Phylogenetics

Paul Lewis lecture series
Q & A

RL-V3 MPP

Rachel Warnock

23.06.25

Q&A Phylogenetics primer part 1 by *Paul Lewis*

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

1. In your own words, how would you describe the terms conditional probability and likelihood?

Conditional probability → the probability of an event, dependent on the value of some other event

Likelihood → the probability of our observations given a set of assumptions (i.e., the model) and parameter values

2. Why do we calculate likelihoods on a log scale?

→ Because likelihoods can get incredibly small — see for yourself using R!

3. How does the probability of transitioning from one character state to another (e.g., from A to T) change over time?

use the Transition Probability tool to explore this further

→ the probability of change increases with time

Felsenstein's pruning algorithm

4a. What do we need this for?

→ to calculate the likelihood of a tree (given an alignment and a substitution model, taking into account all possible ancestral states at every node)

4b. Can you describe the gist of Felsenstein's pruning algorithm?

For a good description of Felsenstein's pruning algorithm see Section 8.8 of Phylogenetic Comparative Methods by *Harmon*

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

Q&A Phylogenetics primer part 3a by *Paul Lewis*

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

1. In your own words can you describe each component of Bayes' rule? Which parts are difficult to understand?

Recap

Bayes' theorem

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

Bayes' theorem

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

Likelihood

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

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

Reflects our combined knowledge based on the likelihood and the priors

posterior

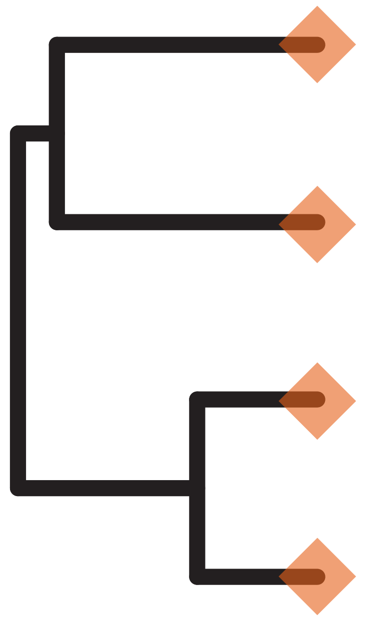
$$\Pr(\text{model} \mid \text{data}) = \frac{\Pr(\text{data} \mid \text{model}) \Pr(\text{model})}{\Pr(\text{data})}$$

Components used to infer trees

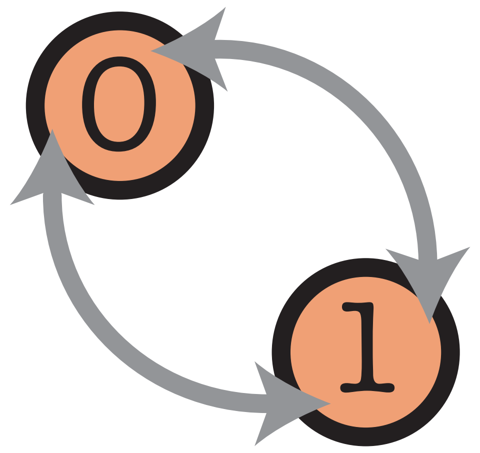
without considering time

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

data
sequences or
characters

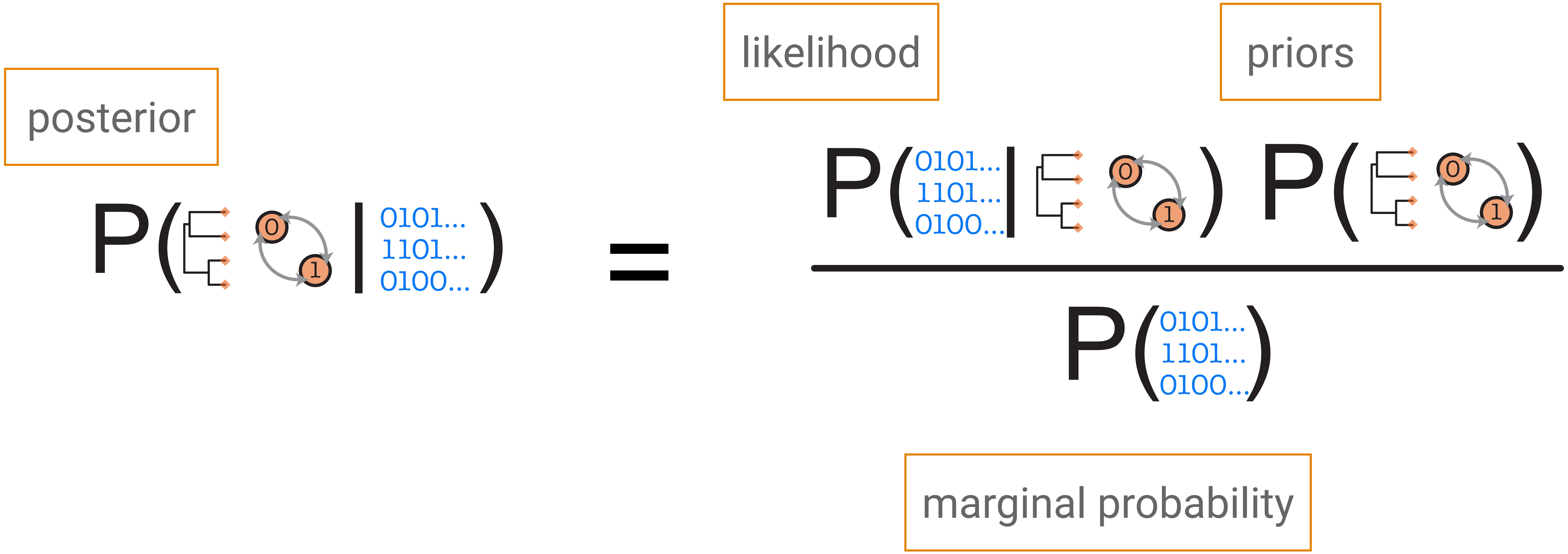


tree
topology and
branch lengths



substitution
model

Bayesian tree inference



1. In your own words can you describe each component of Bayes' rule? Which parts are difficult to understand?

Posterior $\propto$ Likelihood × Priors

The posterior probability is **proportional** to the numerator, i.e., the likelihood **times** the prior

2. Can you describe the difference between discrete and continuous variables?

discrete variables → have a set of predefined values, an integer

e.g., having a tail vs. not

continuous variables → can take on any real number value within a range

e.g., length, body mass

2. Can you describe the difference between probabilities and probability densities?

probabilities → a probability takes a singular value, e.g. $P = 0.5$

probability densities → a range of values represented by a distribution

3. What is the difference between vague vs. informative priors?

→ a vague prior is used for parameters where we have little clue what the true value is

e.g., it could be anything between 0 and infinity

→ an informative prior is used for parameters where we have some good existing knowledge about what the parameter value could be

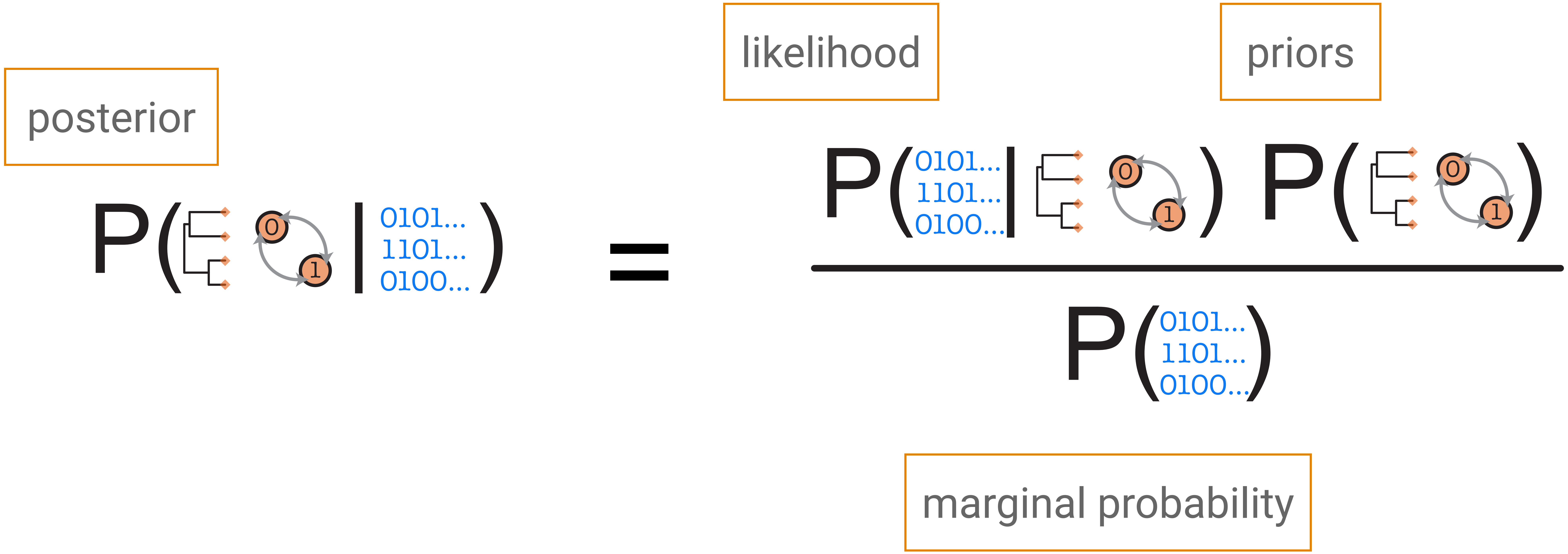
e.g., maybe we already know the rate of evolution among a well studied group, so we could use a prior distribution with a mean equal to the known value and add a small variance

4. What is the aim of MCMC in Bayesian inference?

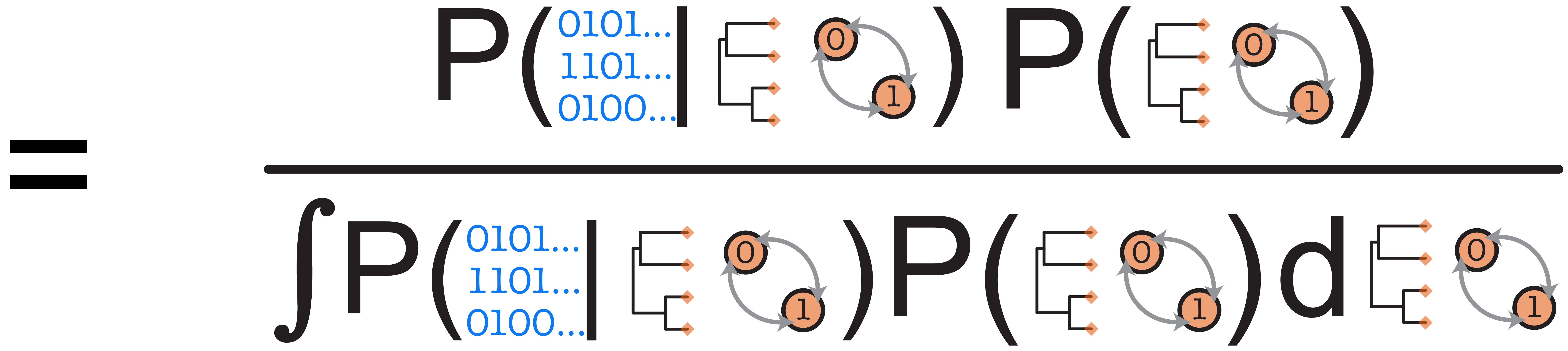
→ the aim is to **approximate** the posterior distribution

The posterior distribution is hard to calculate analytically (i.e., exactly), so we use MCMC to traverse the parameter space and at each step calculate the likelihood *x* the prior, spending time in different regions of the parameter space in proportion to their posterior probability - this means, we spend most time in areas with the highest posterior probability

Bayesian tree inference

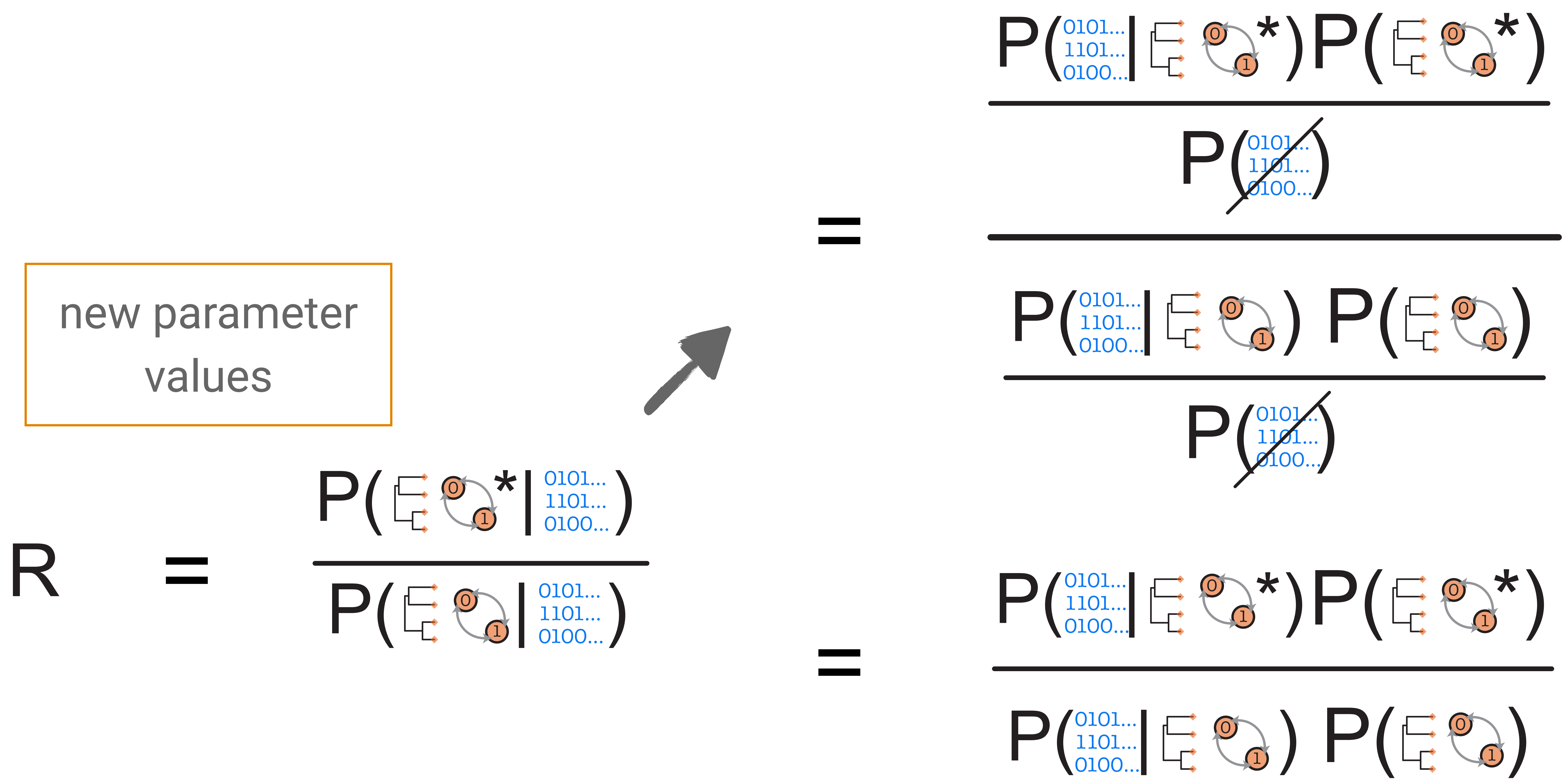


Bayesian tree inference



this part is incredibly difficult to calculate!

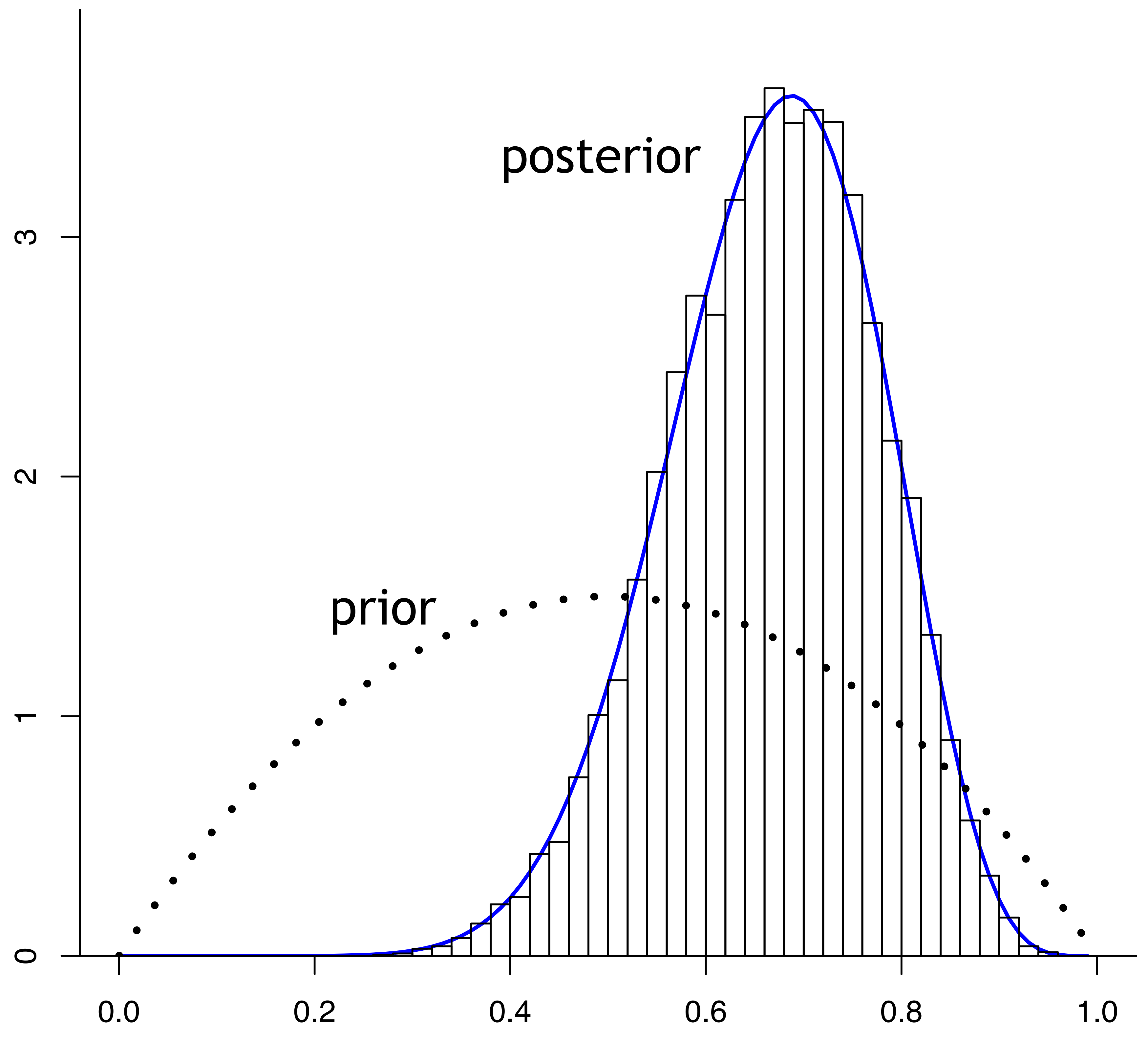
Hastings ratio



The marginal probability of the data cancels out

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

What is Markov chain Monte Carlo (MCMC)?



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

Q&A Phylogenetics primer part 3b by *Paul Lewis*

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

1. How are steps chosen in an MCMC analysis?

→ these depend on the type of parameter and the landscape of the parameters space

2. Give an example of a parameter you would estimate under each of the following prior distributions and try to state why?

Gamma distribution

Lognormal distribution

Beta distribution

Dirichlet distribution

3. Why do we sometimes need to calculate the marginal likelihood?

→ this is required for model testing within a Bayesian framework

4. What is the difference between a hierarchical model and a non-hierarchical model?

→ in a hierarchical model different components of the model are nested, i.e., different models can be joined together to model different processes that apply to the data