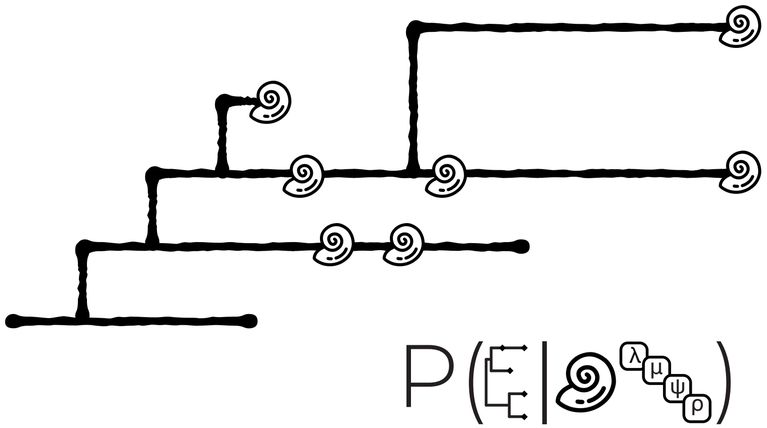
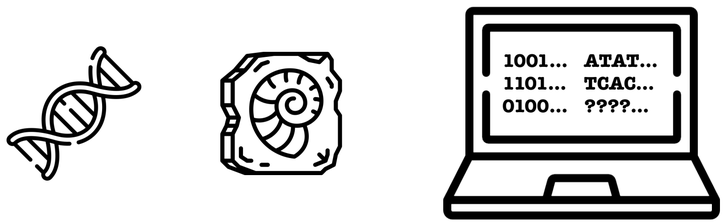


Introduction to statistical phylogenetics

Rachel Warnock

APW 2023 27-08-23



FAU

A bit about me

Worked in both paleo and computational biology groups

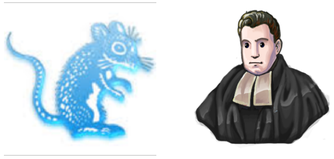
Interested in phylogenetic approaches that can be applied to the fossil record / hypothesis testing in deep time

→ all of the models are applicable to non-paleo problems, e.g., epidemiology, microbiology, archaeology, cell biology


Glasgow
home town
London
BSc, masters
Bristol
PhD
Erlangen
professor
Basel
fellowship
Ames
visitor
Washington
fellowship

development

→ Create and implement new phylogenetic methods



BEAST2 & RevBayes

analysis

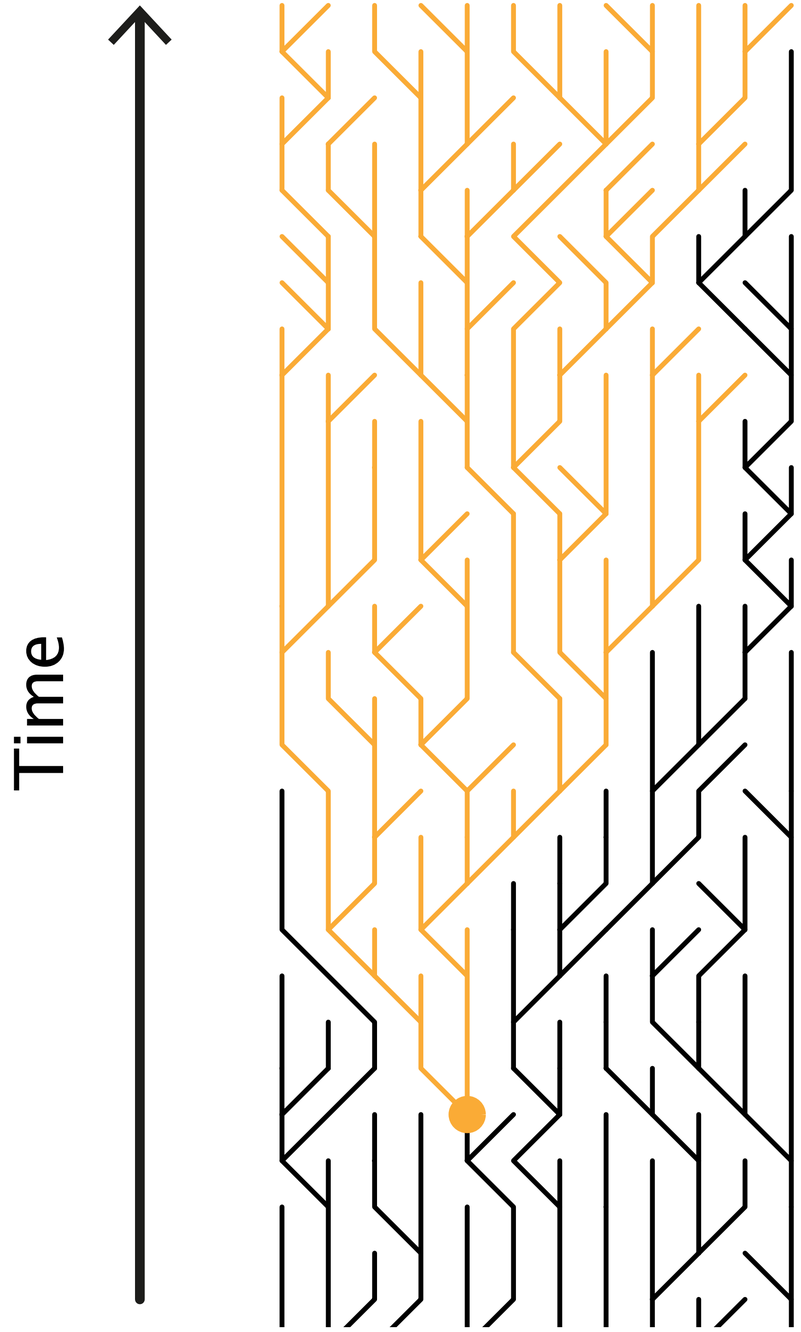
→ Estimate parameters & test hypotheses from real fossil data

The Paleobiology Database

simulations

→ Generate fake data to test methods

Species 1010111...



What is phylogenetics?

populations

species

viruses

cells

languages

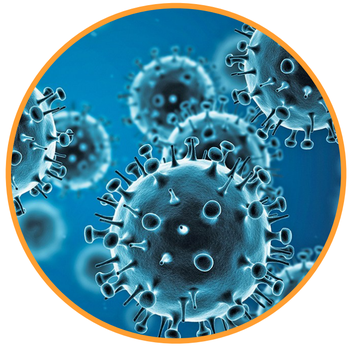
Data

DNA

morphology

words

cells



TO

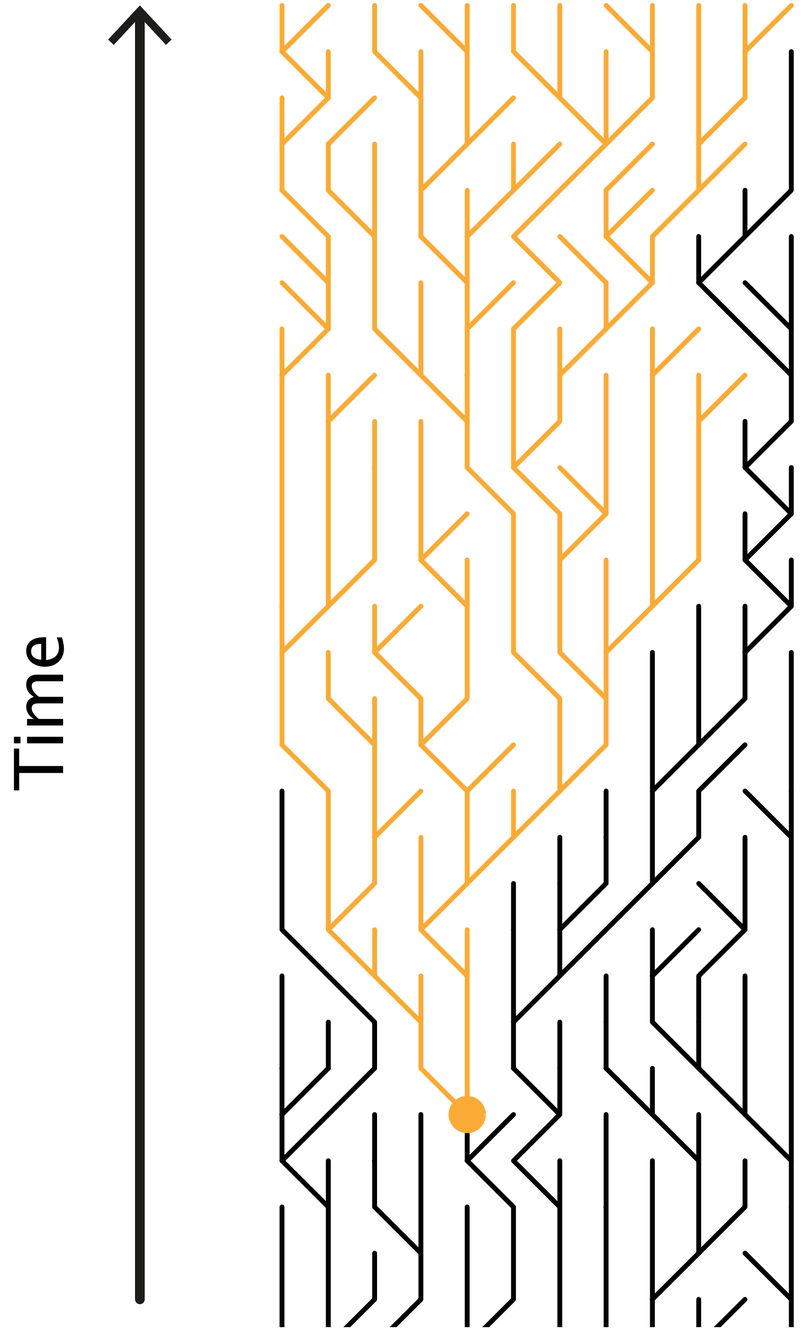
A MOUSE,

On turning her up in her Neſt, with the Plough,

November, 1785.

WEE, ſleeket, cowran, tim'rous *beaſtie*,
O, what a panic's in thy breaſtie!
Thou need na ſtart awa ſae haſty,
Wi' bickering brattle!
I wad be laith to rin an' chaſe thee,
Wi' murd'ring *pattle!*

I'm truly ſorry Man's dominion
Has broken Nature's ſocial union,
An' juſtifies that ill opinion,
Which makes thee ſtartle,
At me, thy poor, earth-born companion,
An' *fellow-mortal!*



What is phylogenetics?

Phylogenetics allows us to study the relationships between entities that are related via an evolutionary process.

We can apply the same principles to any scenario where we have hierarchical (ancestor & descendant) relationships.

The data is anything that can tell us about the relationships between individuals.

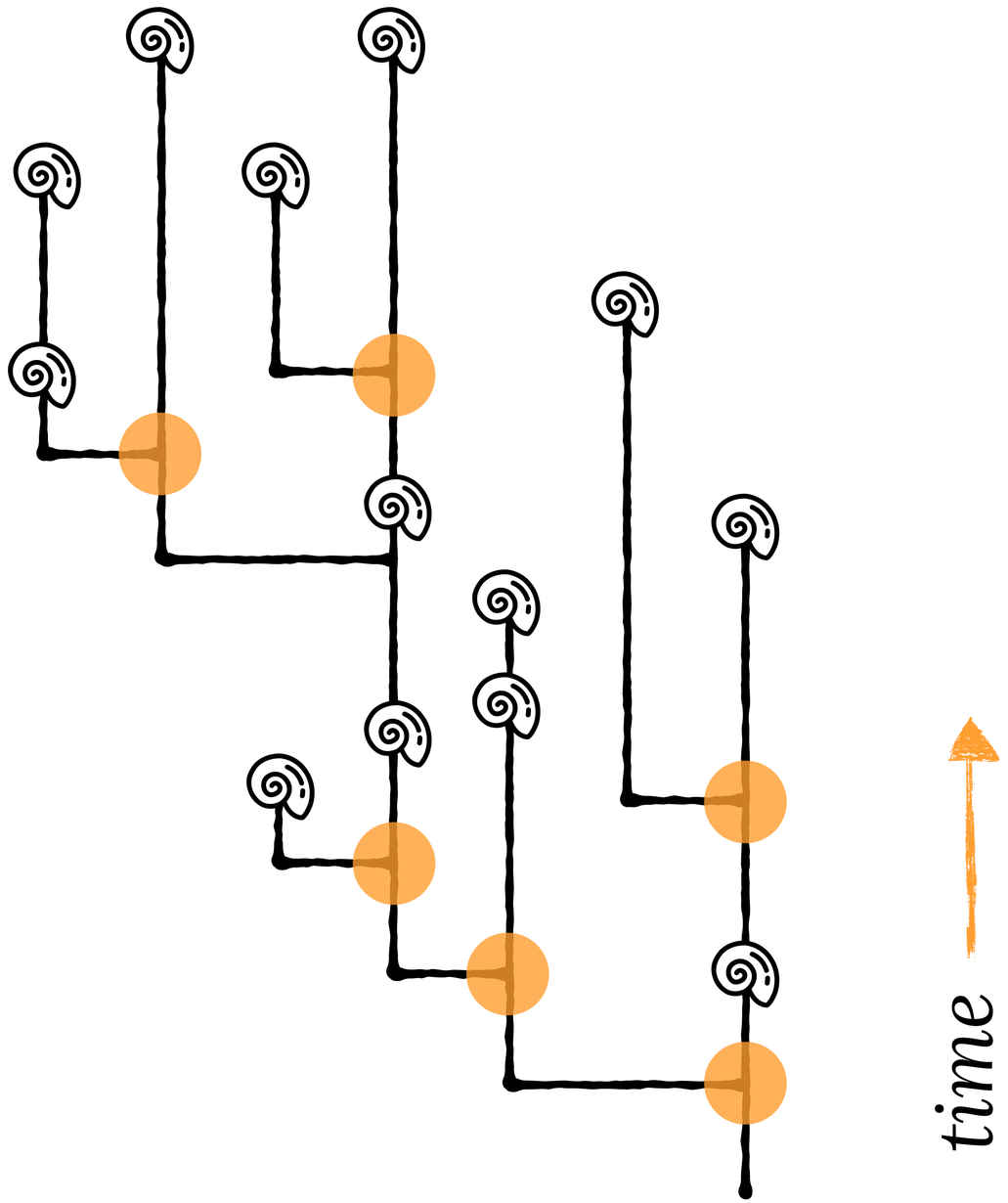
Nothing in biology makes sense except in the light of evolution

— Theodosius Dobzhansky (1973)

Nothing in evolution makes sense except when seen in the light of phylogeny — Jay M. Savage (1997)

A phylogenetic tree captures part of evolutionary history that is otherwise not directly observable.

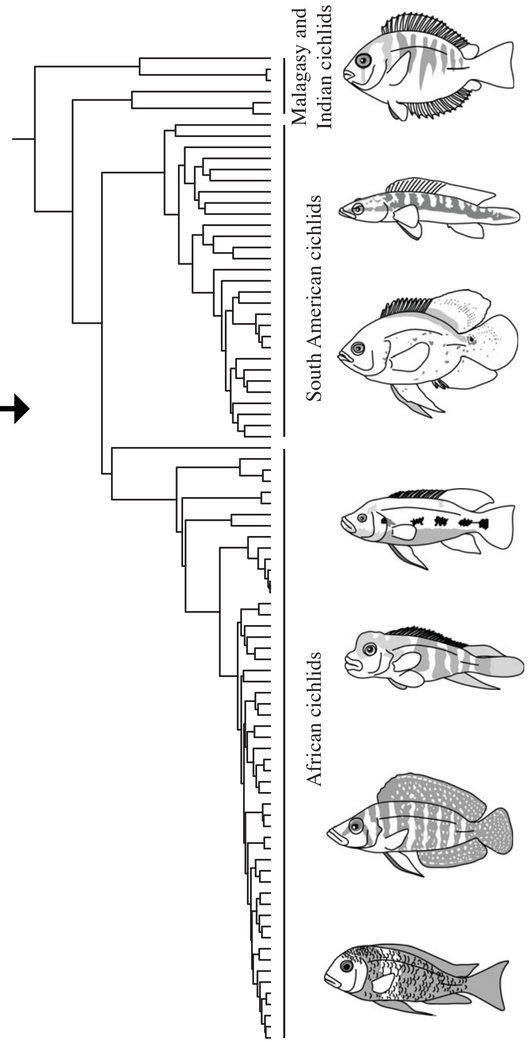
We can date trees by combining character data (molecular or morphological) & temporal evidence.



What can we learn from trees?

Evolutionary relationships

tree topology →



Adapted from Friedmann et al. 2013. PRSB

What can we learn from trees?

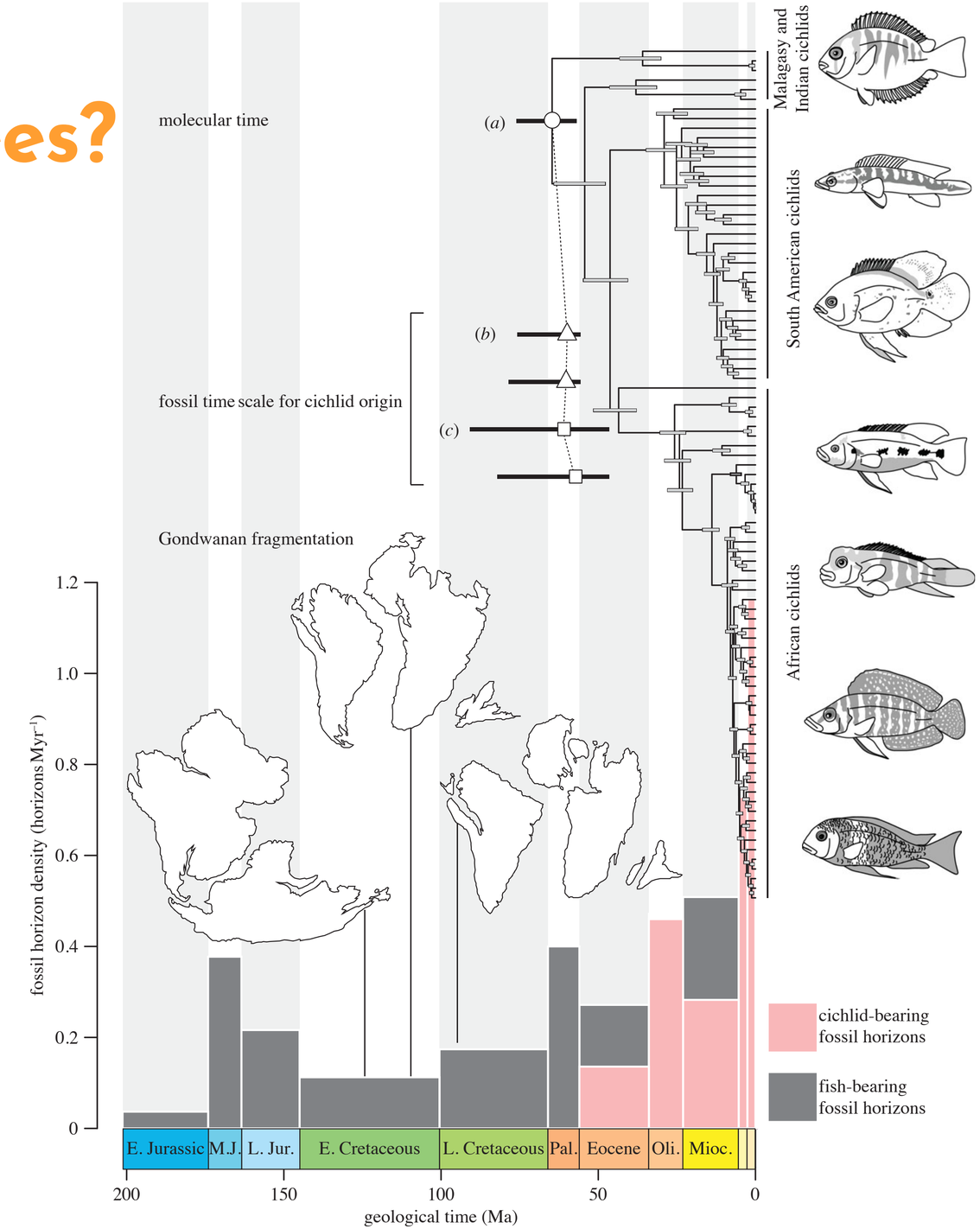
Evolutionary relationships

Timing of diversification events

Geological context

Rates of phenotypic evolution

Diversification rates (origination & extinction)



Adapted from Friedmann et al. 2013. PRSB

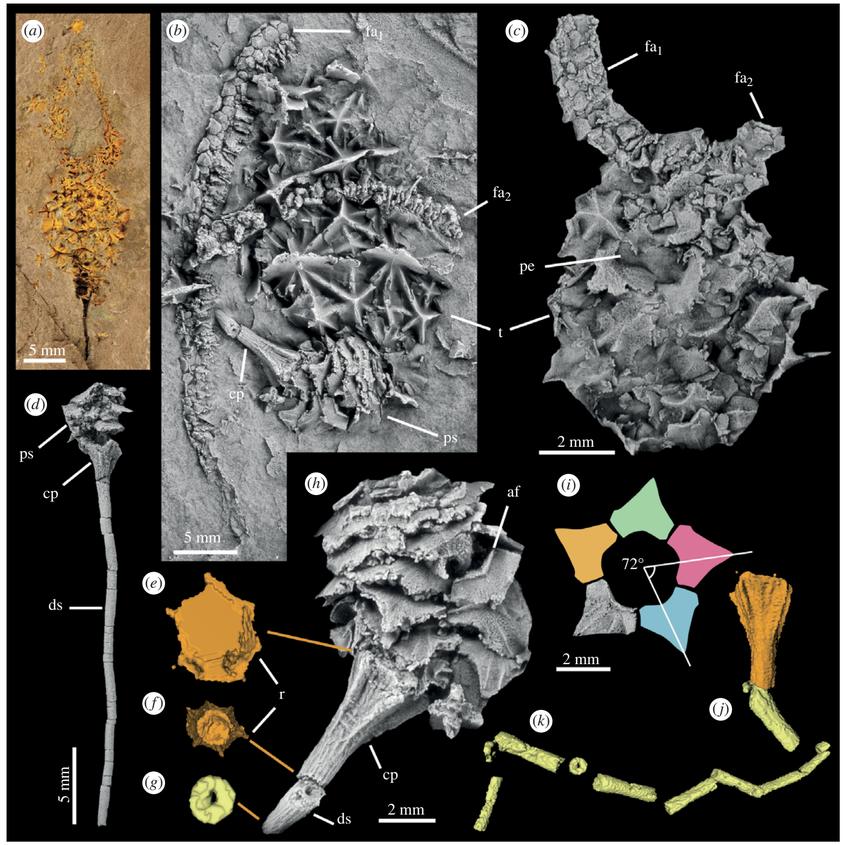
⚠️ Warning — phylogenetics* is really *really* hard.

We're asking a lot of a relative small amount of data, among other reasons.

Phylogenetics is also full of jargon, so don't hesitate to ask questions!

*and palaeobiology in general

```
00151001O?00-100--0000000000
00050001O?200100--0010010000
00250001O?200100--0?10010000
00?5?001O?200100?-0???010110
00150001012010004301000111111
00150001012010104401110111111
??050?????201000440?11011111
01050?01O-210000?501??010110
00020001002101003-1110010110
00020001002110014411210111111
00020111-2100102-??11011121
?103?0?11?1001104-0000010000
10050021101000102--0?00110?20
10050020001010105402?00110020
```



Dibrachicystis purujoensis

Cambrian stalked echinoderms show unexpected plasticity of arm construction
Zamora & Smith. 2012. Proc B

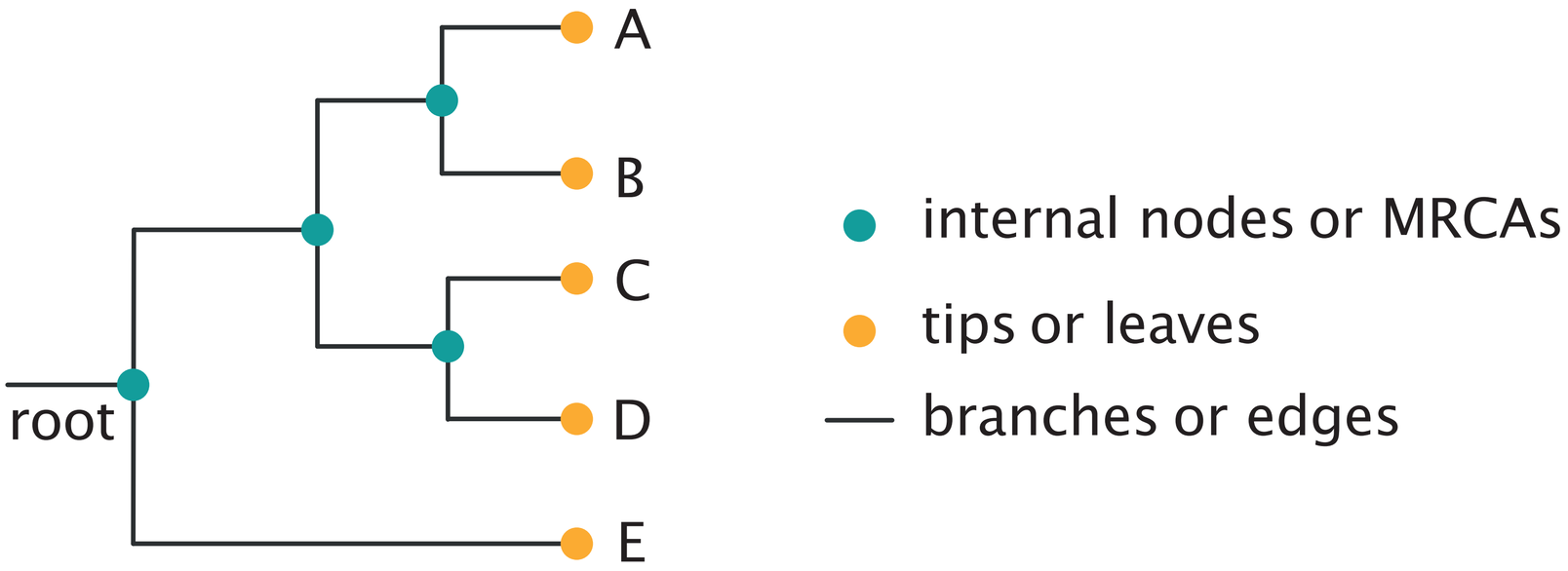
Objectives

Day 1

Day 2

Where do we begin?

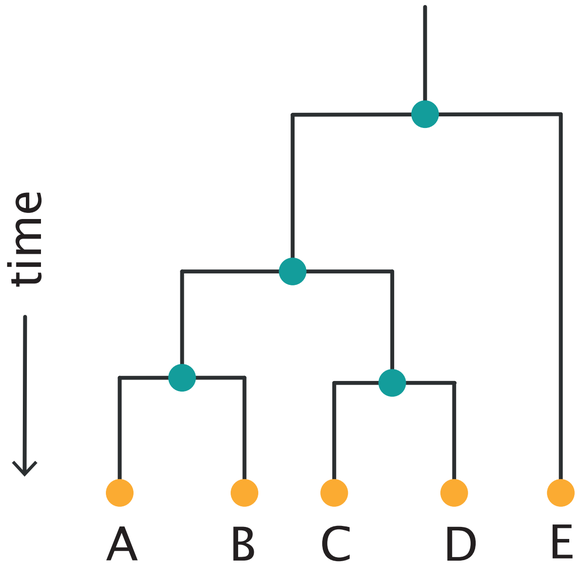
Some useful terms. Note: genetic distance = rate x time



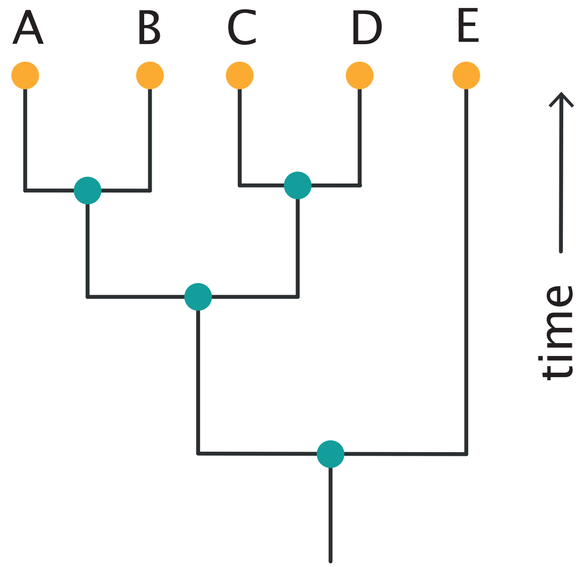
branch lengths = genetic distance OR time

Where do we begin?

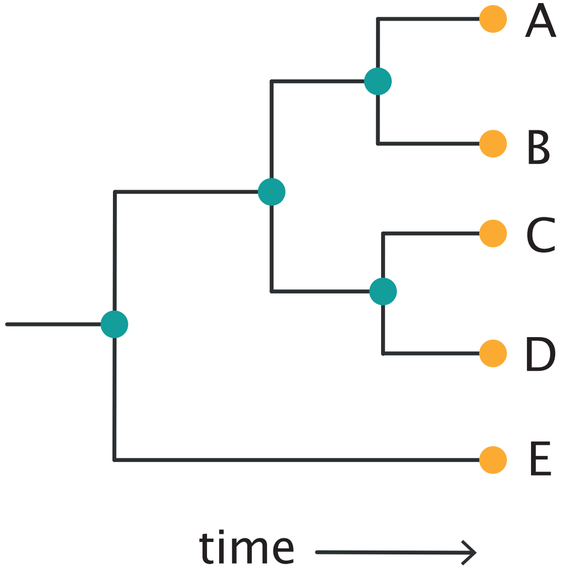
Tip for orientation: look for the root!



Computer science, maths

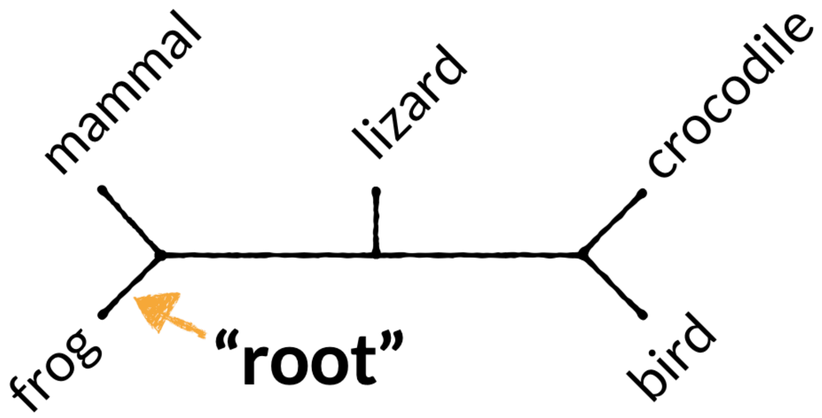


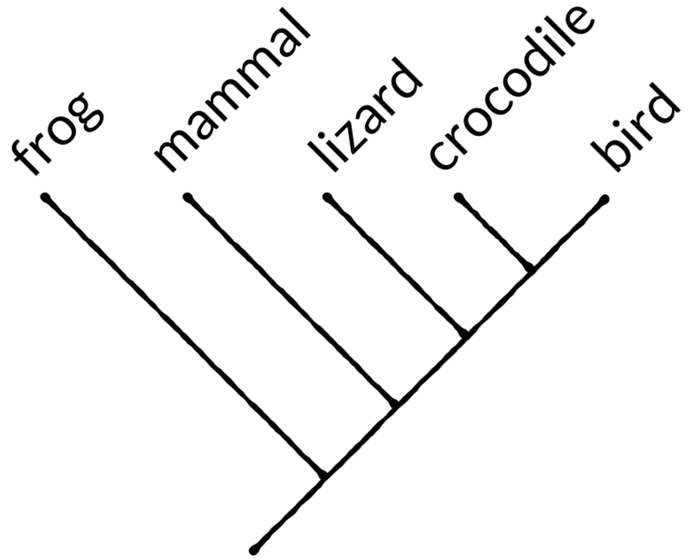
Geology



Evolutionary biology

Rooted versus unrooted trees





Phylogenies are unrooted by default, because phylogenetic characters don't contain information about the direction of time.

Image source Philip Donoghue

Rooted versus unrooted trees

We have to find a way of breaking one of the branches in two, where the break represents the oldest divergence in the tree.

The most common approach is to use an outgroup — a taxon that we know is more distantly related than any of the taxa within the ingroup.

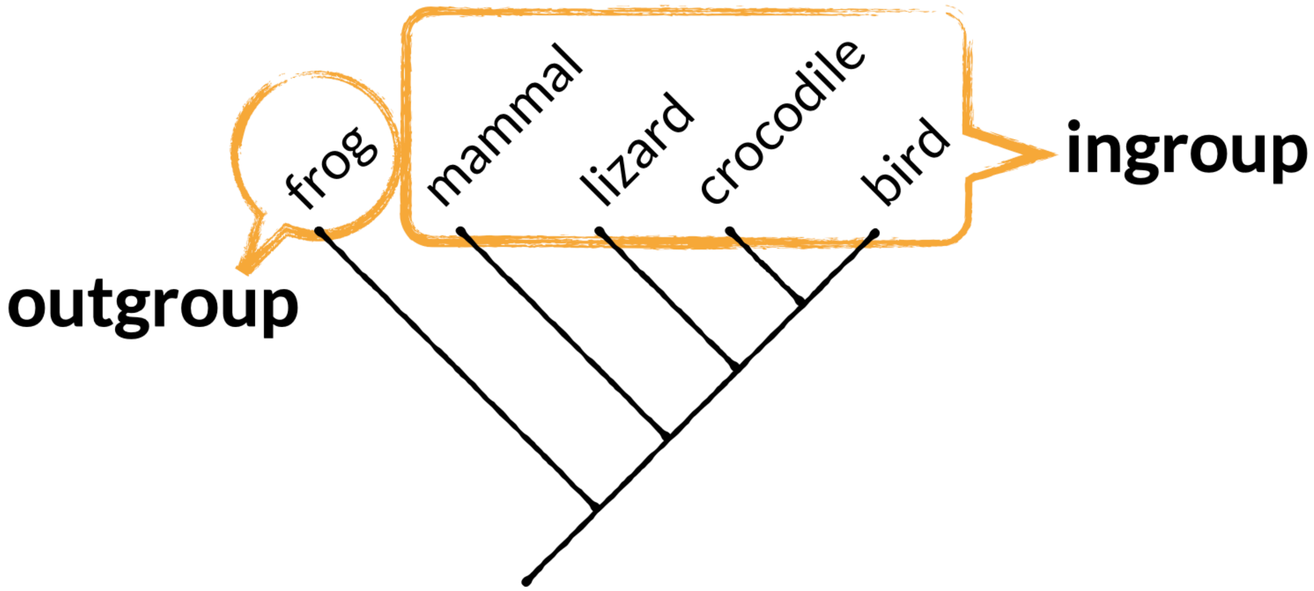
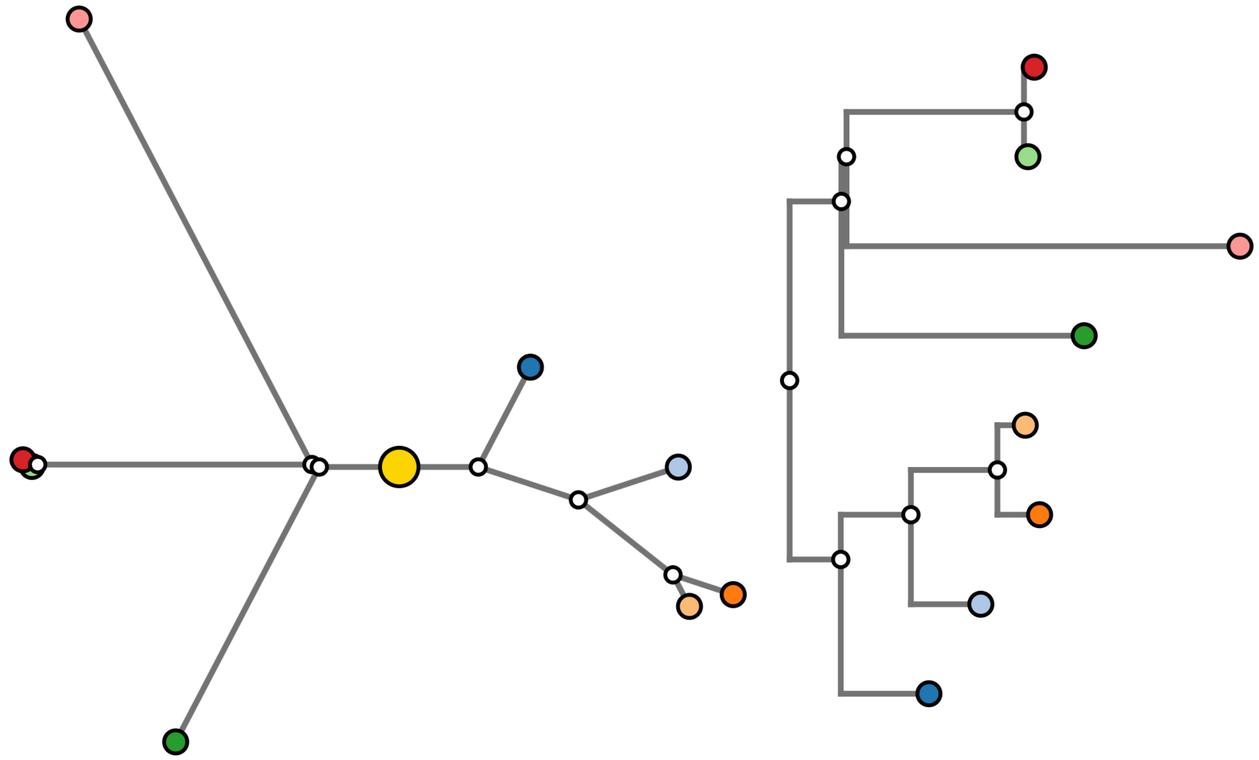


Image source Philip Donoghue

Rooted versus unrooted trees

By default phylogenies are not rooted.

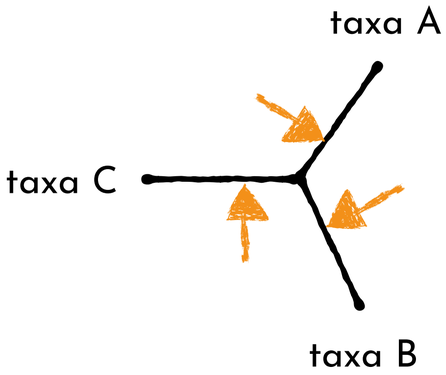
We need an **outgroup** OR a **model** that incorporates **time**.



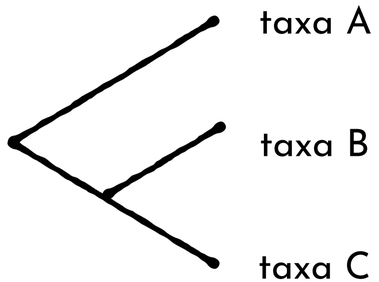
Use Art Poon's online tool to explore this further.
Click here to learn more about reading trees.

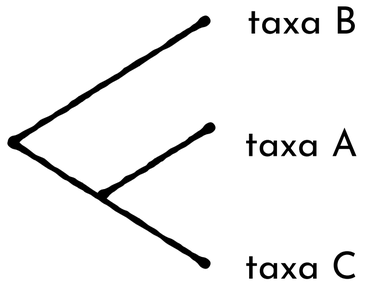
Rooted versus unrooted trees

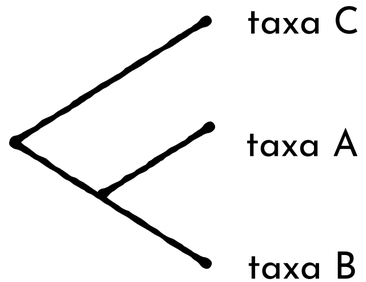
How many possible trees are there for 3 species?



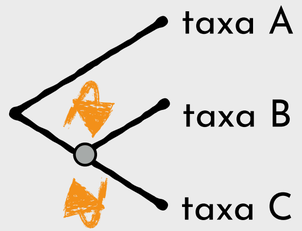
unrooted = 1

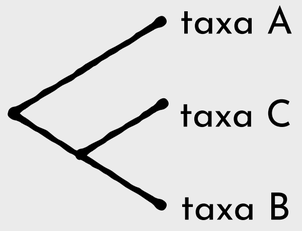






rooted = 3





Note these 2 trees are the same! B and C are more closely related.

Exercise

Character	taxa A	taxa B	taxa C	taxa D	taxa E
Lungs	0	1	1	1	0
Jaws	0	1	1	1	1
Feathers	0	0	1	0	0
Gizzard	0	0	1	1	0
Fur	0	1	0	0	0

How many possible unrooted or rooted trees are there?

What do you think the correct rooted tree should be?

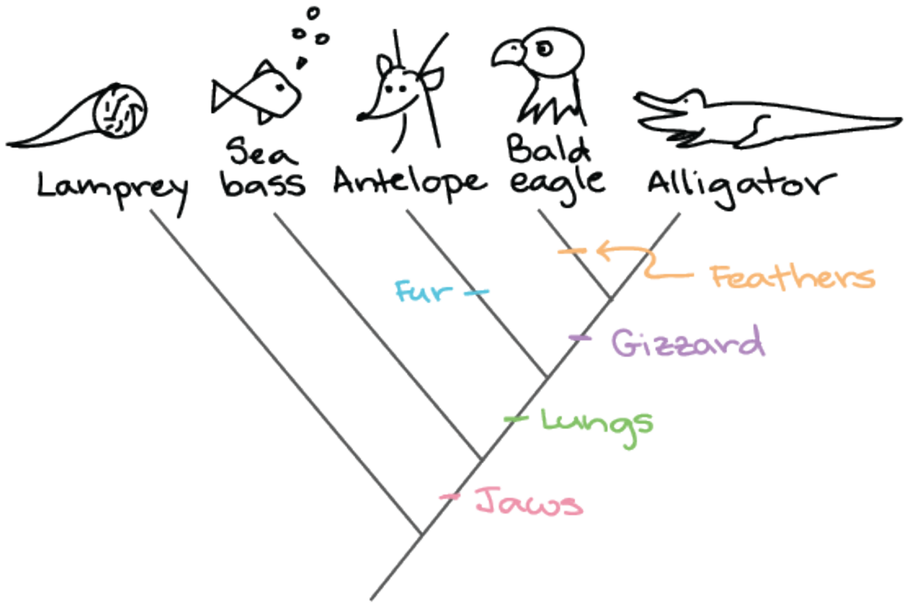
Write down your logic.

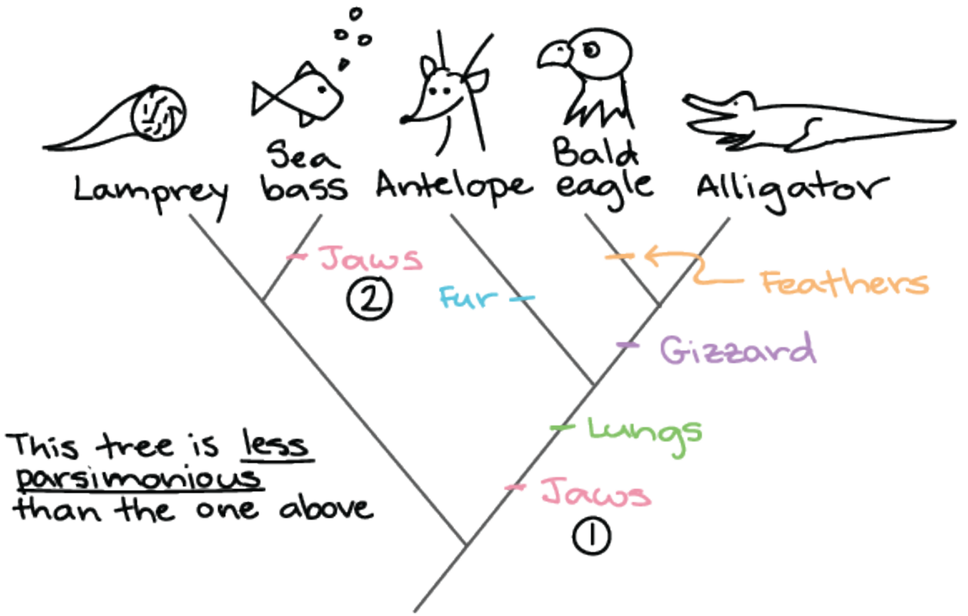
There are a huge number of possible trees!

# species	# unrooted trees	# rooted trees
3	1	3
4	3	15
5	**15**	**105**
6	105	945
7	945	10395
8	10395	135135
9	135135	2027025
10	2027025	34459425

Exercise

What do you think the correct tree should be?





A = Lamprey, B = Antelope, C = Bald eagle, D = Alligator, E = Sea bass

Source Khan Academy

How do we find the “best” tree?

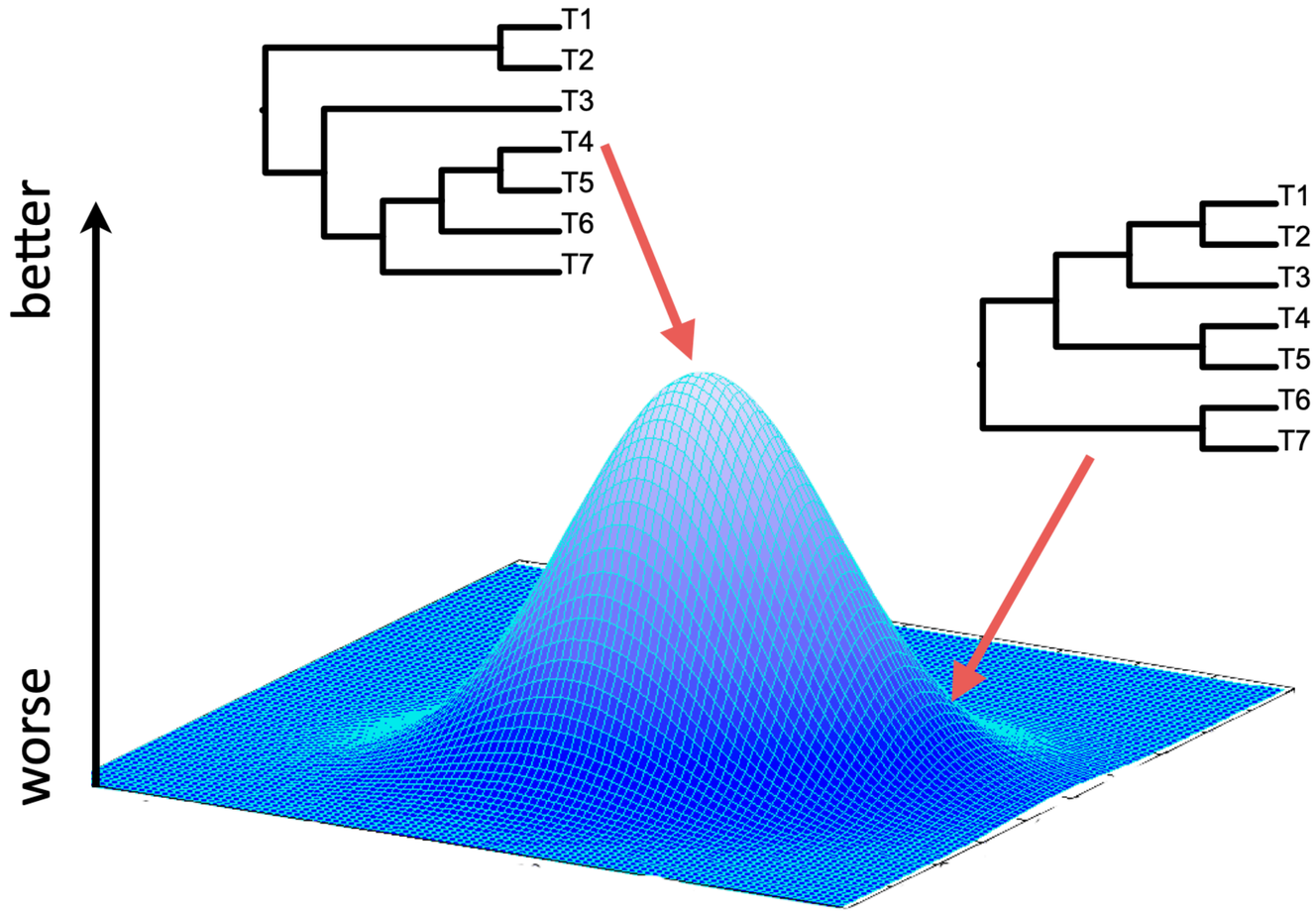


Image source Tracy Heath

Exercise

What do you think the correct tree should be?

Write down your logic.

→ Most people intuitively assume the tree with the **fewest** changes is correct.

→ This approach to tree building is called **parsimony**.

Of course it depends how you measure "best"

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

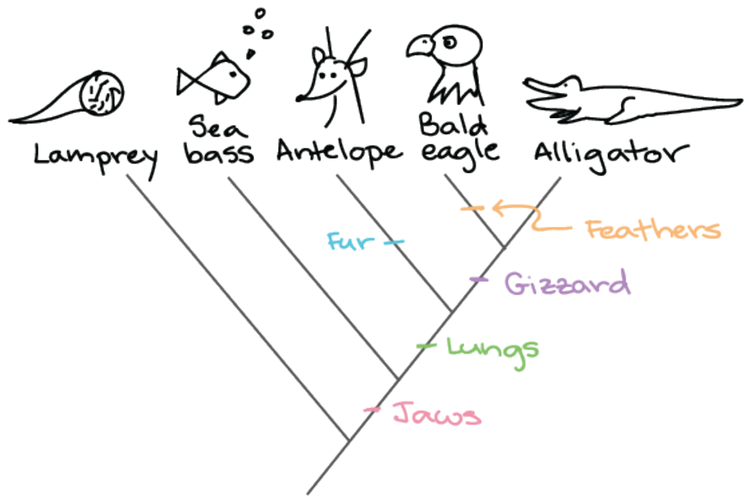
Both maximum likelihood and Bayesian inference are **model-based** approaches.

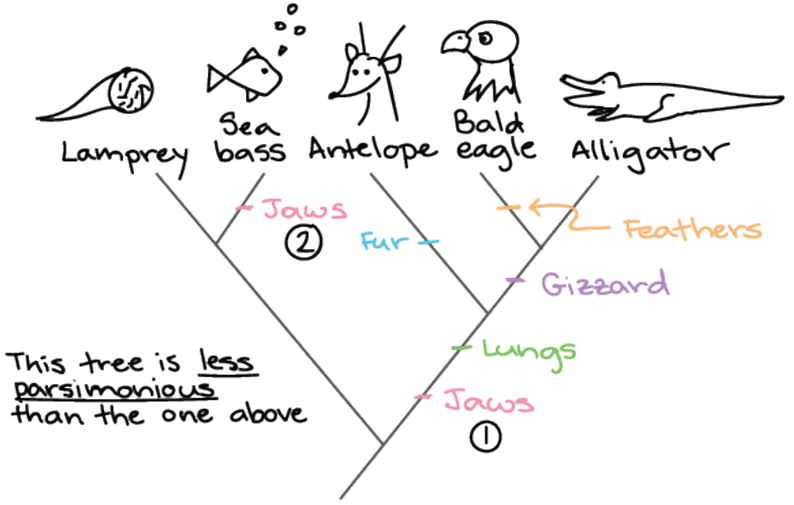
Parsimony

In reality, we **never** know the true tree.

Maximum parsimony selects the tree (or trees) that require the fewest number of changes.

Given two trees, the one minimising the parsimony score (i.e., the minimum number of changes) is the better one.





Branch lengths = number of observed changes or steps.

Parsimony

Based on the parsimony principle: assume simpler explanations are better than complex ones. *The greatest advantage of parsimony is its beautiful simplicity* (Yang, 2014).

It is computationally fast and often produces sensible results.

Parsimony does not make **explicit** assumptions about the evolutionary process that generated the observed data. Some have argued that parsimony is "assumption free" — its not! Parsimony makes **implicit** assumptions.

Exercise 1: intro to phylogenetics using R

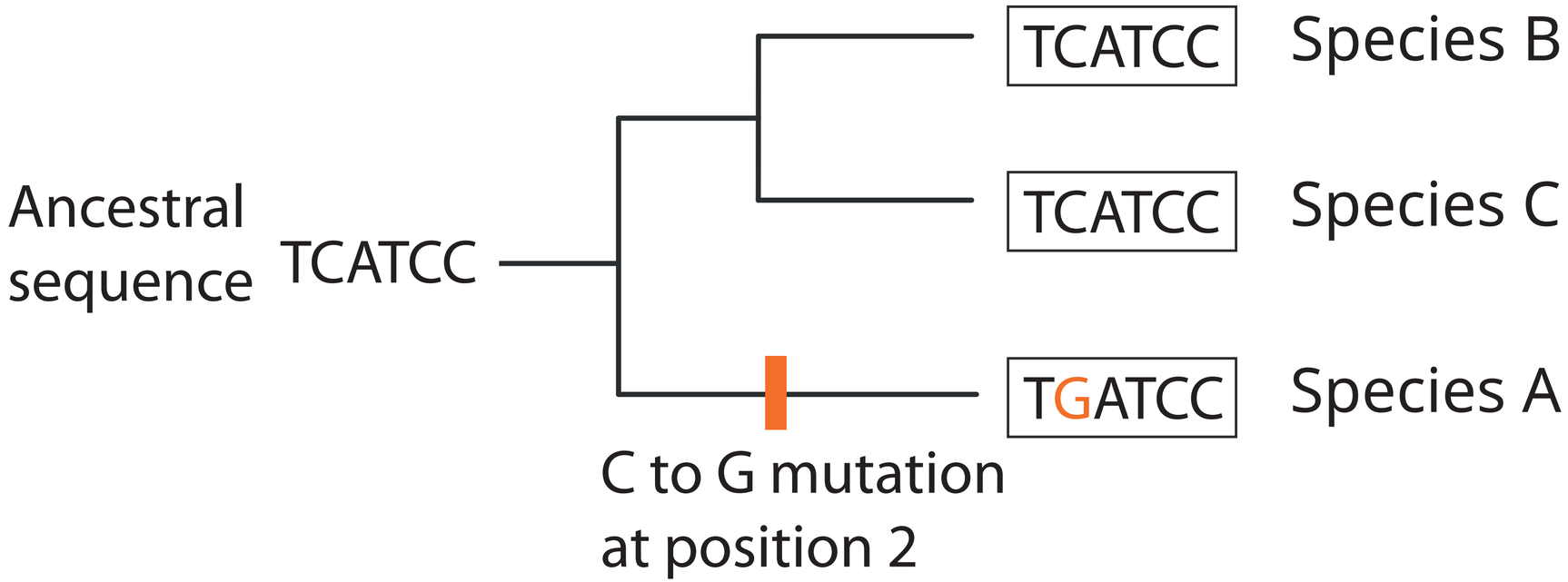
Convergence or homoplasy

Homoplasy: a trait that is found in two species, but not in their common ancestor.

The bluebird, Pterosaur (extinct) and fruit bat: 3 different vertebrates independently lightened bones and transformed hands into wings.

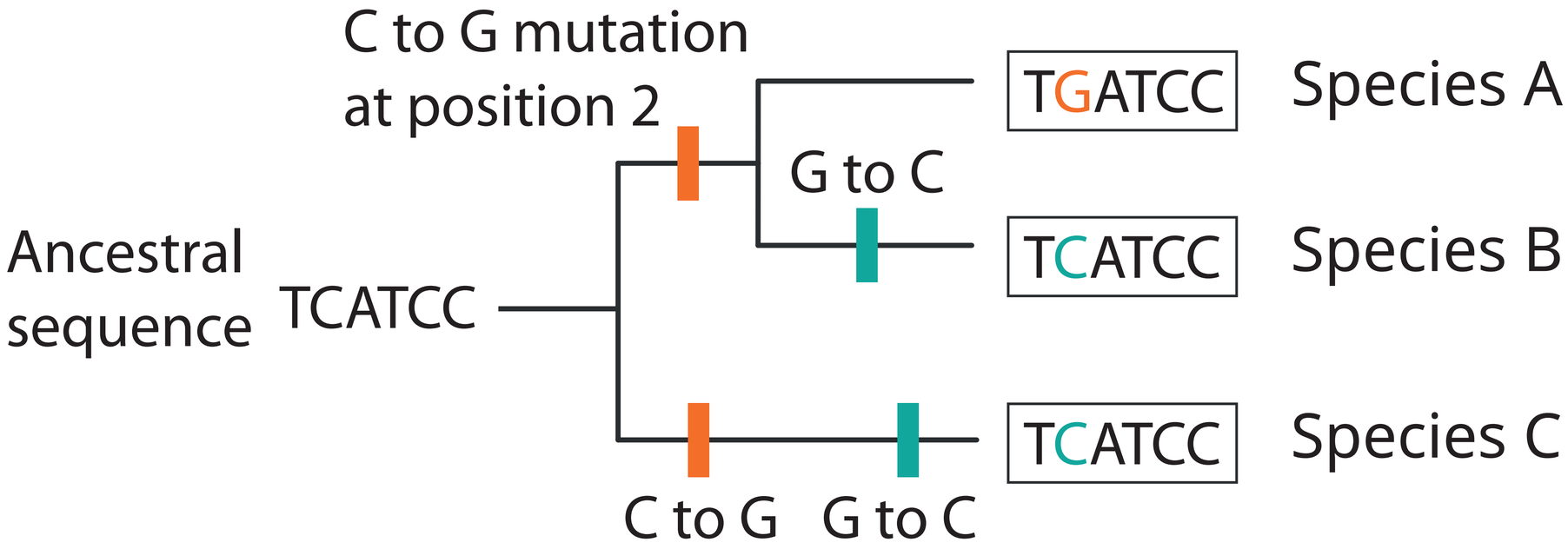
Image source Convergence Evolution: an introduction

Molecular convergence



If we assume the simplest solution is correct, this could mislead our inference if the underlying process is more complex.

Molecular convergence



If we assume the simplest solution is correct, this could mislead our inference if the underlying process is more complex.

Parsimony

When we build a tree using parsimony and observe convergence, **ad hoc** explanations (e.g., convergence, reversals) are required to explain the patterns.

In the case of birds, pterosaurs and bats, we know based on other anatomical features that these taxa are distantly related, but convergence can interfere with our ability to recover the correct tree. In fact, this is very common.

Parsimony has been demonstrated to be **statistically inconsistent**. An estimator is consistent if it is guaranteed to get the correct answer with an infinite amount of data. Felsenstein (1978) demonstrated that in some situations, parsimony is inconsistent, i.e., it will recover the wrong tree, even with an infinite amount of data.

Long branch attraction

If you have long branches (due to higher rates of evolution), the probability of misleading parsimony due to convergence is much higher.

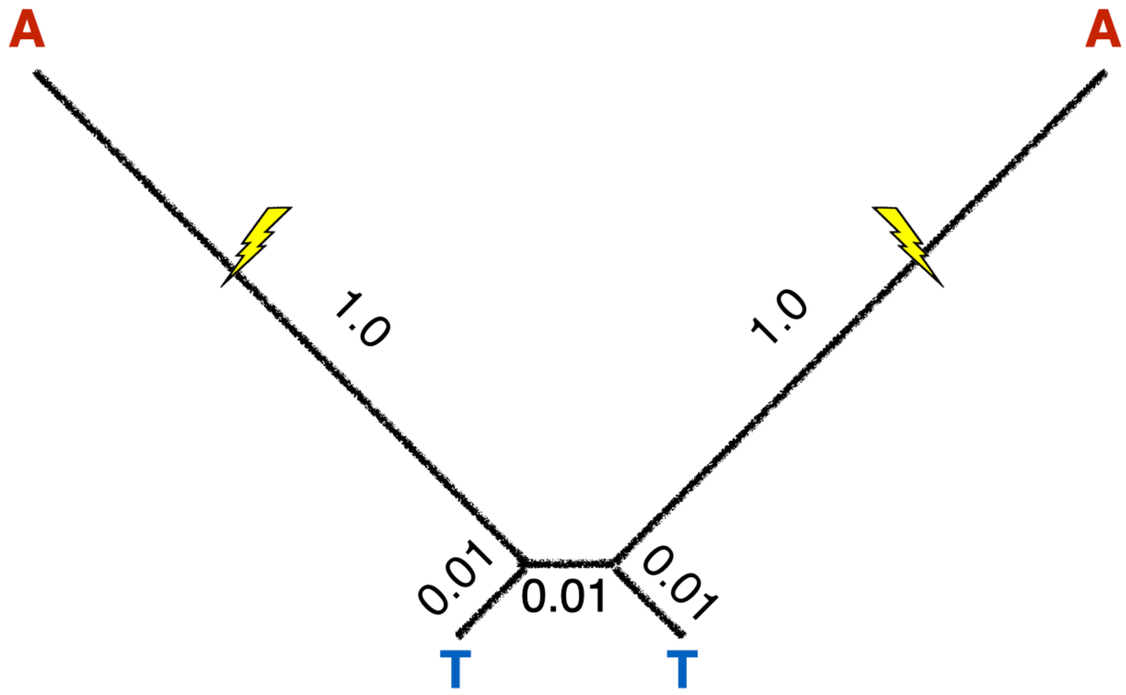


Image source Tracy Heath

Long branch attraction

Parsimony is almost guaranteed to get the tree below wrong. It will incorrectly place two long branches (T1,T3) together as sister lineages. More data will make the problem worse, making this approach statistically inconsistent.

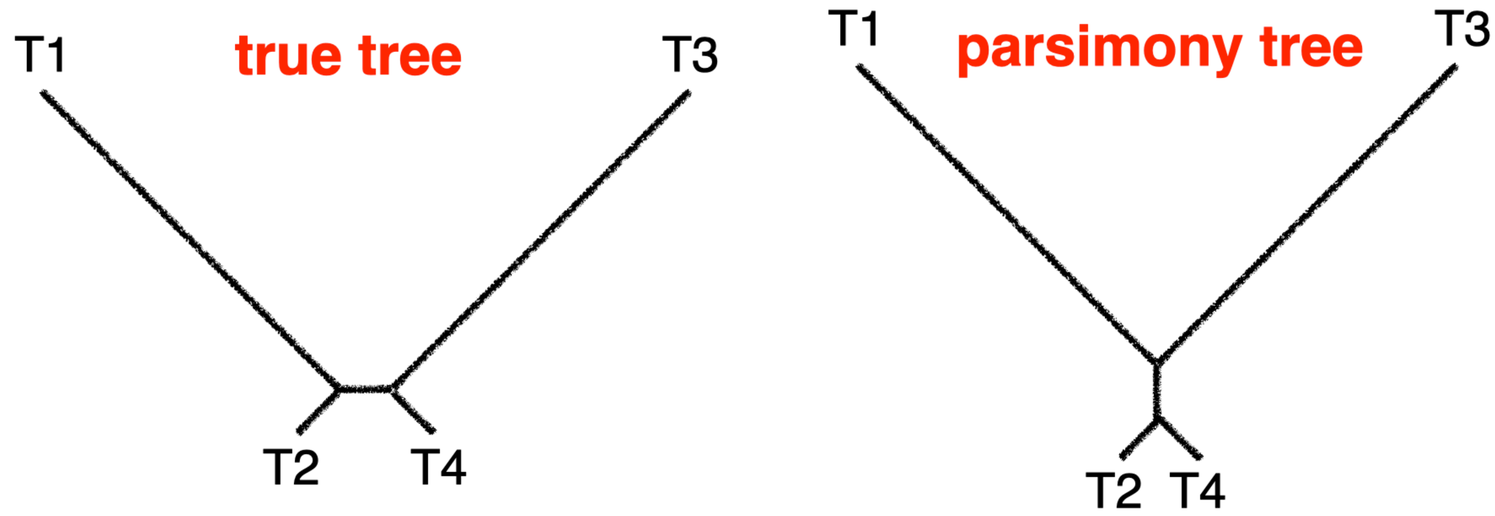
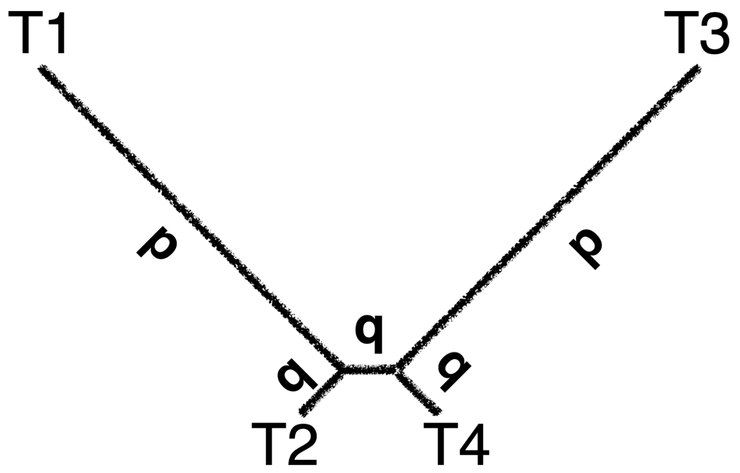
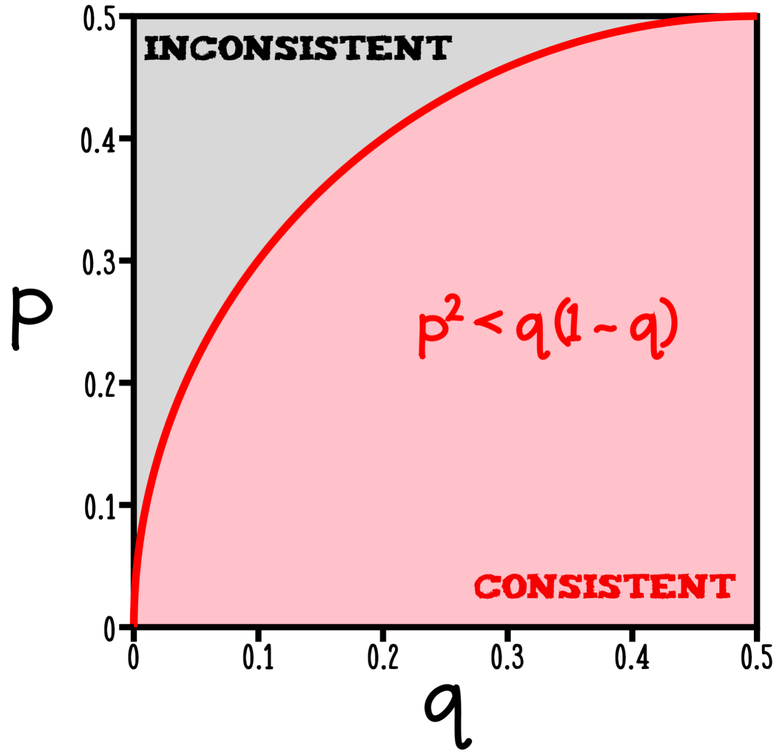


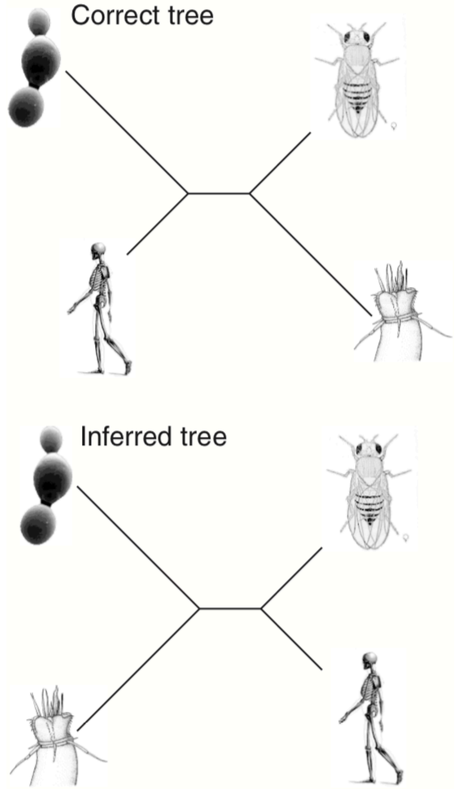
Image source Tracy Heath

Long branch attraction



Here, the branch lengths represent probability (p, q) of change along that branch.





Felsenstein, Inferring Phylogenies, (2004)
Image source Tracy Heath
c.f. Ecdysozoa vs. Coelomata, Telford et al. (2005)

Long branch attraction

Important: this issue can affect all tree building methods! And all types of data (e.g., DNA, morphology).

Things that (sometimes) help: high quality data, increased taxon sampling inc. shorter branching outgroups, models that more reliably capture the variation in evolutionary rates.

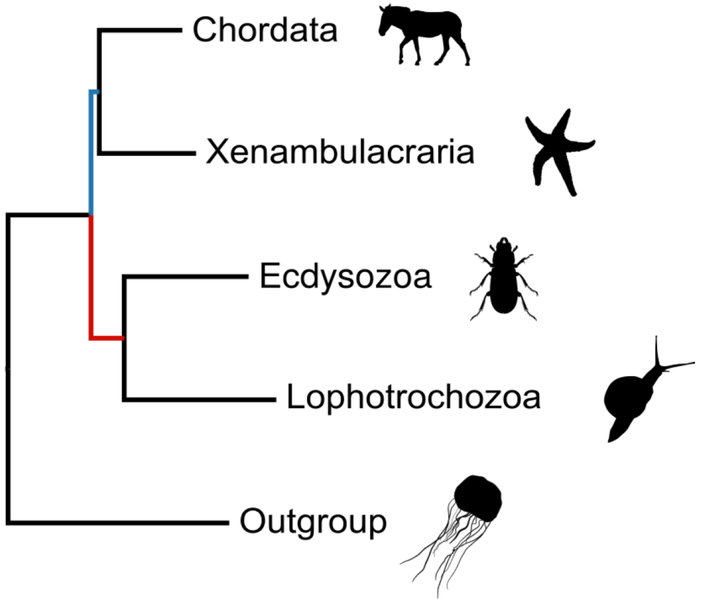
Parsimony vs. model-based approaches

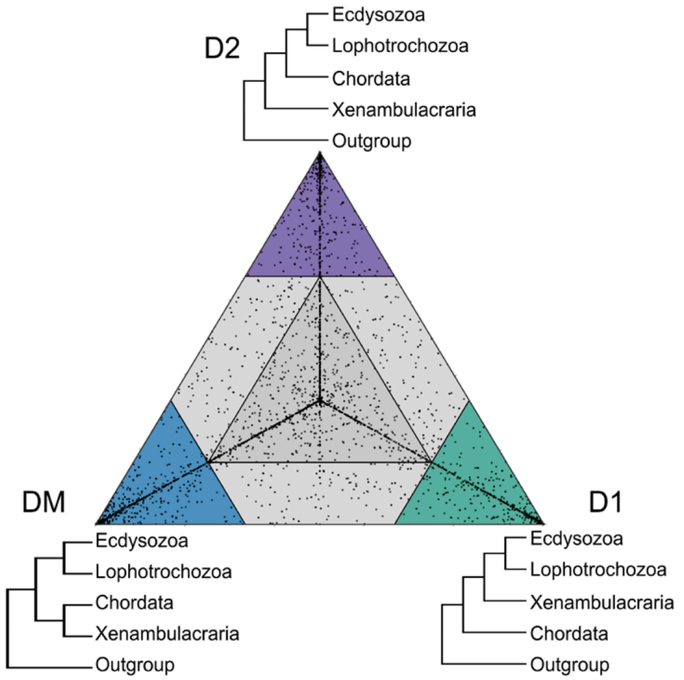
Model-based approaches assume an **explicit** model of molecular or morphological evolution.

If evolutionary distance is relatively small, model based approaches and parsimony will often recover the same tree.

As distance increases, the amount of homoplasy (i.e., convergent or parallel changes) also increases, parsimony is more likely to recover the wrong tree.

Short internal branches pose a huge challenge for any approach





Kapli et al. (2021) Science Advances – support for deuterostomes (chordates + echinoderms) varies across datasets and analyses under different models, probably caused by the extremely short (blue) branch associated with this group.

Summary so far

Parsimony is simple and intuitive but makes **implicit** assumptions about the evolutionary process.

Next, we'll explore model-based approaches — these are more flexible and make **explicit** assumptions → it's very important you to try to understand what these are!

What do we mean by model?

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?

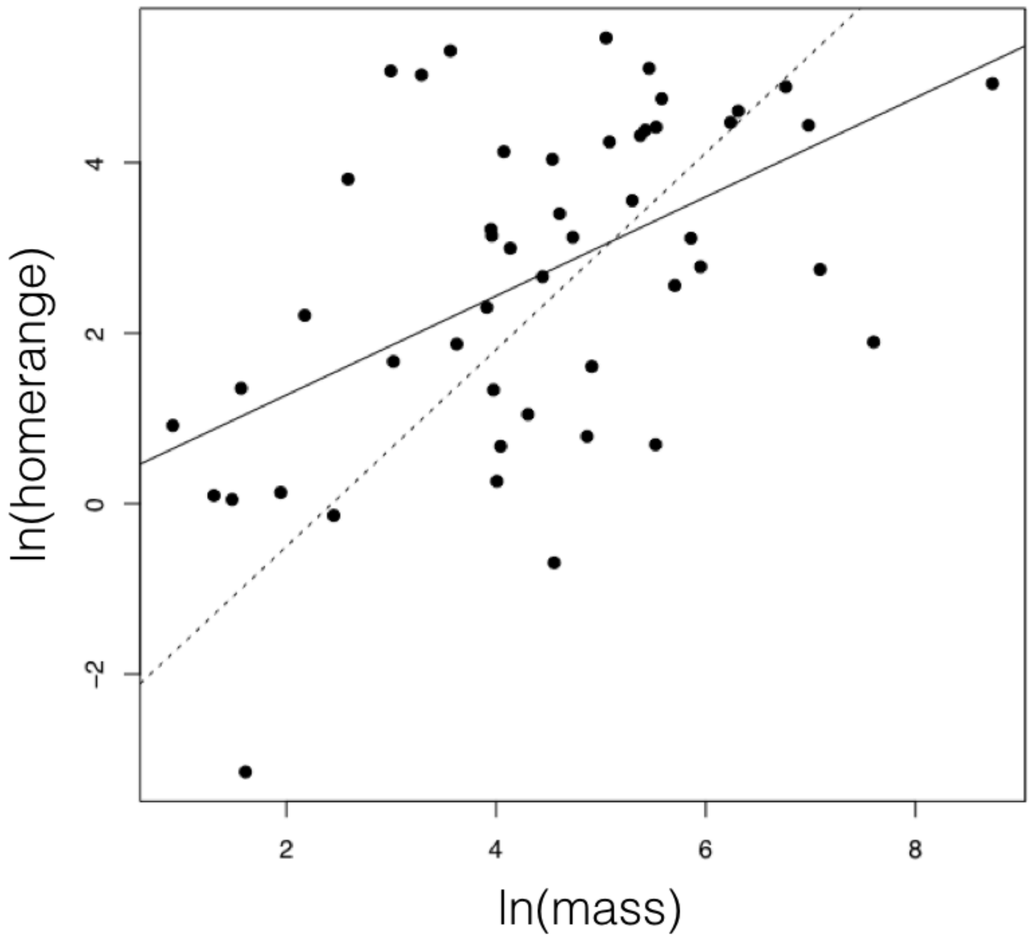
A statistical model is a type of model that includes a set of assumptions about the data-generating process.

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".

Example



The solid black line is a linear regression line.

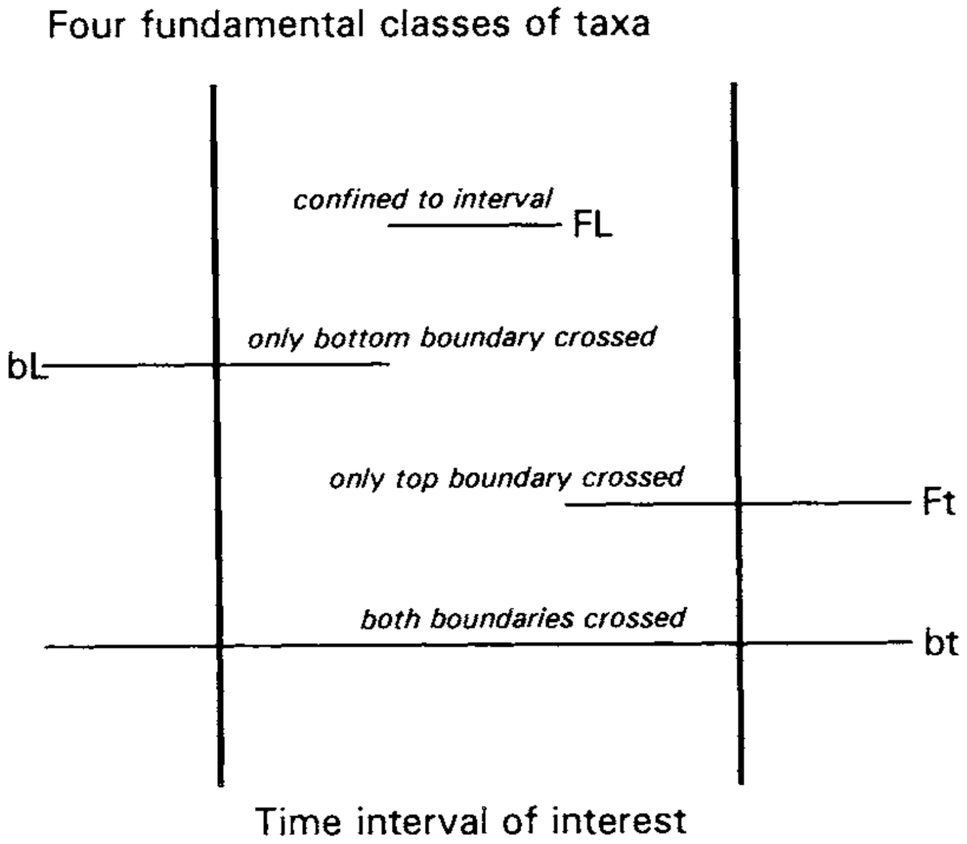
We can estimate the parameters of the regression model.

$$Y = X\beta + \varepsilon$$

It's also straightforward to simulate data under this model.

Image source Harmon (2019)

Example



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 they 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 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 of the models we use in statistical phylogenetics are mechanistic models, e.g. they might include origination, extinction and sampling parameters explicitly.

Note the definition of different model types varies a lot. The above is just my take on things from a very phylogenetics perspective.

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

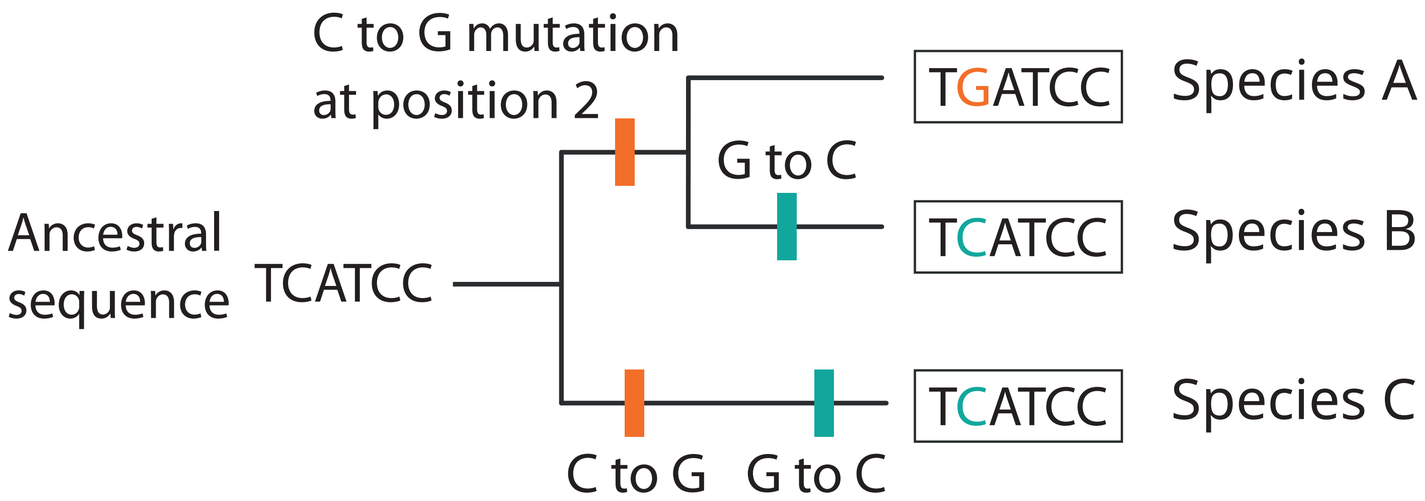
```
i = 1
while  i < 11:
    print(i)
    i = i + 1
```

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

Algorithms are used in phylogenetics for all sorts of tasks, inc. searching tree space or traversing trees.

Model-based phylogenetics

Models can account for the possibility that multiple changes occur at the same site.



Branch lengths = *expected number of changes per site*

In the absence of any information about time, rates are relative, i.e., rates are expected substitutions per site, independent of any time unit.

Model-based methods: advantages and disadvantages

Statistically more sound

Can test and update explicit assumptions

There are many more things we can do with models in palaeobiology!

Computationally slow (often)

Results are sensitive to model choice

Yang (2014) Molecular Evolution: A Statistical Approach

Phylogenetic data

Phylogenetic character data

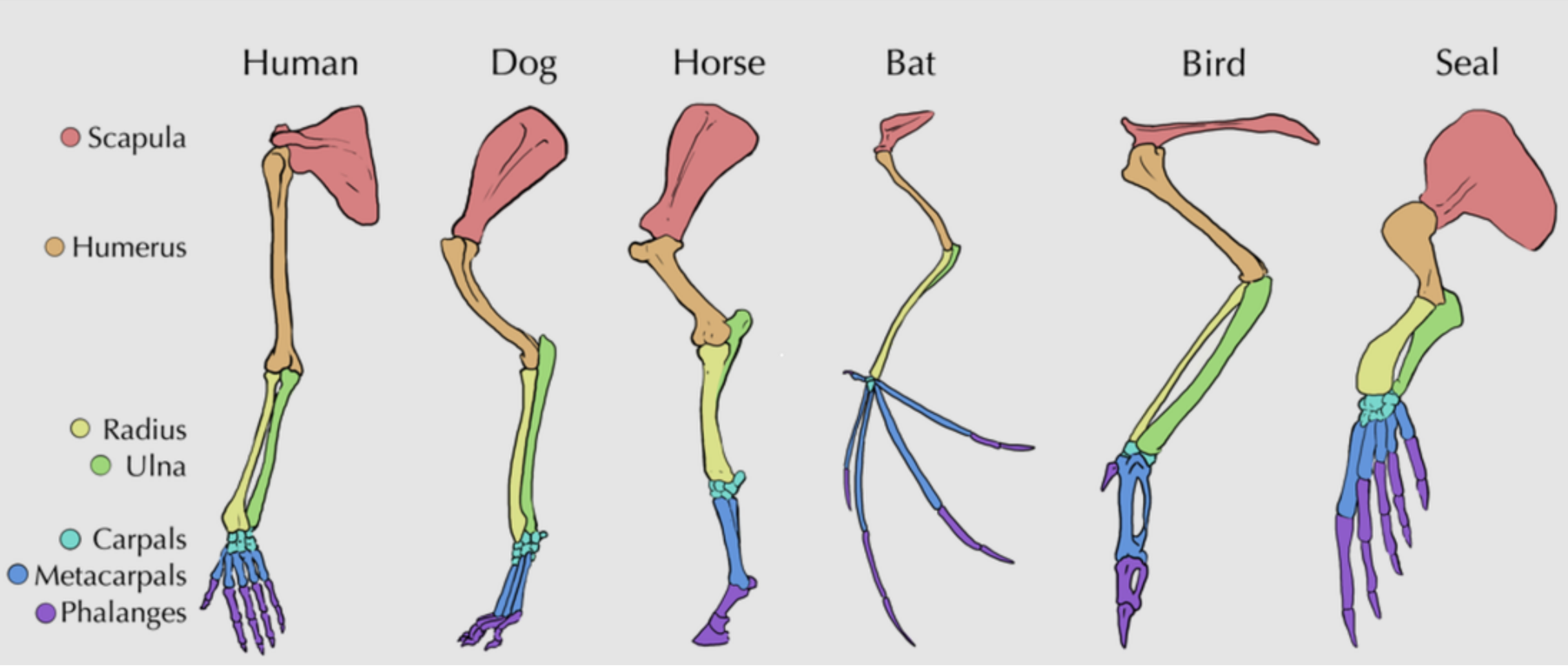
Two main sources of data for building trees:

1. Molecular sequences (nucleotides or proteins)

2. Morphological characters (discrete or continuous)

First we need to collect the data and establish homology.

Homology – similarity due to shared ancestry



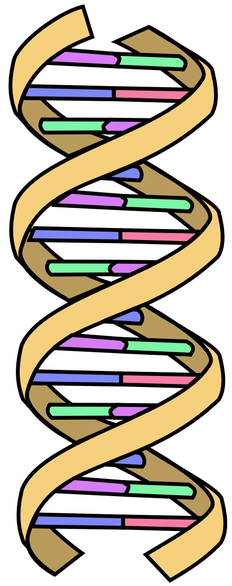
Each coloured bone is a homologous structure.

Molecular sequence data

Nucleotides provide a four letter alphabet we can use to generate trees.

Genes encode amino acids (proteins) that in turn provide a 20 letter alphabet.

Protein sequences are typically used for more distant evolutionary relationships.



DNA

= Adenine

= Thymine

= Cytosine

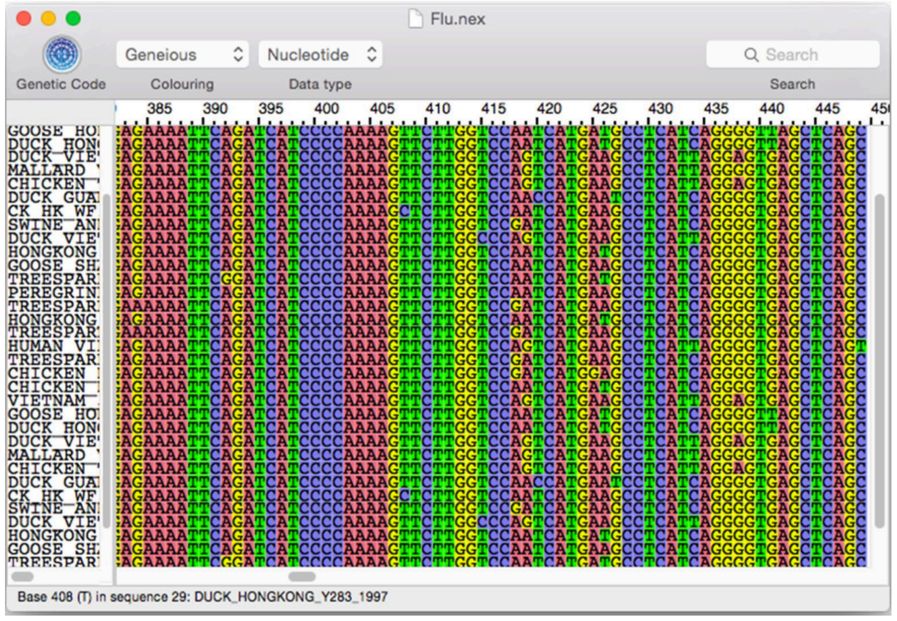
= Guanine

= Phosphate backbone

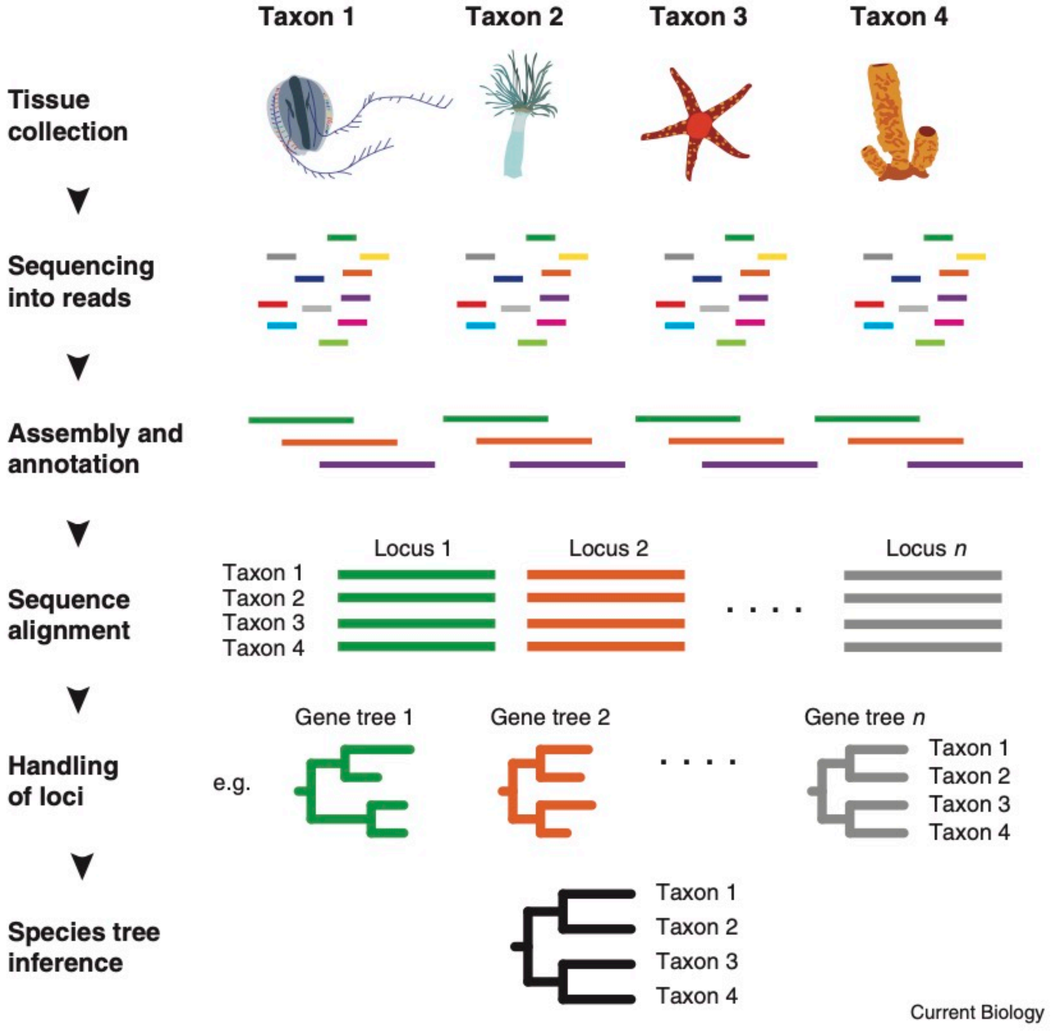
2nd codon position

1st codon position	U	C	A	G	3rd codon position
U	UUU Phe UUC UUA Leu UUG	UCU UCC UCA UCG Ser	UAU Tyr UAC **UAA Stop** **UAG Stop**	UGU Cys UGC **UGA Stop** UGG Trp	U C A G
C	CUU CUC CUA CUG Leu	CCU CCC CCA CCG Pro	CAU His CAC CAA Gln CAG	CGU CGC CGA CGG Arg	U C A G
A	AUU AUC AUA Ile AUG Met	ACU ACC ACA ACG Thr	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G
G	GUU GUC GUA GUG Val	GCU GCC GCA GCG Ala	GAU Asp GAC GAA Glu GAG	GGU GGC GGA GGG Gly	U C A G

Phylogenomics pipeline

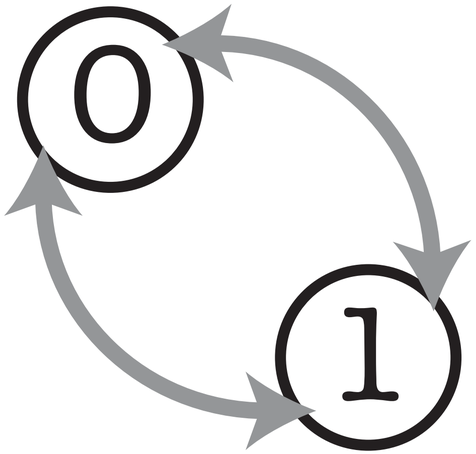


Multiple sequence alignments are the primary input for molecular phylogenetics



Duchêne (2021) Phylogenomics Primer

Models of character evolution



Also known as substitution / site / character models.

They capture the process of character evolution.

Allow us to ask, what is the probability of transitioning from one state to another over time?

What assumptions might you want to incorporate into a model of sequence evolution?

e.g., would all sites evolve at the same rate?

Models of nucleotide evolution: rate matrix

Using the substitution model we can calculate the probability of transitioning between different nucleotides. μ is the substitution rate.

$$Q = \begin{pmatrix} -\mu_A & \mu_{AG} & \mu_{AC} & \mu_{AT} \\ \mu_{GA} & -\mu_G & \mu_{GC} & \mu_{GT} \\ \mu_{CA} & \mu_{CG} & -\mu_C & \mu_{CT} \\ \mu_{TA} & \mu_{TG} & \mu_{TC} & -\mu_T \end{pmatrix}$$

→ We can calculate the the probability of changing between two states over a given branch lengths.

The longer the interval of time has past, the more likely we are to observe a change.

You can explore this principle via this app by Paul Lewis.

The Jukes-Cantor model of sequence evolution

The simplest model of sequence evolution.

Assumptions: equal mutation rates and equal base frequencies.

Base frequencies are the proportion of each nucleotide within the dataset.

$$Q = \begin{pmatrix} * & \mu & \mu & \mu \\ \mu & * & \mu & \mu \\ \mu & \mu & * & \mu \\ \mu & \mu & \mu & * \end{pmatrix}$$

The GTR model of sequence evolution

Nucleotides (ATCG) occur at different frequencies depending on the group of species or gene.

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

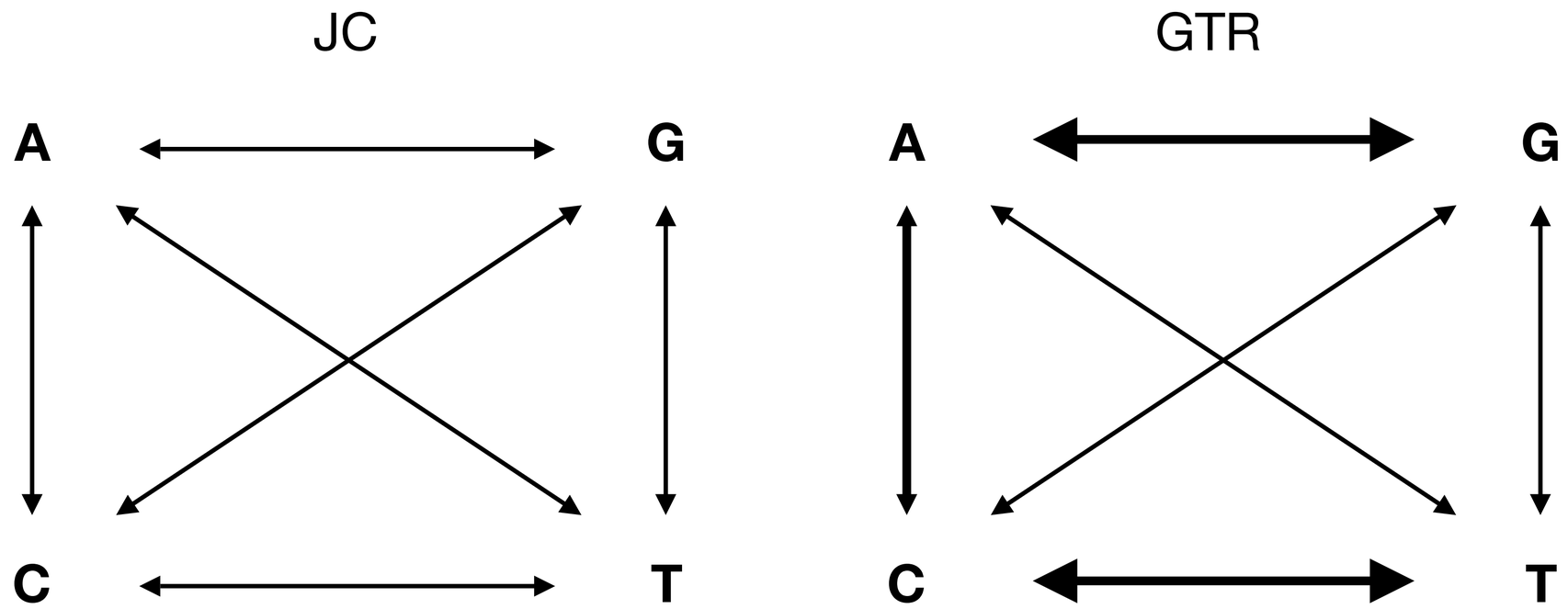
GTR assumptions: unequal mutation rates AND unequal base frequencies.

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

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

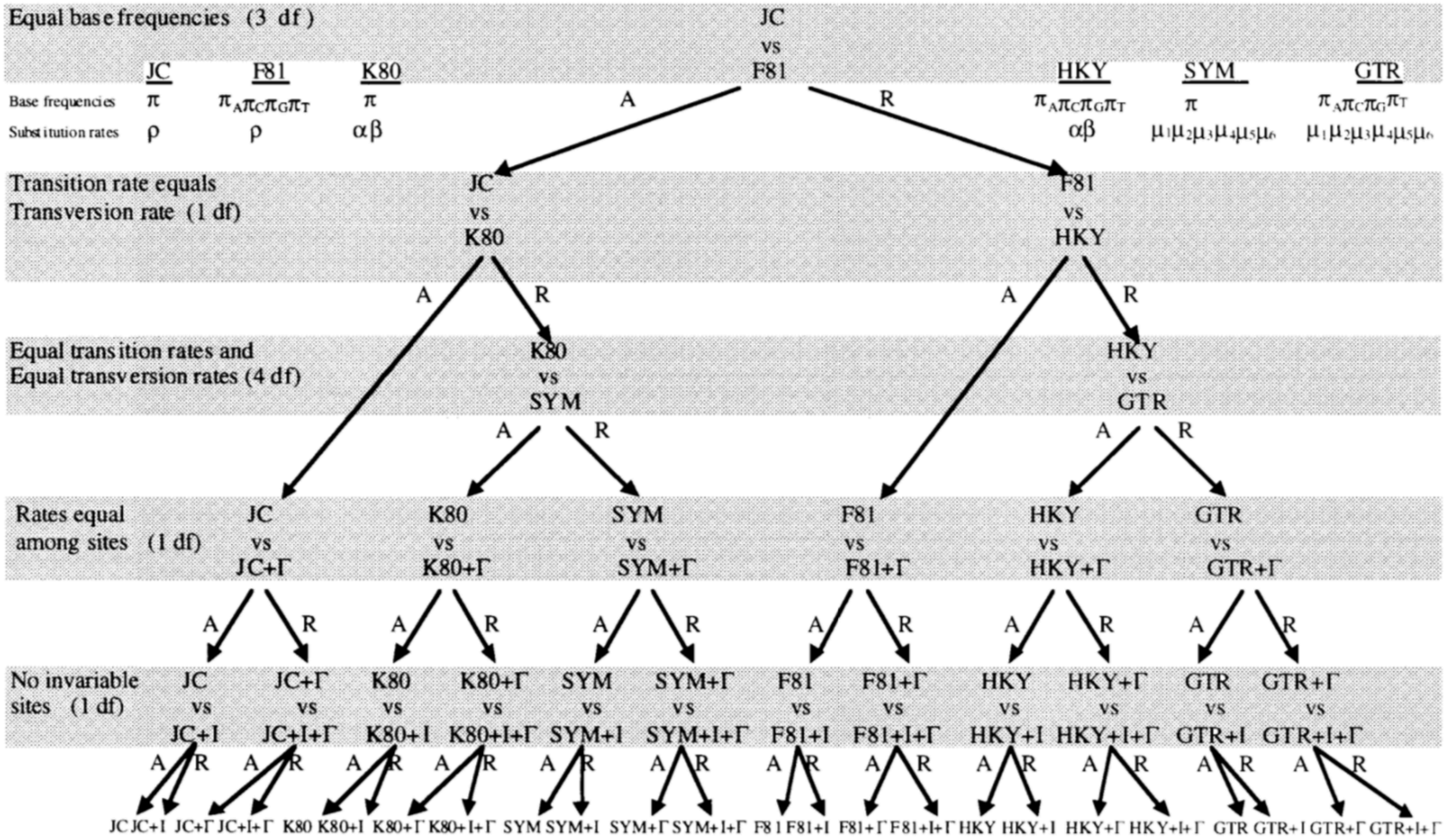
The JC versus GTR models

Another way of visualising substitution models.



Line width represents the relative rate of change between different steps.

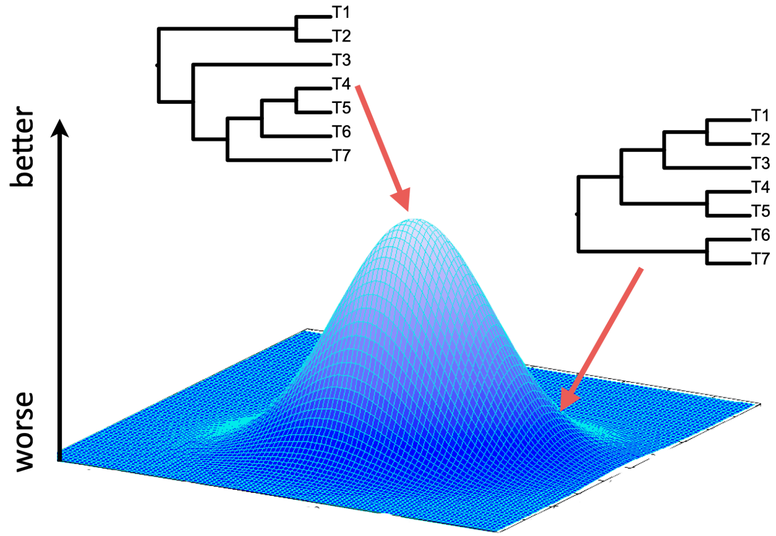
JC & GTR belong to a large family of substitution models



Posada & Crandall (1998) Bioinformatics

A very brief introduction to maximum likelihood

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



Model based phylogenetics

Assume an explicit model of character evolution.

Maximum likelihood is a method for estimating unknown parameters in a model. The tree that maximises the likelihood is the best one.

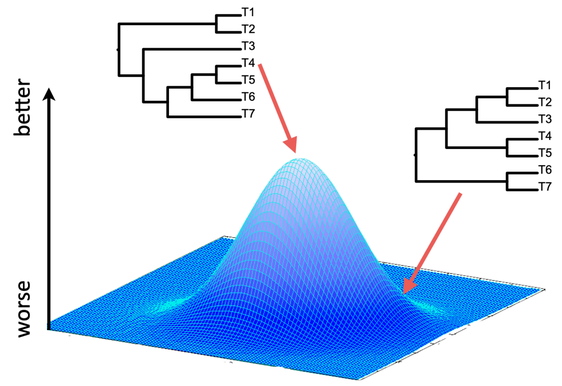
P(data | model, tree)

Maximum likelihood algorithm simplified

1. We first propose a topology with branch lengths and then calculate the likelihood (taking into account all sites).

2. We then propose a new tree or set of branch lengths and recalculate the likelihood. If the likelihood is >, we accept this tree as being better.

3. Proceed until we can't improve the likelihood any further.

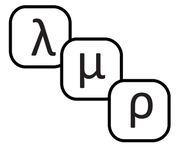


Exercise 2: intro to phylogenetics using R

Introduction to graphical models and RevBayes

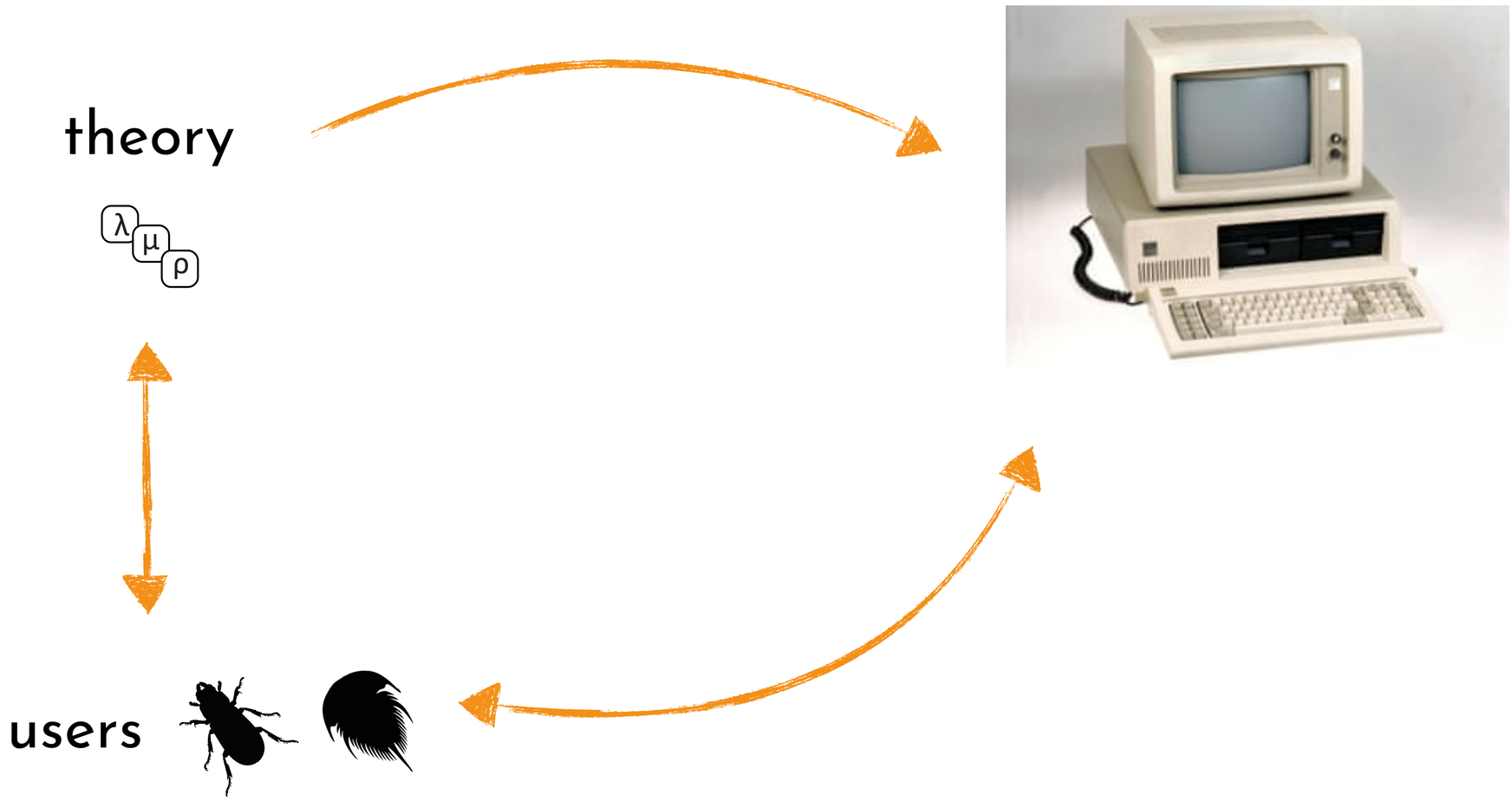
Phylogenetic inference – the old way

theory



users

Is there a better way?



Aims for RevBayes

Flexible model specification

– Availability of (common) models
– Extendability

Easy to learn

– Well structured model specification
– Explicit models
– Documentation, examples and tutorials

Computational efficiency

– Fast likelihood calculators
– Efficient (MCMC) algorithms

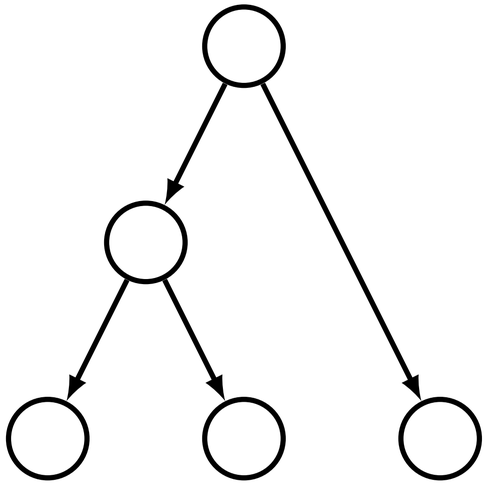
There's a huge team behind the scenes.

Höhna et al. 2016. Sys Bio

RevBayes uses a graphical model framework

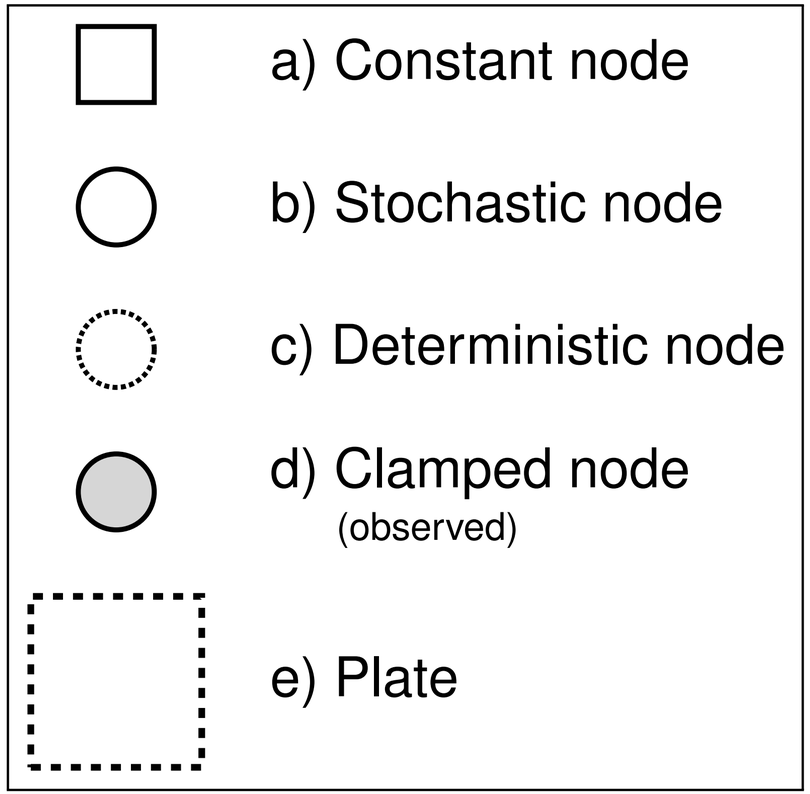
Graphical models provide tools for visually and computationally representing complex, parameter-rich probabilistic models.

We can depict the conditional dependence structure of various parameters and other random variables.



Höhna et al. 2014. Sys Bio

Graphical models – types of variables (nodes)



a) fixed-value variables

b) random variables that depend on other variables

c) variables determined by a specific function applied to another variable (transformations)

d) observed stochastic variables (data)

e) replication over a set of variables

Specifying graphical models using the Rev syntax

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

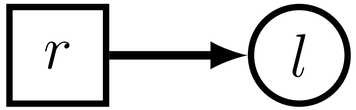
Operator	**Variable**
`<-`	constant variable
`~`	stochastic variable
`:=`	deterministic variable
`node.clamp(data)`	clamped variable
`=`	inference (*i.e.*, non-model) variable
`for(i in 1:N){...}`	plate

a)


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

a)

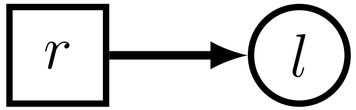
b)



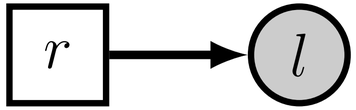

```
# constant node
r <- 10


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

a)

b)



c)

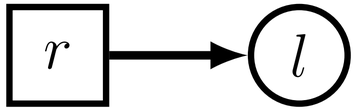



```
# constant node
r <- 10


# stochastic node
l ~ dnExp(r)


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

a)

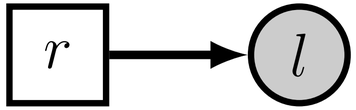

```
# constant node
r <- 10
```

b)



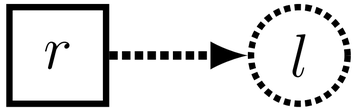

```
# stochastic node
l ~ dnExp(r)
```

c)




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

d)

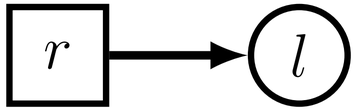



```
# deterministic node
l := exp(r)
```

a)

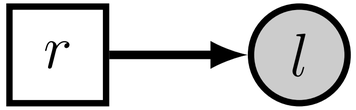

```
# constant node
r <- 10
```

b)



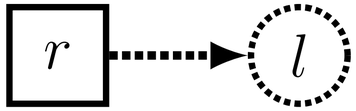

```
# stochastic node
l ~ dnExp(r)
```

c)



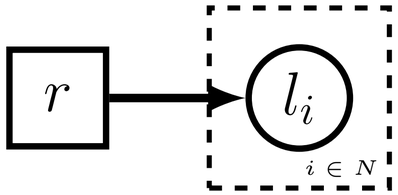

```
# stochastic node (observed)
l.clamp(0.1)
```

d)




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

e)

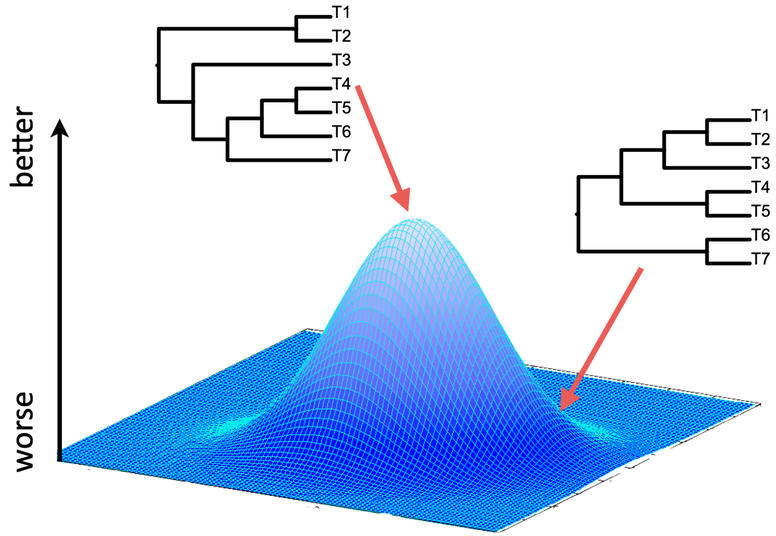



```
# stochastic nodes (iid)
for (i in 1:N) {
    l[i] ~ dnExp(r)
}
```

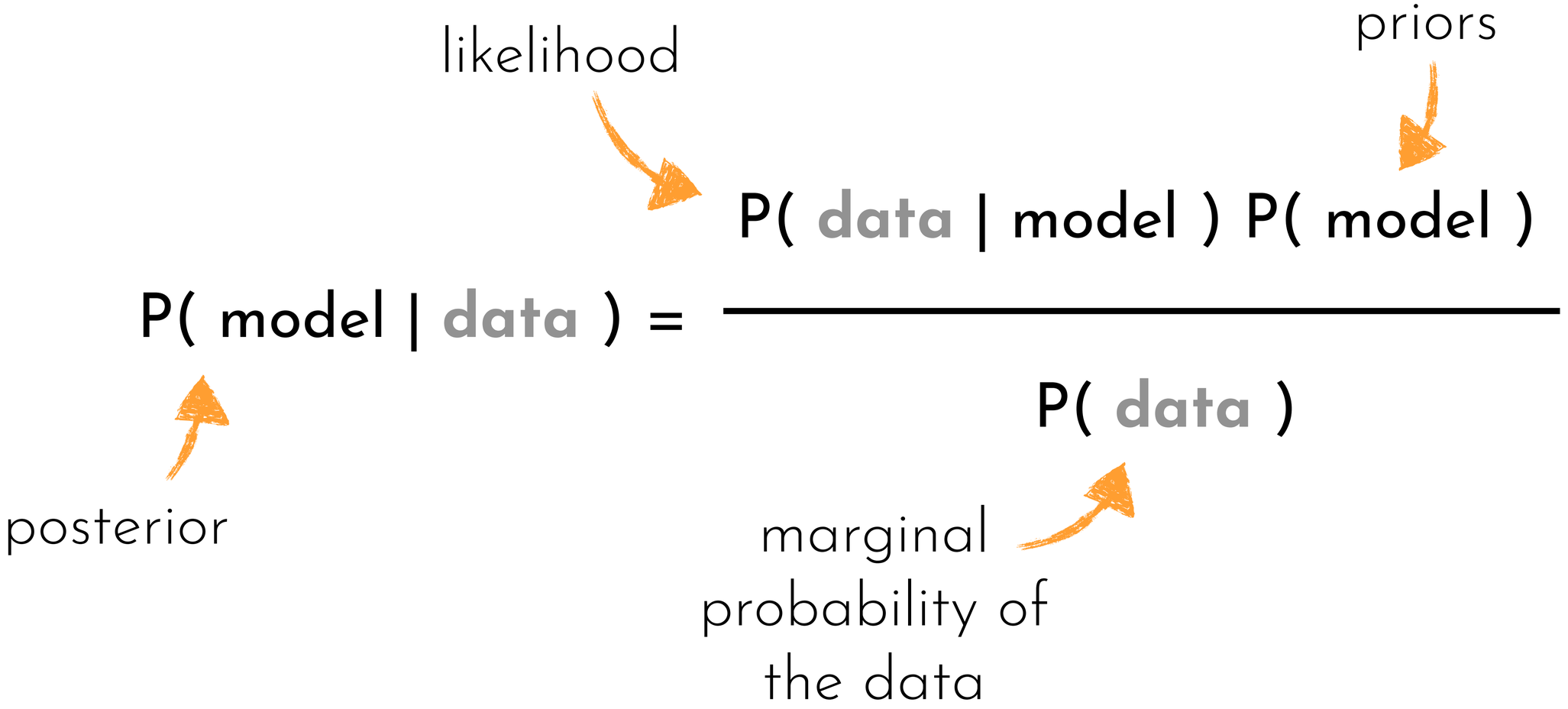
Exercise 3: intro to the Rev language

Introduction to Bayesian inference and MCMC

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



Bayes' theorem



Bayes' theorem

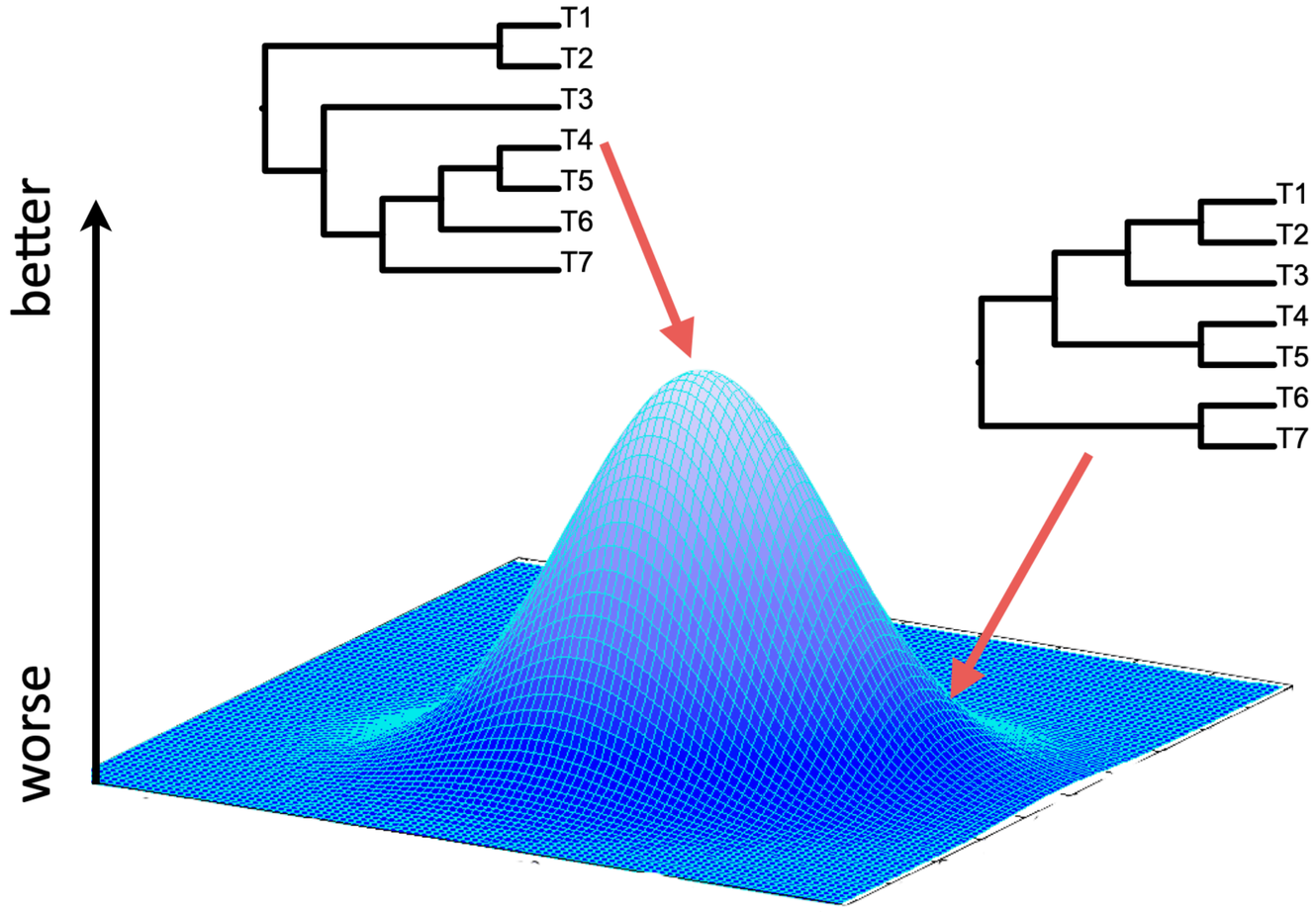
P (**data** | parameters, model) ← the model used to calculate the likelihood.

P (parameters | model) ← this represents our prior knowledge of the model parameters.

P (**data** | model) ← the probability of the data integrated over all possible parameter values. Can be thought of as a normalising constant (i.e., ensuring the posterior sums to one).

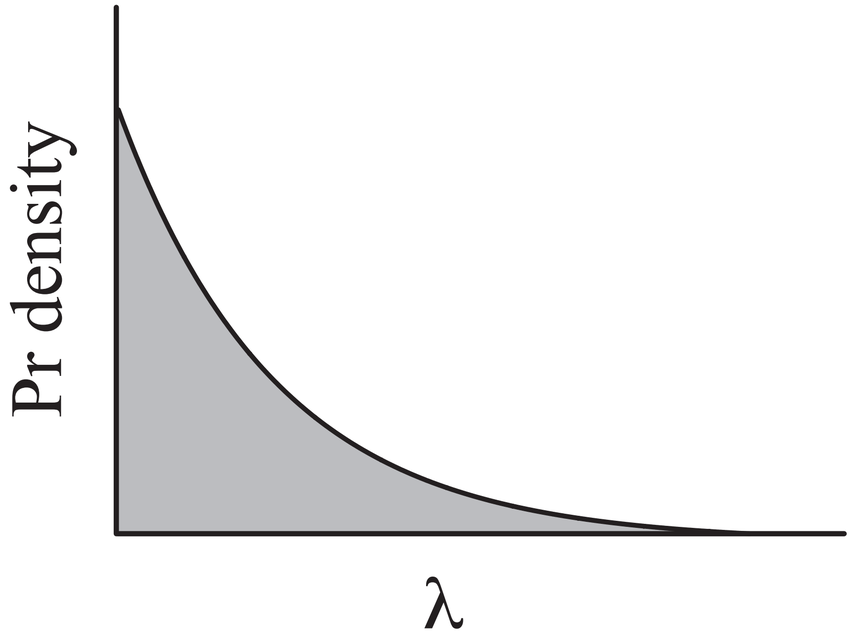
P (parameters | **data**, model) ← the posterior reflects our combined knowledge based on the likelihood and the priors.

The output of a Bayesian phylogenetic analysis is a distribution of trees (+ any other estimated parameters)



Probabilities vs probability densities

In phylogenetics, probabilities are not normally discrete (i.e., represented by a single value) and we're often dealing with a lot of uncertainty (esp. in the fossil record). Instead we typically work with **probability densities**.

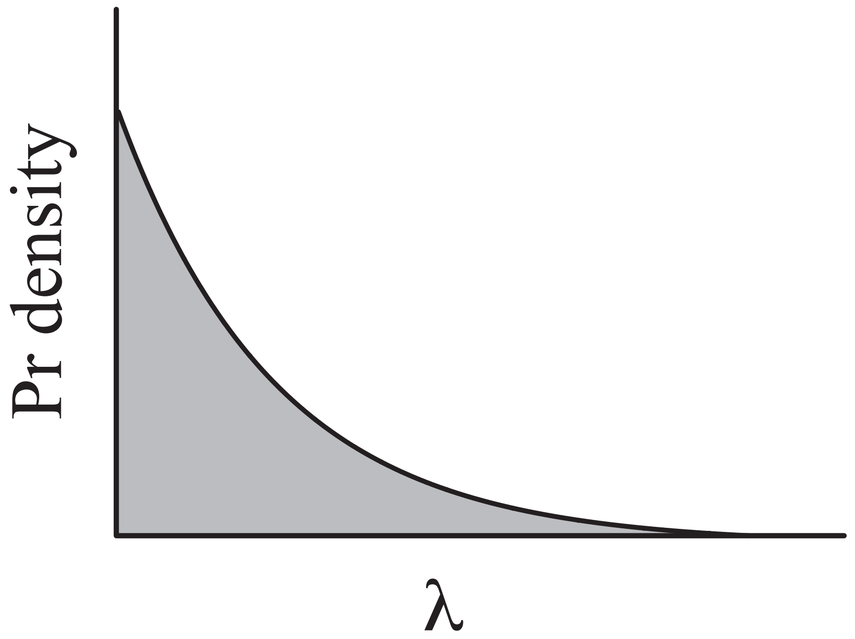


Probability densities

The x-axis represents the value of our parameter λ.

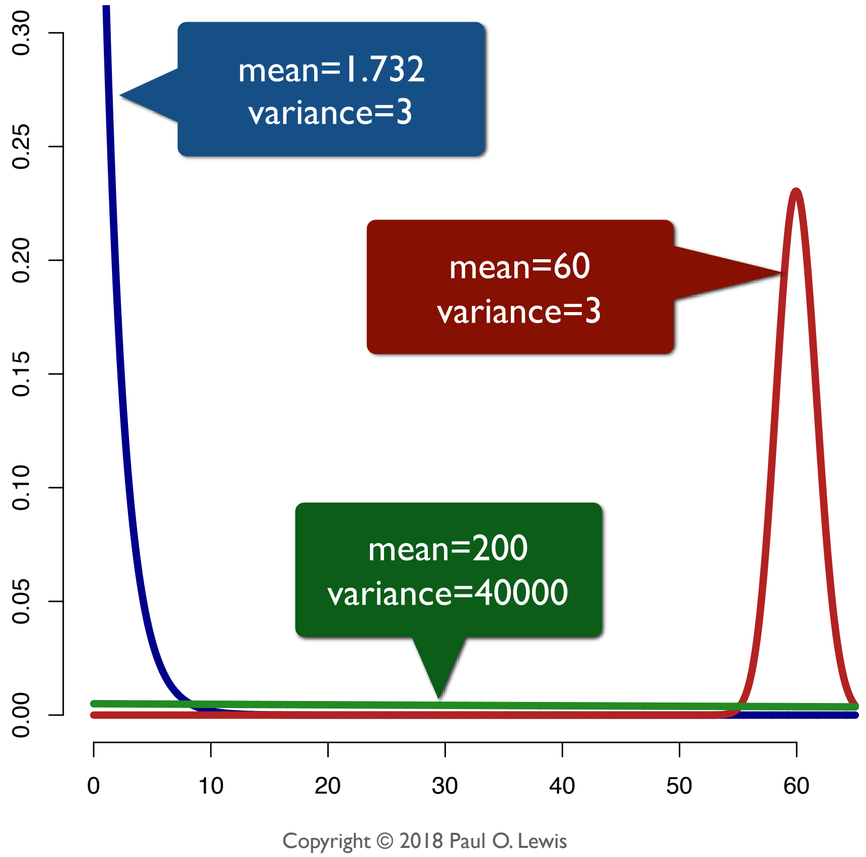
The y-axis is relative probability.

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

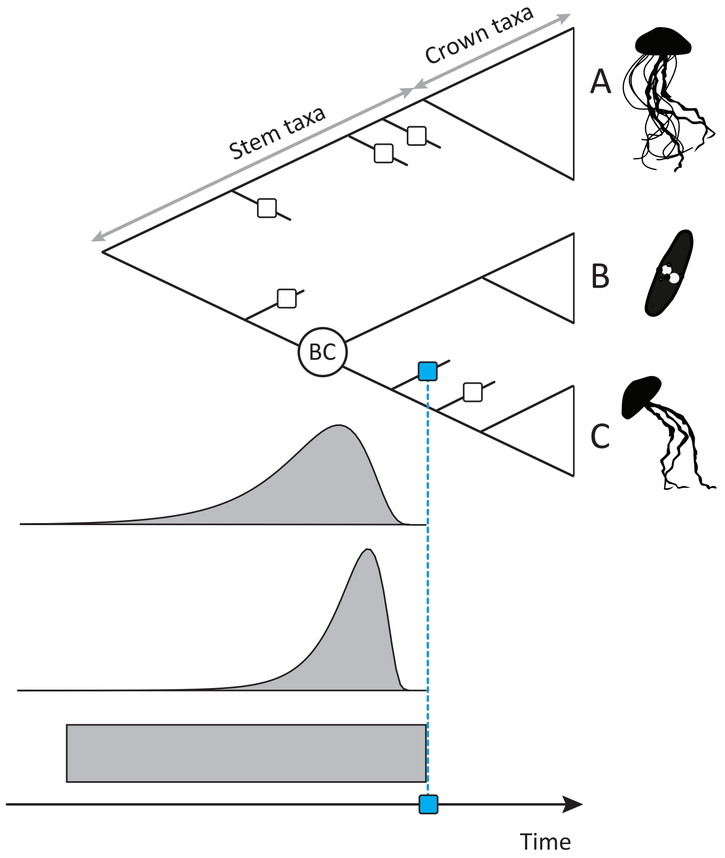


λ. is drawn from an exponential distribution

Probability densities



λ. drawn from an gamma distribution



Node age (calibration) densities

Why do we need Markov chain Monte Carlo?

Probability densities already introduce some complexity. Remember the posterior is not usually a point estimate (i.e., a single value) but a range of values.

The marginal probability of the data is also very tricky to calculate.

P (**data** | model)

Calculating this requires taking into account all possible alternative parameter combinations (e.g., all possible trees).

This makes it challenging to calculate the posterior analytically (i.e., exactly).

What is Markov chain Monte Carlo (MCMC)?

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

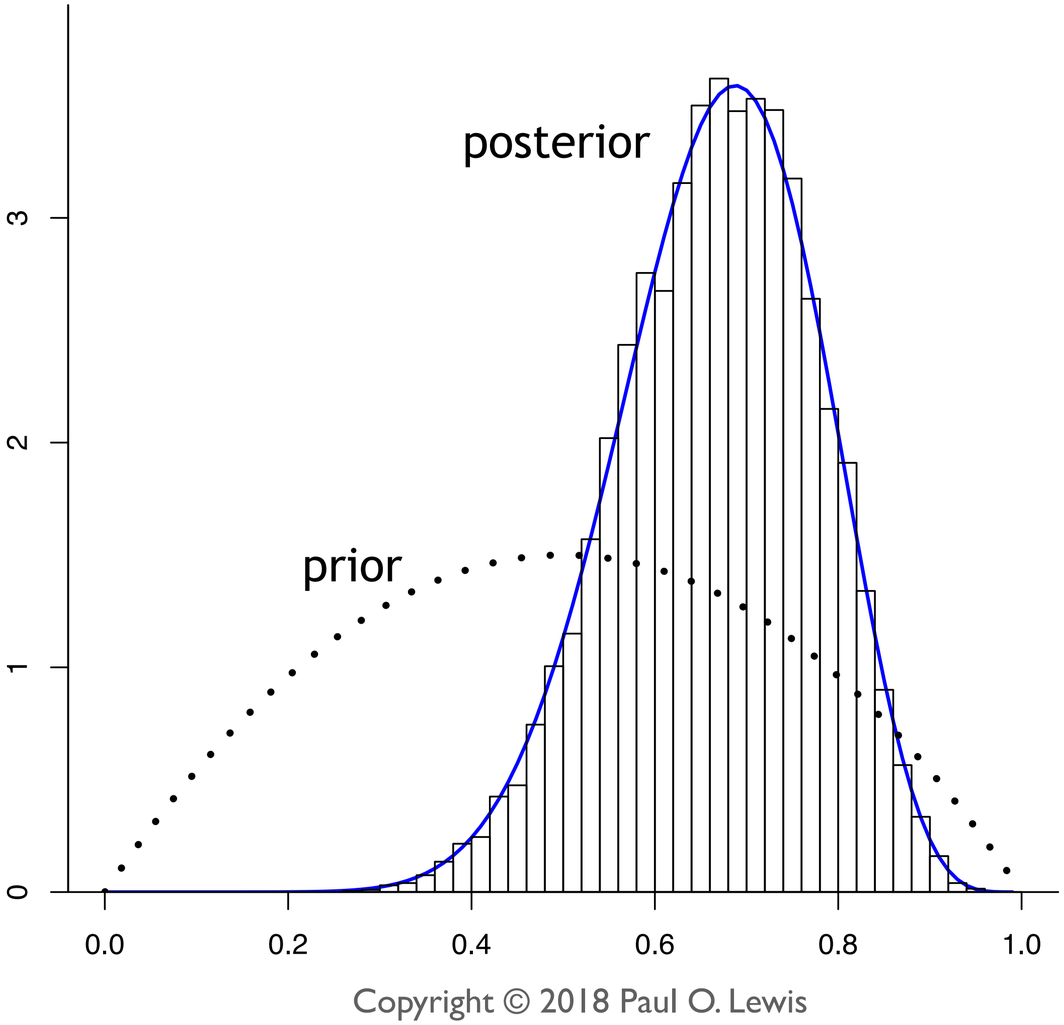
Markov chain means the progress of the algorithm doesn't depend on its past.

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.

Click here for a little bit of history.

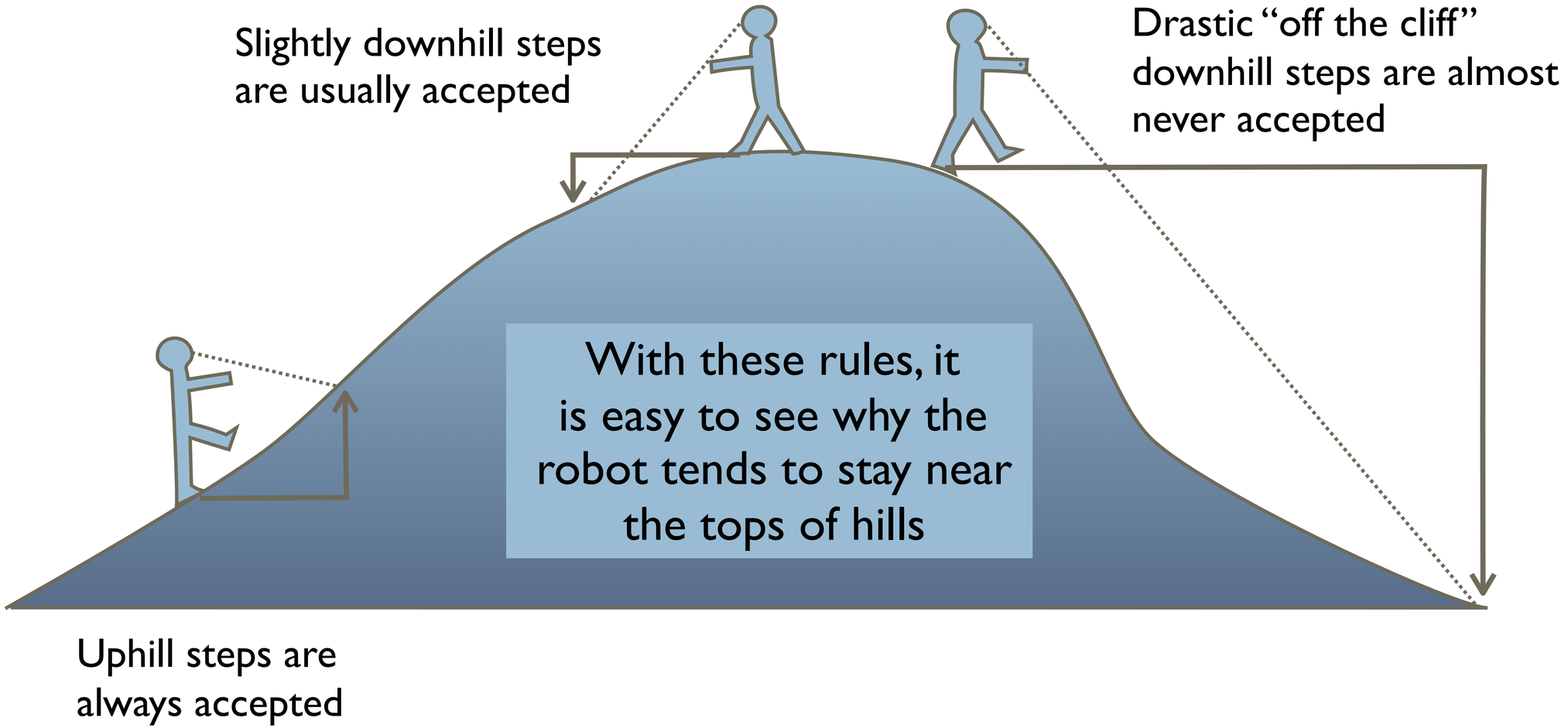
What is Markov chain Monte Carlo (MCMC)?



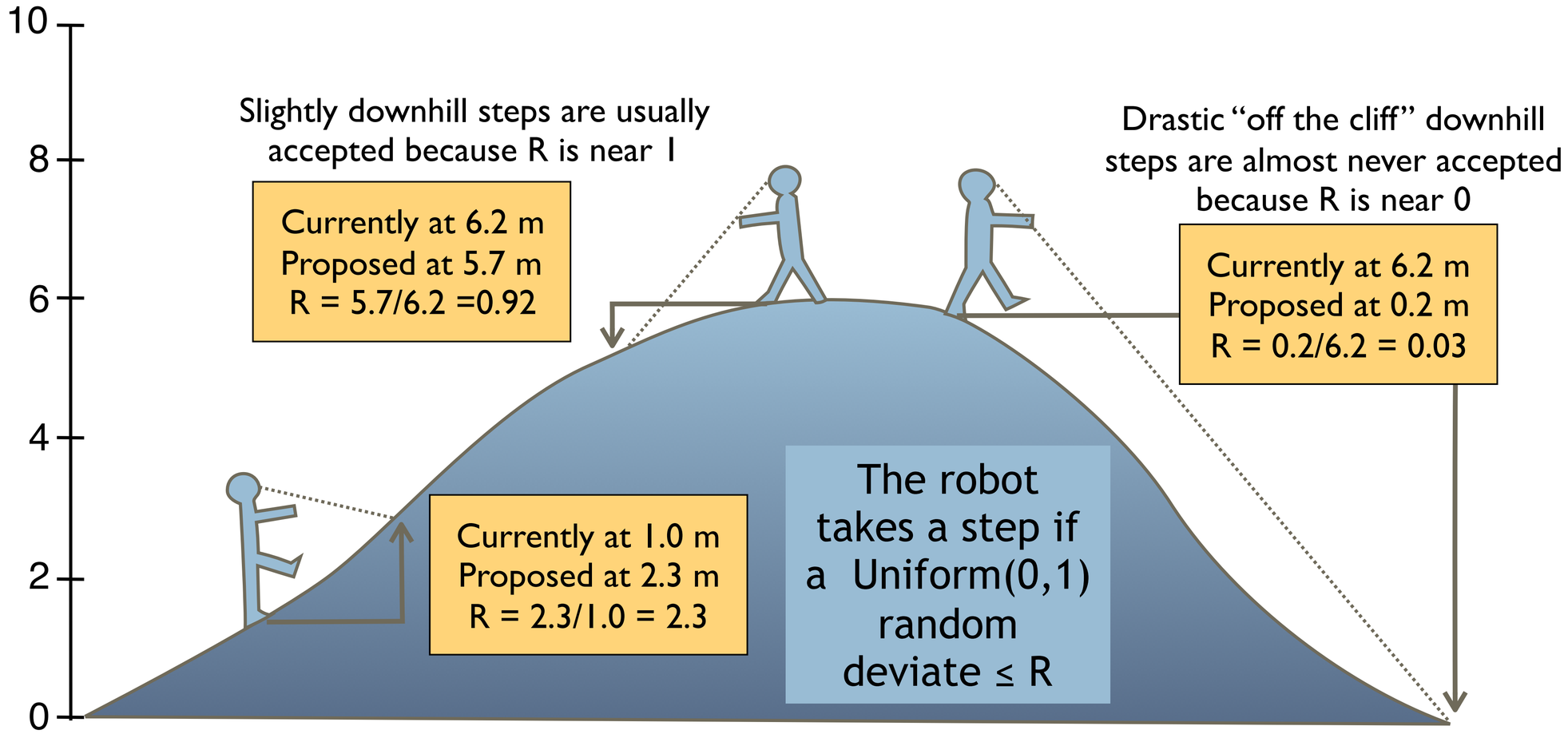
The aim is to produce a **histogram** that provides a good approximation of the posterior.

The most widely used MCMC algorithm in phylogenetics is the Metropolis Hastings algorithm.

MCMC robot's rules



Actual rules (Metropolis algorithm)



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

The marginal likelihood is cancelled out

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

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

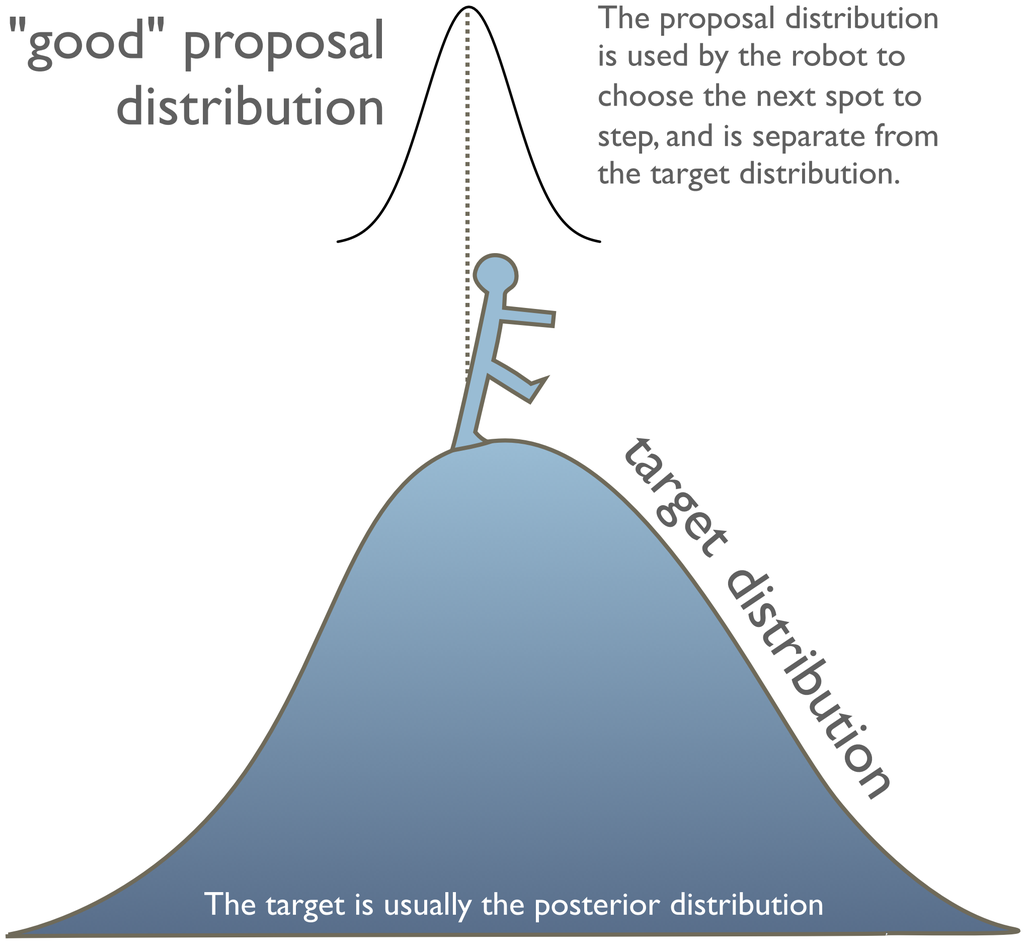
Posterior odds

Apply Bayes' rule to both top and bottom

Likelihood ratio

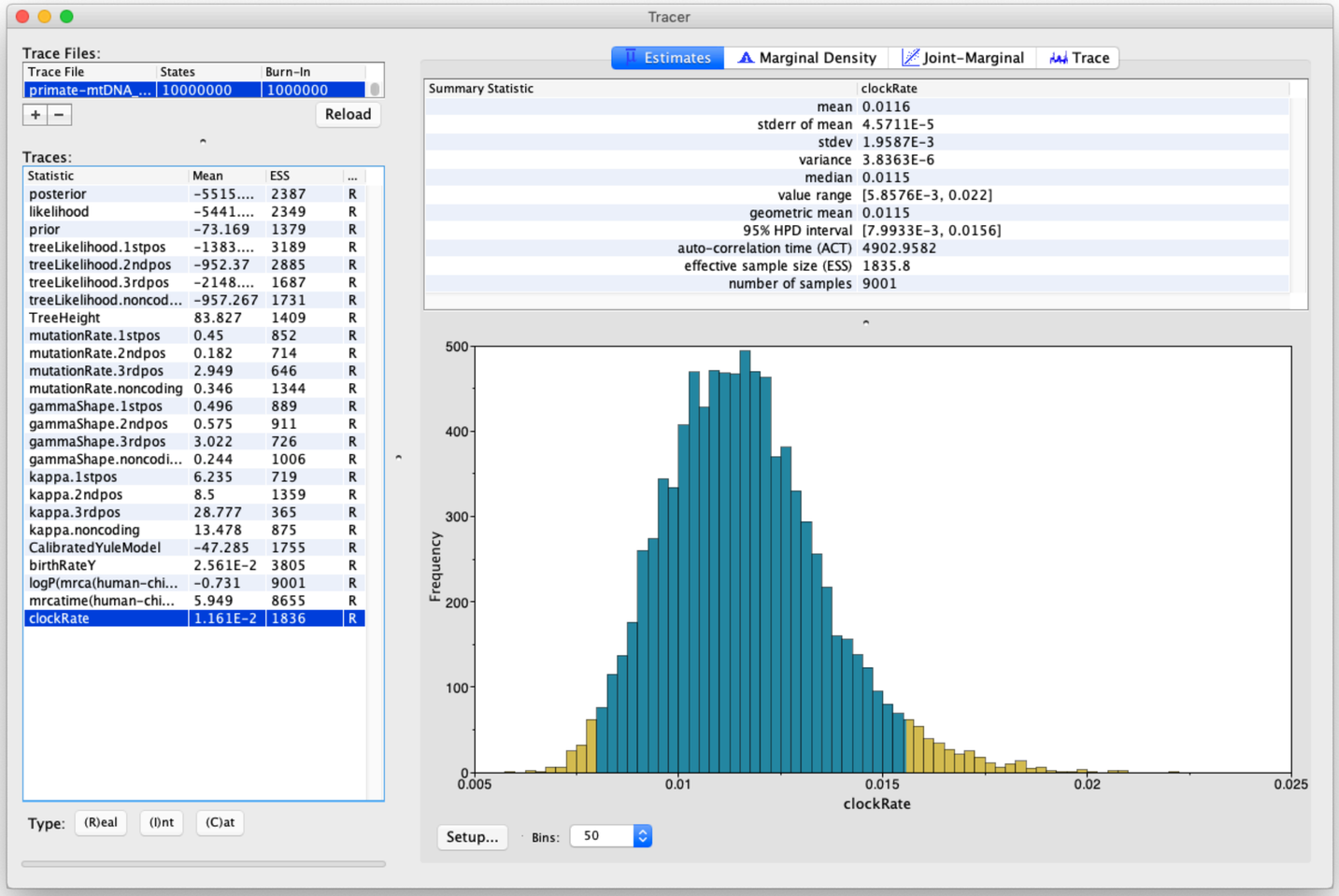
Prior odds

MCMC proposals, steps or moves



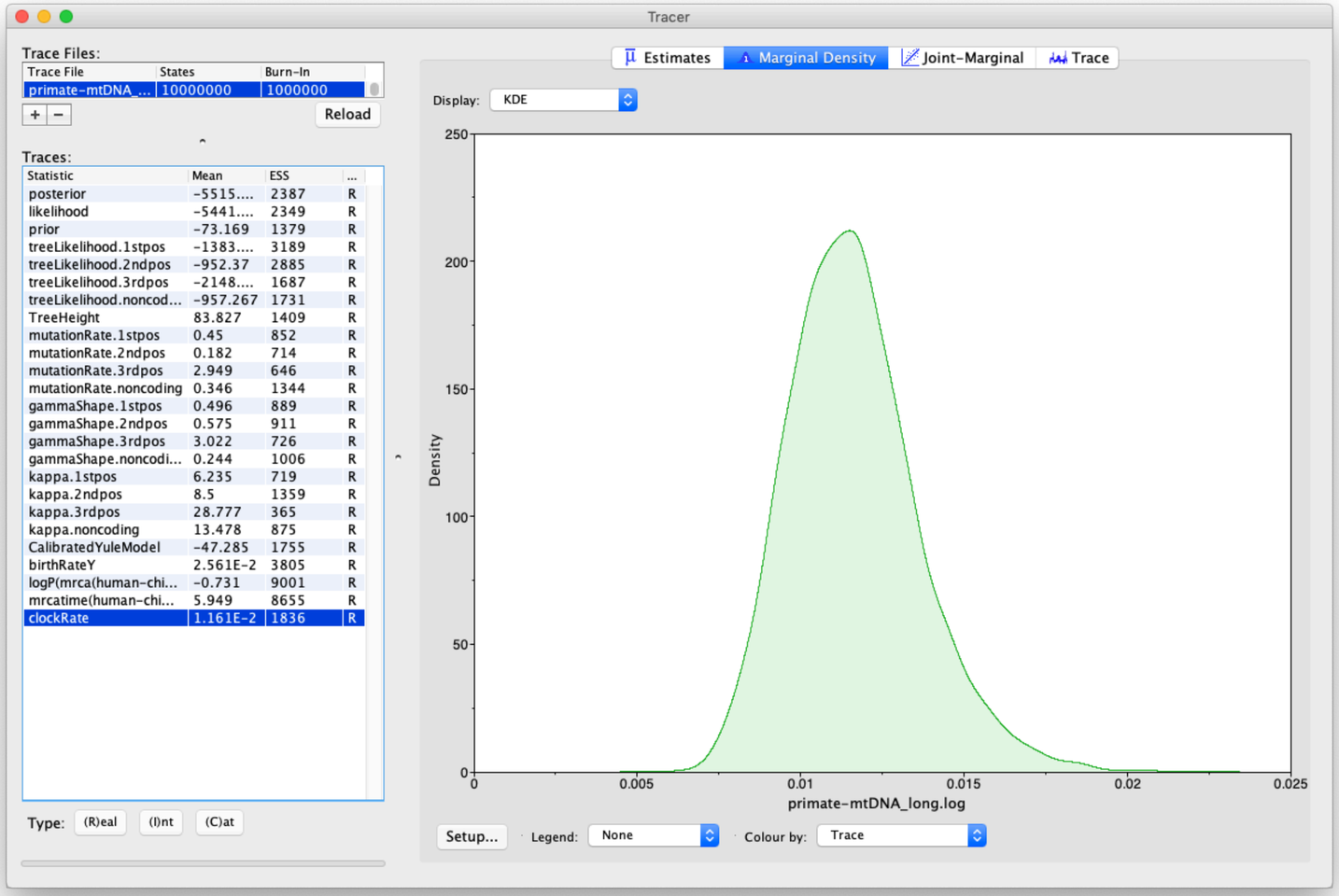
Summarising the posterior

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 sometimes be misleading.

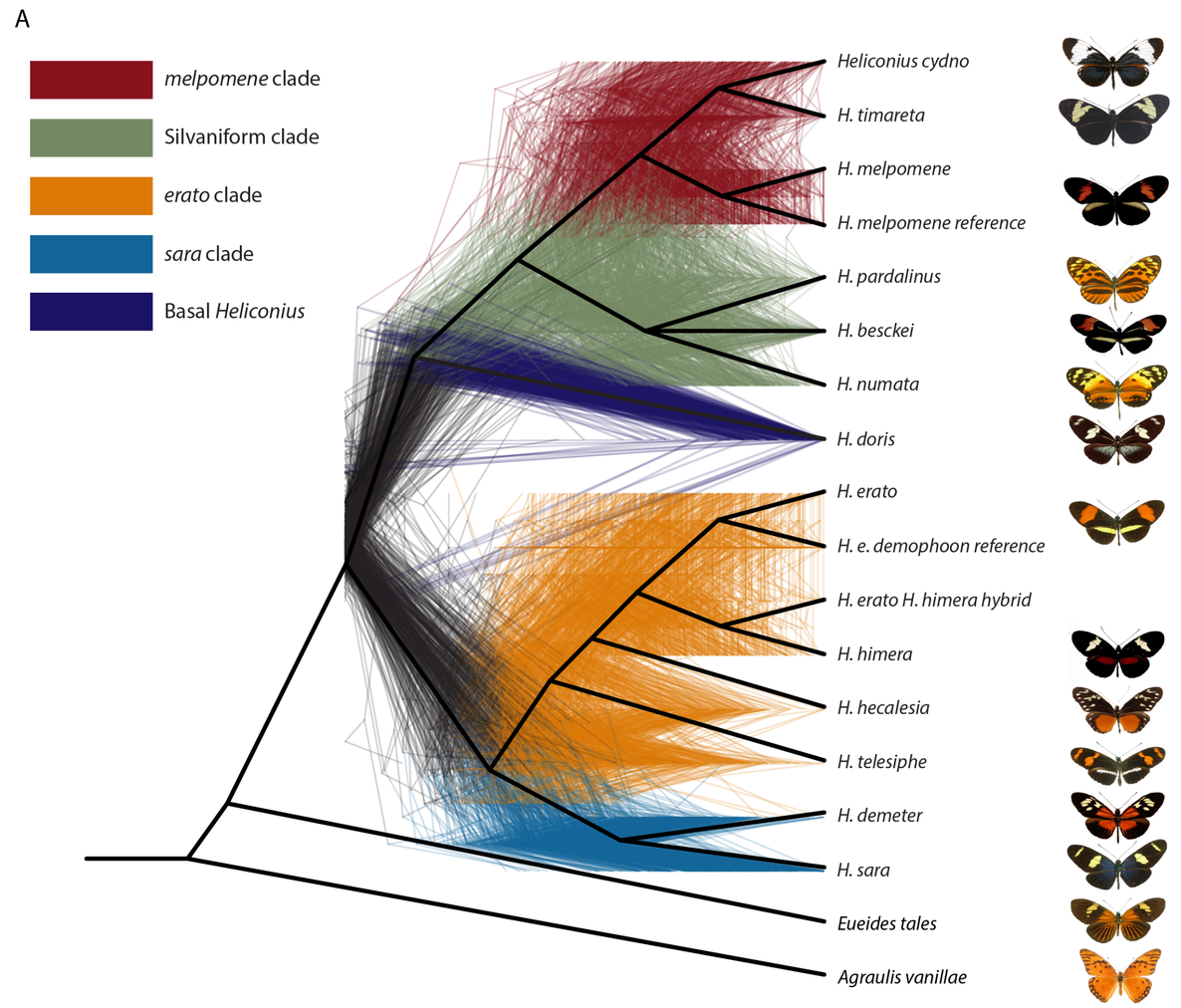


Image source Edelman et al. (2019) Science

Summarising the posterior

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.

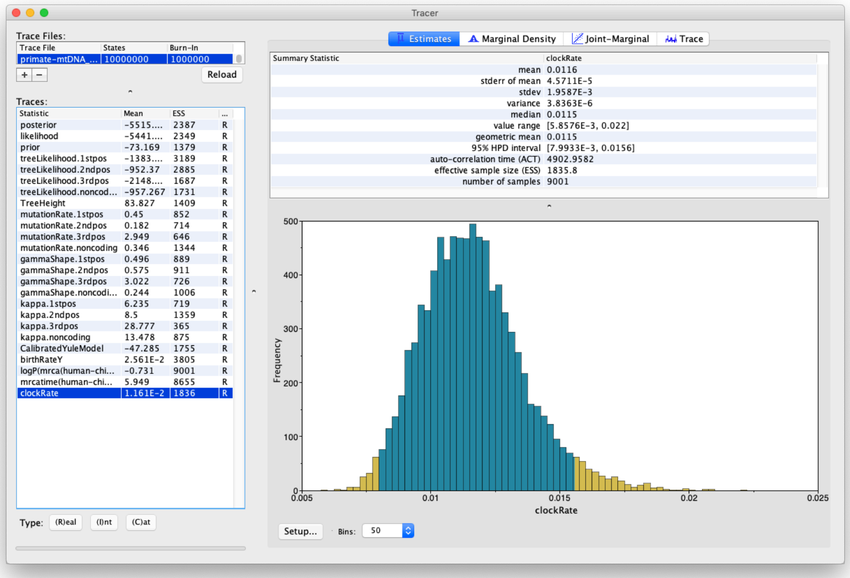
Maximum clade credibility (MCC) tree: the tree in the posterior sample that has the highest posterior probability (i.e. clade support) across all nodes.

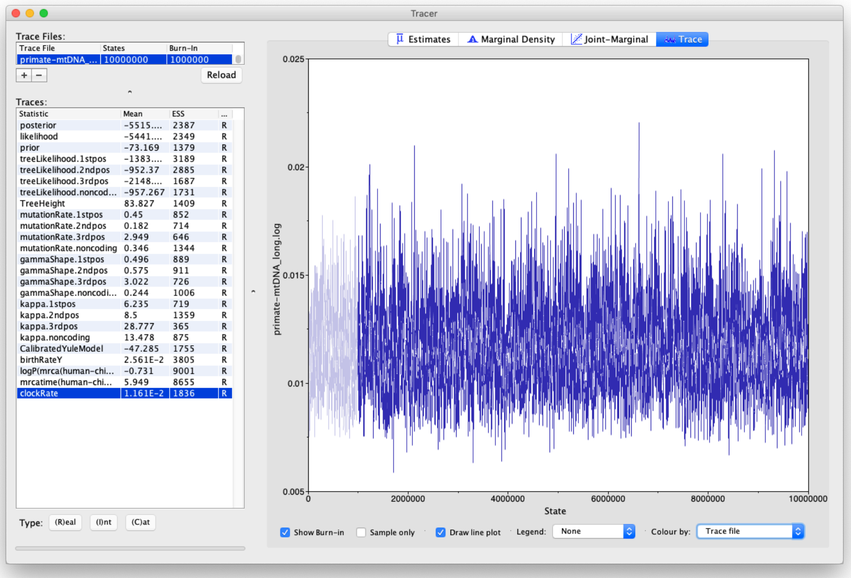
For more on issues associated with summary tree methods see O'Reilly & Donoghue (2018) Sys Bio.

Convergence

How do you know if you've run the run the chain long enough?

You don't! But there are some clues.



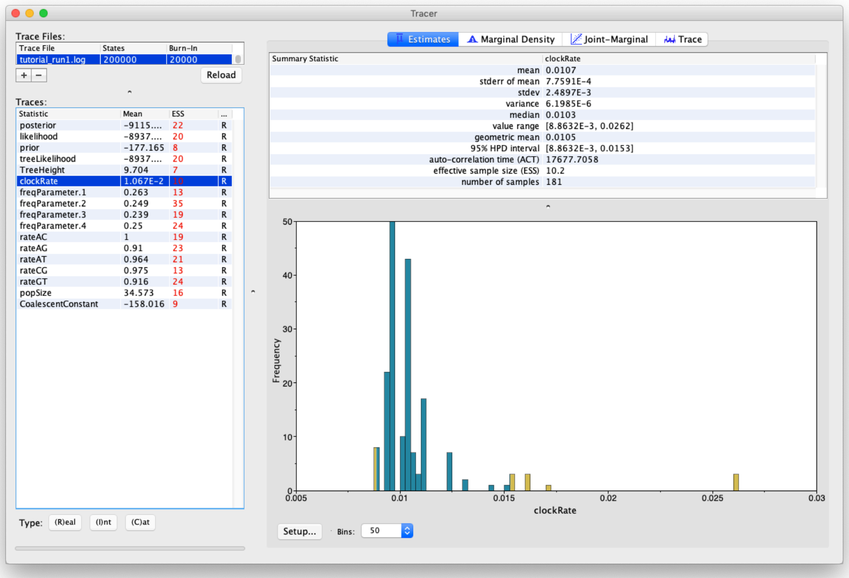


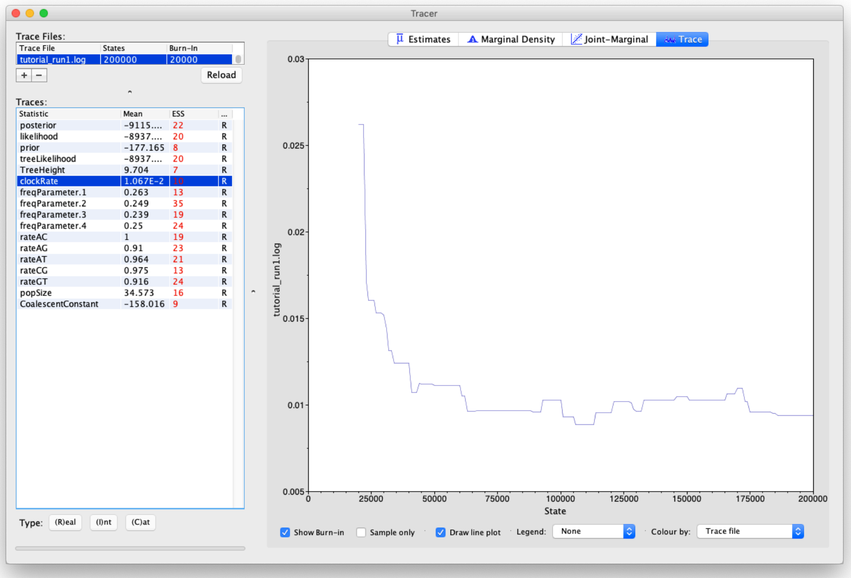
Good mixing

Convergence

How do you know if you've run the run the chain long enough?

You don't! But there are some clues.



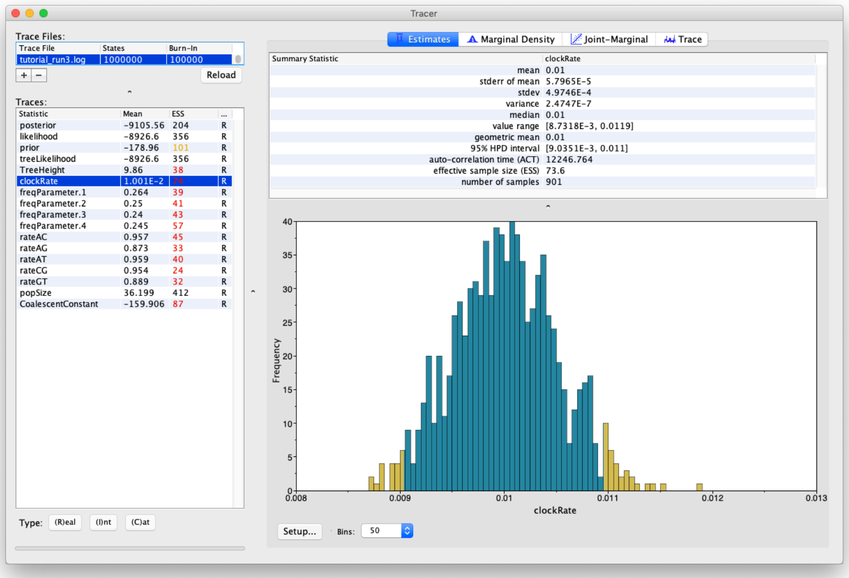


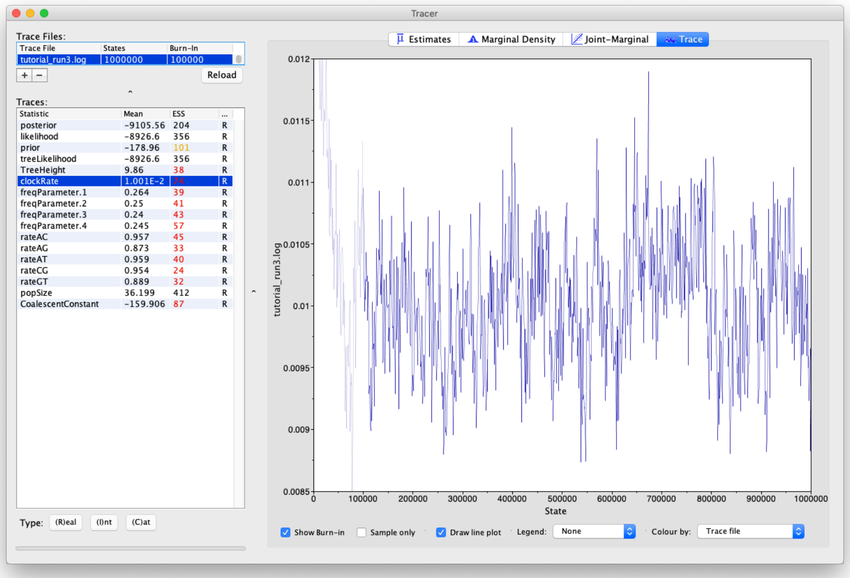
Bad mixing

Convergence

How do you know if you've run the run the chain long enough?

You don't! But there are some clues.





Better mixing

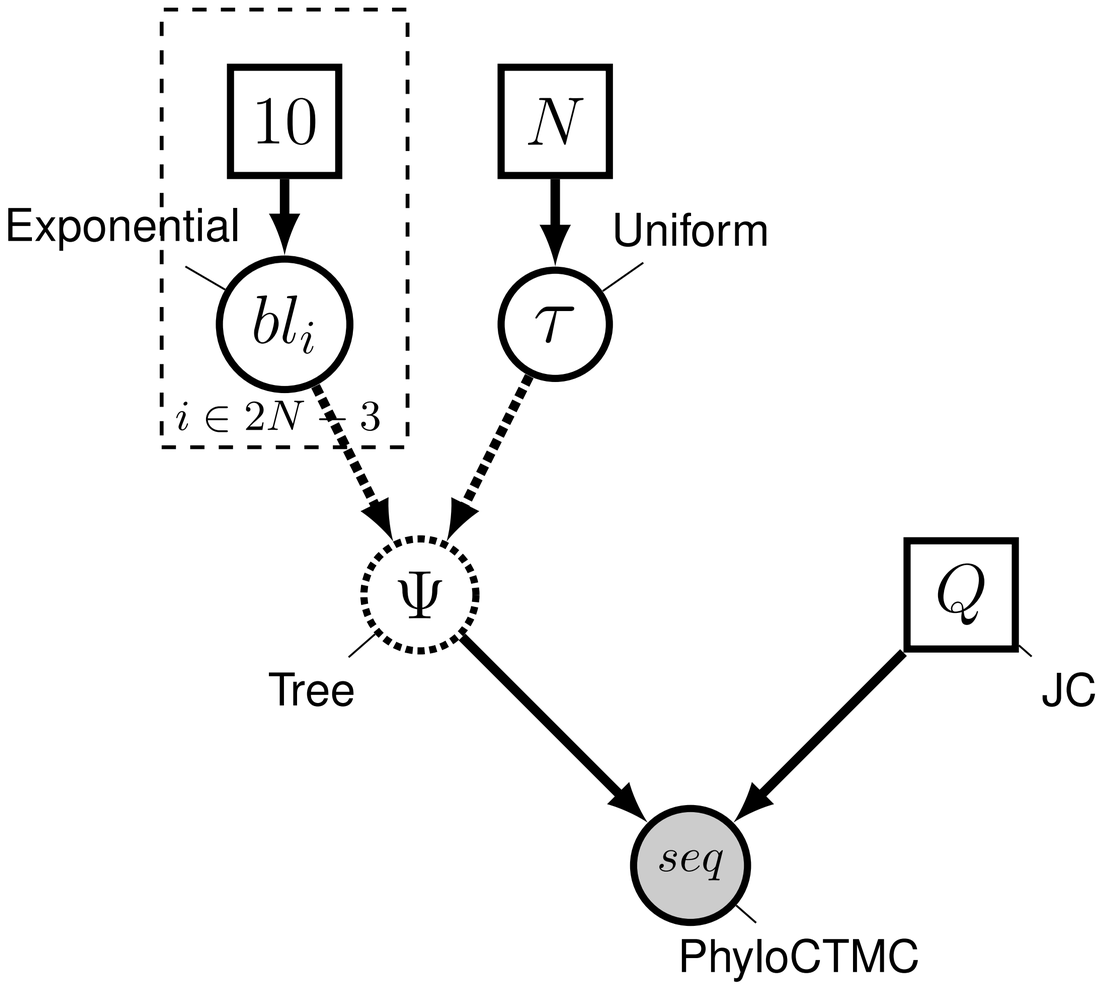
Summary from this part

Bayesian inference provides a flexible and intuitive way to incorporate and represent uncertainty.

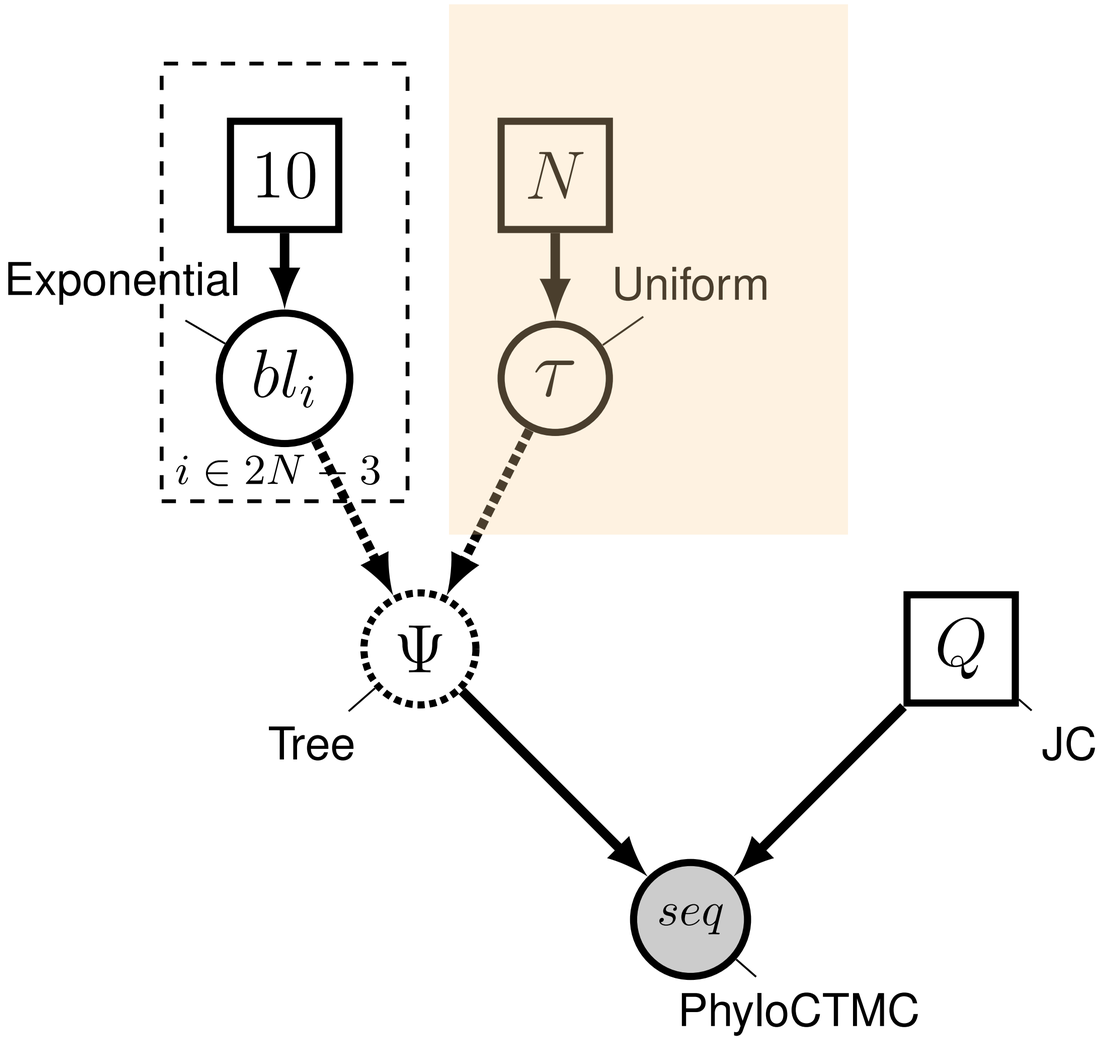
MCMC is an elegant algorithm trick to infer the posterior distribution.

It samples values directly from posterior in proportion to how probable they are, resulting in a histogram, which provides a good approximation of the posterior.

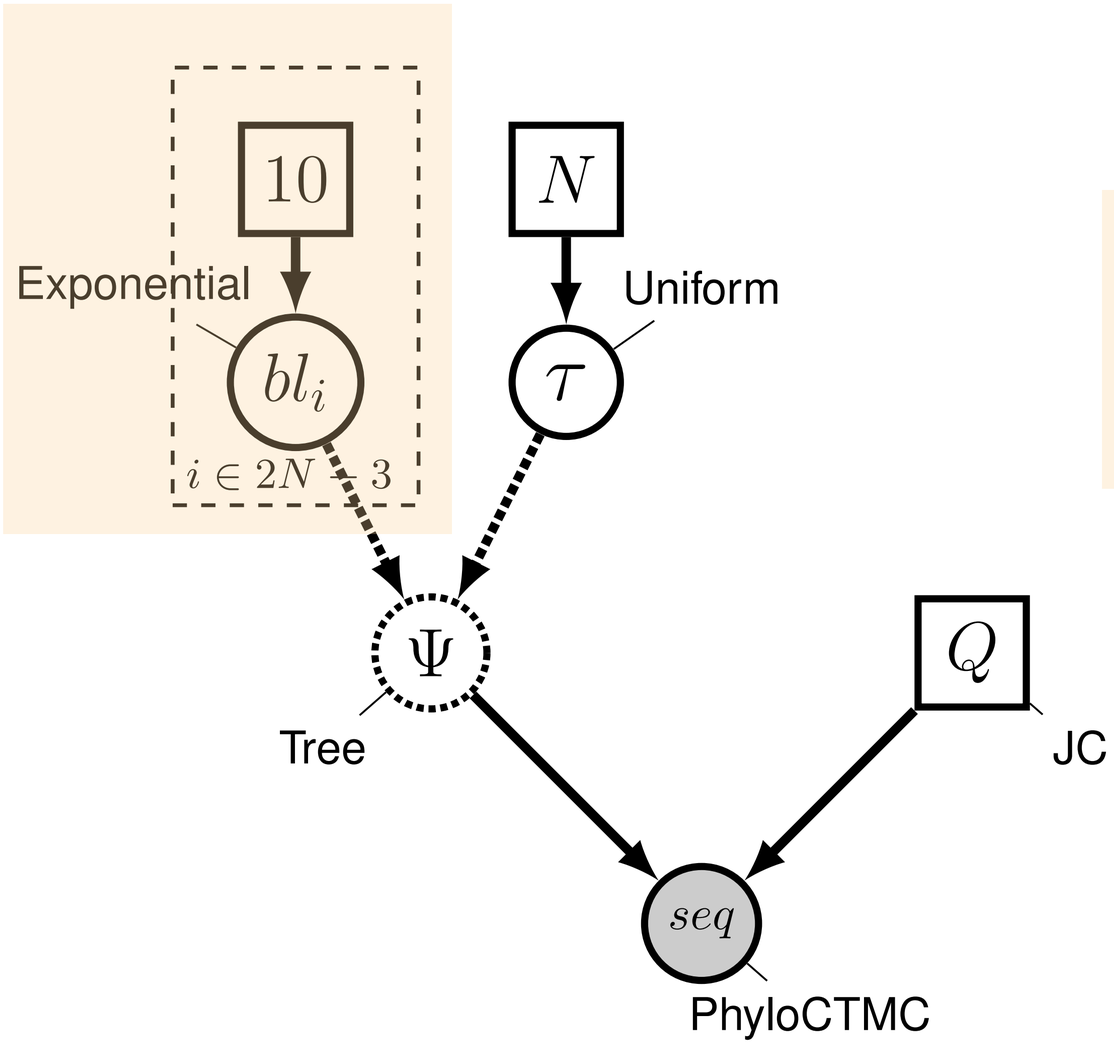
Bayesian tree inference using RevBayes



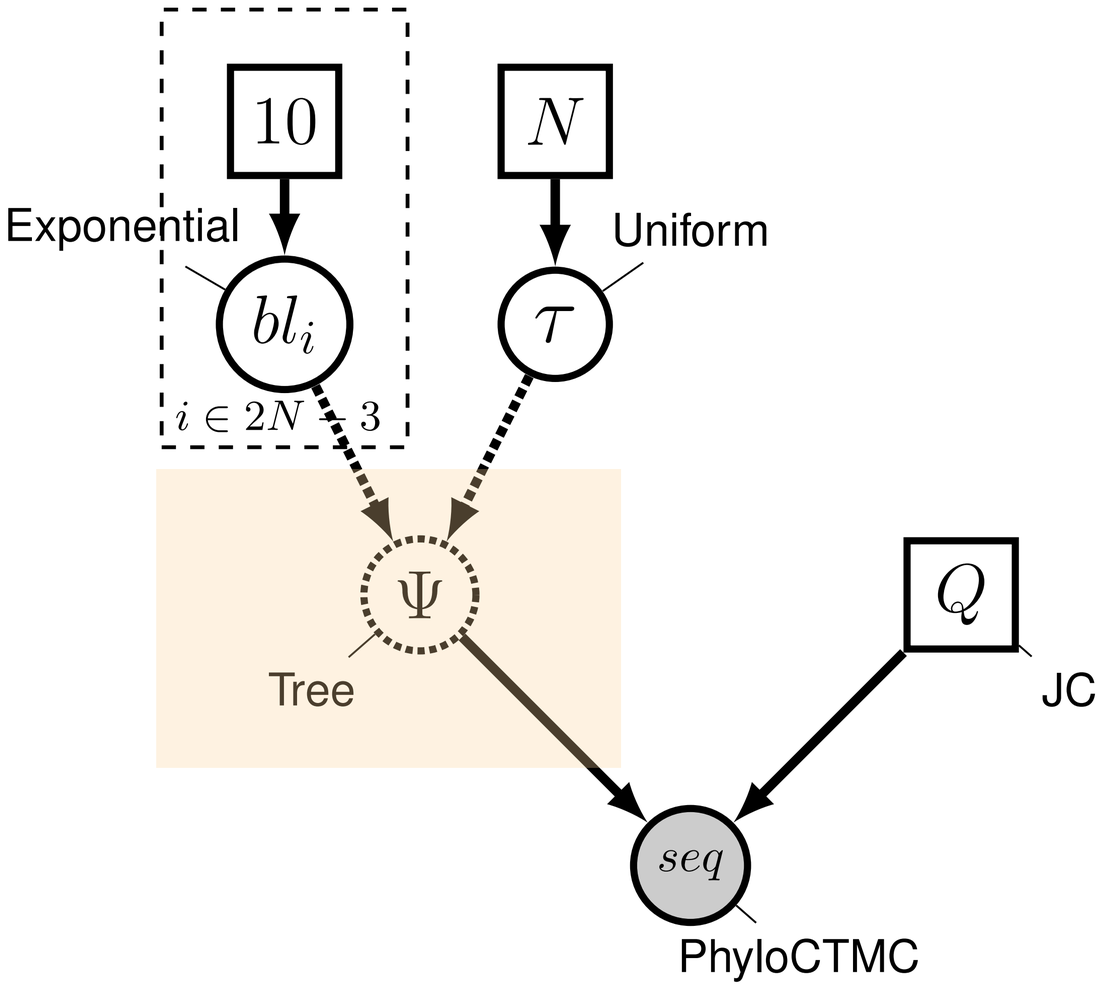

```
# prior on the tree topology
topology ~ dnUniformTopology(taxa)

# prior on the branch lengths
for (i in 1:num_branches) {
   br_lens[i] ~ dnExponential(10)
   moves.append( mvScale(br_lens[i]) )
}

tree := treeAssembly(topology, br_lens)

TL := sum(br_lens)

# 4 state rate maxtrix (JC model)
Q <- fnJC(4)

# attach the model to your sequence data
seq ~ dnPhyloCTMC(tree = tree, Q = Q, type = "DNA")
seq.clamp(data)
```



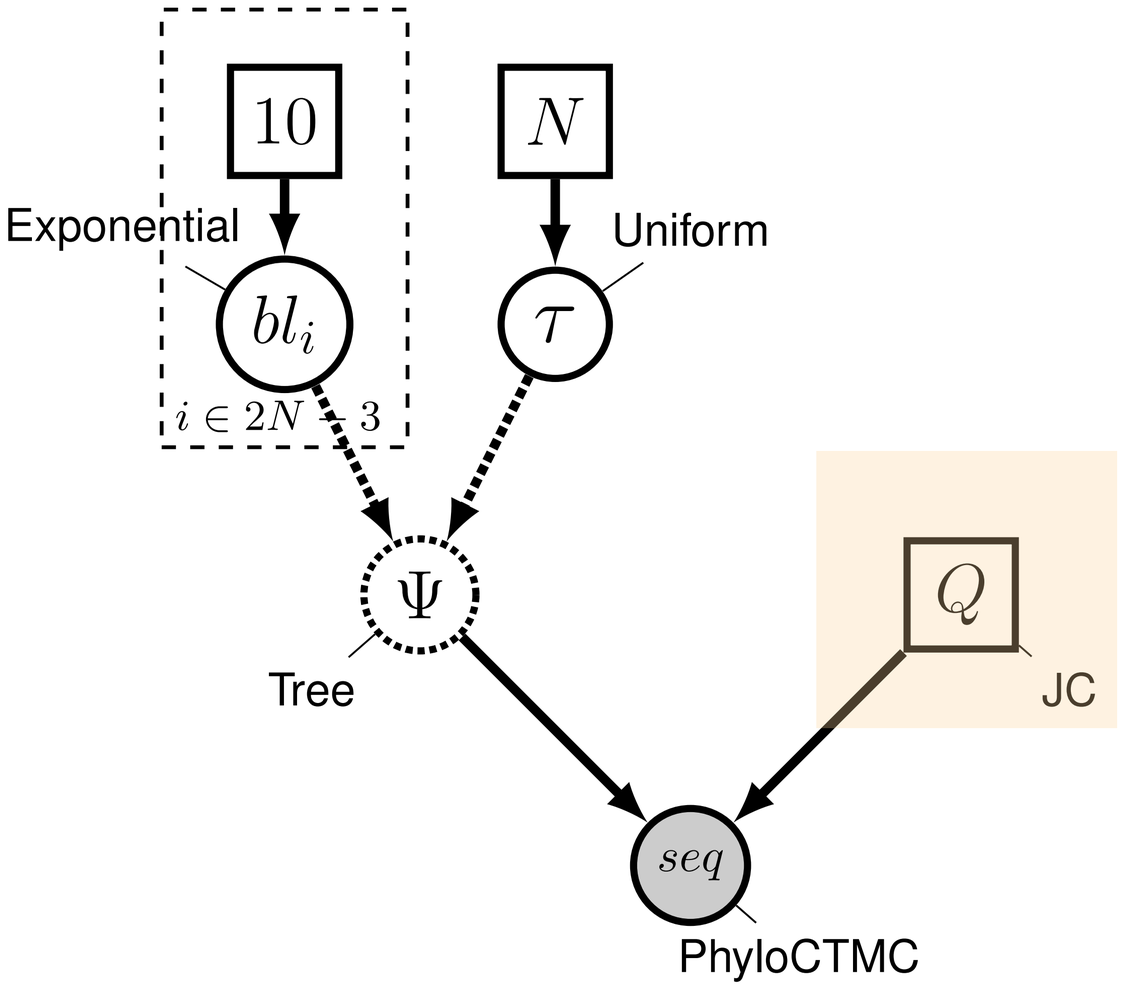

```
# prior on the tree topology
topology ~ dnUniformTopology(taxa)

# prior on the branch lengths
for (i in 1:num_branches) {
   br_lens[i] ~ dnExponential(10)
   moves.append( mvScale(br_lens[i]) )
}

tree := treeAssembly(topology, br_lens)

TL := sum(br_lens)

# 4 state rate maxtrix (JC model)
Q <- fnJC(4)

# attach the model to your sequence data
seq ~ dnPhyloCTMC(tree = tree, Q = Q, type = "DNA")
seq.clamp(data)
```



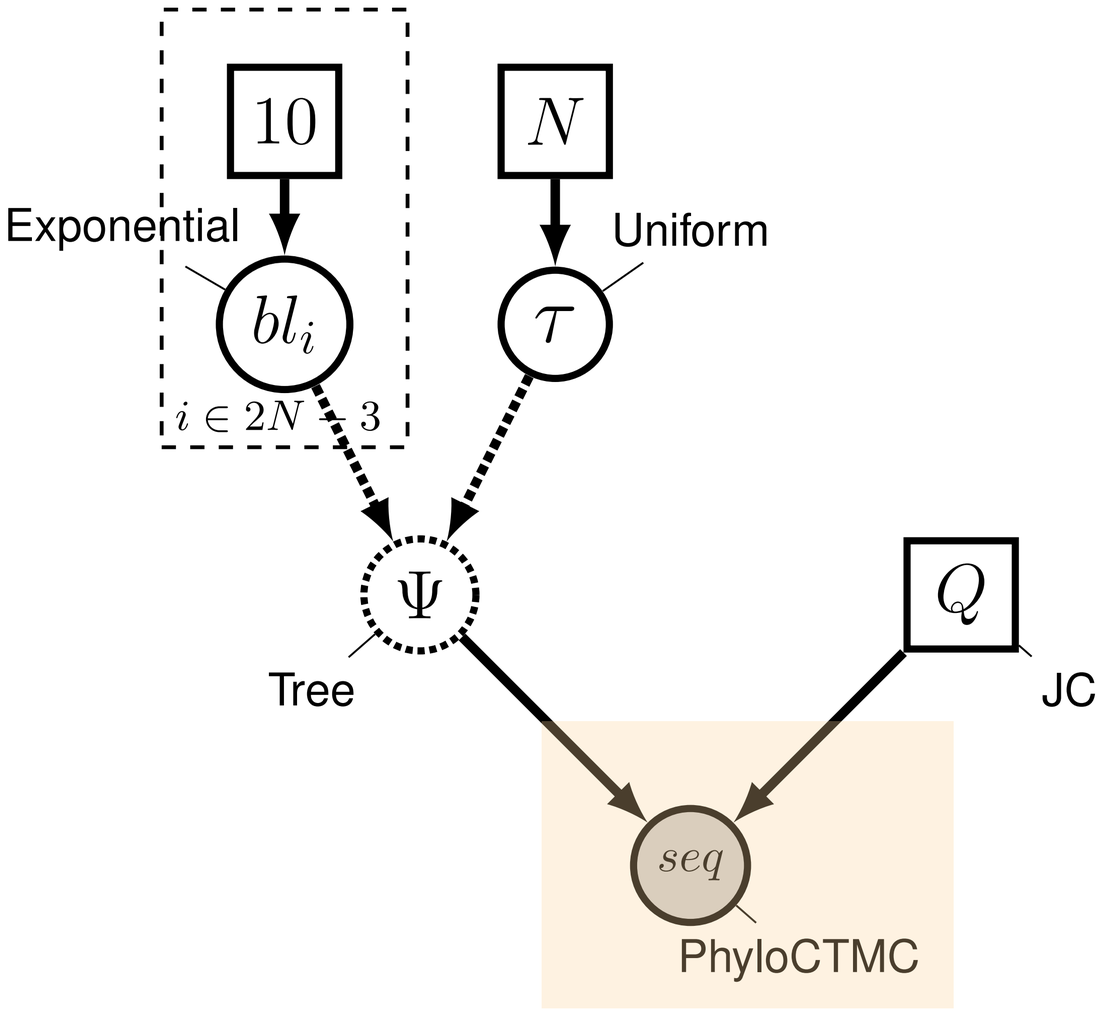

```
# prior on the tree topology
topology ~ dnUniformTopology(taxa)

# prior on the branch lengths
for (i in 1:num_branches) {
   br_lens[i] ~ dnExponential(10)
   moves.append( mvScale(br_lens[i]) )
}

tree := treeAssembly(topology, br_lens)

TL := sum(br_lens)

# 4 state rate maxtrix (JC model)
Q <- fnJC(4)

# attach the model to your sequence data
seq ~ dnPhyloCTMC(tree = tree, Q = Q, type = "DNA")
seq.clamp(data)
```




```
# prior on the tree topology
topology ~ dnUniformTopology(taxa)

# prior on the branch lengths
for (i in 1:num_branches) {
   br_lens[i] ~ dnExponential(10)
   moves.append( mvScale(br_lens[i]) )
}

tree := treeAssembly(topology, br_lens)

TL := sum(br_lens)

# 4 state rate maxtrix (JC model)
Q <- fnJC(4)

# attach the model to your sequence data
seq ~ dnPhyloCTMC(tree = tree, Q = Q, type = "DNA")
seq.clamp(data)
```




```
# prior on the tree topology
topology ~ dnUniformTopology(taxa)

# prior on the branch lengths
for (i in 1:num_branches) {
   br_lens[i] ~ dnExponential(10)
   moves.append( mvScale(br_lens[i]) )
}

tree := treeAssembly(topology, br_lens)

TL := sum(br_lens)

# 4 state rate maxtrix (JC model)
Q <- fnJC(4)

# attach the model to your sequence data
seq ~ dnPhyloCTMC(tree = tree, Q = Q, type = "DNA")
seq.clamp(data)
```




```
# prior on the tree topology
topology ~ dnUniformTopology(taxa)

# prior on the branch lengths
for (i in 1:num_branches) {
   br_lens[i] ~ dnExponential(10)
   moves.append( mvScale(br_lens[i]) )
}

tree := treeAssembly(topology, br_lens)

TL := sum(br_lens)

# 4 state rate maxtrix (JC model)
Q <- fnJC(4)

# attach the model to your sequence data
seq ~ dnPhyloCTMC(tree = tree, Q = Q, type = "DNA")
seq.clamp(data)
```

Exercise 4: Bayesian tree inference