

Phylogenetics

Introduction to Bayesian
inference II
RL-V3 MPP

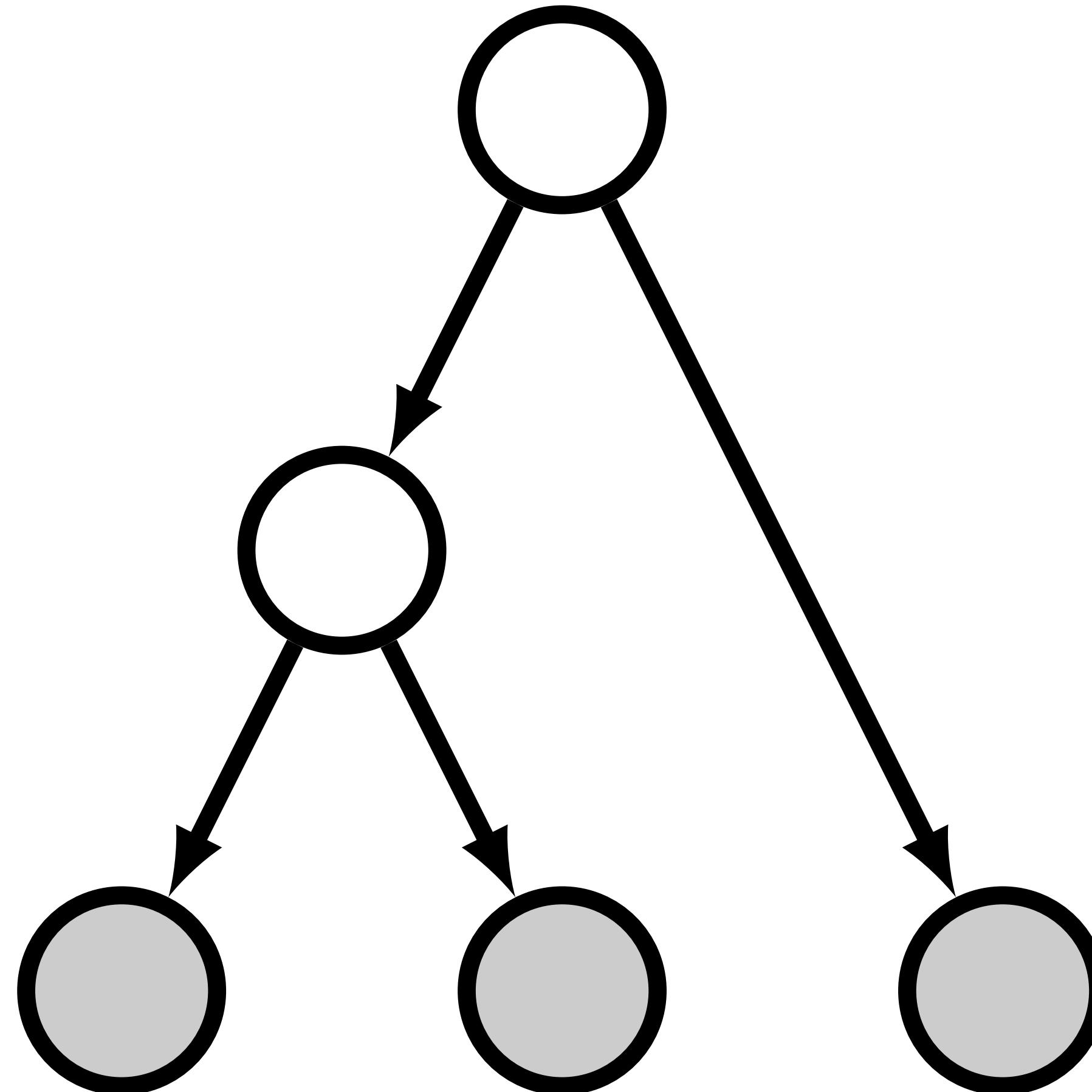
Rachel Warnock

07.05.24



Today's objectives

- Homework
- **Recap**
 - Bayesian inference
 - MCMC



Homework

Install the software [software FigTree](#) and [Tracer](#)



[Phylogenetics primer part 2: tree likelihood and rate heterogeneity](#)

Paul Lewis

→ See the question guide

Q&A Phylogenetics primer part 2 by *Paul Lewis*

(the answers provided here are my interpretation of these concepts – answers may vary!)

1. What are the assumptions of the following substitution models? Consider the rate of change between character states and the state frequencies.

JC →

HKY →

GTR →

1. What are the assumptions of the following substitution models? Consider the rate of change between character states and the state frequencies.

JC → equal frequencies, equal rates

HKY → unequal frequencies, unequal rates between transversions & transitions

GTR → unequal frequencies, unequal rates

transversions: A ↔ T or G ↔ C, transitions: A ↔ G or C ↔ T

2. Can you briefly describe the following approaches to account for rate variation among characters?

2a. Site specific rates

→ assign sites to separate partitions and allow each partition to have its own set of parameters

2b. Invariant sites model

→ assign a subset of sites to a “constant” (i.e., non-variable) category

2. Can you briefly describe the following approaches to account for rate variation among characters?

2c. Discrete Gamma model

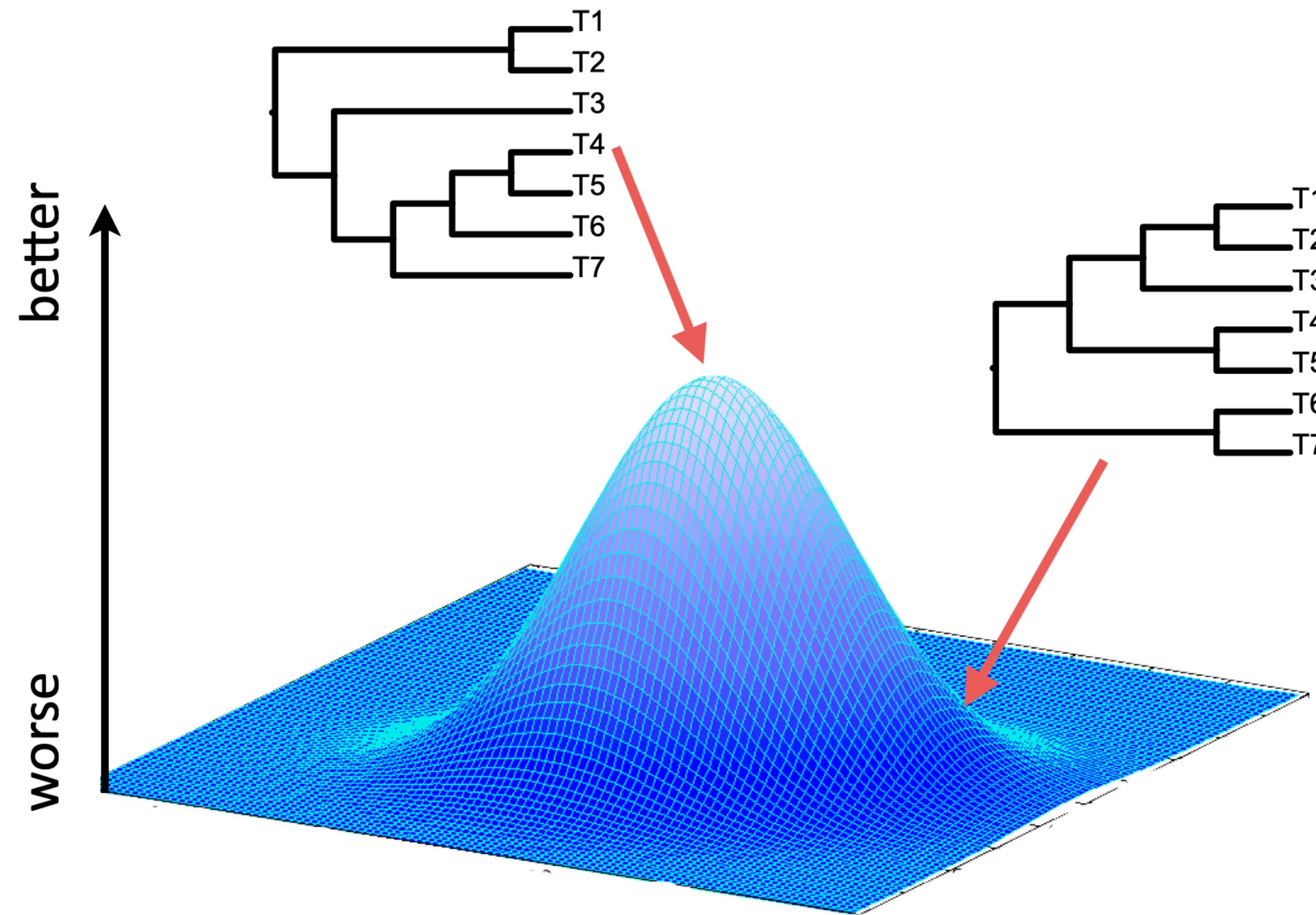
→ calculate the likelihood assuming there are discrete rate categories, e.g., 4. Variation in rate categories is represented by a gamma distribution and the parameters of the gamma distribution are calculated from the data

3. For a given substitution model, what do the values in the **Q matrix** and the **P matrix** represent?

- the Q matrix is the instantaneous rate matrix, i.e., the instantaneous rates of change between each character state combination
- the P matrix is the transition probability matrix, i.e., the probabilities of change between character states after relative time t , which is represented by the branch lengths

Recap

How do we find the ‘best’ tree?



It depends how you measure ‘best’

Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes
.....
Maximum likelihood	Likelihood score (probability), optimised over branch lengths and model parameters
.....
Bayesian inference	Posterior probability, integrating over branch lengths and model parameters

Both maximum likelihood and Bayesian inference are model-based approaches

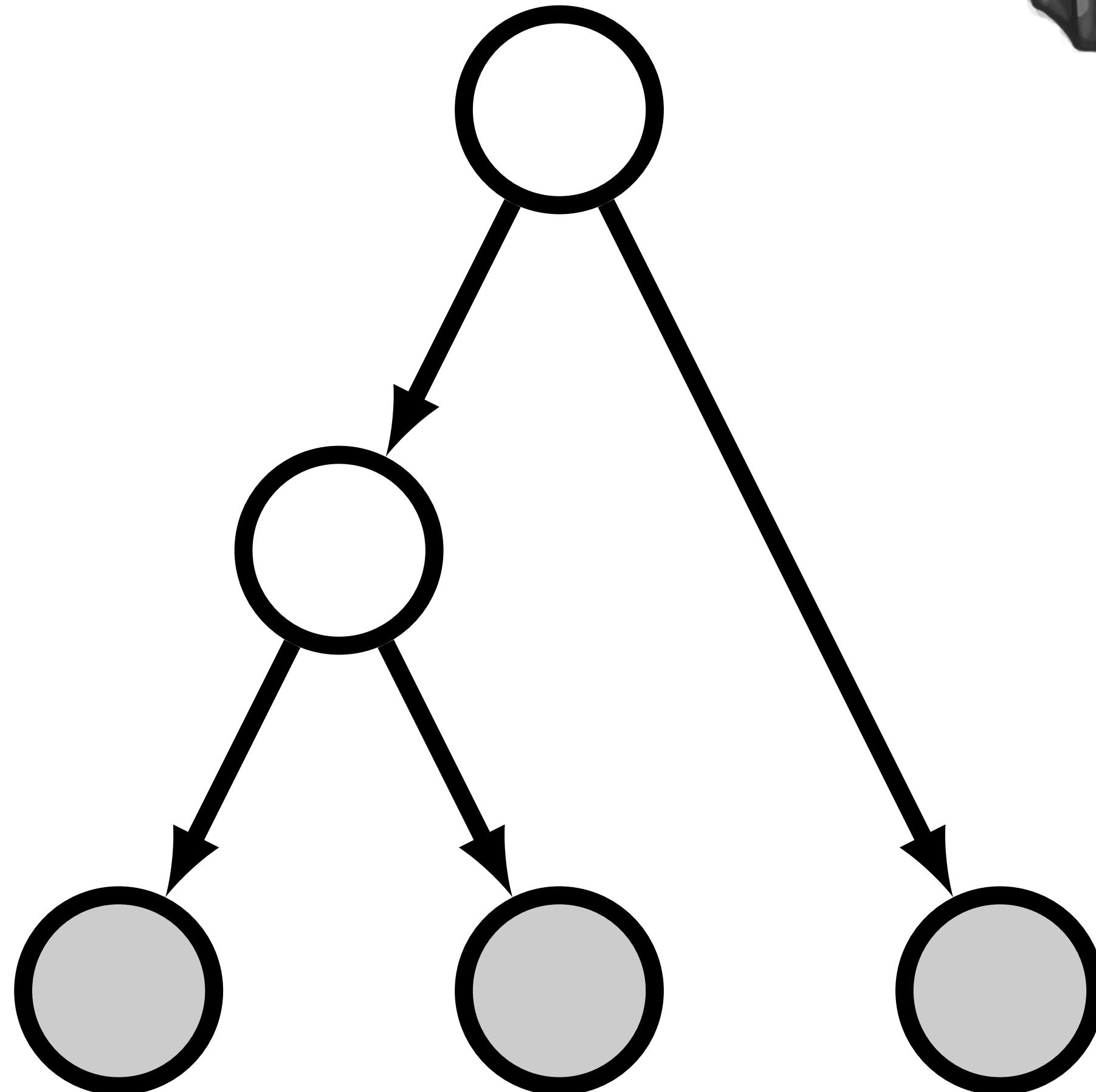
Note these are not the only approaches to tree-building but they are the most widely used

Graphical models

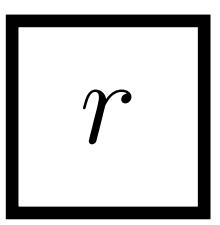


Provide tools for visually and computationally representing complex, parameter-rich models

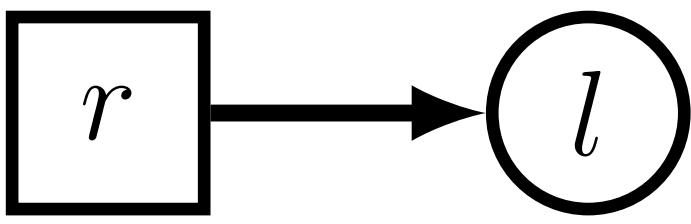
Depict the conditional dependence structure of parameters and other random variables



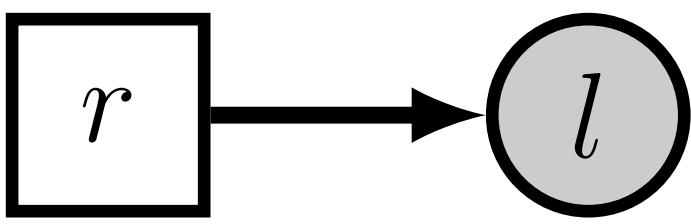
a)



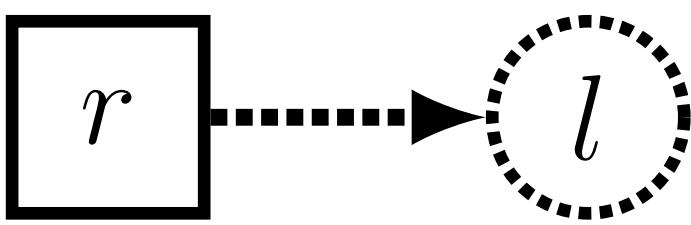
b)



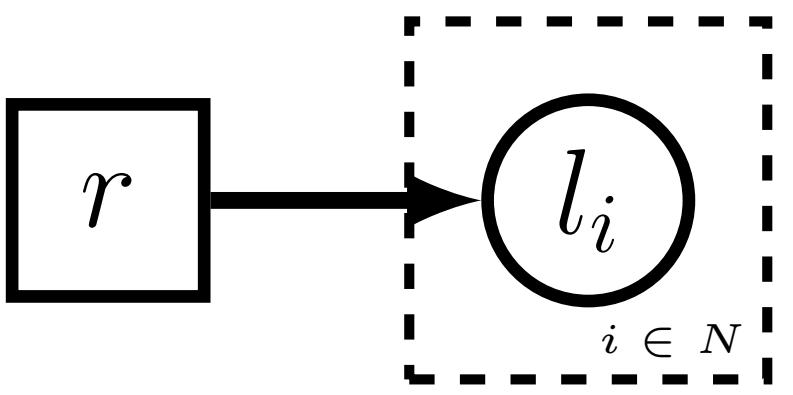
c)



d)



e)



```
# constant node  
r <- 10
```

```
# stochastic node  
l ~ dnExp(r)
```

```
# stochastic node (observed)  
l.clamp(0.1)
```

```
# deterministic node  
l := exp(r)
```

```
# stochastic nodes (iid)  
for (i in 1:N) {  
  l[i] ~ dnExp(r)  
}
```

Bayes' theorem

$$\Pr(\text{model} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{model}) \Pr(\text{model})}{\Pr(\text{data})}$$

Bayes' theorem

Likelihood

The probability of the data given the model assumptions and parameter values

$$\Pr(\text{model} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{model}) \Pr(\text{model})}{\Pr(\text{data})}$$

Bayes' theorem

Priors

This represents our prior knowledge of the model parameters

$$\Pr(\text{model} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{model}) \Pr(\text{model})}{\Pr(\text{data})}$$

Bayes' theorem

$$\Pr(\text{model} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{model}) \Pr(\text{model})}{\Pr(\text{data})}$$

Marginal probability

The probability of the data, given all possible parameter values. Can be thought of as a normalising constant

Bayes' theorem

posterior

Reflects our combined knowledge based on the likelihood and the priors

$\Pr(\text{model} \mid \text{data}) =$

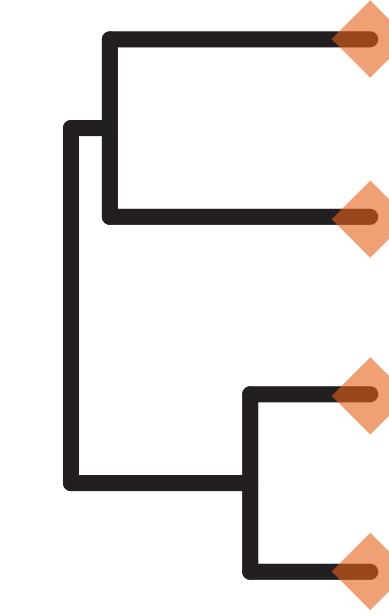
$$\frac{\Pr(\text{data} \mid \text{model}) \Pr(\text{model})}{\Pr(\text{data})}$$

Components used to infer trees

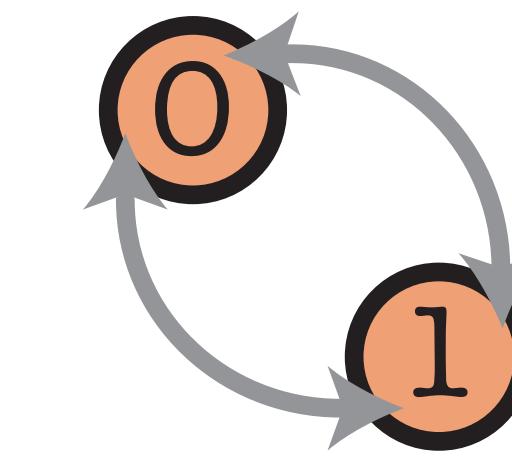
without considering time

0101...
1101...
0100...

data
sequences or
characters



tree
topology and
branch lengths



substitution
model

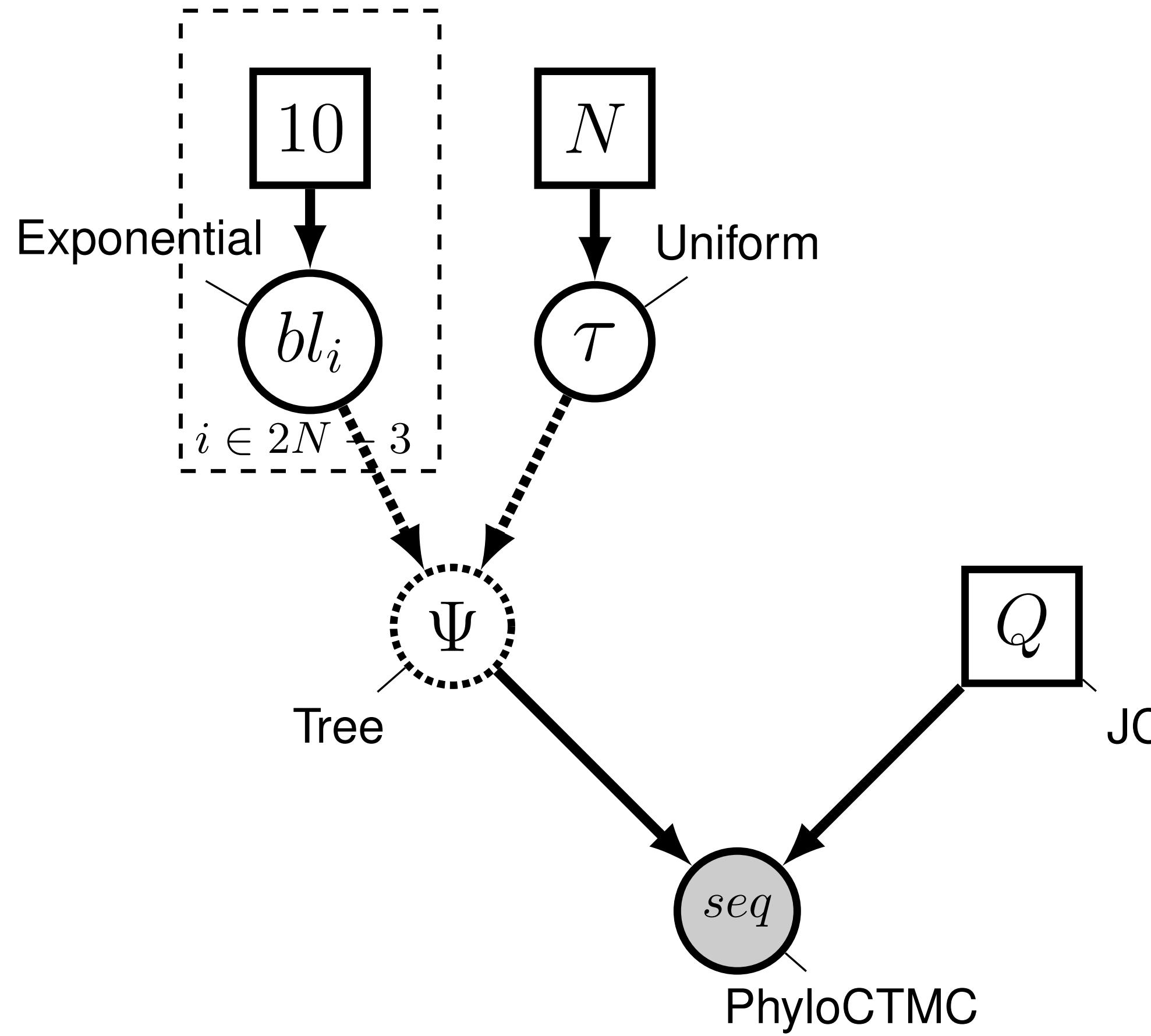
Bayesian tree inference

$$\text{posterior} \quad P(E \mid \text{0101...}, \text{1101...}, \text{0100...}) = \frac{\text{likelihood} \quad P(\text{0101...} \mid E) \quad P(E)}{\text{priors} \quad P(\text{0101...}, \text{1101...}, \text{0100...})}$$

Diagram illustrating the components of Bayesian tree inference:

- posterior**: $P(E \mid \text{0101...}, \text{1101...}, \text{0100...})$
- likelihood**: $P(\text{0101...} \mid E)$
- priors**: $P(E)$
- marginal probability**: $P(\text{0101...}, \text{1101...}, \text{0100...})$

The diagram shows a phylogenetic tree with two terminal nodes. The left node is labeled '0' and the right node is labeled '1'. Arrows indicate the direction of evolution from root to leaves. The likelihood term $P(\text{0101...} \mid E)$ corresponds to the probability of observing the sequence '0101...' at the first node given the tree E . The prior term $P(E)$ corresponds to the probability of the tree E itself.



```

for (I in 1:n_branches) {
  bl[I] ~ dnExponential(10.0)
}
topology ~ dnUniformTopology(taxa)
psi := treeAssembly(topology, bl)

Q_morpho <- fnJC(2)

phyMorpho ~ dnPhyloCTMC( tree=psi,
siteRates=rates_morpho, Q=Q_morpho,
type="Standard", coding="variable" )
phyMorpho.clamp( data )

```

Bayesian tree inference

$$\text{posterior} \quad P(E \mid \text{0101...}, \text{1101...}, \text{0100...}) = \frac{\text{likelihood} \quad P(\text{0101...} \mid E) \quad P(E)}{\text{priors} \quad P(\text{0101...}, \text{1101...}, \text{0100...})}$$

Diagram illustrating the components of Bayesian tree inference:

- posterior**: $P(E \mid \text{0101...}, \text{1101...}, \text{0100...})$
- likelihood**: $P(\text{0101...} \mid E)$
- priors**: $P(E)$
- marginal probability**: $P(\text{0101...}, \text{1101...}, \text{0100...})$

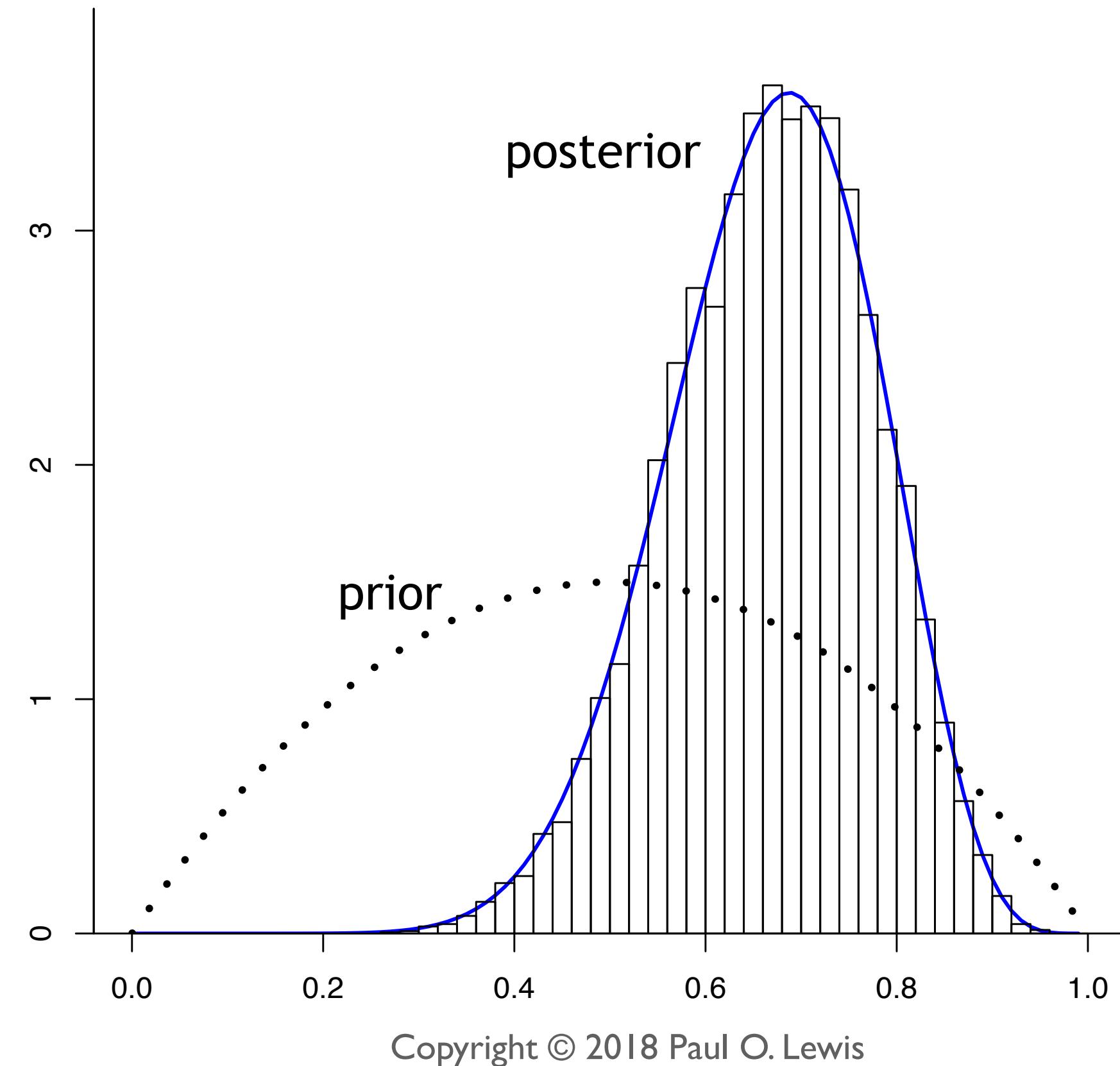
The diagram shows a phylogenetic tree with two terminal nodes. The left node is labeled '0' and the right node is labeled '1'. Arrows indicate the direction of evolution from root to leaves. The likelihood term $P(\text{0101...} \mid E)$ corresponds to the probability of observing the sequence '0101...' at the first node given the tree E . The prior term $P(E)$ corresponds to the probability of the tree E itself.

Bayesian tree inference

$$= \frac{P(\text{0101...} | \text{E} \circlearrowleft \text{O} \rightarrow \text{1}) P(\text{E} \circlearrowleft \text{O} \rightarrow \text{1})}{\int P(\text{0101...} | \text{E} \circlearrowleft \text{O} \rightarrow \text{1}) P(\text{E} \circlearrowleft \text{O} \rightarrow \text{1}) d\text{E} \circlearrowleft \text{O} \rightarrow \text{1}}$$

this part is incredibly difficult to calculate!

What is Markov chain Monte Carlo (MCMC)?

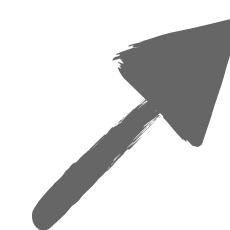


The aim is to produce a
histogram that provides a good
approximation of the posterior

Hastings ratio

new parameter
values

$$R = \frac{P(\text{E}^* | \text{0101... 1101... 0100...})}{P(\text{E} | \text{0101... 1101... 0100...})}$$



=

=

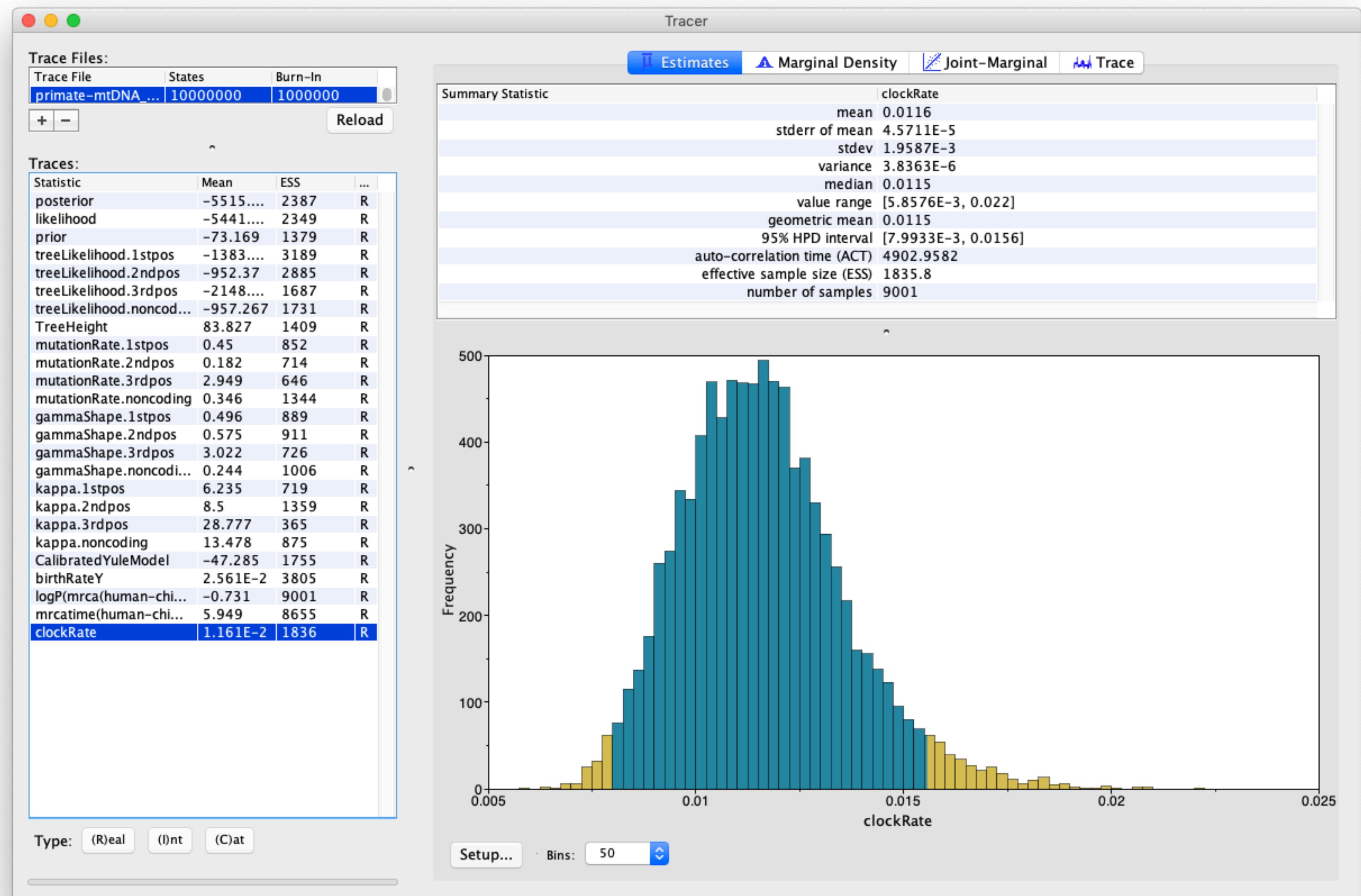
$$\frac{\cancel{P(\text{0101... 1101... 0100...})} P(\text{E}^* | \text{0101... 1101... 0100...})}{\cancel{P(\text{0101... 1101... 0100...})} P(\text{E} | \text{0101... 1101... 0100...})}$$
$$\frac{P(\text{0101... 1101... 0100...}) P(\text{E}^* | \text{0101... 1101... 0100...})}{P(\text{0101... 1101... 0100...}) P(\text{E} | \text{0101... 1101... 0100...})}$$

The marginal
probability of the
data cancels out

All we're left to
calculate is the
likelihood ratio and
the prior odds ratio

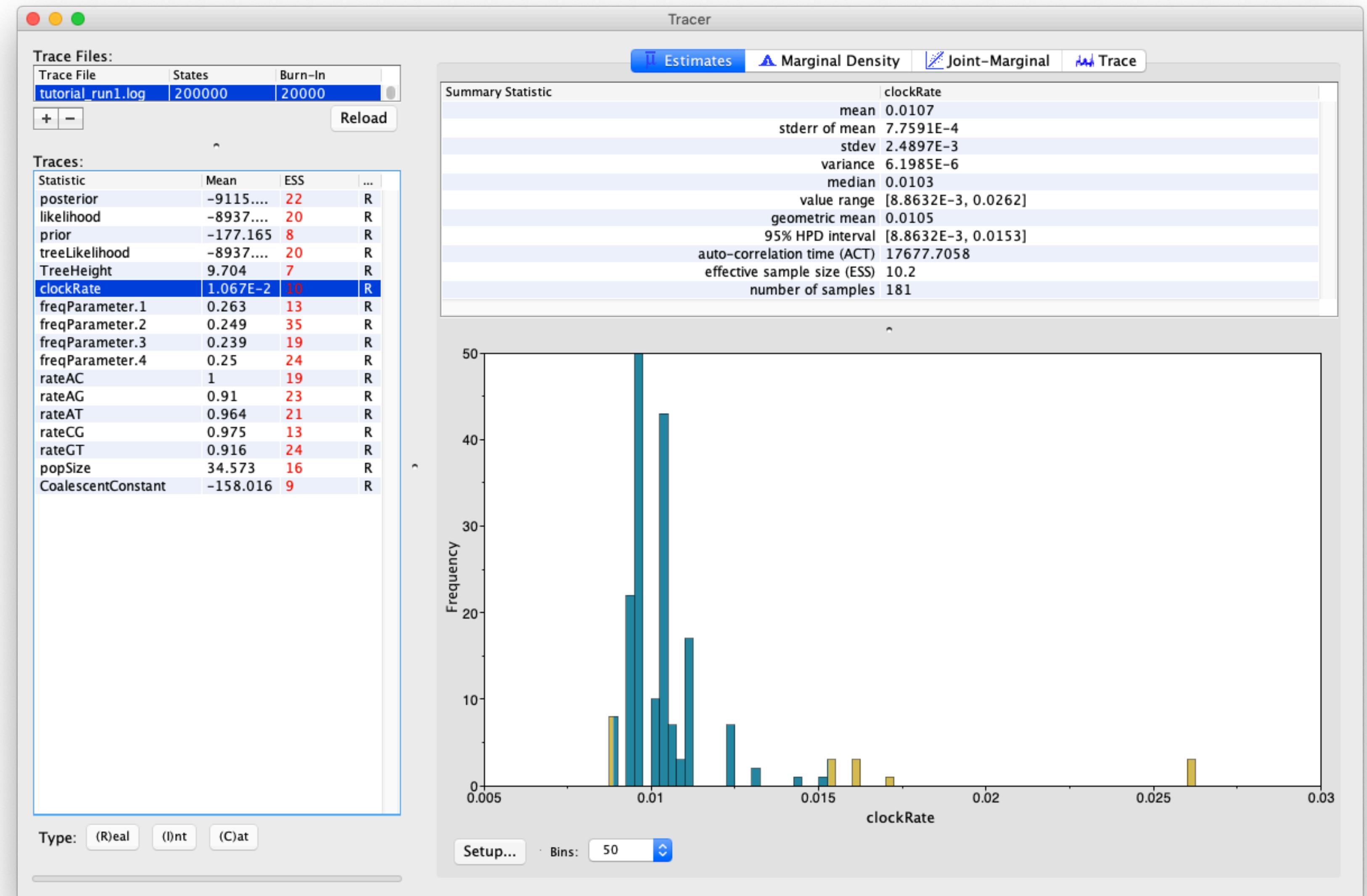
Summarising the posterior

Tracer is an amazing program for exploring MCMC output



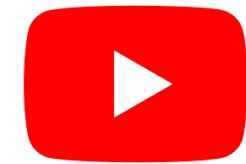
Summarising the posterior

Tracer is an amazing program for exploring MCMC output



Exercise

Homework



[Phylogenetics primer part 3a: Introduction to Bayesian statistics Paul Lewis](#)



See previous week's reading