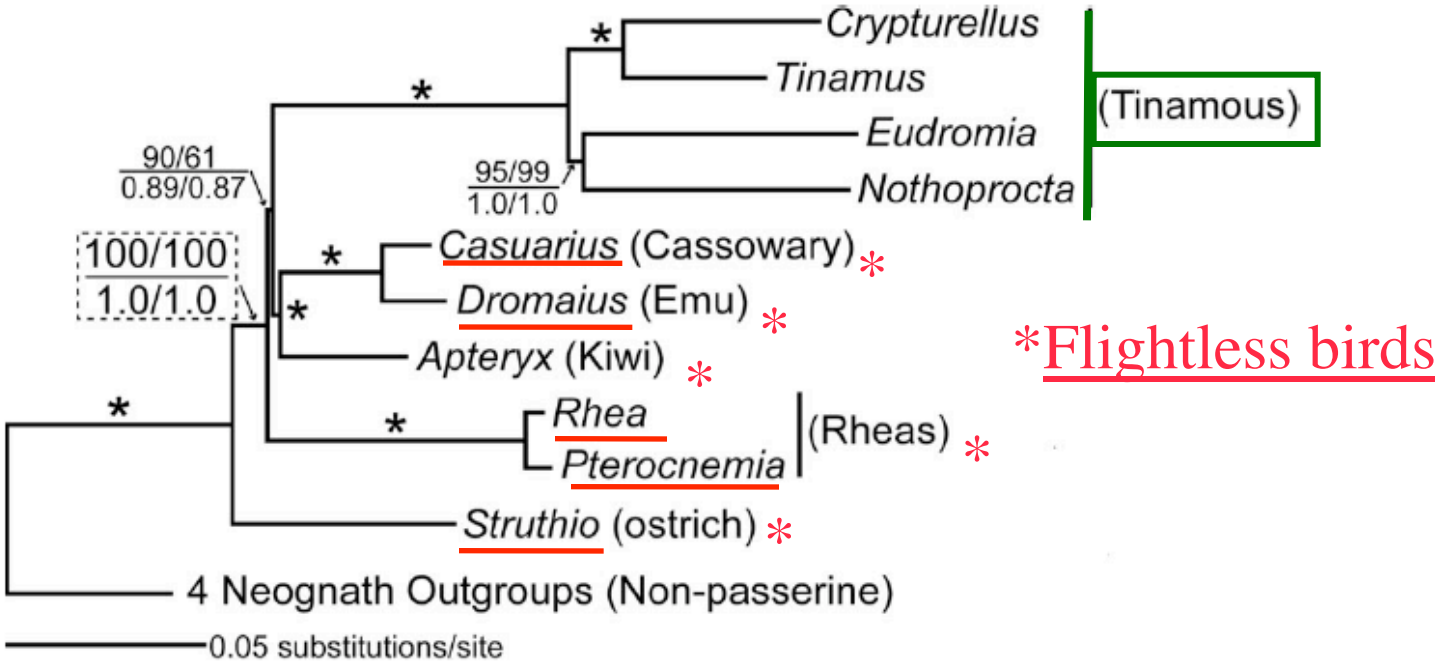
Fig. 3

Example 3: Single vs Multiple Losses of Flight

- Study of 20 nuclear genes by Harshman et al. *PNAS*, 2008 105: 13462-13467.
- Sought to resolve controversy about loss of flight in Ratites – using multiple methods of concatenated analyses.
- Assume odds of gaining flight are much lower than odds of loss.

Supermatrix (concatenation) analysis of paleognath relationships

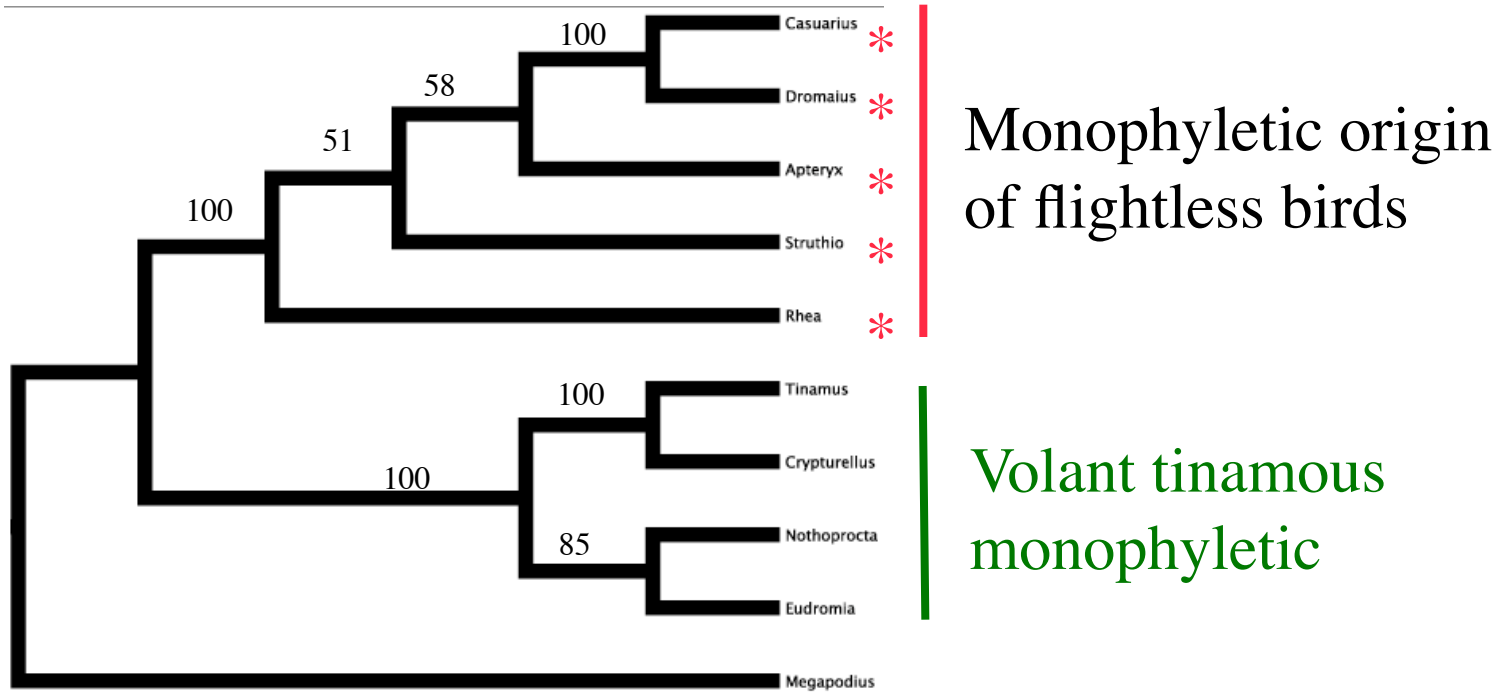
Harshman et al. 2008. *PNAS* 105: 13462



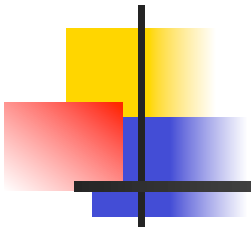
*Flightless birds

Bayesian analysis:10 million generations
Bootstrap analysis: 1000 repetitions

Species tree (BEST) analysis of paleognath relationships



100 million MCMC cycles



Summary

- Uncertainty in the evolutionary history can be summarized as a distribution of trees given the data (Bayesian approach)
- Multi-locus data is becoming increasingly available.
- Construction of species phylogenies need to combine this data in meaningful ways.
- An estimate of a phylogeny must be made together with an analysis of its variability.
- Key variables still need to be integrated into the methodology.
- See www.stat.osu.edu/~dkp/BEST