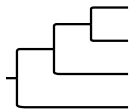
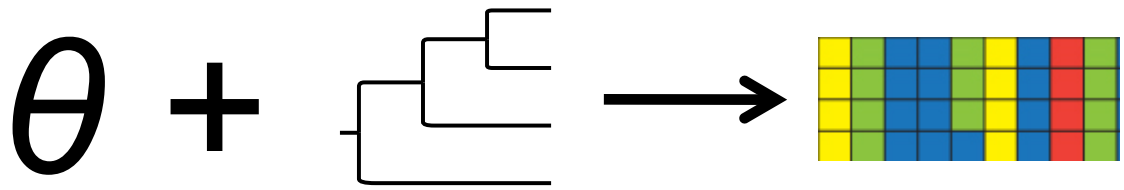
The Players

θ = Sequence Evolution Model Parameters

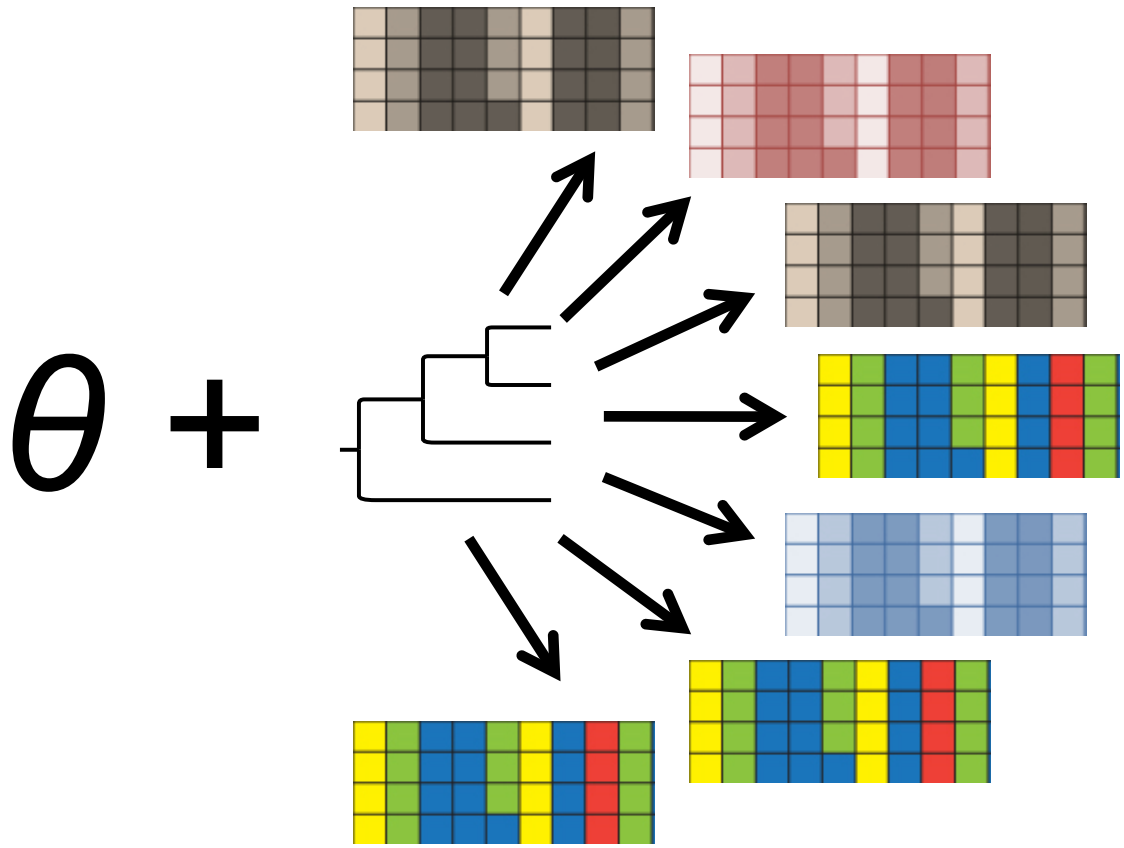
 = Tree Topology and Branch Lengths

 = Sequence Alignment

Simulation View of Models

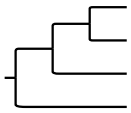


Simulation View of Models



How frequently is some data observed when datasets are repeatedly generated with a particular tree and set of model parameters?

The Likelihood Function

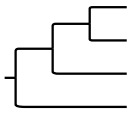
$P($  $|$  $, \theta)$

Read as “the probability of the sequence data given a tree and a set of model parameter values”.

The quantity by which the data provide information.

Compares how well different trees and models predict the observed data.

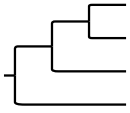
The Likelihood Function

$P($  $|$  $, \theta)$

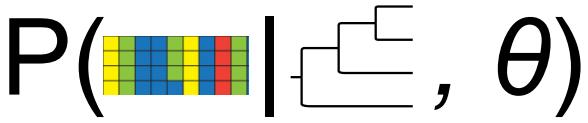
Read as “the probability of the sequence data given a tree and a set of model parameter values”.

The quantity by which the data provide information.

Compares how well different trees and models predict the observed data or as a “measure of surprise”.

NOT the same as $P($  $|$  $)$

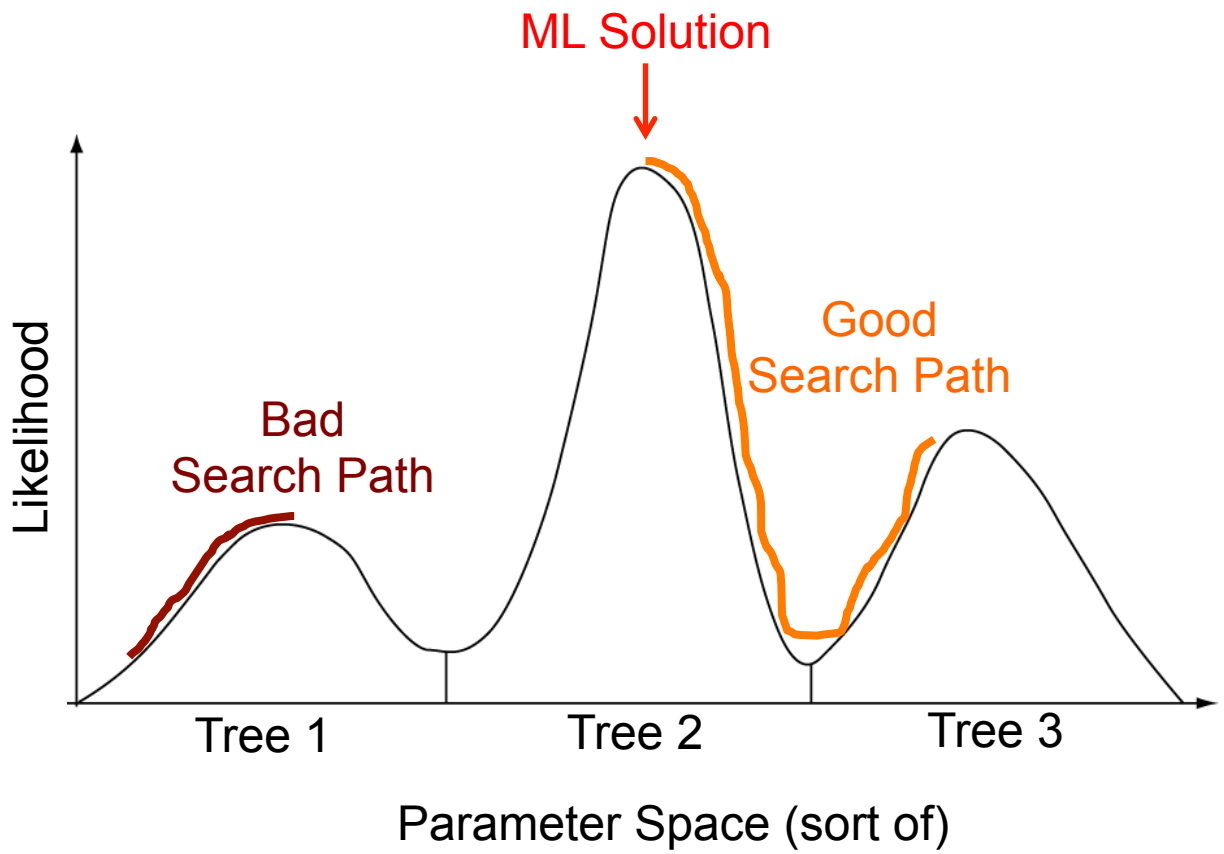
Maximum Likelihood



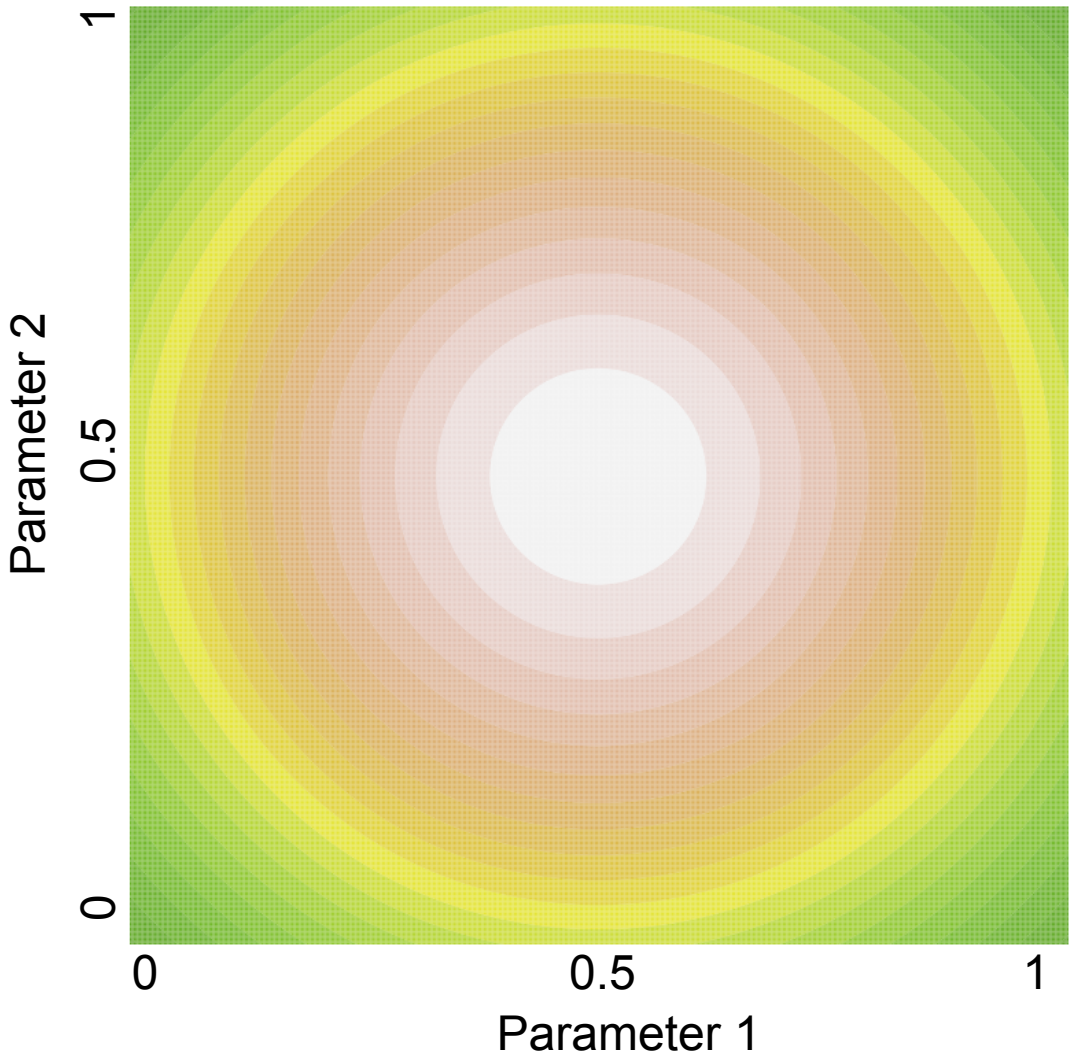
What tree and parameter values give the highest likelihood?

ML scores are just relative, so alone they don't tell us how confident we are in this solution, just that this is the preferred solution.

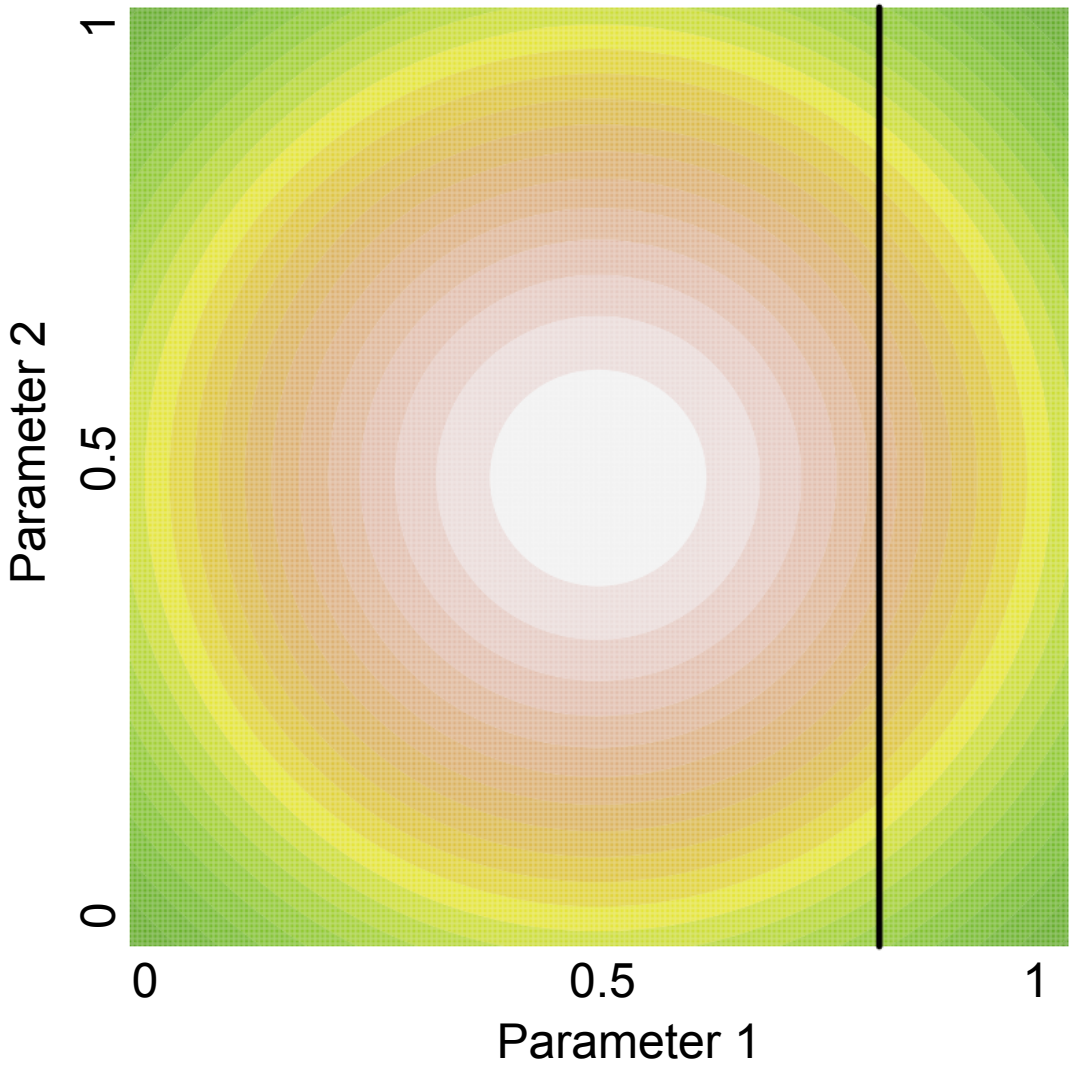
Maximum Likelihood



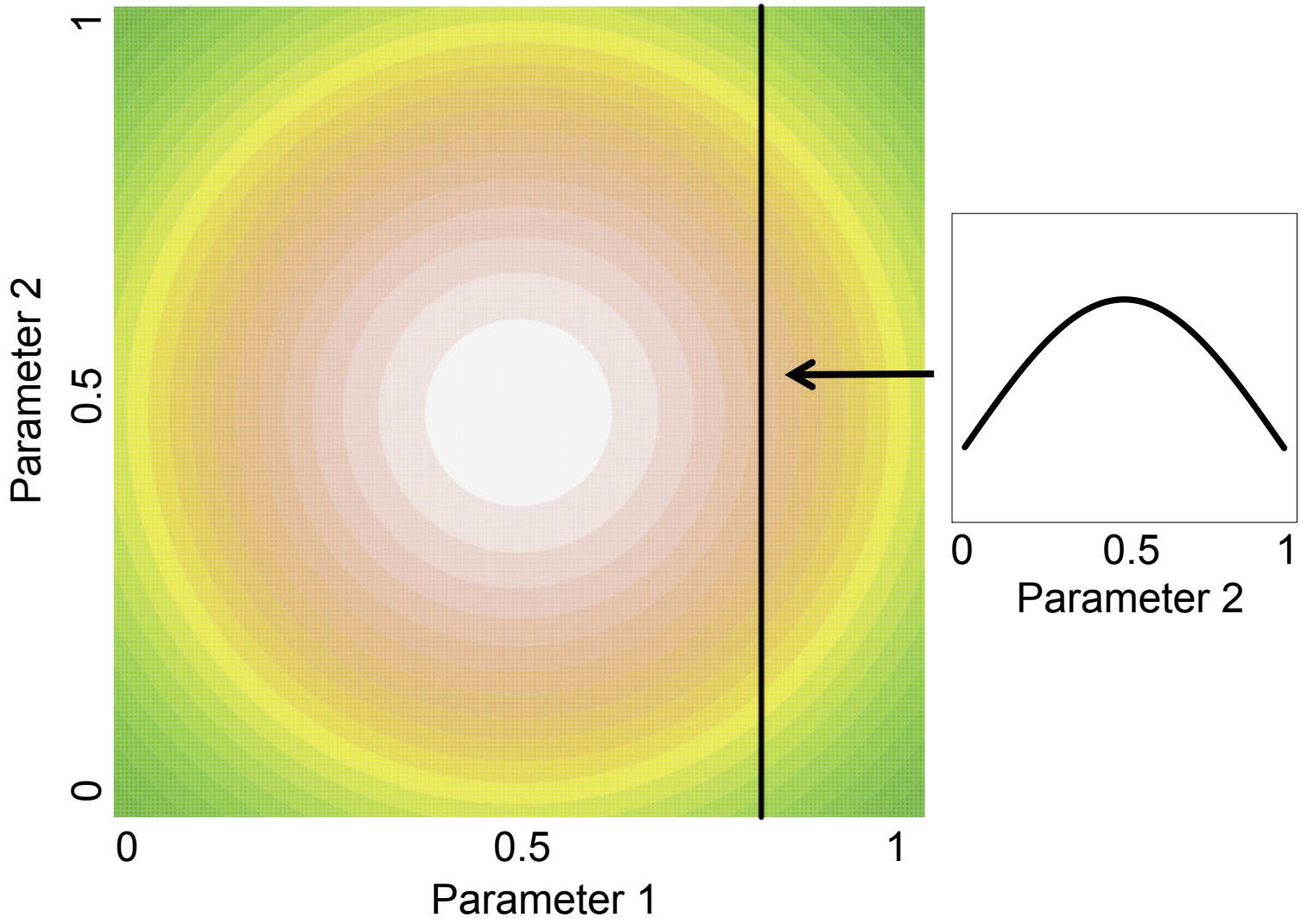
More Parameters = Better ML Score



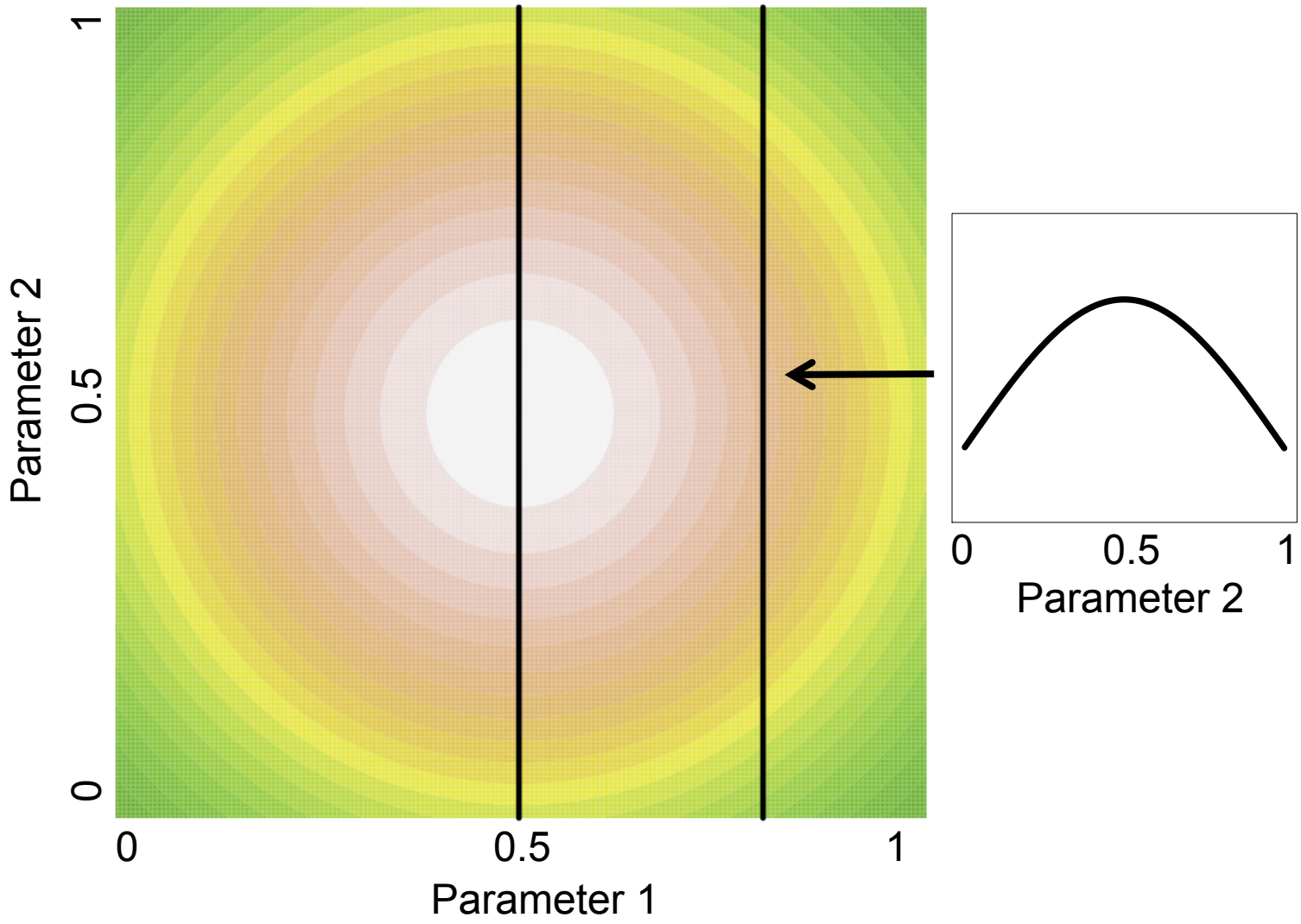
More Parameters = Better ML Score



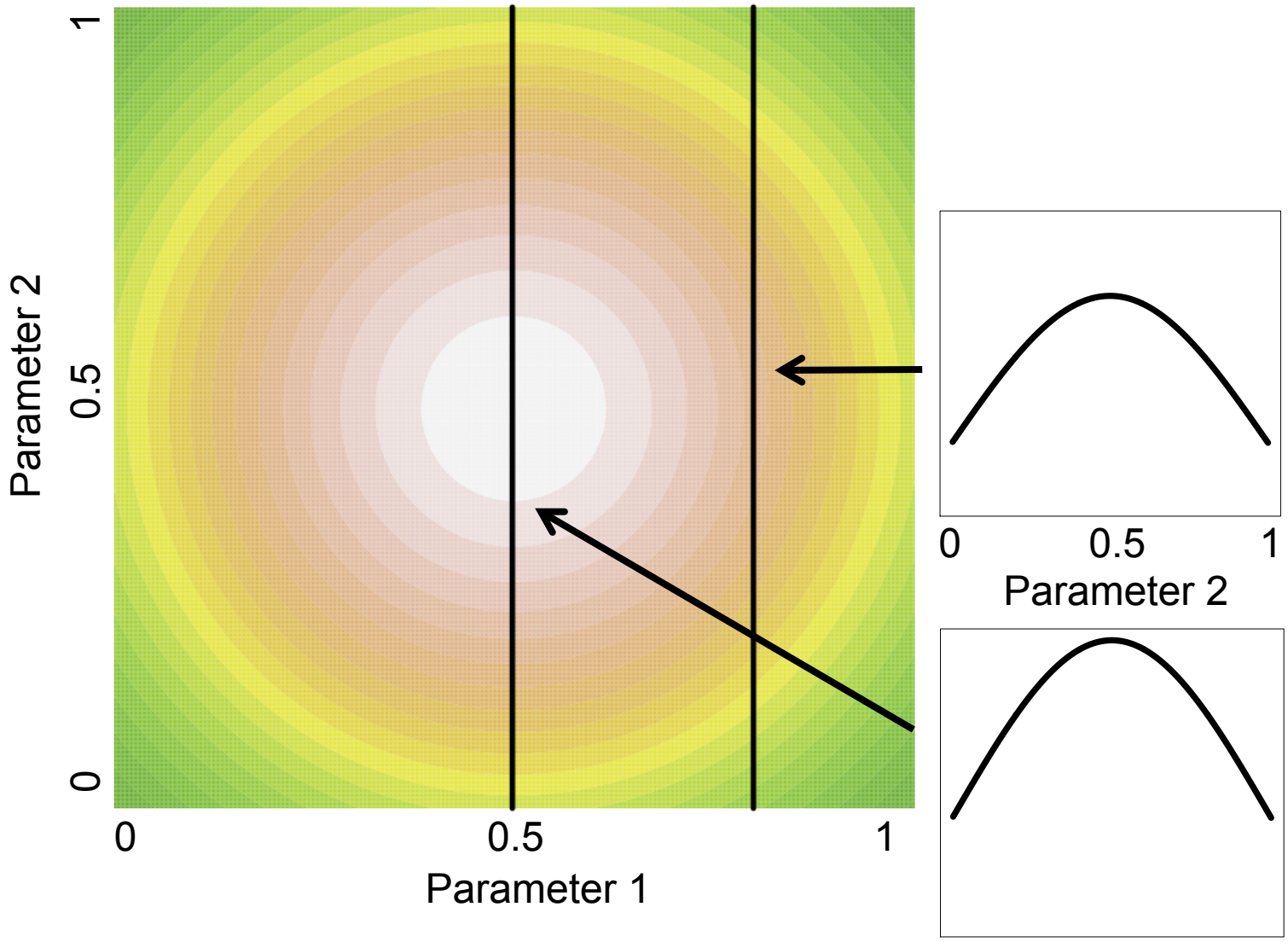
More Parameters = Better ML Score



More Parameters = Better ML Score

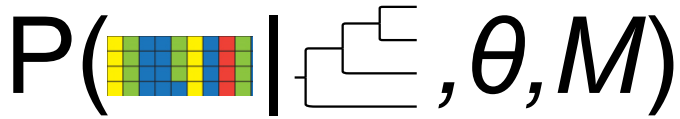


More Parameters = Better ML Score



Maximum Likelihood

Usually implicit



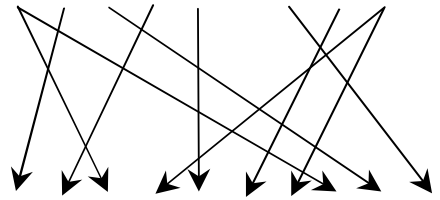
What tree and parameter values give the highest likelihood?

ML scores are just relative, so alone they doesn't tell us how confident we are in this solution, just that this is the preferred solution.

Bootstrapping to Assess Confidence

Original alignment

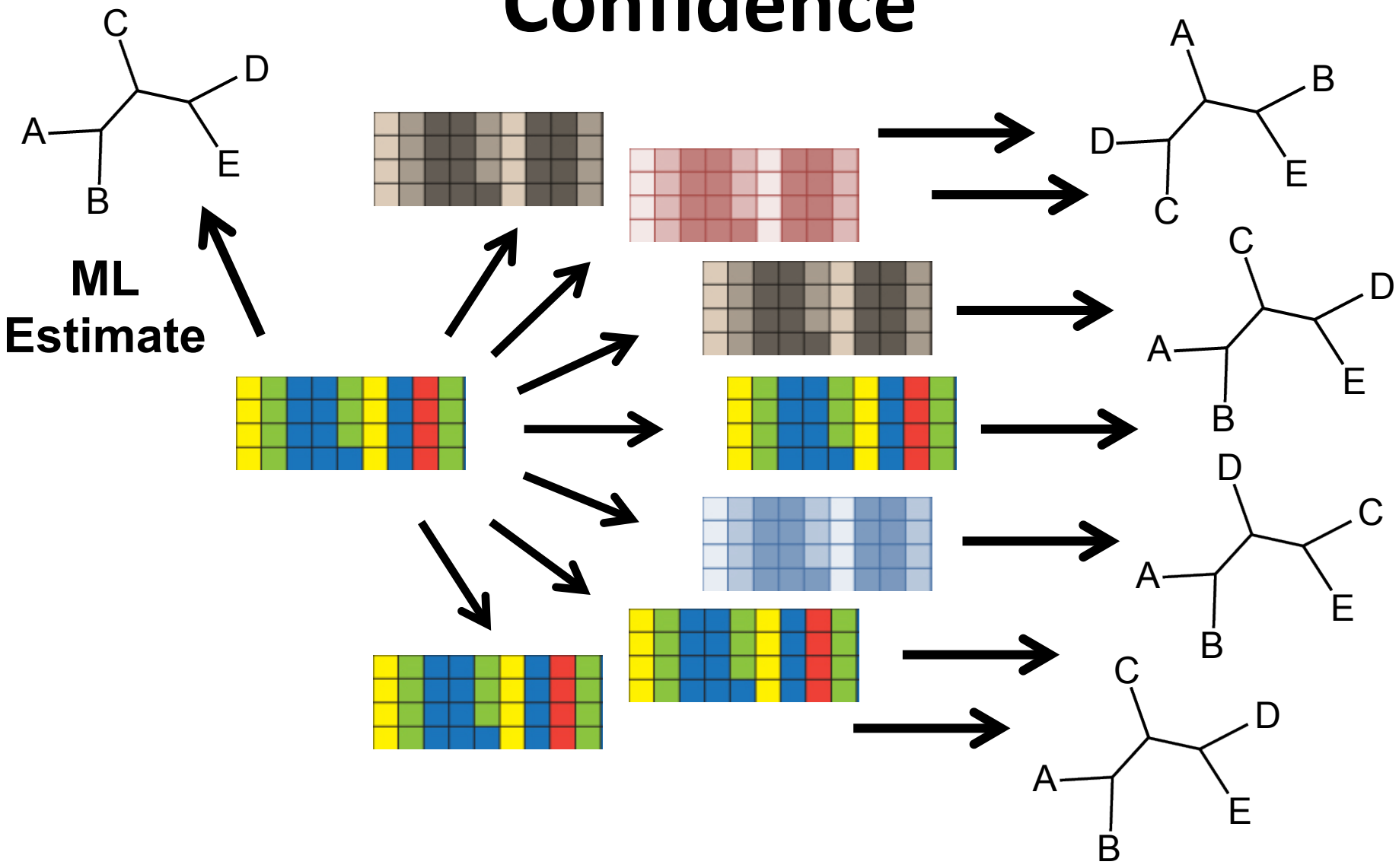
Site	1	2	3	4	5	6	7	8	9	10
human	N	E	N	L	F	A	S	F	I	A
chimpanzee	N	E	N	L	F	A	S	F	A	A
bonobo	N	E	N	L	F	A	S	F	A	A
gorilla	N	E	N	L	F	A	S	F	I	A
orangutan	N	E	D	L	F	T	P	F	T	T
Sumatran	N	E	S	L	F	T	P	F	I	T
gibbon	N	E	N	L	F	T	S	F	A	T



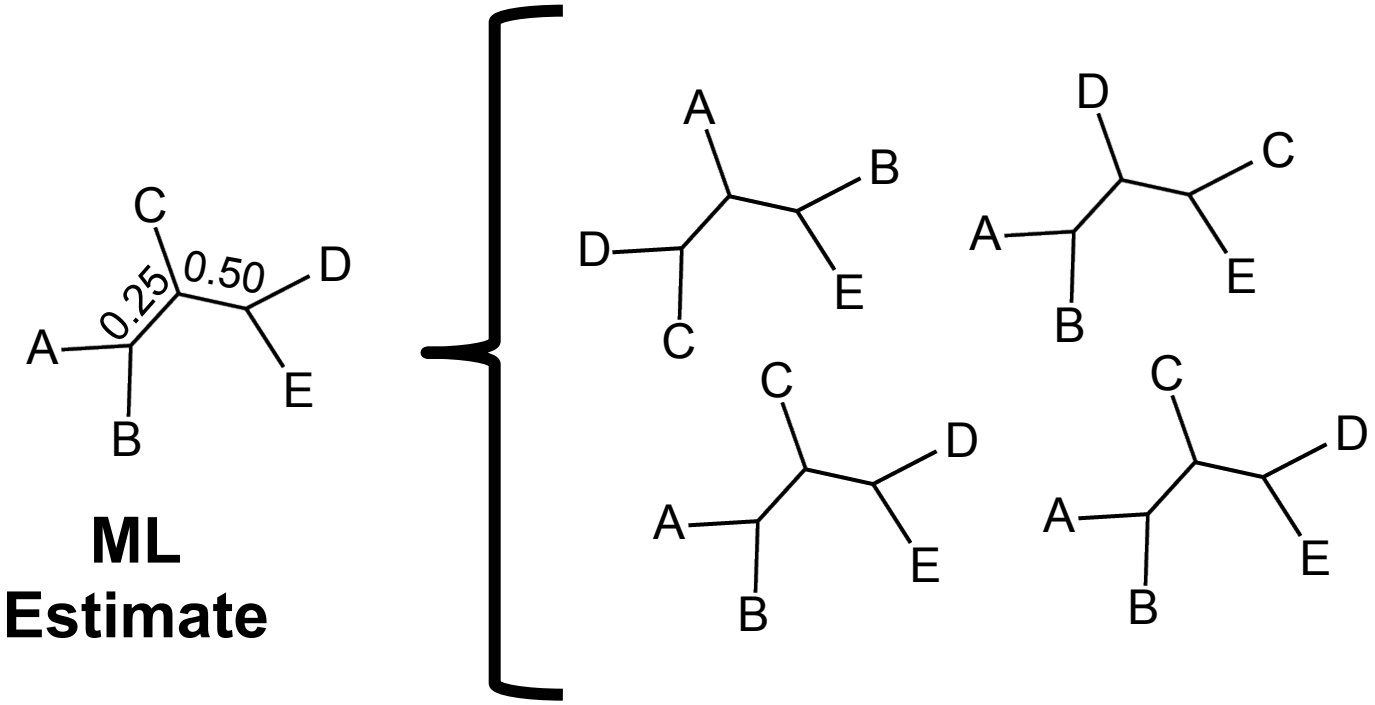
Bootstrap sample

Site	2	4	1	9	5	8	9	1	3	7
human	E	L	N	I	F	F	I	N	N	S
chimpanzee	E	L	N	A	F	F	A	N	N	S
bonobo	E	L	N	A	F	F	A	N	N	S
gorilla	E	L	N	I	F	F	I	N	N	S
orangutan	E	L	N	T	F	F	T	N	D	P
Sumatran	E	L	N	I	F	F	I	N	S	P
gibbon	E	L	N	A	F	F	A	N	N	S

Bootstrapping to Assess Confidence



Bootstrapping to Assess Confidence



Bootstrapping to Assess Confidence

Interpretations of the bootstrap:

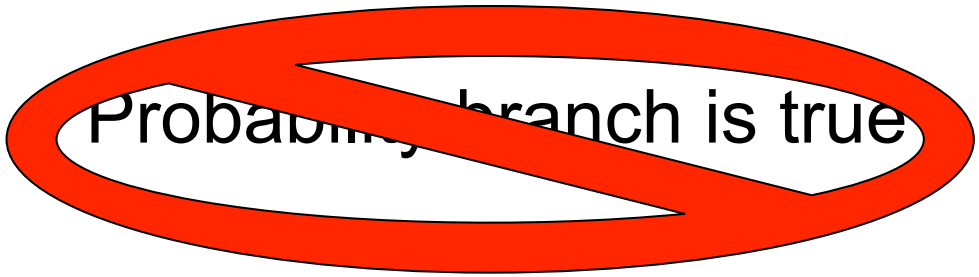
- Repeatability
- 1 – False Positive Rate from a polytomy
- Probability branch is true

Bootstrapping to Assess Confidence

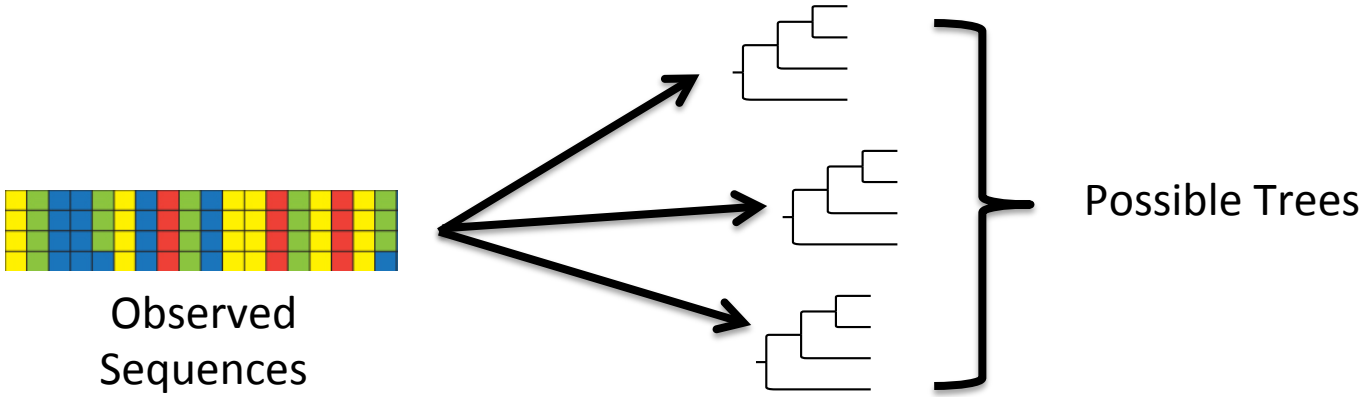
Interpretations of the bootstrap:

- Repeatability

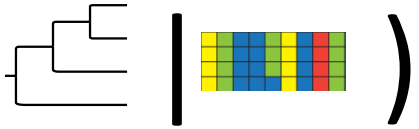
- 1 – False Positive Rate from a polytomy



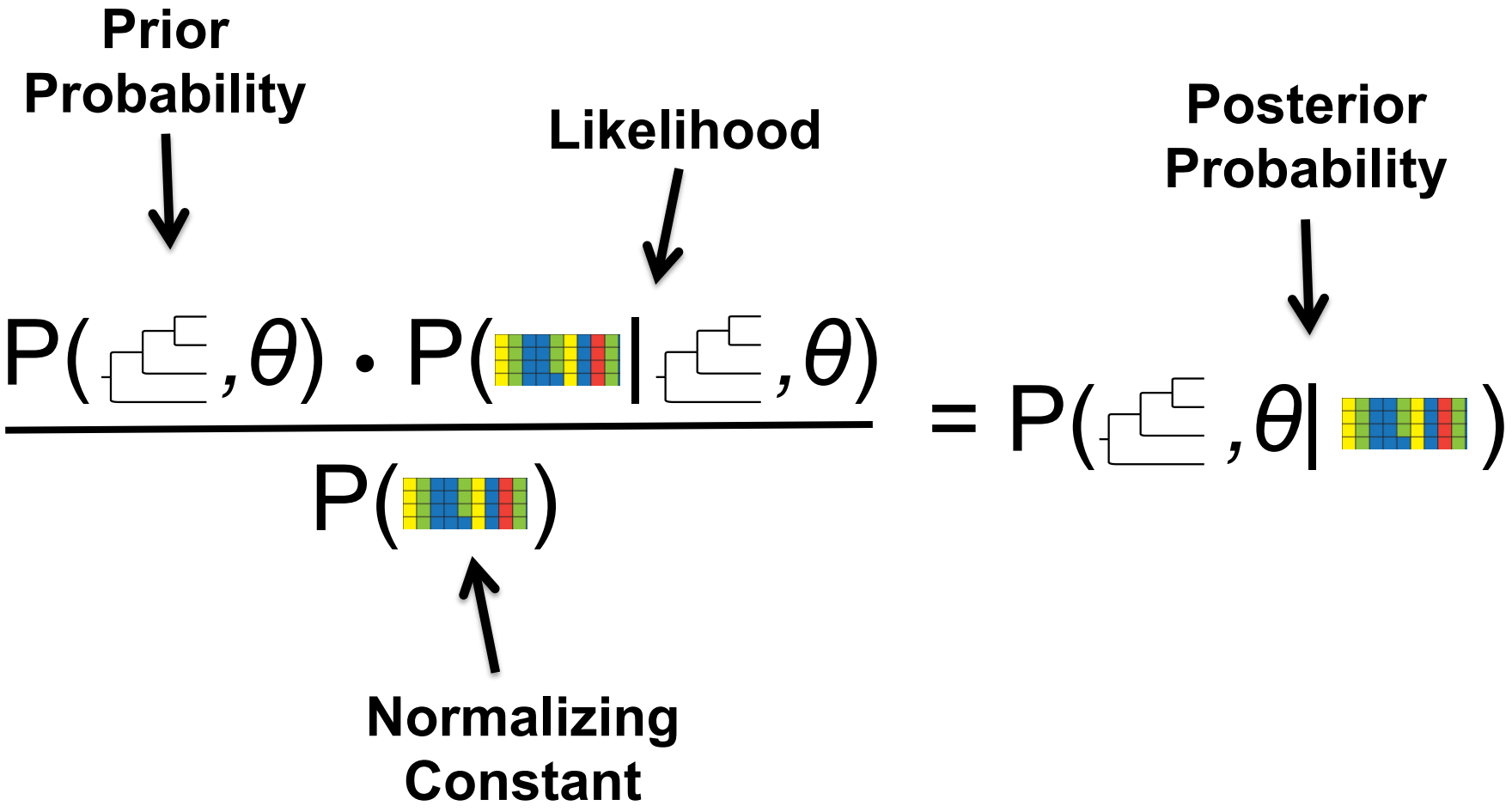
Bayesian Inference



Posterior Probability: Conditional on observed data (alignment), the probability that a particular tree (or part of a tree) is true.

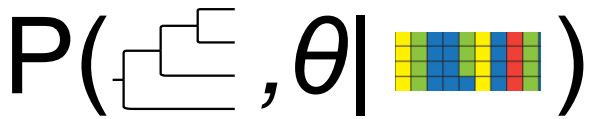


Bayes' Theorem



Marginalizing

Joint Posterior Probability



Marginalizing

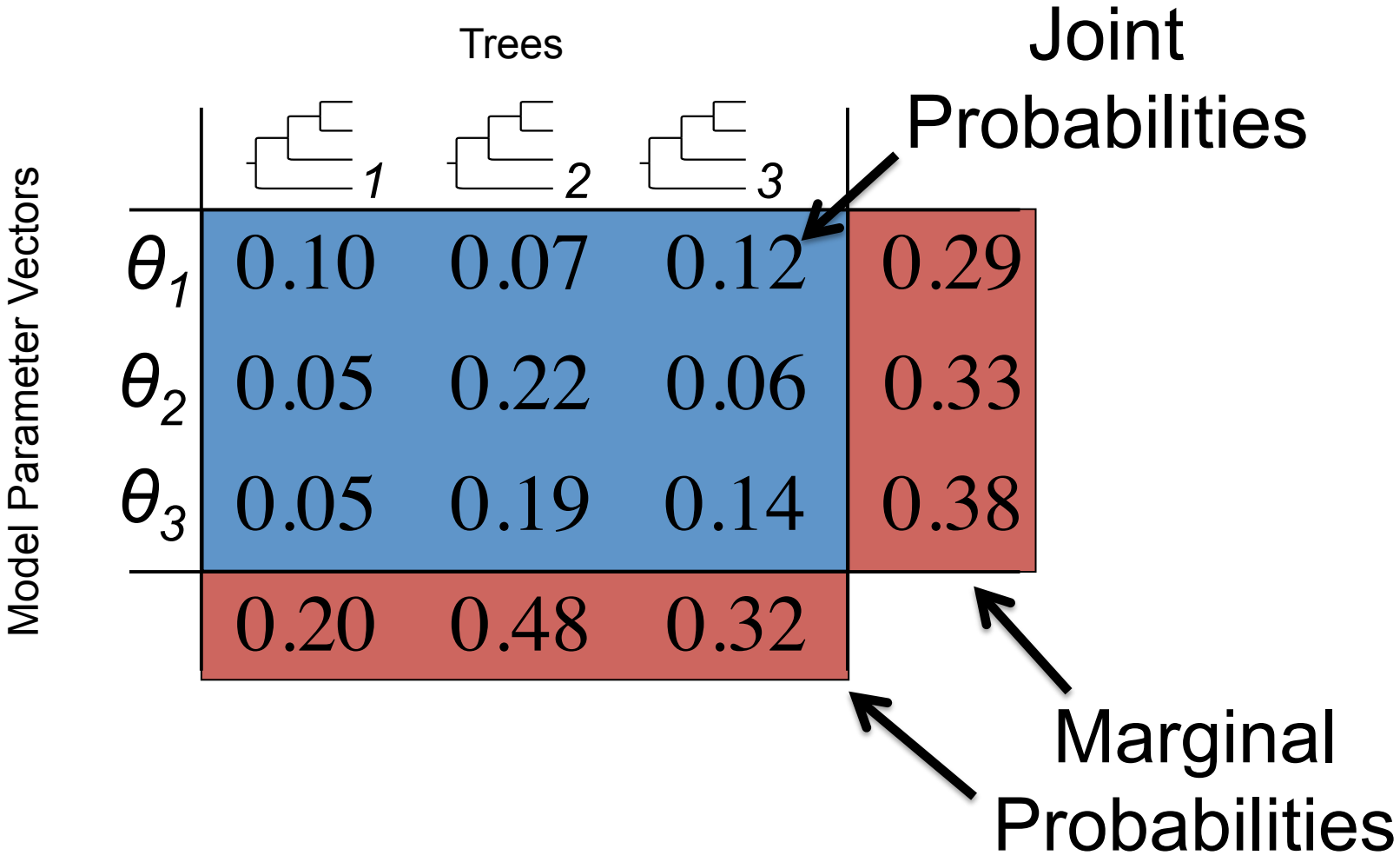
Joint Posterior Probability

Marginal Posterior Probability

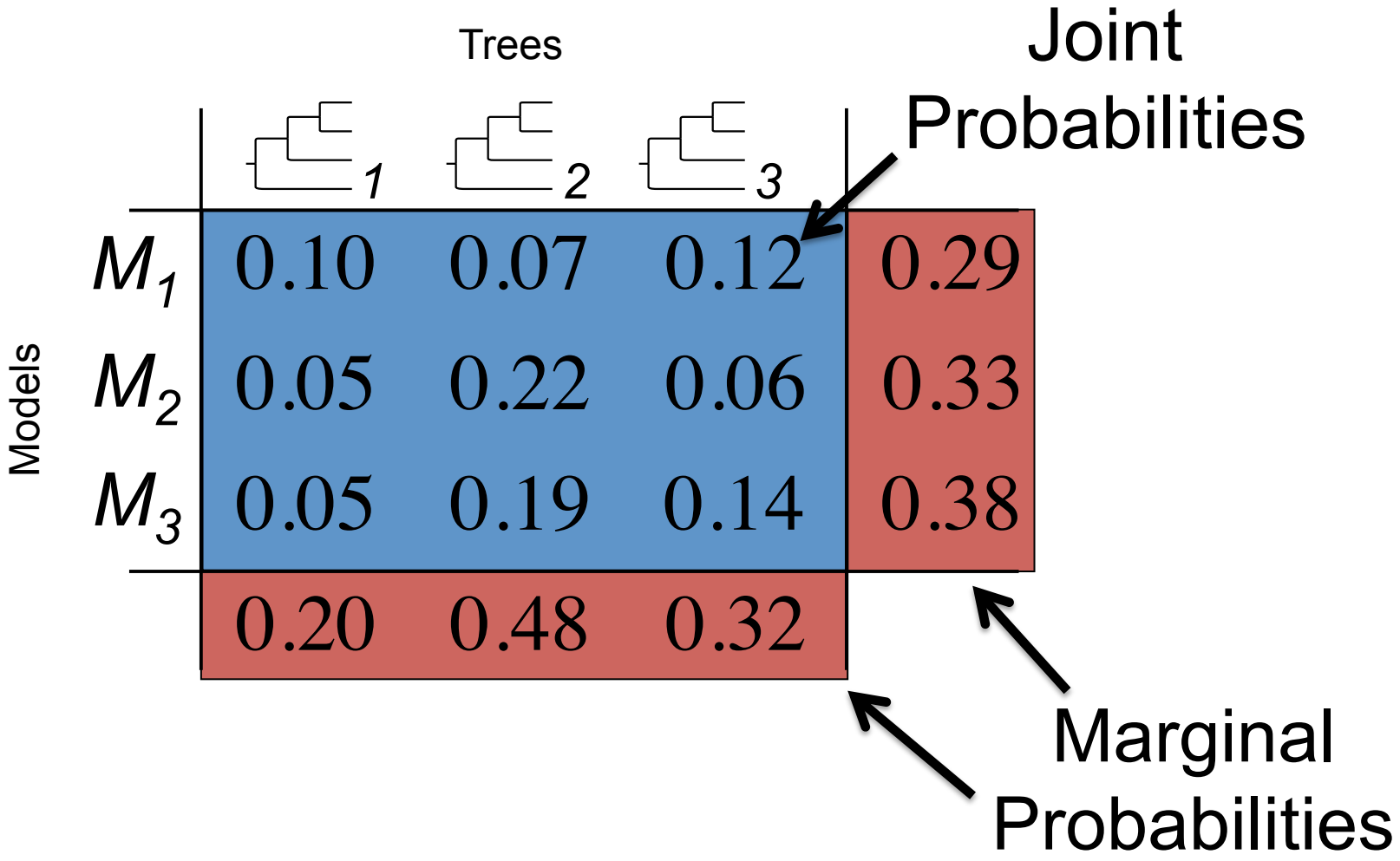
$$\int P(\text{tree}, \theta | \text{data}) \, d\theta = P(\text{tree} | \text{data})$$

Integrating across all values of model parameters

Marginalizing



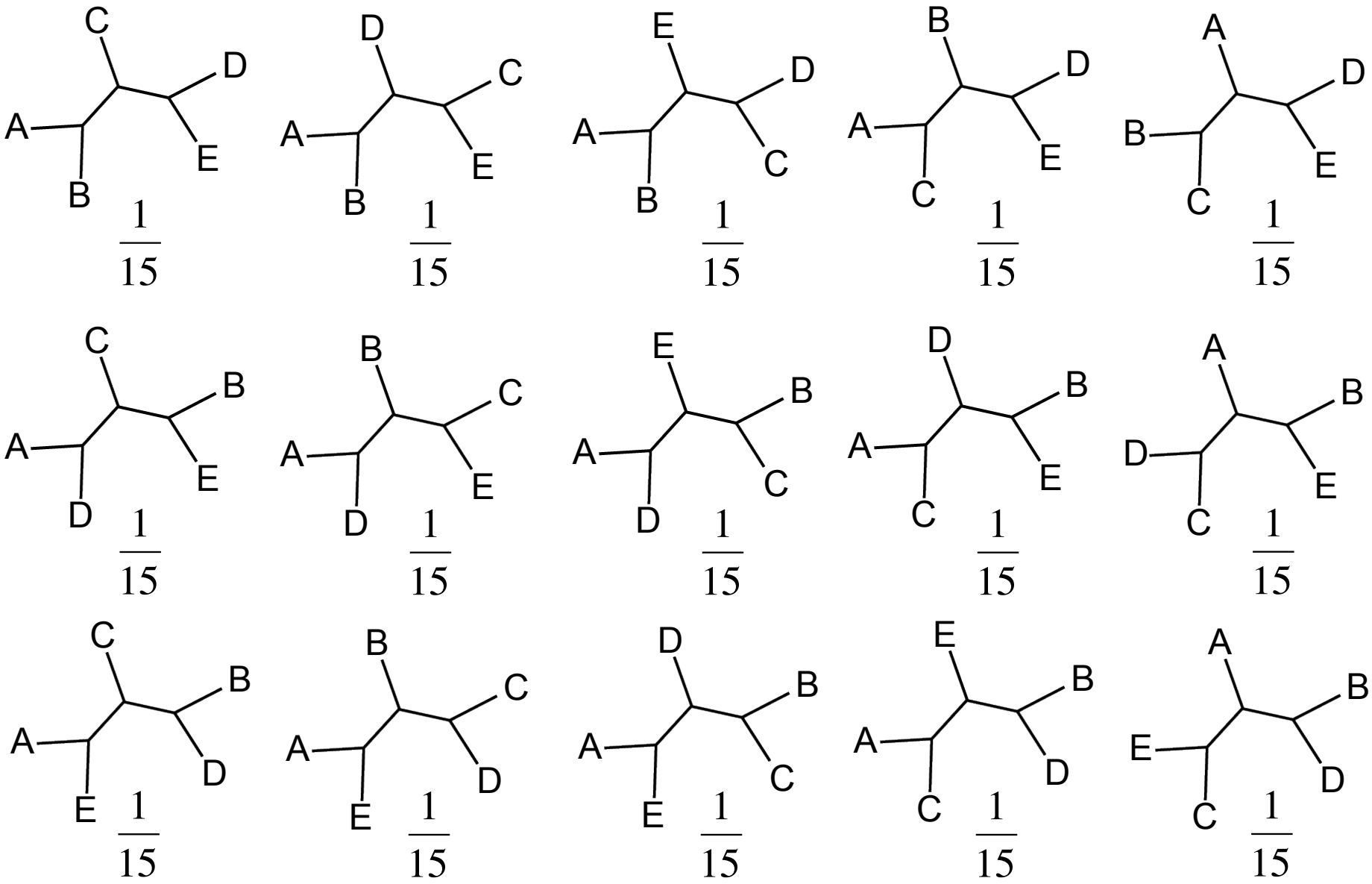
Marginalizing Across Models



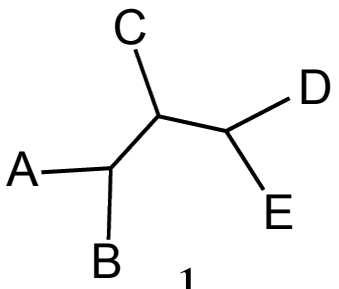
What Priors to Use?

- The controversial part of Bayesian analysis
- Choice can vary by researcher
- Often chosen in an attempt to reflect prior ignorance
- Analysis can be run under several priors to assess sensitivity

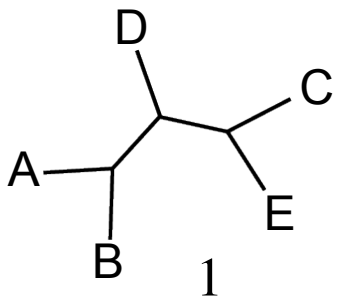
Uniform Topology Prior



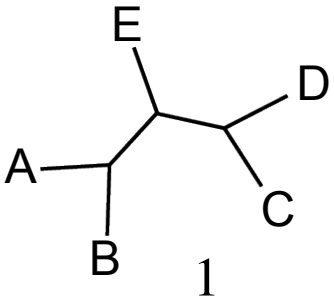
(AB|CDE) Constraint Prior



$\frac{1}{3}$



$\frac{1}{3}$

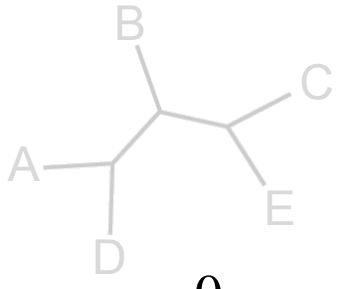


$\frac{1}{3}$

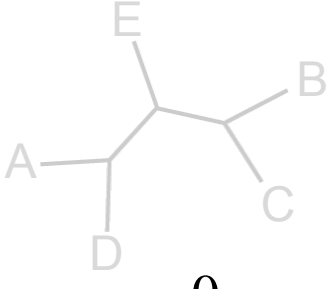
0

0

0

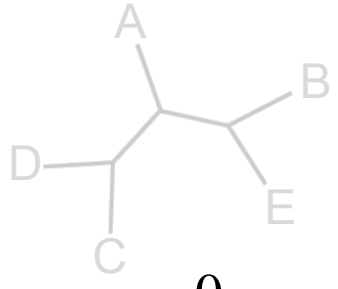


0

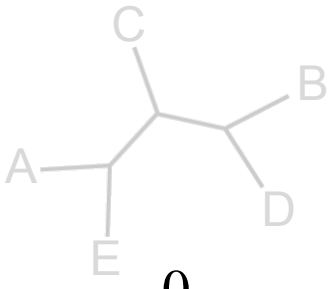


0

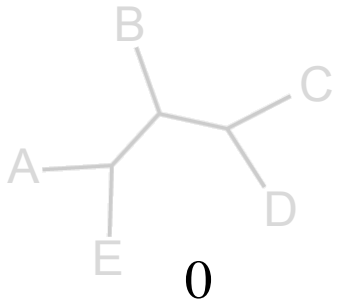
0



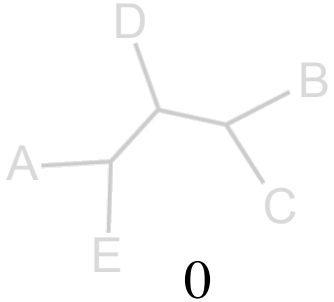
0



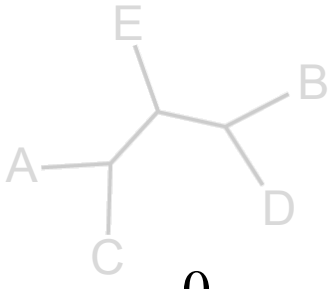
0



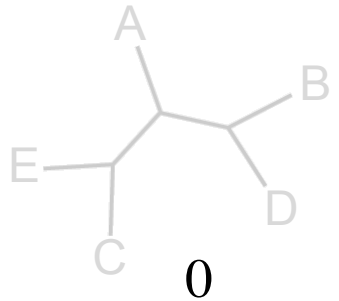
0



0

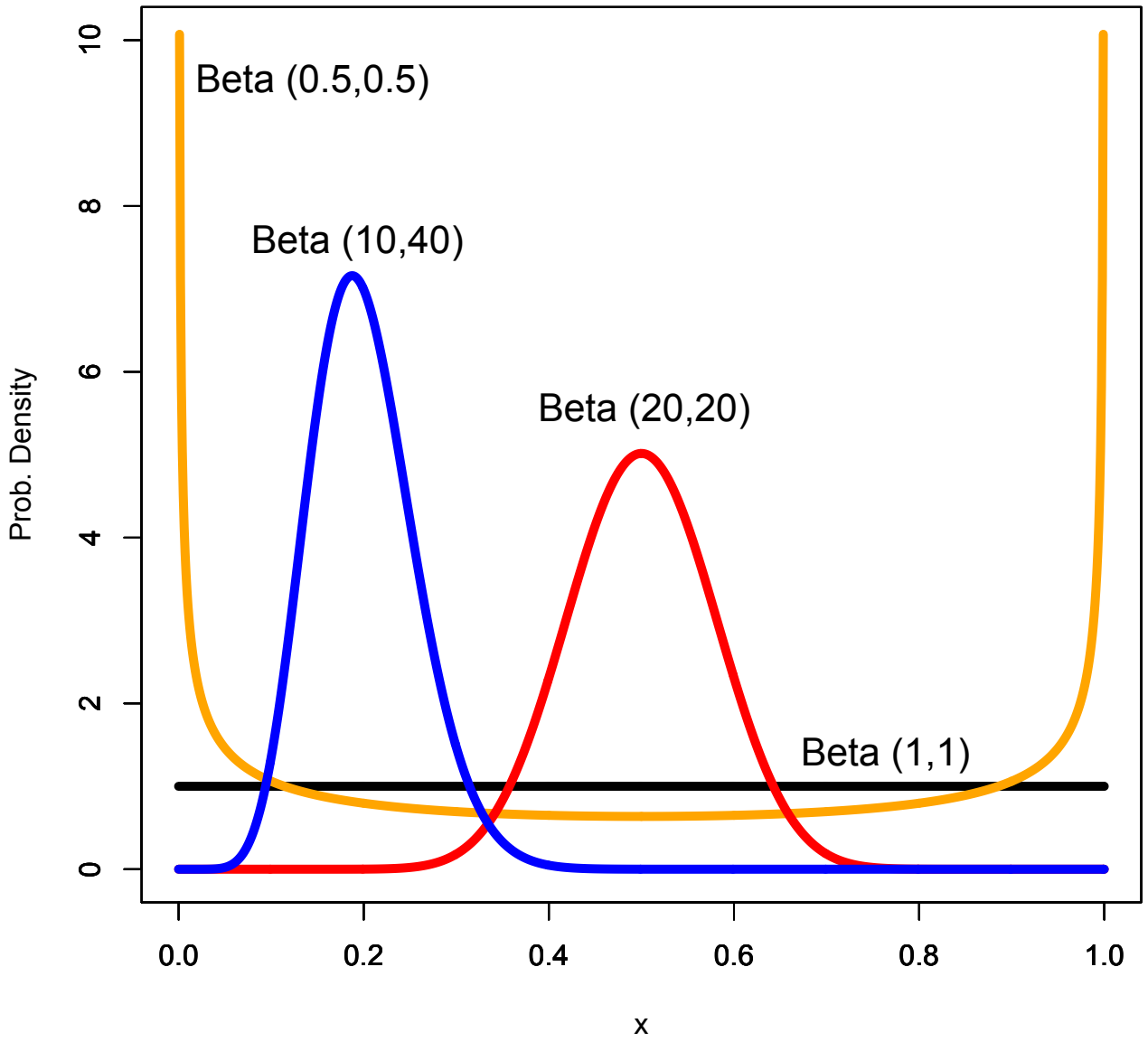


0

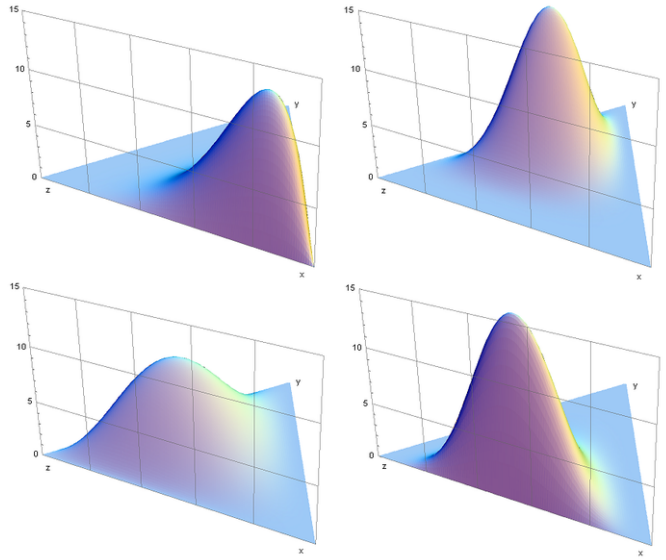


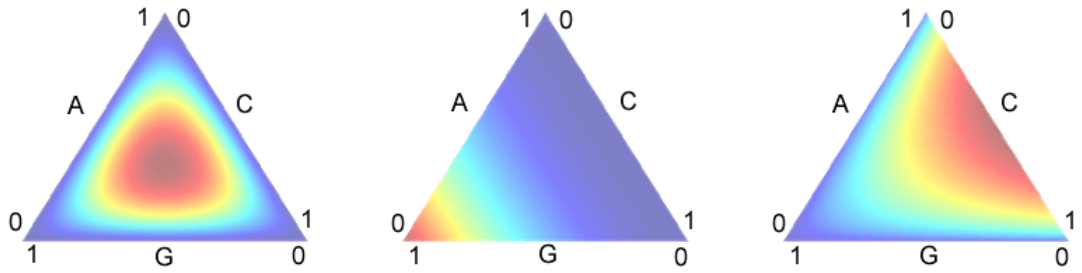
0

Beta Distributions

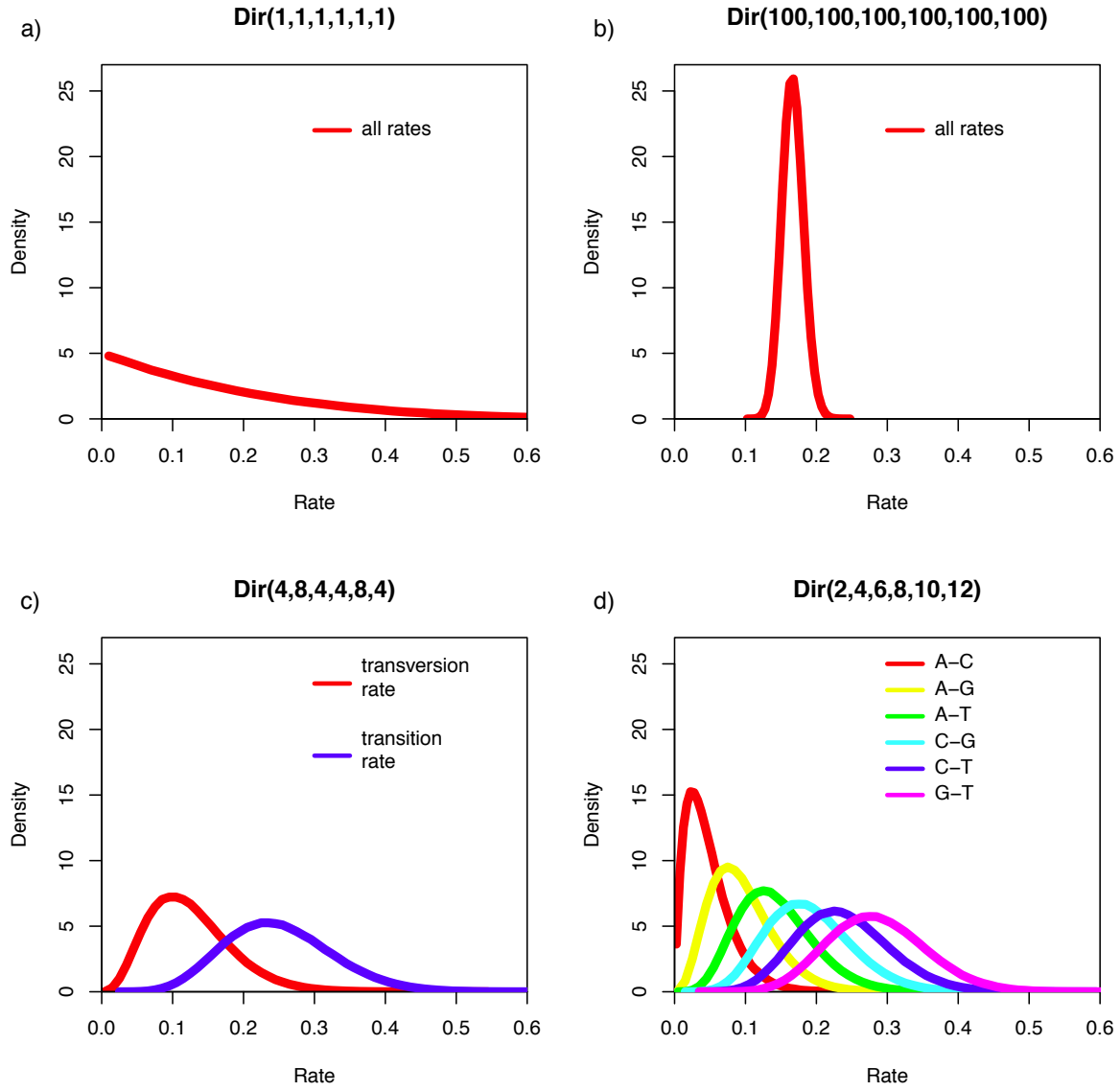


Prior on Sets - Dirichlet Distribution



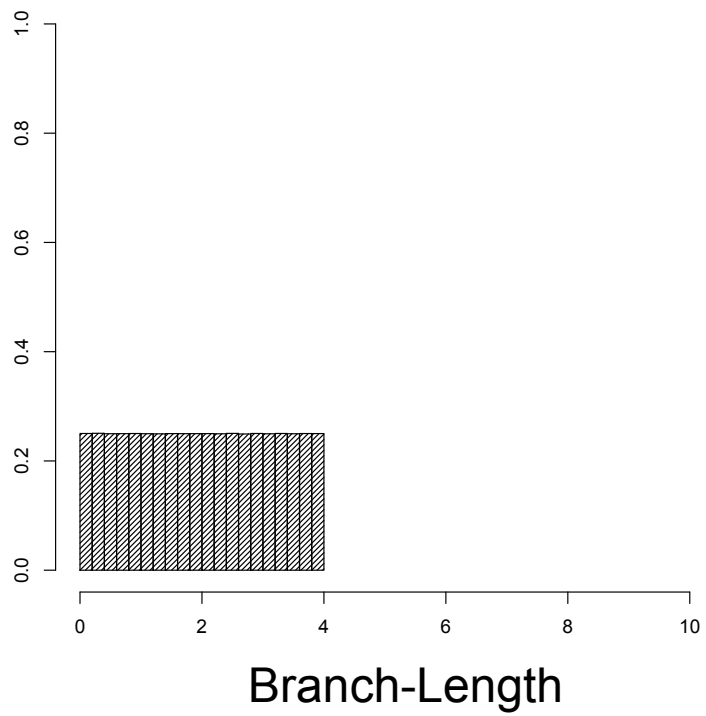


Prior on Sets - Dirichlet Distribution

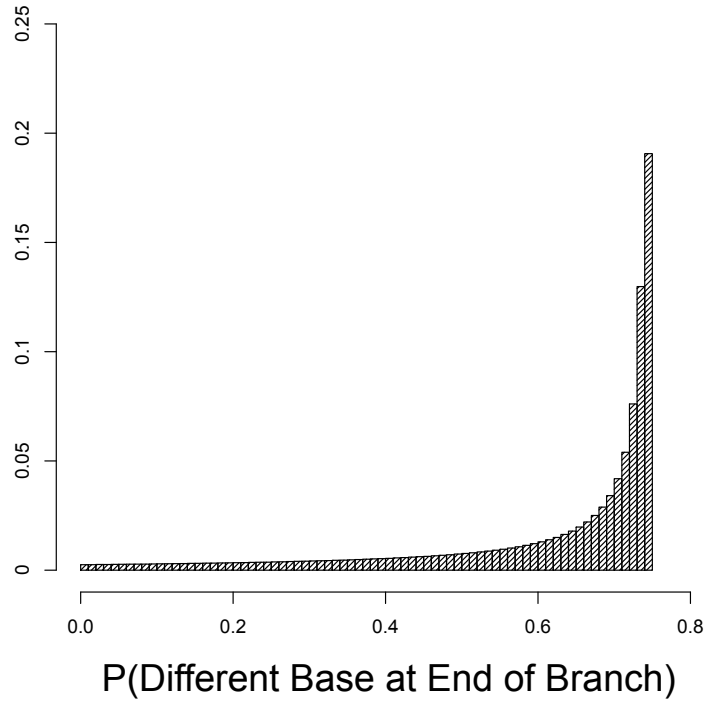


Branch-Length Priors

Uniform(0,4) Branch-Length Prior



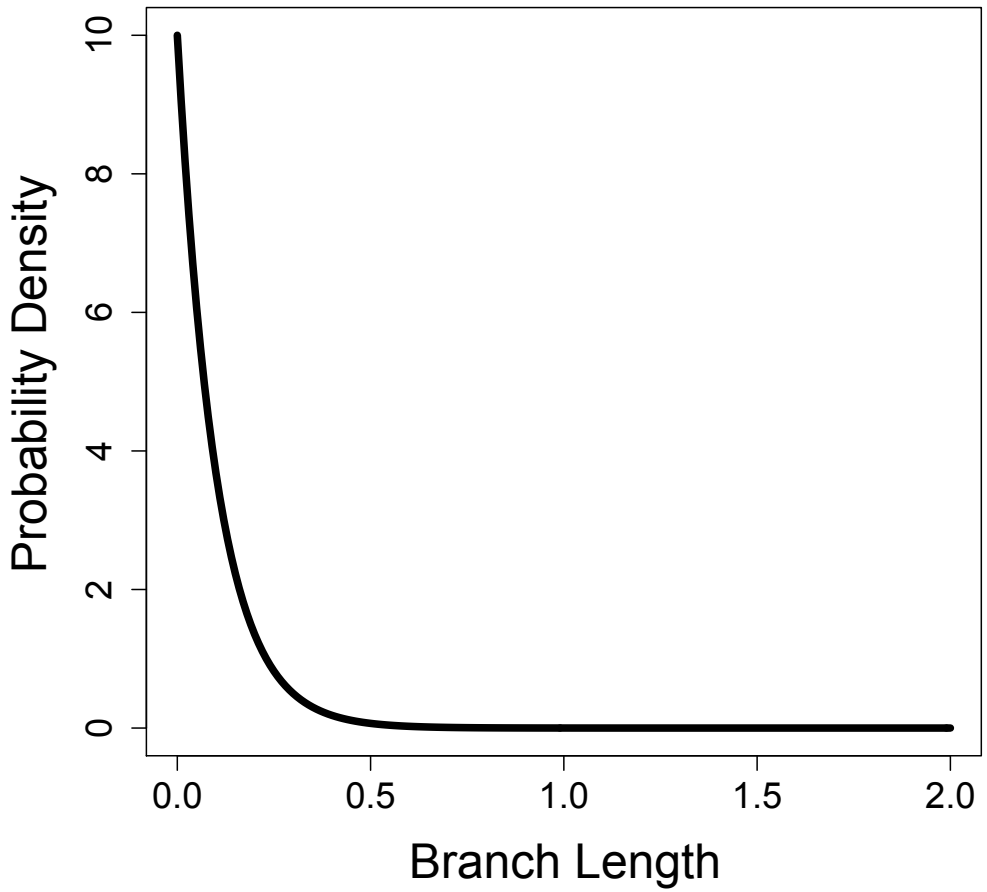
Strongly Informative Prior on Change



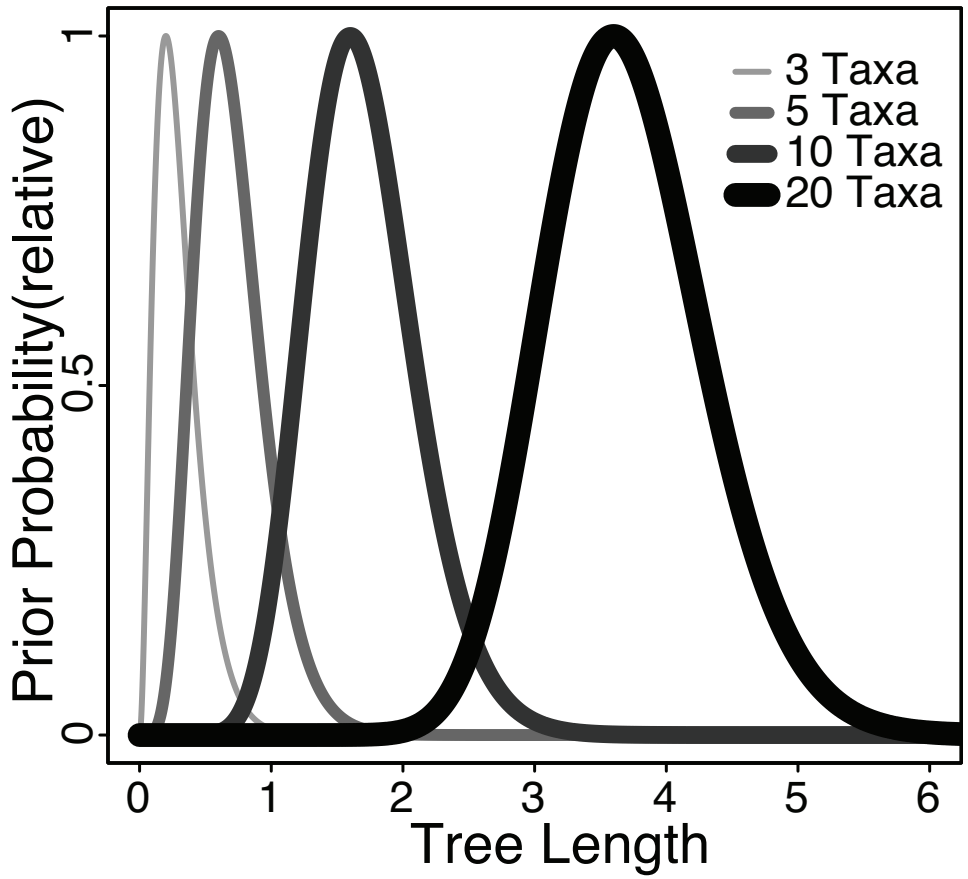
Idea from P.O. Lewis

Branch-Length Priors

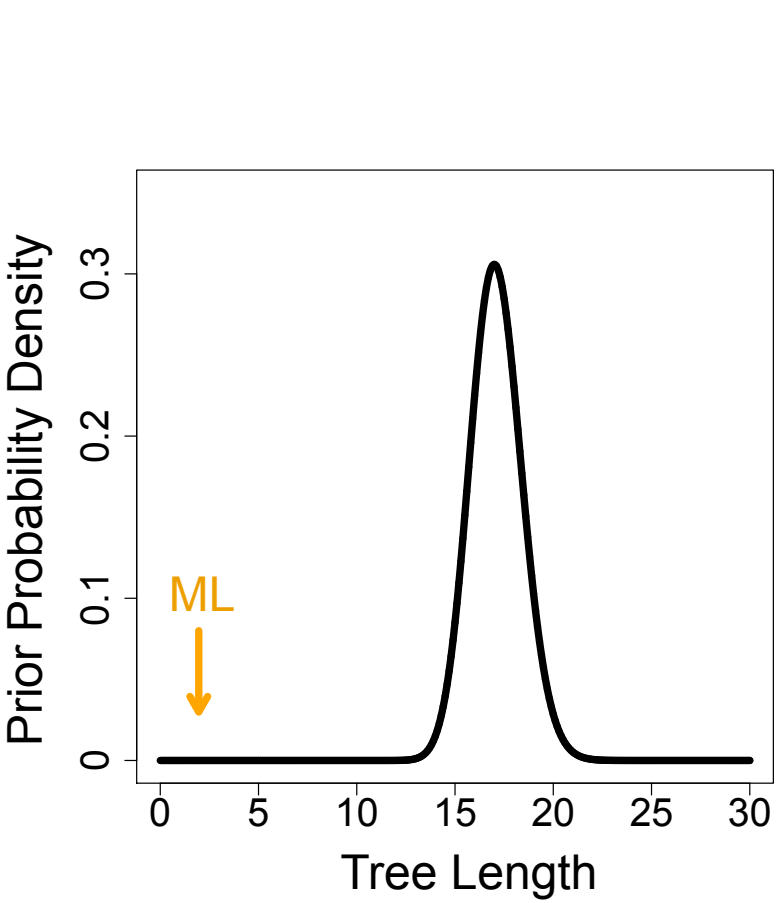
Default Exponential Branch-Length Prior (λ=10, mean=0.1)

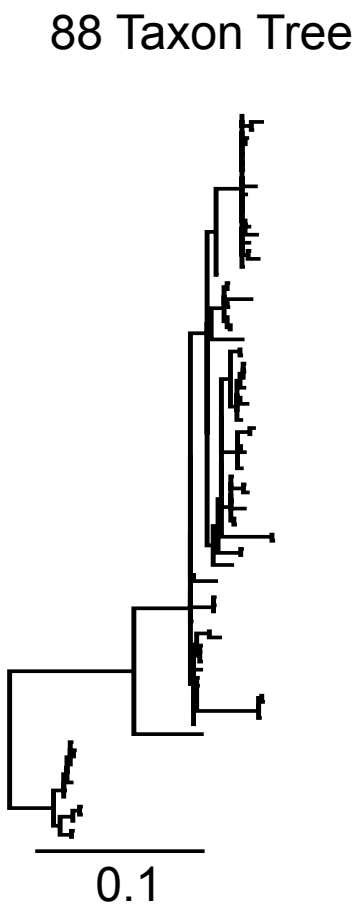


Implied Tree-Length Prior

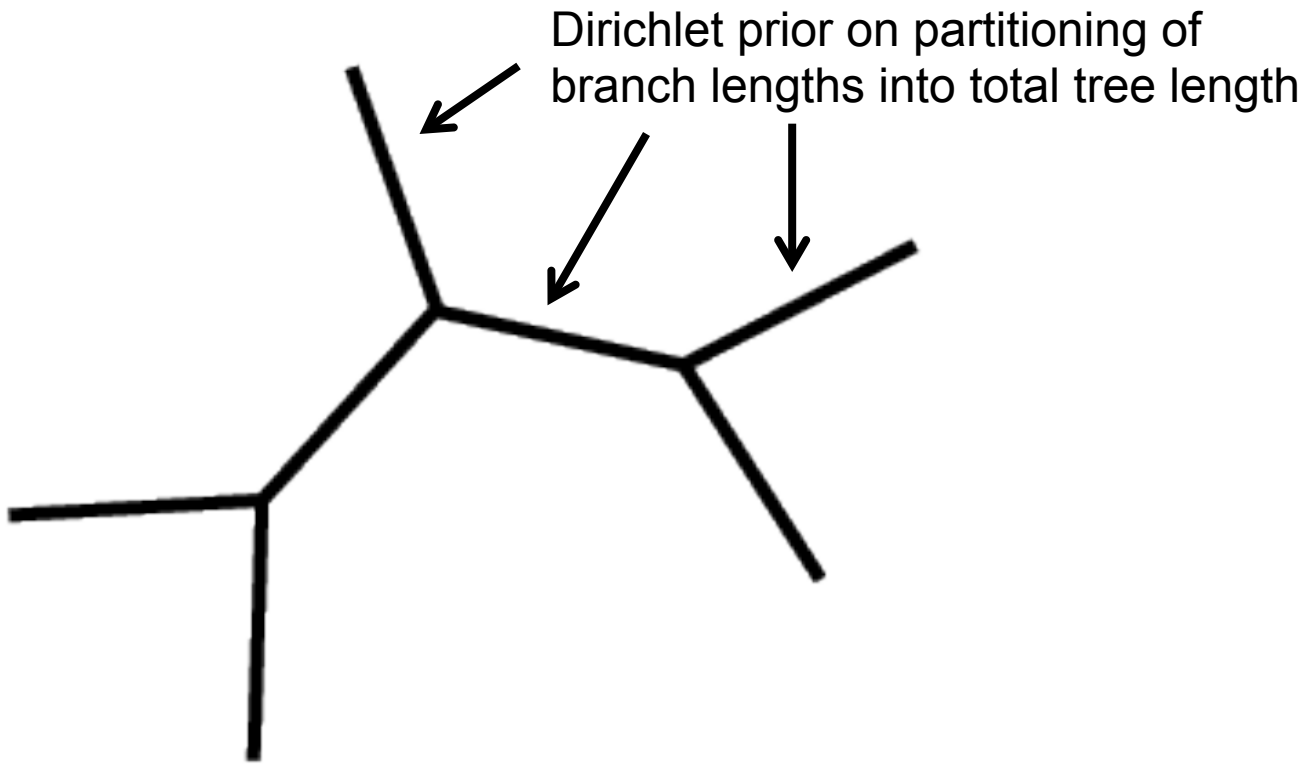


Implied Tree-Length Prior





Compound Branch-Length Prior



Total Tree Length Prior: (Inv) Gamma

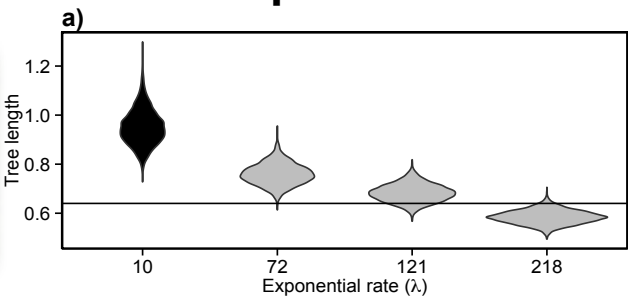
Rannala et al. 2011. Mol. Biol. Evol.
Zhang et al. 2012. Syst. Biol

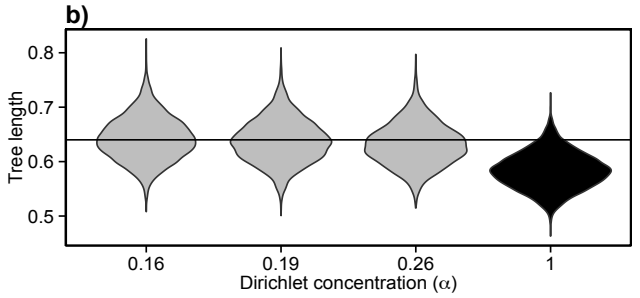
Outside Info Improves Estimates

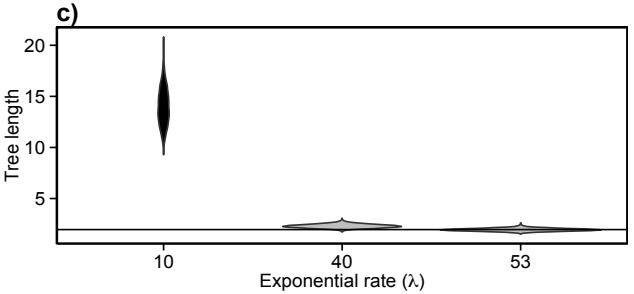
Dataset	Clams	Frogs
ML TL Estimate	1.96	0.55
Bayes TL Interval (MB Default)	10.7 - 17.7	1.77-3.29
Bayes TL Interval (Informed)	1.15 – 1.43	0.32 – 0.38
Bayes TL Interval (Compound Prior)	0.9 – 1.3	0.44 – 0.57

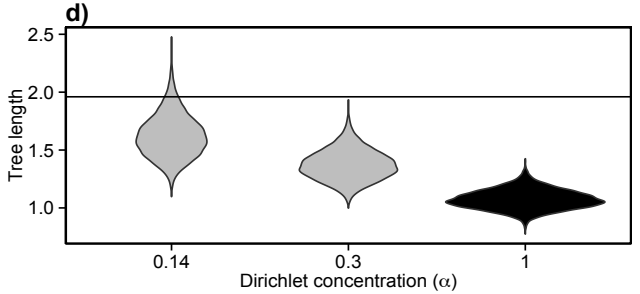
Exponential

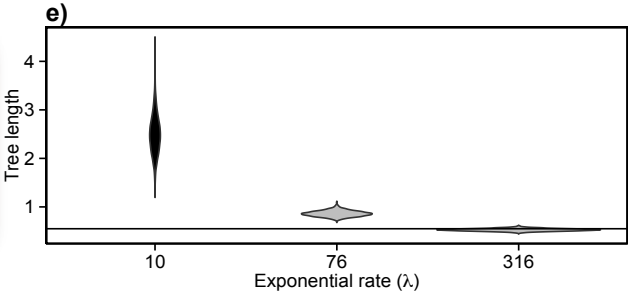
Compound Dirichlet

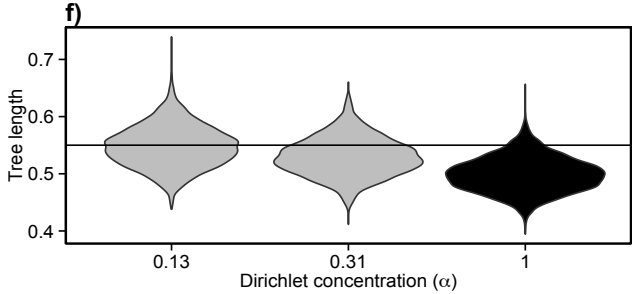




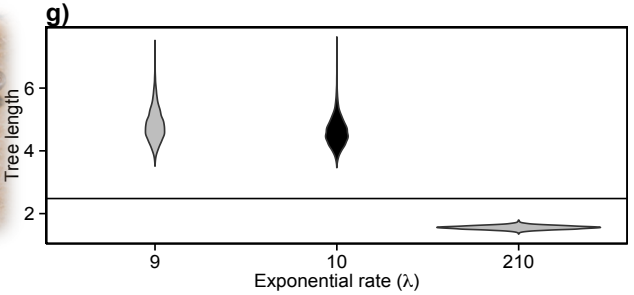


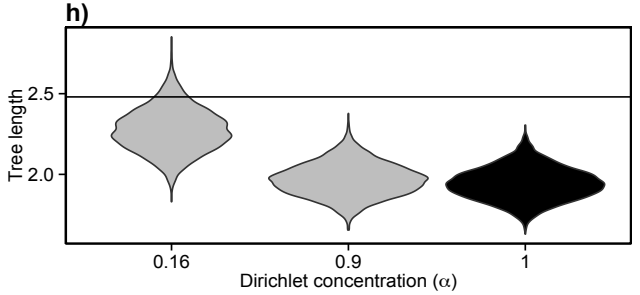






Brad Nelson





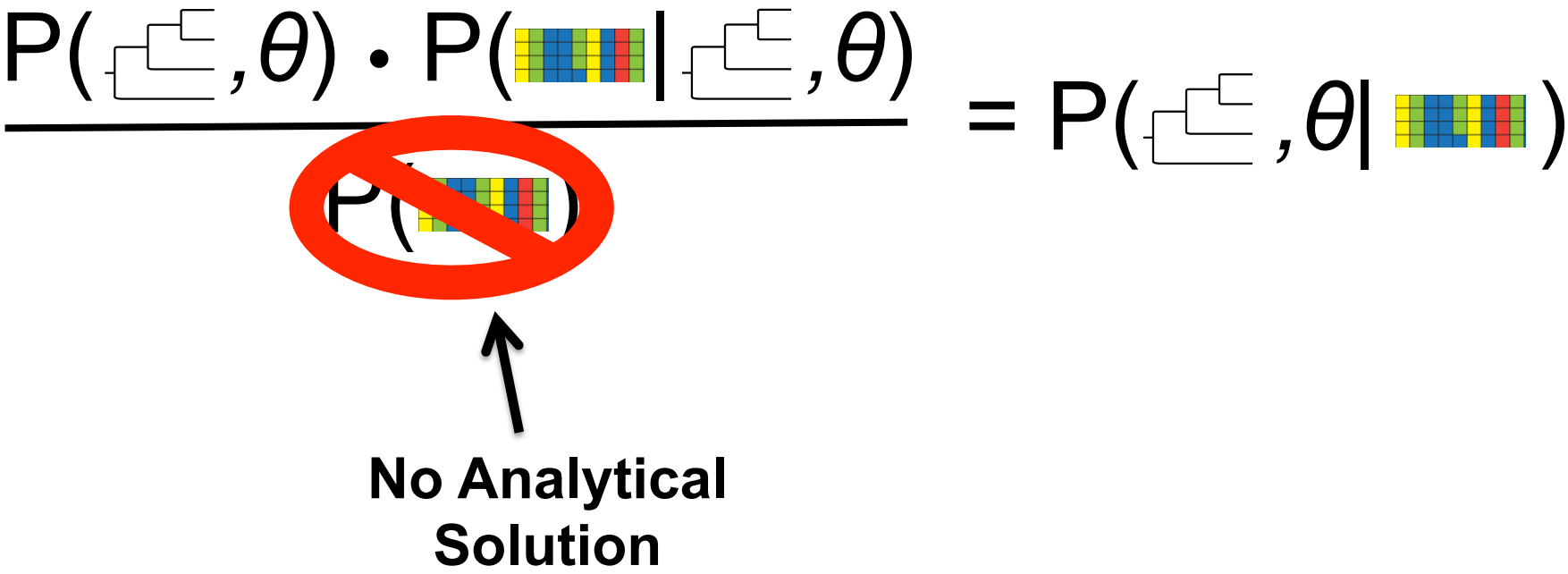
EmpPrior

https://code.google.com/p/empprior/

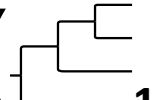
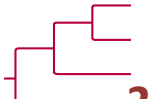
Searches TreeBase to find other phylogenetic studies that have used the same gene at roughly the same taxonomic depth, infers an ML tree, and fits the parameters of a branch-length distribution using these related data sets.

Written by John Andersen and Brad Nelson

Bayes' Theorem



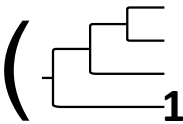
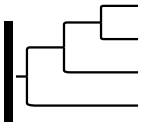
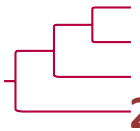
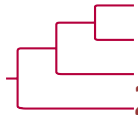
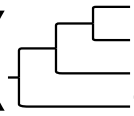
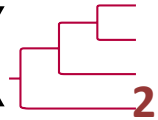
Posterior Odds Ratio

$$\frac{P(\text{  }_1, \theta_1 | \text{  })}{P(\text{  }_2, \theta_2 | \text{  })}$$

Posterior Odds Ratio

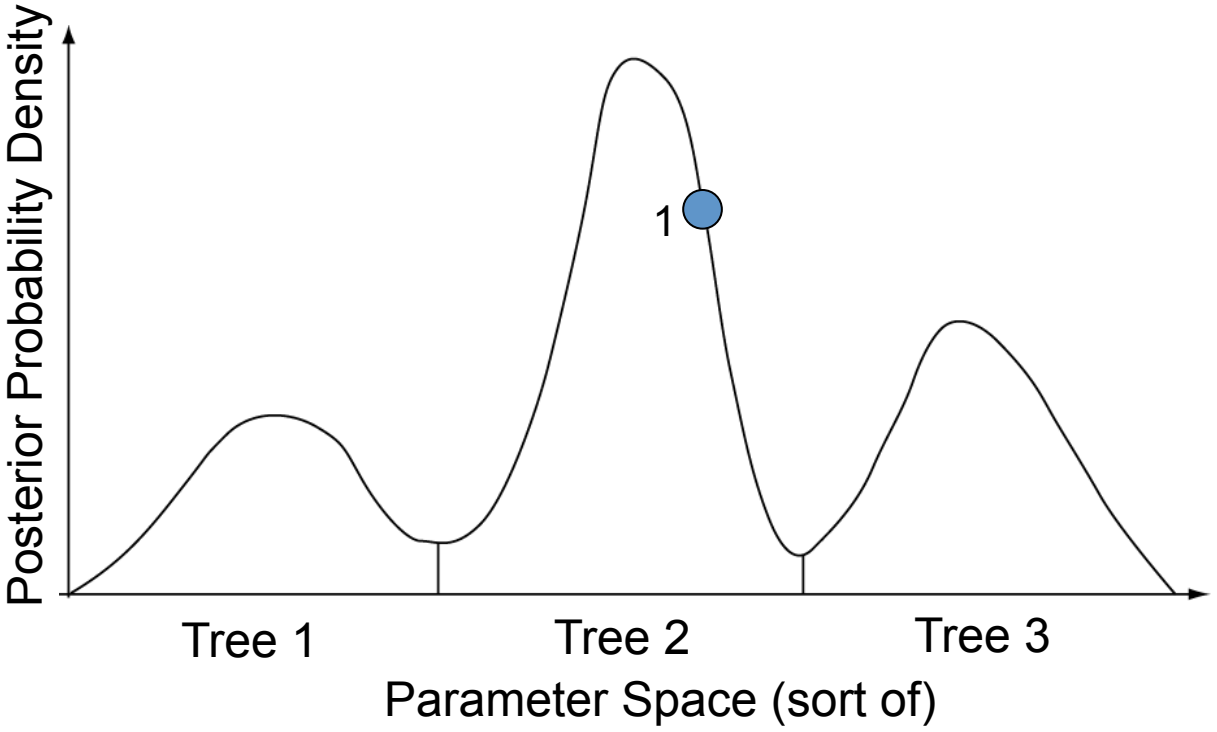
$$\frac{P(\text{tree}_1, \theta_1) \cdot P(\text{data} \mid \text{tree}_1, \theta_1)}{\cancel{P(\text{data})}} \Bigg/ \frac{P(\text{tree}_2, \theta_2) \cdot P(\text{data} \mid \text{tree}_2, \theta_2)}{\cancel{P(\text{data})}} = \frac{P(\text{tree}_1, \theta_1 \mid \text{data})}{P(\text{tree}_2, \theta_2 \mid \text{data})}$$

Posterior Odds Ratio

$$\frac{P(\text{  }_1, \theta_1) \cdot P(\text{  } \mid \text{  }_1, \theta_1)}{P(\text{  }_2, \theta_2) \cdot P(\text{  } \mid \text{  }_2, \theta_2)} = \frac{P(\text{  }_1, \theta_1 \mid \text{  })}{P(\text{  }_2, \theta_2 \mid \text{  })}$$

Markov chain Monte Carlo

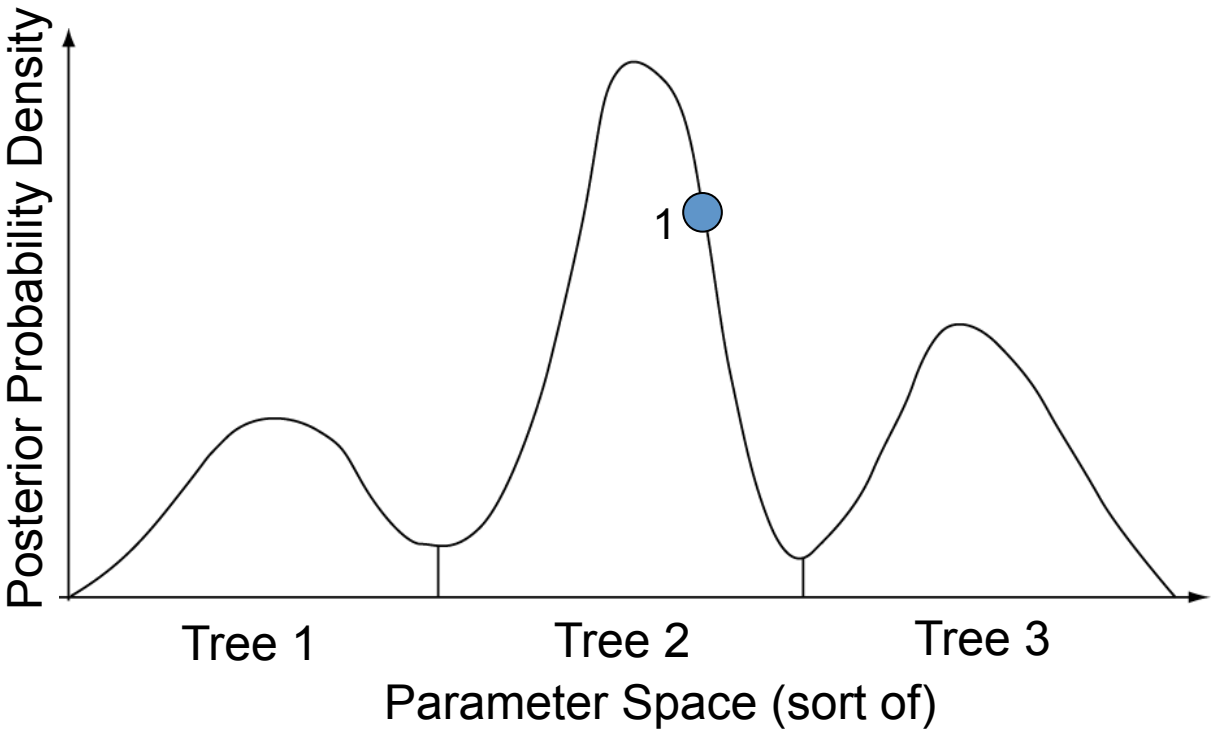
1. Start at an arbitrary point



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

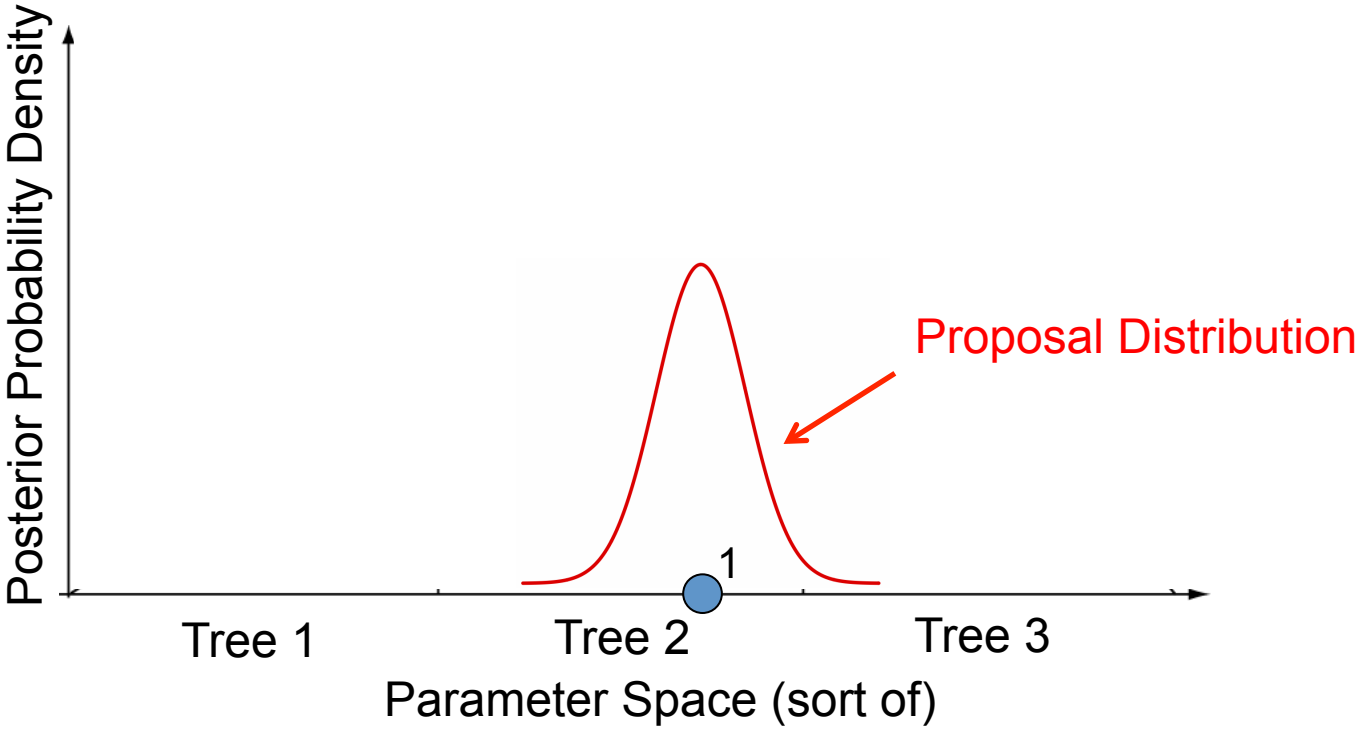
1. Start at an arbitrary point
2. Make a small random move



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

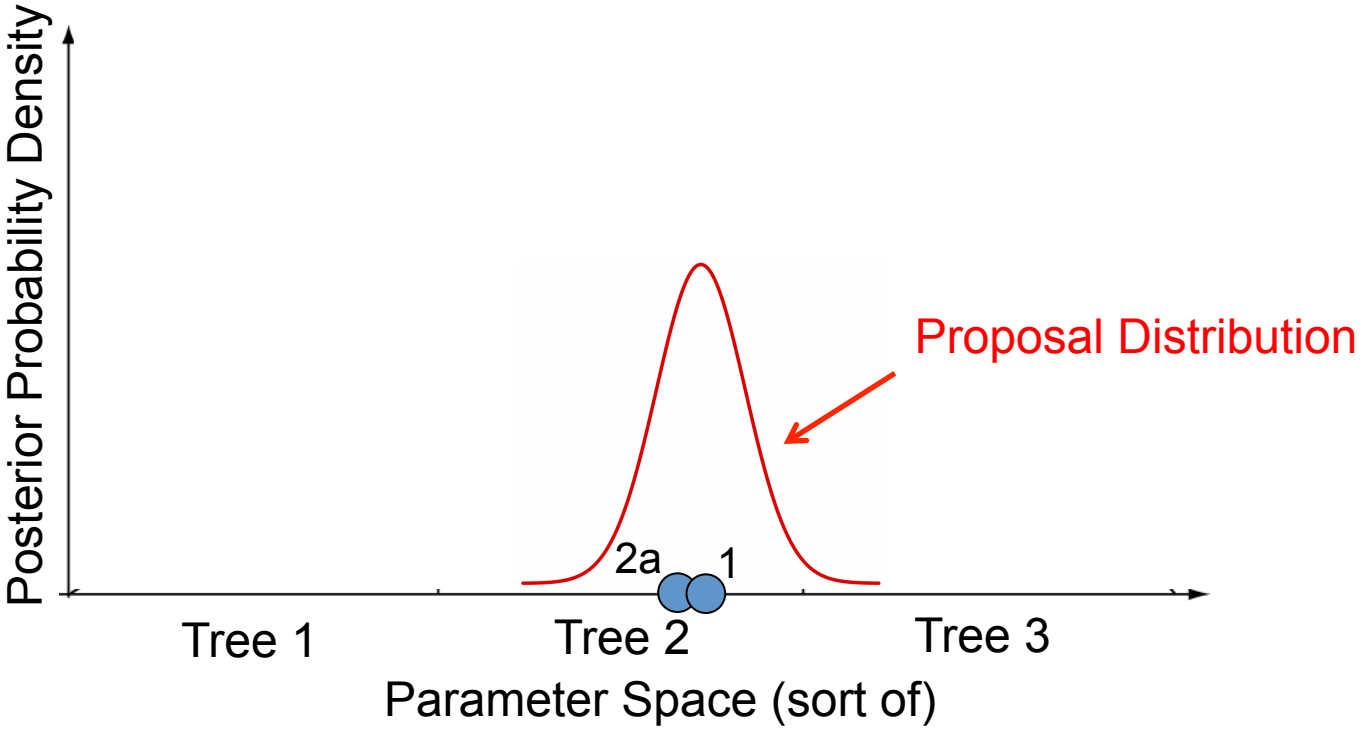
1. Start at an arbitrary point
2. Make a small random move



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

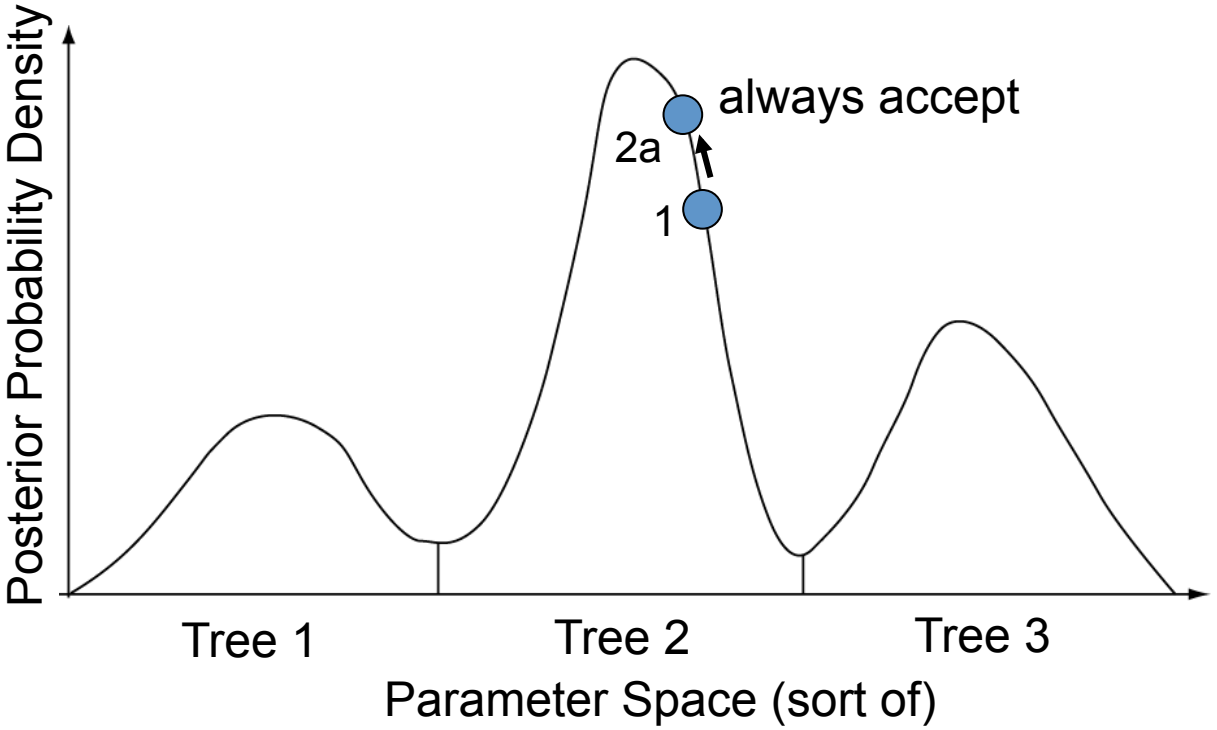
1. Start at an arbitrary point
2. Make a small random move



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

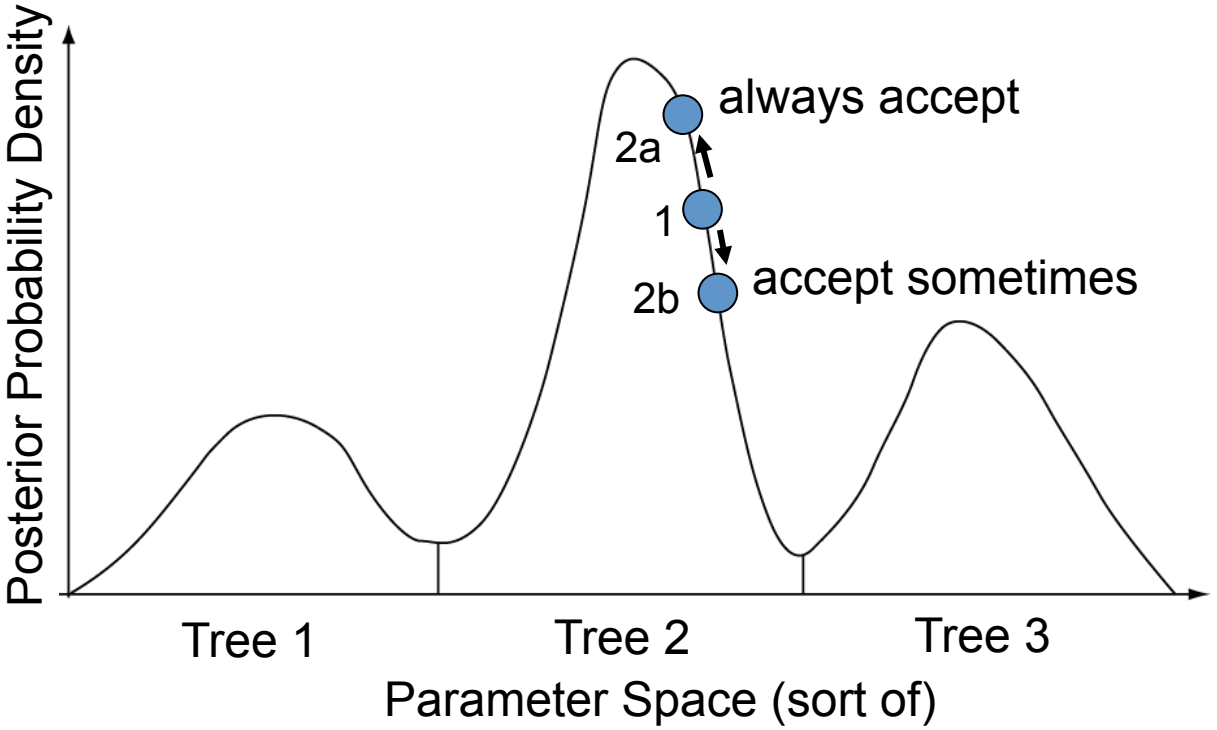
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (*r*) of new state to old state:
   a) r > 1 -> new state accepted
   b) r < 1 -> new state accepted with probability *r*. If new state not accepted, stay in the old state



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

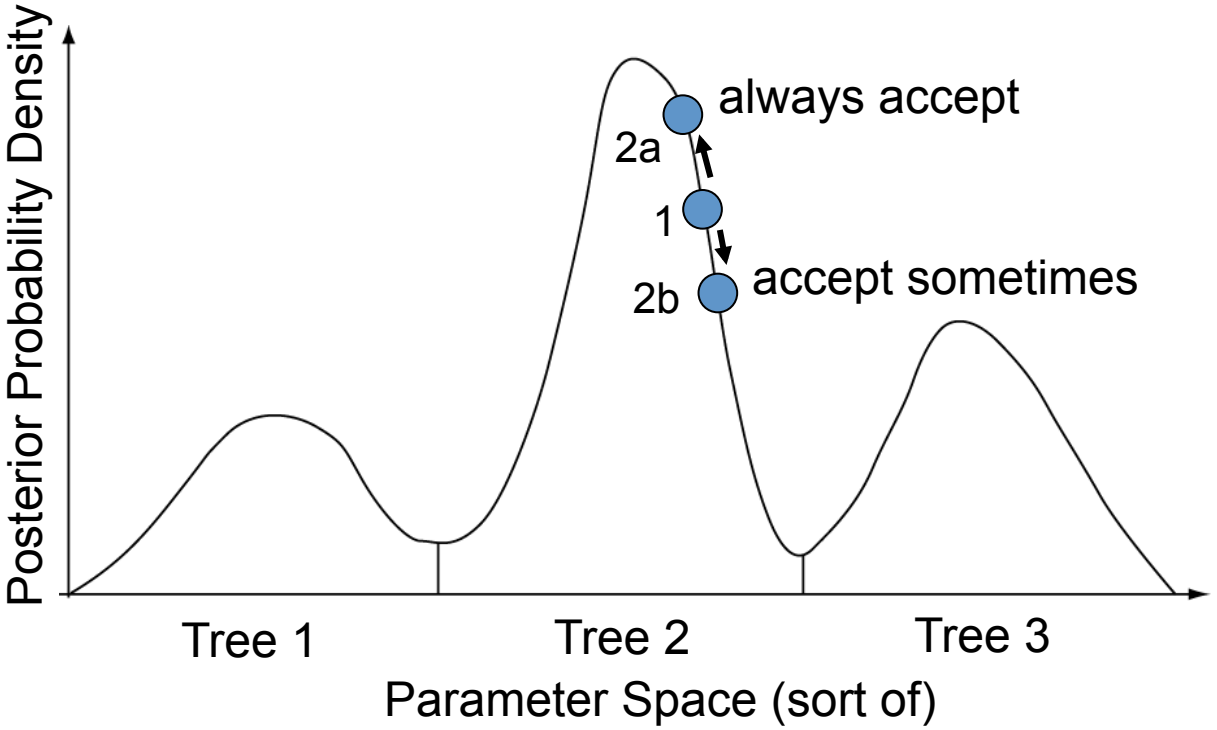
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (*r*) of new state to old state:
   a) r > 1 -> new state accepted
   b) r < 1 -> new state accepted with probability *r*. If new state not accepted, stay in the old state



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

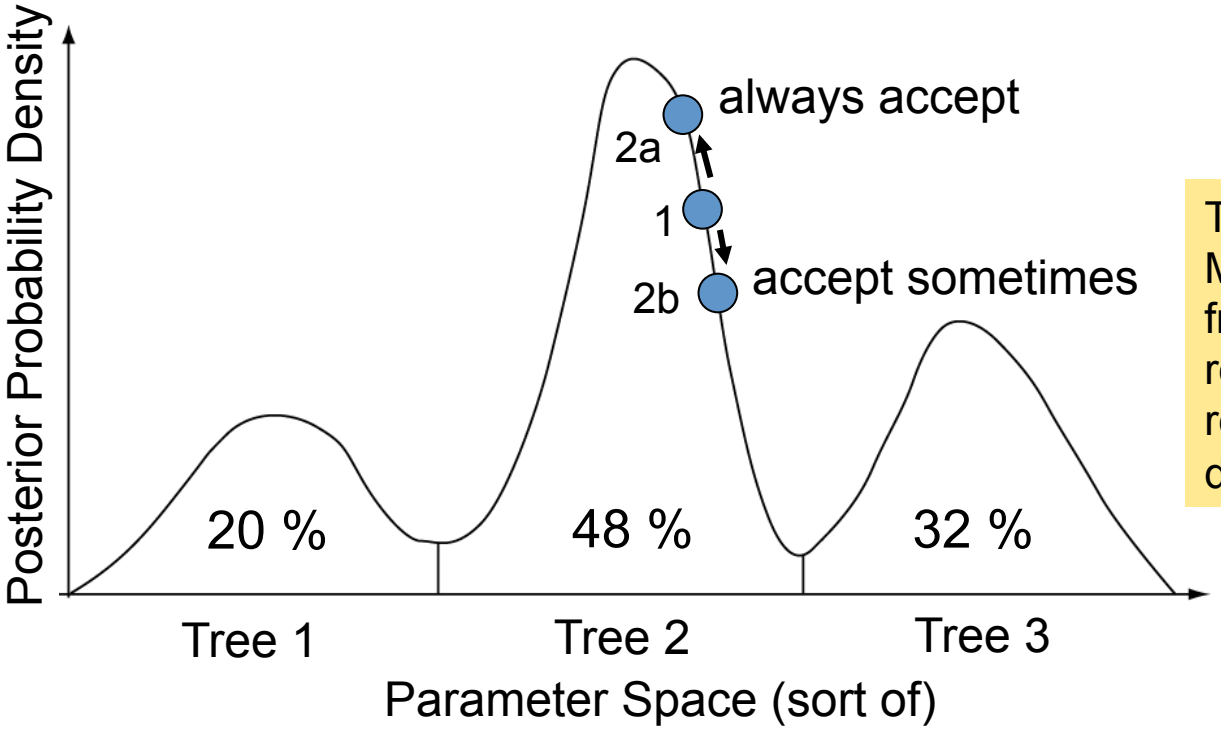
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (*r*) of new state to old state:
   a) r > 1 -> new state accepted
   b) r < 1 -> new state accepted with probability *r*. If new state not accepted, stay in the old state
4. Go to step 2 a BUNCH (x 10,000's – x 10,000,000's)



This slide "borrowed" from F. Ronquist

Markov chain Monte Carlo

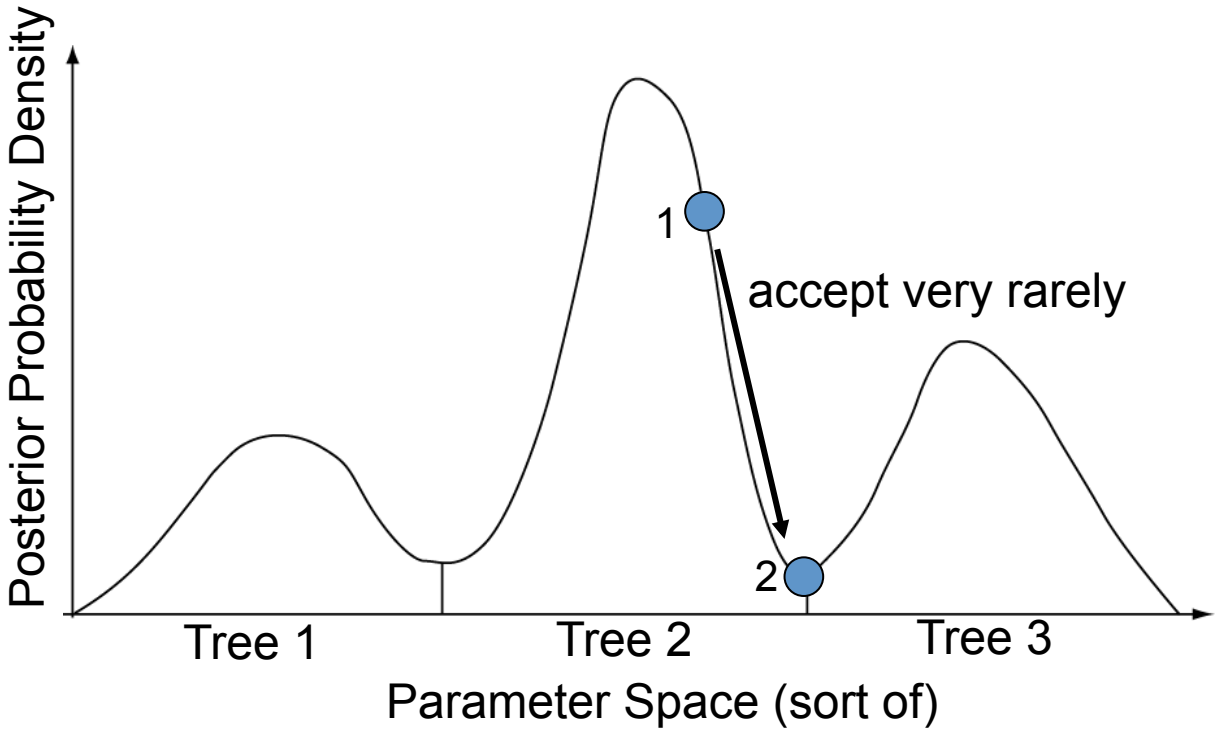
1. Start at an arbitrary point
2. Make a small random move
3. Calculate posterior density ratio (*r*) of new state to old state:
   a) r > 1 -> new state accepted
   b) r < 1 -> new state accepted with probability *r*. If new state not accepted, stay in the old state
4. Go to step 2 a BUNCH (x 10,000's – x 10,000,000's)



The proportion of time the MCMC procedure samples from a particular parameter region is an estimate of that region's posterior probability density

This slide "borrowed" from F. Ronquist

Metropolis Coupling

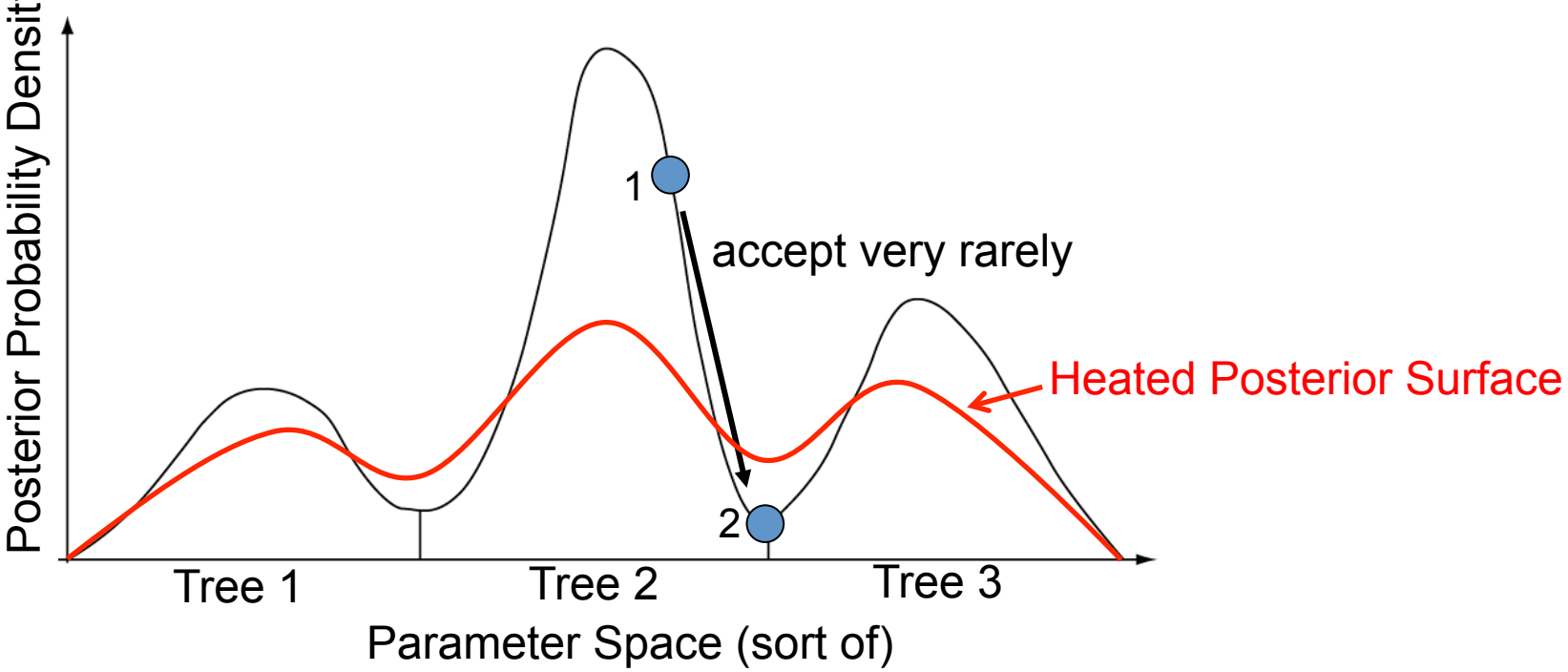


This slide "borrowed" from F. Ronquist

Metropolis Coupling

- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to (1-temp) when deciding whether to accept a move.

$r^{(1\text{-temp})}$ approaches 1 (flat surface) as temp. increases

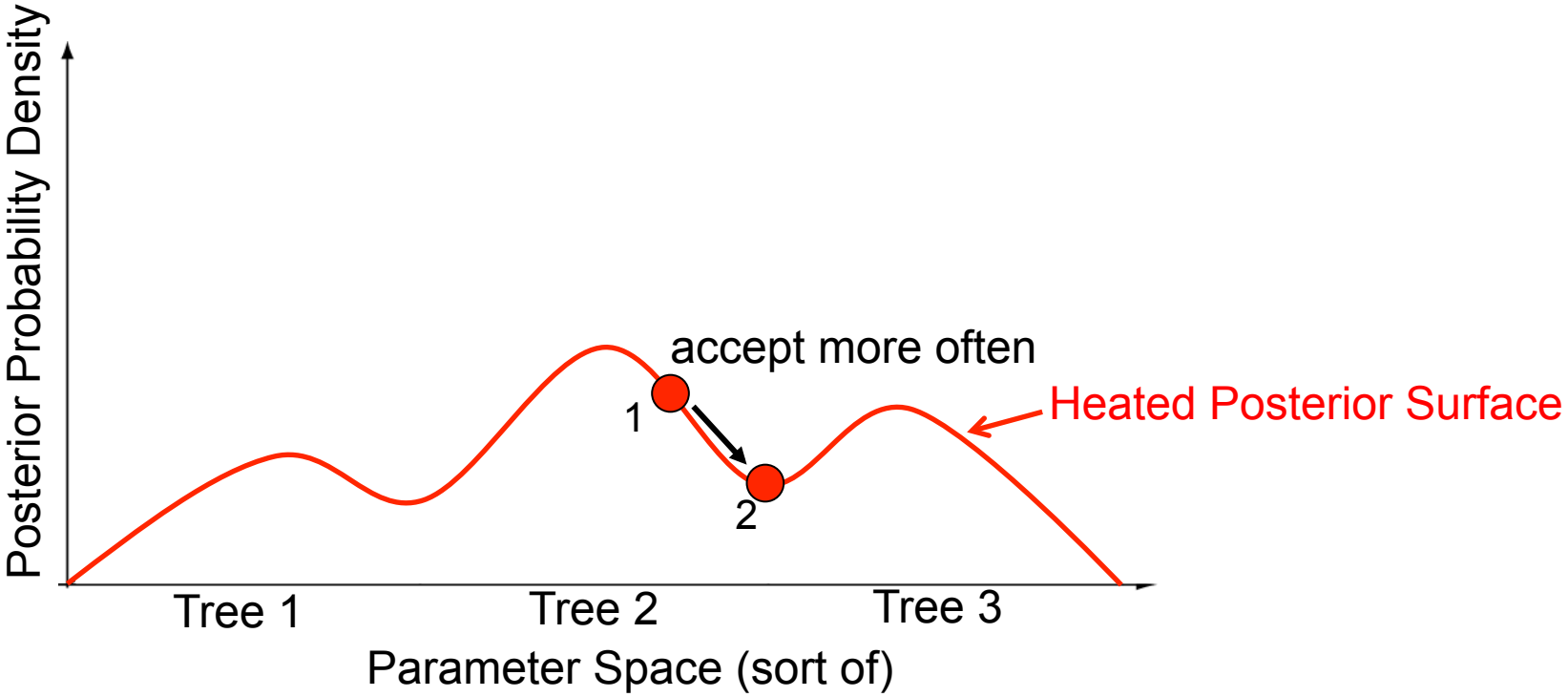


This slide “borrowed” from F. Ronquist

Metropolis Coupling

- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to (1-temp) when deciding whether to accept a move.

$r^{(1\text{-temp})}$ approaches 1 (flat surface) as temp. increases



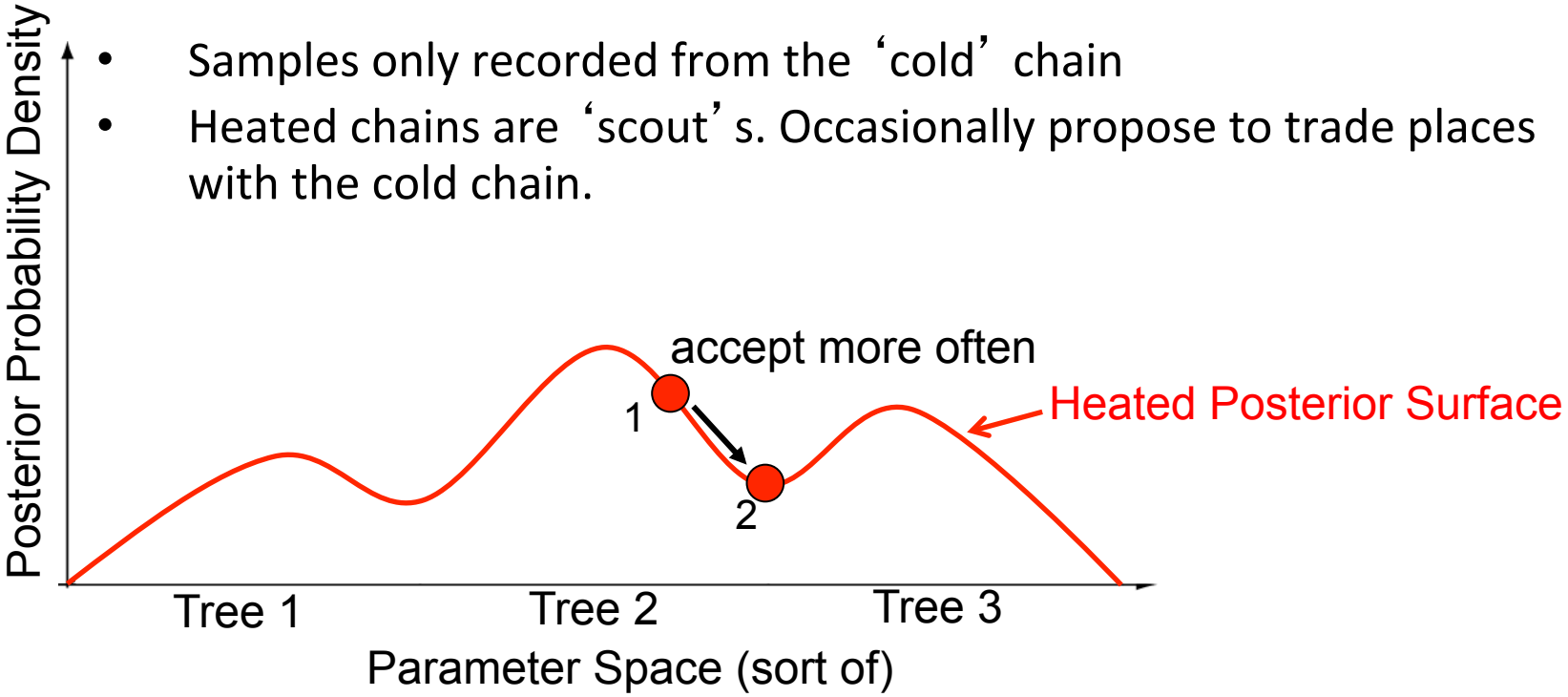
This slide “borrowed” from F. Ronquist

Metropolis Coupling

- Same rules as regular MCMC, but now there are multiple chains with different ‘temperatures’.
- ‘Heated’ chains sample a ‘melted’ version of the posterior
- Only difference is that heated chains raise the ratio of posterior densities to (1-temp) when deciding whether to accept a move.

$r^{(1\text{-temp})}$ approaches 1 (flat surface) as temp. increases

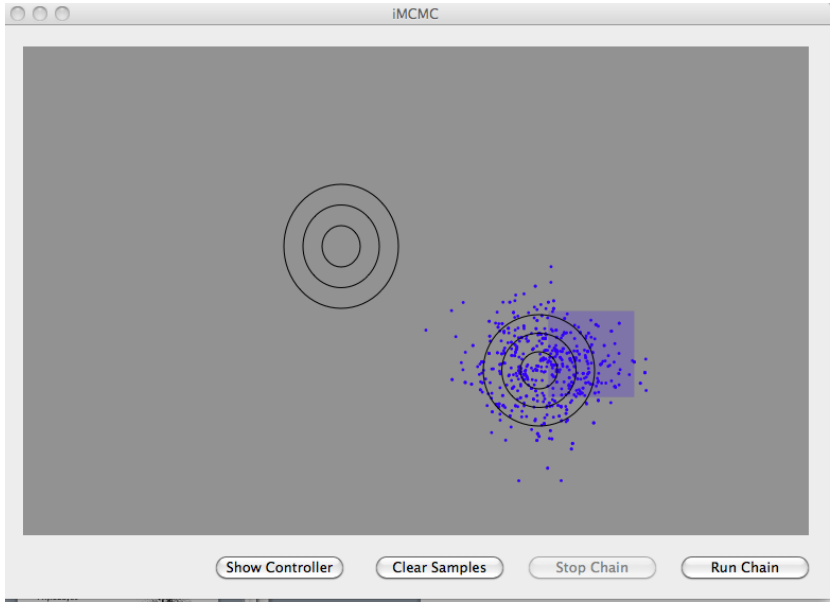
- Samples only recorded from the ‘cold’ chain
- Heated chains are ‘scout’s. Occasionally propose to trade places with the cold chain.

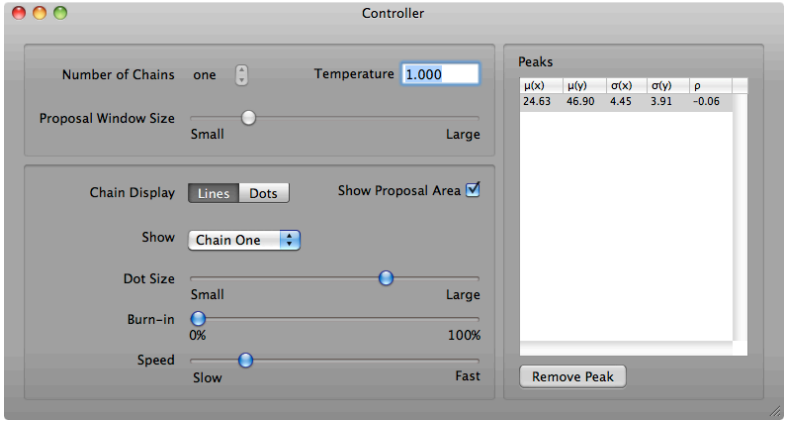


This slide “borrowed” from F. Ronquist

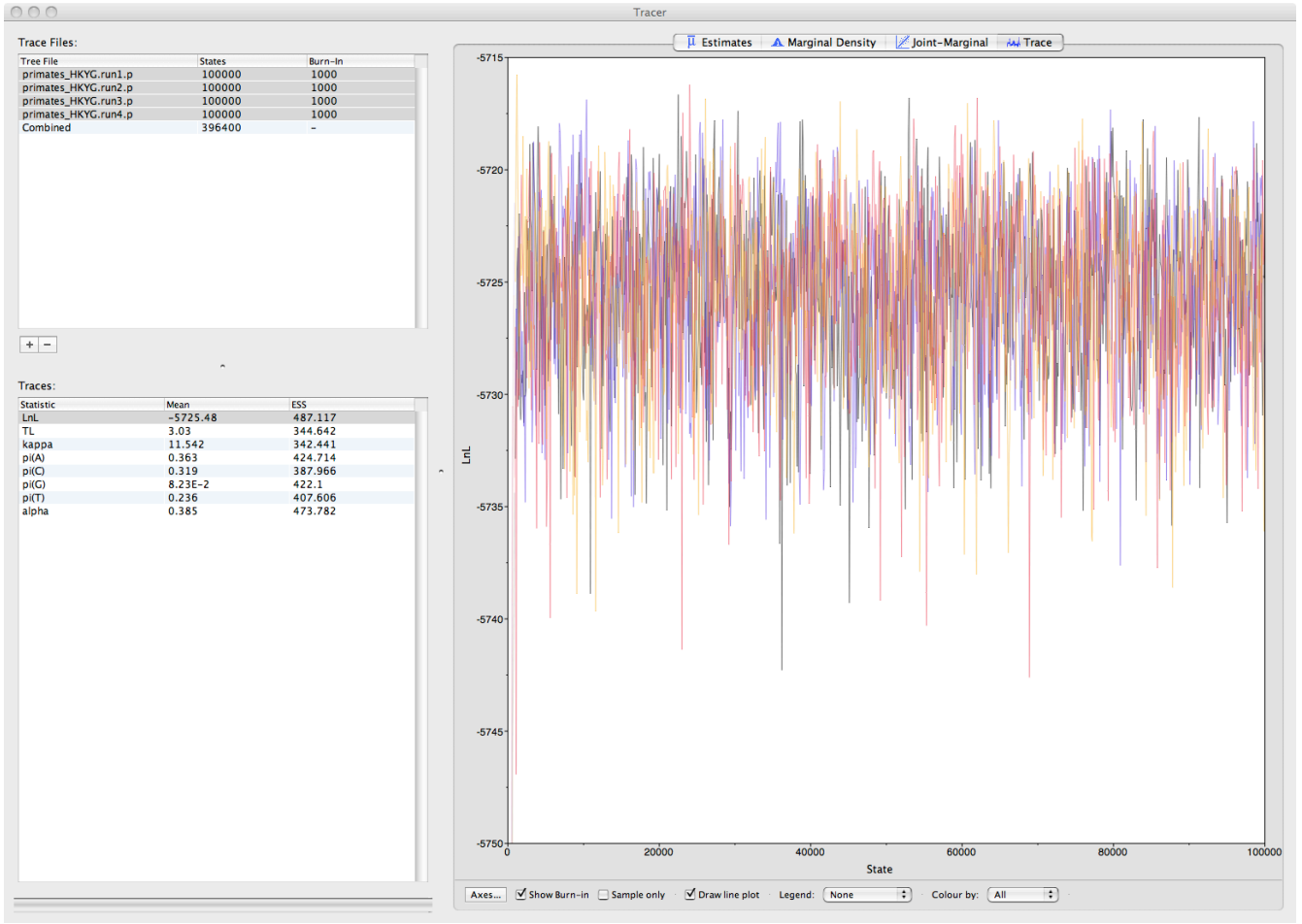
Toy MCMC Exploration

- MCRobot – PC (Lewis)
  - http://www.eeb.uconn.edu/people/plewis/software.php
- iMCMC – Mac (Huelsenbeck)
  - http://fisher.berkeley.edu/cteg/software.html



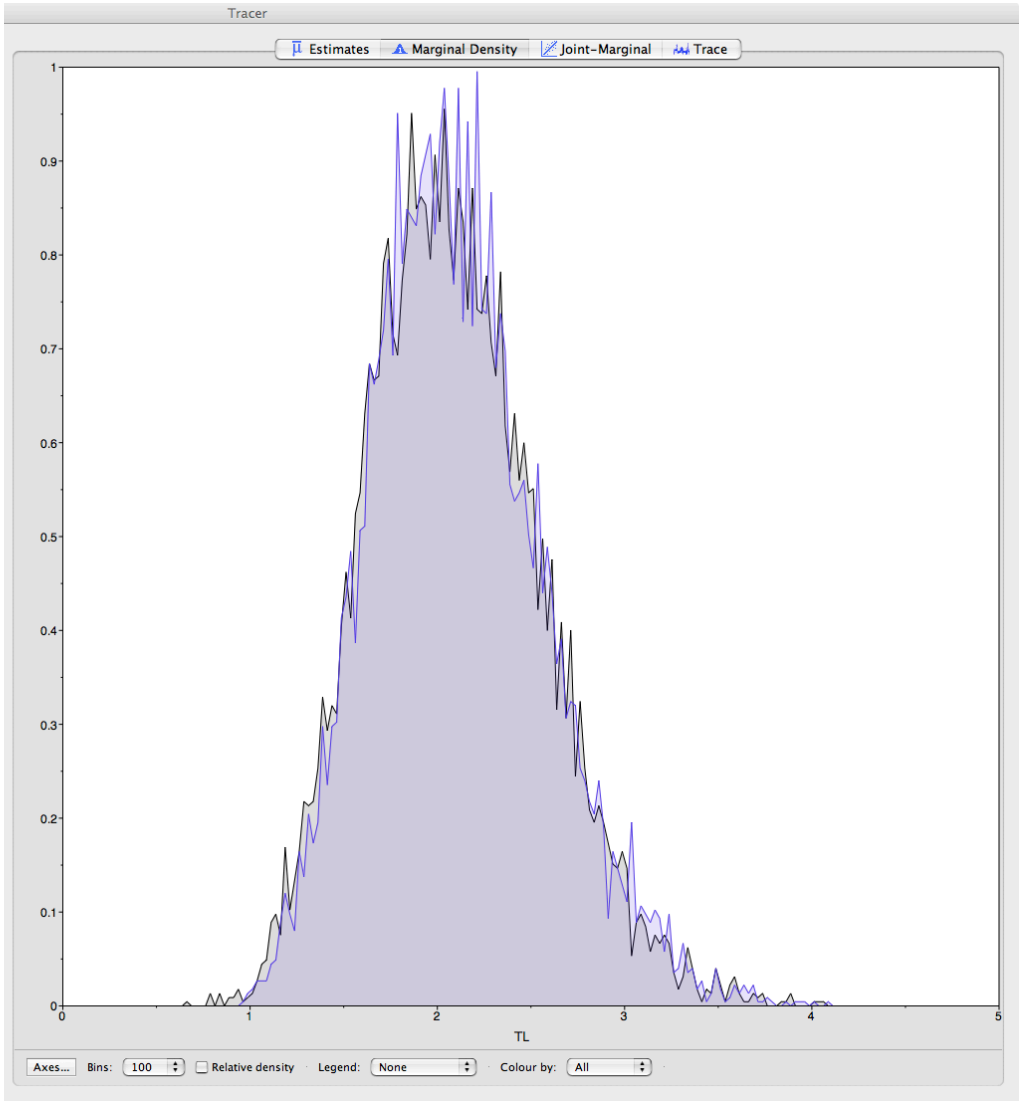


Convergence of Scalars - Tracer



Running on Empty

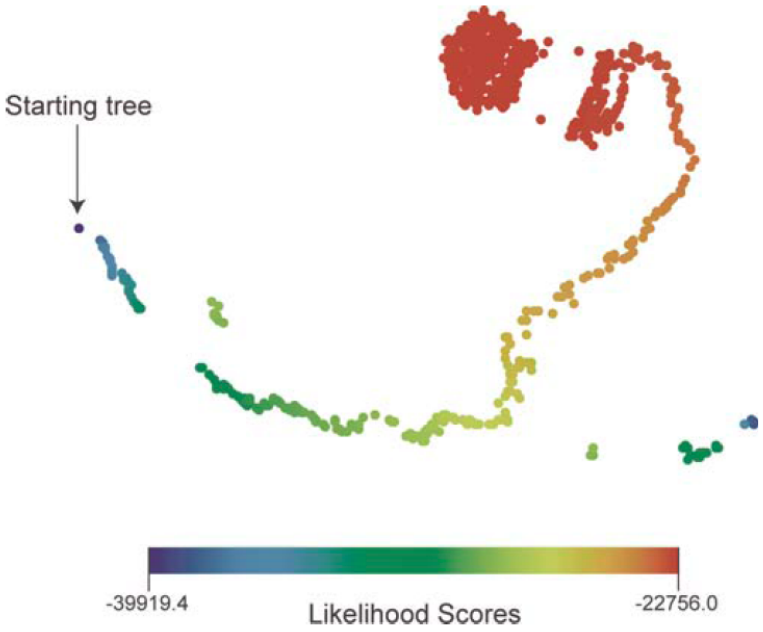
```
#NEXUS
begin data;
dimensions ntax=12 nchar=5;
format datatype=dna interleave=no gap=- missing=?;
matrix
Tarsius_syrichta    ?????
Lemur_catta         ?????
Homo_sapiens        ?????
Pan                 ?????
Gorilla             ?????
Pongo               ?????
Hylobates           ?????
Macaca_fuscata      ?????
M_mulatta           ?????
M_fascicularis      ?????
M_sylvanus          ?????
Saimiri_sciureus    ?????
;
end;
```

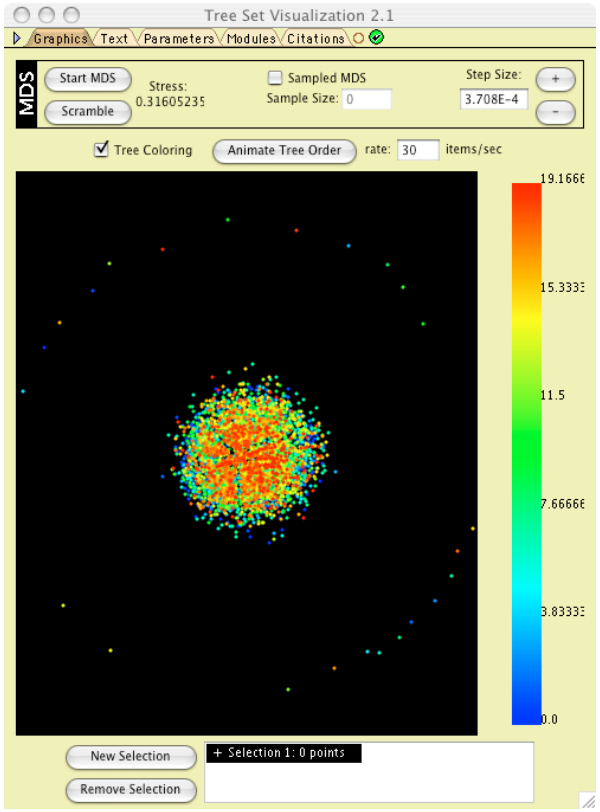


Or now in MrBayes:

mcmc data=no

Topological Convergence - TreeSetViz



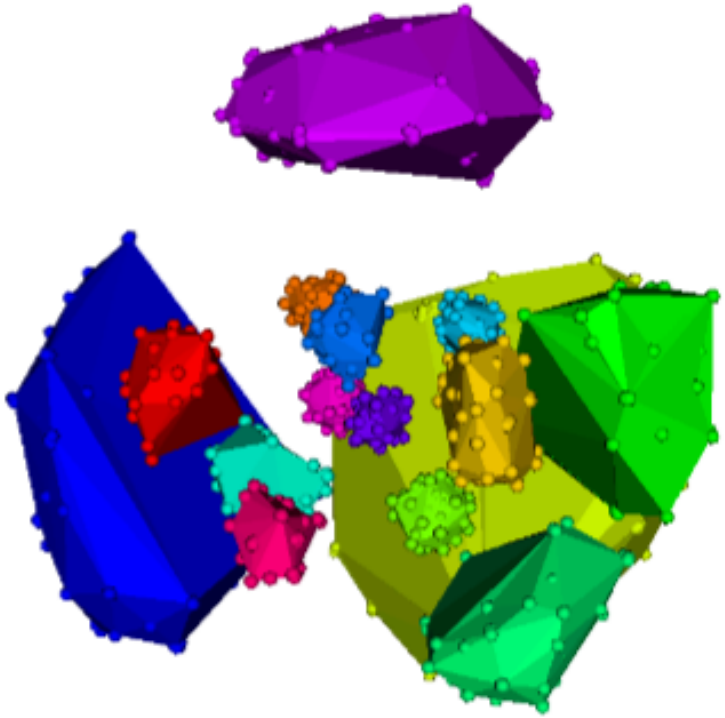


TreeSetViz: http://comet.lehman.cuny.edu/treeviz/
Mesquite: http://www.mesquiteproject.org/mesquite/mesquite.html
Hillis et al., Analysis and Visualization of Tree Space, *Syst. Biol.*, 54(3): 471-482.

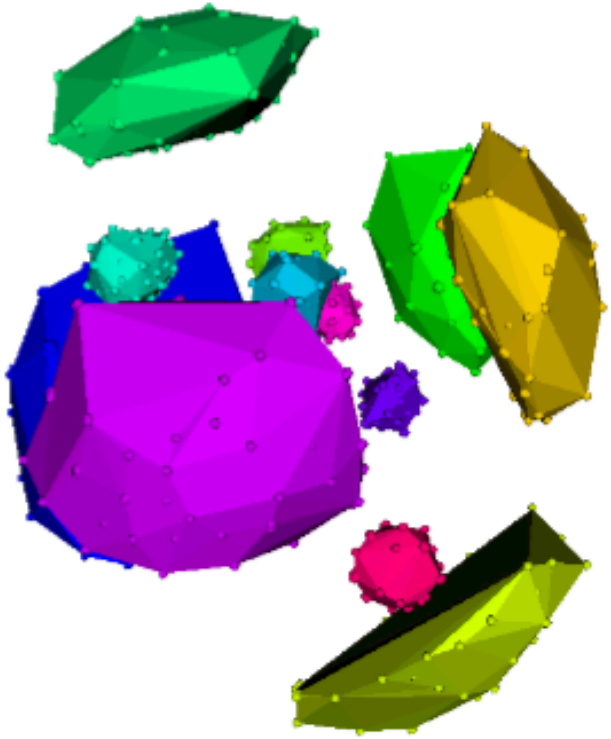
Topological Convergence - TreeScaper

Colors = Trees from Different Genes

Salamanders



Mammals

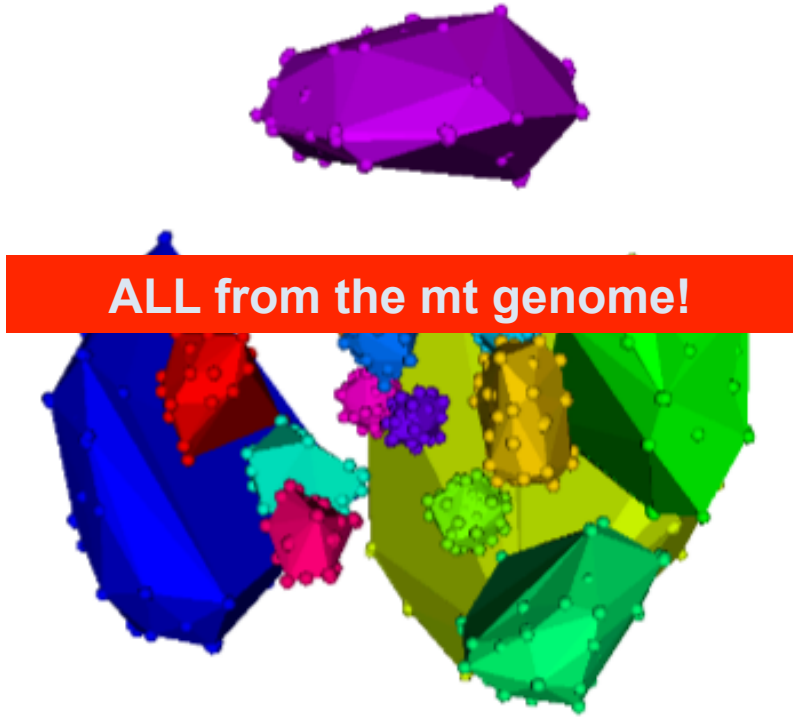


http://bpd.sc.fsu.edu/index.php/diagnostic-software/104-treescaper

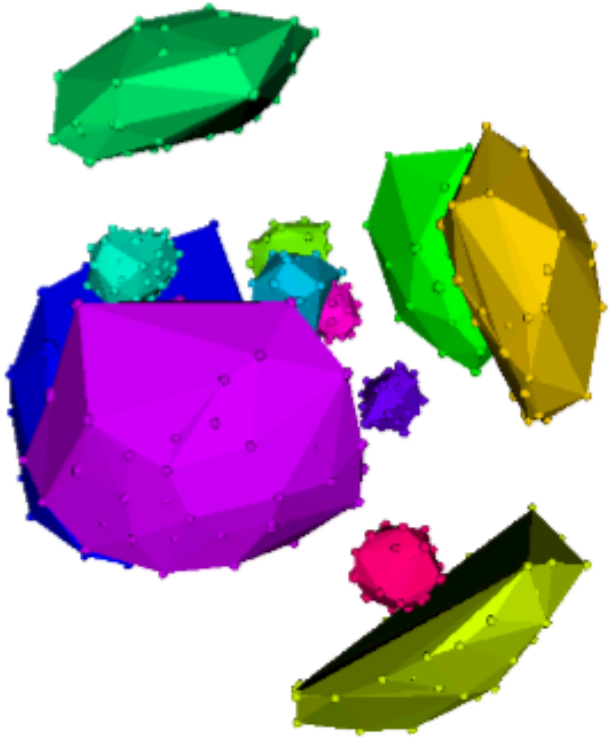
Topological Convergence - TreeScaper

Colors = Trees from Different Genes

Salamanders

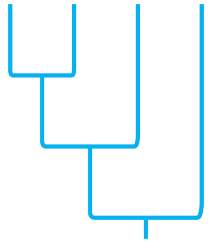


Mammals



http://bpd.sc.fsu.edu/index.php/diagnostic-software/104-treescaper

Posterior Odds Ratio

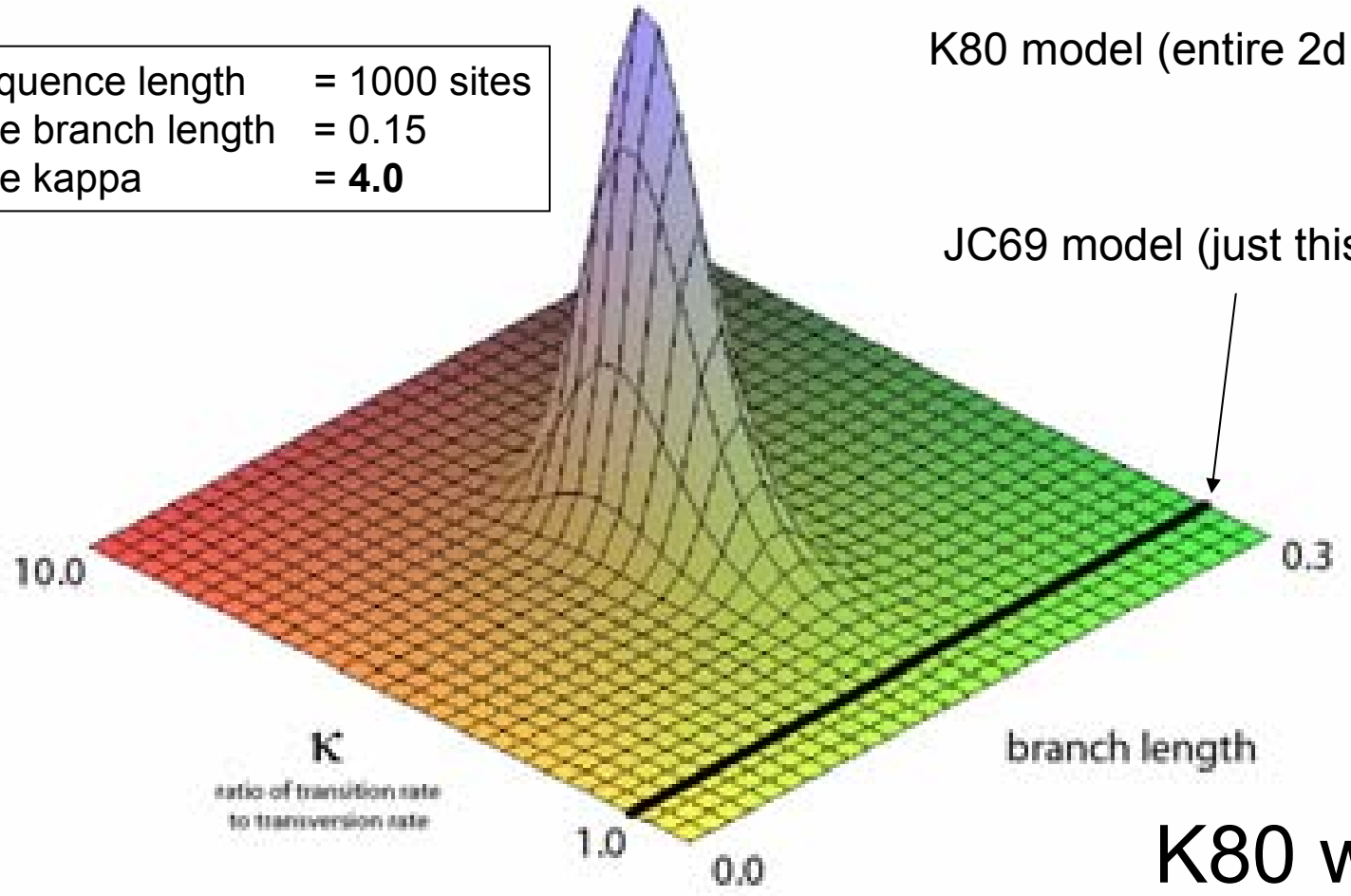


Prior Odds | Bayes Factor | Posterior Odds

$$\frac{P(M_1)}{P(M_2)} \cdot \frac{P(\text{data} \mid M_1)}{P(\text{data} \mid M_2)} = \frac{P(M_1 \mid \text{data})}{P(M_2 \mid \text{data})}$$

Marginal Likelihood of a Model

sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= **4.0**



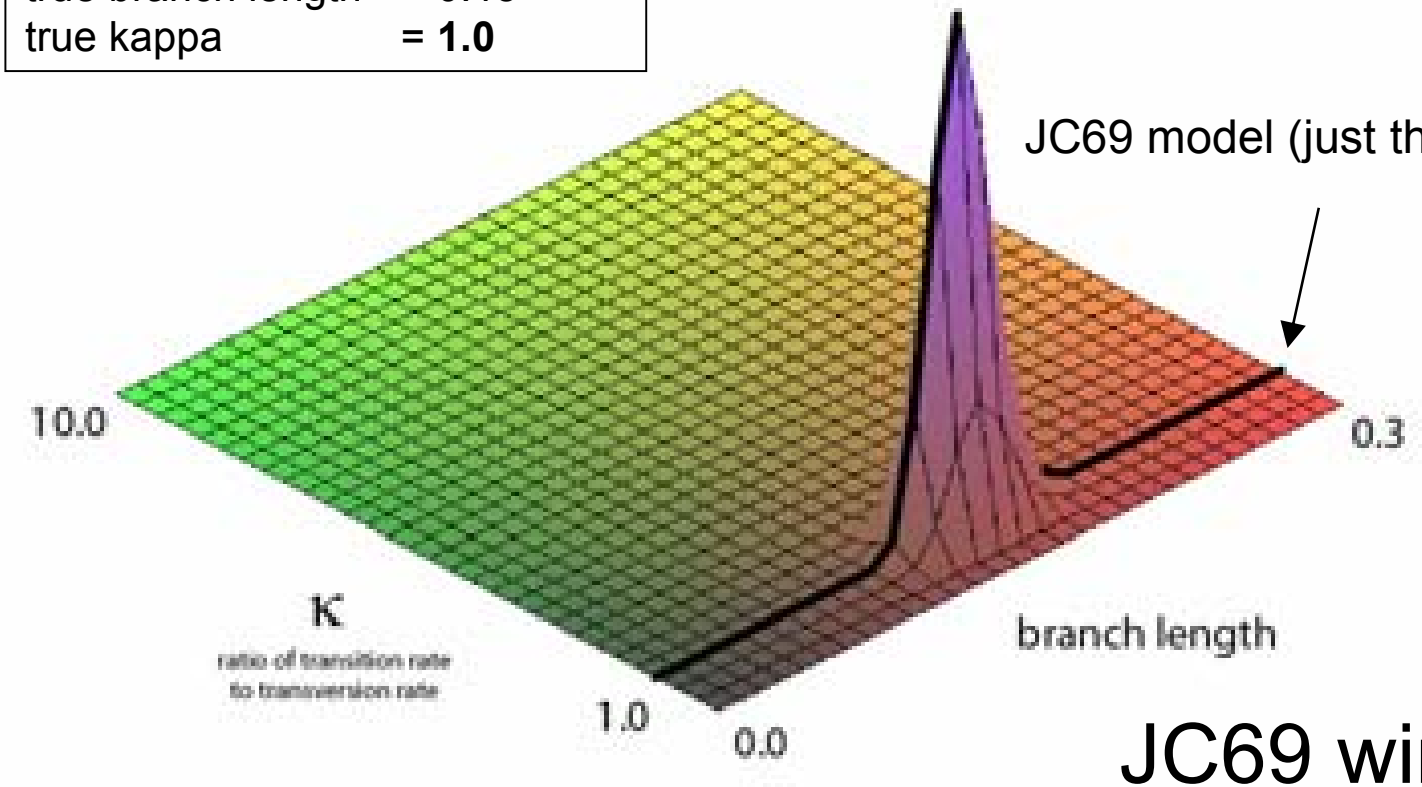
K80 wins

Marginal Likelihood of a Model

sequence length	= 1000 sites
true branch length	= 0.15
true kappa	= **1.0**

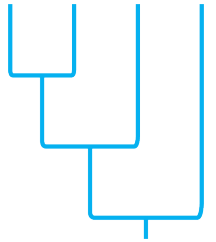
K80 model (entire 2d space)

JC69 model (just this 1d line)



JC69 wins

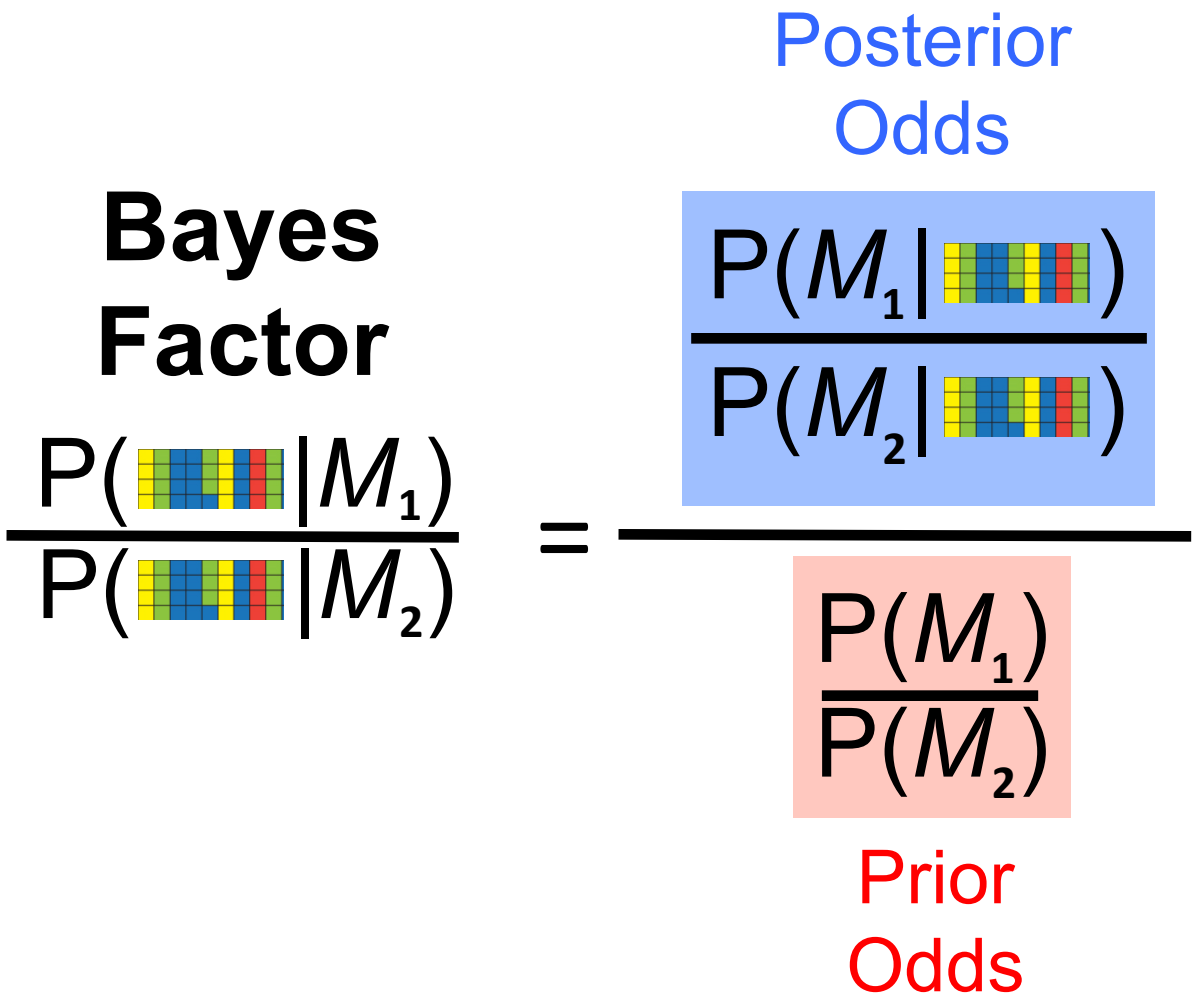
Posterior Odds Ratio



Prior Odds | Bayes Factor | Posterior Odds

$$\frac{P(M_1)}{P(M_2)} \cdot \frac{P(\text{data} \mid M_1)}{P(\text{data} \mid M_2)} = \frac{P(M_1 \mid \text{data})}{P(M_2 \mid \text{data})}$$

Bayes Factor



Bayes Factor

Bayes Factor

$$\frac{P(\text{  }|M_1)}{P(\text{  }|M_2)}$$

$=$

Posterior Odds

$$\frac{P(M_1|\text{  })}{P(M_2|\text{  })}$$

$$\frac{P(M_1)}{P(M_2)}$$

Prior Odds

What should be true in the case of equal prior probabilities?

Interpreting Bayes Factors

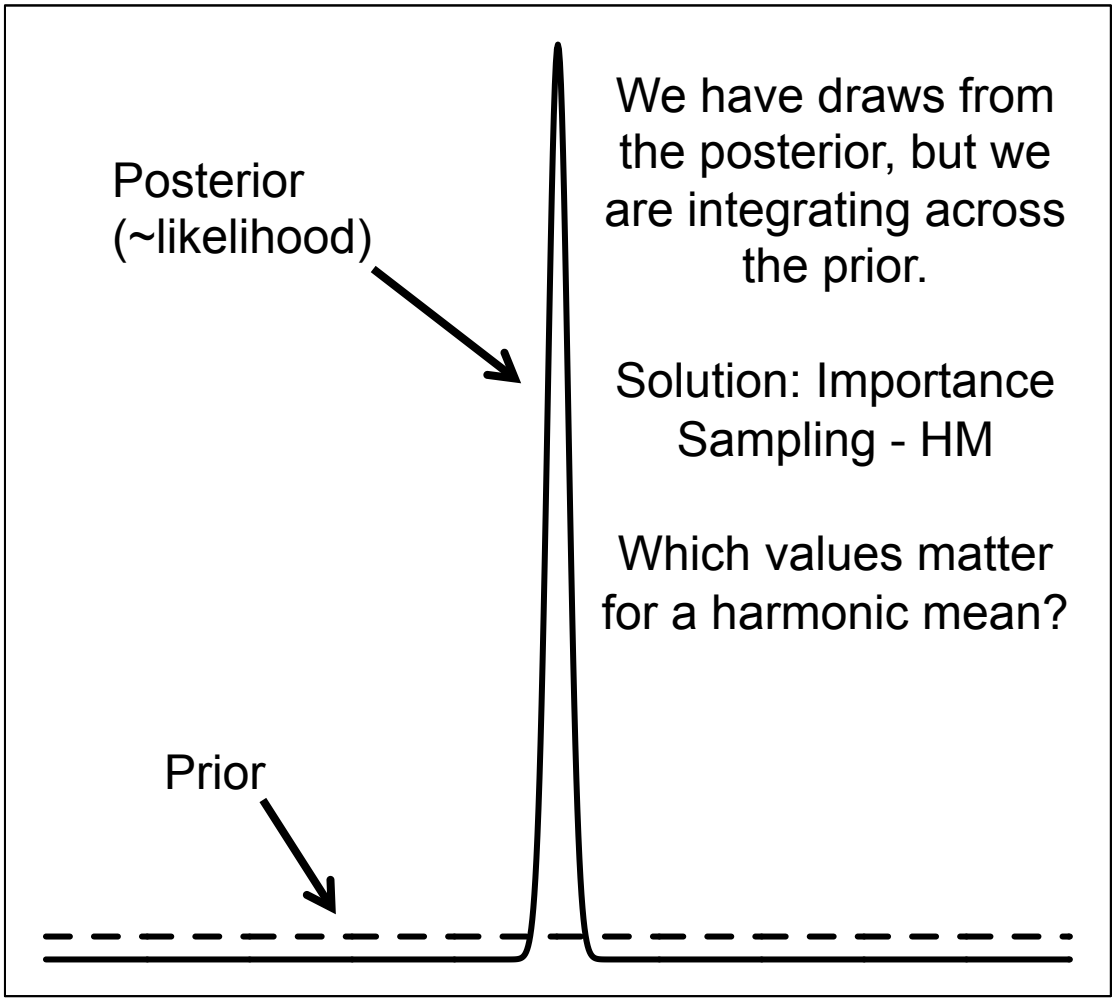
- BFs greater than 1 indicate support for model 1 (numerator), less than 1 indicate support for model 2 (denominator)
- log(BF) values greater than 0 support model 1 and less than 0 support model 2
- Practically, log(BF) values of > 5 or < -5 are often to indicate very strong support for the corresponding model.

Estimating Bayes Factors

- With RJ, use posterior & prior odds ratios

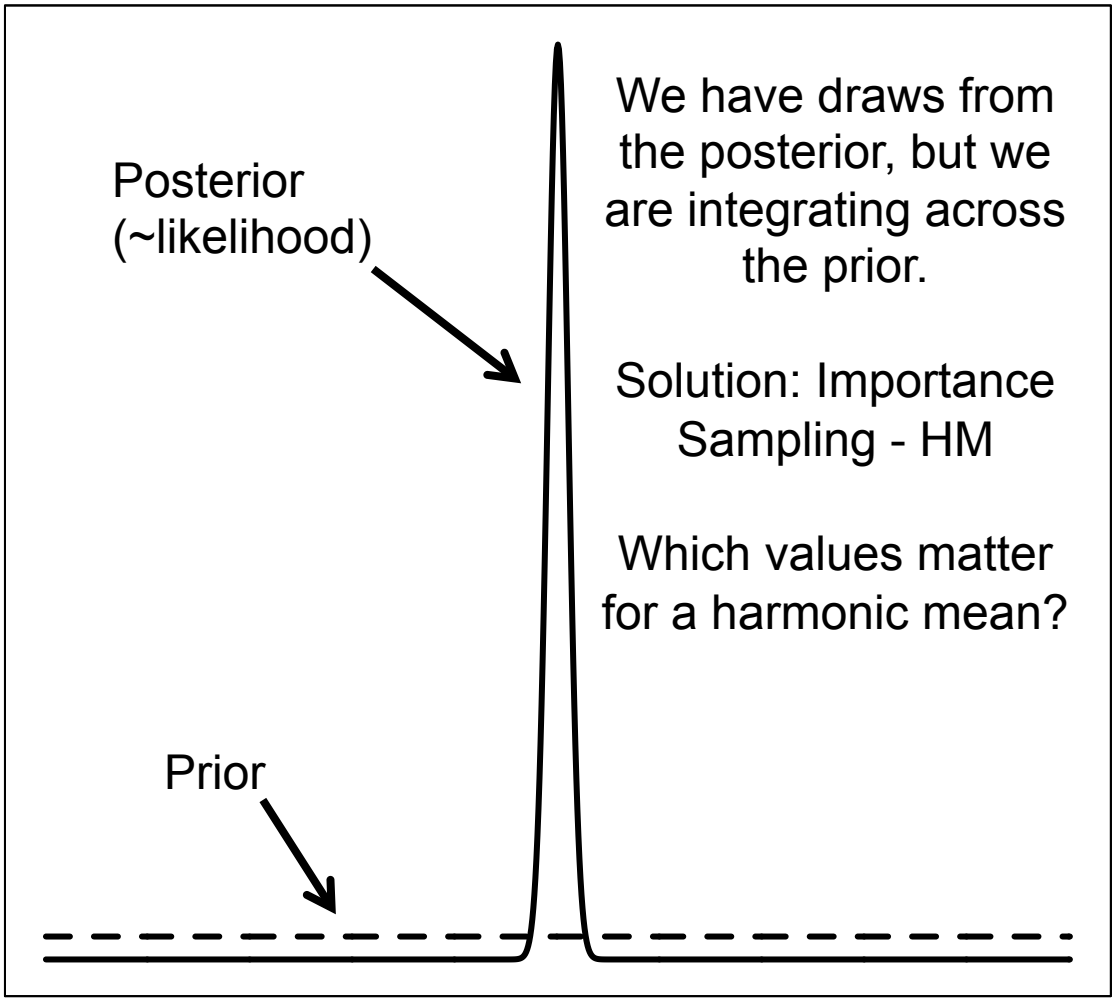
- Can also calculate marginal likelihoods directly
  - Harmonic mean (easy, but inaccurate)
  - Stepping stone (more accurate, but harder)
  - Thermodynamic integration (similar to SS)
  - MORE SOON!

Harmonic Mean Estimator



Newton and Raftery. 1994.
J. R. Statistic. Soc. B 56:3-48.

Harmonic Mean Estimator



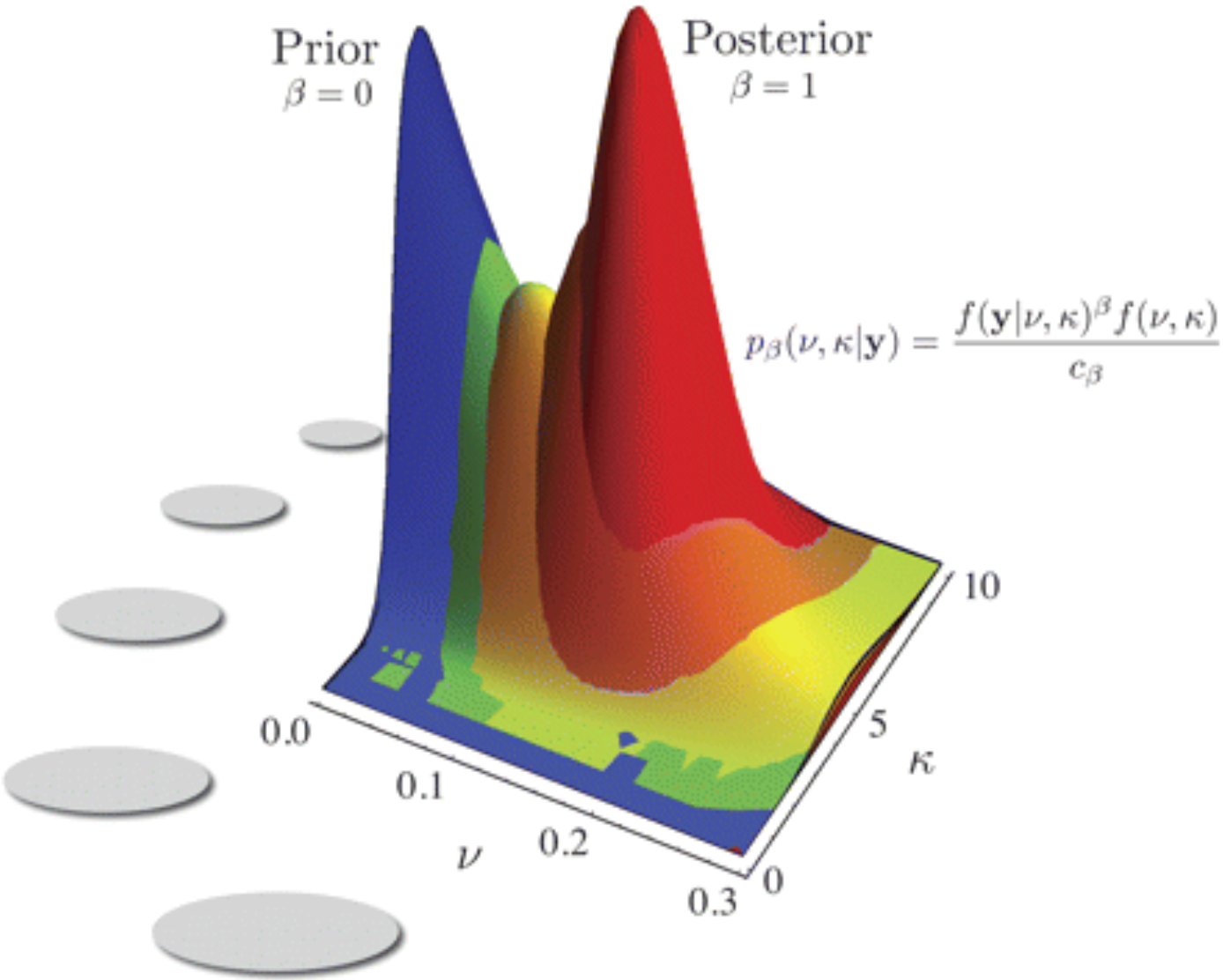
Look up the HM estimate of the marginal likelihood in the .lstat file from your MrBayes run with a Jukes-Cantor model (after running sump) and record it.

Newton and Raftery. 1994.
J. R. Statistic. Soc. B 56:3-48.

Harmonic Mean

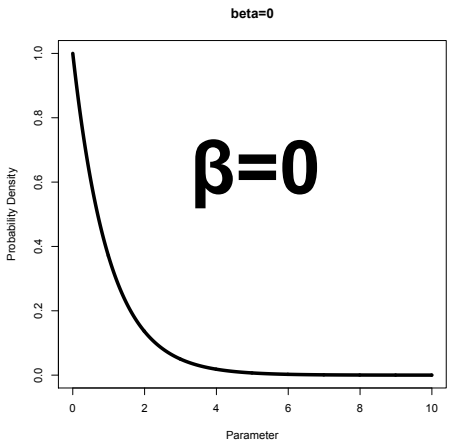
$$H = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \ldots + \frac{1}{x_n}}$$

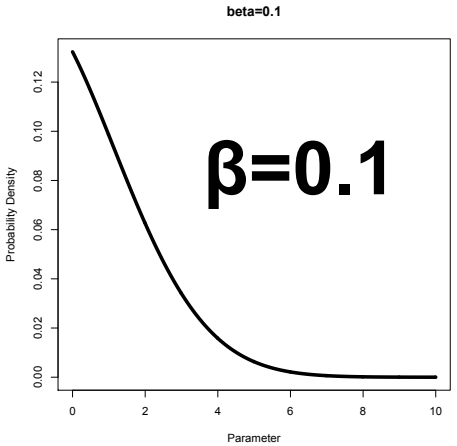
Stepping Stone Sampling

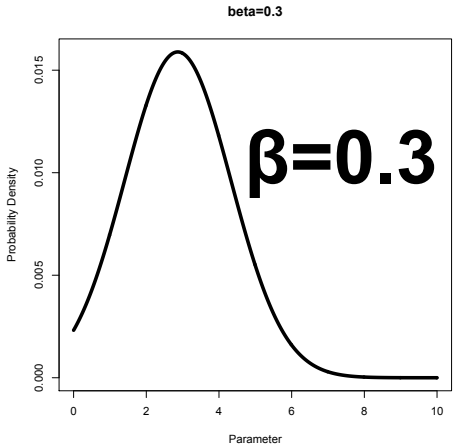


Power Posteriors

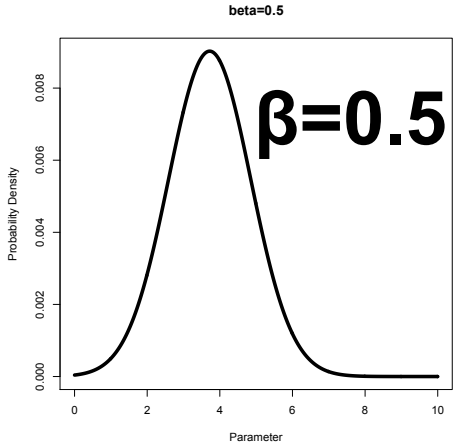
Prior

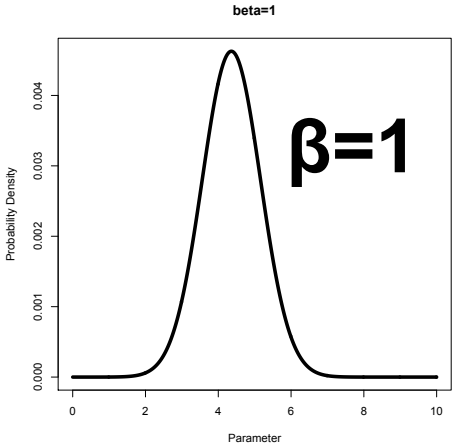




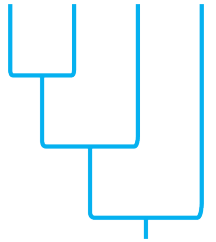


Posterior



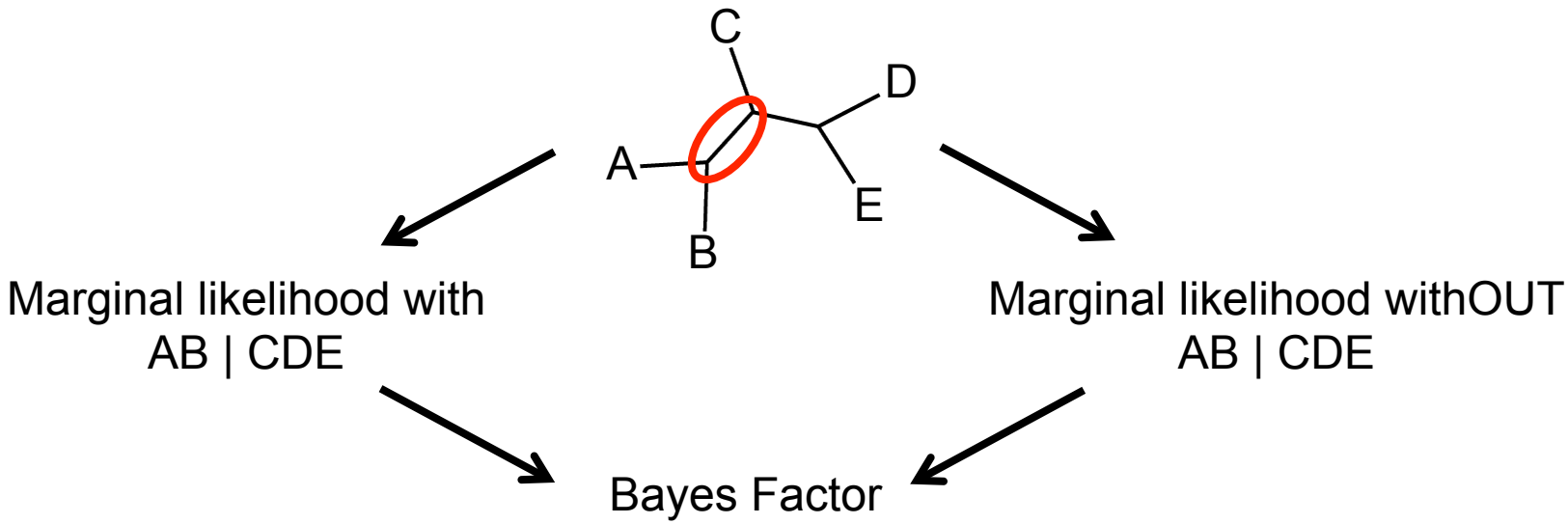


Topology BFs



Can also use stepping stone to calculate marginal likelihoods for sets of trees subject to particular constraints.

For instance, comparing the marginal likelihoods for a set of trees that all contain a branch to the set of trees that all do NOT contain that branch would give you a Bayes factor supporting that branch.



Model Choice v “Adequacy”

Which of the available models will perform best?

vs.

Is any given model sufficient to provide unbiased inferences?

Model Choice v "Adequacy"

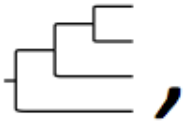
Or, better, **Plausibility**

Which of the available models will perform best?

vs.

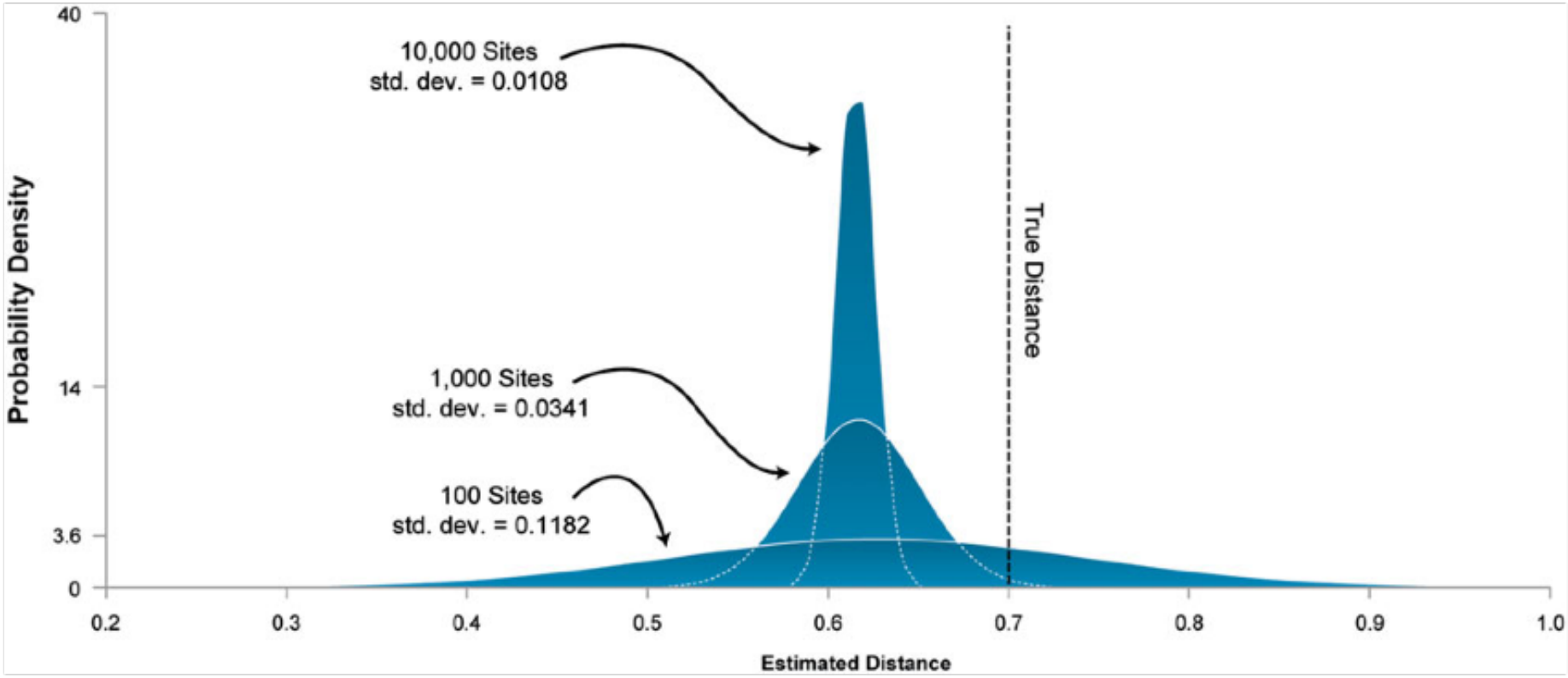
Is any given model sufficient to provide unbiased inferences?

Posterior Prediction

Could  have come from $P($  $,\theta|$  $)$?

Could the model and priors plausibly have given rise to the data?

If not, phylogenies may be **biased**



Kumar et al. 2012. *Mol. Biol. Evol.*

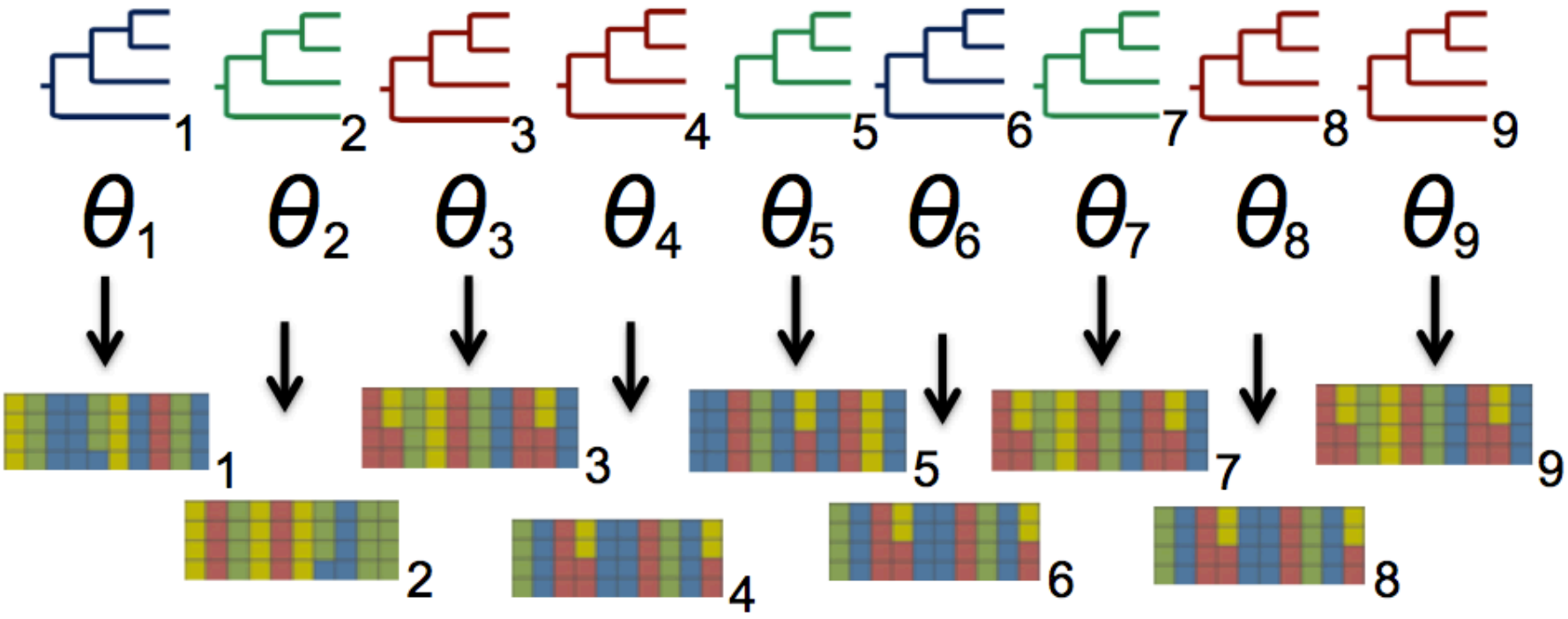
Posterior Prediction

"We do not like to ask, 'Is our model true or false?', since probability models in most analyses will not be perfectly true…The more relevant question is, 'Do the model's deficiencies have a noticeable effect on the substantive inferences?'"

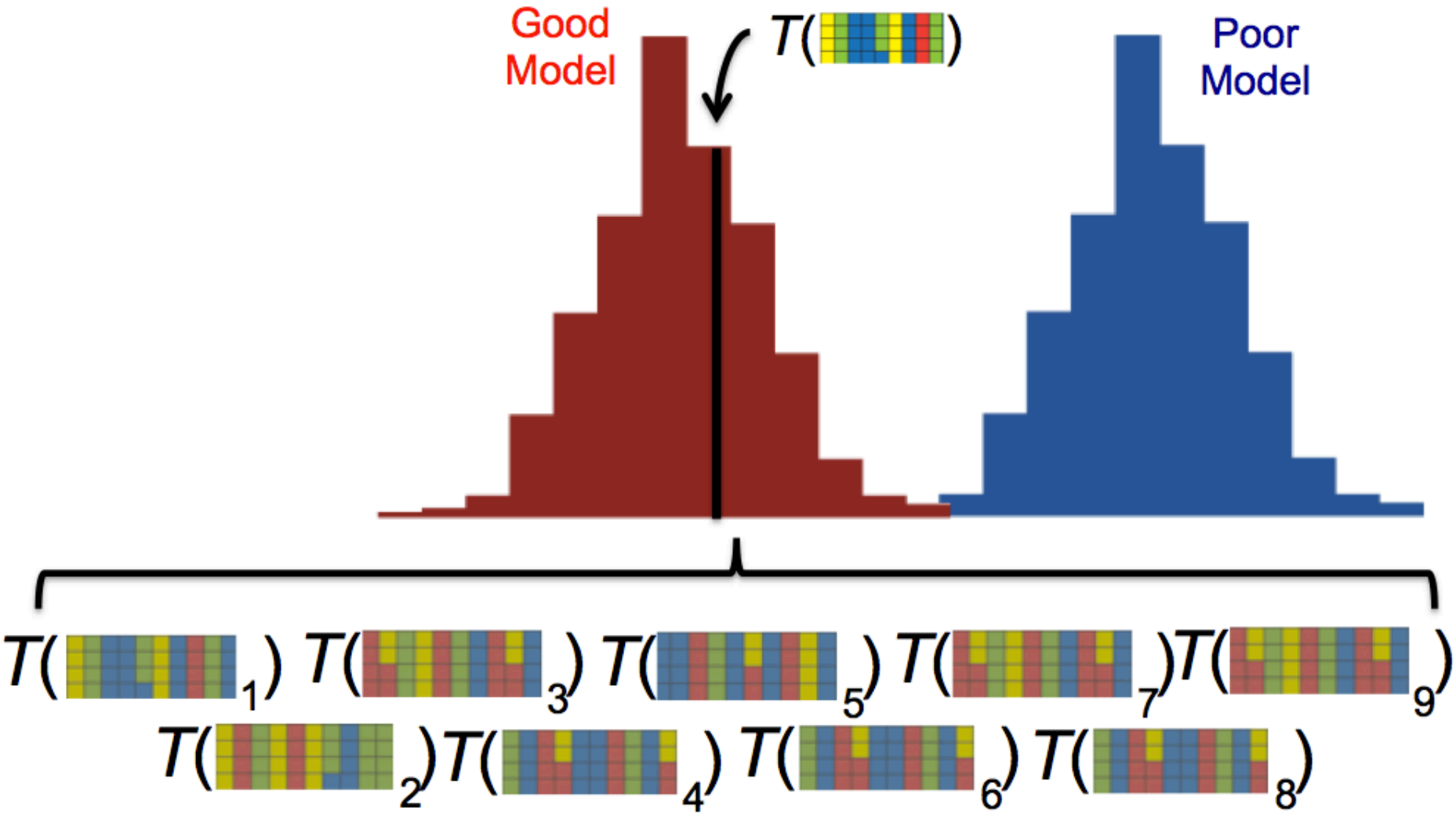
-A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin
Bayesian Data Analysis

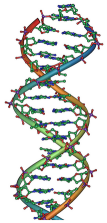
Posterior Prediction

$P(\text{tree}, \theta \mid \text{data})$

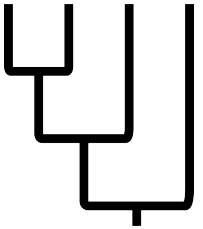


Posterior Prediction

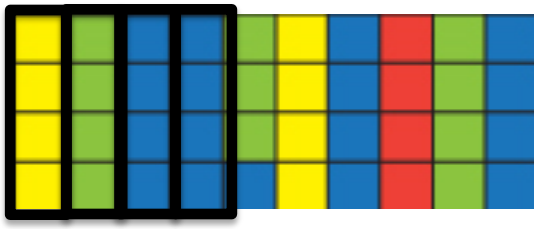




Posterior Predictive Simulation

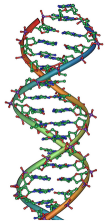


Previously proposed statistic*: Multinomial Likelihood

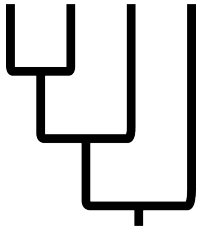


Based on the frequency of different site patterns

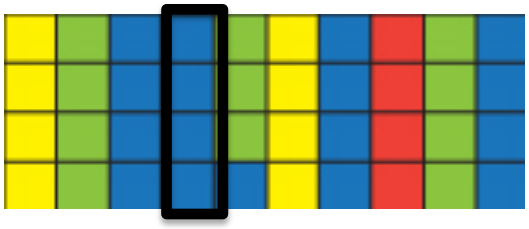
*Goldman, 1993; Bollback, 2002



Posterior Predictive Simulation



Previously proposed statistic*: Multinomial Likelihood

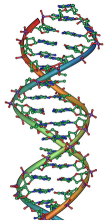


Based on the frequency of different site patterns

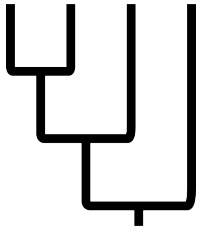
Tests if the assumed and generating models
produce data with similar site pattern frequencies

Intuitively appealing, but very sensitive to branch-length biases.
Can reject adequacy, **even when inferred phylogeny is correct**

*Goldman, 1993; Bollback, 2002



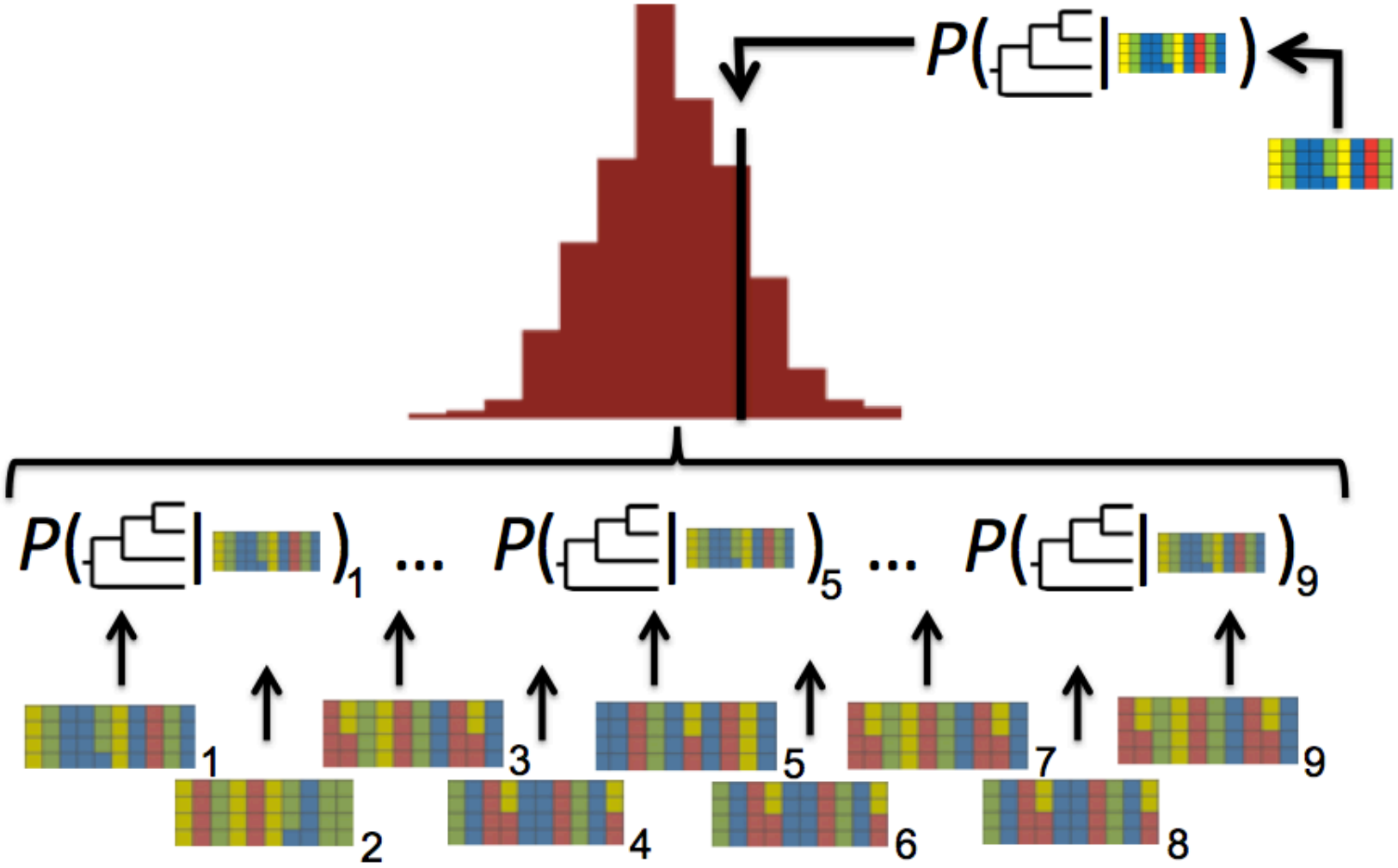
Posterior Predictive Simulation

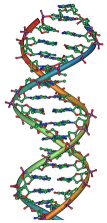


Previously proposed statistics:

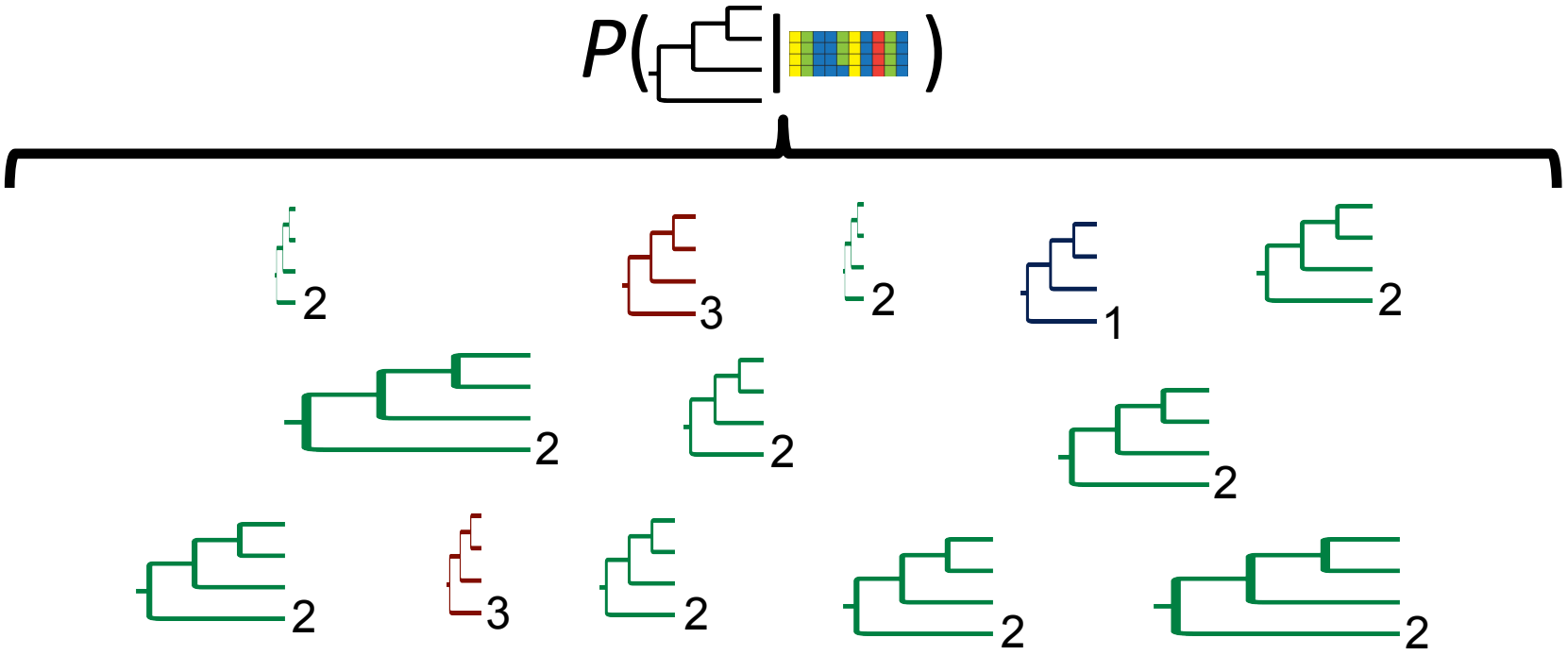
- Multinomial Likelihood
- Number of Unique Site Patterns
- Frequency of Invariant Sites
- Heterogeneity of Base Frequencies
- Number of parsimony-inferred "parallel" sites

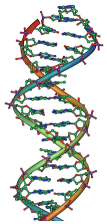
Posterior Prediction



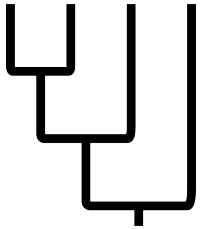


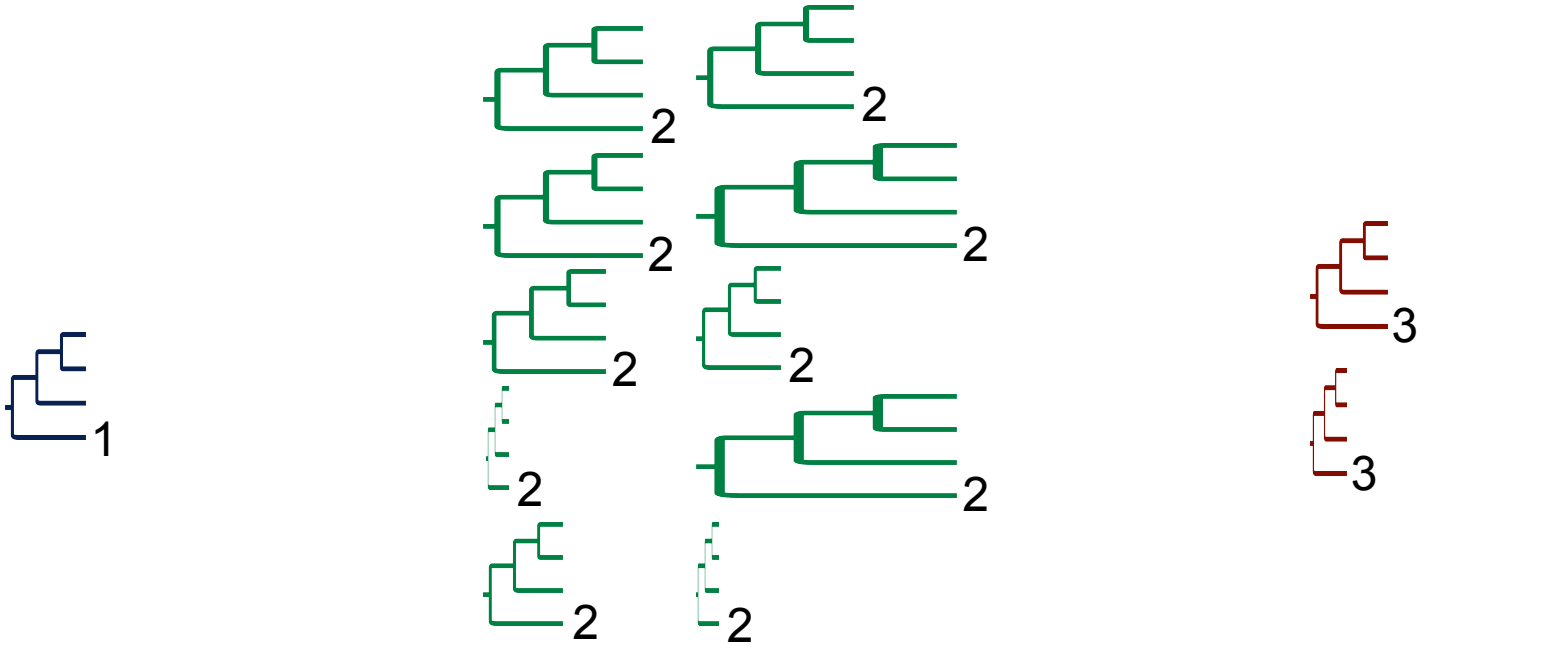
Marginal Test Quantities





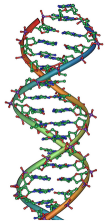
Marginal Test Quantities - Topology



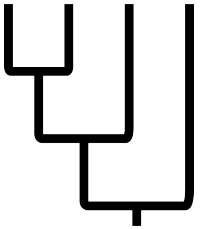


1/13 10/13 2/13

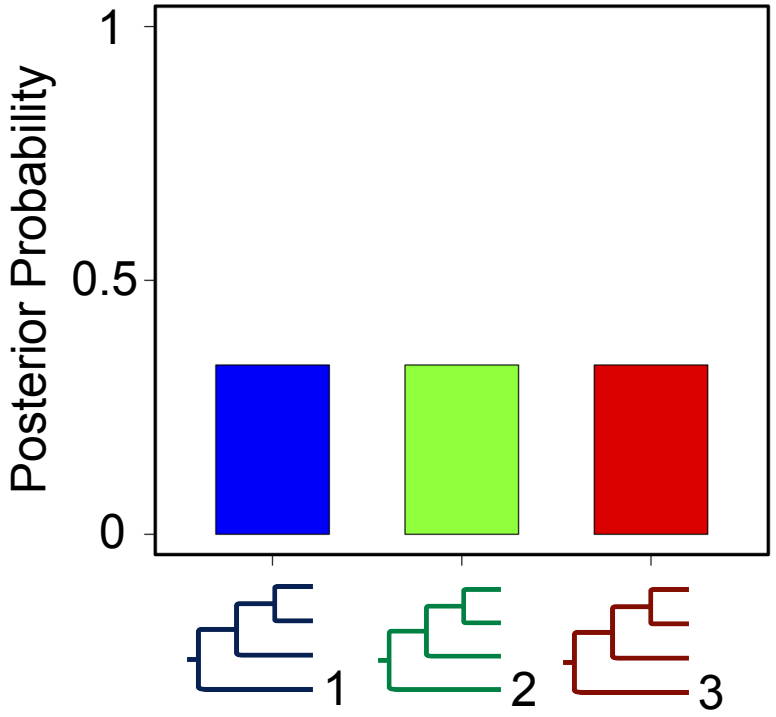
Integrating across variation in branch lengths (nuisance parameter)



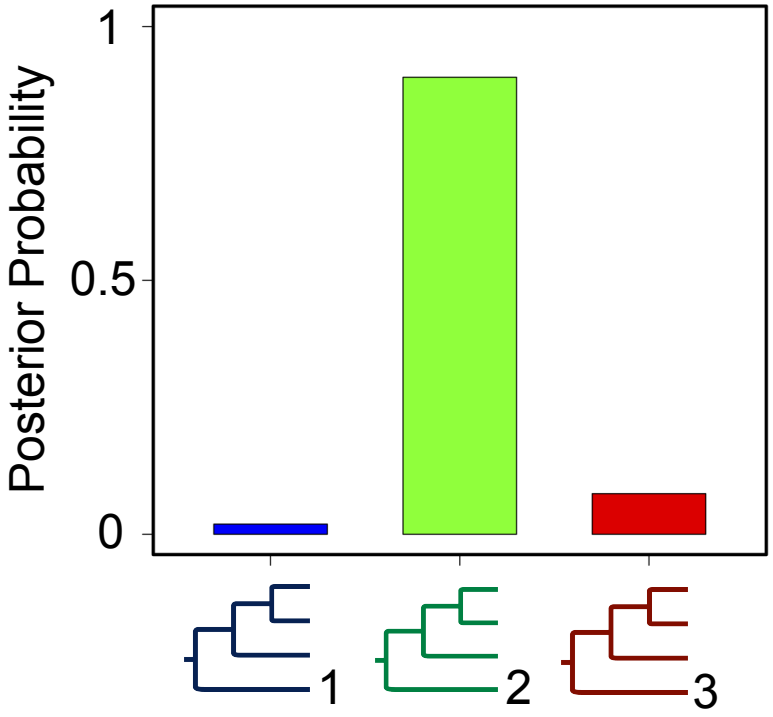
Entropy Quantifies Support Across Topologies



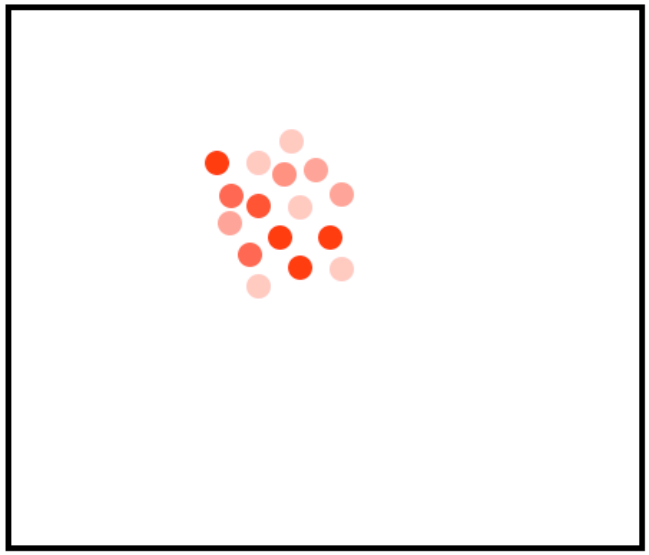
Higher Entropy



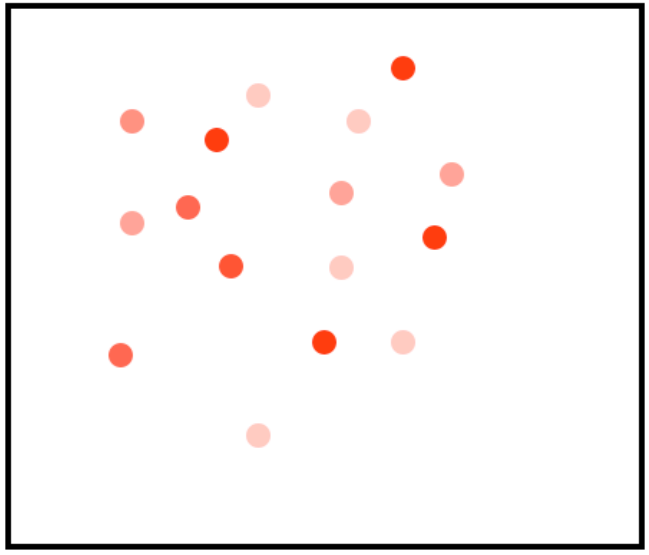
Lower Entropy



Topological Clustering Statistics

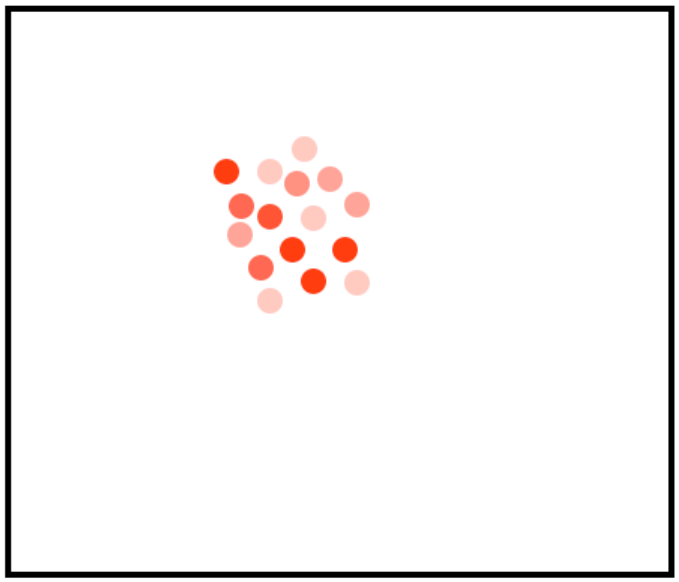


Tree Space

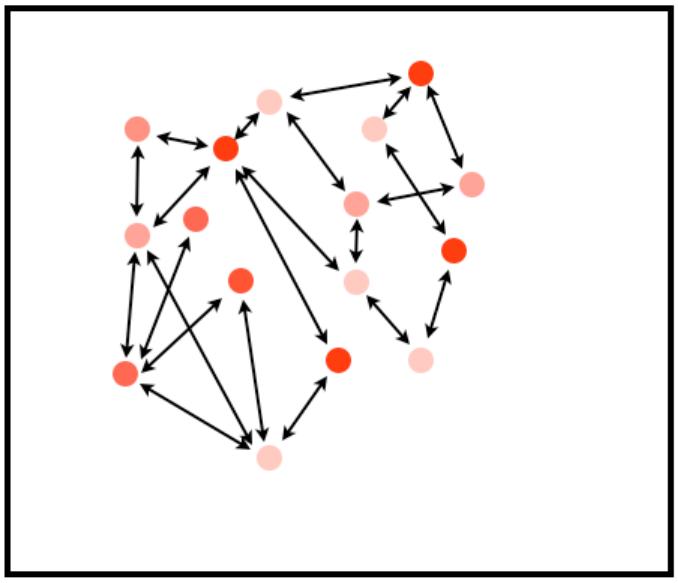


Tree Space

Topological Clustering Statistics

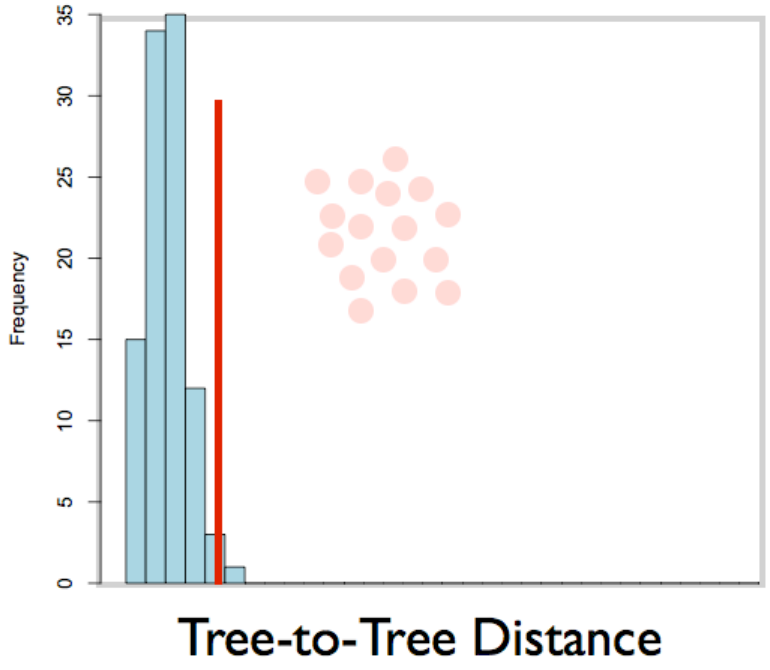


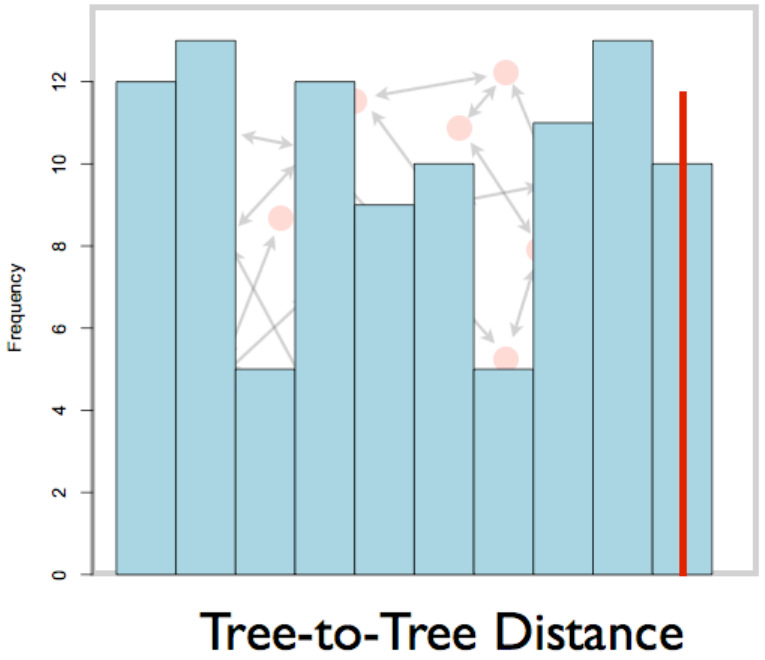
Tree Space

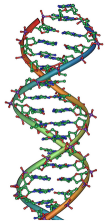


Tree Space

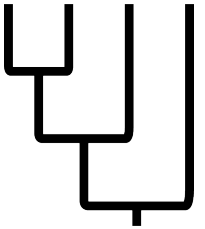
Topological Clustering Statistics

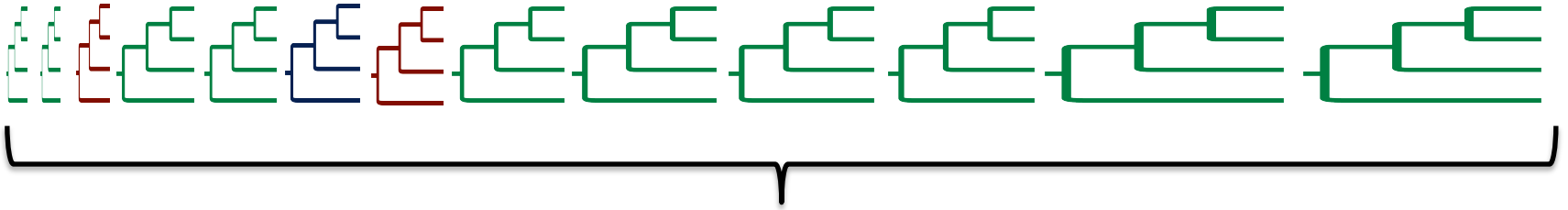






Marginal Test Quantities – Tree Length





Mean Tree Length = 3.15

Integrating across topologies (nuisance parameter)

Posterior Prediction

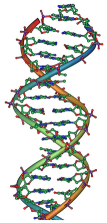
What kinds of inferences might we care about?

- Overall topological inference

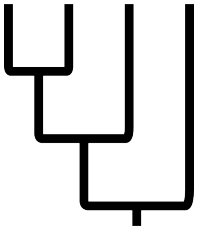
- Branch-specific support (posteriors)

- Tree (or branch) length

- Support from individual sequence positions (Identify specific biased sites)



Try It Out With AMP

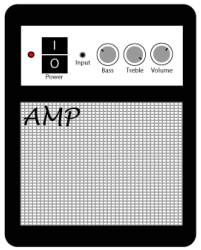


1. Perform Bayesian data analysis (e.g., MrBayes)

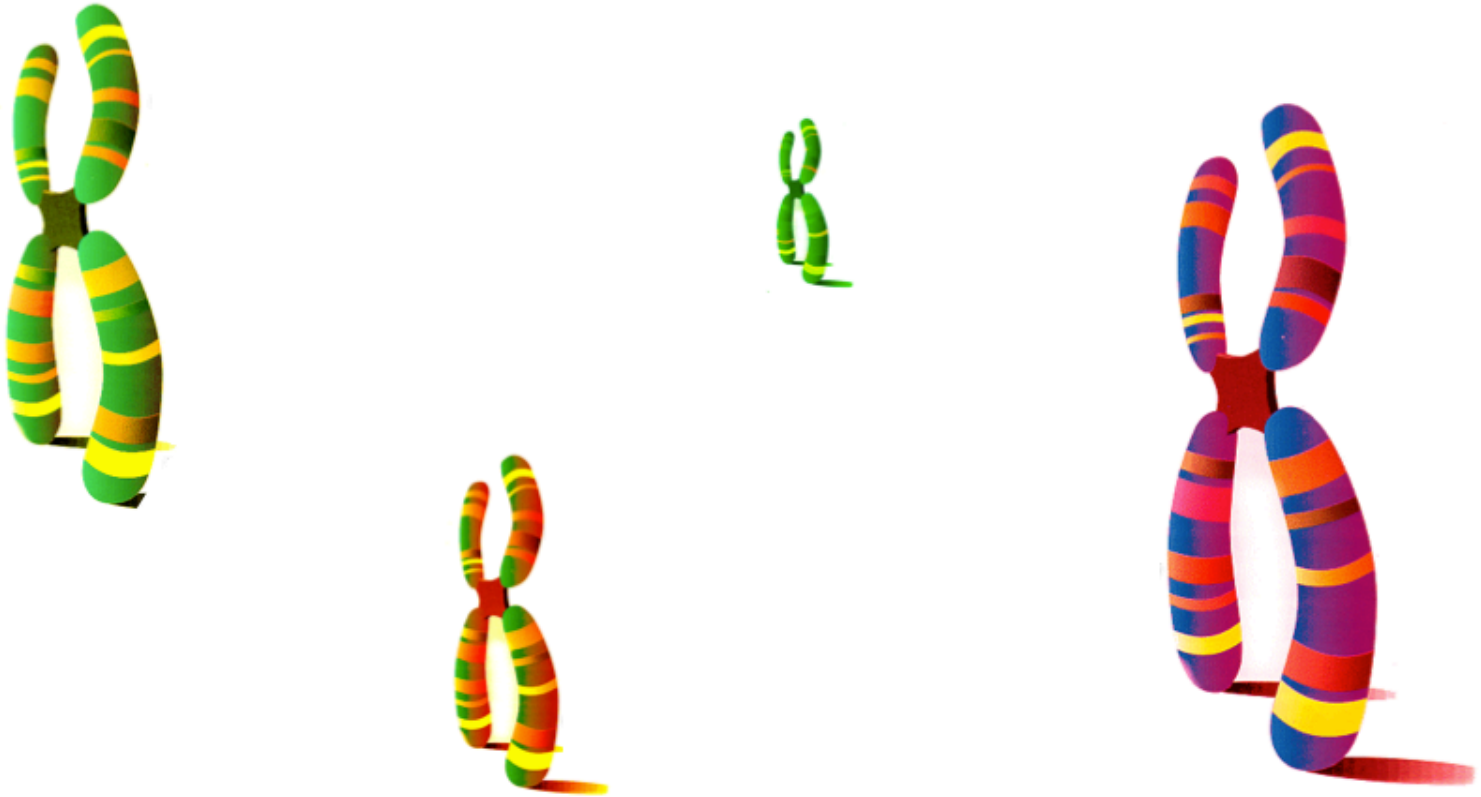
2. Simulate posterior predictive data (e.g., PuMA or MAPPS)

3. Analyze posterior predictive data sets (e.g., MrBayes)

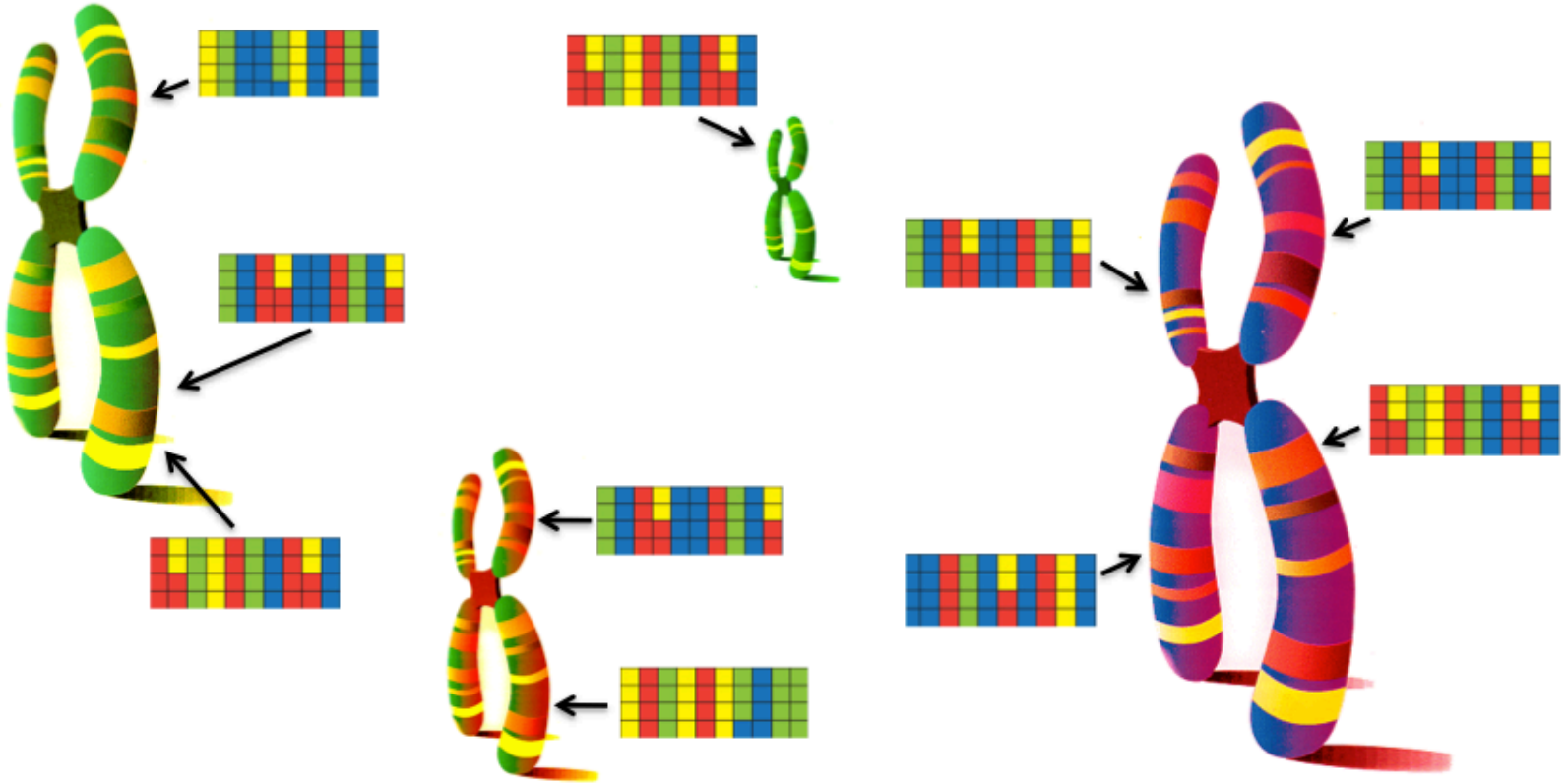
4. Calculate marginal posterior predictive *P*-values
   AMP (http://code.google.com/p/phylo-amp)



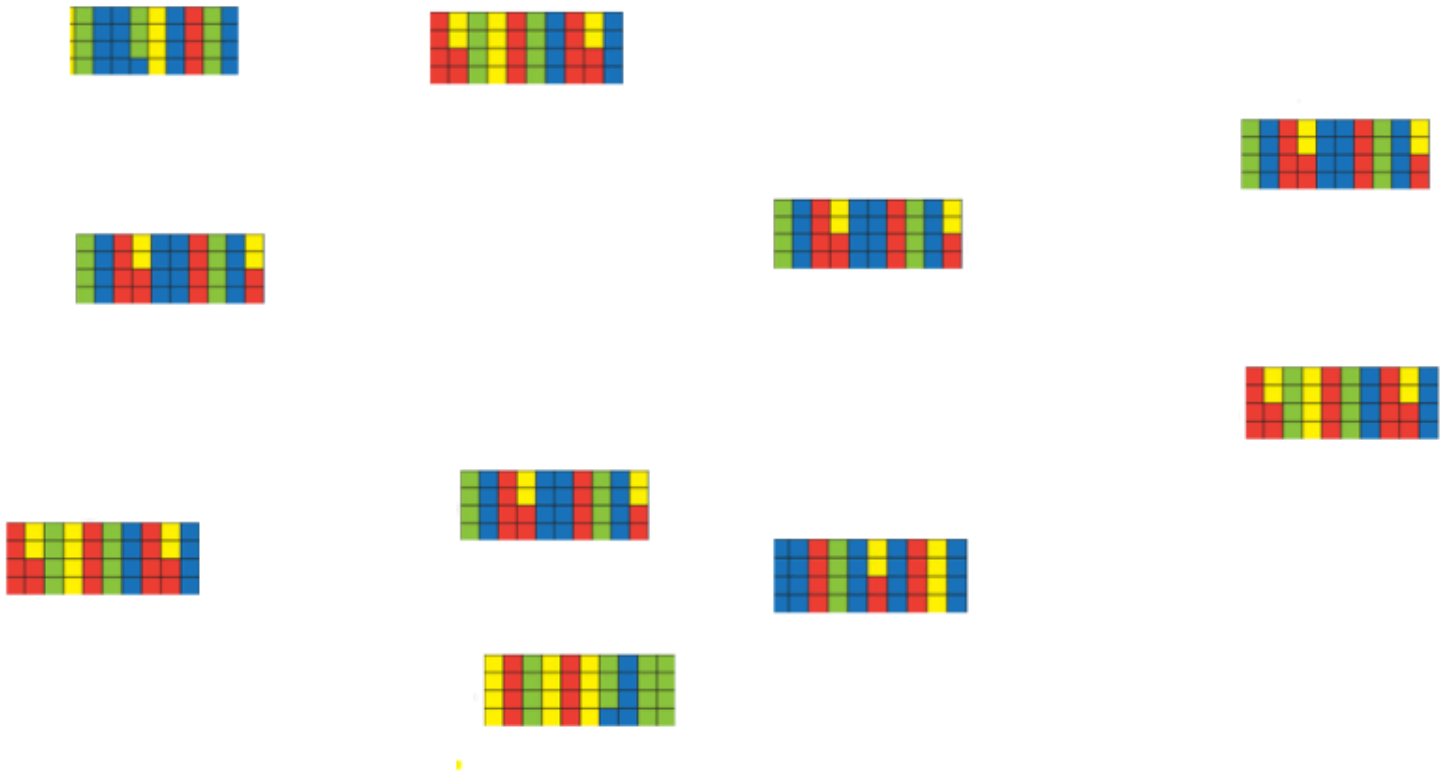
Bottom-Up Phylogenomics



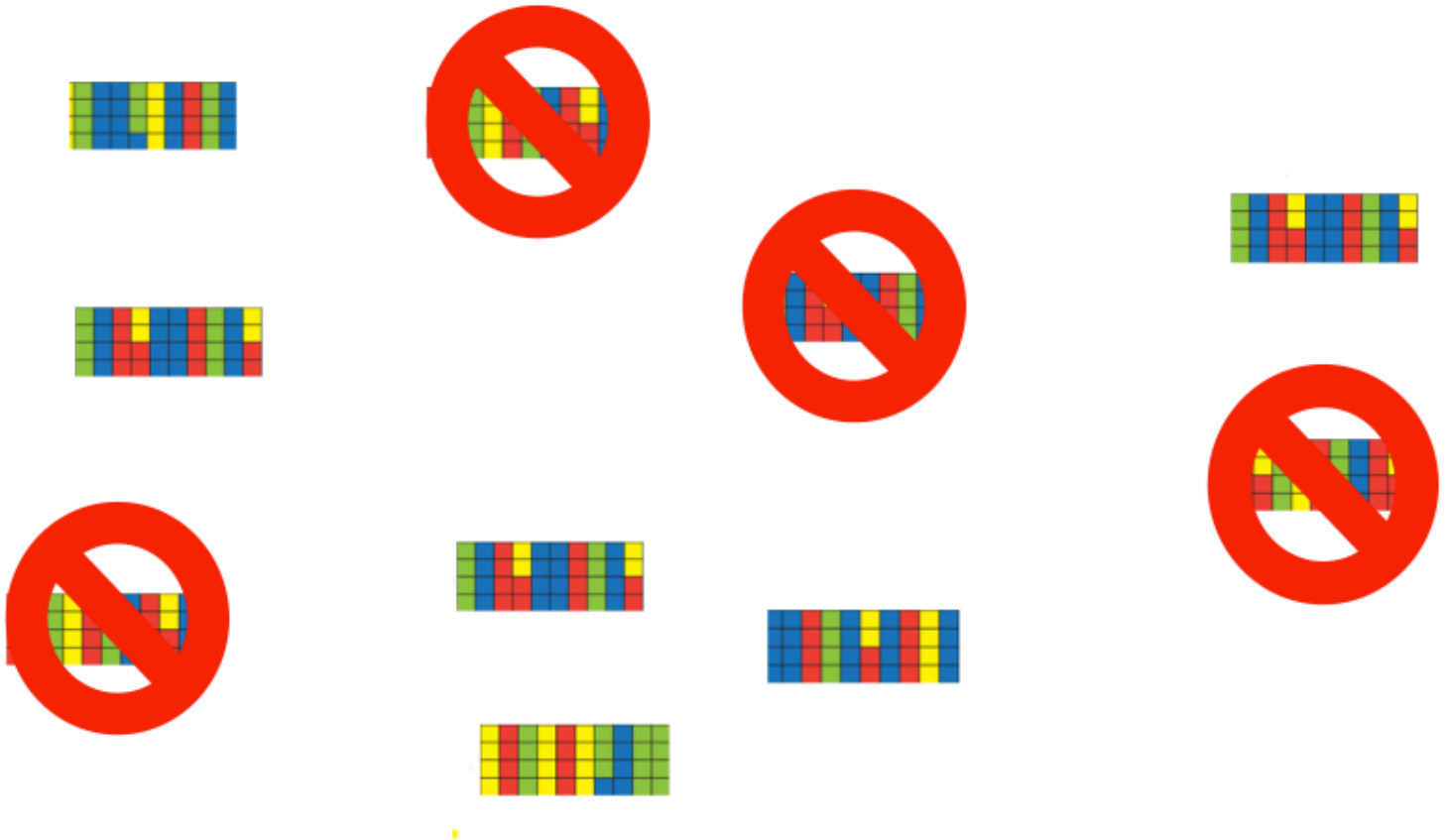
Bottom-Up Phylogenomics



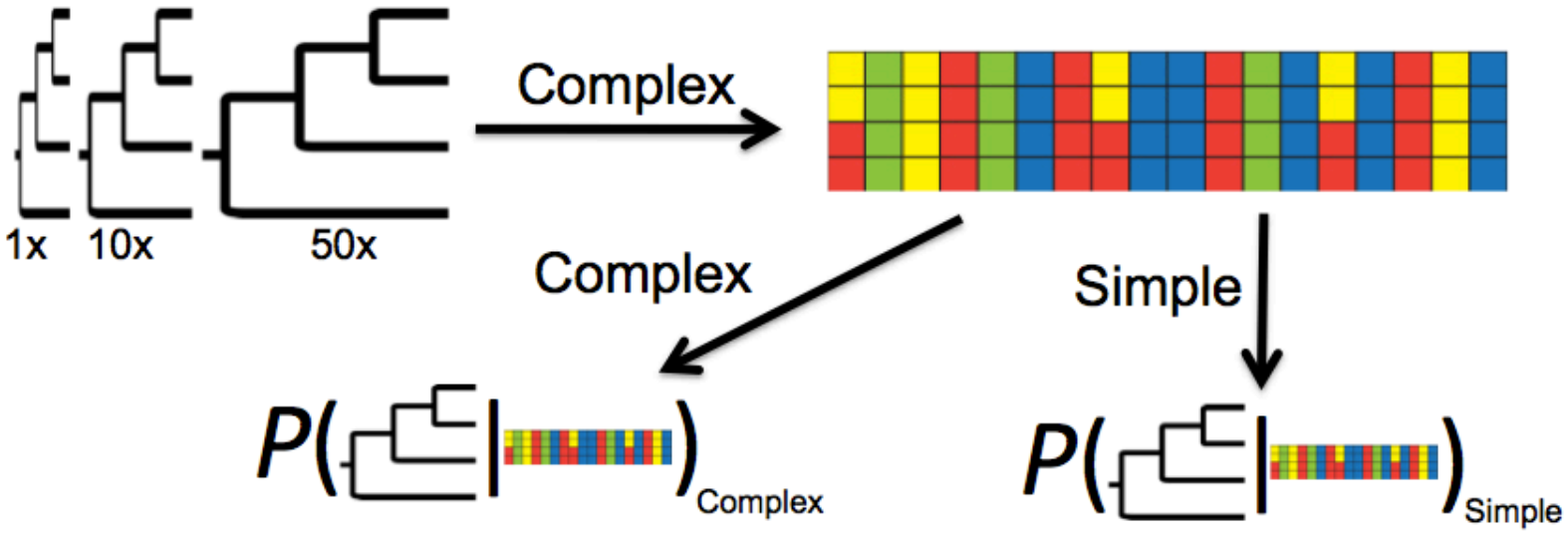
Bottom-Up Phylogenomics

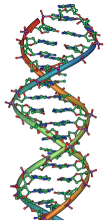


Bottom-Up Phylogenomics

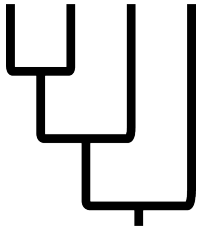


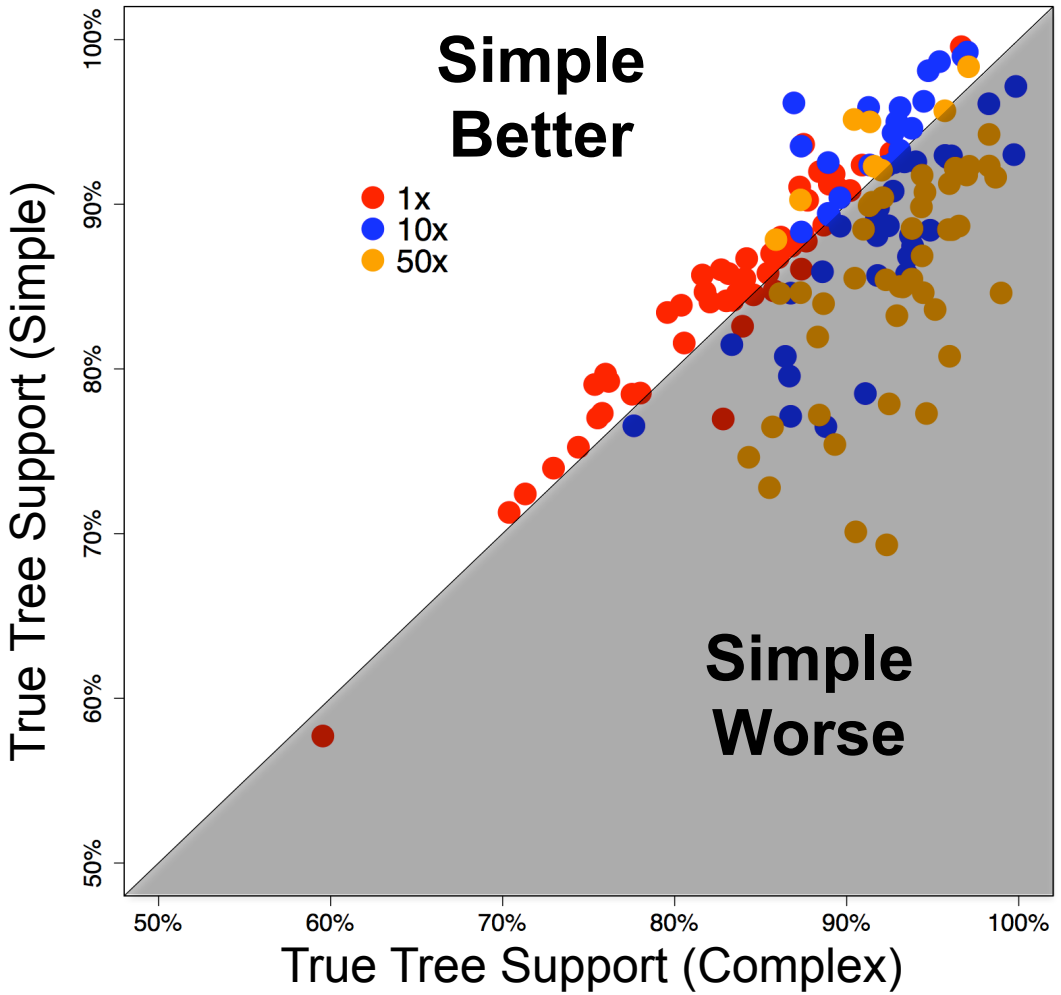
Simulation Test

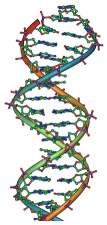




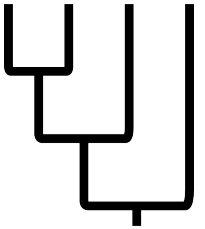
Simple Model Biased with Longer Trees

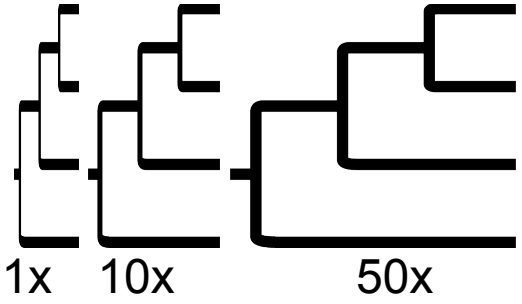






Testing Performance with Simulated Data

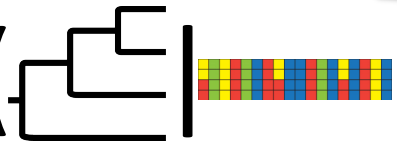




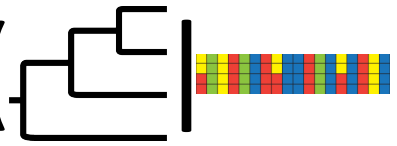
Complex

Complex

Simple

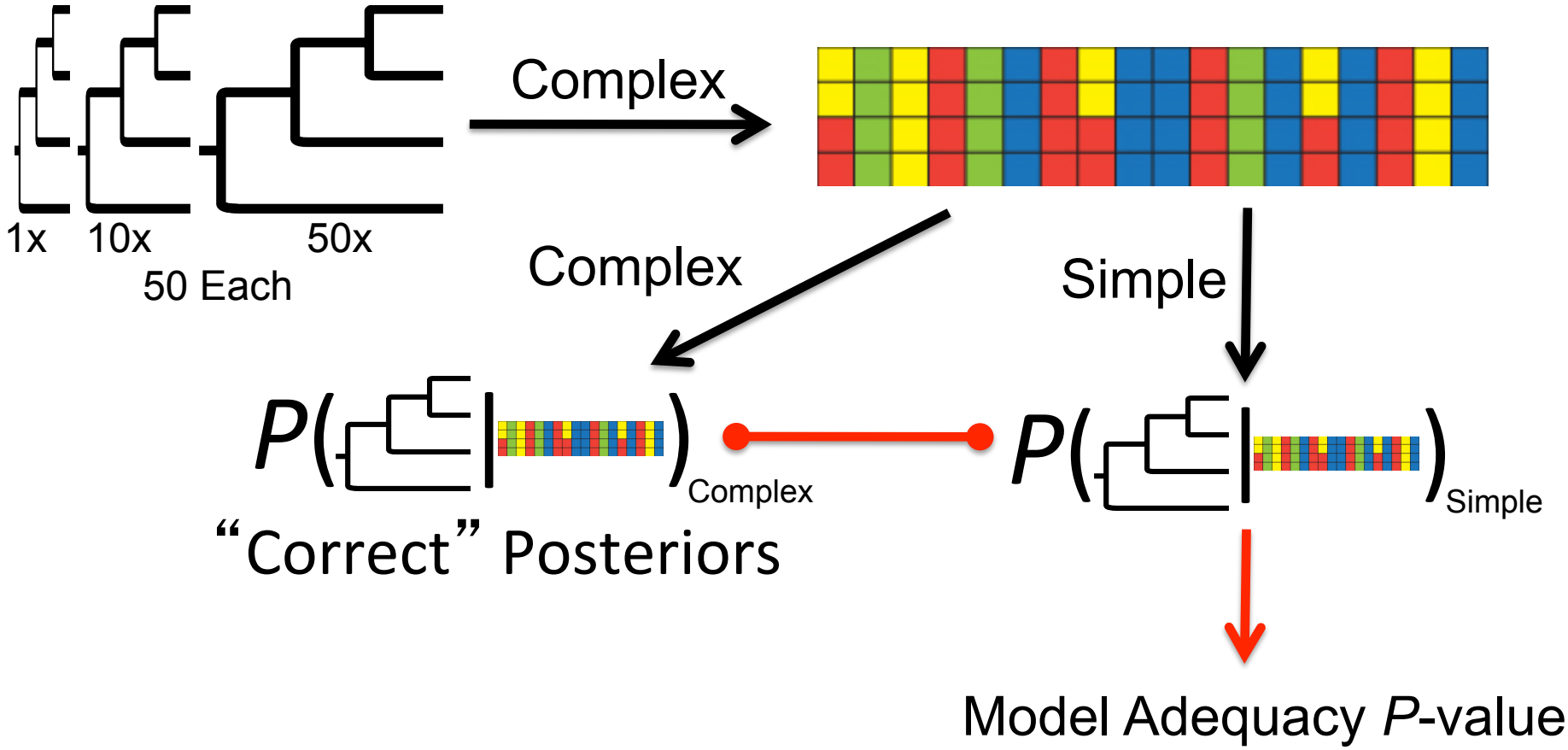
$P($  $)_{Complex}$

"Correct" Posteriors

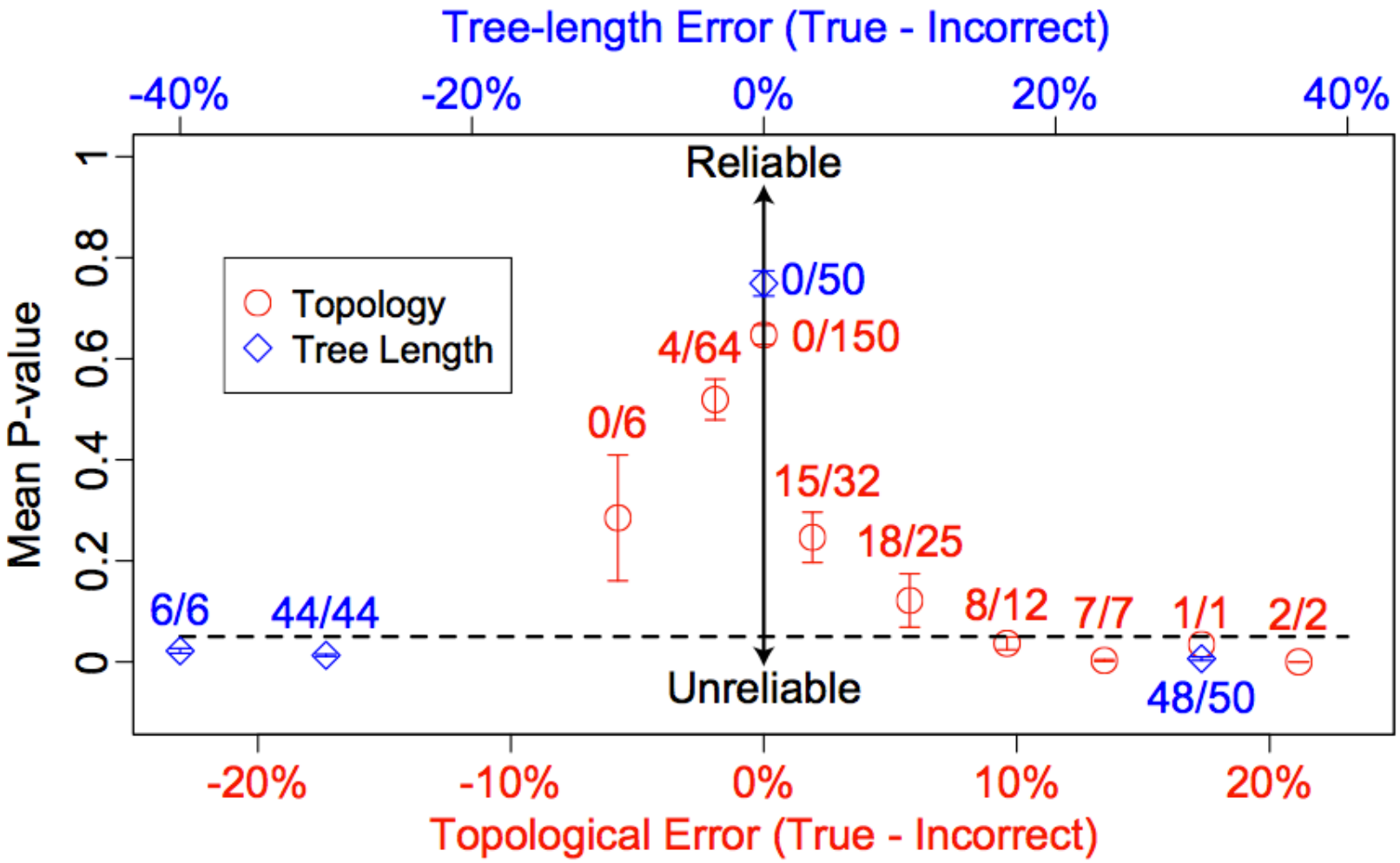
$P($  $)_{Simple}$

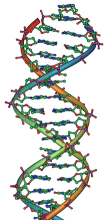
Model Adequacy *P*-value

Testing Performance with Simulated Data

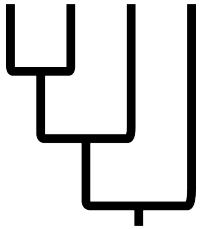


Simulation Test

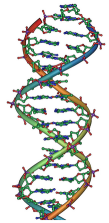




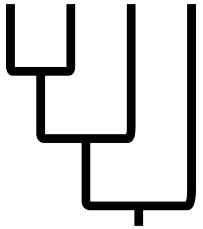
Treelength Test Quantity
Detects Biased Branch-Length Inference



- 50 replicate datasets
- Bipartition posteriors nearly identical to true branch-length prior
- Tree Lengths overestimated



Treelength Test Quantity Detects Biased Branch-Length Inference



- 50 replicate datasets
- Bipartition posteriors nearly identical to true branch-length prior
- Tree Lengths overestimated