Bayesian Phylogenetic Inference and RevBayes: Introduction

Fred(rik) Ronquist

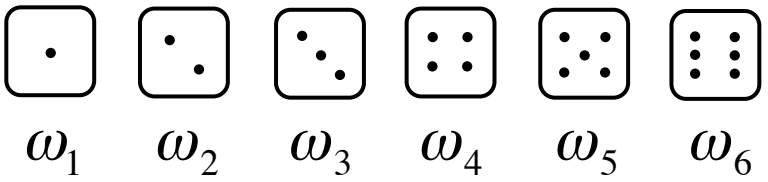
Swedish Museum of Natural History, Stockholm, Sweden

Topics

- Probability 101
- Bayesian Phylogenetic Inference
- Graphical Models

1. Probability 101

Random variable *X*



ω_1 ω_2 ω_3 ω_4 ω_5 ω_6

Sample space

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_k\}$$

$\mathrm{m}(\omega)$

Distribution function
(E.g., uniform)

$$E = \{\omega_1, \omega_3, \omega_5\}$$

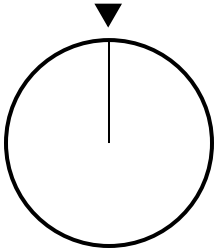
Event (subset of outcomes;
e.g., face with odd number)

$$\Pr(E) = \sum_{\omega \in \mathrm{E}} \mathrm{m}(\omega)$$

Probability

Random variable X

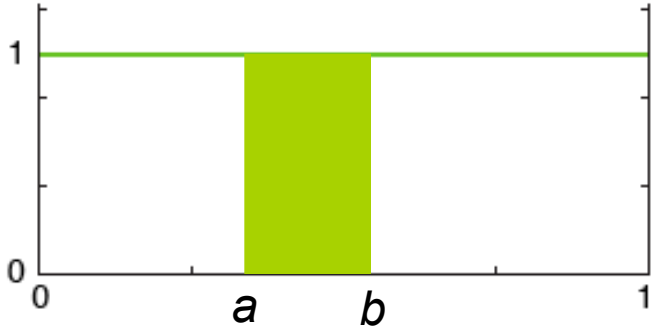
Disc with circumference 1



$\Omega = [0,1)$

Sample space (an interval)

$f(x)$



Probability density function (pdf) (e.g. Uniform(0,1))

$E = [a,b)$

Event (a subspace of the sample space)

$$\Pr(E) = \int_{x \in E} f(x)\,\mathrm{d}x$$

Probability

Continuous Distributions

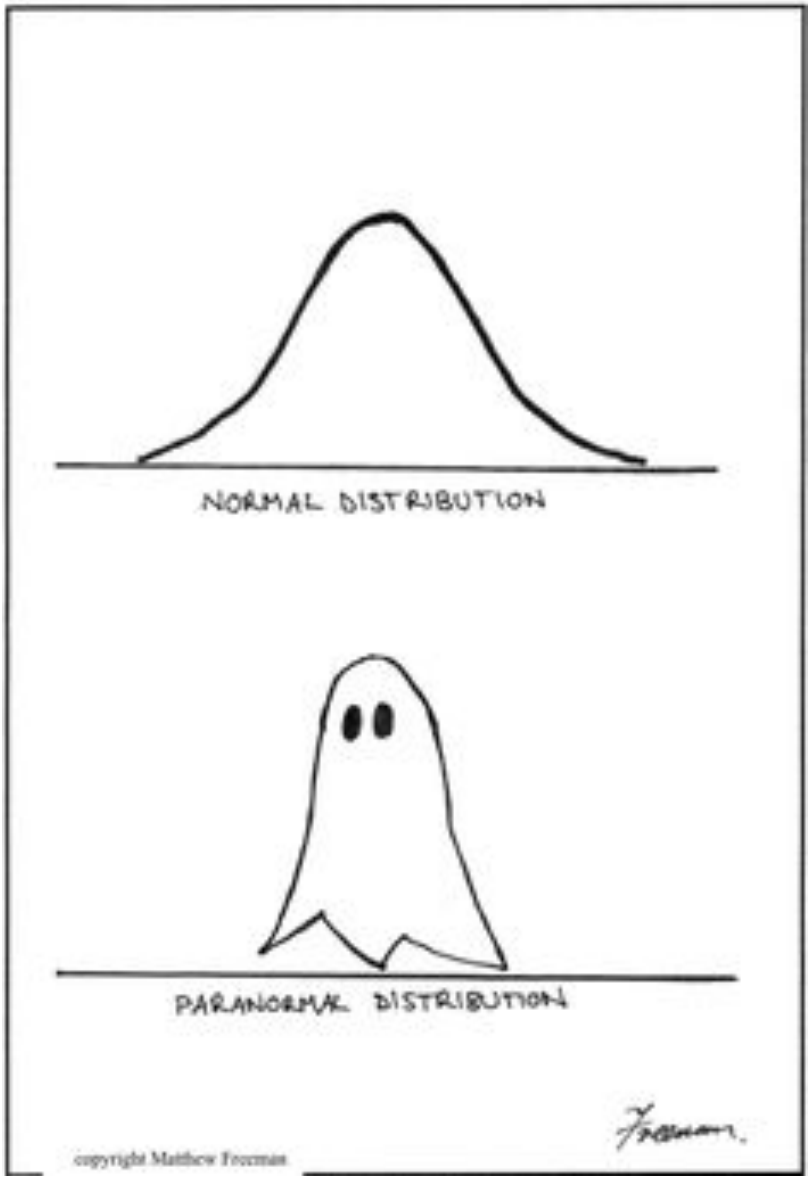
- Uniform distribution
- Beta distribution
- Gamma distribution
- Dirichlet distribution
- Exponential distribution
- Normal distribution
- Lognormal distribution
- Multivariate normal distribution

Discrete Distributions

- Bernoulli distribution
- Categorical distribution
- Binomial distribution
- Multinomial distribution
- Poisson distribution

Stochastic Processes

- Markov chain
- Poisson process
- Birth-death process
- Coalescence
- Dirichlet Process Mixture

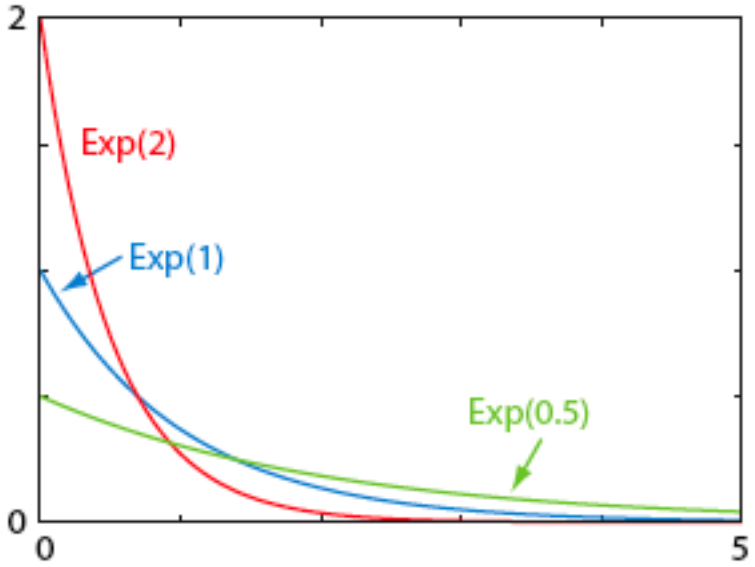

NORMAL DISTRIBUTION
PARANORMAL DISTRIBUTION
copyright Matthew Freeman

Exponential distribution $X \sim \text{Exp}(\lambda)$

Parameters: λ = rate (of decay)

Probability density function: $f(x) = \lambda e^{-\lambda x}$

Mean: $1/\lambda$



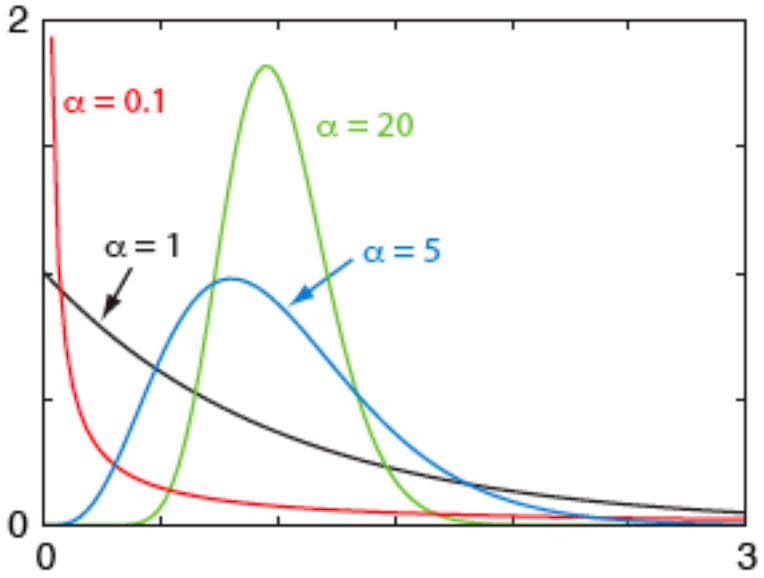
Gamma distribution $X \sim \text{Gamma}(\alpha, \beta)$

Parameters: α = shape β = inverse scale

Probability density function: $f(x) \propto x^{\alpha-1} e^{-\beta x}$

Mean: α / β

Scaled gamma: $\alpha = \beta$



Scaled Gamma

Beta Distribution

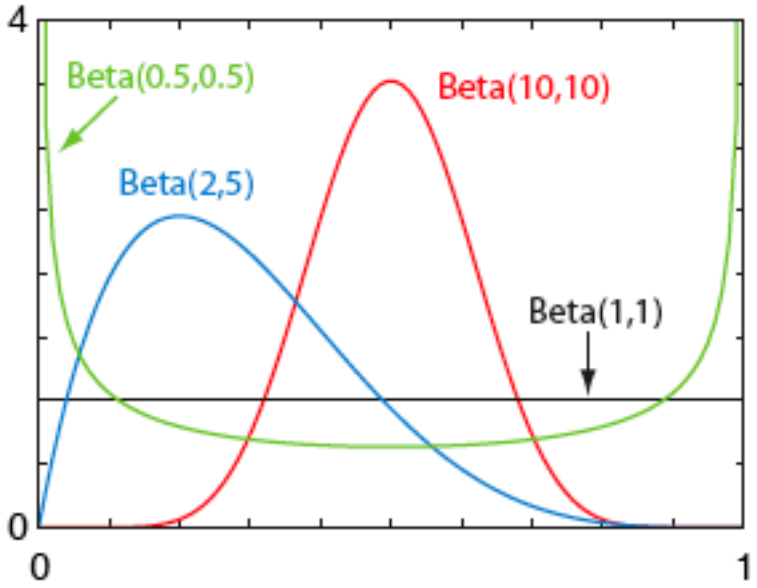
Beta distribution $X \sim \text{Beta}(\alpha_1, \alpha_2)$

Parameters: α_1, α_2 = shape parameters

Probability density function: $f(x) \propto x^{\alpha_1 - 1}(1-x)^{\alpha_2 - 1}$

Mode: $\dfrac{\alpha_1 - 1}{\sum_i (\alpha_i - 1)}$

Defined on two proportions of a whole (a simplex)



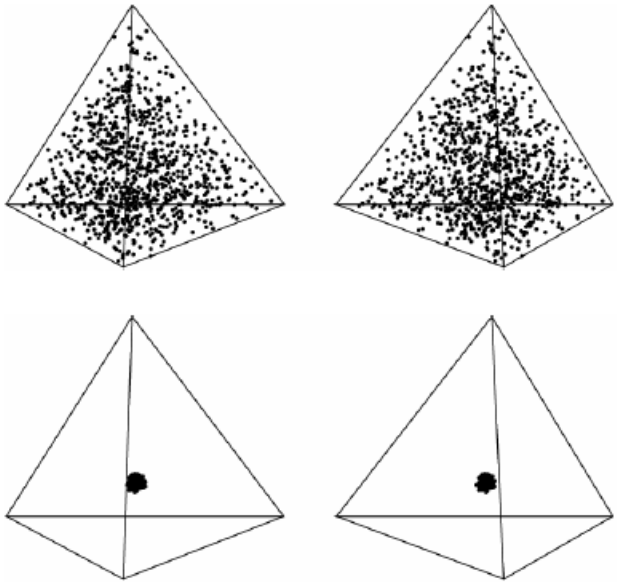
Dirichlet Distribution

Dirichlet distribution $X \sim \mathrm{Dir}(\alpha) : \alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_k\}$

Parameters: α = vector of *k* shape parameters

Probability density function: $f(x) \propto \prod_i x_i^{\alpha_i - 1}$

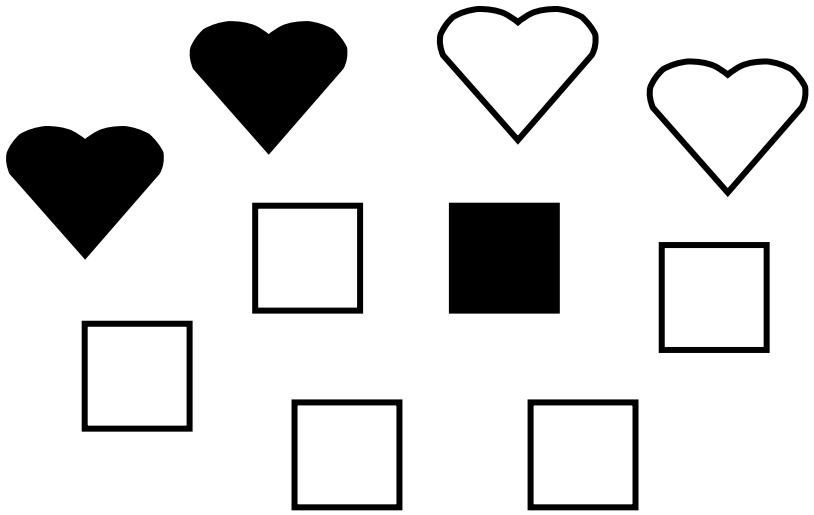
Defined on *k* proportions of a whole



Dir(1,1,1,1)

Dir(300,300,300,300)

Conditional Probability



$$\Pr(H) = \frac{4}{10} = 0.4$$

$$\Pr(D) = \frac{3}{10} = 0.3$$

Joint probability: $\Pr(D,H) = \frac{2}{10} = 0.2$

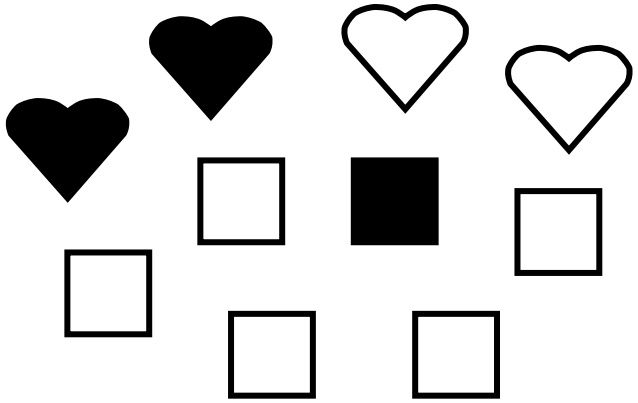
Conditional probability: $\Pr(D \mid H) = \frac{2}{4} = 0.5$

Bayes' Rule

Reverend Thomas Bayes
(1701-1760)

$$\Pr(A \mid B) \Rightarrow \Pr(B \mid A) \ ?$$

Bayes' Rule



$$\Pr(D,H) = \Pr(D)\Pr(H \mid D) = \frac{3}{10} \times \frac{2}{3} = \frac{2}{10} = 0.2$$

$$= \Pr(H)\Pr(D \mid H) = \frac{4}{10} \times \frac{2}{4} = \frac{2}{10} = 0.2$$

$$\Pr(D)\Pr(H \mid D) = \Pr(H)\Pr(D \mid H)$$

$$\boxed{\Pr(H \mid D) = \frac{\Pr(H)\Pr(D \mid H)}{\Pr(D)}}$$

Bayes' rule

Maximum Likelihood Inference

Data *D*; Model *M* with parameters θ

We can calculate $\Pr(D \mid \theta)$ or $f(D \mid \theta)$

Define the likelihood function $\mathrm{L}(\theta) \propto f(D \mid \theta)$

Maximum likelihood: find the value of θ that maximizes L(θ)

Confidence: asymptotic behavior, more samples, bootstrapping

Bayesian Inference

Data *D*; Model *M* with parameters θ

We can calculate $\Pr(D \mid \theta)$ or $f(D \mid \theta)$

We are actually interested in $\Pr(\theta \mid D)$ or $f(\theta \mid D)$

Bayes' rule:

Posterior density

Prior density

"Likelihood"

$$f(\theta \mid D) = \frac{f(\theta) f(D \mid \theta)}{f(D)}$$

Normalizing constant

Marginal likelihood of the data

Model likelihood

$$f(D) = \int f(\theta) f(D \mid \theta) \, \mathrm{d}\theta$$

Coin Tossing Example

DID THE SUN JUST EXPLODE?

(IT'S NIGHT, SO WE'RE NOT SURE.)

FREQUENTIST STATISTICIAN:

BAYESIAN STATISTICIAN:

Four Nations 2014: New Zealand champions after defeating Australia 22-18 in Wellington

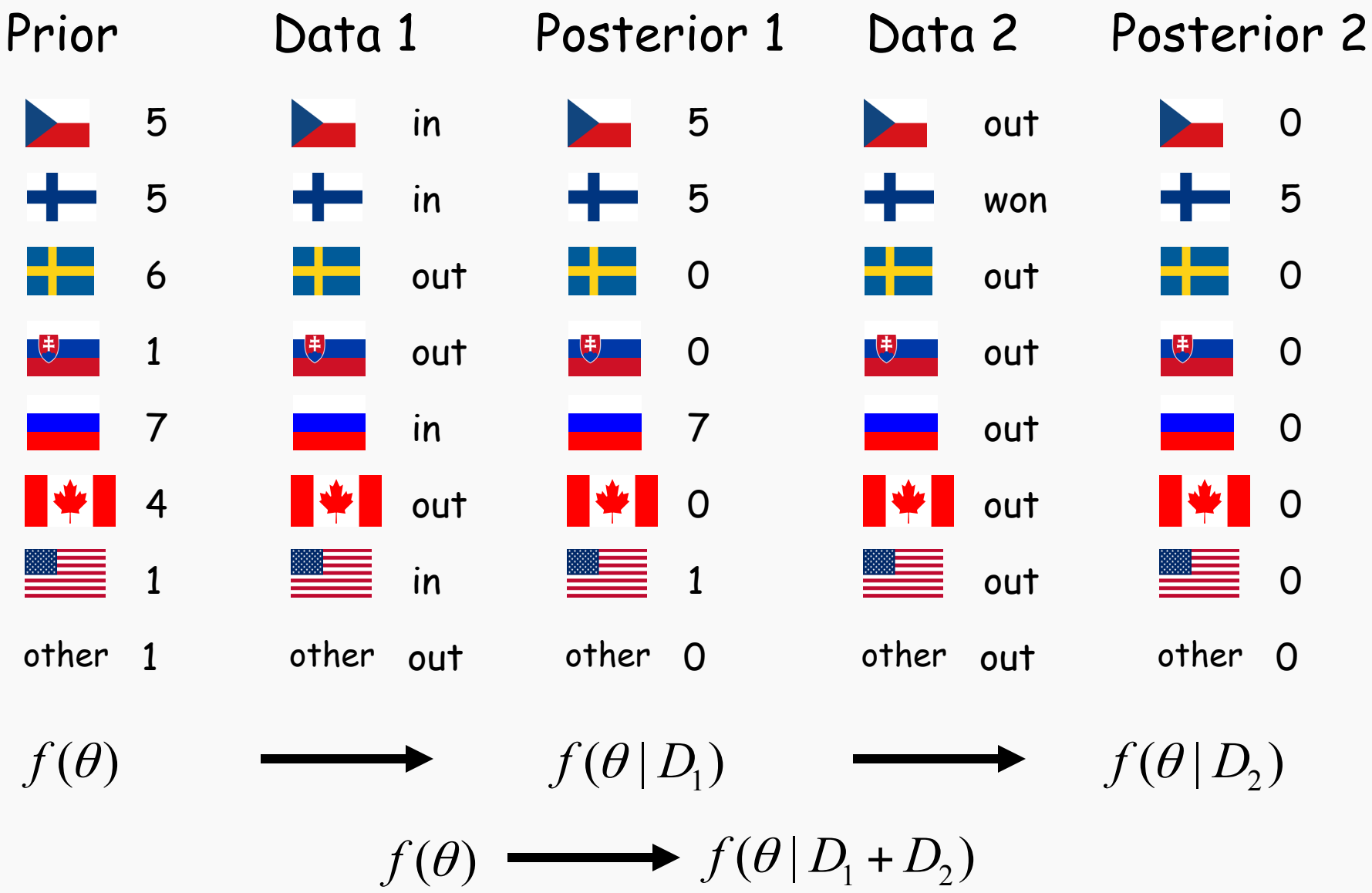
What is the probability of your favorite team winning the next ice hockey World Championships?

World Championship Medalists

	Gold	Silver	Bronze
2005			
2006			
2007			
2008			
2009			
2010			
2011			
2012			
2013			
2014			

Medals

	5
	5
	6
	1
	7
	4
	1
other	1

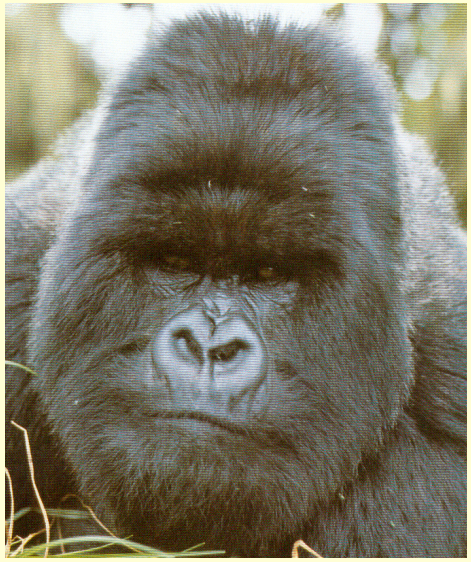

Prior
Data 1
Posterior 1
Data 2
Posterior 2
5
in
5
out
0
5
in
5
won
5
6
out
0
out
0
1
out
0
out
0
7
in
7
out
0
4
out
0
out
0
1
in
1
out
0
other 1
other out
other 0
other out
other 0
$f(\theta)$
$f(\theta \mid D_1)$
$f(\theta \mid D_2)$
$f(\theta)$
$f(\theta \mid D_1 + D_2)$

Learn more:

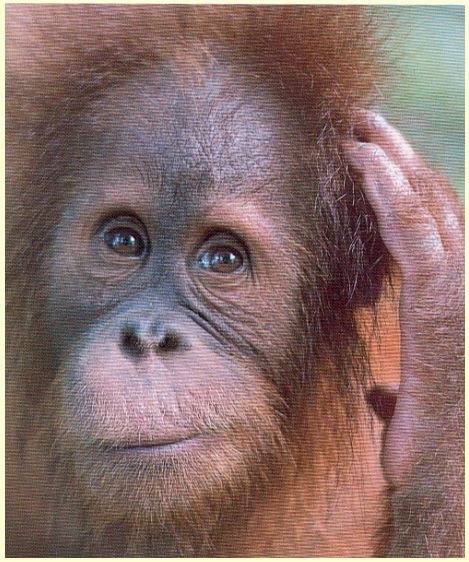
- Wikipedia (good texts on most statistical distributions, sometimes a little difficult)
- Grinstead & Snell: Introduction to Probability. American Mathematical Society. Free pdf available from: http://www.dartmouth.edu/~chance/teaching_aids/books_articles/probability_book/amsbook.mac.pdf
- Team up with a statistician or a computational / theoretical evolutionary biologist!

2. Bayesian Phylogenetic Inference

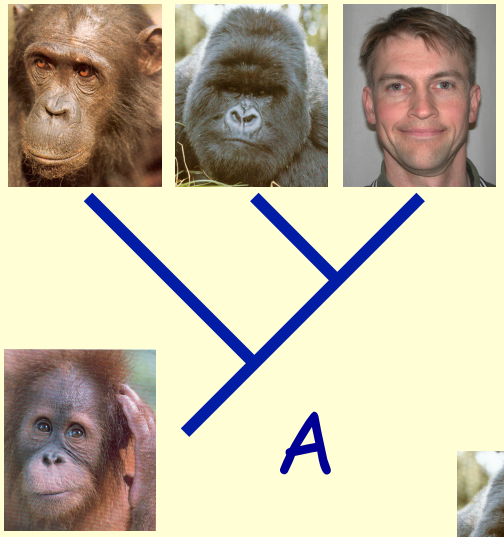
Infer relationships among three species:



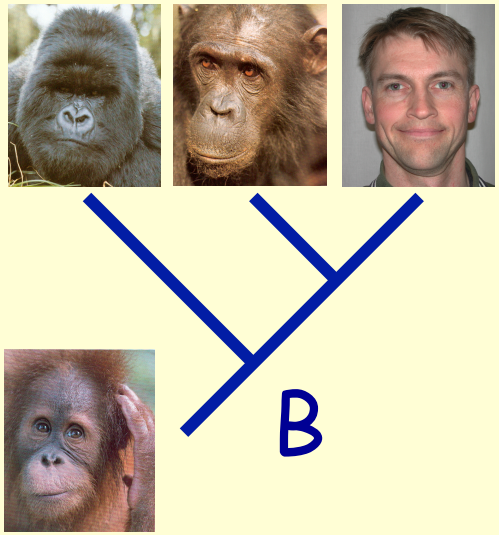
Outgroup:



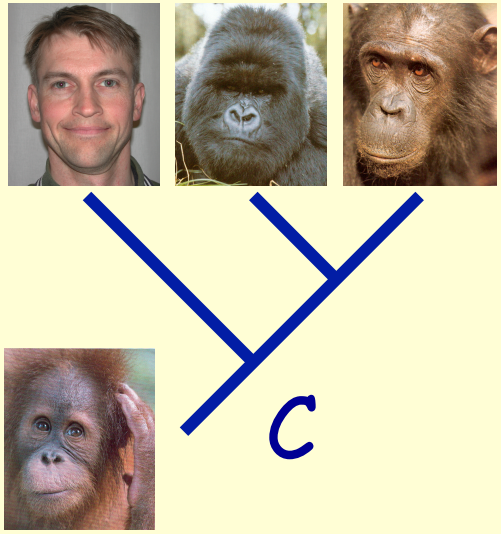
Three possible trees (topologies):



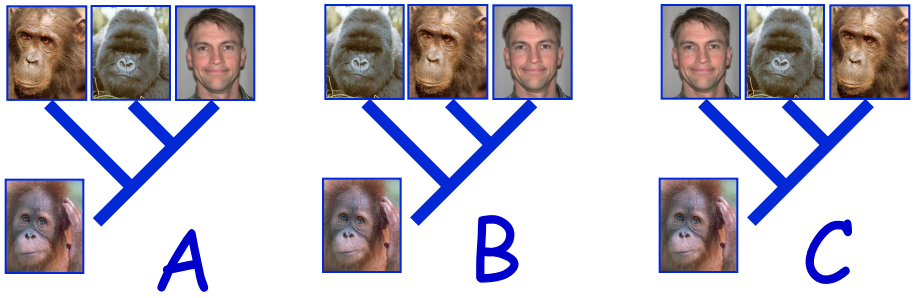
A



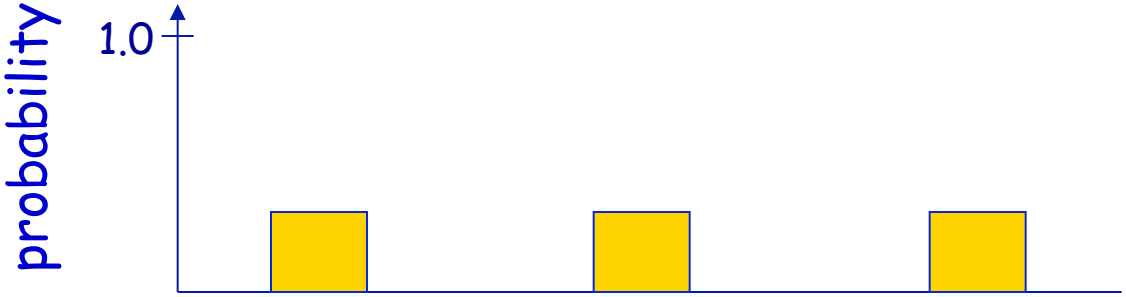
B



C

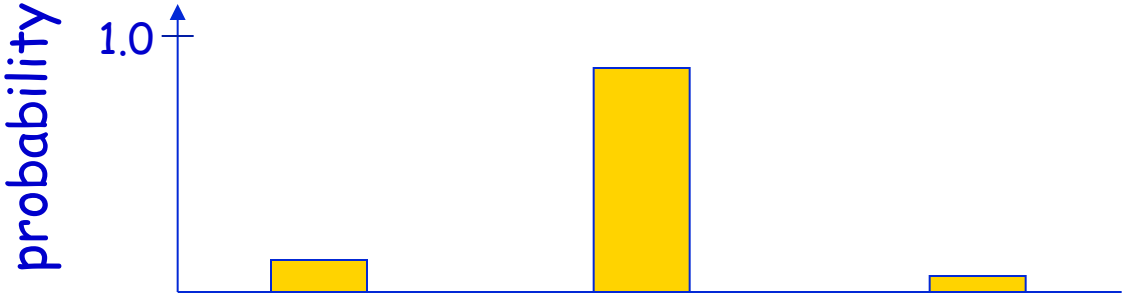


Model



Prior distribution

Data (observations)



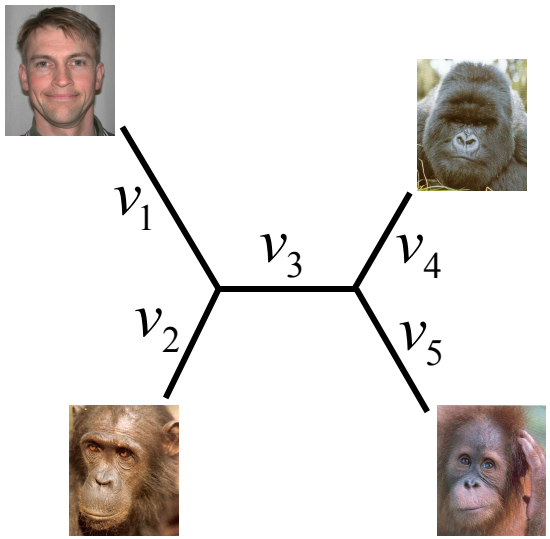
Posterior distribution

D The data

Taxon	Characters
	ACG TTA TTA AAT TGT CCT CTT TTC AGA
	ACG TGT TTC GAT CGT CCT CTT TTC AGA
	ACG TGT TTA GAC CGA CCT CGG TTA AGG
	ACA GGA TTA GAT CGT CCG CTT TTC AGA

Model: topology AND branch lengths

θ Parameters



topology (τ)

branch lengths (v_i)
(expected amount of change)

$$\theta = (\tau, v)$$

Model: molecular evolution

θ Parameters

$$Q = \begin{pmatrix} & [A] & [C] & [G] & [T] \\ [A] & - & \mu & \mu & \mu \\ [C] & \mu & - & \mu & \mu \\ [G] & \mu & \mu & - & \mu \\ [T] & \mu & \mu & \mu & - \end{pmatrix}$$

instantaneous rate matrix
(Jukes-Cantor)

Model: molecular evolution

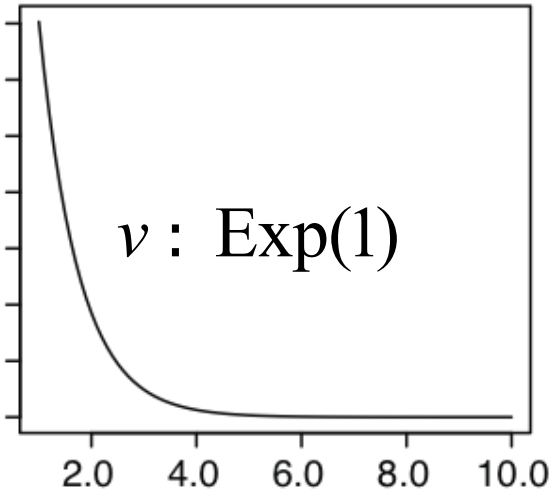
Probabilities are calculated using the transition probability matrix P

$$P(v) = e^{Qv} = \begin{cases} \frac{1}{4} - \frac{1}{4}e^{-4v/3} & \text{(change)} \\ \frac{1}{4} + \frac{3}{4}e^{-4v/3} & \text{(no change)} \end{cases}$$

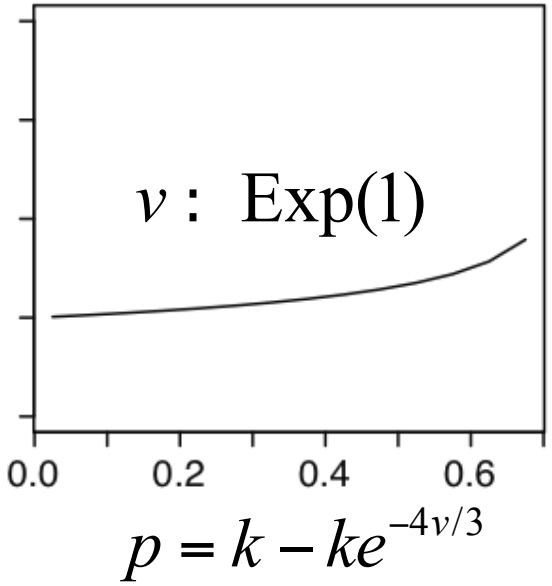
Priors on parameters

- Topology
  - all unique topologies have equal probability
- Branch lengths
  - exponential prior (puts more weight on small branch lengths)

Scale matters in priors



Branch length



Prob. of substitution

The effect on data likelihood is most important

Jeffrey's uninformative priors formalize this

Bayes' theorem

D = Data
θ = Model parameters

Posterior distribution

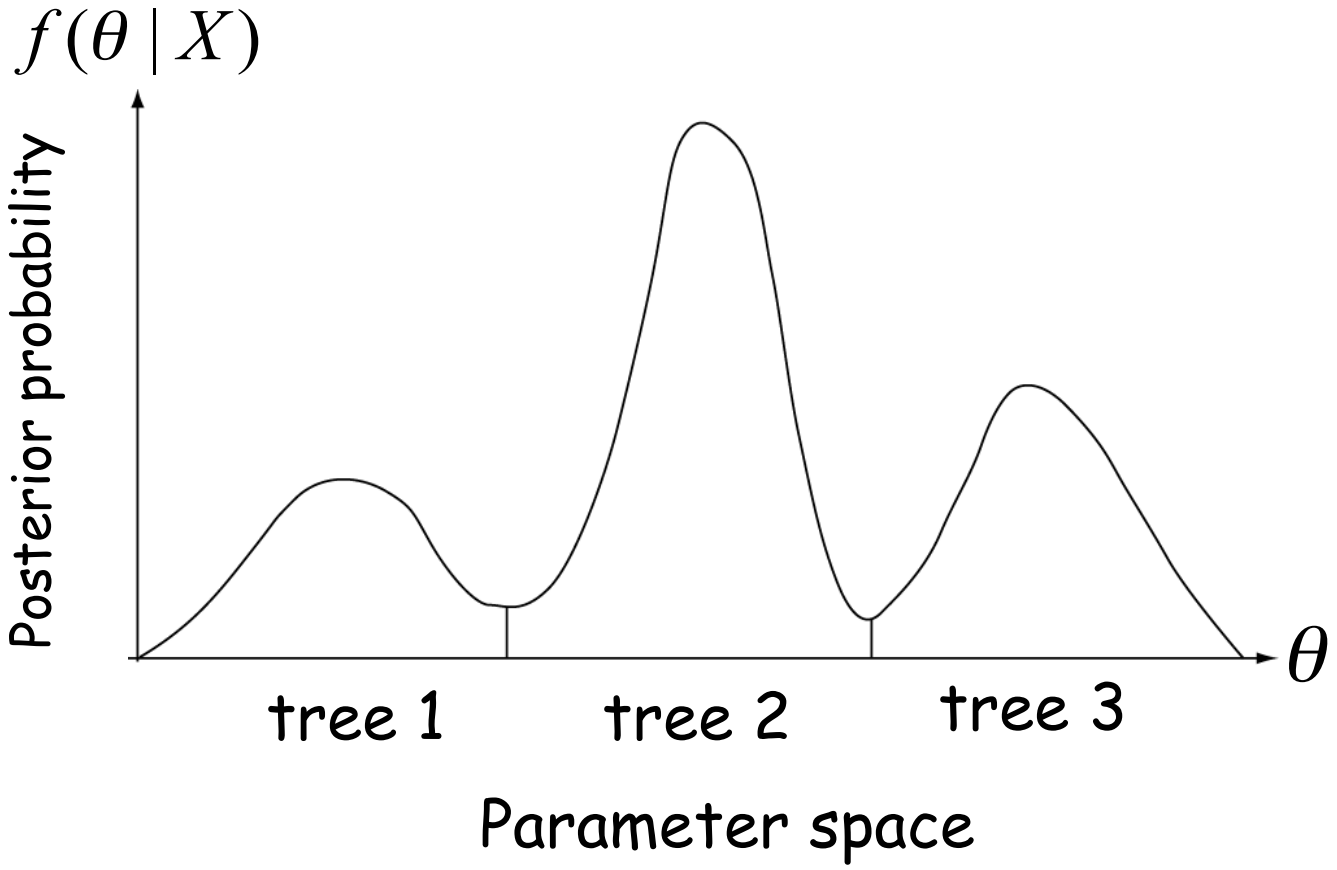
Prior distribution

"Likelihood"

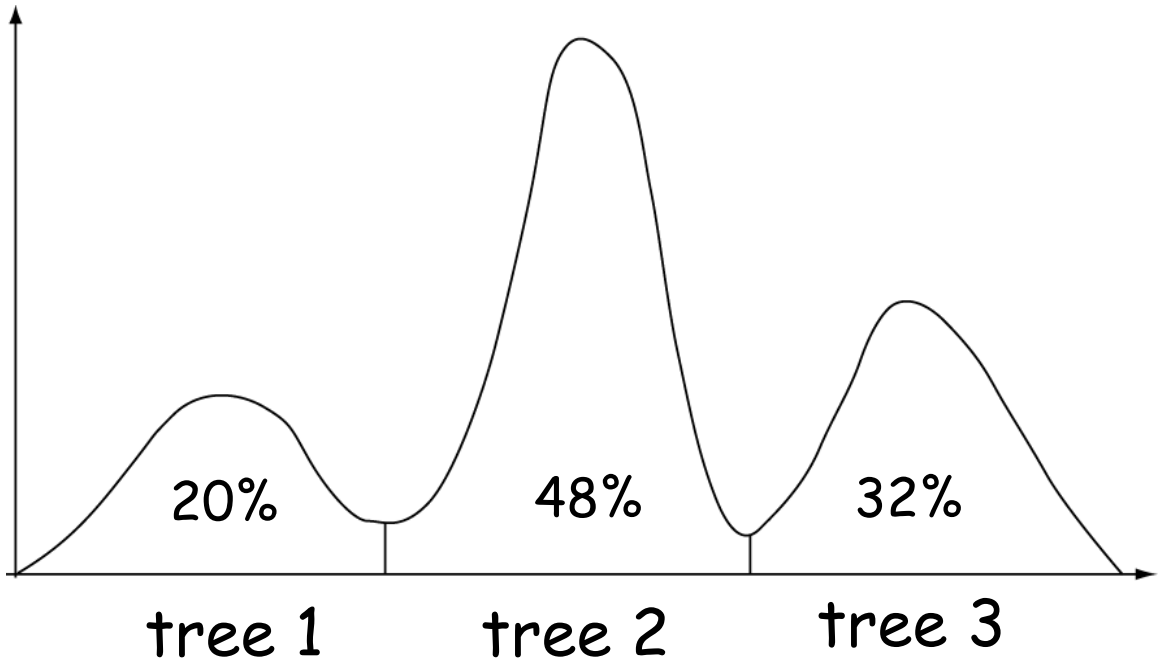
$$f(\theta \mid D) = \frac{f(\theta) f(D \mid \theta)}{\int f(\theta) f(D \mid \theta)\, \mathrm{d}\theta}$$

Normalizing constant

Posterior probability distribution

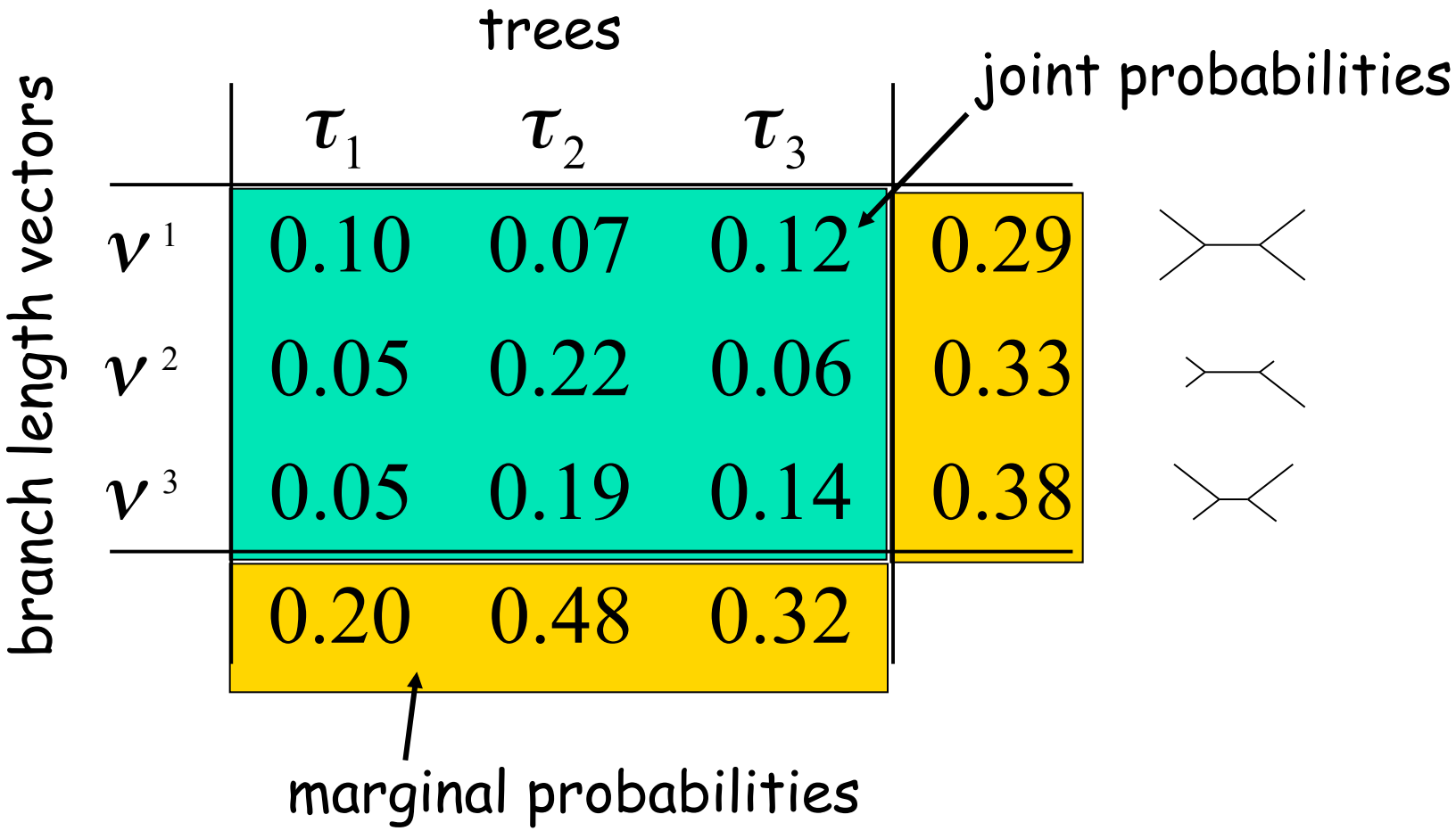


We can focus on any parameter of interest (there are no nuisance parameters) by marginalizing the posterior over the other parameters (integrating out the uncertainty in the other parameters)



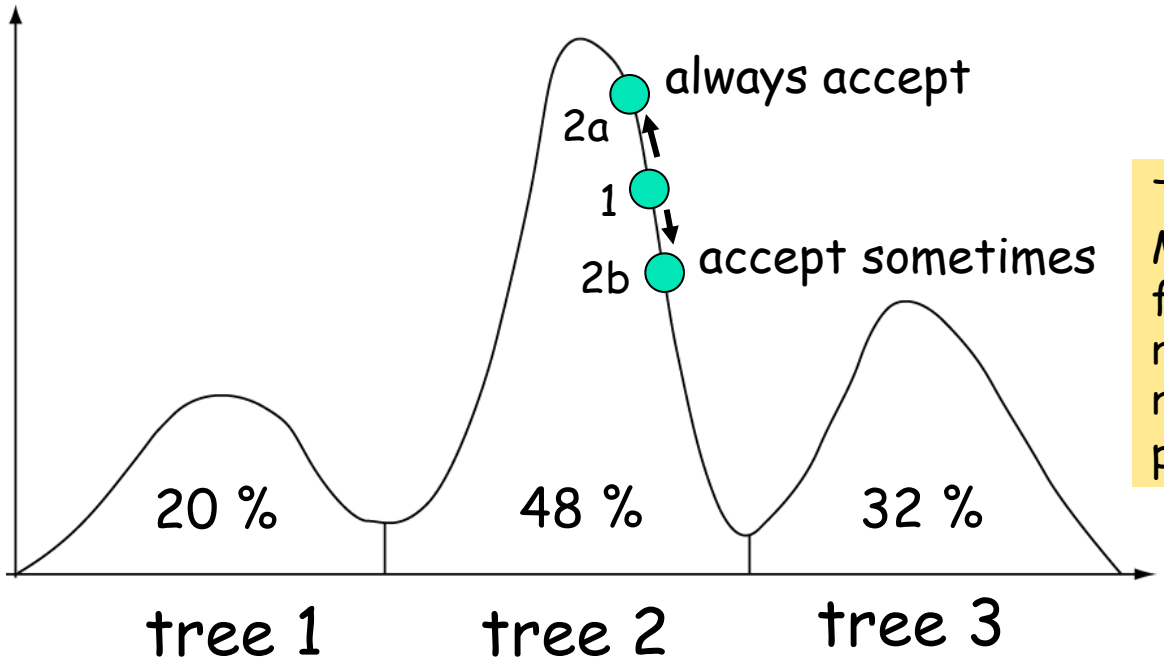
(Percentages denote marginal probability distribution on trees)

Why is it called marginalizing?



Markov chain Monte Carlo

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



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

Metropolis algorithm

Assume that the current state has parameter values θ

Consider a move to a state with parameter values θ^*

The height ratio *r* is

$$r = \frac{f(\theta^* \mid D)}{f(\theta \mid D)} = \frac{f(\theta^*) f(D \mid \theta^*) / f(D)}{f(\theta) f(D \mid \theta) / f(D)} = \frac{f(\theta^*)}{f(\theta)} \times \frac{f(D \mid \theta^*)}{f(D \mid \theta)}$$

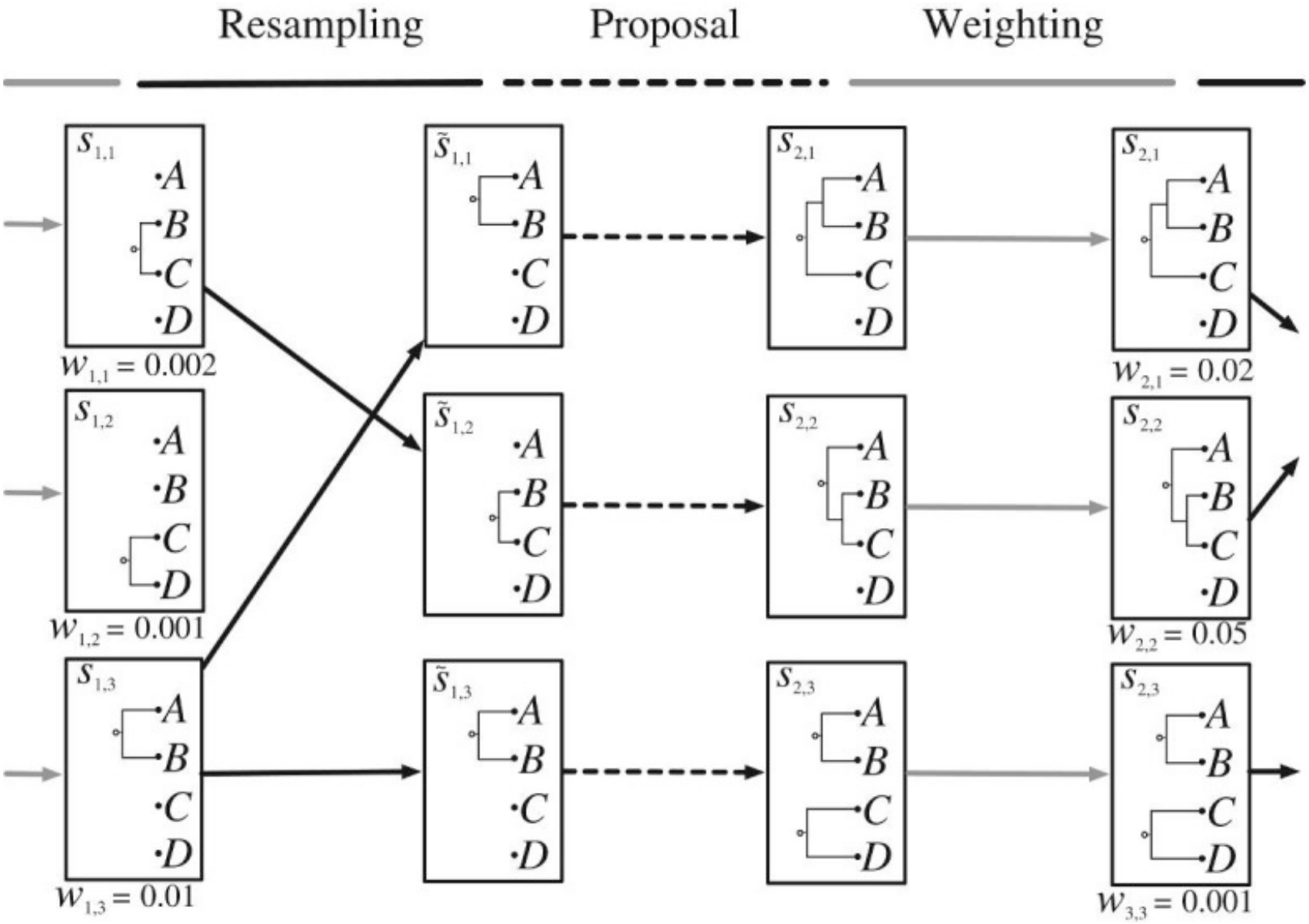
(prior ratio x likelihood ratio)

MCMC Sampling Strategies

- Great freedom of strategies:
  - Typically one or a few related parameters changed at a time
  - You can cycle through parameters systematically or choose randomly
  - One "generation" or "iteration" or "cycle" can include a single randomly chosen proposal (or move, operator, kernel), one proposal for each parameter, a block of randomly chosen proposals

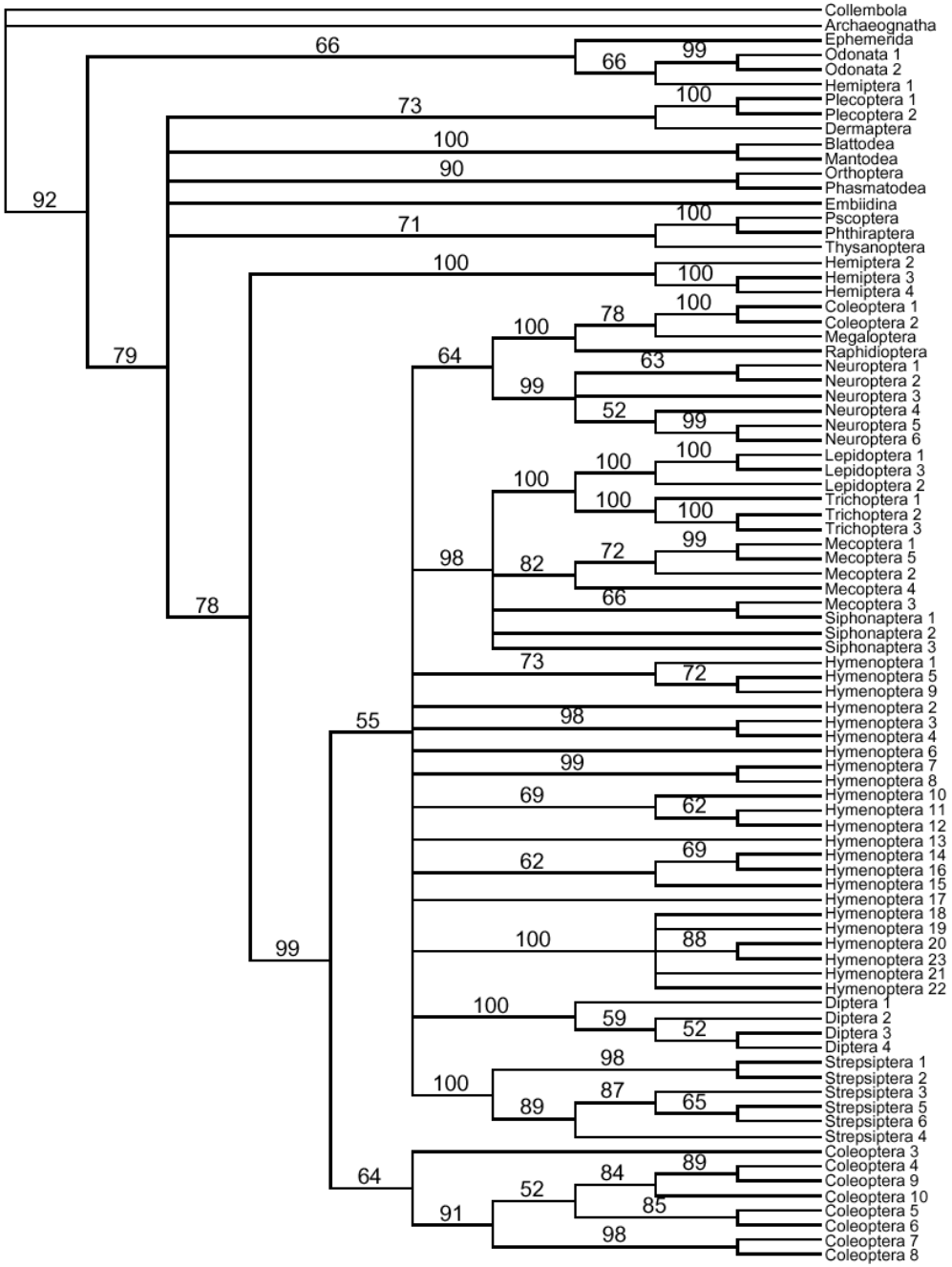
Other Sampling Methods

- **Gibbs sampling**: sample from the conditional posterior (a variant of the Metropolis algorithm)
- **Metropolized Gibbs sampling**: more efficient variant of Gibbs sampling of discrete characters
- **Slice sampling**: less prone to get stuck in local optima than the Metropolis algorithm
- **Hamiltonian sampling**. A technique for decreasing the problem with sampling correlated parameters.
- **Simulated annealing**: increase "greediness" during the burn-in phase of MCMC sampling
- **Data augmentation techniques**: add parameters to facilitate probability calculations
- **Sequential Monte Carlo techniques**: generate a sample of complete state by building sets of particles from incomplete states



Sequential Monte Carlo Algorithm for Phylogenetics

Bouchard et al. 2012. Syst. Biol.

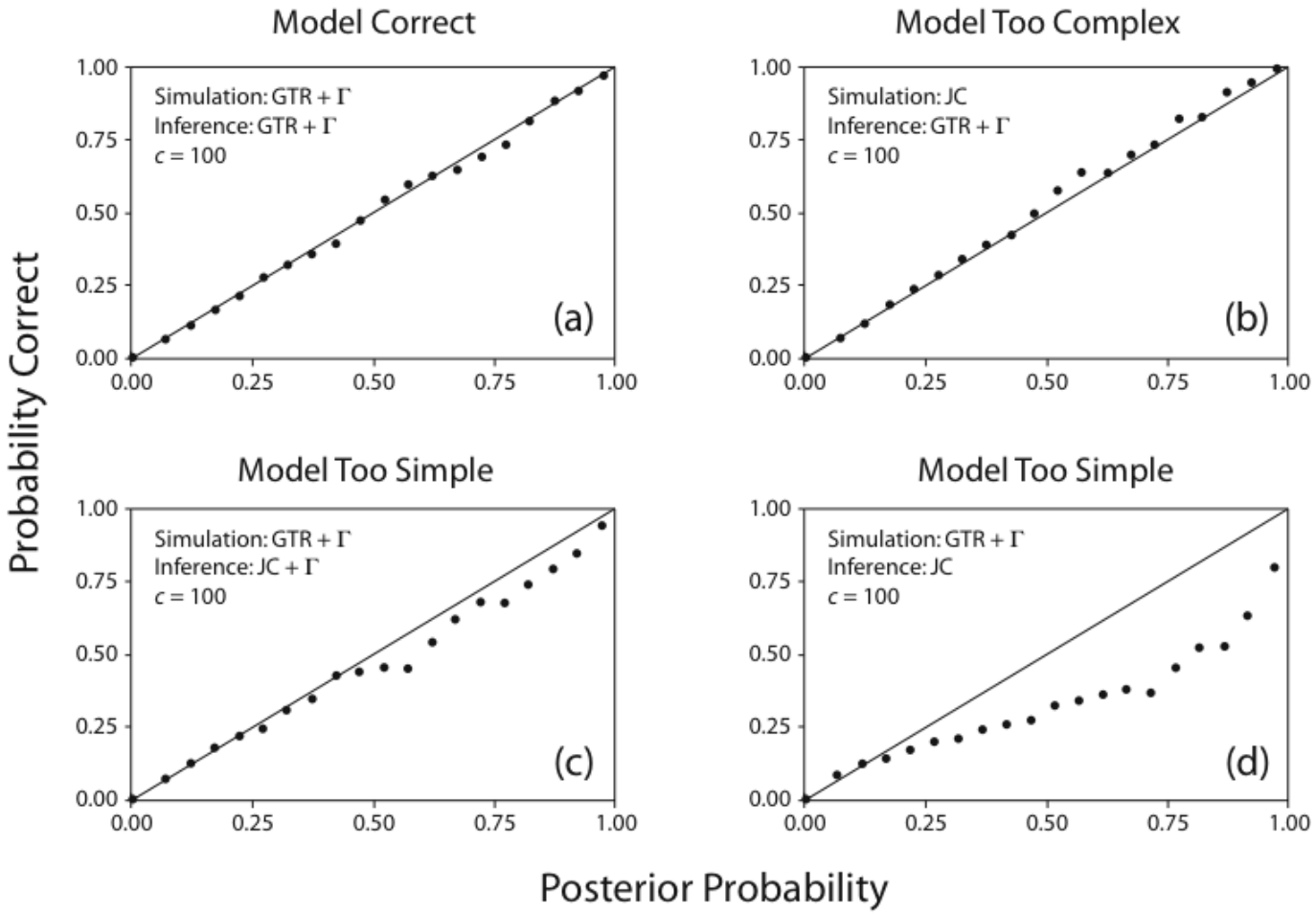


Majority rule consensus tree

Frequencies represent the posterior probability of the clades

Probability of clade being true given data and model

Bayesian Model Sensitivity



Models in the Bayesian World

- **Underfitting**: really bad because posterior probabilities can be misleading
- **Overfitting**: usually not a problem:
  - Bayesian robustness to overparameterization
  - Computational efficiency
  - Standard MCMC machinery easy to extend to large numbers of parameters
- … but overfitting can lead to:
  - Problems sampling from the posterior
  - Posterior largely reflects the prior

3. Graphical Models

Statistical Phylogenetics

- Statistical approaches increasingly important:
  - Difficult problems requiring accurate and unbiased inference (e.g., structure of rapid radiations)
  - More aspects of molecular evolution being examined (structural dependencies, etc)
  - Combination of background knowledge and sequence information (e.g., divergence time estimation)
- Modeling explosion, especially in the Bayesian context
- Challenging for empiricists to communicate and correctly understand models
- Challenging for developers of inference software

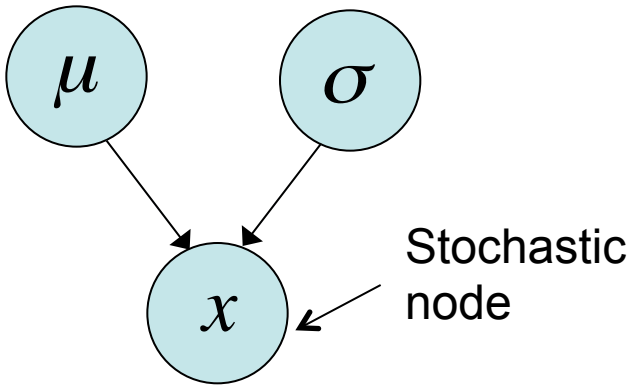
Probabilistic Graphical Models

- Theoretical framework for specifying dependencies in complex statistical models
- Allows a complex model to be broken down into conditionally independent distributions
- Closely related to standard statistical model formulae:

$$x \sim \text{Norm}(\mu, \sigma)$$

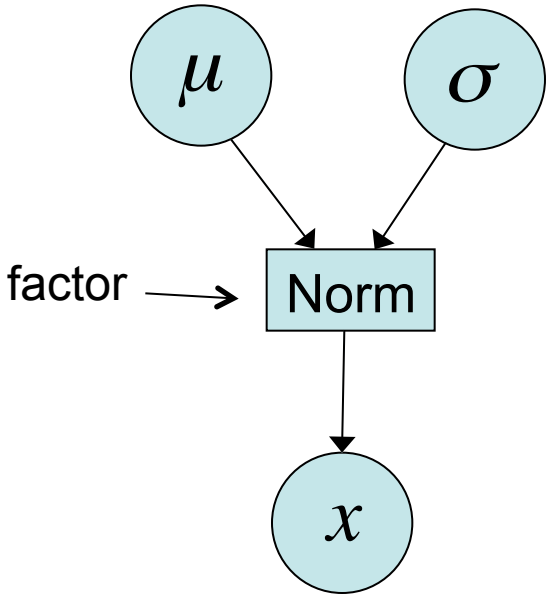
- Extensive literature on generic algorithms that apply to model graphs

$$x \sim \text{Norm}(\mu, \sigma)$$

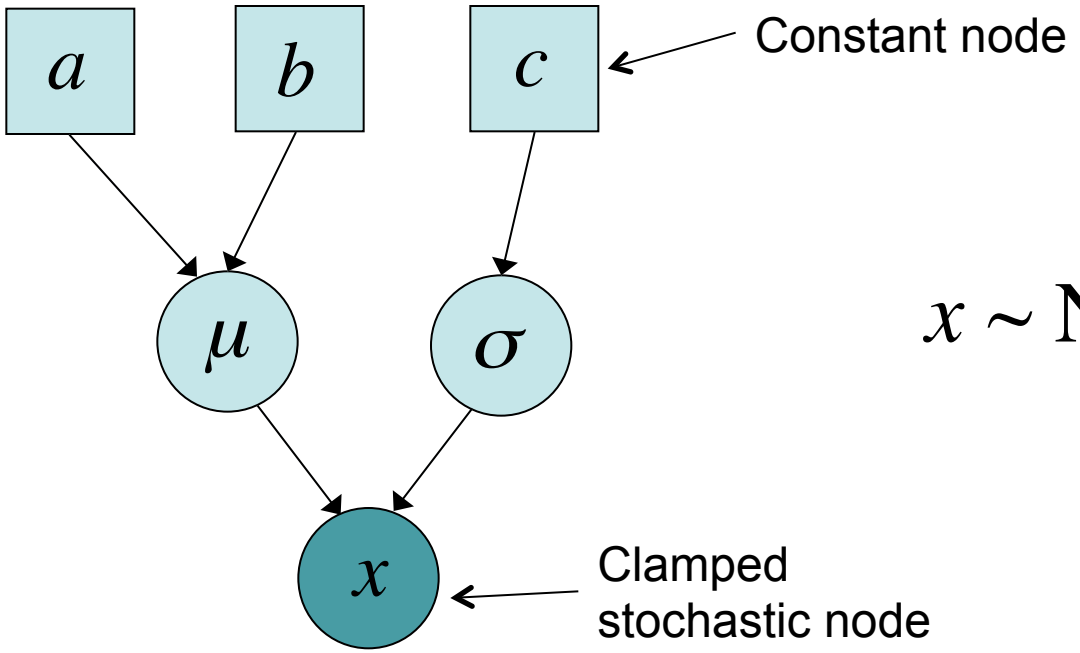


Graphical Model

Compact Form

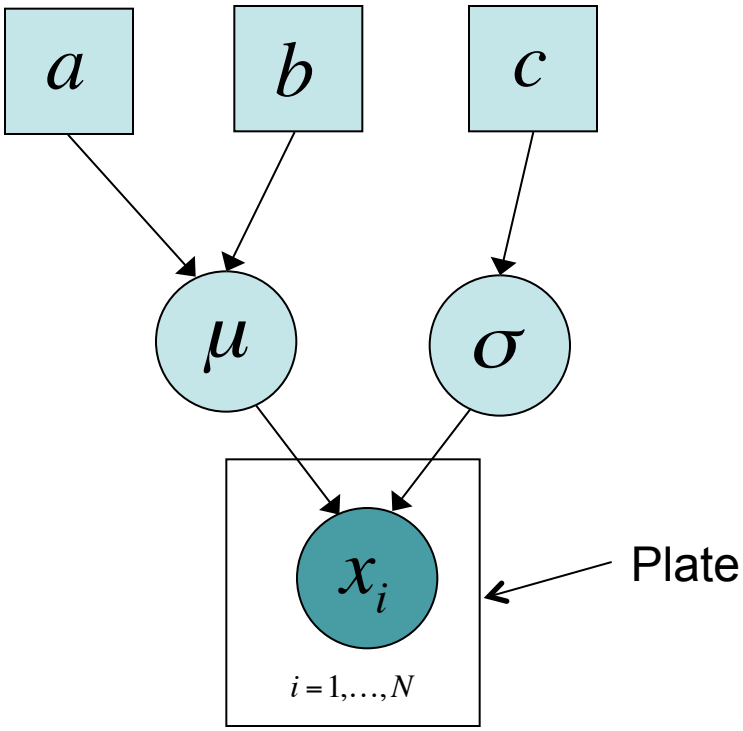


Factor Graph



$$x \sim \text{Norm}(\mu, \sigma)$$

Hierarchical Graphical Model



$$x \sim \text{Norm}(\mu, \sigma)$$

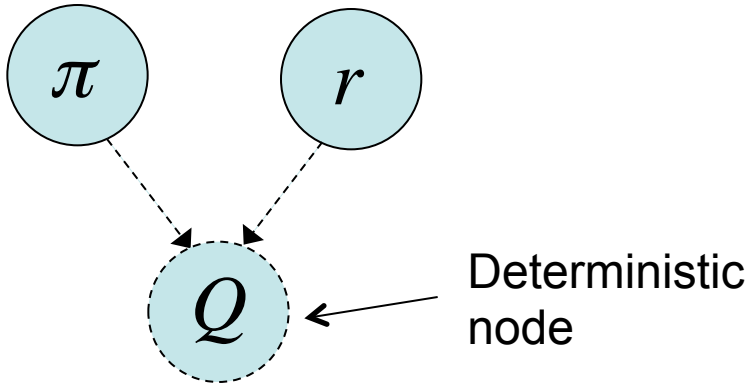
Hierarchical Graphical Model

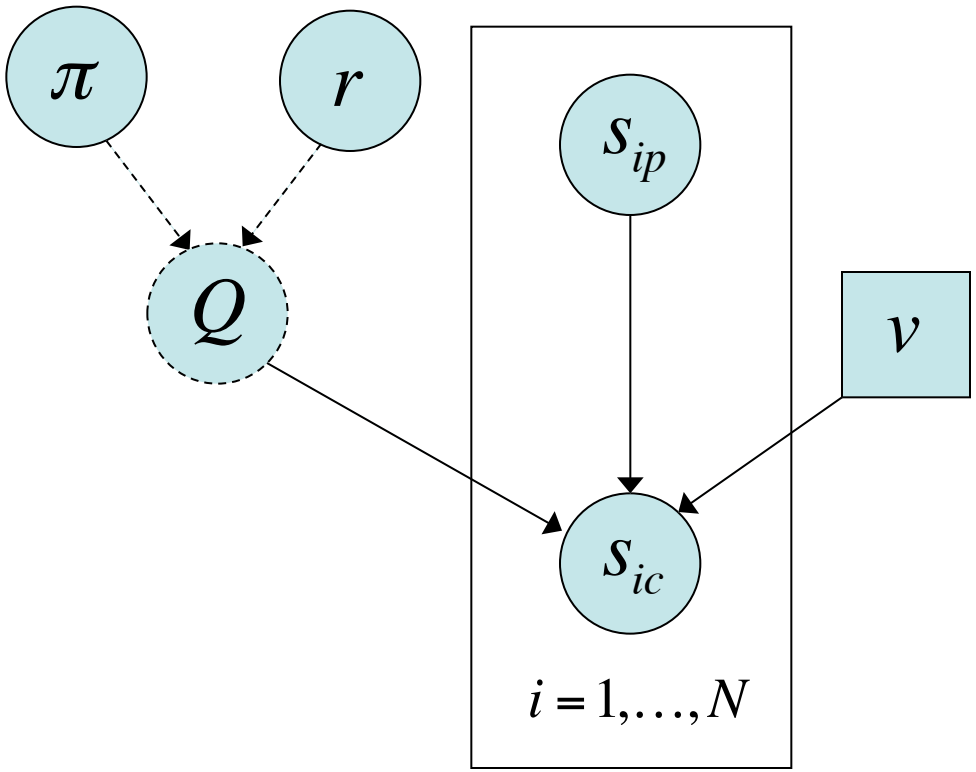
$$Q = \begin{pmatrix} - & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & - & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & - & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & - \end{pmatrix}$$

General Time Reversible (GTR) substitution model

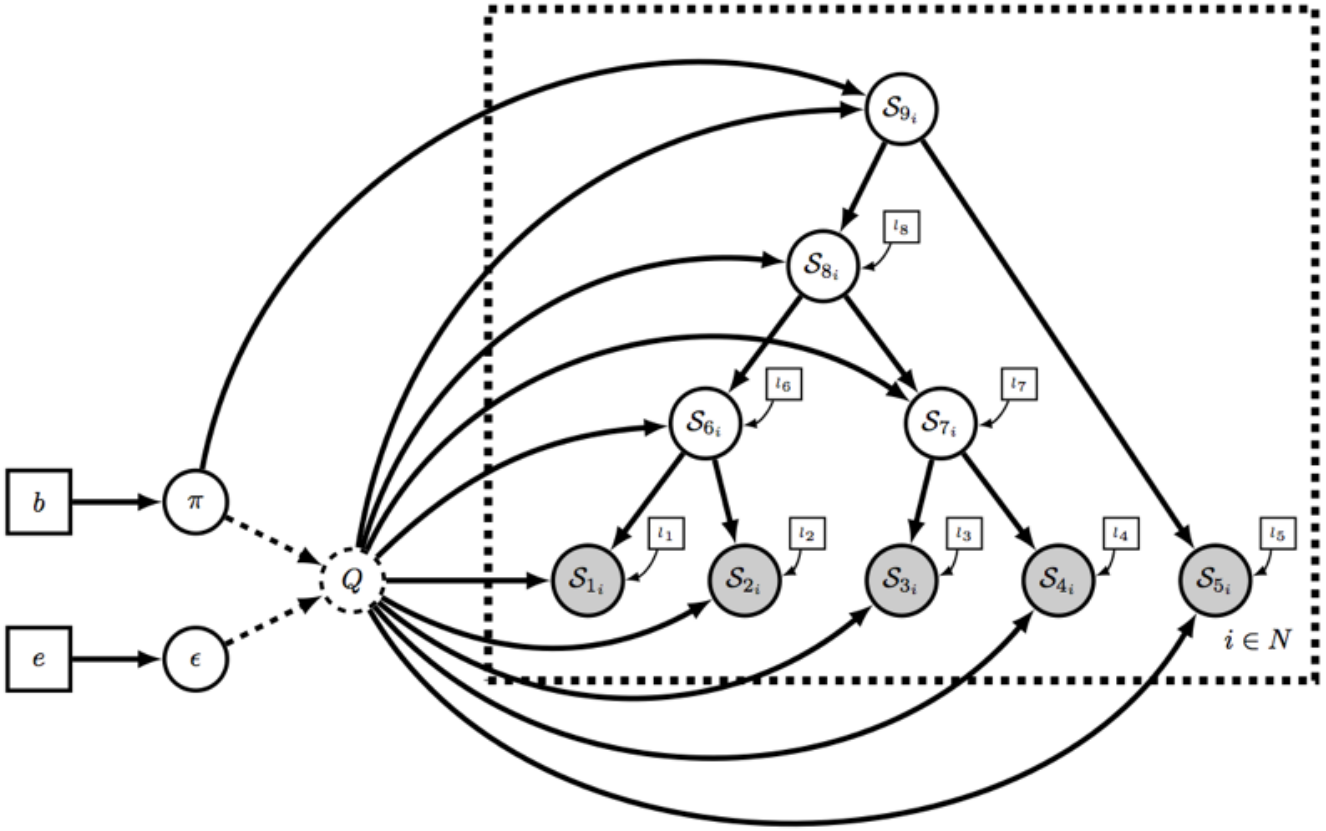
π Stationary state frequencies

r Exchangeability rates

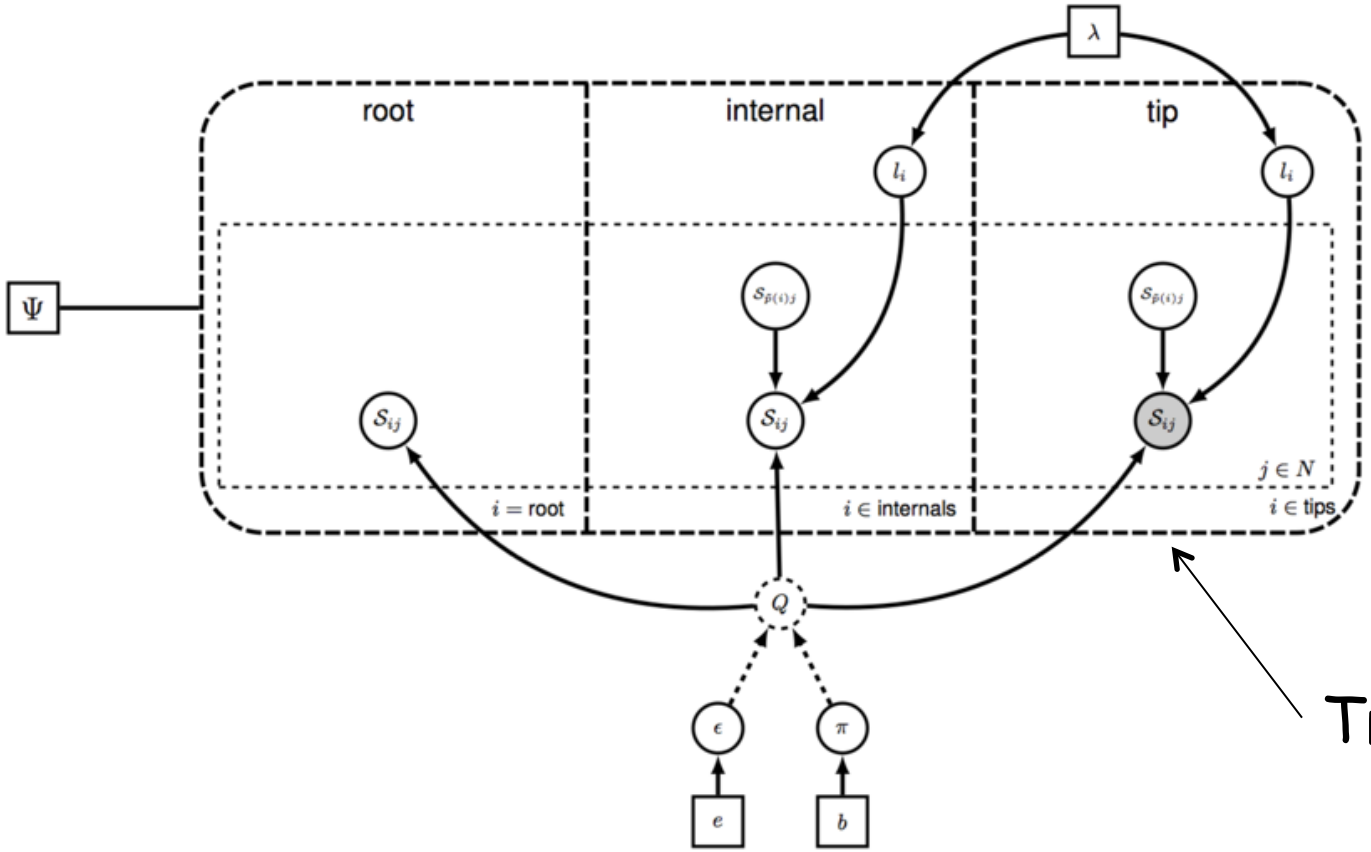

π
r
Q
Deterministic
node

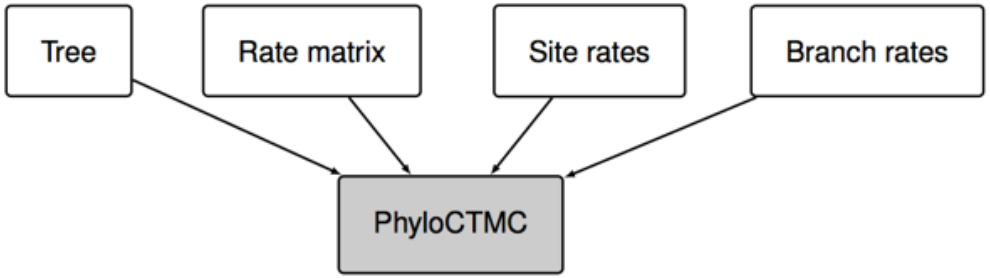

π
r
Q
s_{ip}
v
s_{ic}
$i = 1,\ldots,N$

GTR Phylogeny Model

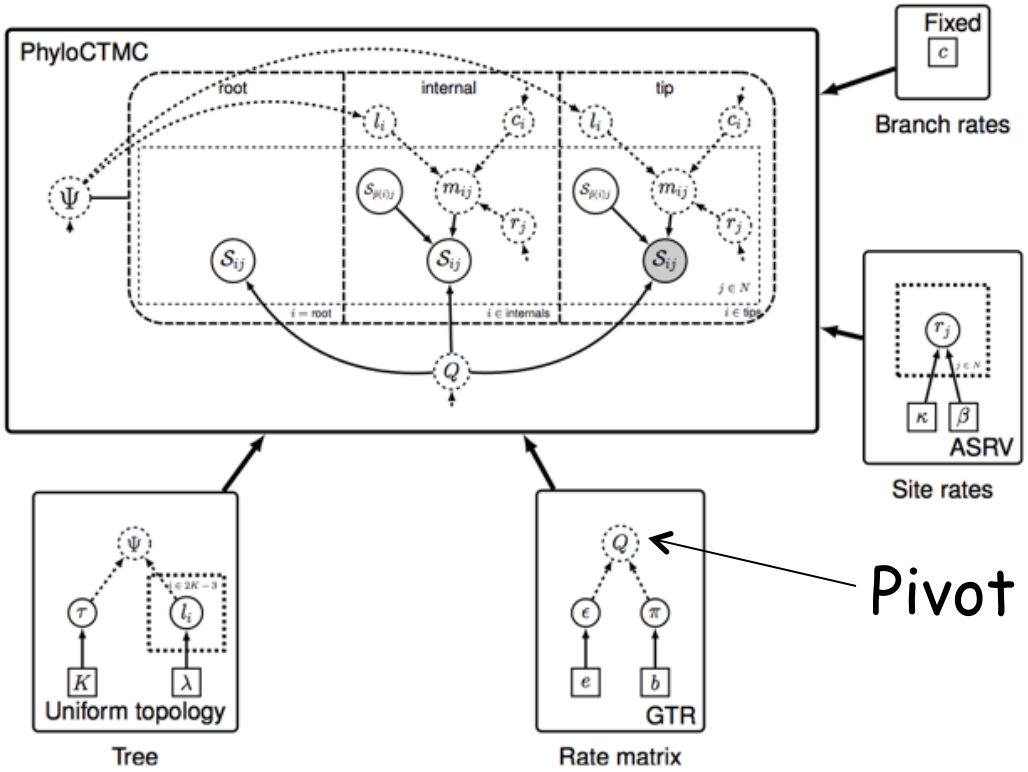


Tree Plate Representation





Modular Representation



Pivot variable

RevBayes Project

- Interactive computing environment intended primarily for Bayesian phylogenetic inference
- Uses a special language, Rev, for constructing probabilistic phylogenetic and evolutionary graphical models interactively, step by step
- Rev is similar to R and the BUGS modeling language
- RevBayes provides generic computing machinery for simulation, inference and model testing

Basic properties of the Rev language

```
# There are three kinds of statements in the language


# 1. Arrow assignment (value assignment, create constant nodes)

> a <- 4                # Give a the value 4
> b <- sqrt(a)          # Giva b the value of sqrt(a), that is, 2
> b                     # Print the value of b
 2



# 2. Equation assignment (create deterministic nodes)

> c := sqrt(a)          # Make c a dynamic function node evaluating sqrt(a)
> c
 2
> a <- 9                # Give a the value 9
> b                     # Print the value of b
 2
> c                     # Print the value of c
 3
```

Basic properties of the Rev language

```
# 3. Tilde assignment (create stochastic variables (nodes))

> a ~ dnExp( rate = x )              # a is drawn from exp dist with rate = x
```

Basic properties of the Rev language

```
# -------------------------------
# Declaring and defining functions
# -------------------------------

> function foo ( x ) { x * x }
> foo( 2 )
 4

# If you wish, you can specify types as well

> function PosReal foo ( Real x ) { x * x }

# Without explicit types, RevObject is the assumed type


# -------------------------------
# Declaring and defining new types
# -------------------------------

> class myclass : Move {
+     Real myTuningParam;
+     procedure Real move( Real x ) { myTuningParam * x }
+ }

# Inheritance, function overriding and overloading
```

A complete MCMC analysis in Rev

```
a <- -1.0
b <- 1.0

mu    ~ dnUnif(a, b)
sigma ~ dnExp(1.0)

for (i in 1:10) {
    x[i] ~ dnNorm(mu, sigma)
    x[i].clamp(0.5)
}

mymodel = model(mu)   # Any stochastic node in the model works

mymcmc =- mcmc(mymodel)

mymcmc.run(1000)
```

```
# definition of the myGTR function (“Ziheng's favorite”)
function model myGTR (CharacterMatrix data) {

   # describe Q matrix
   pi ~ dflatdir(4);
   r  ~ dflatdir(6);
   Q := gtr(pi, r);

   # describe tree
   tau ~ dtopuni(data.taxa(), rooted=false);

   # gamma shape
   alpha ~ dunif(0.0, 50.0);

   # discrete gamma mixture
   for (i in 1:4)
        catRate[i] := qgamma(i*0.25-0.125, alpha, alpha);
   for (i in 1:data.size())
        ratecat[i] ~ dcat(simplex(0.25,0.25,0.25,0.25));

   # associate distributions with tree parts
   for (i in 1:data.size()) {
        for (n in 1:tau.numNodes()) {
             if (tau.isTerminal(n)) {
                  tau.length[n] ~ exp(1.0);
                  tau.state[n] ~ ctmc(Q, e.length*catRate[ratecat[i]],
                       tau.state[tau.parent(n));
                  tau.state[n] <- data[i][tau.tipIndex(n)];
             }
             else {
                  tau.length[n] ~ exp(10.0);
                  tau.state[n] ~ ctmc(Q, e.length*catRate[ratecat[i]],
                       tau.state[n]);
             }
        }
   }

   # return model
   return model( Q );
}
```

Definition of a new phylogenetic model

Appr. 20 lines

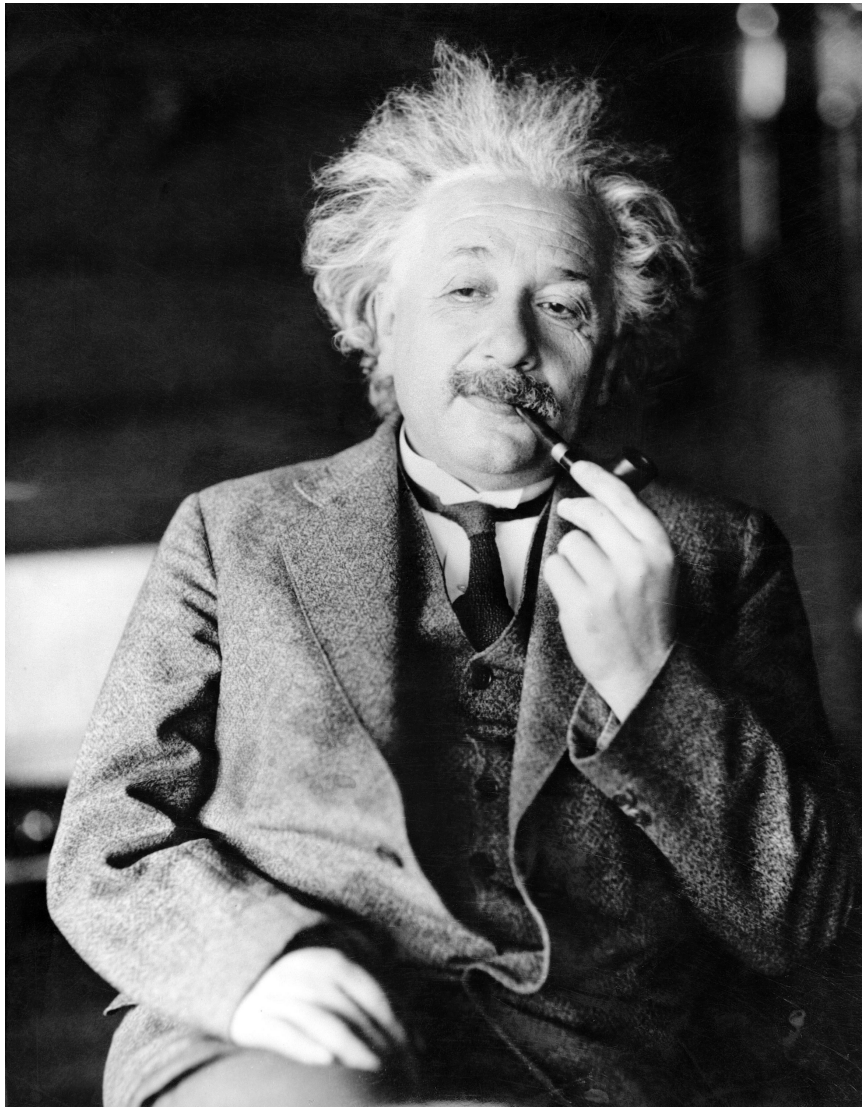
Complexity hidden from normal user

```
# Read in data
myData <- read( “data.nex” )

# Apply model
myModel = zihengGTR( myData )

# Construct mcmc
myMCMC = mcmc( myModel )

# Run mcmc
myMCMC.run(10000)
```



Listening to lectures,
after a certain age,
diverts the mind too much
from its creative pursuits.
Any scientist who attends
too many lectures and
uses her own brain too
little falls into lazy habits
of thinking.

after *Albert Einstein*