Bayesian Phylogenetics

Analytical Paleobiology

Rachel Warnock

13.08.24

Objectives for today and tomorrow

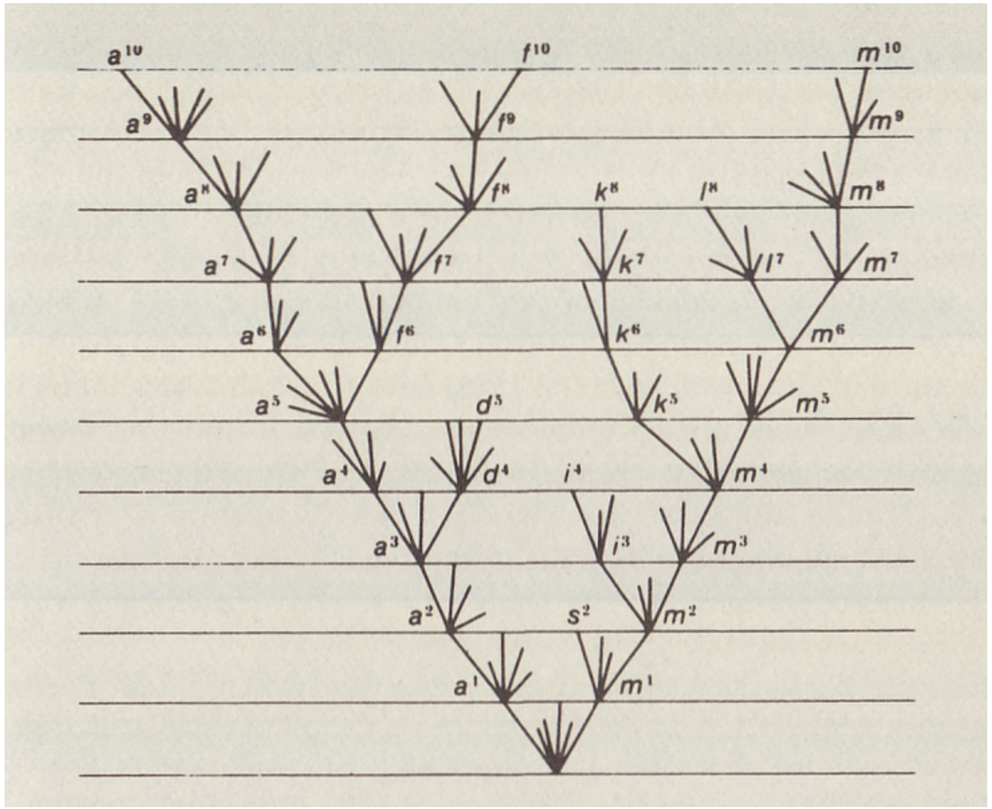
Intro to Bayesian phylogenetics in paleobiology

- Tree building
- Substitution models
- Dating trees
- Clock models
- Tree models
- Diversification rates
- Morphological models

Please ask questions!

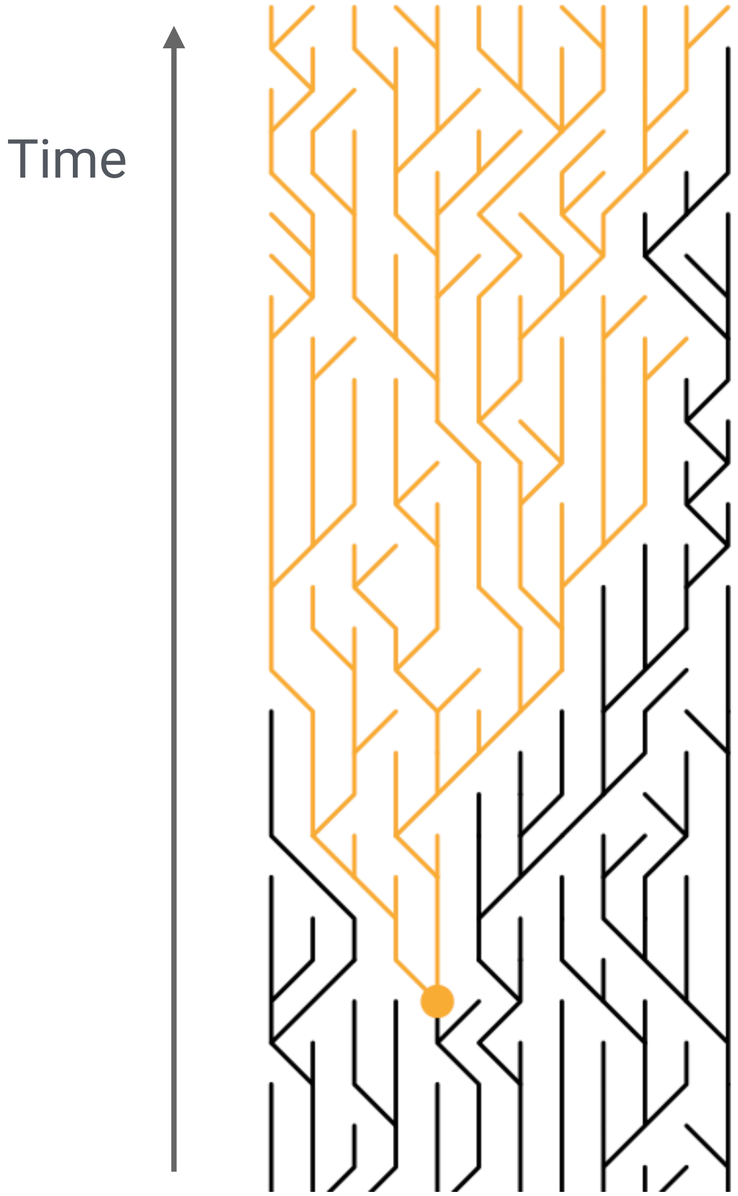
Today’s objectives

- Intro to RevBayes
- Bayesian tree inference
- Morphological models



Time tree from Darwin’s *Origin of Species*

What is phylogenetics?

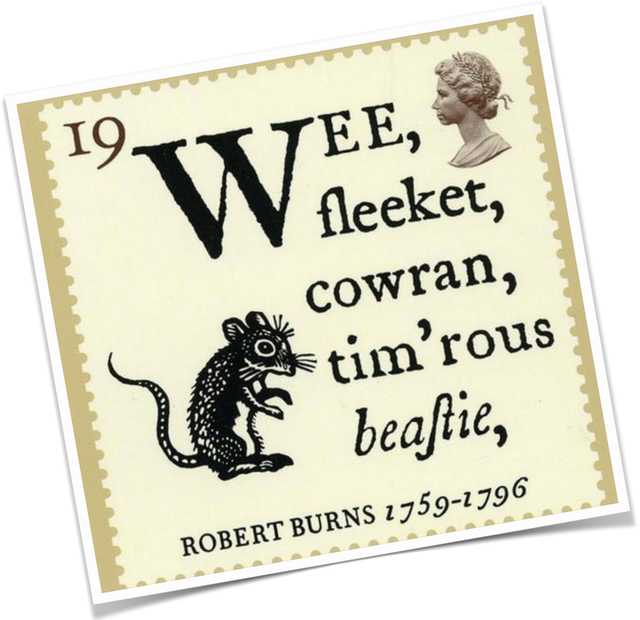


- populations
- species
- viruses
- cells
- languages

Data

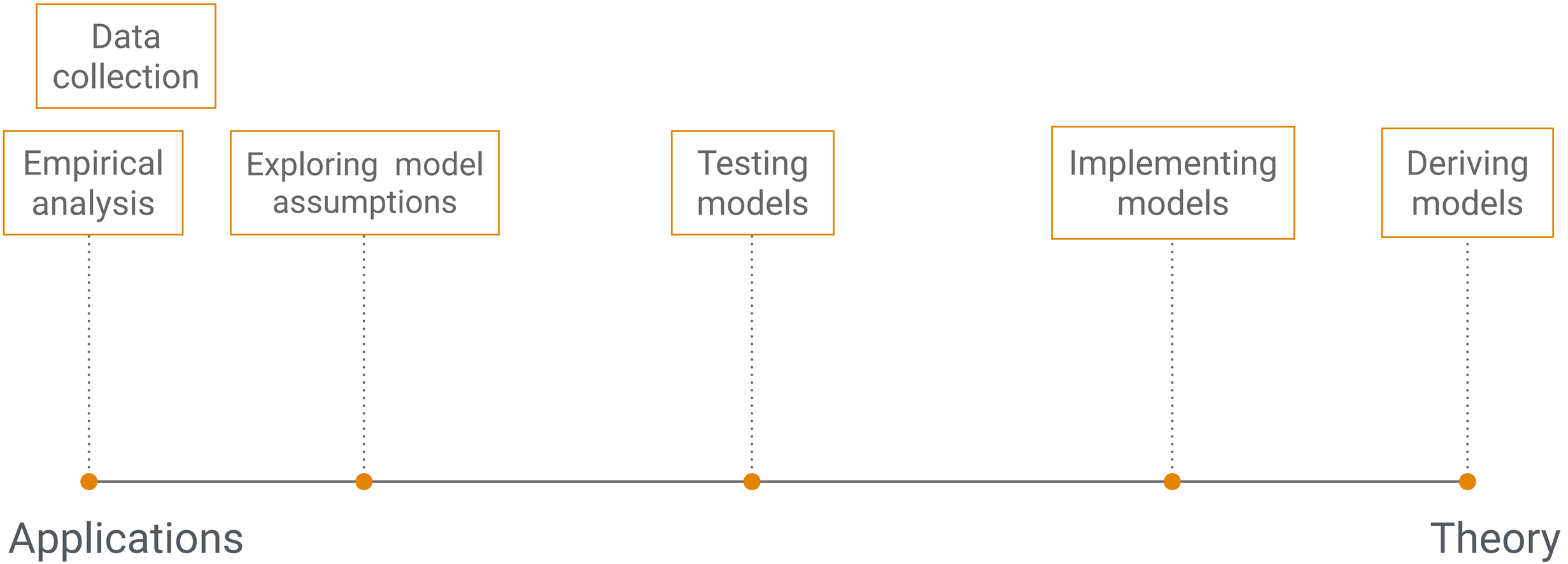
- DNA
- morphology
- words

In this course we mainly focus on trees that include one representative per species



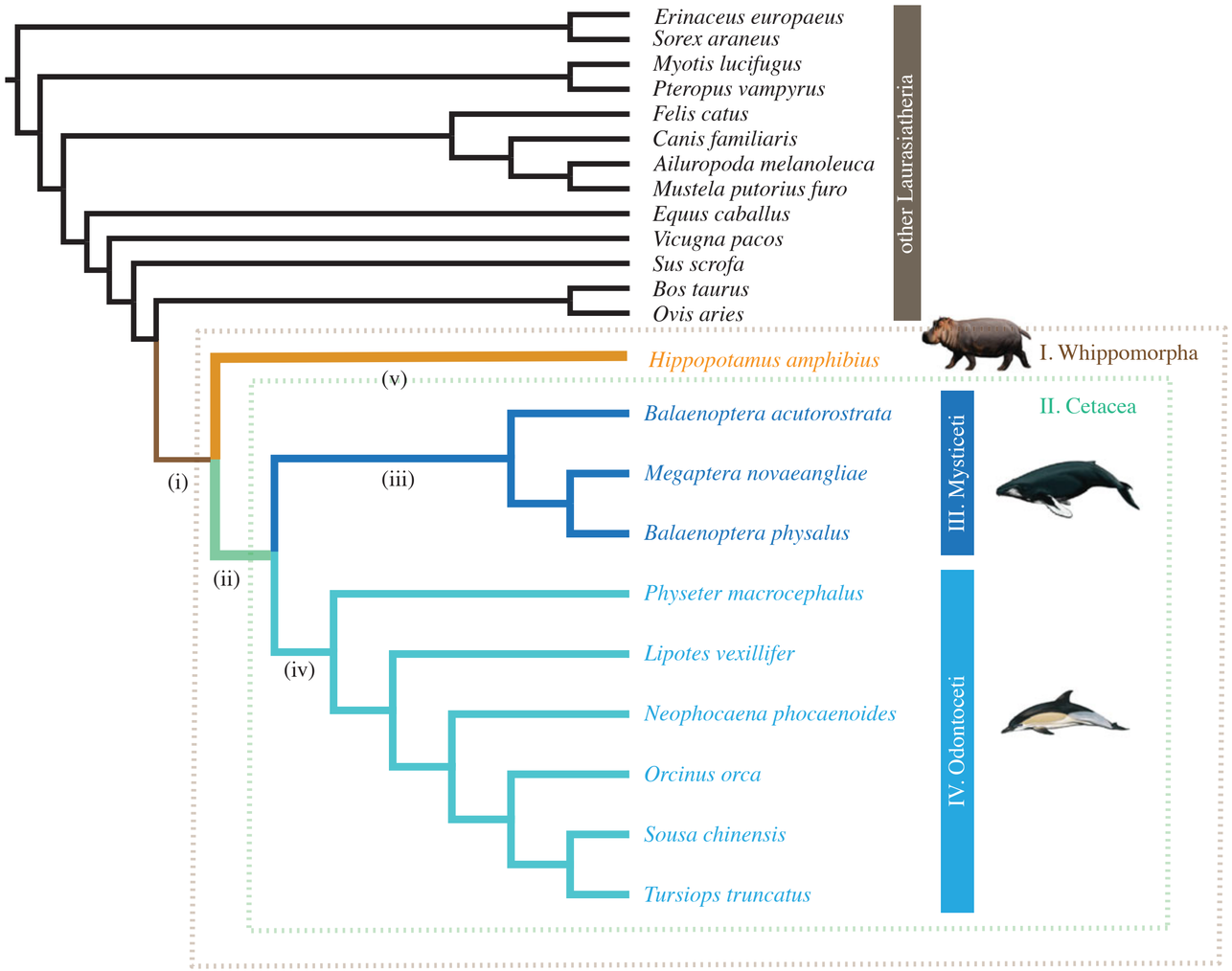
Scots poem - also the BEAST2 logo!

Research topics in phylogenetics



Trees in paleobiology

What can we learn from trees?



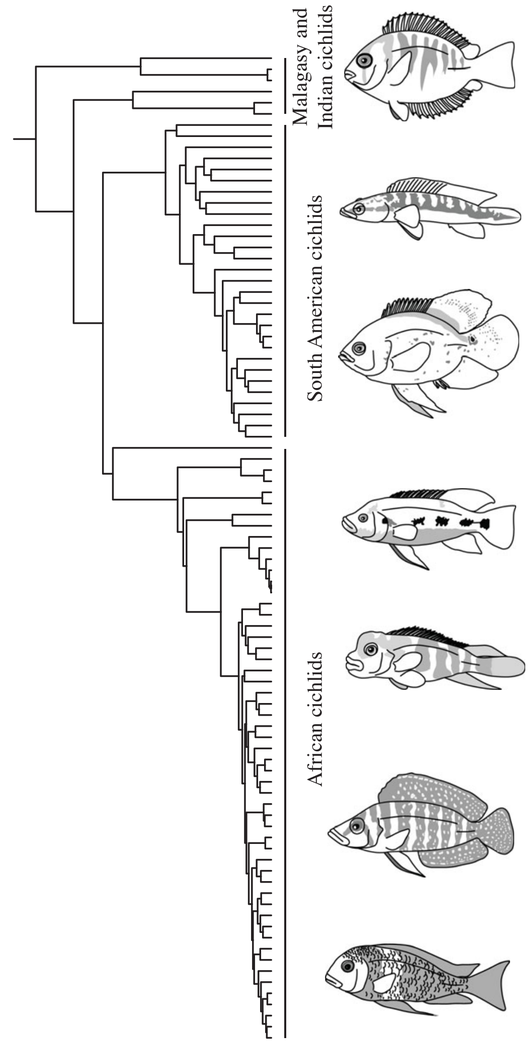
How are our favourite species related?

Does the phylogeny support the taxonomy?

Tsagkogeorga et al. (2015)

What can we learn from trees?

- Evolutionary relationships



What can we learn from trees?

- Evolutionary relationships
- Timing of diversification events
- Geological context
- Rates of phenotypic evolution
- Diversification rates

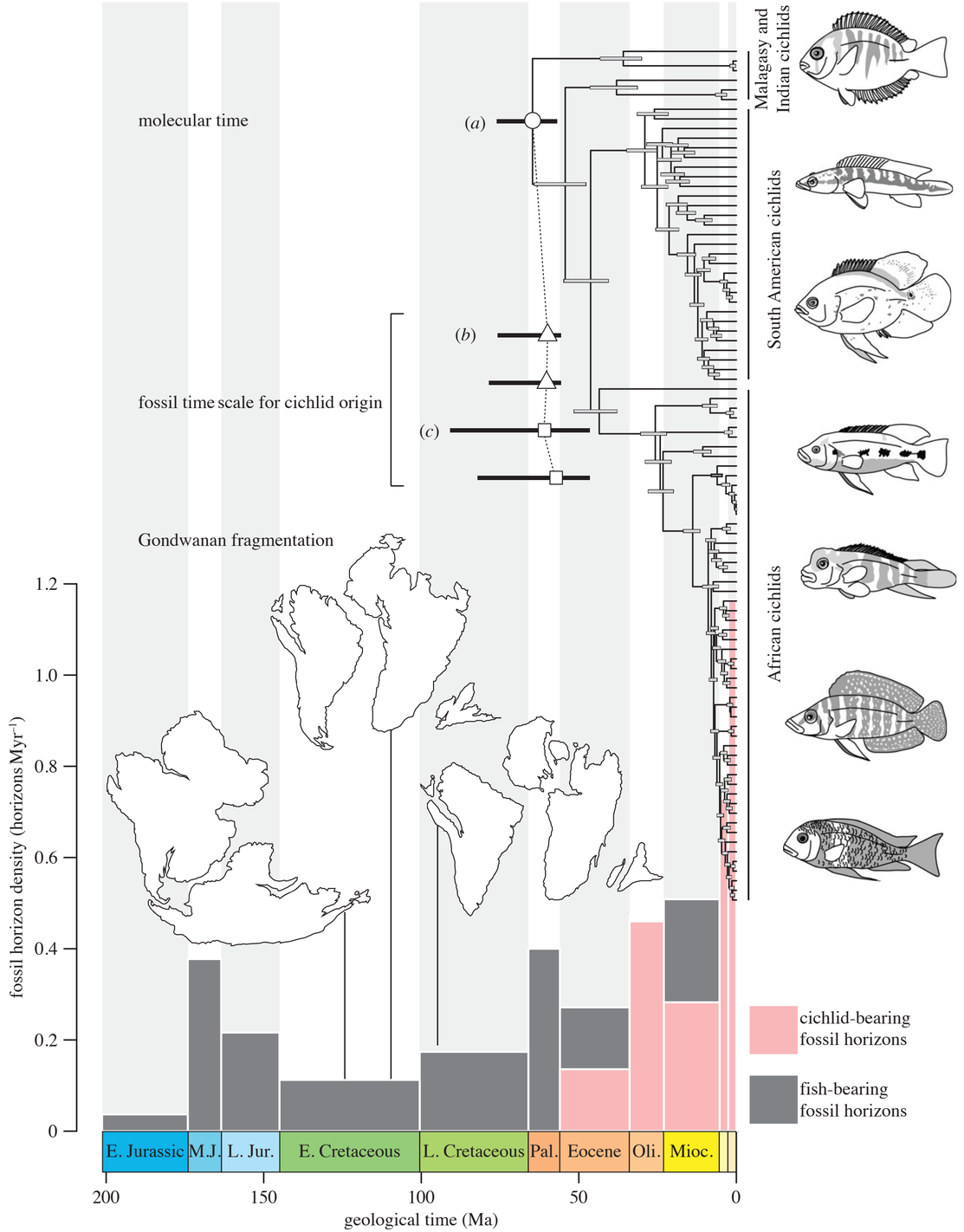


Image adapted from *Friedmann et al.* (2013)

Phylogenetics

Phylogenetics aims to reconstruct the phylogeny of individual samples based on molecular or morphological character data

A phylogeny captures part of evolutionary history that is otherwise not directly observable

Phylodynamics aims to quantify the processes that gave rise to the tree, e.g., speciation, extinction

What do we mean by model?

(the following is my take on things — intended to be useful but not definitive)

What is a statistical model? When is an equation a model?

What is a mechanistic model?

What is the difference between an algorithm and a model?

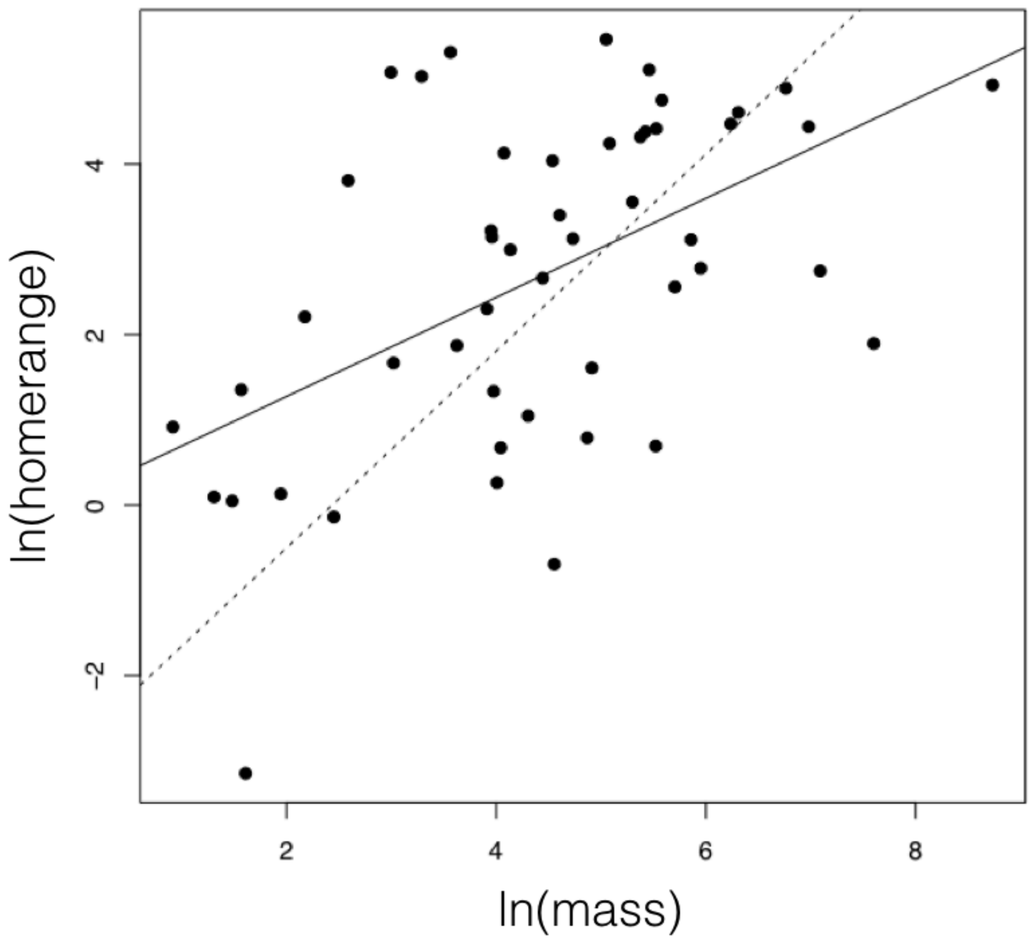
A statistical model is a type of model that includes a set of assumptions about the data-generating process

It should be possible to simulate data under the assumptions of the model

If we're lucky, we might *also* be able to estimate parameters under the model*. This isn't always possible because some models are too complex

*A fancy way of saying this is, "we can perform inference under the model"

An example



The solid black line is a linear regression line

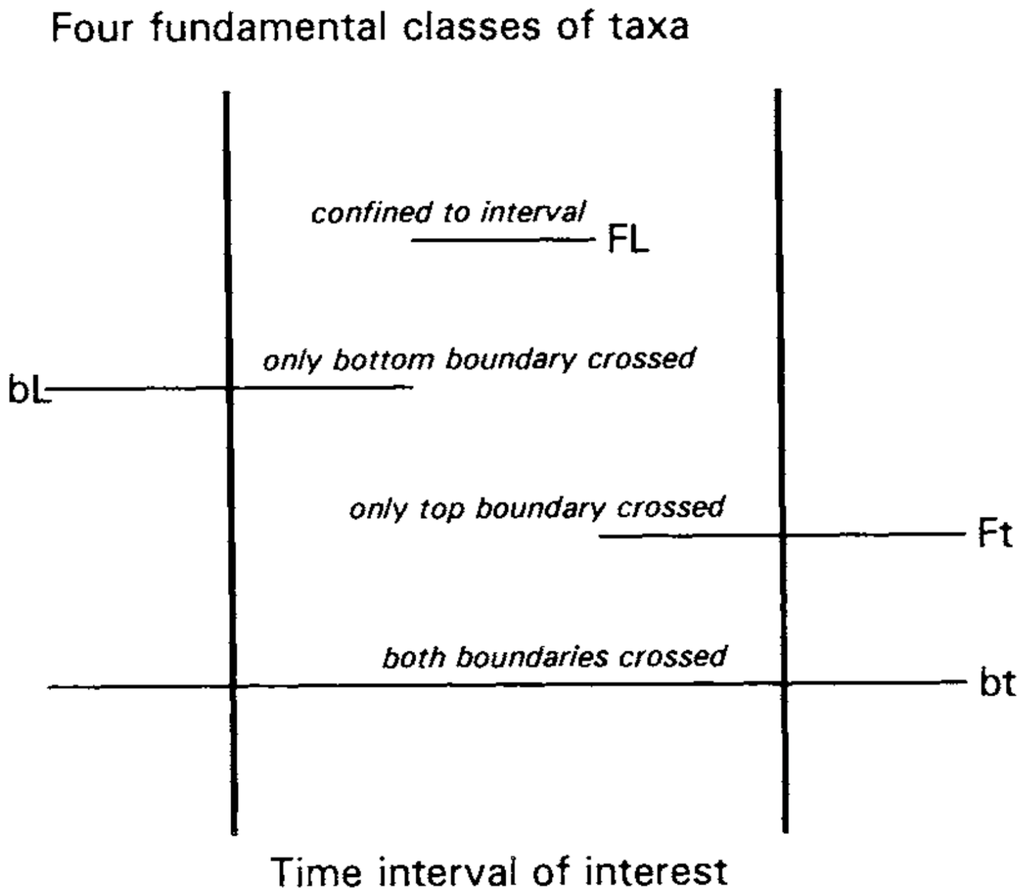
We can estimate the parameters of the regression model

$y = X\beta + \varepsilon$

It's also straightforward to simulate data under this model

Image source *Harmon* (2019)

Non model-based approaches are still useful



Foote (2000)

The boundary-crosser and three-timer metrics are *not* models

They provide a clever way of approximating origination and extinction rates (and often perform well), but don't describe the data generating processes

Mechanistic or process based models are based on ‘physical principles’. They describe the data as a function of a set of parameters that have a tangible biological or geological meaning

A regression model is not mechanistic – it describes the relationship between *x* and *y* but the parameters don’t have a biological meaning

Many models used in phylogenetics are mechanistic, e.g., they might include parameters for origination, extinction, or sampling

An algorithm is a precise rule (or set of rules) specifying how to solve some problem

```
i = 1
while  i < 11:
    print(i)
    i = i + 1
```

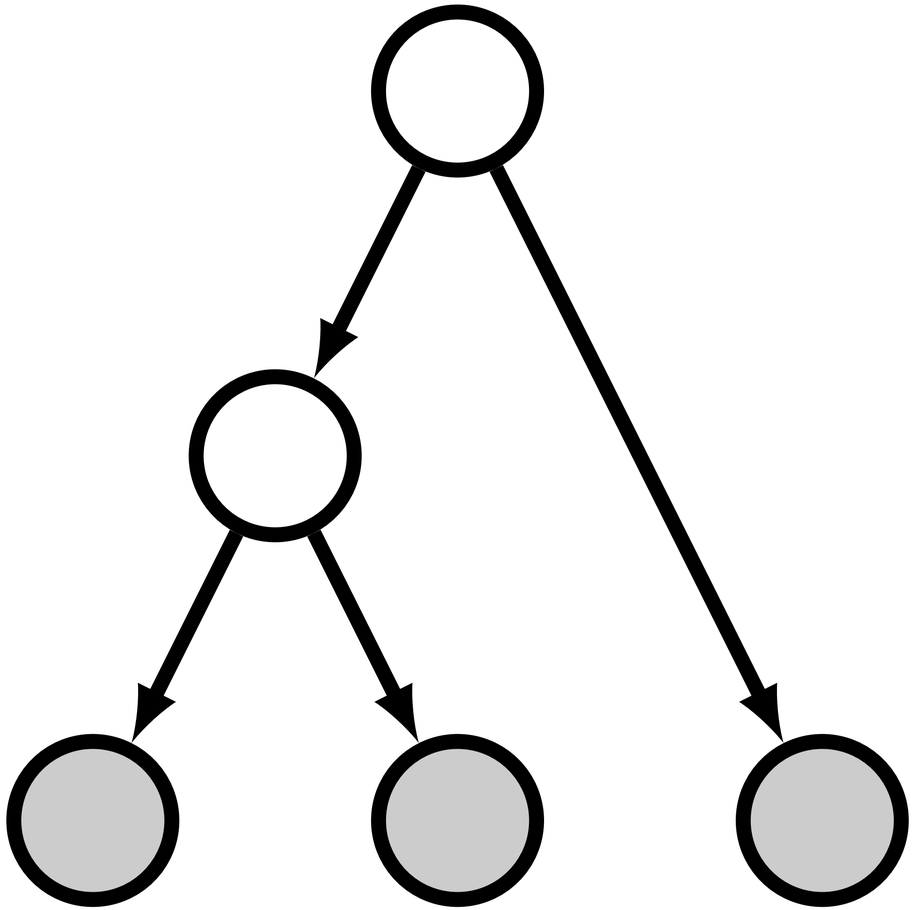
```
for i in range(1,11):
  print(i)
```

Used in phylogenetics for all sorts of tasks, e.g., traversing tree space

Mini reading group 

Next

- graphical models
- RevBayes
- Bayesian inference
- MCMC



How do we find the ‘best’ tree?

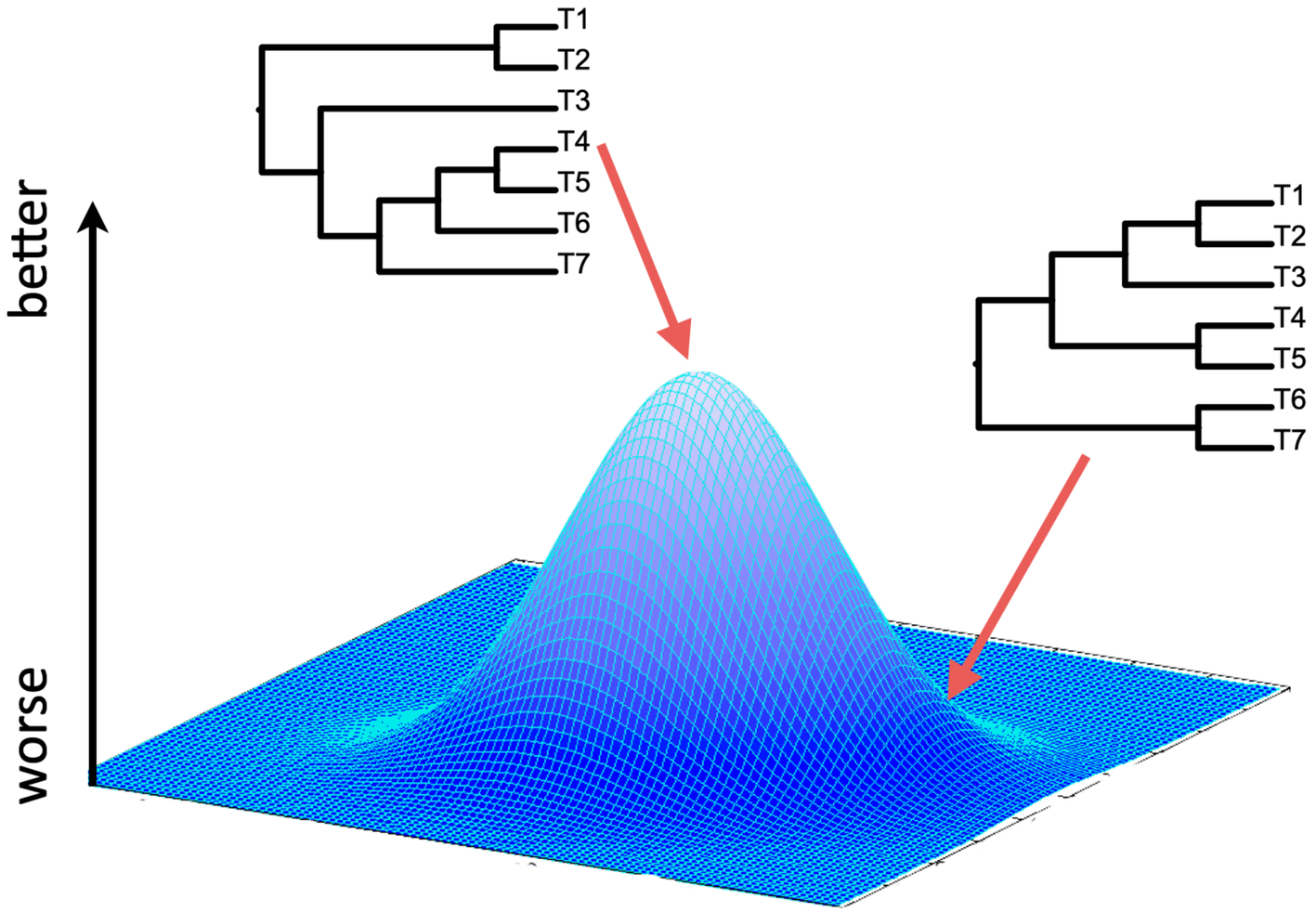


Image source Tracy Heath

It depends how you measure ‘best’

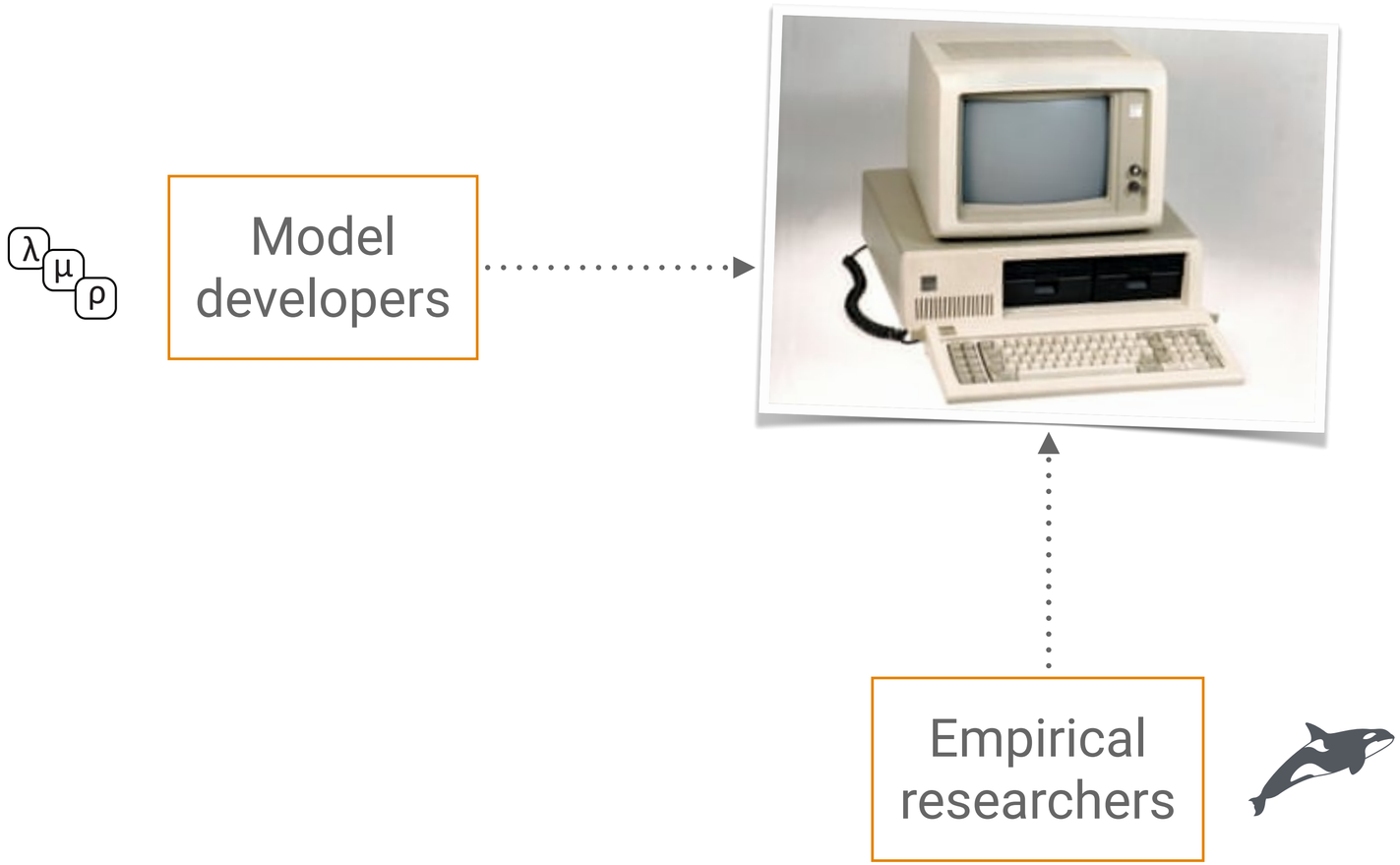
Method	Criterion (tree score)
Maximum parsimony	Minimum number of changes
Maximum likelihood	Likelihood score (probability), optimised over branch lengths and model parameters
Bayesian inference	Posterior probability, integrating over branch lengths and model parameters

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

RevBayes

Phylogenetic inference — the old way



What we might call a **“black box”** approach

Phylogenetic inference — a better way?

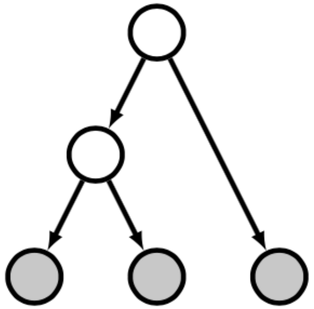
Model developers

Empirical researchers

The goal is to bring researchers with different expertise together, increase transparency, and do better research

revbayes.github.io

Download Tutorials Documentation Interfaces Workshops Jobs Developer



RevBayes

Bayesian phylogenetic inference using probabilistic graphical models and an interpreted language

About

RevBayes provides an interactive environment for statistical computation in phylogenetics. It is primarily intended for modeling, simulation, and Bayesian inference in evolutionary biology, particularly phylogenetics. However, the environment is quite general and can be useful for many complex modeling tasks.

RevBayes uses its own language, Rev, which is a probabilistic programming language like JAGS, STAN, Edward, PyMC3, and related software. However, phylogenetic models require inference machinery and distributions that are unavailable in these other tools.

The Rev language is similar to the language used in R. Like the R language, Rev is designed to support interactive analysis. It supports both functional and procedural programming models, and makes a clear distinction between the two. Rev is also more strongly typed than R.

RevBayes is a collaboratively developed software project.

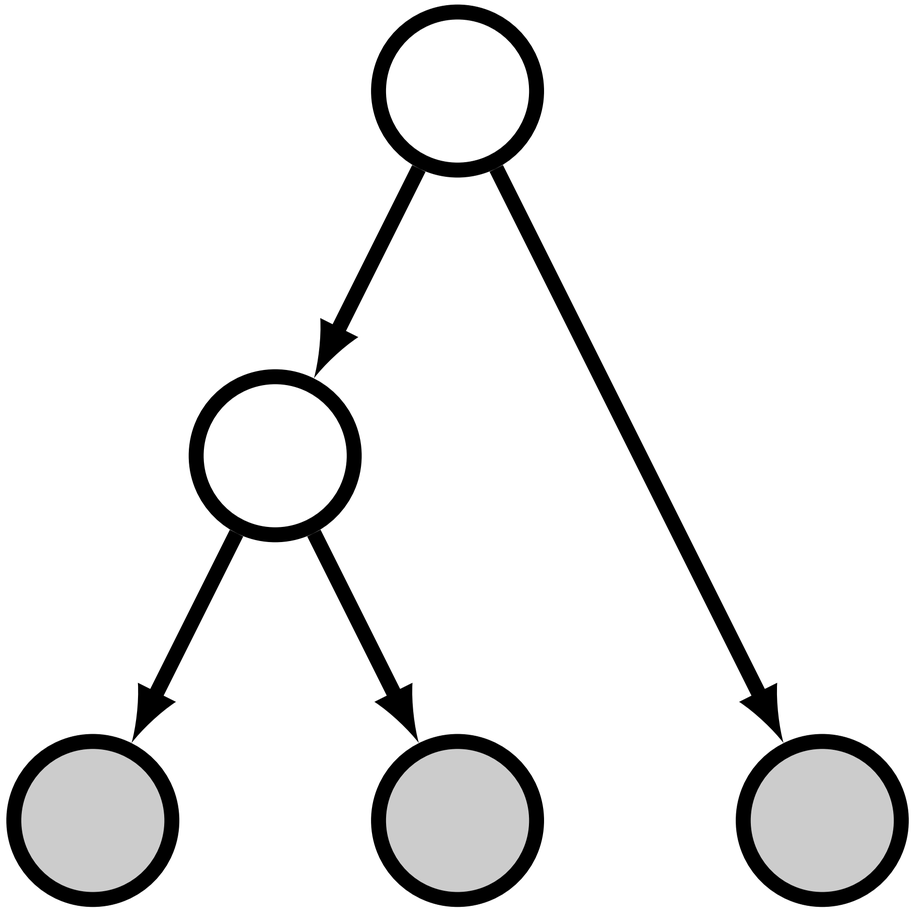
GitHub | License | Citation | Users Forum

Graphical models

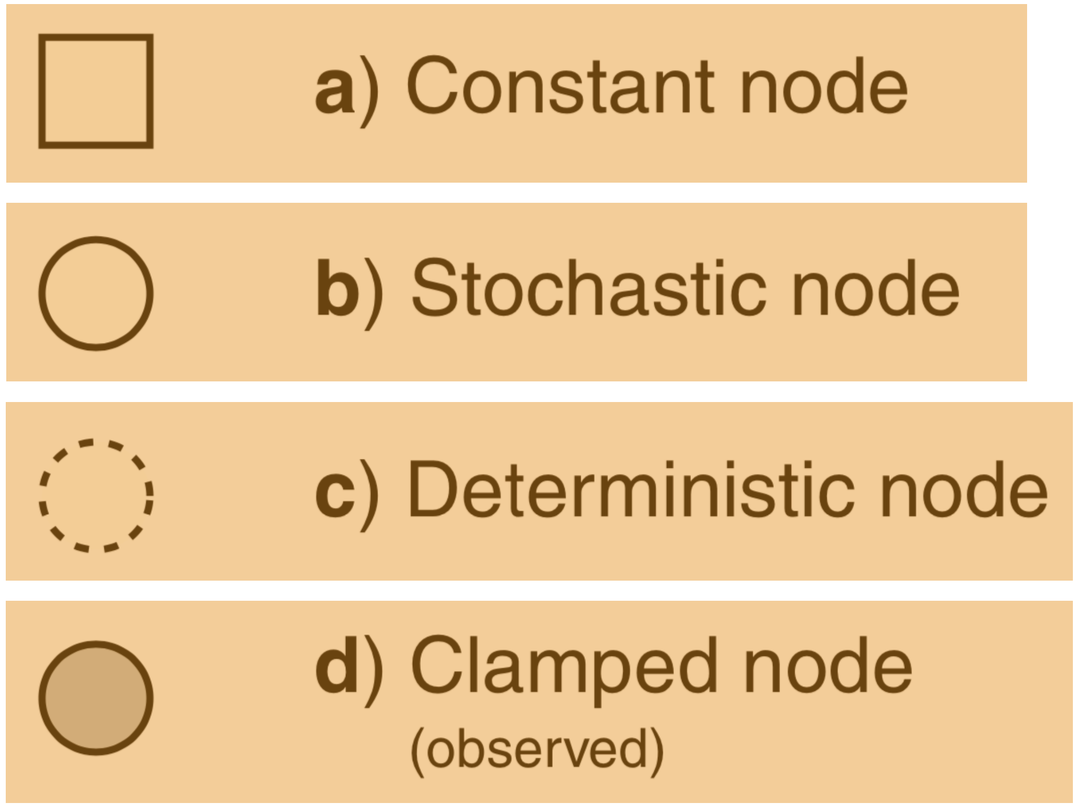
Graphical models

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

Depict the conditional dependence structure of parameters and other random variables



Types of variables (nodes)

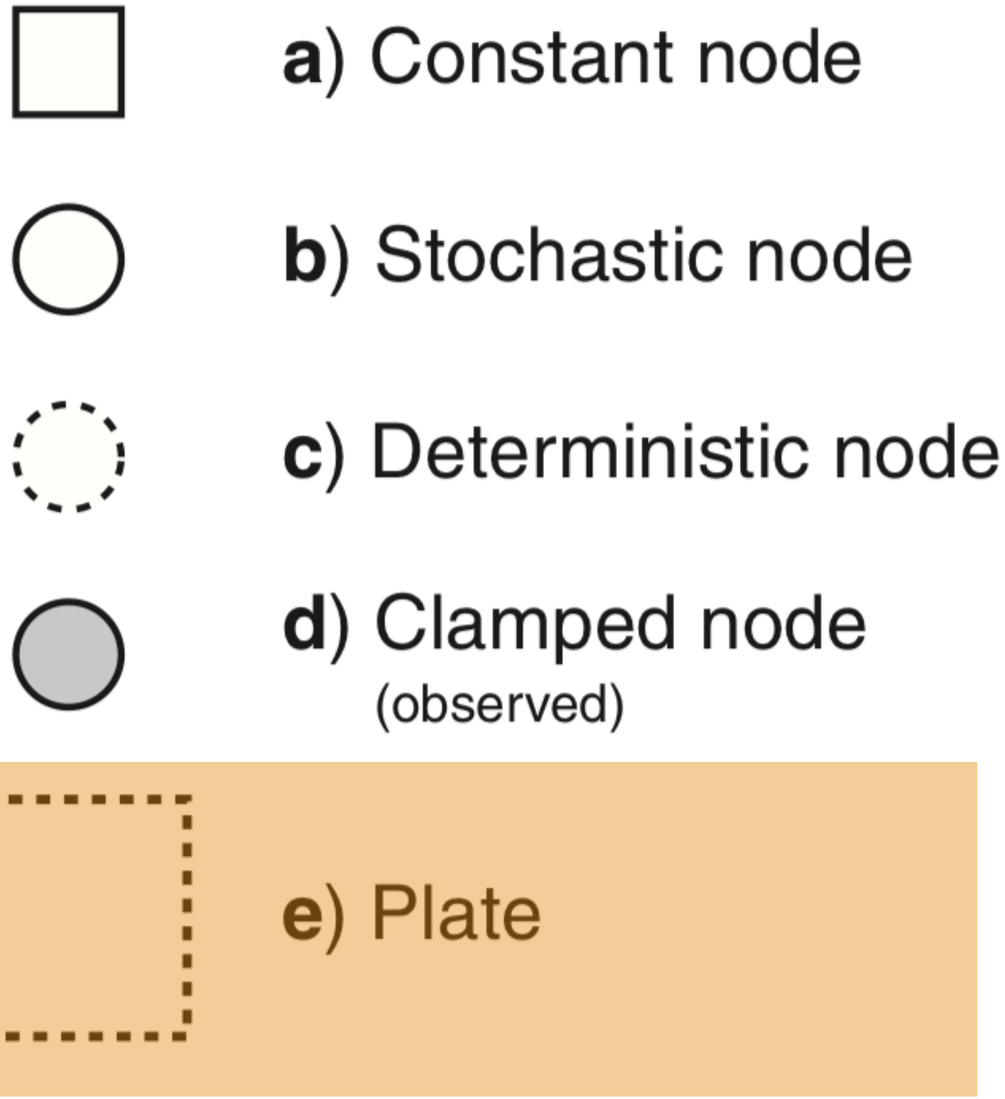


a. fixed value variables

b. random variables that depend on other variables

c. variables determined by a function applied other variables (transformations)

d. observed stochastic variables (data)



a. fixed value variables

b. random variables that depend on other variables

c. variables determined by a function applied other variables (transformations)

d. observed stochastic variables (data)

e. repetition over multiple variables (equivalent to a loop)

Specifying graphical models using the Rev syntax

Table 1: `Rev` assignment operators, clamp function, and plate/loop syntax.

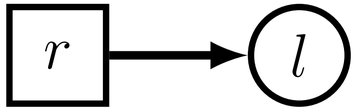
Operator	**Variable**
`<-`	constant variable
`~`	stochastic variable
`:=`	deterministic variable
`node.clamp(data)`	clamped variable
`=`	inference (*i.e.*, non-model) variable
`for(i in 1:N){...}`	plate

a)


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

a)

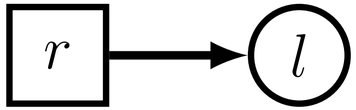
b)



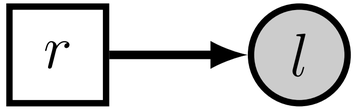

```
# constant node
r <- 10


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

a)

b)



c)

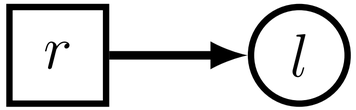



```
# constant node
r <- 10


# stochastic node
l ~ dnExp(r)


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

a)

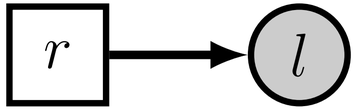

```
# constant node
r <- 10
```

b)



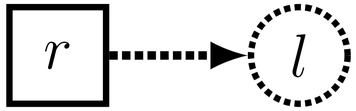

```
# stochastic node
l ~ dnExp(r)
```

c)




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

d)

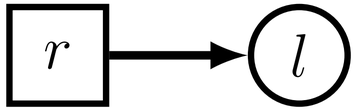



```
# deterministic node
l := exp(r)
```

a)

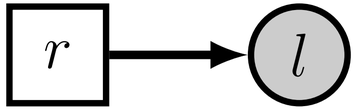

```
# constant node
r <- 10
```

b)



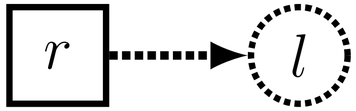

```
# stochastic node
l ~ dnExp(r)
```

c)



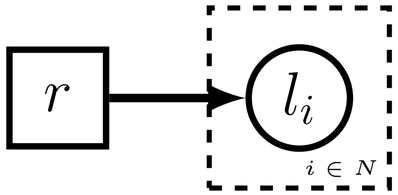

```
# stochastic node (observed)
l.clamp(0.1)
```

d)




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

e)




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

Exercise

Bayesian tree inference

Bayes' theorem

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

Bayes' theorem

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

Likelihood

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

Bayes' theorem

This represents our prior knowledge of the model parameters

Priors

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

Bayes' theorem

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

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

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

How do we find the ‘best’ tree?

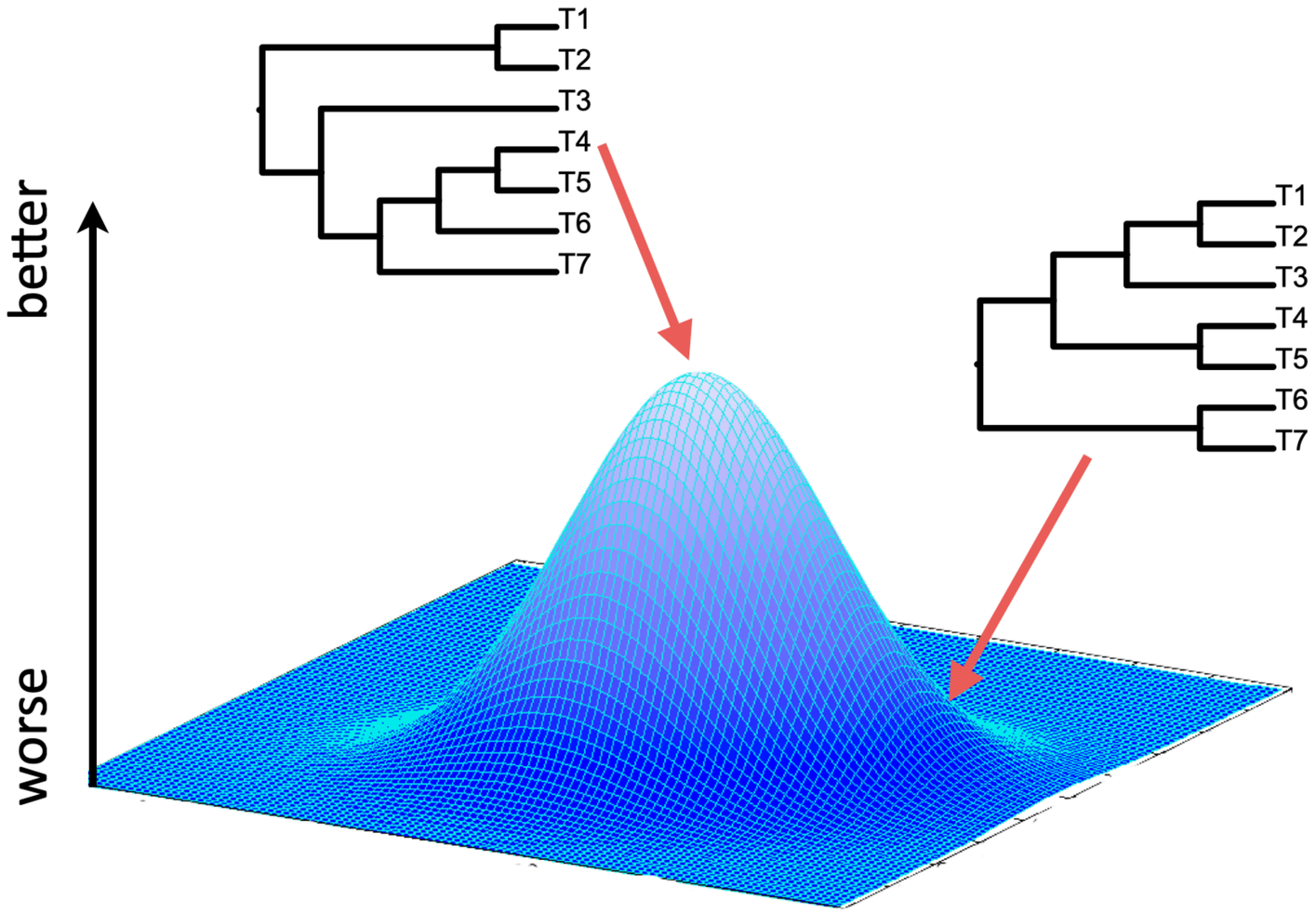


Image source Tracy Heath

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

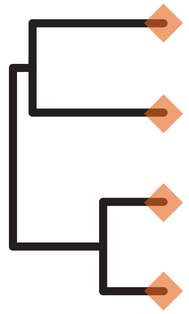
Note these are not the only approaches to tree-building but they are the most widely used

Components used to infer trees

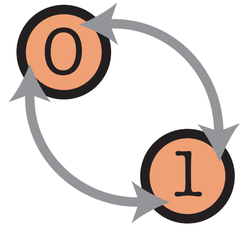
without considering time

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

data
sequences or characters

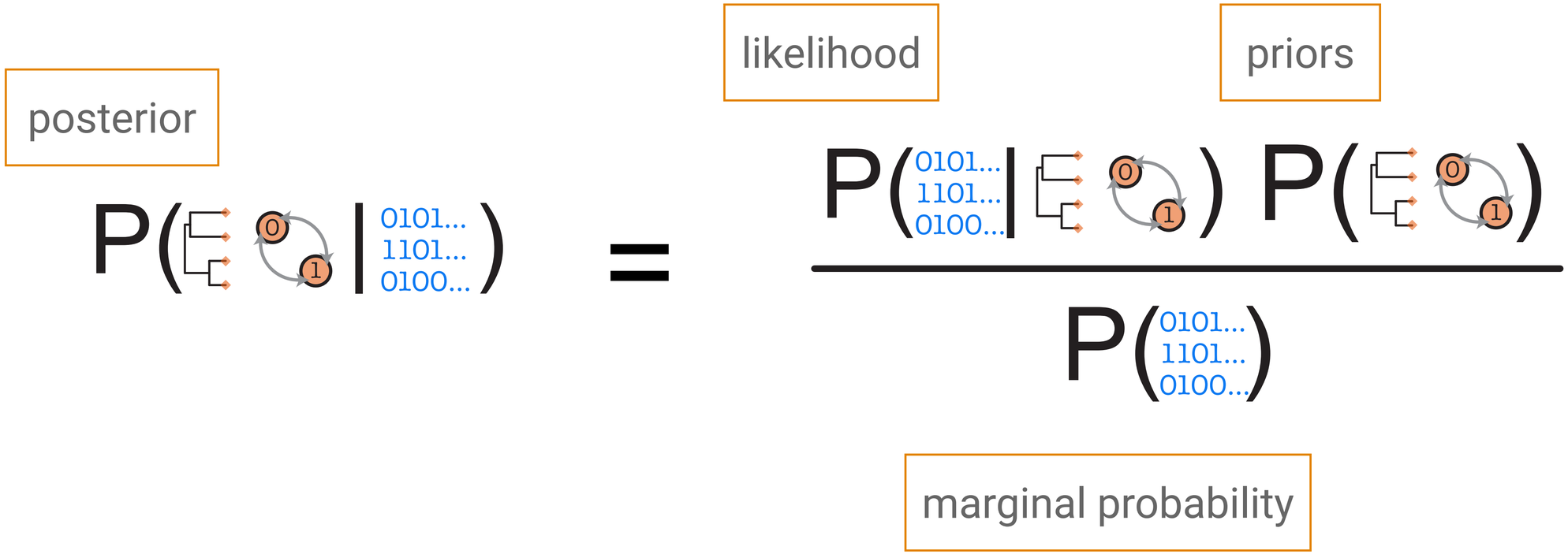


tree
topology and branch lengths



substitution model

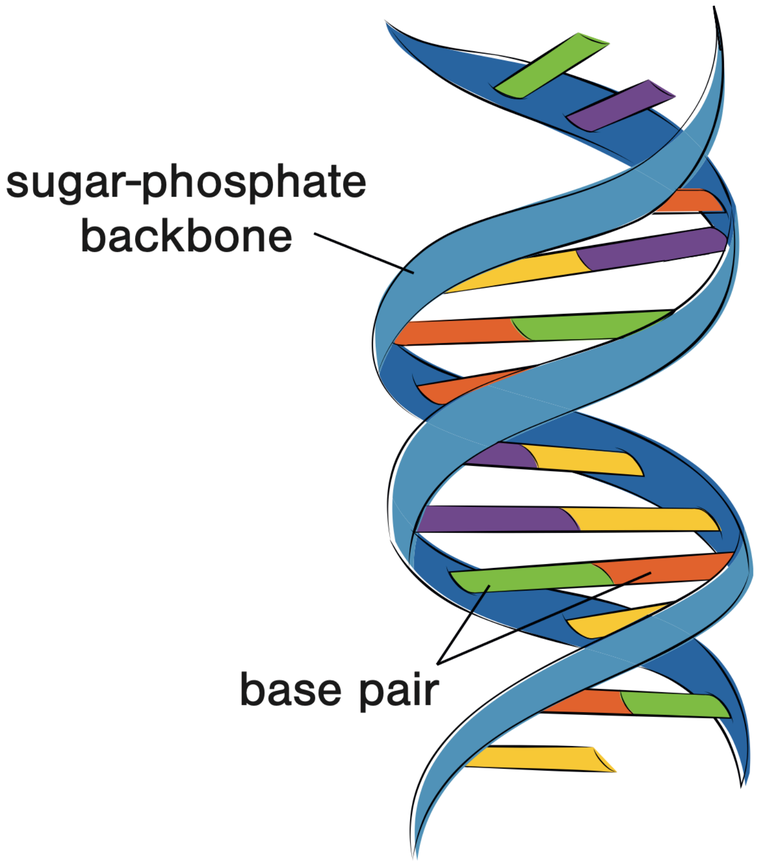
Bayesian tree inference

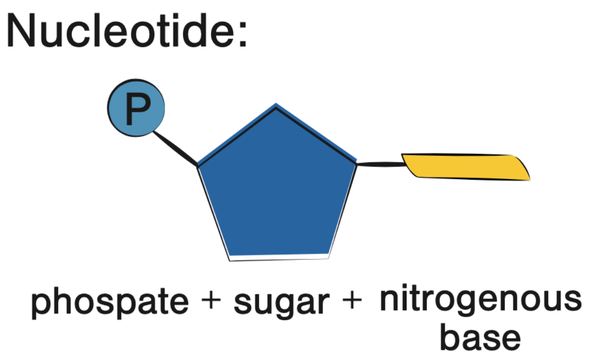


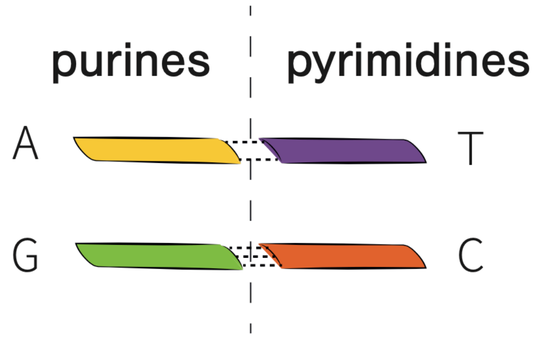
Likelihood and substitution models

Molecular evolution

Deoxyribonucleic acid (DNA)







Purines

- Adenine (**A**)
- Guanine (**G**)

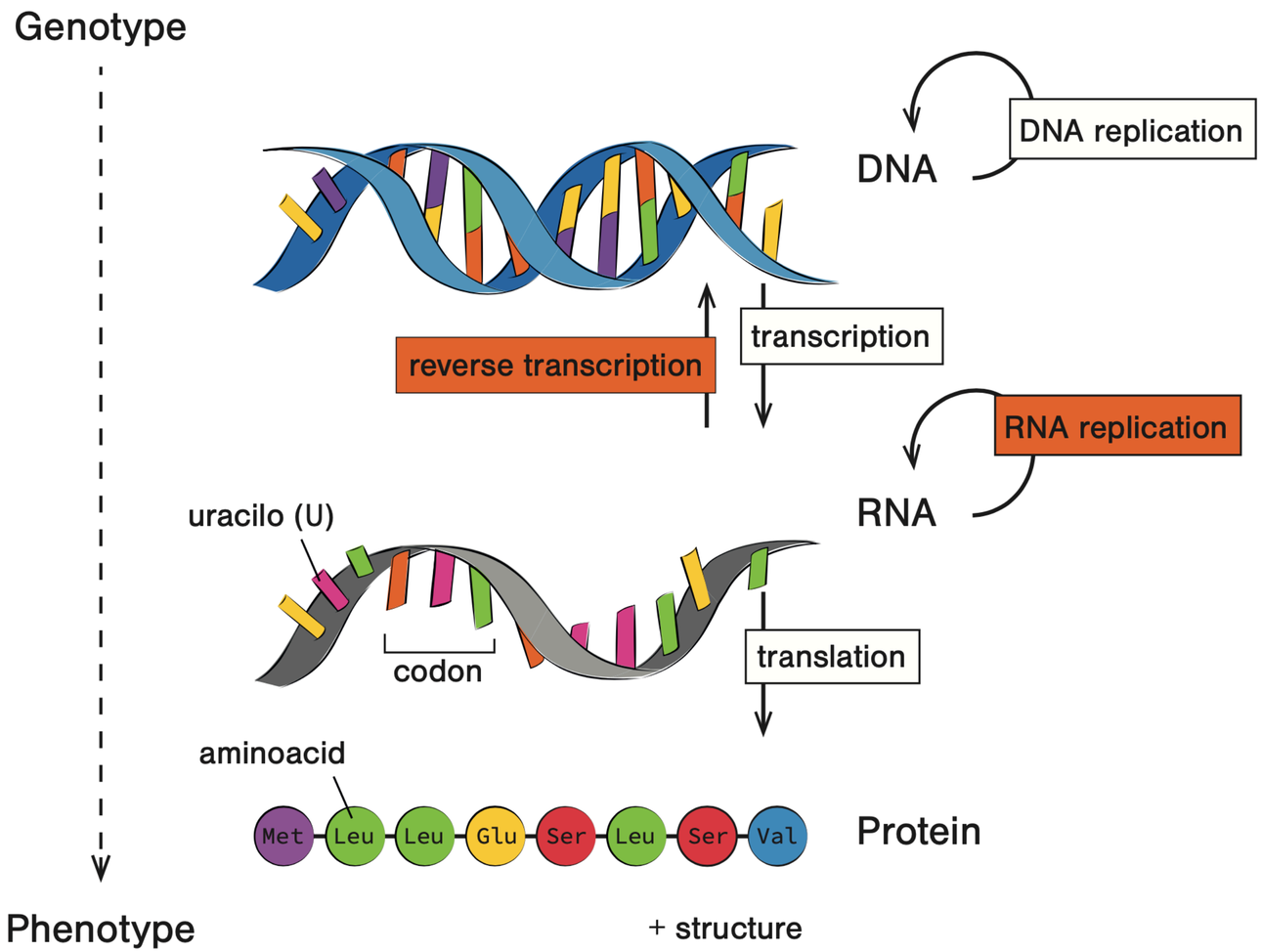
Pyrimidines

- Cytosine (**C**)
- Thymine (**T**)*

Image source Decoding Genomes *Stadler et al.* (2024)

*Replaced by uracil (U) in RNA

The central dogma of biology

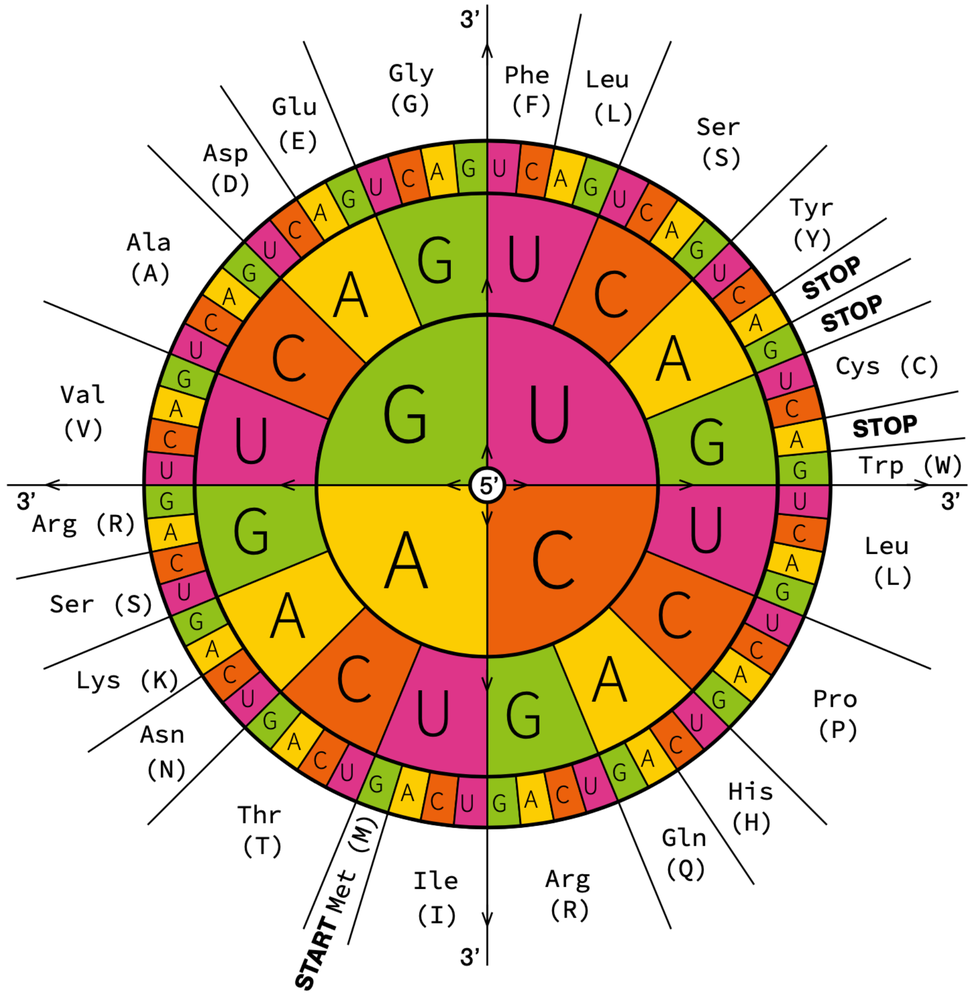


DNA → RNA → protein

Each group of 3 successive nucleotides in a gene is a codon that encodes an amino acid (or terminate translation)

Image source Decoding Genomes *Stadler et al.* (2024)

The universal genetic code



Amino acid	3-letter code	1-letter code
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

4^3 = 64 combinations

3 terminate translation

21 amino acids

Image source Decoding Genomes *Stadler et al.* (2024)

Mutation vs. substitution

Variation in genotypes (and in phenotypes) is due to errors that arise during DNA replication

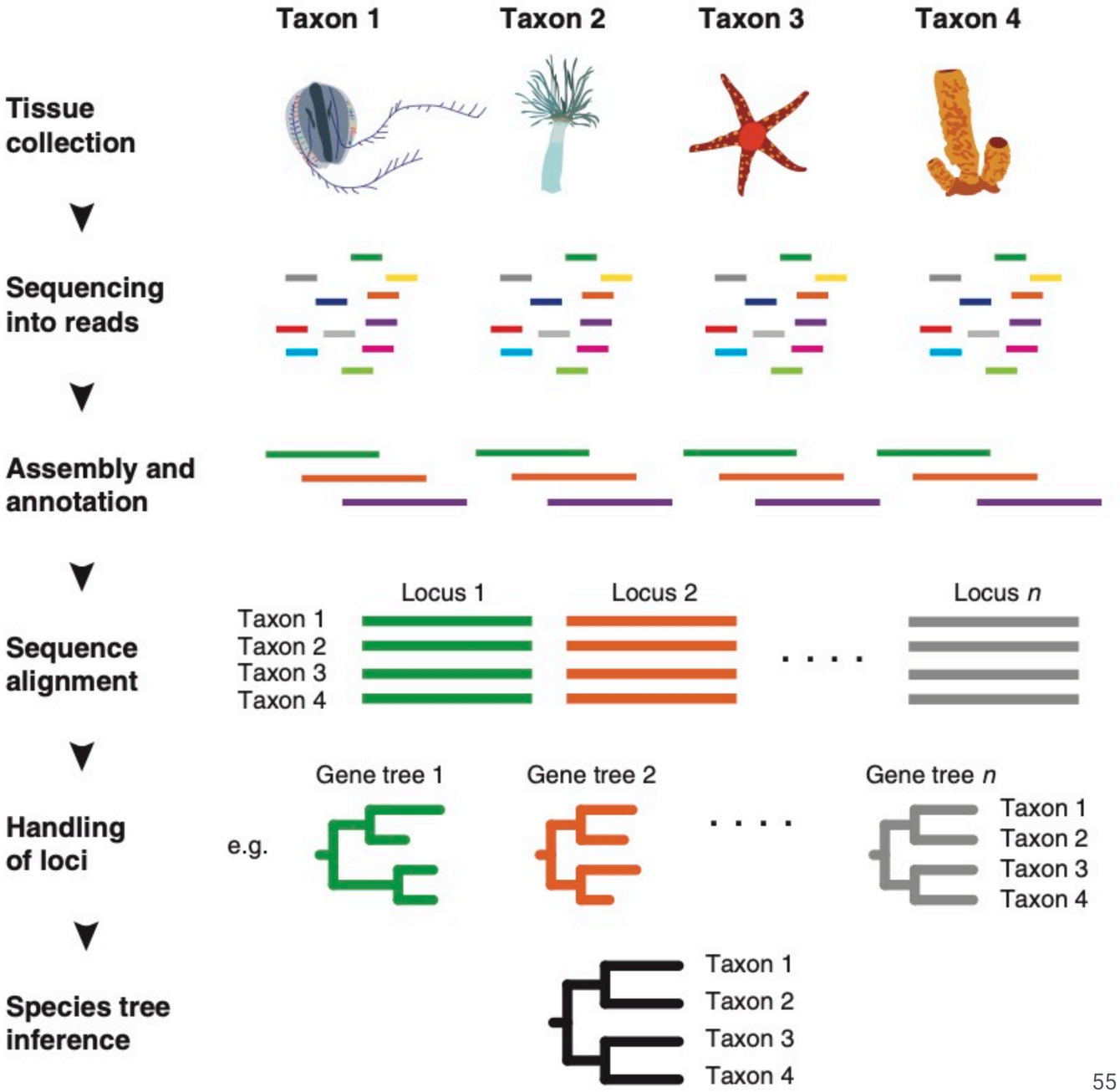
Individuals of the same species have identical characters at *most* positions in their genome (only 0.1% vary among humans)

Most mutations are repaired but can persist across generations

Mutations that spread throughout a population and become ‘fixed’ called substitutions

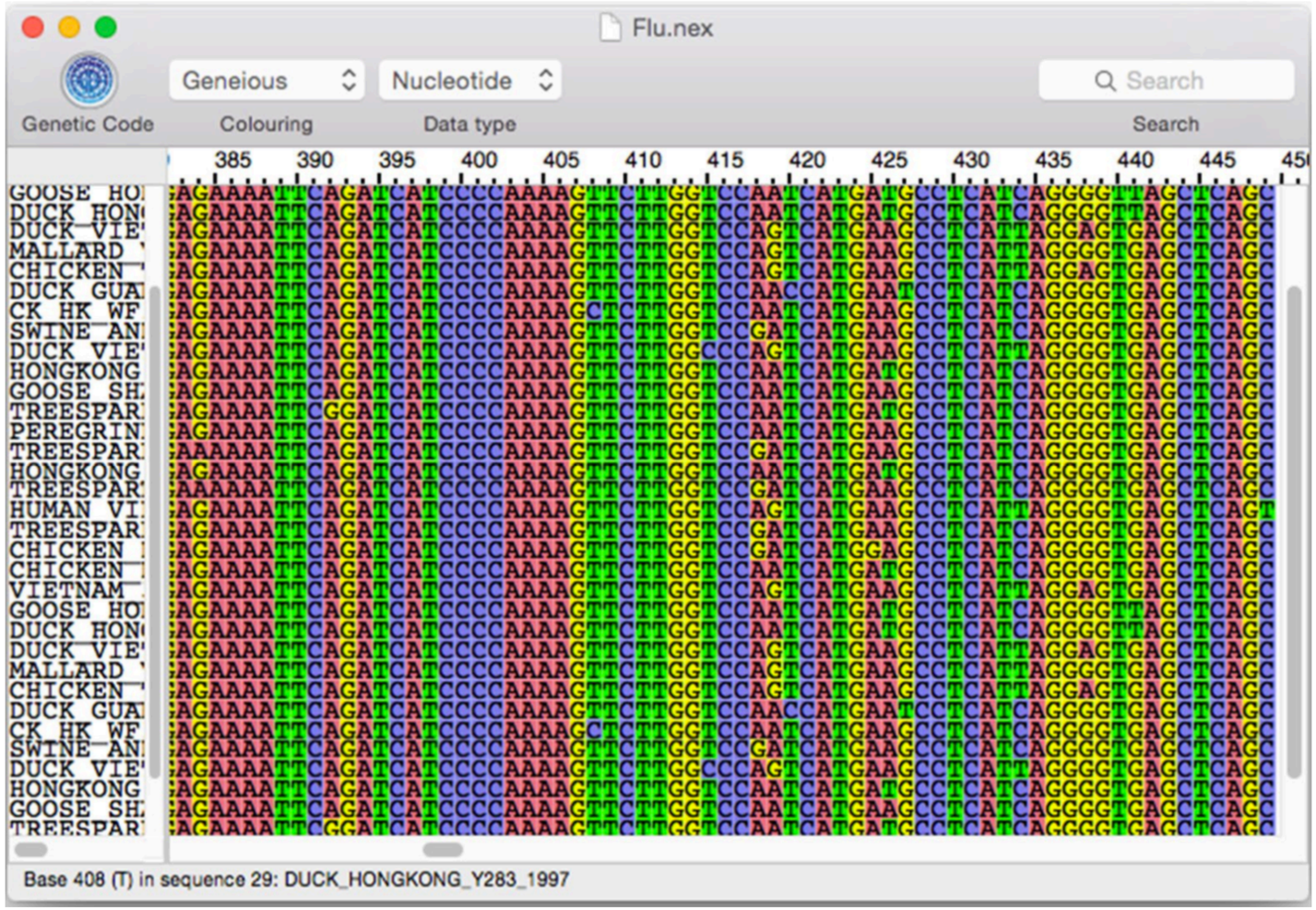
DNA sequencing

Multiple sequence alignment software establishes homology across sites from different species



Duchêne (2021) Phylogenomics Primer

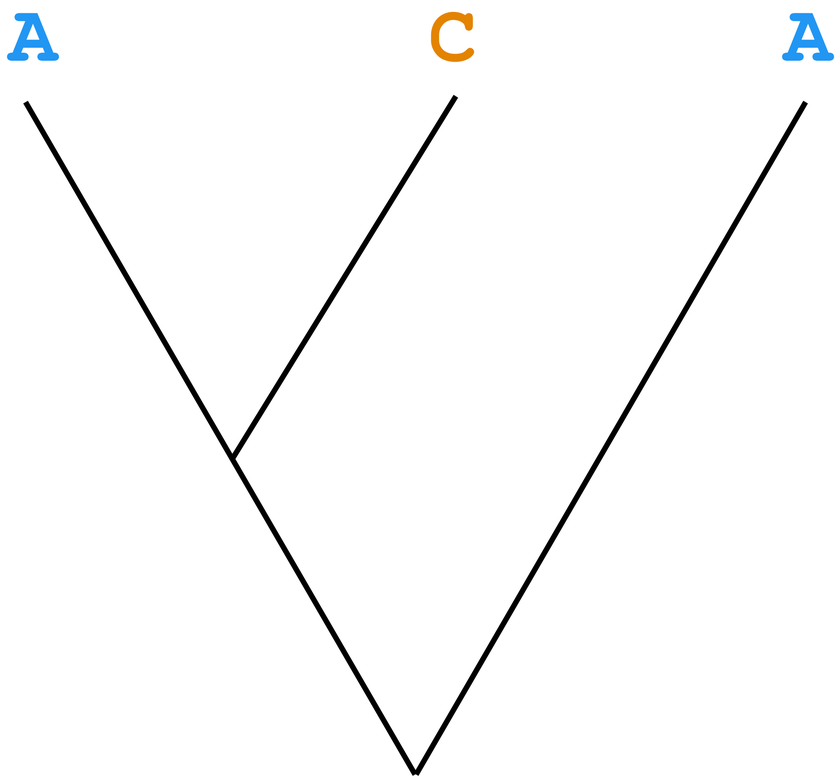
Multiple sequence alignment



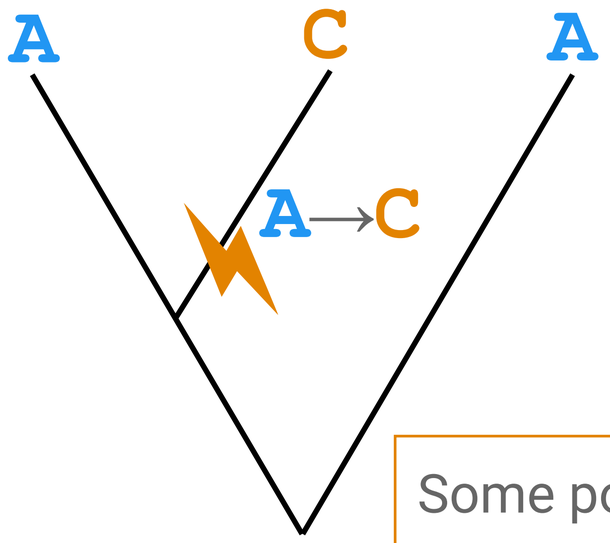
```
#NEXUS

[Cytochrome oxidase B genes - bears]
[Data source: https://revbayes.github.io/tutorials/dating/]


BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=1000;
    FORMAT DATATYPE = DNA MISSING=? GAP=- ;


MATRIX

Ailuropoda_melanoleuca     ATGATCAACATCCGAAAAACTCATCCATTAGTTAAAATTATCAACAACTCATTCATTGACCT...
Arctodus_simus             ATGACCAACATCCGAAAGACTCACCCACTGGCCAAAATTATCAATAACTCATTCATCGACCT...
Helarctos_malayanus        ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATTAACAACTCACTTATTGACCT...
Melursus_ursinus           ATGACCAACATCCGAAAAACCCACCCACTAGCTAAAATCATTAACAACTCACTCATTGACCT...
Ursus_americanus           ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCACTTATTGATCT...
Ursus_arctos               ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCATTTATTGACCT...
Ursus_maritimus            ATGACCAACATCCGAAAAACCCACCCATTAGCTAAAATCATCAACAACTCATTTATTGATCT...
Ursus_thibetanus           ATGACCAACATCCGAAAAACCCATCCATTAGCCAAAATCATCAACAACTCACTCATTGATCT...
Ursus_spelaeus             ATGACCAACATCCGAAAAACCCATCCACTAGCTAAAATCATCAACAACTCATTCATTGACCT...
Tremarctos_ornatus         ATGACCAACATCCGAAAAACTCACCCACTAGCTAAAATCATCAACAACTCATTCATCGACCT...

;

END;
```

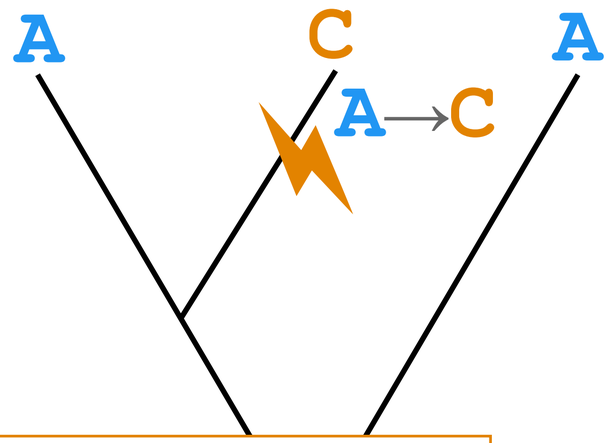
The data are the observed states at the tips

How probable is our data, given my tree?

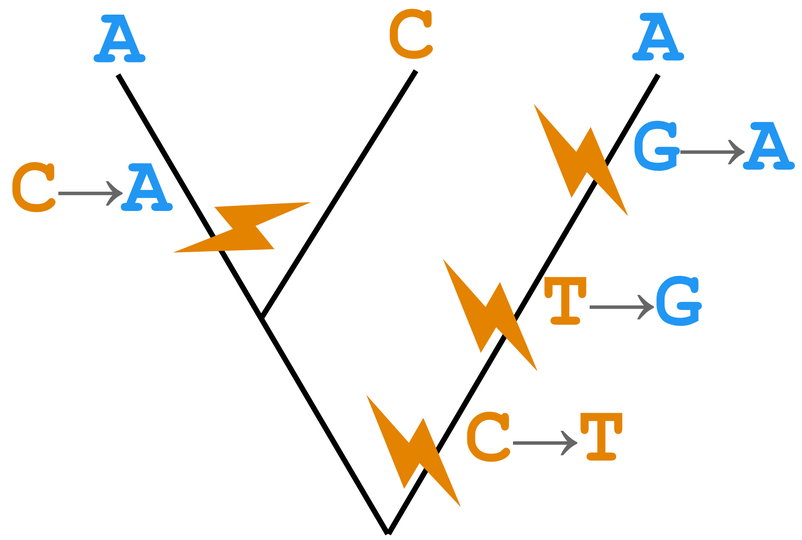


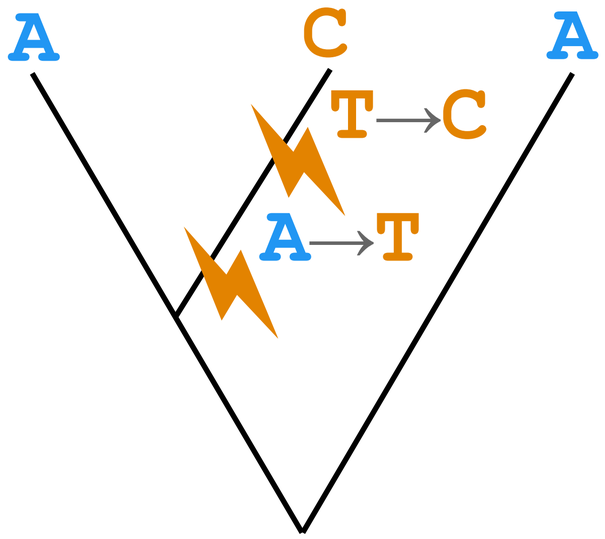
To apply a model based approach we need to be able to compute the probability of our sequence alignment (or character matrix)

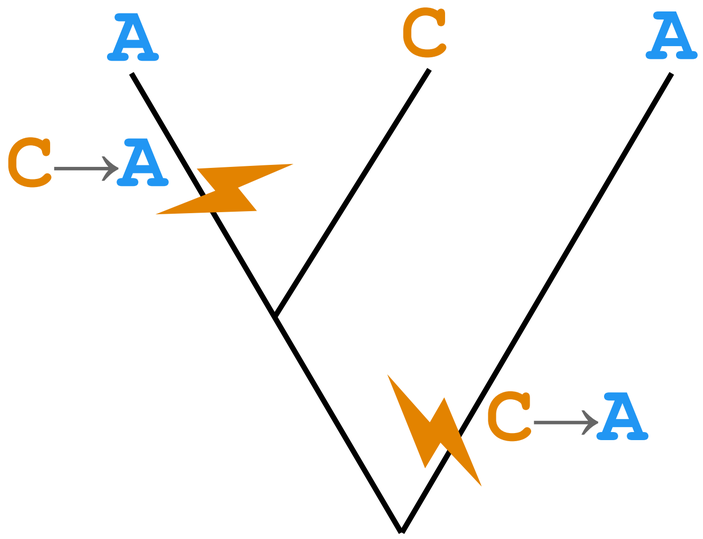


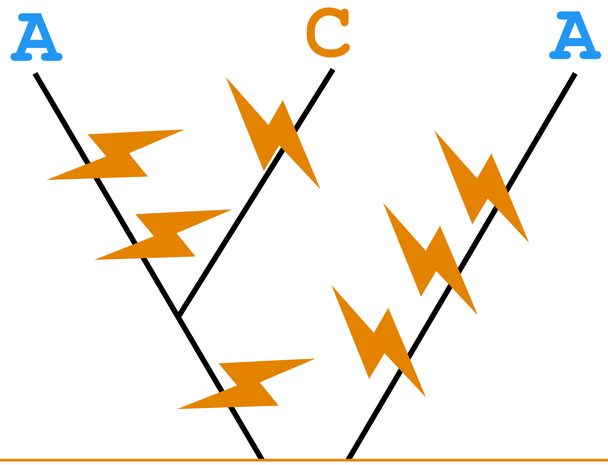


Some possible character histories...





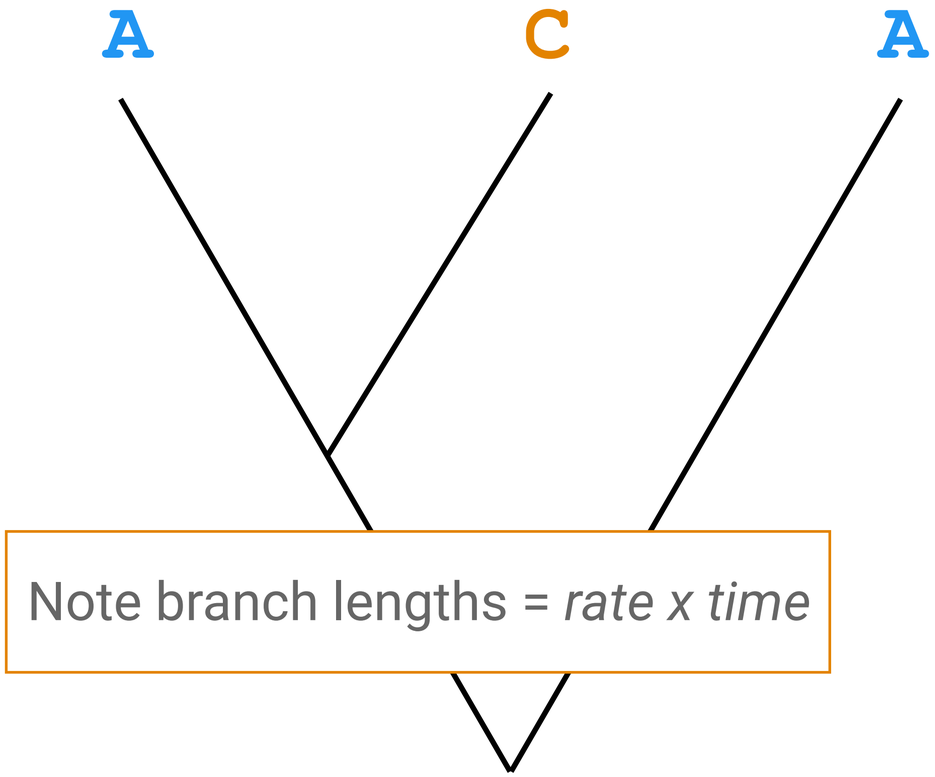




There are **infinitely** many histories leading to the same observed data

The data are the observed states at the tips

How probable is our data, given my tree?



To compute *P*, we need:

- A model of sequence (or character) evolution
- A way of calculating the probability for given a phylogeny (tree topology + branch lengths)

Substitution models

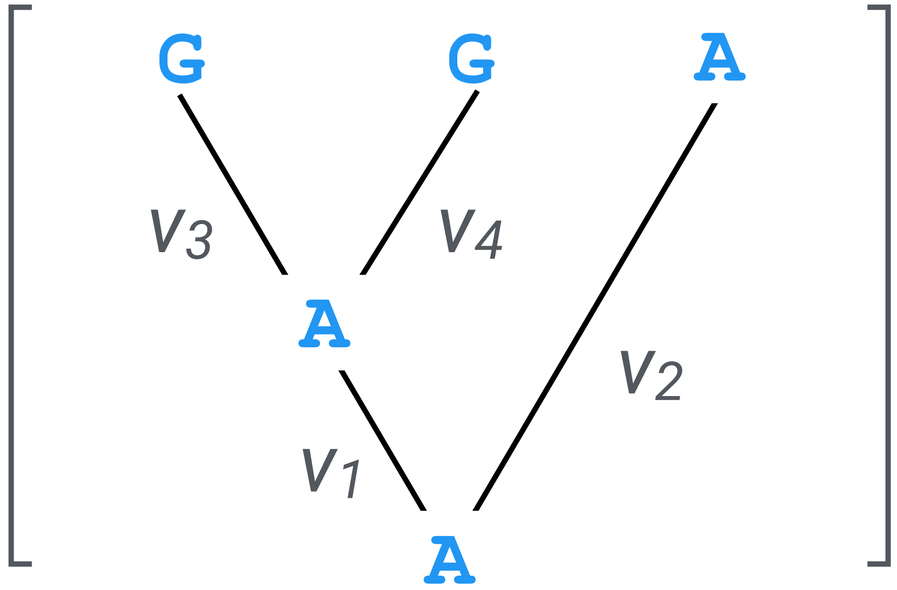
Models of molecular sequence evolution

Also known as substitution / site / character models

They allow us to compute the probability of changing from one state to another over branch length *v*

Computing the probability of the observed data

$P =$



Just suppose for now we know the ancestral states at internal nodes

$P_{AA}(v_1) \times P_{AA}(v_2) \times P_{AG}(v_3) \times P_{AG}(v_1)$

$P_{ij}(v)$ — transition probabilities

Rate matrix

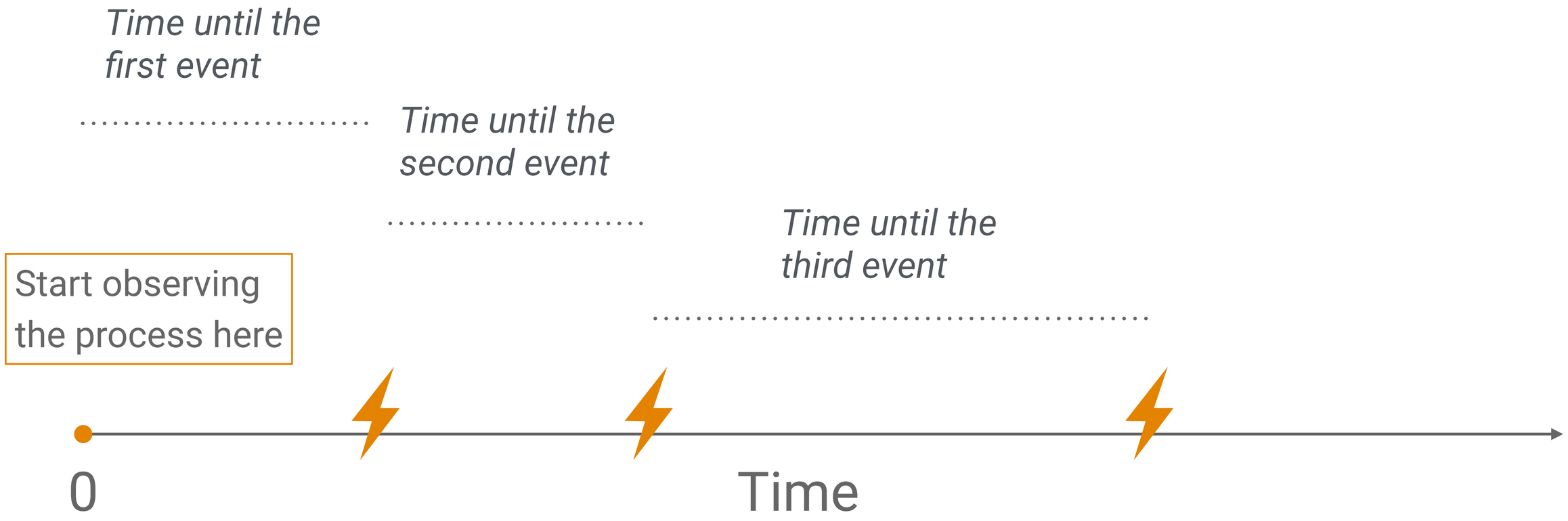
$$Q = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{T} \\ \mathtt{G} \\ \mathtt{C} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{T} & \mathtt{G} & \mathtt{C} \end{array} \\ \begin{bmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{bmatrix} \end{array}$$

In this model, we only have one parameter, substitution rate parameter λ

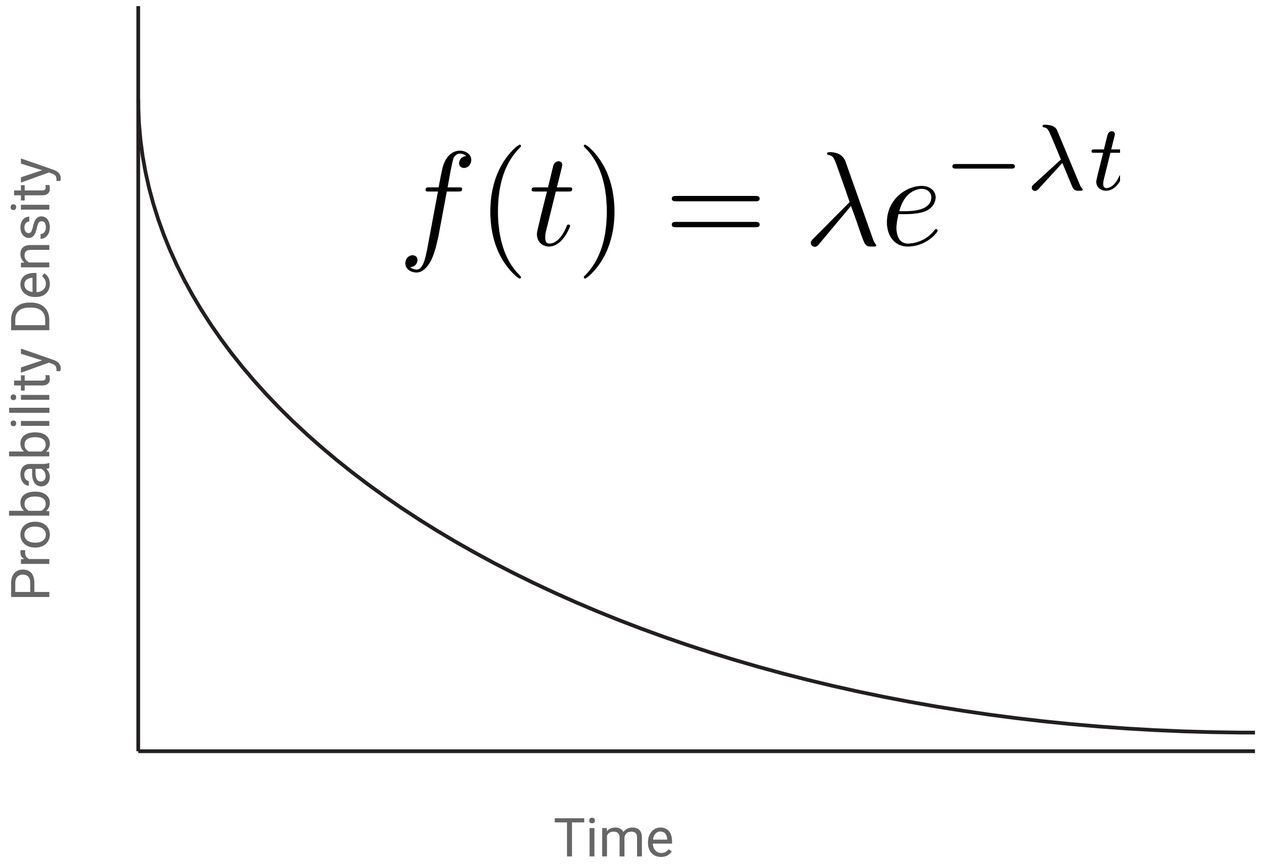
This is the Jukes-Cantor (1969) or JC69 model

Continuous time Markov chain

Nucleotide substitutions (events) occur at a constant rate



The poisson process



The waiting times are exponentially distributed random variables

We can use this to calculate the probability of change over time (or branch length *v*)

The longer the interval of time, the more likely we are to observe change

Exercise

Jukes-Cantor model transition probability applet

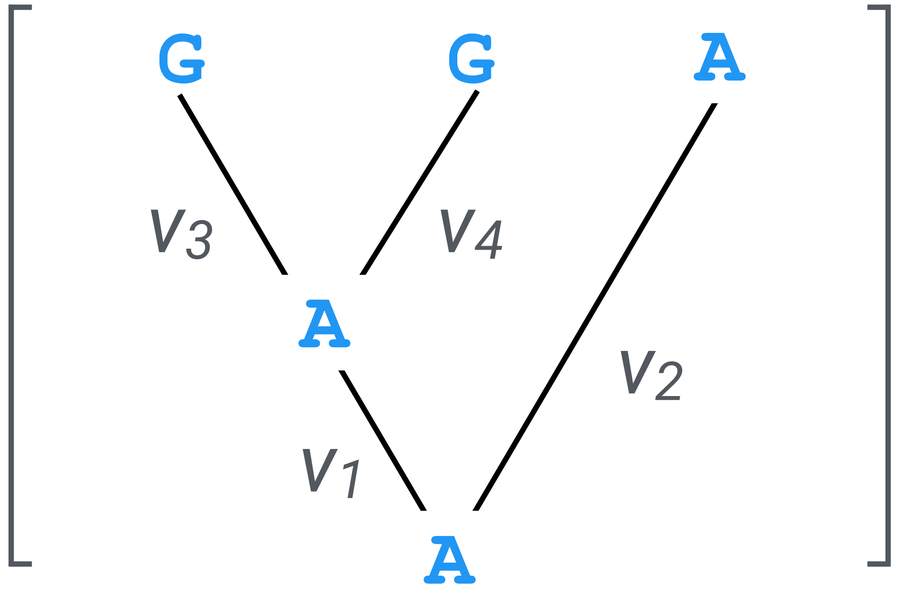
Written by Paul Lewis

Felsenstein’s pruning algorithm

The following slides are adapted from John Huelsenbeck (c/o Sebastian Höhna)

Computing the probability of the observed data

$P =$

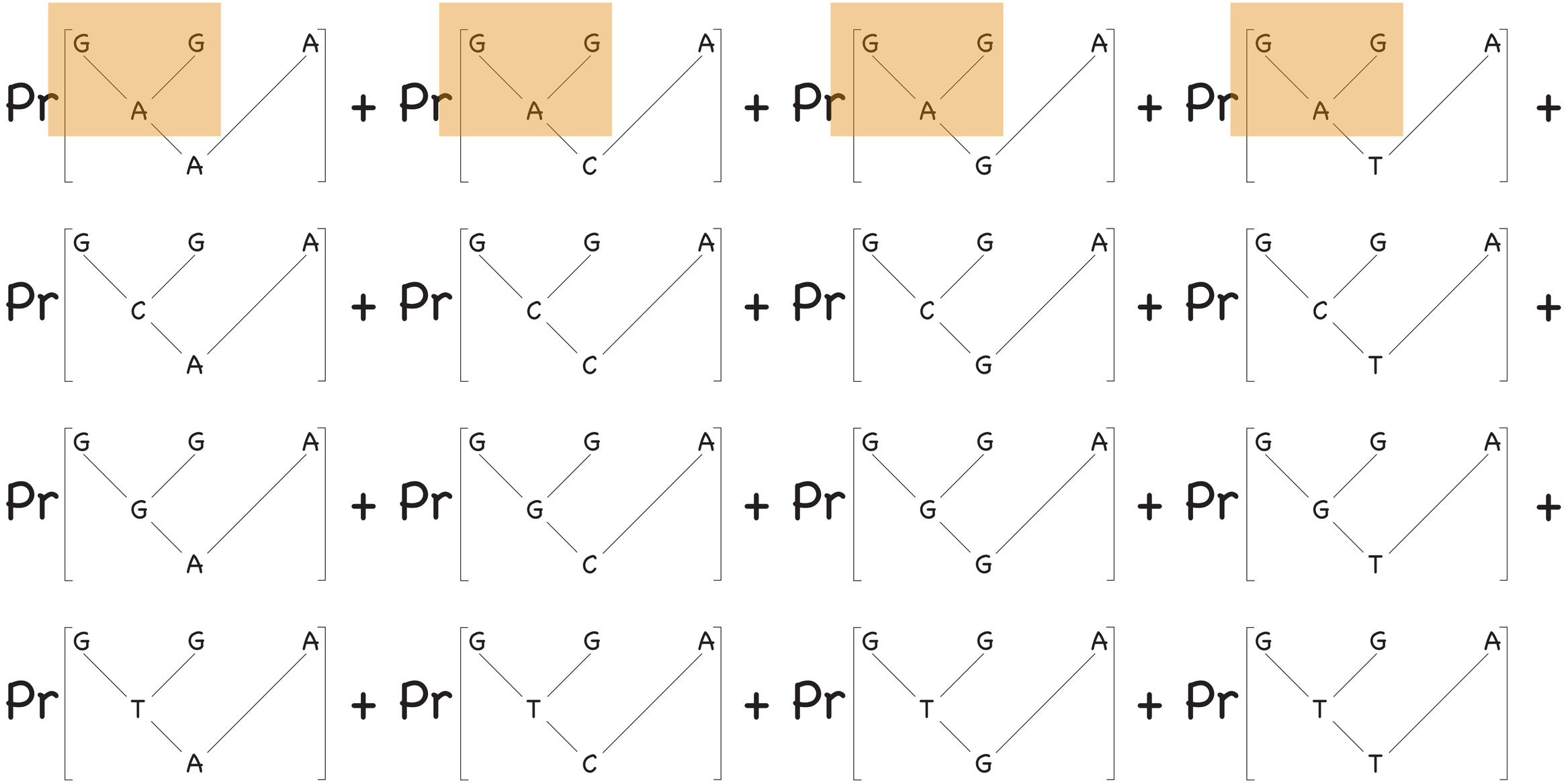


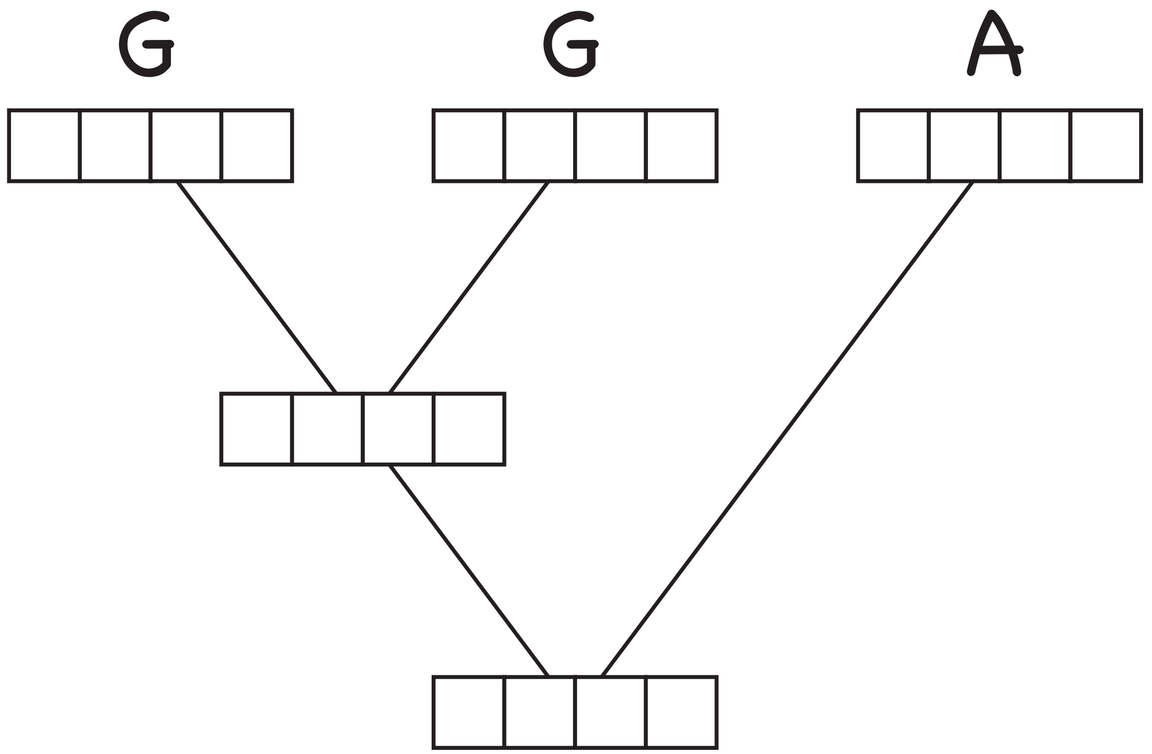
Just suppose for now we know the ancestral states at internal nodes

$\pi_A \times P_{AA}(v_1) \times P_{AA}(v_2) \times P_{AG}(v_3) \times P_{AG}(v_1)$

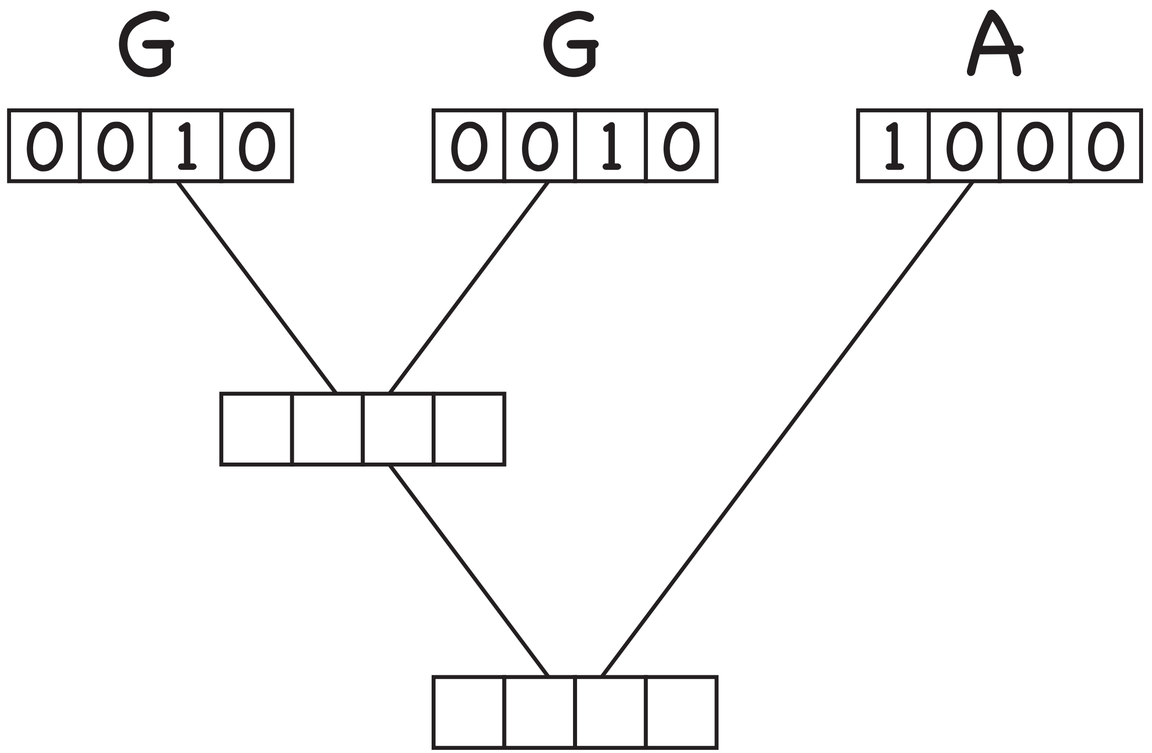
$P_{ij}(v)$ — transition probabilities

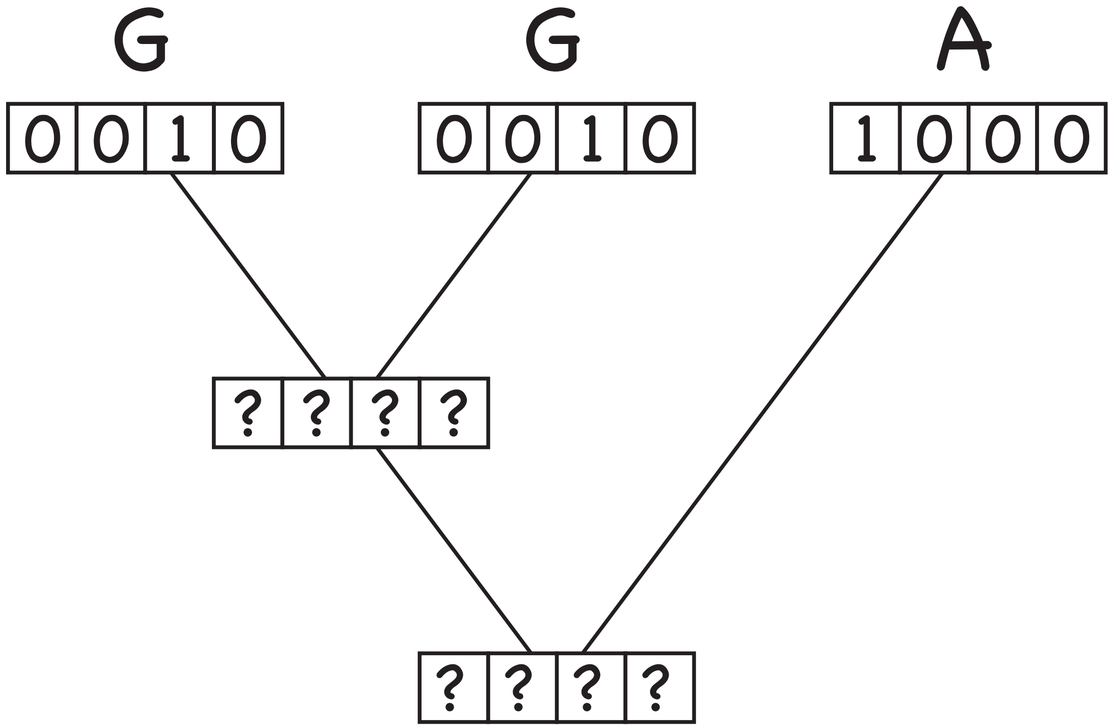
π_i — stationary frequencies

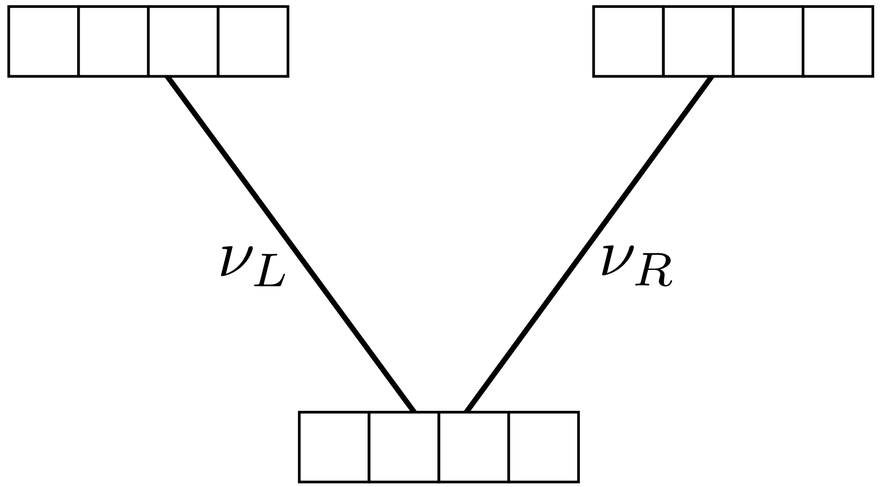

Pr [G G A / A / A] + Pr [G G A / A / C] + Pr [G G A / A / G] + Pr [G G A / A / T] +
Pr [G G A / C / A] + Pr [G G A / C / C] + Pr [G G A / C / G] + Pr [G G A / C / T] +
Pr [G G A / G / A] + Pr [G G A / G / C] + Pr [G G A / G / G] + Pr [G G A / G / T] +
Pr [G G A / T / A] + Pr [G G A / T / C] + Pr [G G A / T / G] + Pr [G G A / T / T]



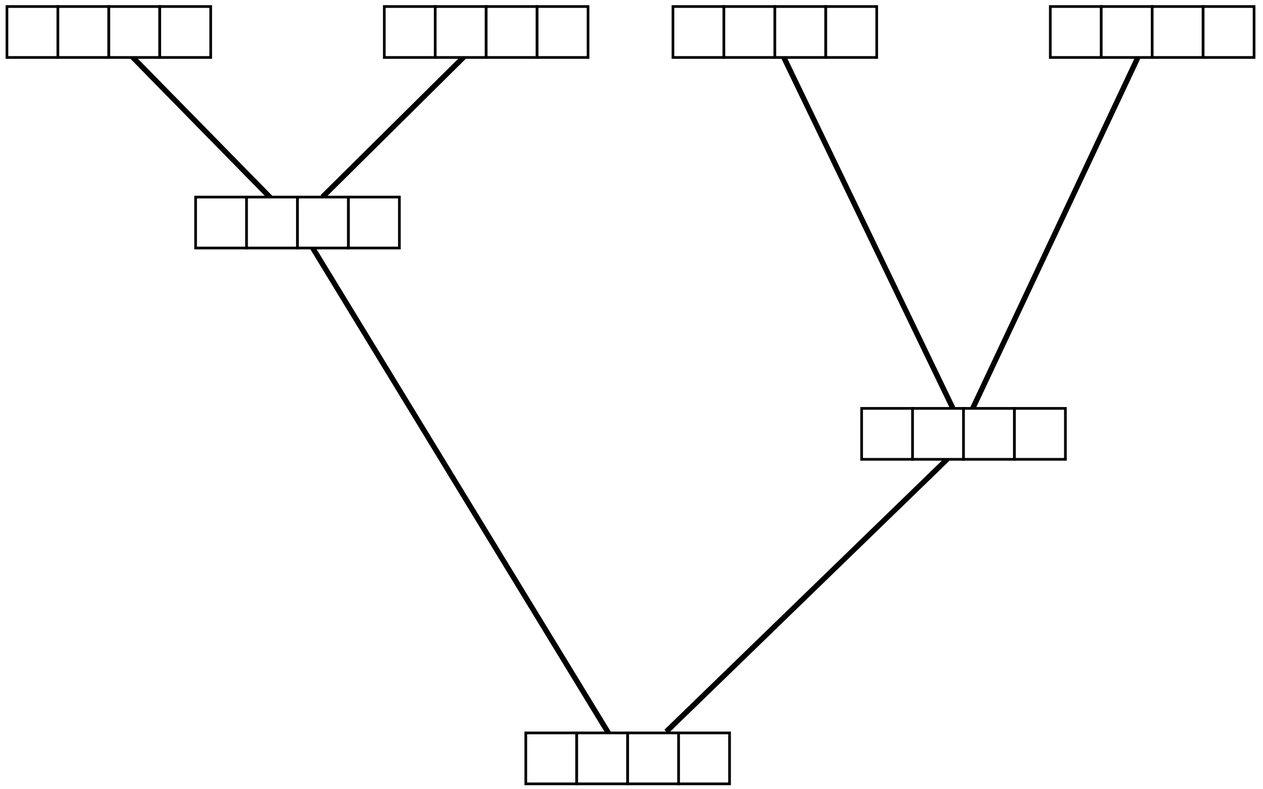
Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach.


G
G
A
0 0 1 0
0 0 1 0
1 0 0 0

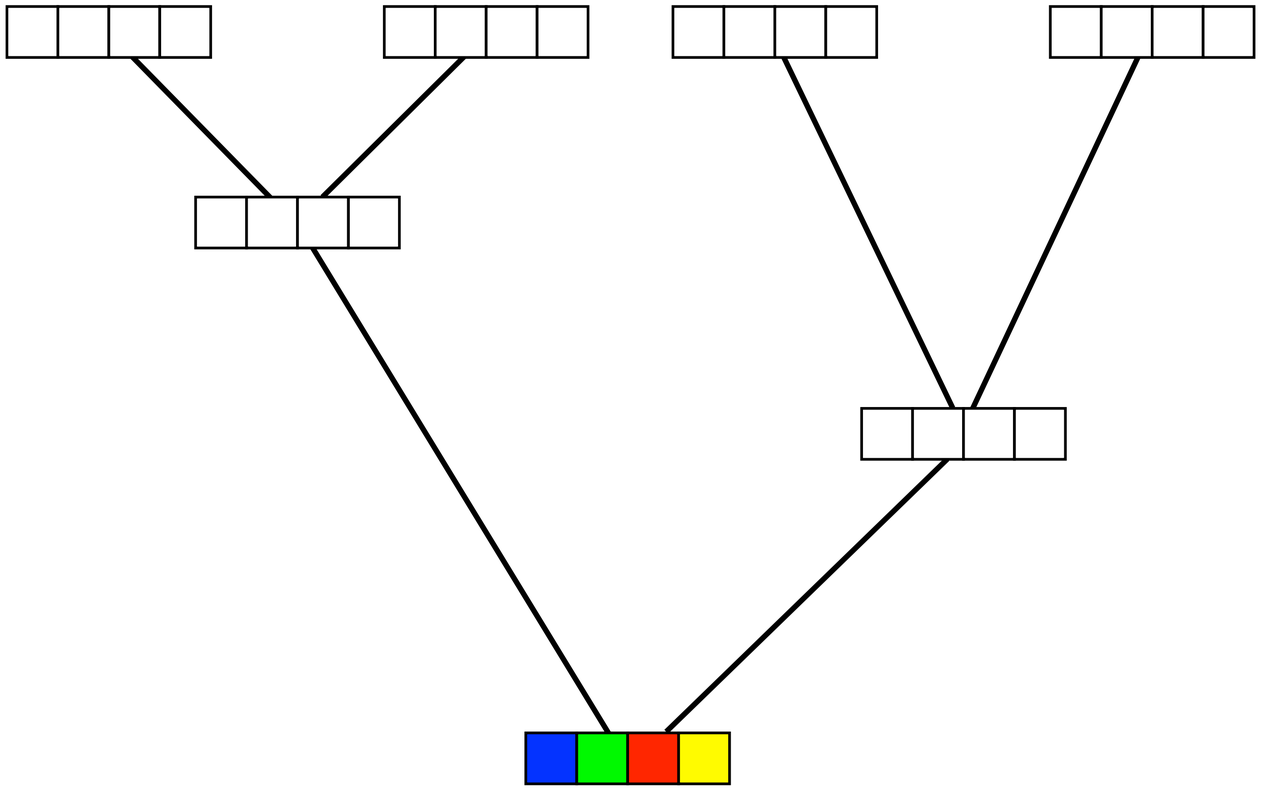

G
G
A
0 0 1 0
0 0 1 0
1 0 0 0
? ? ? ?
? ? ? ?



$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$



$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$



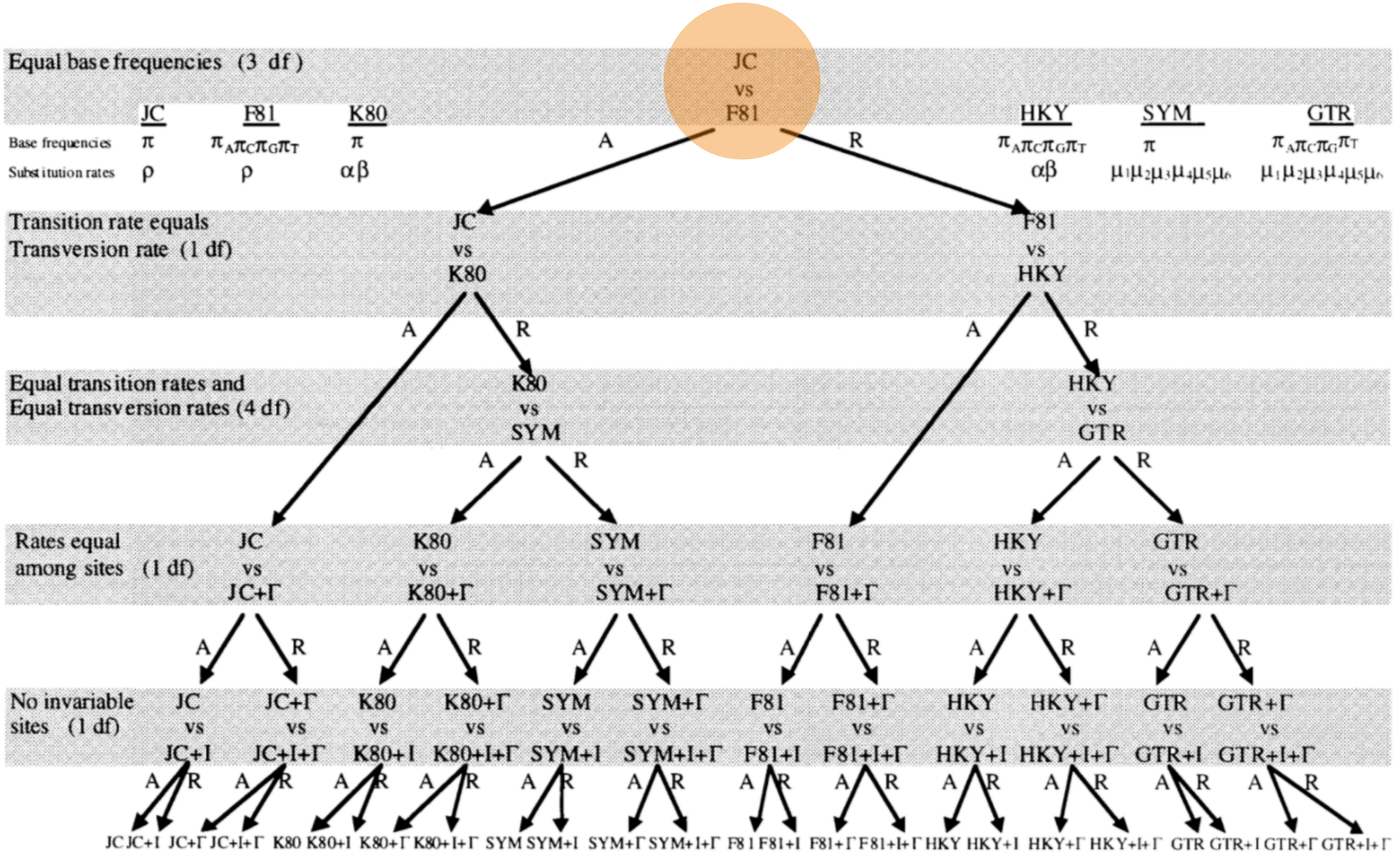
$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$

$$\ell_i = \left(\sum_j p_{ij}(\nu_L)\, \ell_j^L \right) \times \left(\sum_j p_{ij}(\nu_R)\, \ell_j^R \right)$$

$$\ell_{\text{Site}} = \pi_A \times \ell_A^{\text{Root}} + \pi_C \times \ell_C^{\text{Root}} + \pi_G \times \ell_G^{\text{Root}} + \pi_T \times \ell_T^{\text{Root}}$$

Another nice description of the pruning algorithm: *Harmon* (2019) Phylogenetic Comparative Methods, Chapter 8

Other substitution models



Posada & Crandall (1998)

Base frequencies

The JC69 model assumes equal transition rates and equal base frequencies

Base frequencies are the proportion of each nucleotide in the dataset

If a given nucleotide appears in our dataset at a low frequency, we are less likely to observe a transition to that state

The general time reversible model

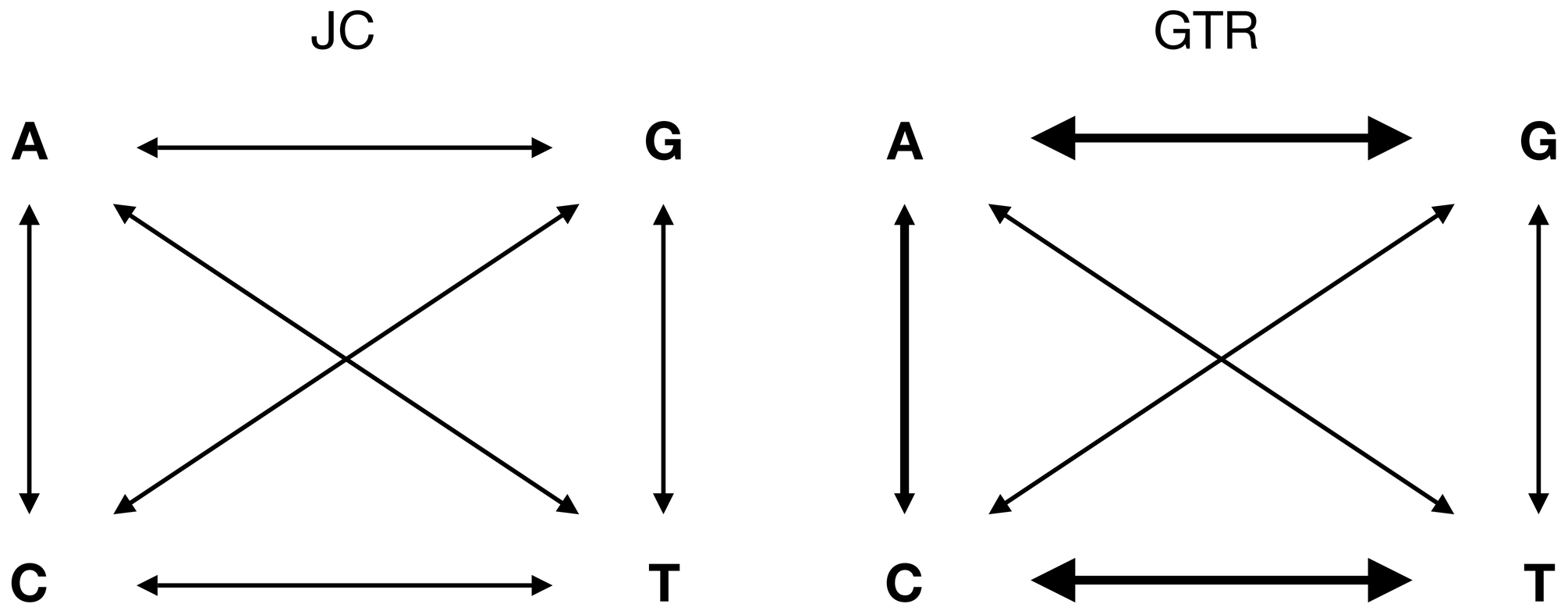
$$Q = \begin{pmatrix} * & \mu_{AG}\pi_G & \mu_{AC}\pi_C & \mu_{AT}\pi_T \\ \mu_{GA}\pi_A & * & \mu_{GC}\pi_C & \mu_{GT}\pi_T \\ \mu_{CA}\pi_A & \mu_{CG}\pi_G & * & \mu_{CT}\pi_T \\ \mu_{TA}\pi_A & \mu_{TG}\pi_G & \mu_{TC}\pi_C & * \end{pmatrix}$$

Allows for unequal transition rates (μ) *and* unequal base frequencies (π)

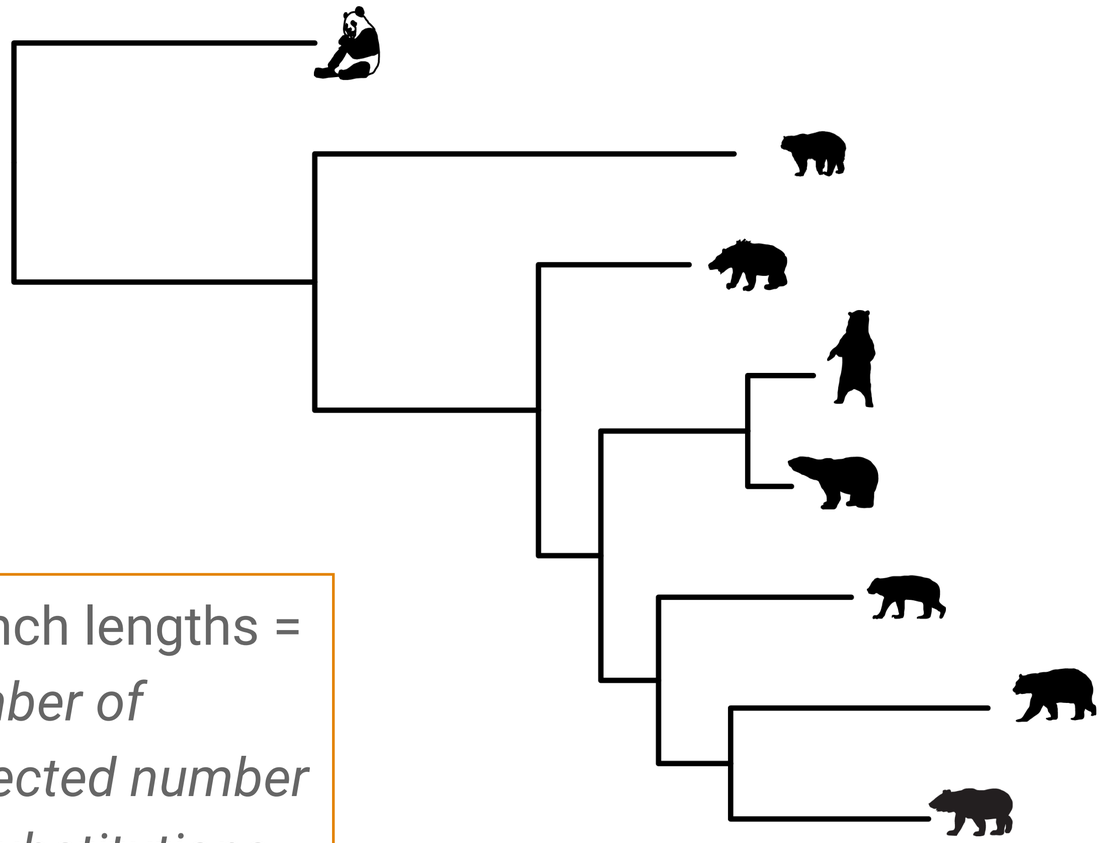
Note the rates are symmetric – e.g., the rate of change between A and T, is the same in both directions – but the frequency of each character state also affects the probability of change

Tavaré (1986)

The JC versus GTR models



Branch lengths



Branch lengths = *number of expected number of substitutions per site*

Branch lengths are a product of rate and time

Without temporal information we can only measure relative genetic distance

Maximum likelihood

How do we find the ‘best’ tree?

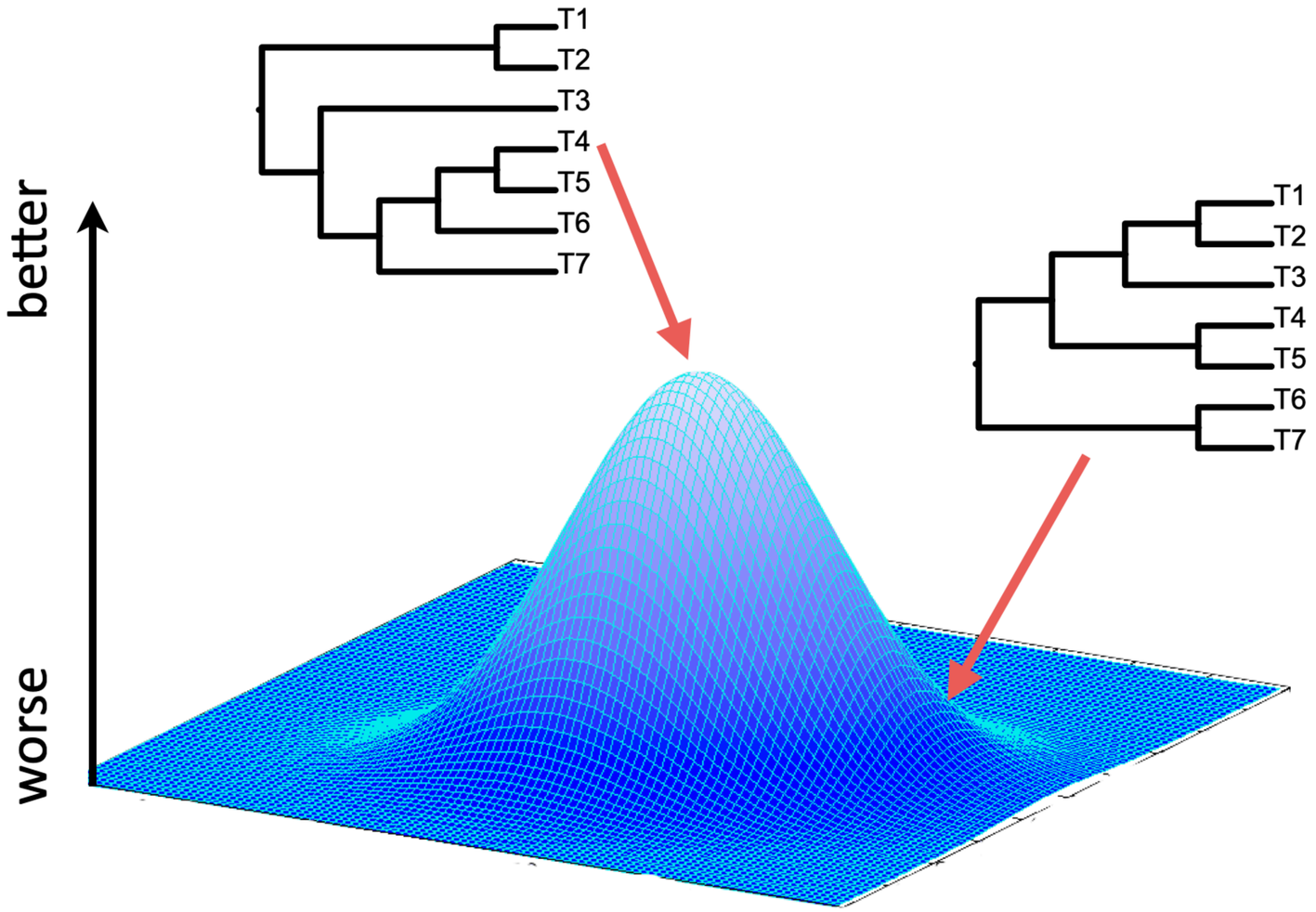


Image source Tracy Heath

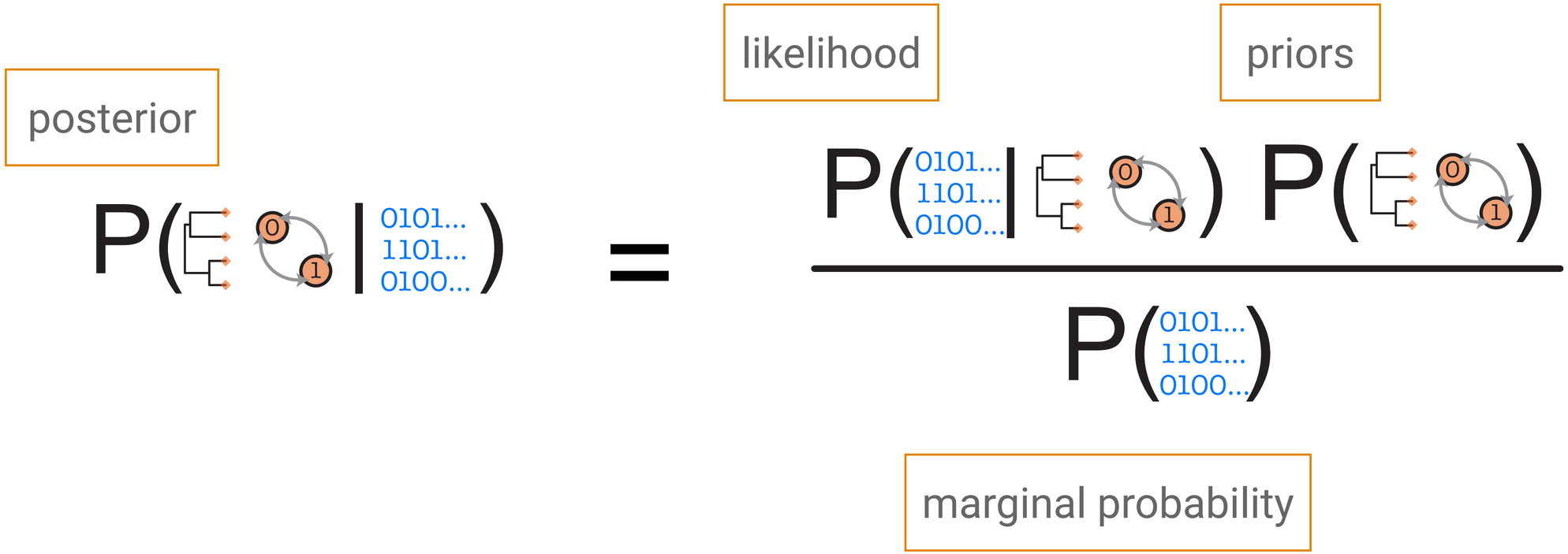
It depends how you measure ‘best’

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

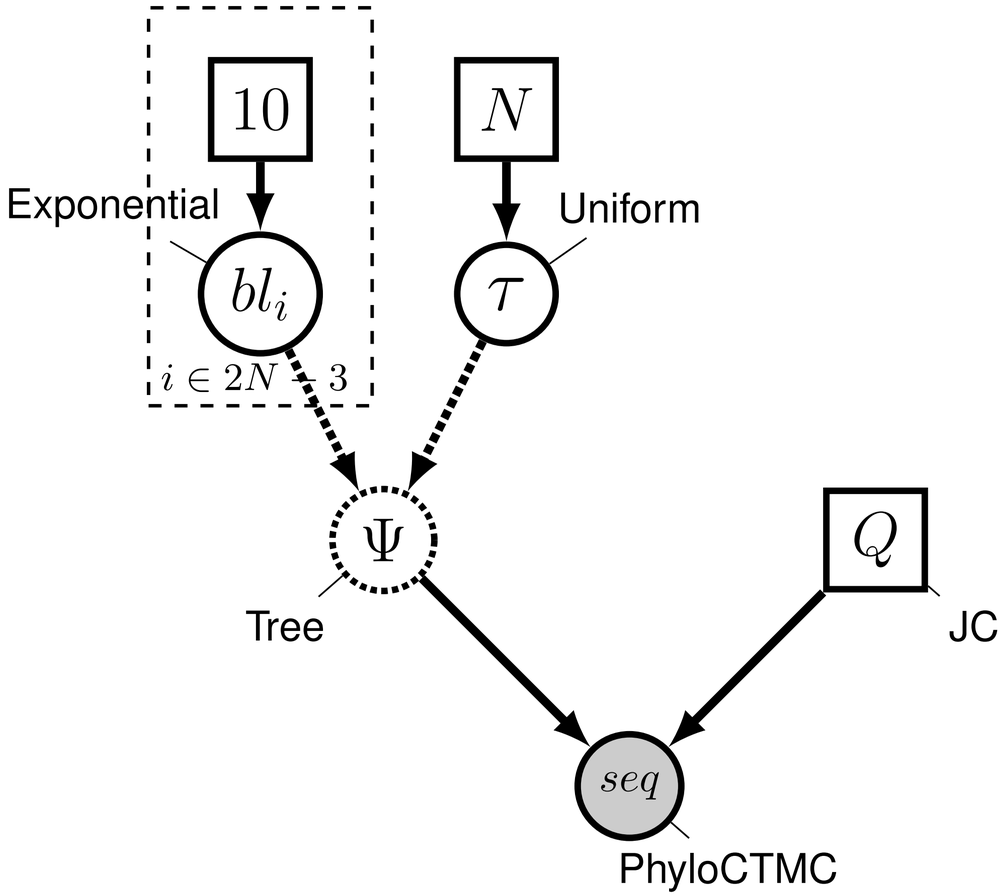
Both maximum likelihood and Bayesian inference are model-based approaches

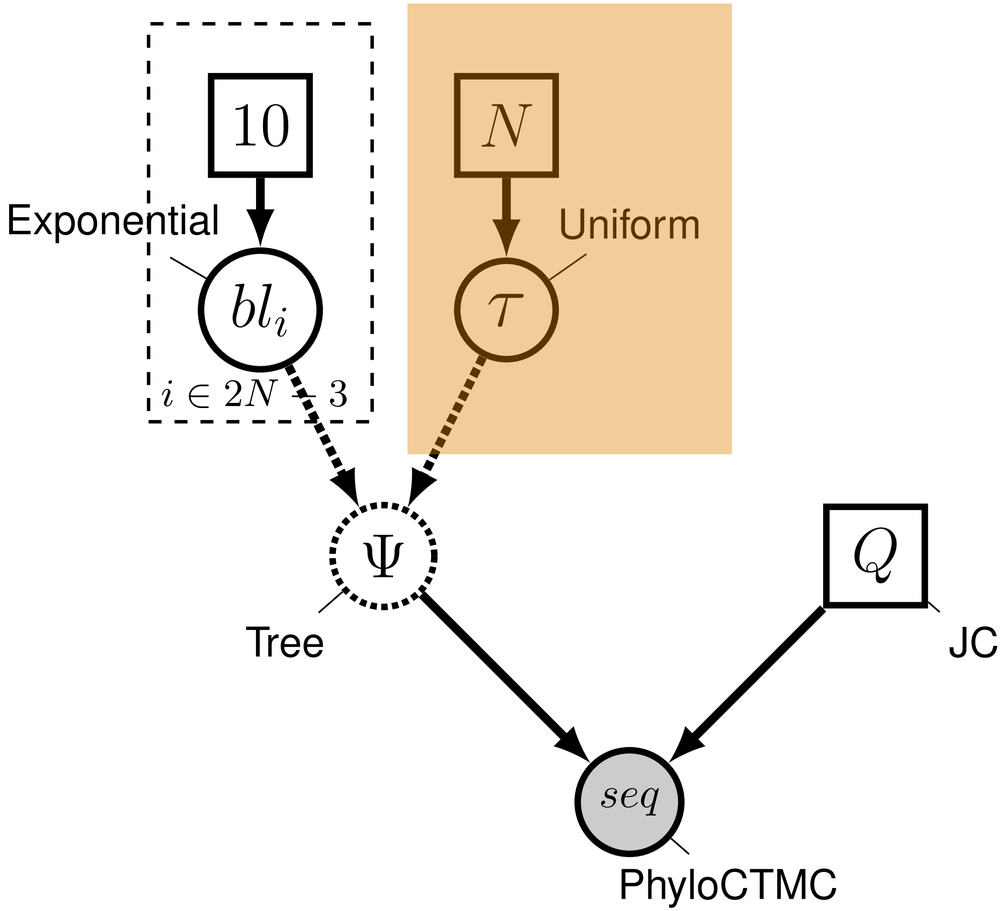
Note these are not the only approaches to tree-building but they are the most widely used

Bayesian tree inference

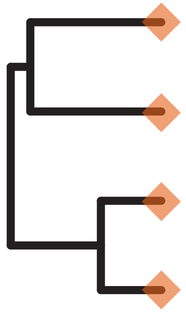


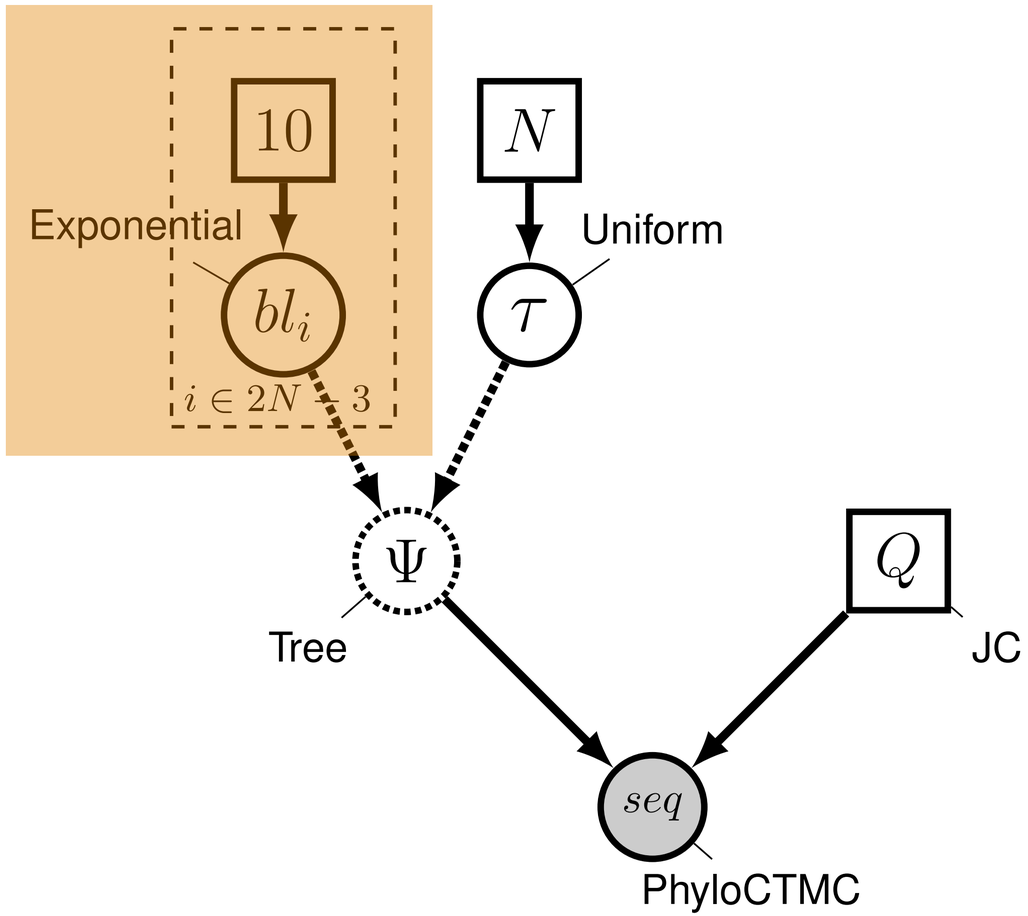
$


10
N
Exponential
Uniform
bl_i
τ
$i \in 2N - 3$
Ψ
Q
Tree
JC
seq
PhyloCTMC

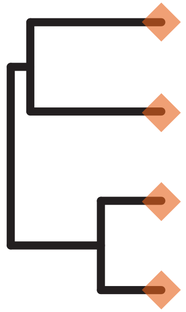


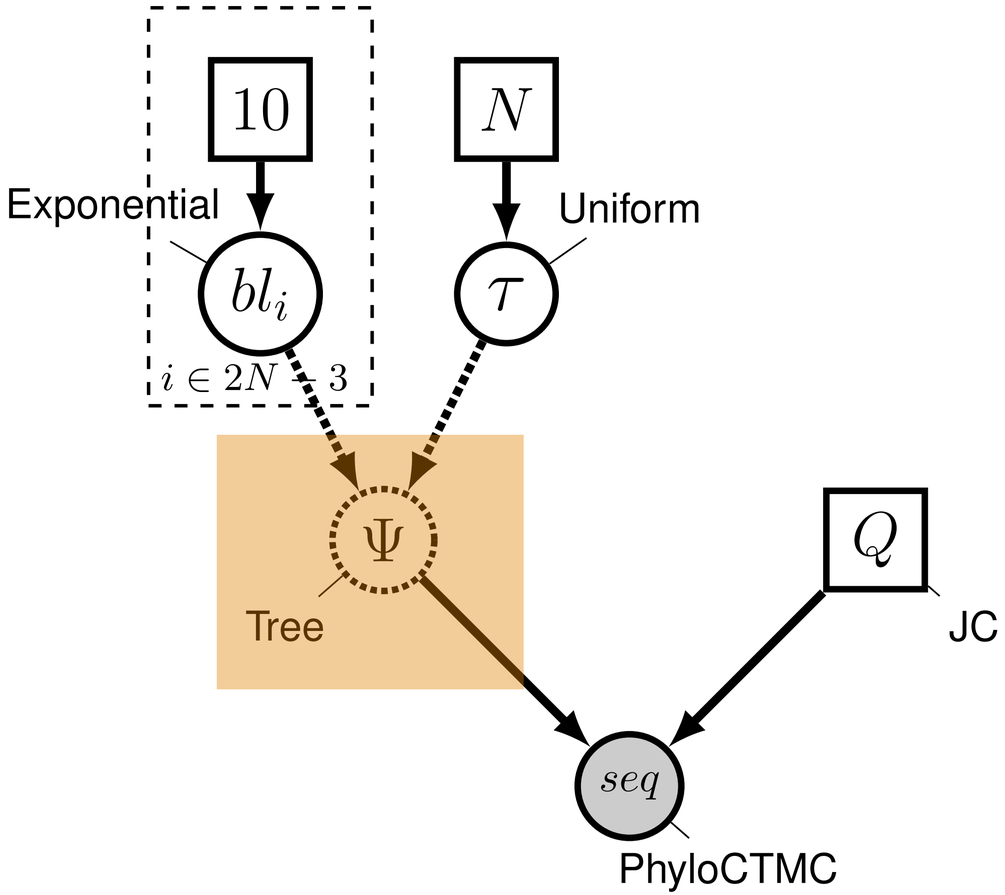
prior on the tree topology



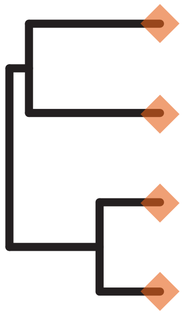


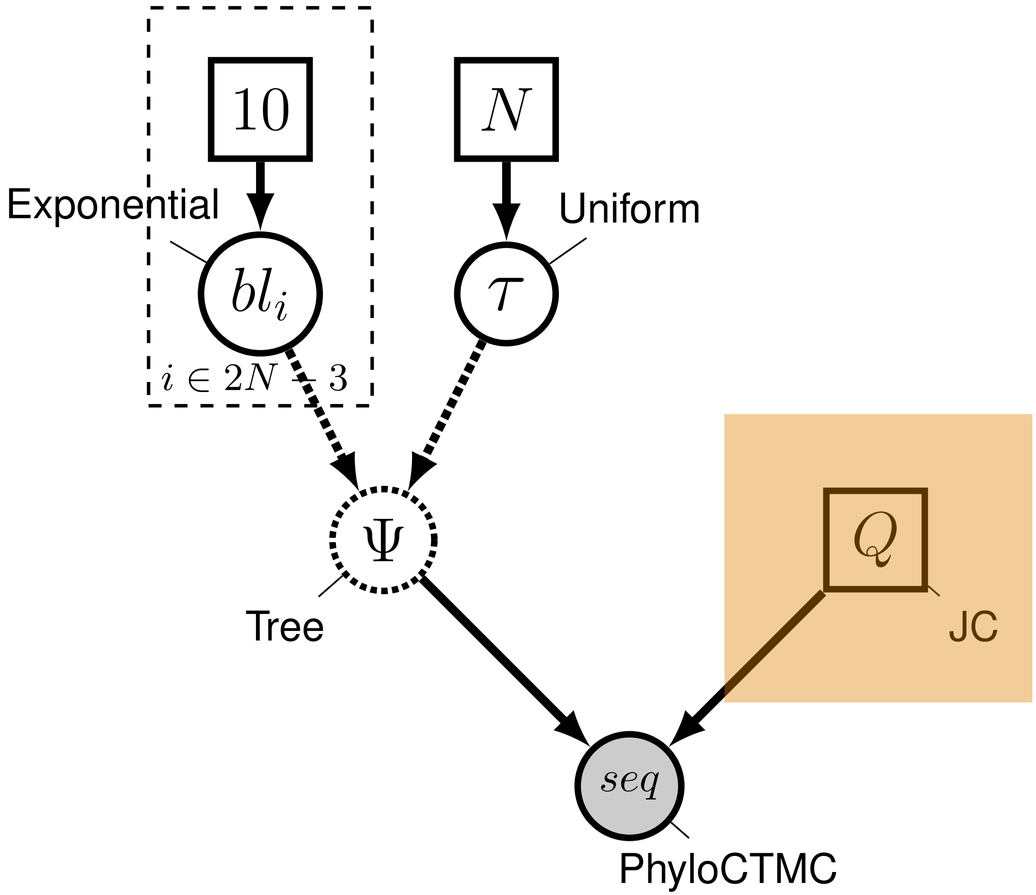
prior on the branch lengths



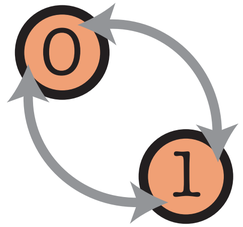


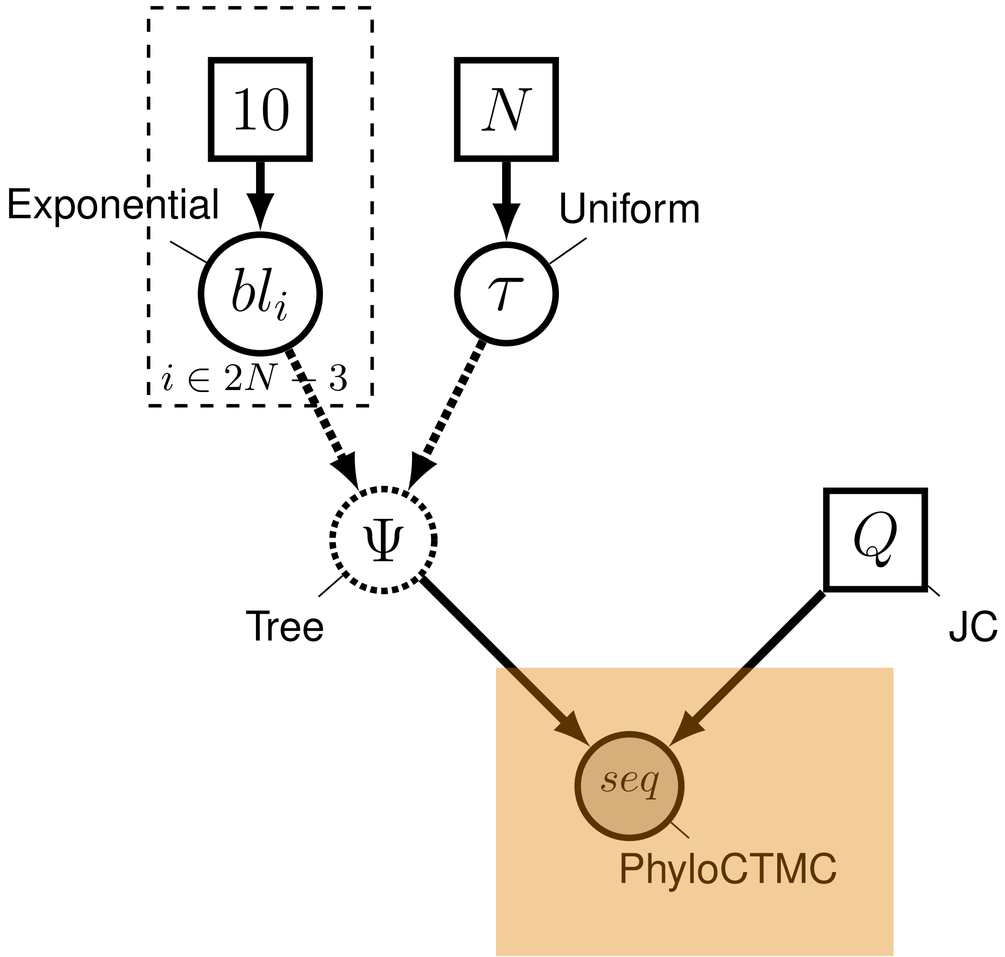
we can combine the topology and branch lengths





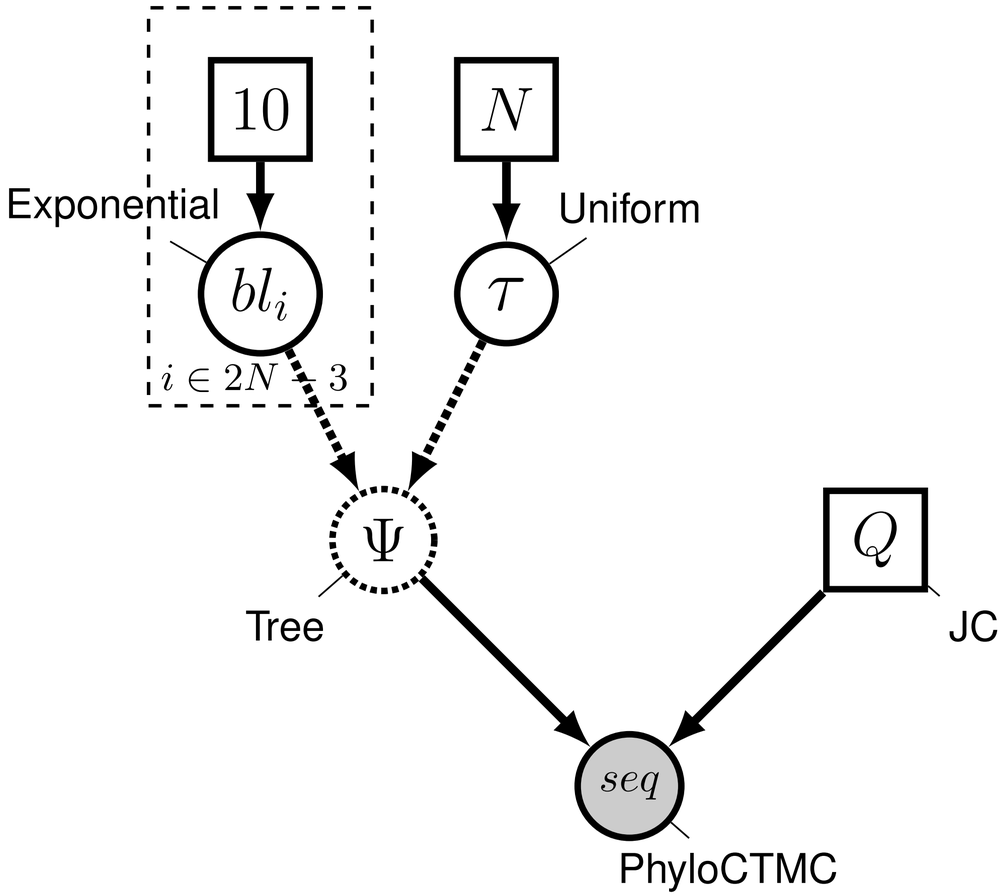
Substitution model



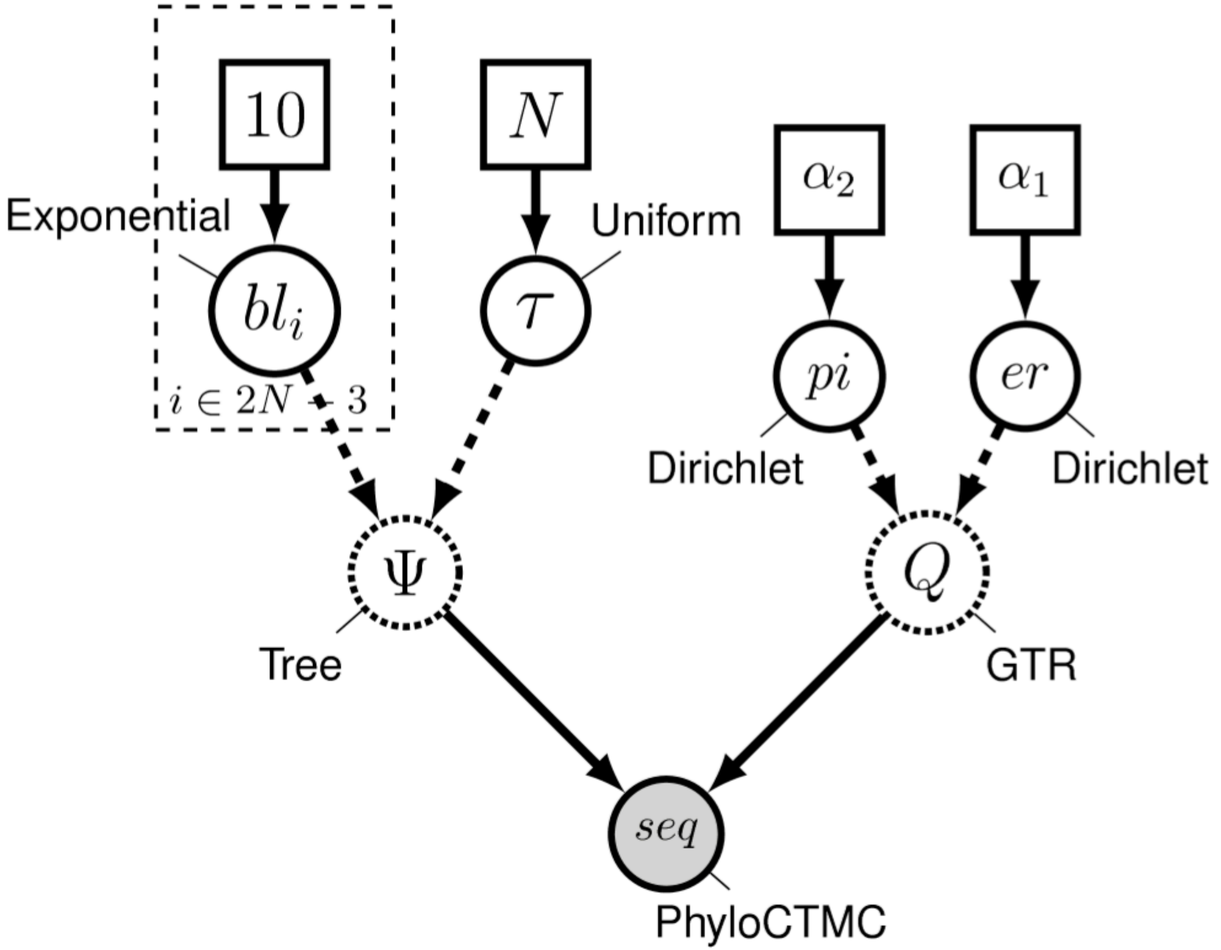


Observed data

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




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


Q <- fnJC(4)


seq ~ dnPhyloCTMC( tree=psi, Q=Q, type="DNA" )

seq.clamp( data )
```




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


alpha1 <- v(1,1,1,1,1,1)
alpha2 <- v(1,1,1,1)
er ~ dnDirichlet( alpha1 )
pi ~ dnDirichlet( alpha2 )
Q := fnGTR(er, pi)
```

Introduction to MCMC

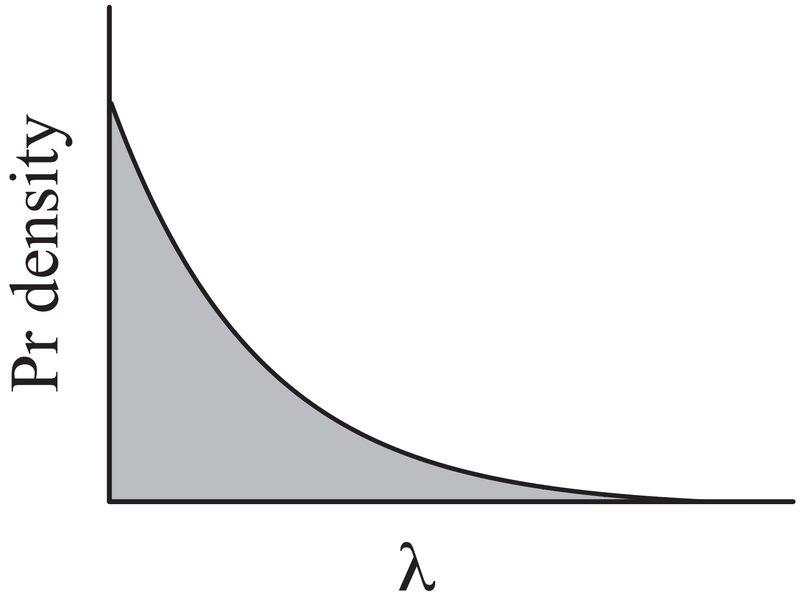
Probabilities vs probability densities

In phylogenetics, probabilities are not normally discrete (i.e., represented by a single value)

We’re often dealing with a lot of uncertainty and typically work with probability densities

Probability densities introduce some complexity

Probabilities vs probability densities



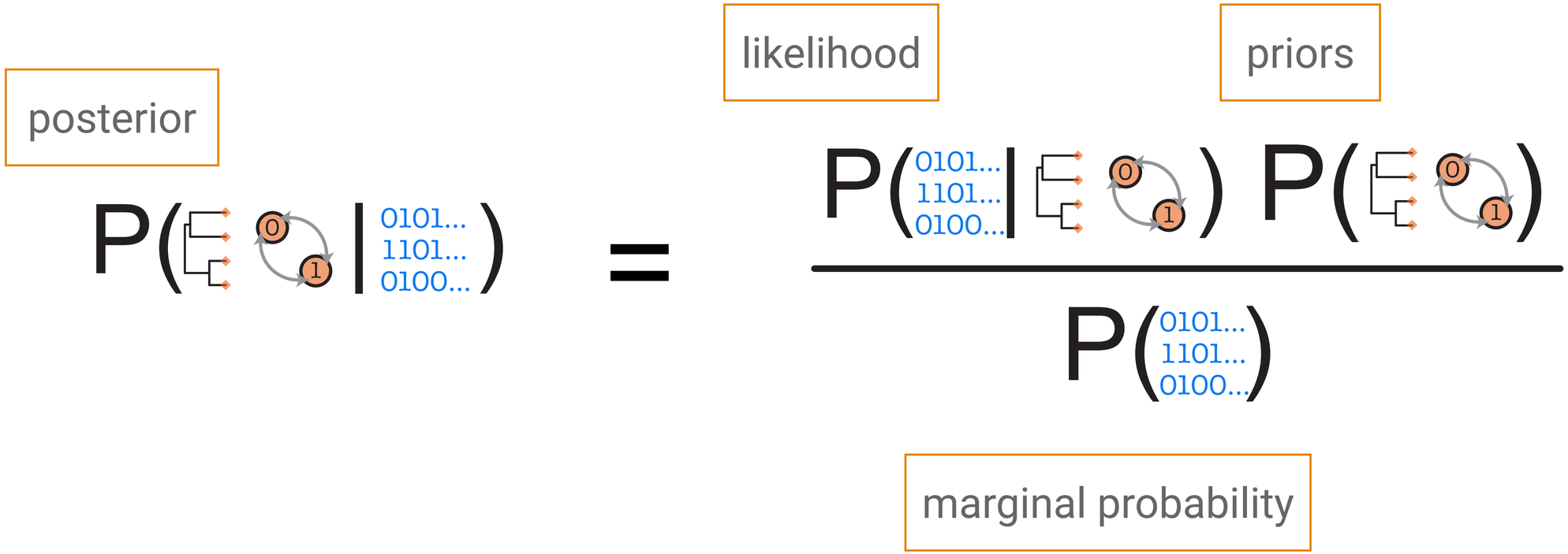
λ is drawn from an exponential distribution with mean δ

The x-axis represents the value of our parameter λ

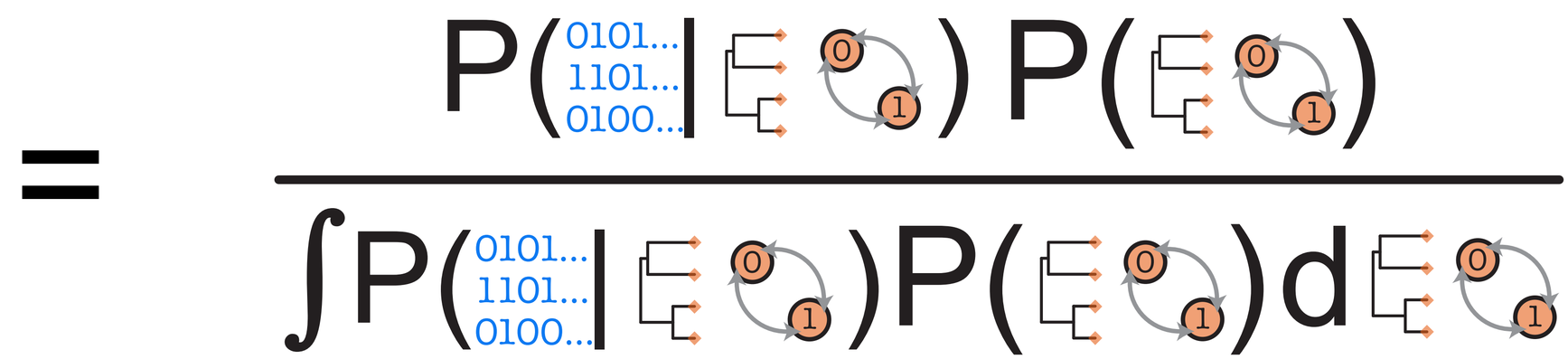
The y-axis does have a value but it is not so easily interpretable

The distribution height reflects the relative probability of a given range of values

Bayesian tree inference



Bayesian tree inference



this part is incredibly difficult to calculate!

What is Markov chain Monte Carlo (MCMC)?

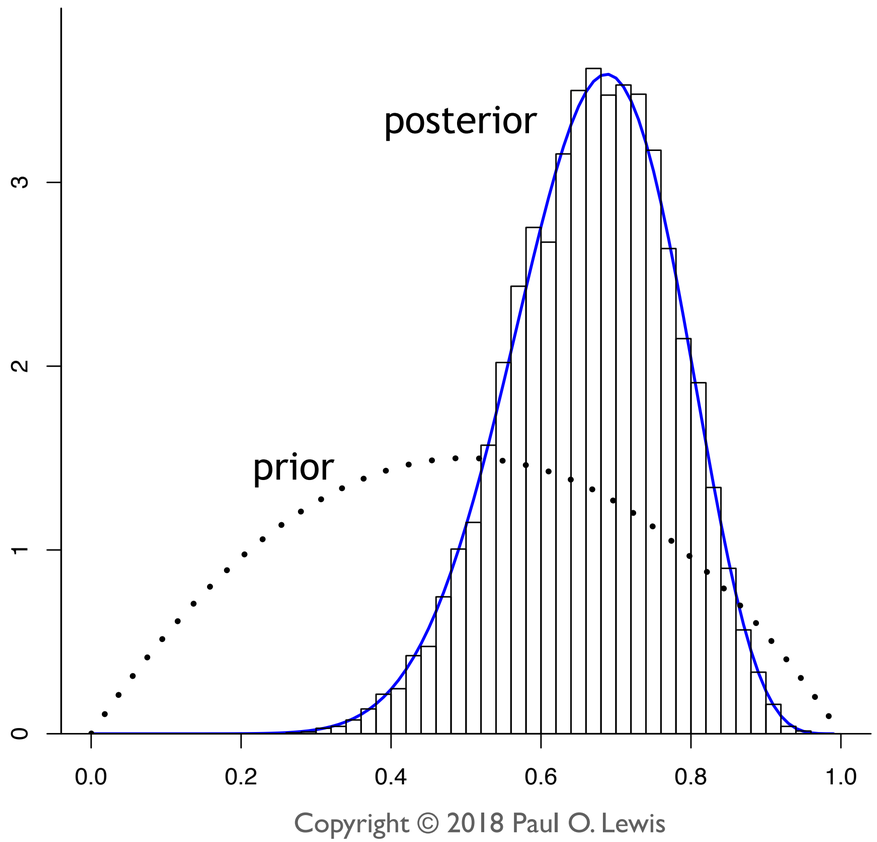
A group of algorithms for approximating the posterior distribution (also known as samplers)

Markov chain means the progress of the algorithm doesn't depend on its past

Monte Carlo (named for the casino in Monaco) methods estimate a distribution via random sampling

We use this algorithm to visit different regions the parameter space. The number of times a given region is visited will be in proportion to its posterior probability

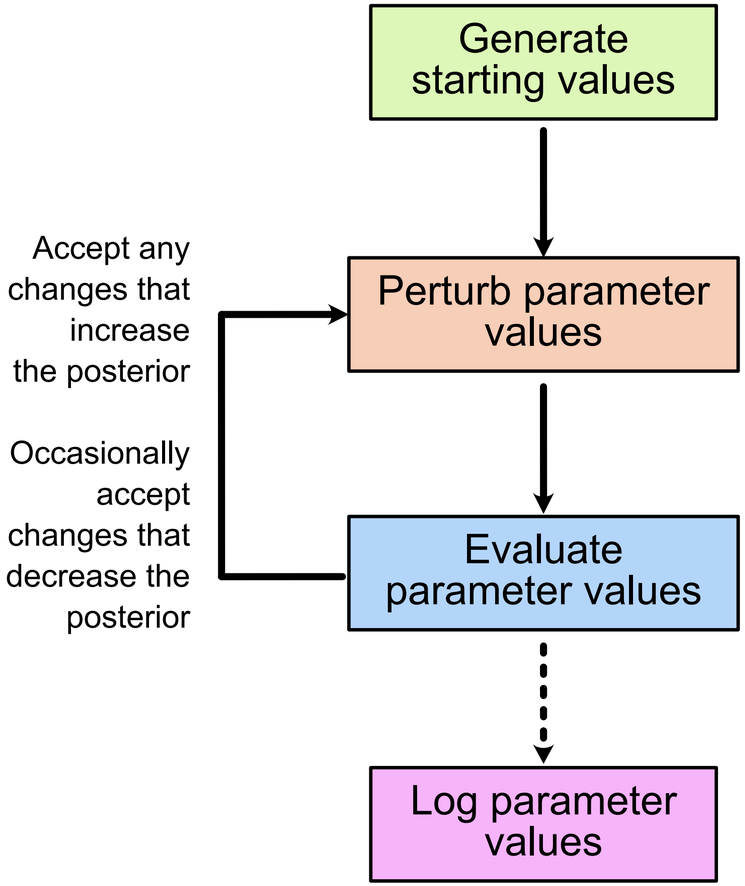
What is Markov chain Monte Carlo (MCMC)?



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

The Metropolis-Hastings algorithm

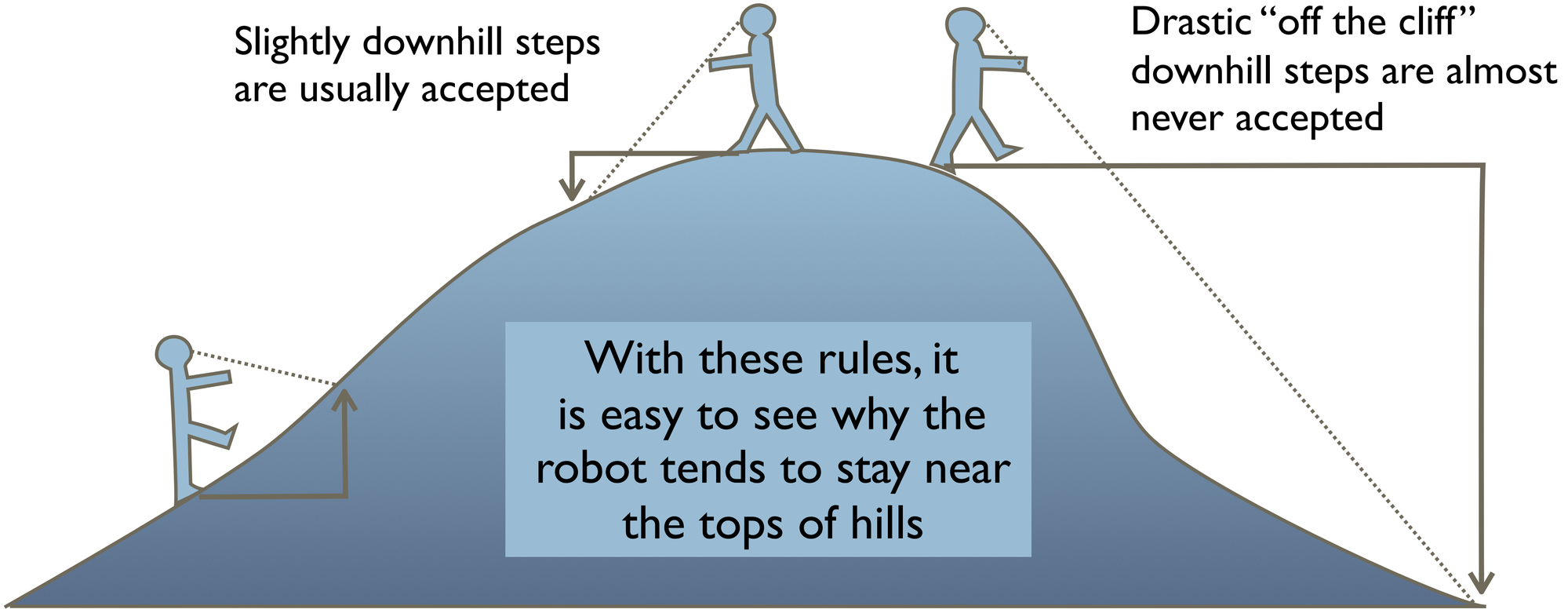
Flowchart

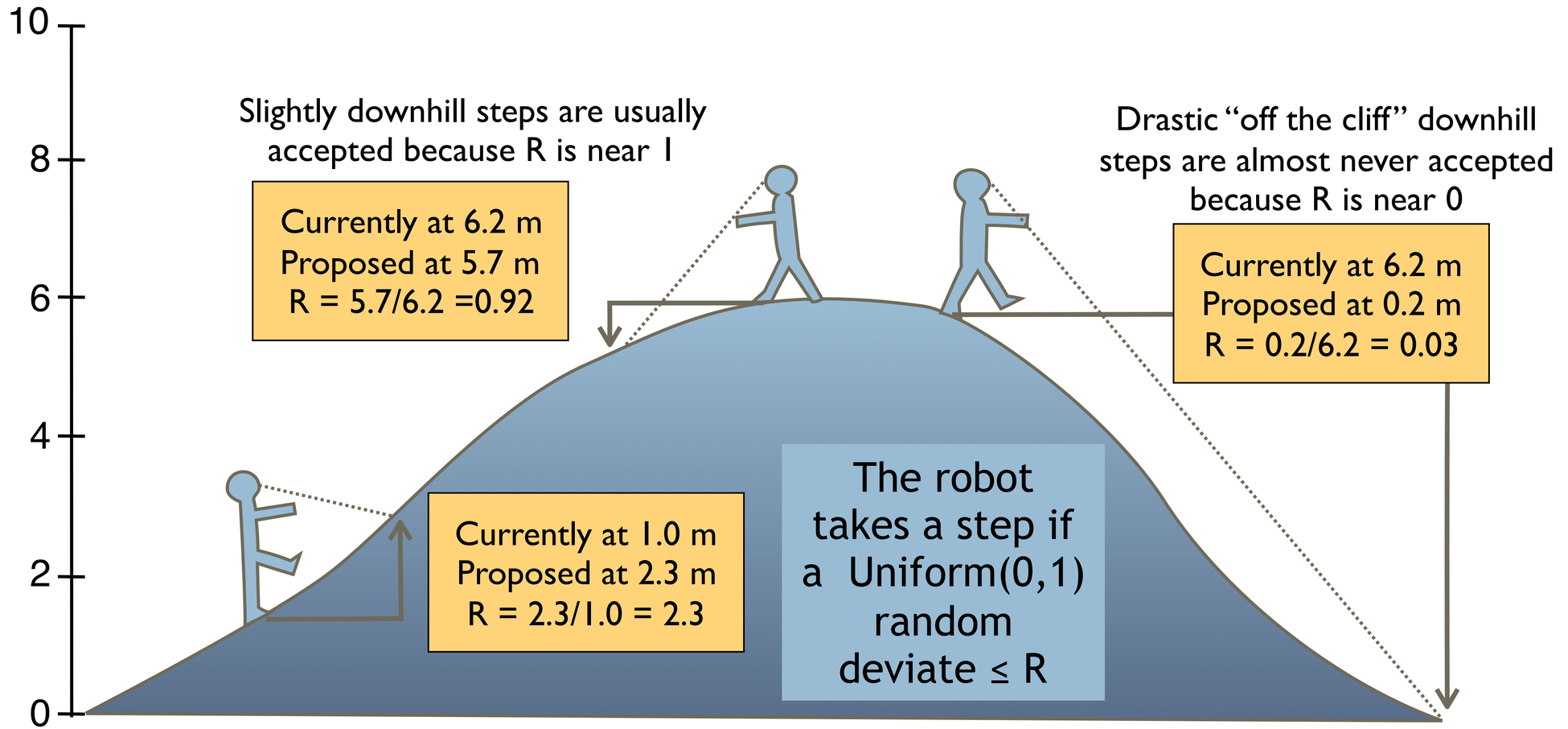


Pseudocode

```
initialize starting values;

for i in mcmc steps
do
     propose new parameter values;
     calculate the Hastings ratio R;

     if( R > 1 )
          accept the new values;
     else
          accept the new values with Pr = R;

     store the values with frequency j;
done
```

MCMC robot's rules





Metropolis et al. 1953. Equation of state calculations by fast computing machines. *J. Chem. Physics* 21(6):1087-1092.

When calculating the ratio (*R*) of posterior densities, the marginal probability of the data cancels.

$$\frac{p(\theta^* \mid D)}{p(\theta \mid D)} = \frac{\dfrac{p(D \mid \theta^*)\, p(\theta^*)}{\cancel{p(D)}}}{\dfrac{p(D \mid \theta)\, p(\theta)}{\cancel{p(D)}}} = \frac{p(D \mid \theta^*)\, p(\theta^*)}{p(D \mid \theta)\, p(\theta)}$$

Posterior odds

Apply Bayes' rule to both top and bottom

Likelihood ratio

Prior odds

Hastings ratio

new parameter values

$$R = \frac{P(\ast \mid 0101\ldots\ 1101\ldots\ 0100\ldots)}{P(\mid 0101\ldots\ 1101\ldots\ 0100\ldots)}$$

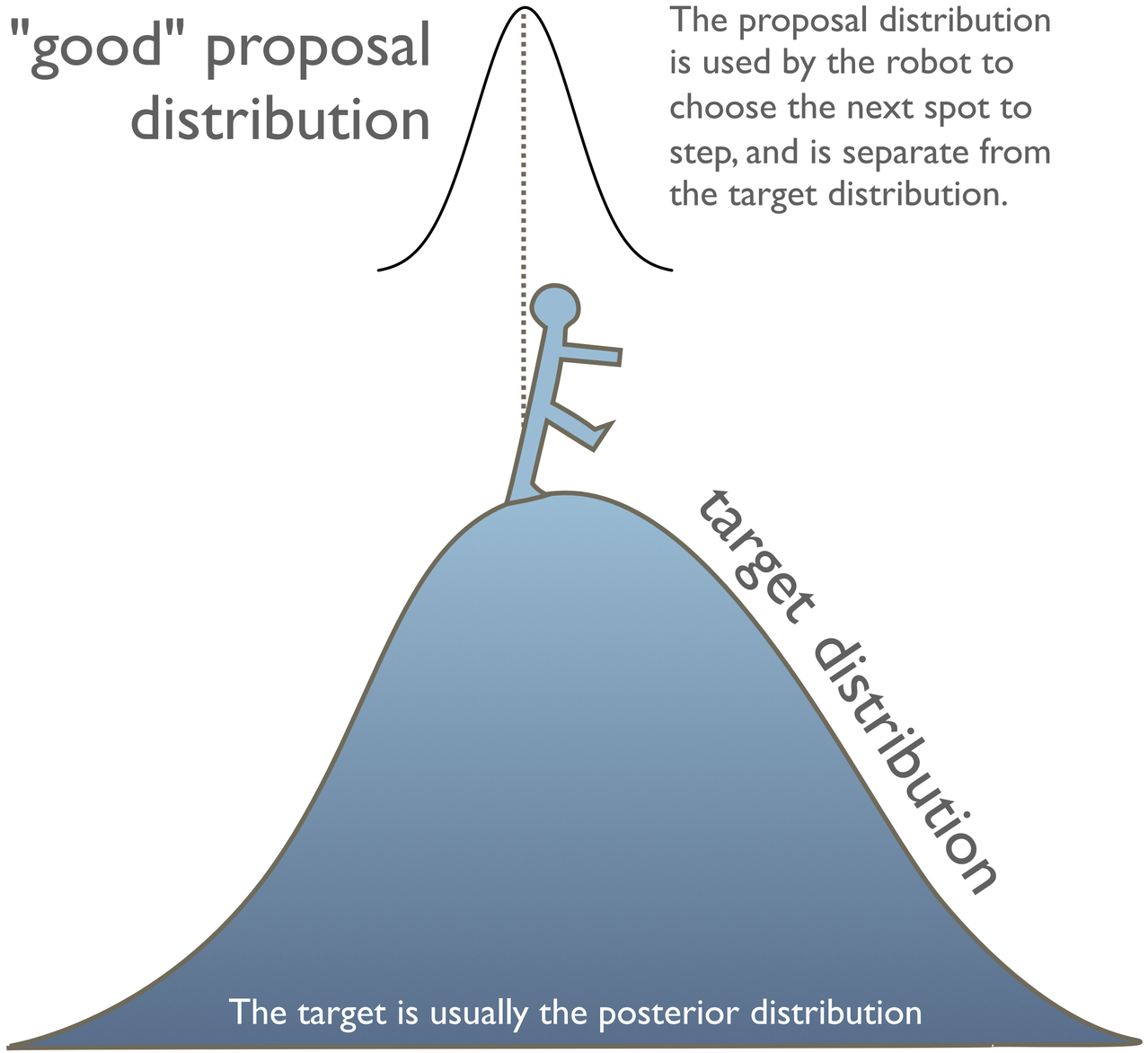
$$= \frac{P(0101\ldots\ 1101\ldots\ 0100\ldots \mid \ast)P(\ast)}{\cancel{P(0101\ldots\ 1101\ldots\ 0100\ldots)}}$$

$$= \frac{P(0101\ldots\ 1101\ldots\ 0100\ldots \mid \ast)P(\ast)}{P(0101\ldots\ 1101\ldots\ 0100\ldots \mid)\ P(\ast)}$$

The marginal probability of the data cancels out

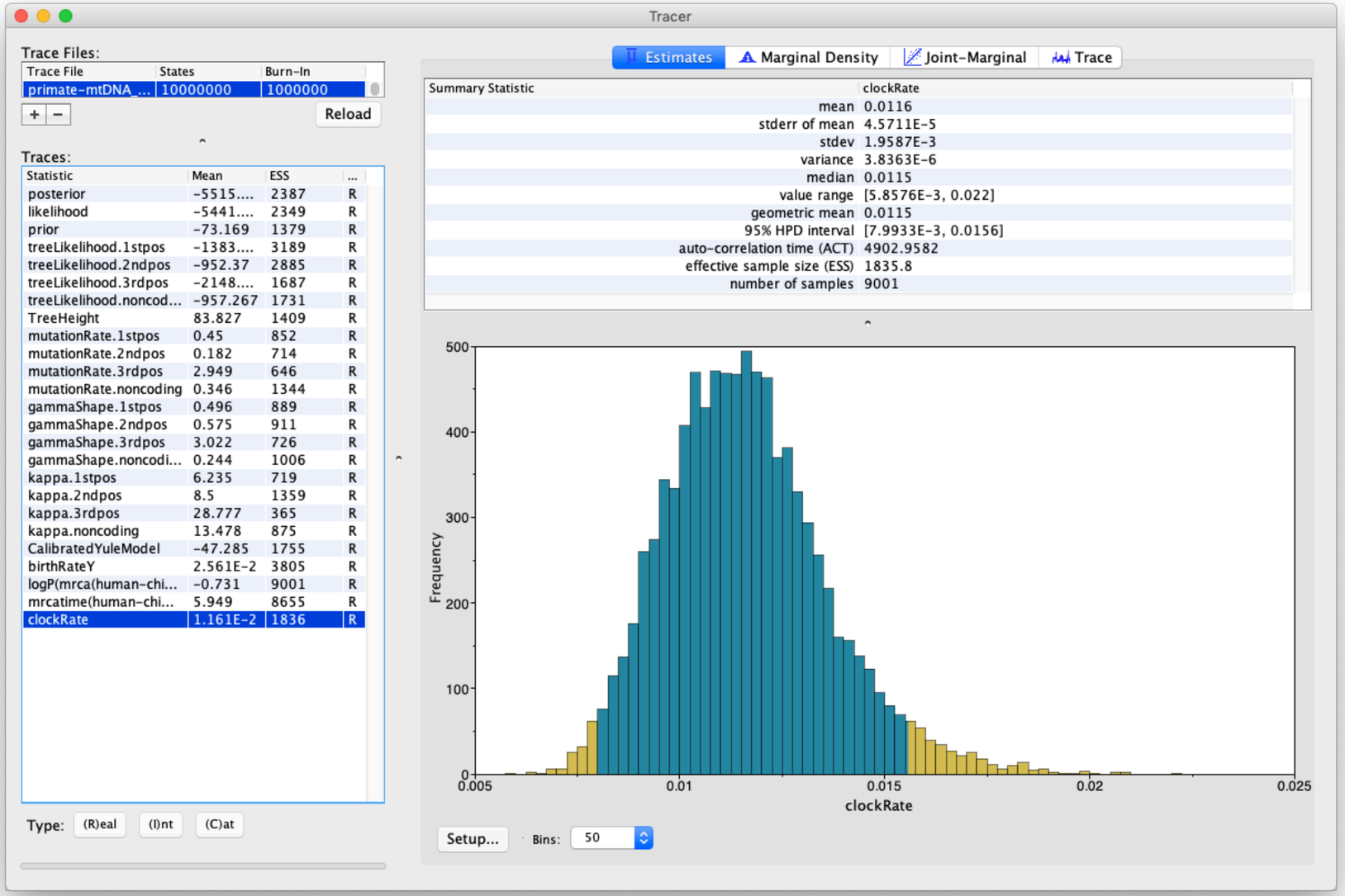
All we're left to calculate is the likelihood ratio and the prior odds ratio

Proposals



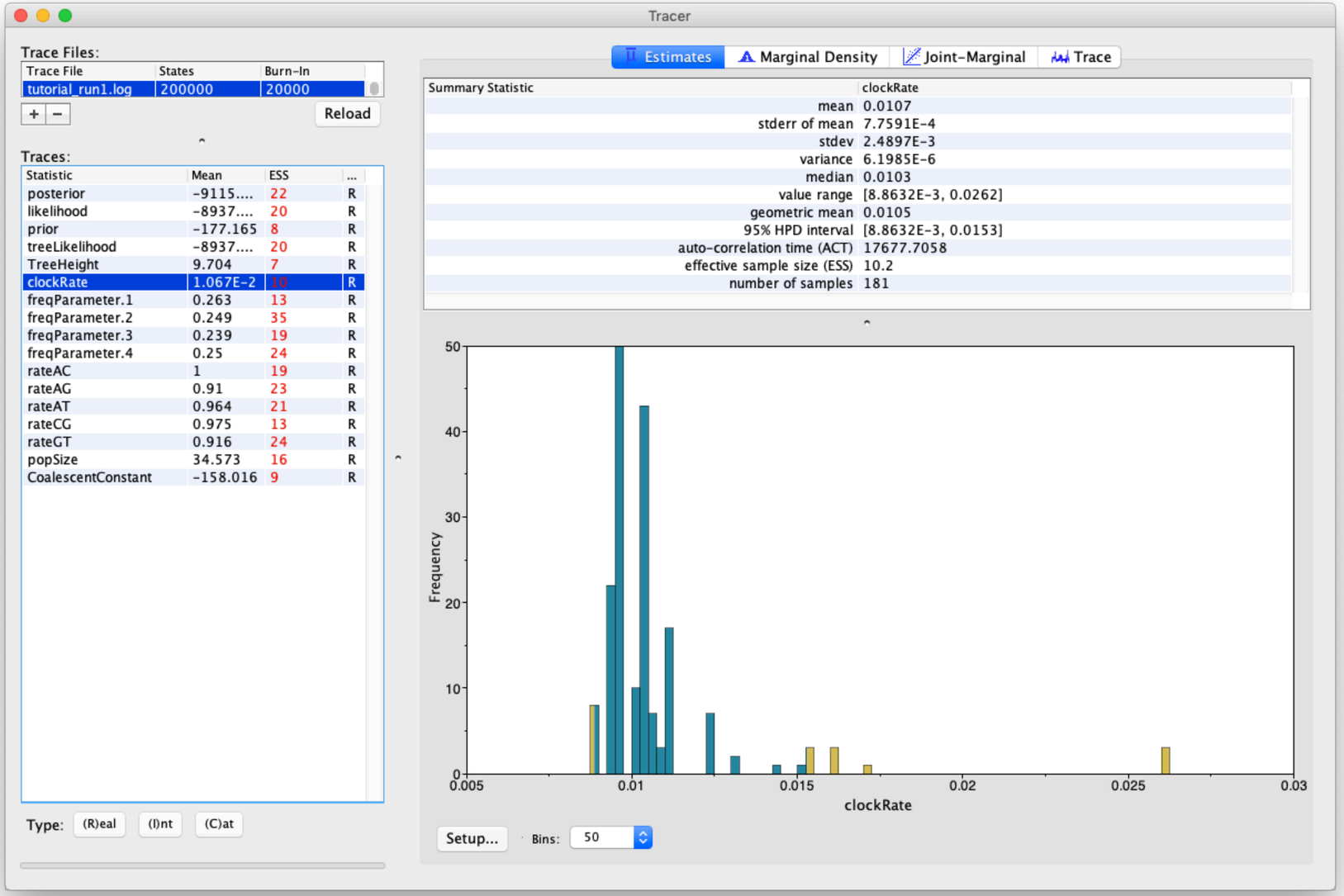
Summarising the posterior

Tracer is an amazing program for exploring MCMC output



Summarising the posterior

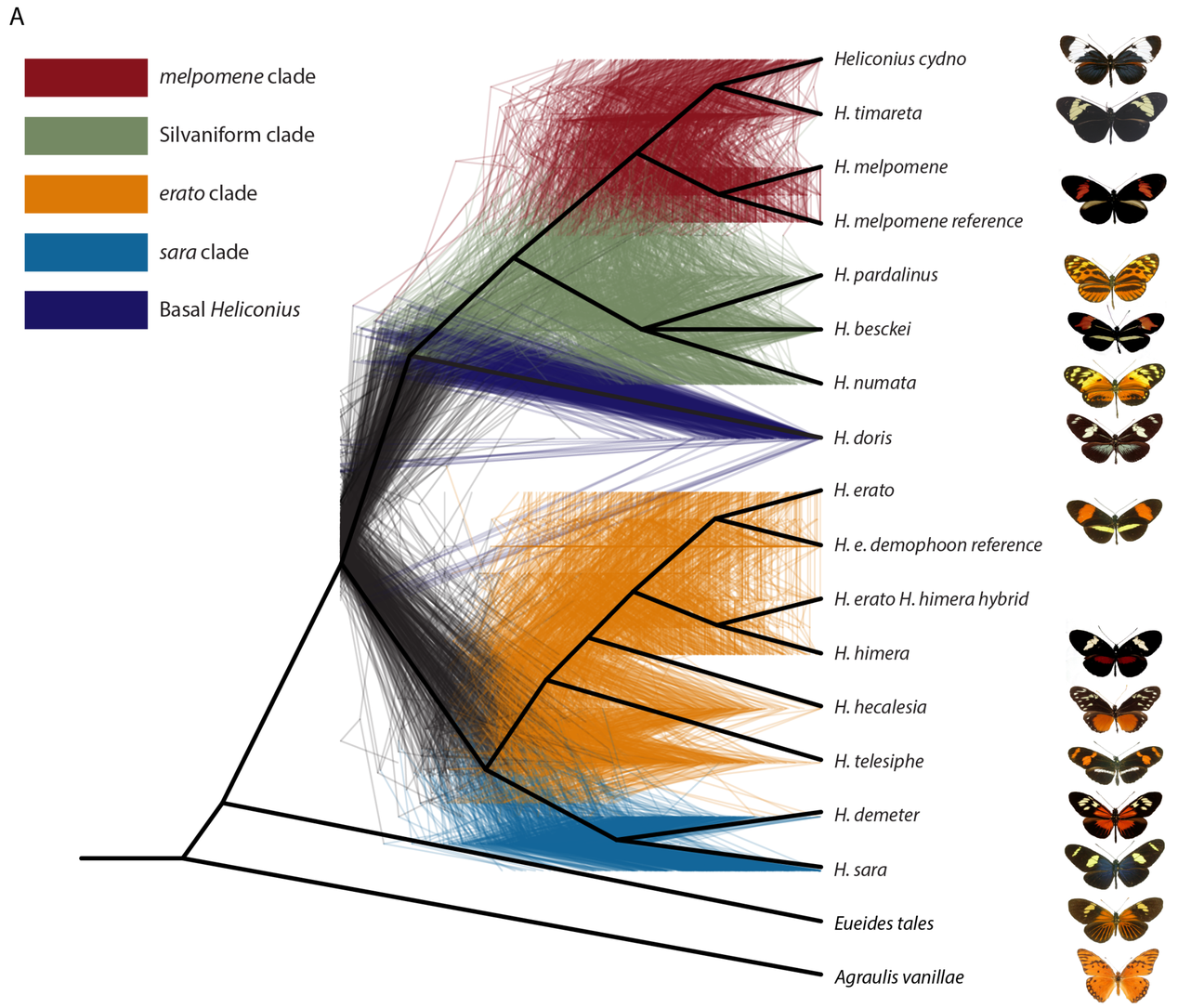
Tracer is an amazing program for exploring MCMC output



Summarising the posterior

Summarising trees is much more challenging

Presenting a single summary tree can be misleading



Summarising the posterior

Maximum clade credibility (MCC) tree — the tree in the posterior sample that has the highest posterior probability (i.e., clade support) across all nodes

The 95% highest posterior density (HPD) — the shortest interval that contains 95% of the posterior probability. The Bayesian equivalent of the 95% confidence interval

Marginal posterior density — the probability of a parameter regardless of the value of the others, represented by the histogram

Exercise