BI694

Bioinformatics & Phylogenetics

How about maximum likelihood?

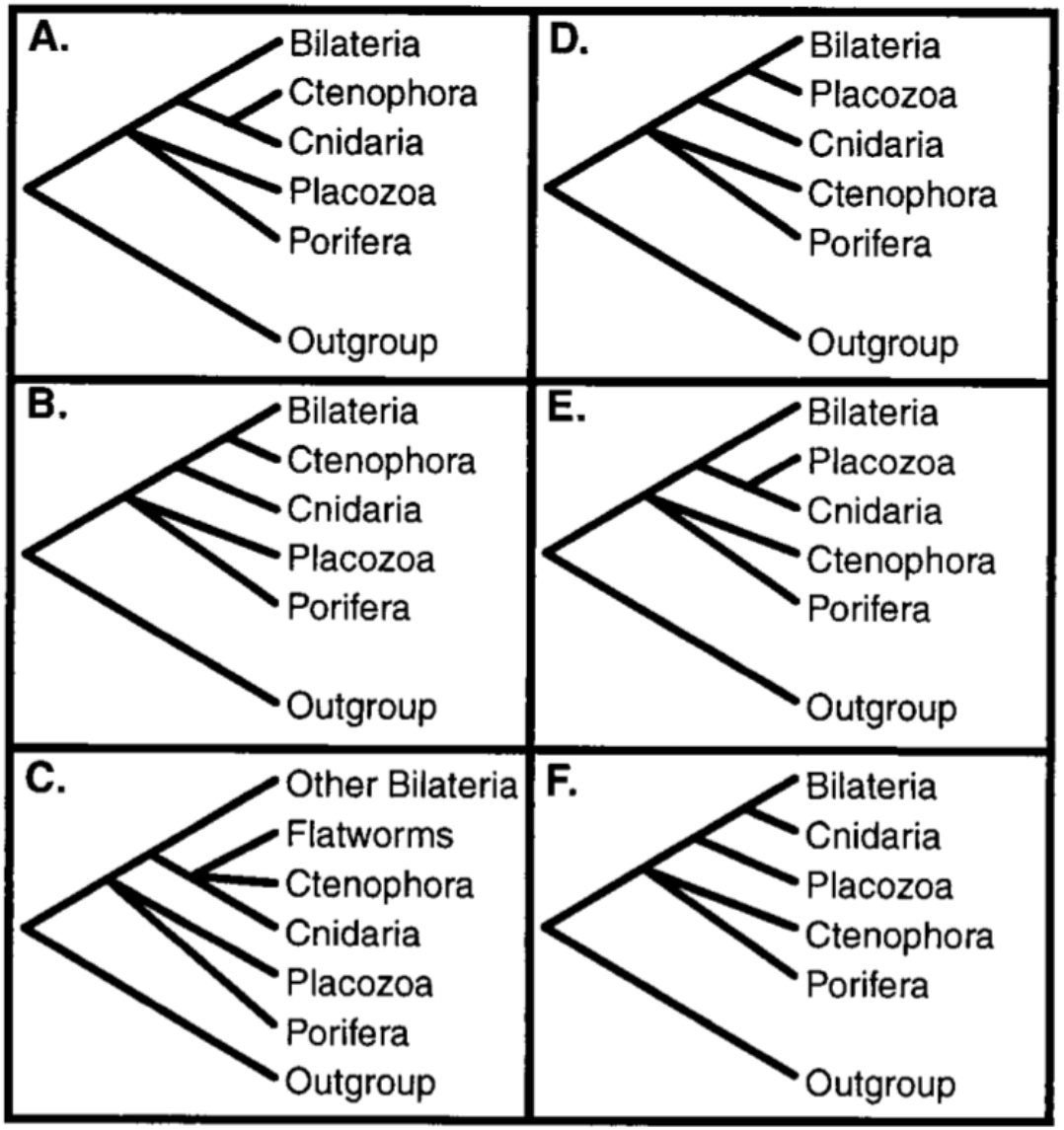


FIG. 1. Six alternative hypotheses for the origin of the Bilateria.

Collins (1998) PNAS

How about maximum likelihood?

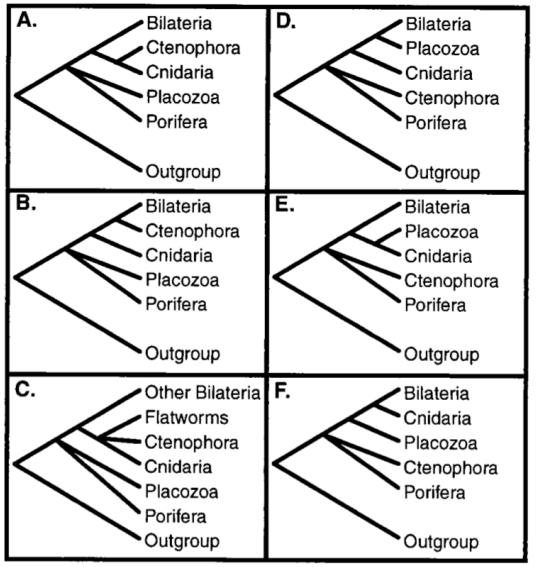


FIG. 1. Six alternative hypotheses for the origin of the Bilateria.

Bootstrapping ML phylogenies: RAxML implements fast algorithm for ML search

```
$ raxmlHPC   -f a \ # rapid bootstrap and search for best ML tree
             -m GTRGAMMA \ # model of evolution
             -p 12345 \ # random no seed for tree search
             -x 12345 \ # random number seed for bootstrap
             -# 100 \ # no of bootstrap replicates
             -s RootOfBilateria_RAxML.phy \ # alignment file
             -n Bilatria_RAxML # name of run
```

How about maximum likelihood?

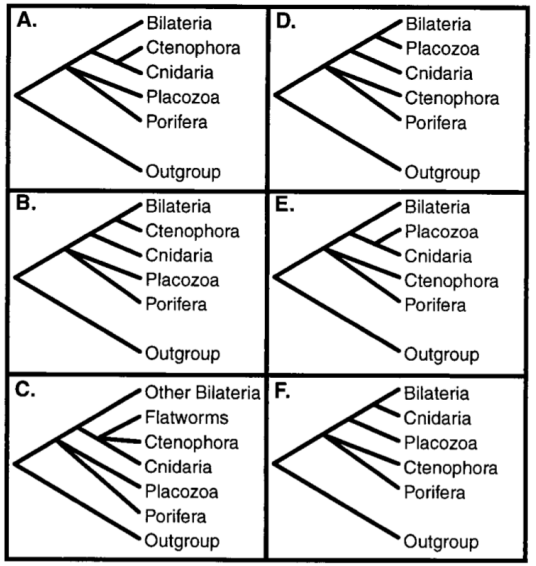


FIG. 1. Six alternative hypotheses for the origin of the Bilateria.

Did we use the best fitting model of evolution, ie, the best substitution model?

Testing Model Fit

Pick a more complex model only when the gain in likelihood is higher than expected if the simpler model were true.

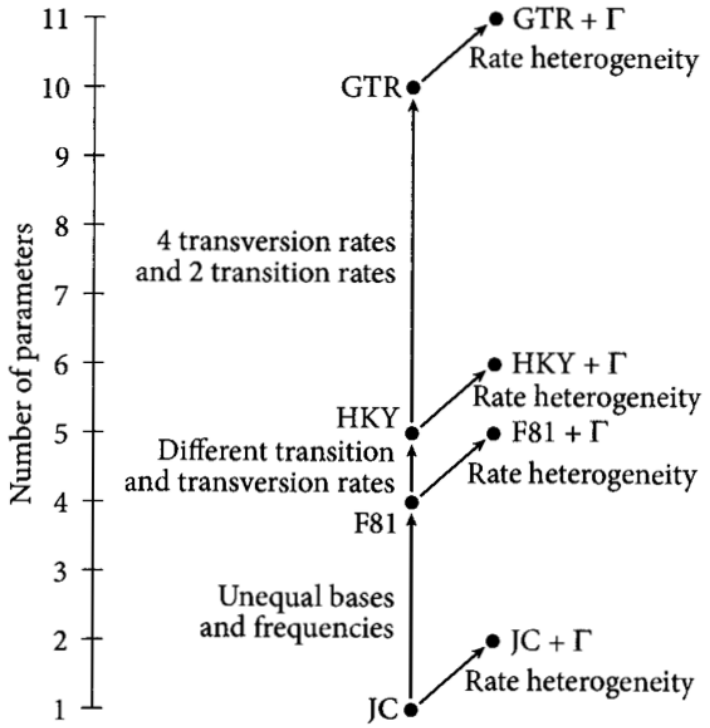
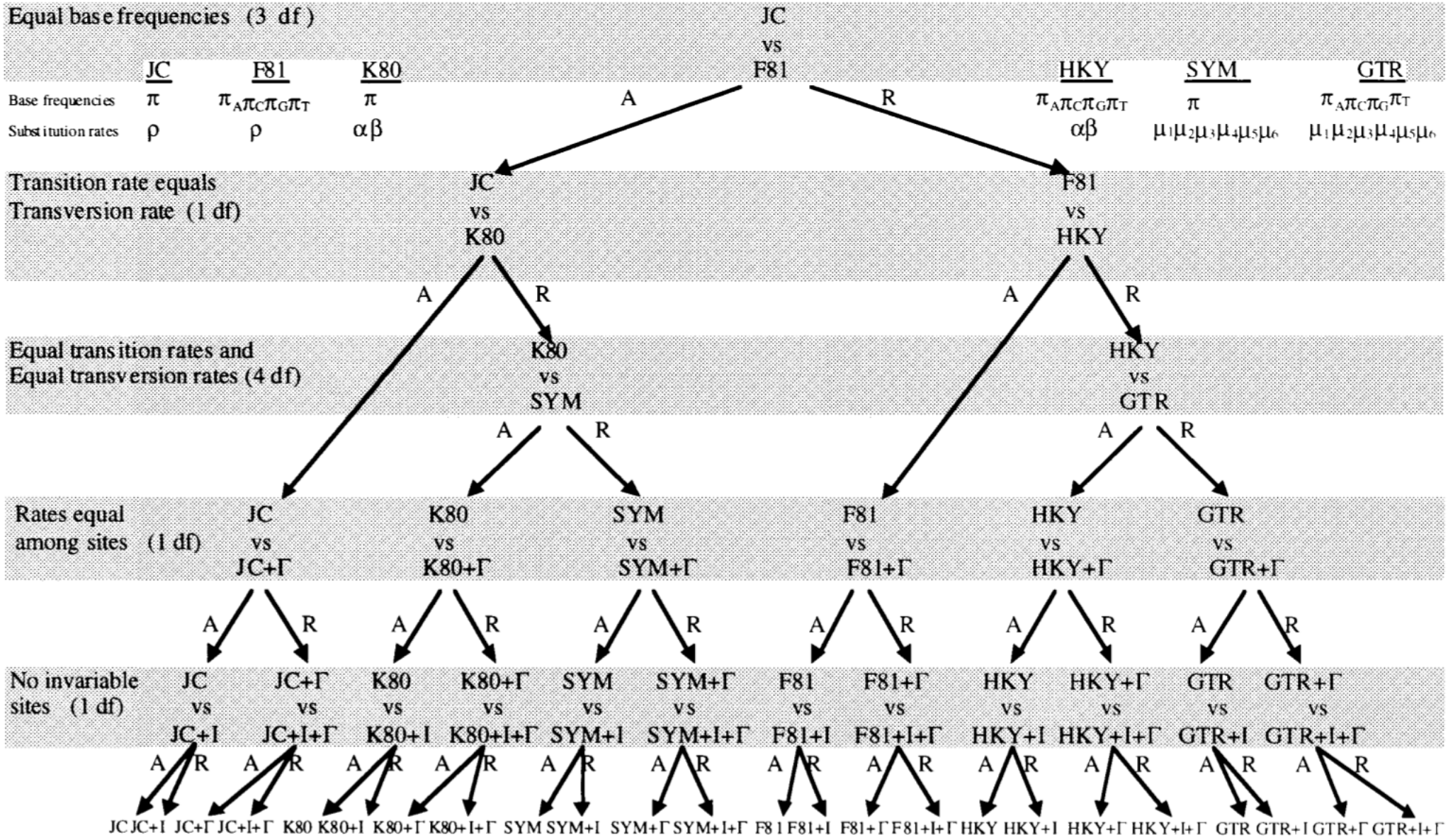


FIGURE 8.10 Depiction of the relationship between some commonly used models of evolution. Any two models that are connected by arrows that proceed in the same direction are nested: the simpler model (closer to the bottom of the chart) contains a subset of the free parameters of the more complex model. The axis on the left shows the number of free rate parameters in the model. The figure assumes that site-to-site rate heterogeneity is modeled using a discrete approximation to a gamma (Γ) distribution, which adds one free parameter to the model.

Testing Model Fit



Posada & Crandall (1998)

Testing Model Fit

Likelihood Ratio:

Chi-square can be used as approximation for likelihood ratio test

$$\Lambda = \frac{\max\ [L_0\ (Null\ \ Model \mid Data)]}{\max\ [L_1\ (Alternative\ \ Model \mid Data)]}$$

Testing Model Fit

Akaike Information Criterion (AIC)

Suppose data generated by some unknown process **f**. Two candidate models represent **f**: **g1** and **g2**.

If we knew **f**, then we could find the information lost from using **g1**; similarly, the information lost from using **g2** to represent **f** could be found. We would then choose the candidate model that minimized the information loss.

We do not know **f**. We can estimate how much more (or less) information is lost by g1 than by g2; ie, which model fits the data better (Akaike 1974)

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

k is the number of estimated parameters; ln(L) the log likelihood

Smaller AIC indicates better fit of the model to the data

Testing Model Fit

MrModeltest:

Execute your data file in PAUP*.

Execute the file MrModelblock in PAUP*. A file called mrmodel.scores will appear in the current directory.

This file is the input for mrmodeltest2.

```
$ mrmodeltest2 < mrmodel.scores > out
```

Bayesian Phylogenetics

In almost all cases posterior probability distribution cannot be derived analytically.

Cannot estimate it by random sampling: optima are concentrated in small areas of vast parameter space ← run MCMC robot without taking landscape into account

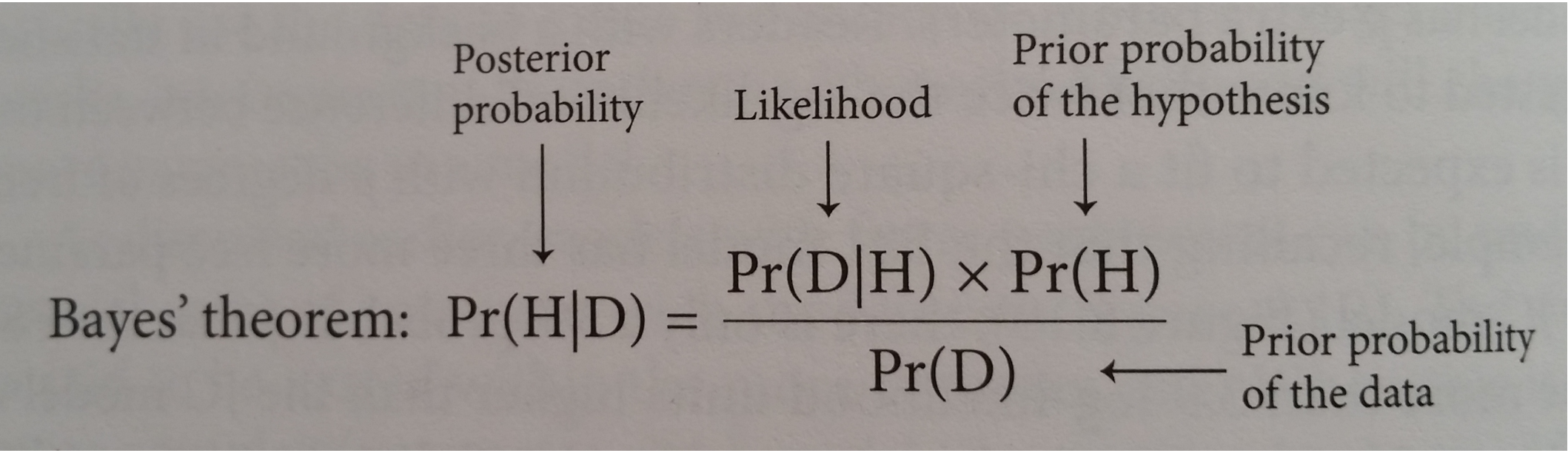
Solution: estimate posterior using Markov chain Monte Carlo (MCMC) sampling

Bayesian Phylogenetics

We inferred the best fitting substitution model earlier... So let's run a bayesian phylogenetic analysis using MrBayes.

```
$ mb RootOfBilateria_Bayes.nxs
```

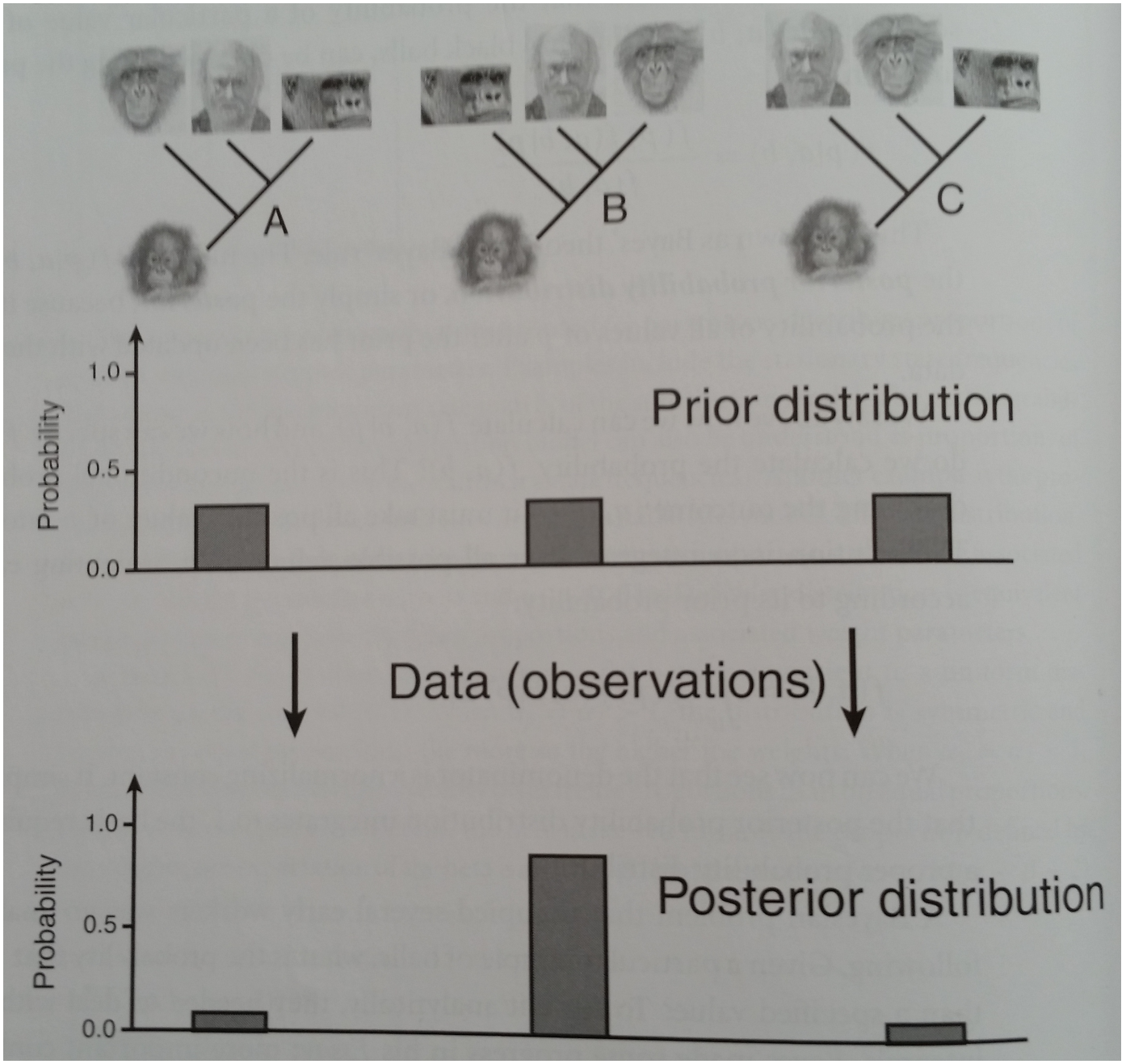
Bayesian Phylogenetics



Bayesian Phylogenetics

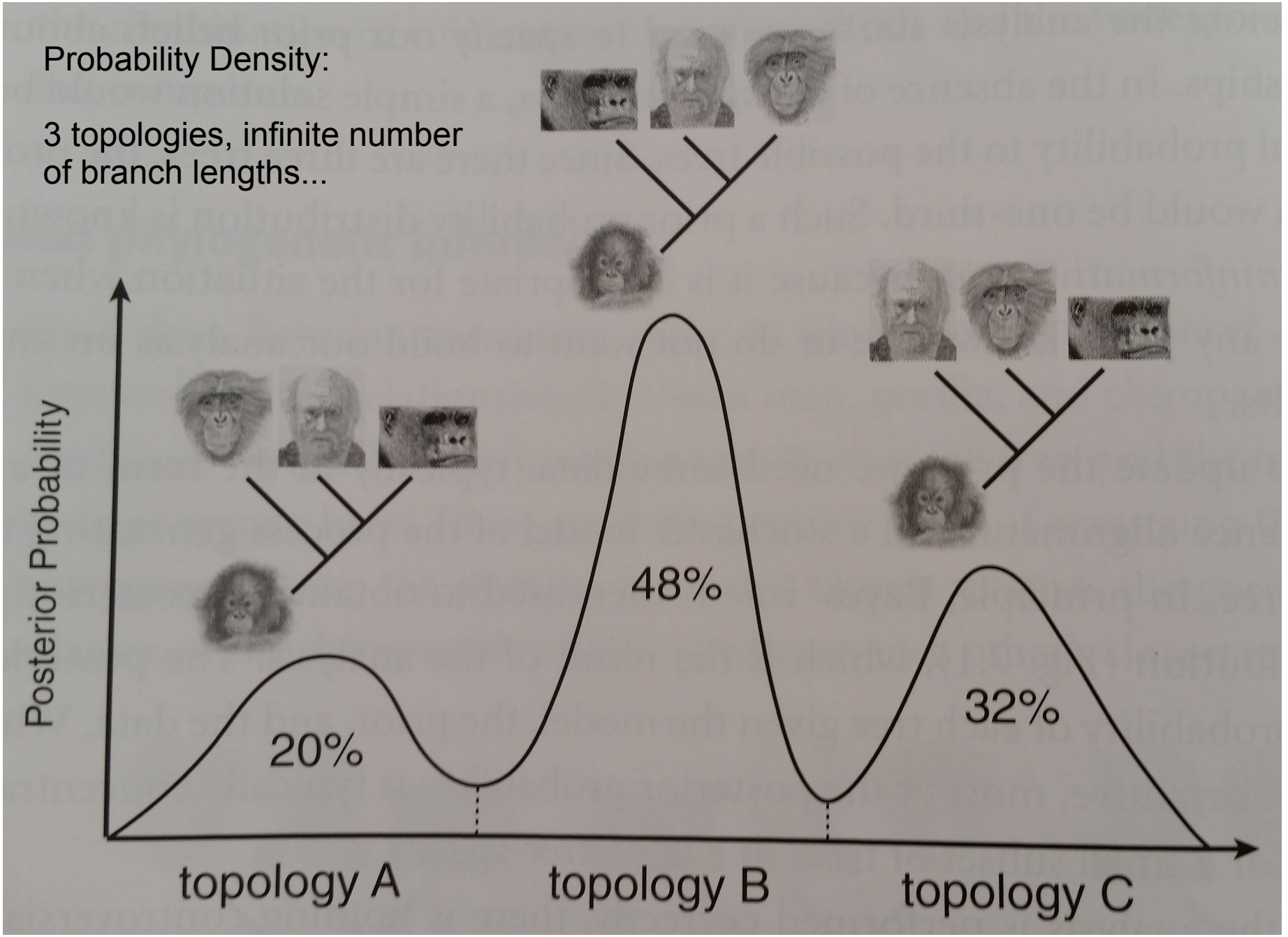
$$\text{Prob}(H \mid D) = \frac{\text{Prob}(H)\ \text{Prob}(D \mid H)}{\sum_H \text{Prob}(H)\ \text{Prob}(D \mid H)}$$

Bayesian Phylogenetics

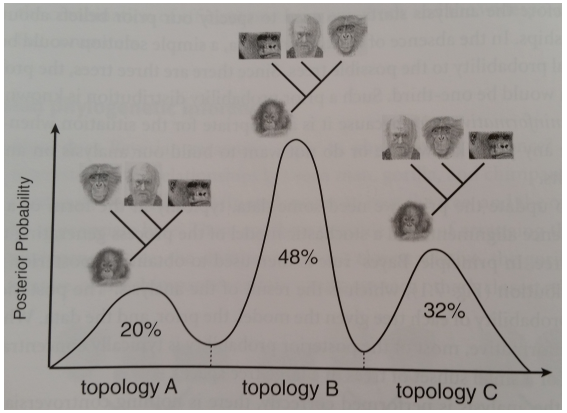


Uninformative, flat, or uniform prior

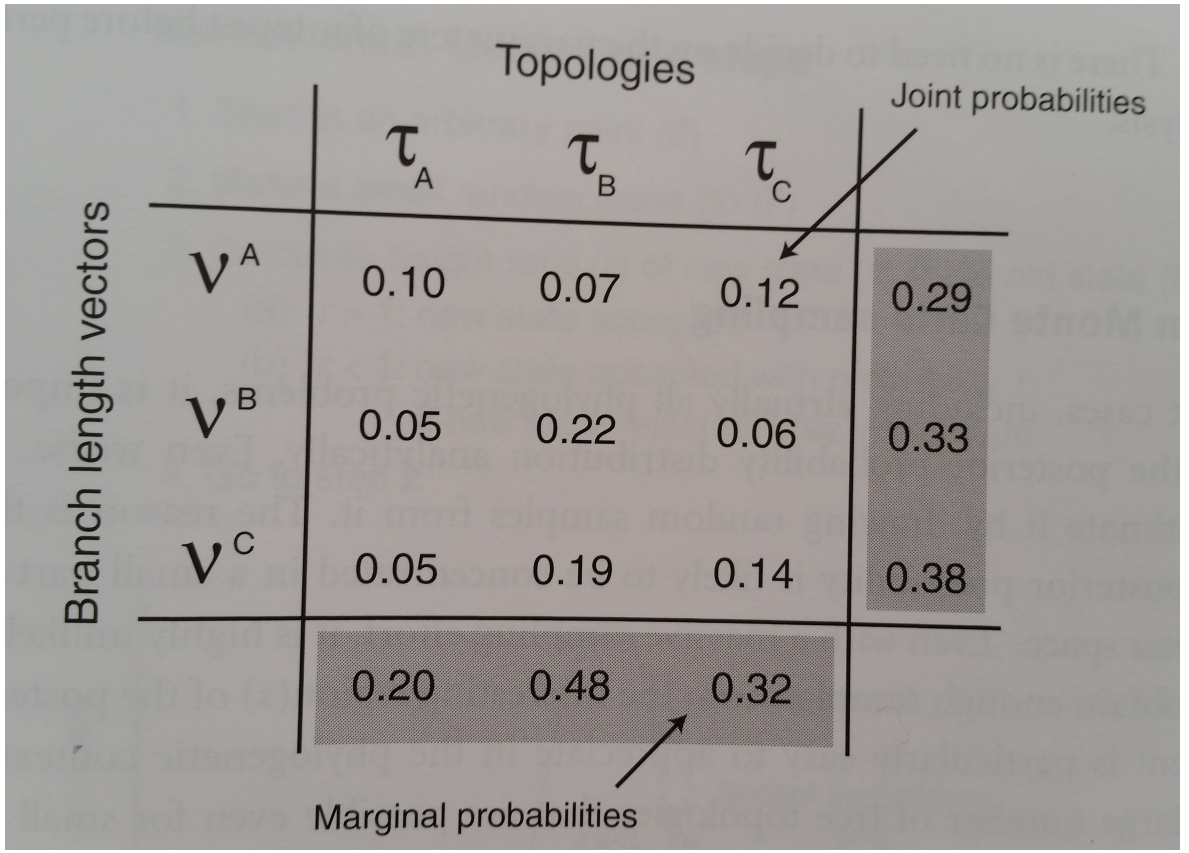
Bayesian Phylogenetics



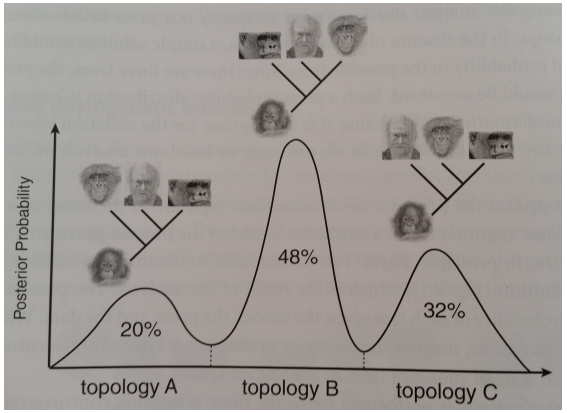
Bayesian Phylogenetics



Integrate over branch lengths for each topology to derive the (marginal) probability of the topology



Bayesian Phylogenetics



Integrate over branch lengths for each topology to derive the (marginal) probability of the topology

However, the branch lengths vector is literally infinitely long!

Topologies

Joint probabilities

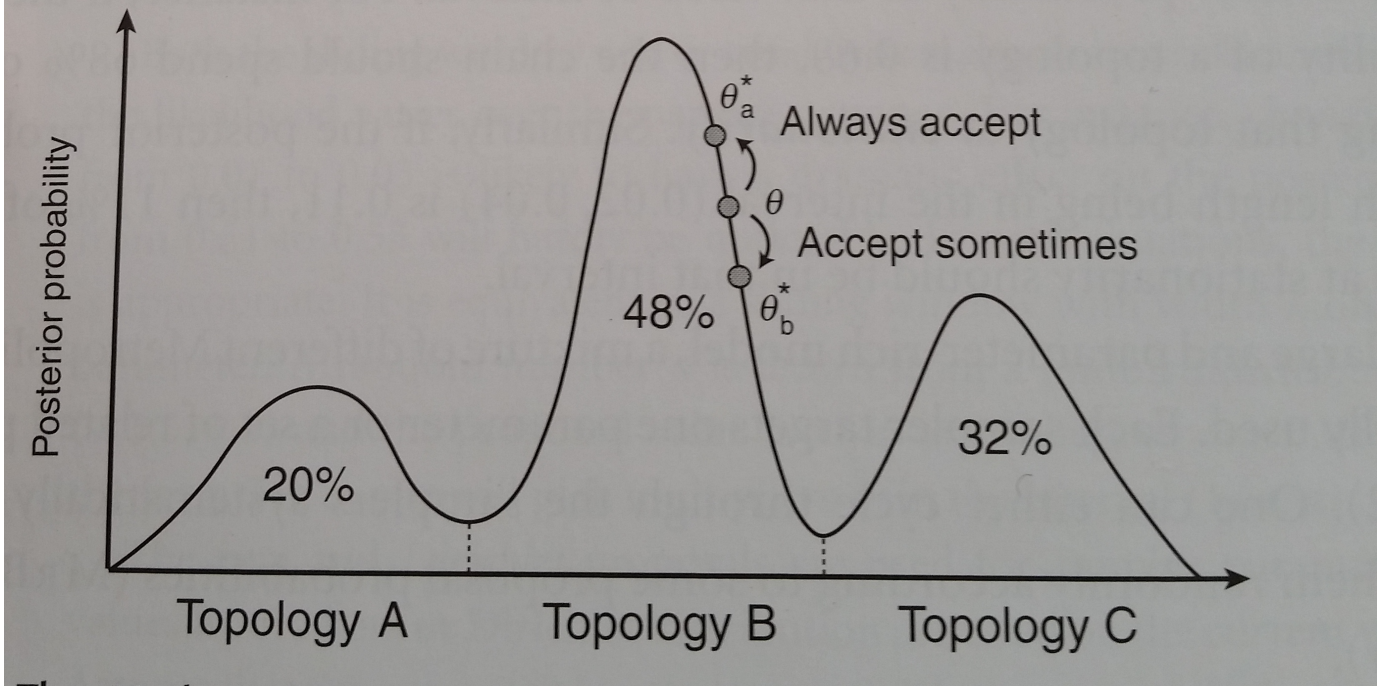
Branch length vectors	τ_A	τ_B	τ_C	
ν^A	0.10	0.07	0.12	0.29
ν^B	0.05	0.22	0.06	0.33
ν^C	0.05	0.19	0.14	0.38
	0.20	0.48	0.32	

Marginal probabilities

Bayesian Phylogenetics

Markov chain Monte Carlo steps

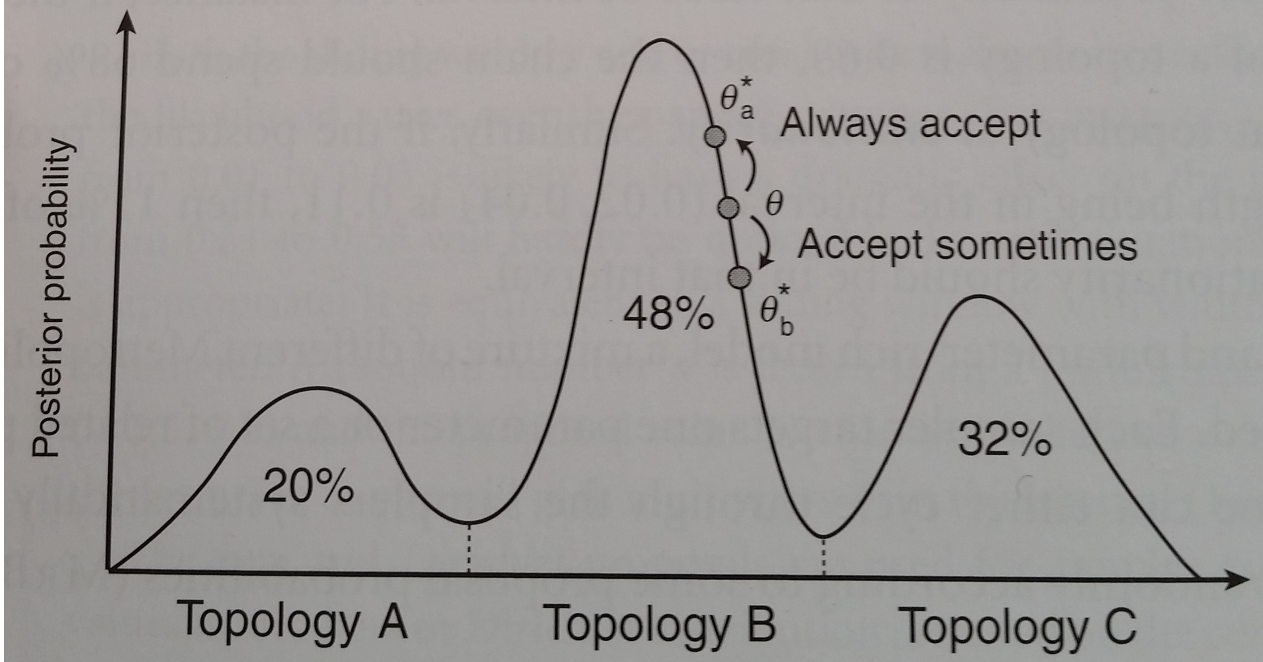
1. Start at an arbitrary point (θ)
2. Make a small random move (to θ^*)
3. Calculate height ratio (r) of new state (to θ^*) to old state (θ)
   (a) $r > 1$: new state accepted
   (b) $r < 1$: new state accepted with probability r
   if new state rejected, stay in old state
4. Go to step 2



Bayesian Phylogenetics

Markov chain Monte Carlo steps

1. Start at an arbitrary point (θ)
2. Make a small random move (to θ^*)
3. Calculate height ratio (r) of new state (to θ^*) to old state (θ)
   (a) $r > 1$: new state accepted
   (b) $r < 1$: new state accepted with probability r
   if new state rejected, stay in old state
4. Go to step 2



Better fitting tree topologies are will be more frequent in the posterior sample than ill-fitting ones...

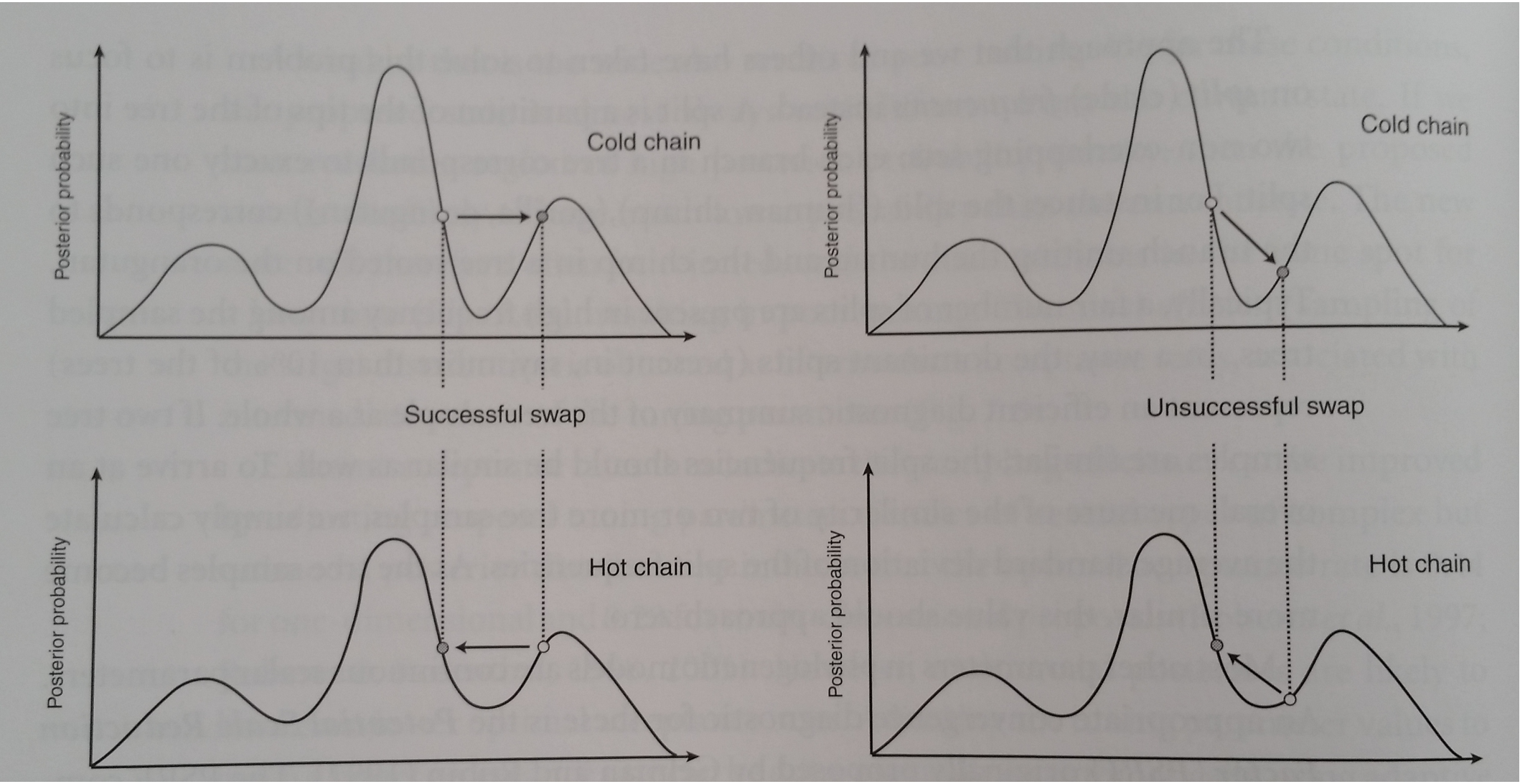
MCMC guaranteed to find posterior if run long enough

MCMC robot with 1 chain

Bayesian Phylogenetics

Exploring the posterior more effectively with “hot” and “cold” Markov chains

Posterior of “hot”chain raised to a power < 1; leads to a flattening of valleys between peaks and thus easier moves between peaks



Bayesian Phylogenetics

Let's investigate our Bayesian run in R (start R by typing: R)

```
$ library(rwty)

# load trees and parameter estimates
$ mb <- load.multi('./', format='mb')

# Let's look at how we are sampling treespace...
$ mb.rwty.topology <- makeplot.topology(mb, burnin=0)

$ mb.rwty.param <- makeplot.param(mb, burnin=0)

$ stats.rwty.treespace <- makeplot.treespace(mb)
```