Analytical paleo

Introduction to Bayesian phylogenetic inference and MCMC

Rachel Warnock

20.08.26



A bit about me

Worked in both paleo and comp bio groups

Interested in phylogenetic approaches that can be applied to the fossil record

Also interested in interdisciplinary training (>20 workshops)



Hidden instructors



Joëlle
Barido-Sottani
ENS Paris



Laura Mulvey
QMU London



April Wright
SLU (Louisiana) & NSF

Plus many other members of the BEAST2 and RevBayesian communities

Who are you?

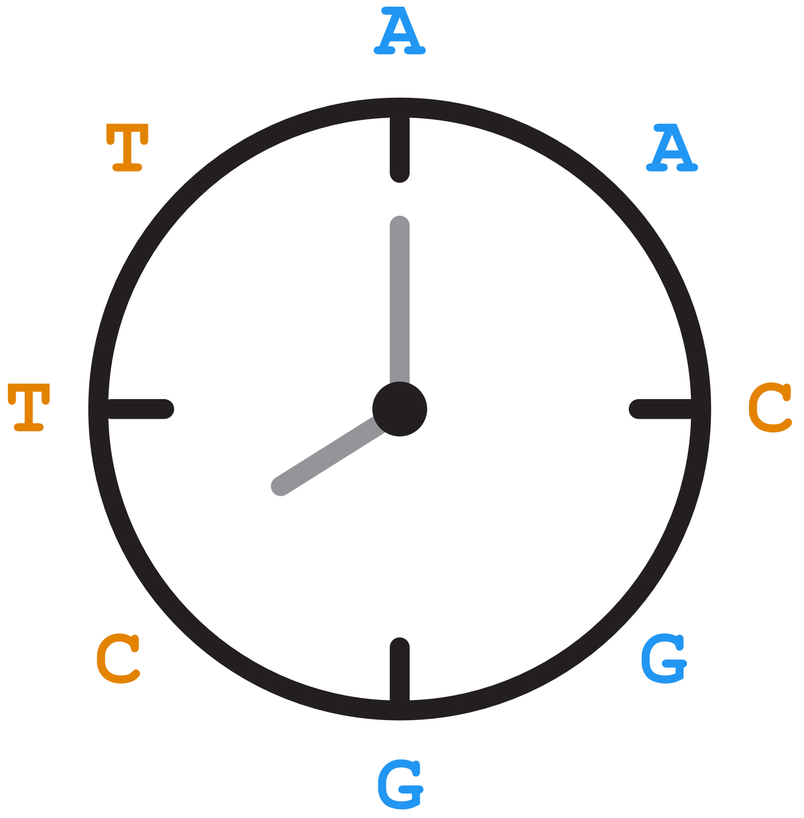
Where are you based? What do you work on?
What’s the last book you enjoyed? (optional)



Please ask questions!

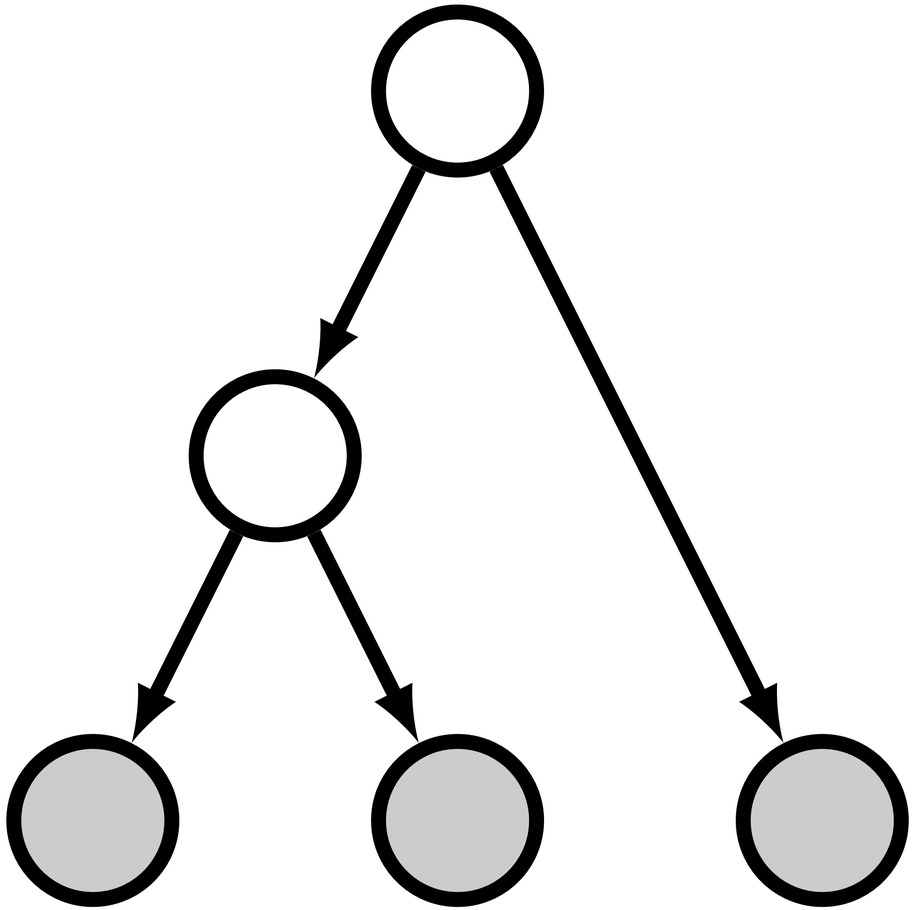
Today (day 5)

- Introduction to Bayesian (time)tree inference and MCMC
- Introduction to BEAST2
- The fossilised birth-death process



Today (day 6)

- Introduction to RevBayes
- Morphological models and model adequacy
- FBD model extensions



Software

You will need:


BEAST2


FigTree


Tracer


RevBayes (for tomorrow)

What do I 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?

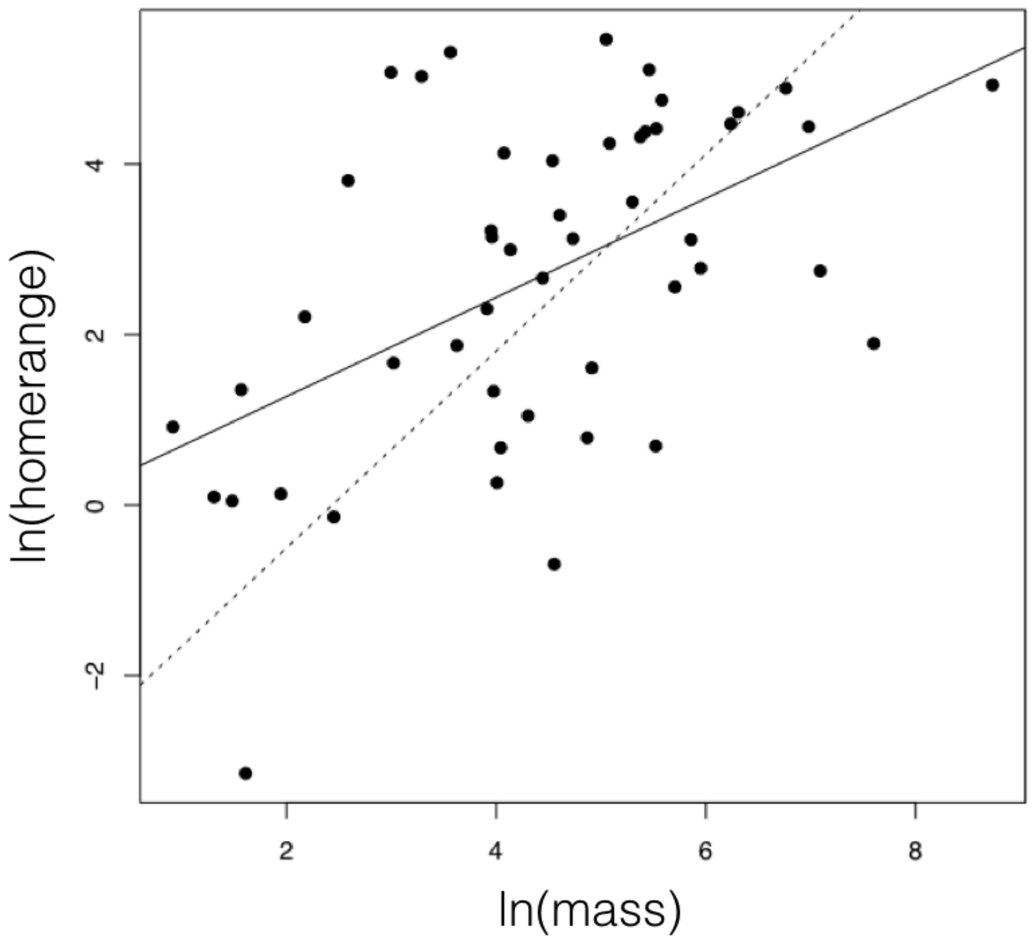
A statistical model is a type of model that includes a set of assumptions about the data-generating process

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

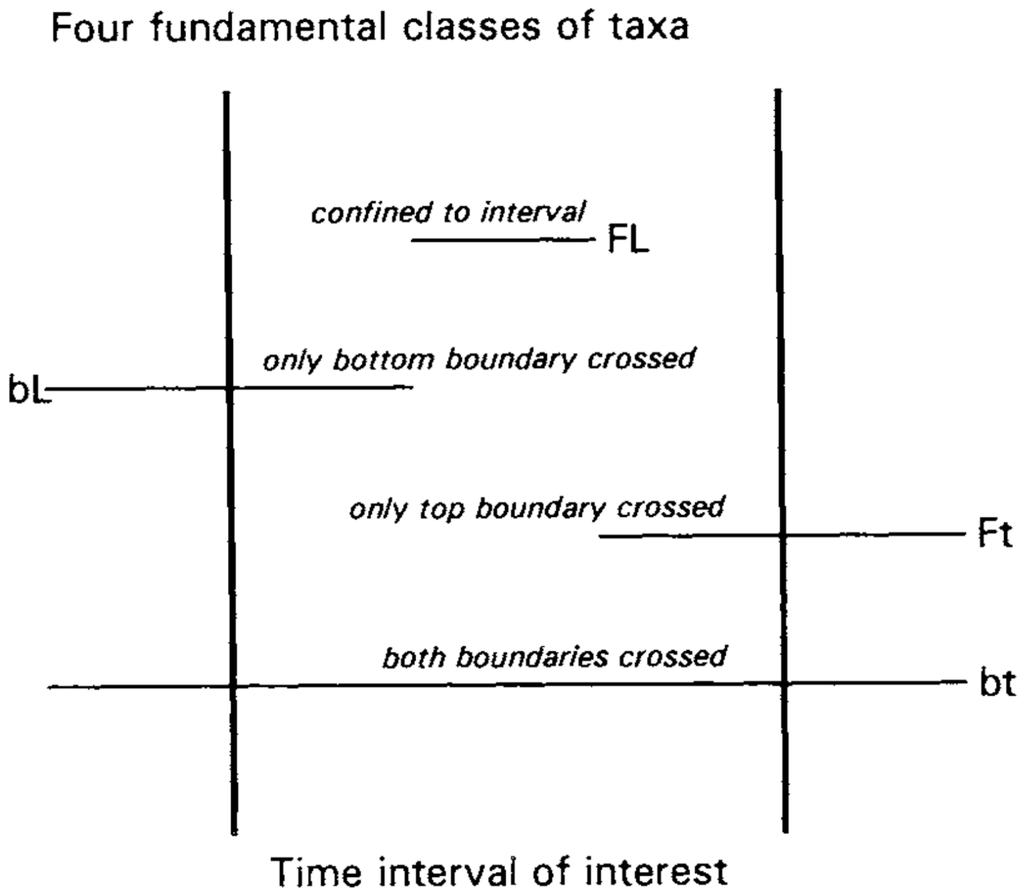
Image source *Harmon* (2019)

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

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

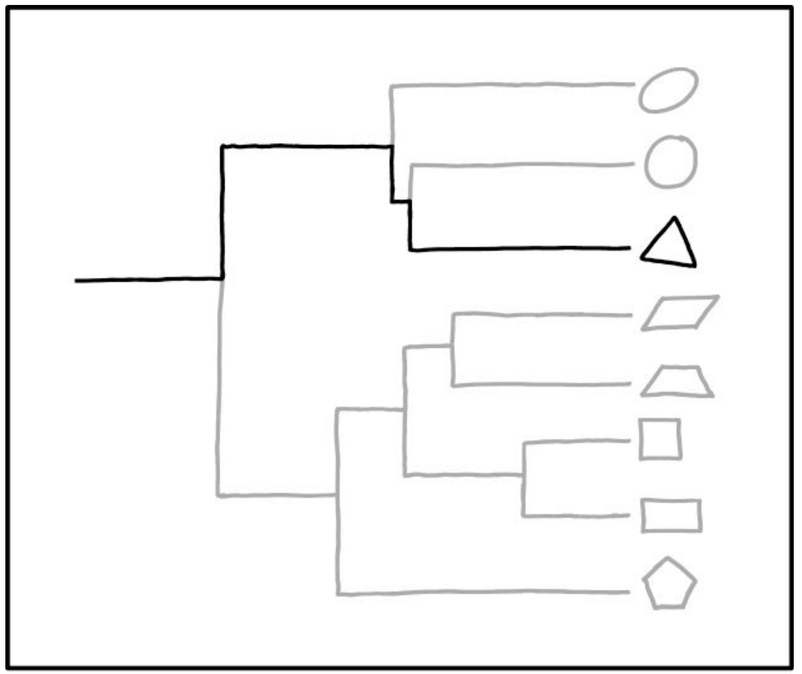
An algorithm is a precise rule (or set of rules) specifying how to solve some problem

```
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

What is phylogenetics?



THE PHYLOGENETIC REVOLUTION CONTINUES:

TRIANGLES WERE LONG BELIEVED TO BE RELATED TO SQUARES, BUT GENETIC ANALYSIS PROVES THAT THEY ARE ACTUALLY VERY POINTY CIRCLES.

XKDC

Phylogenetics

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

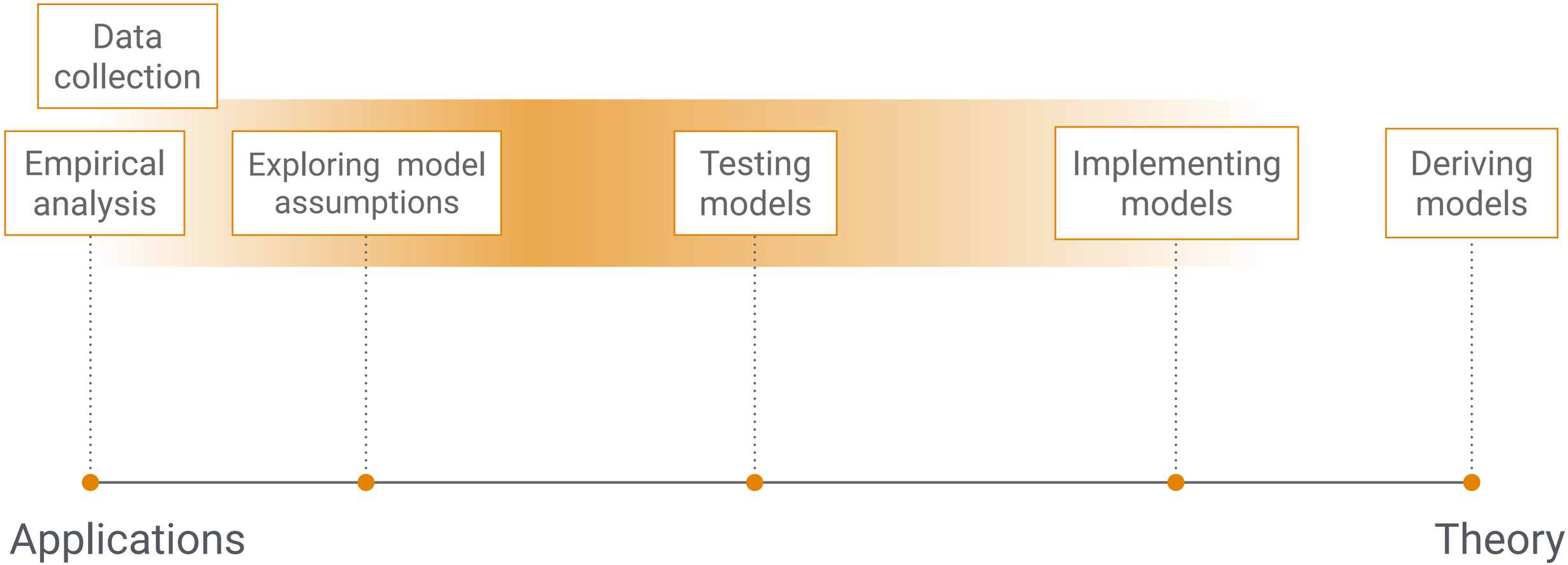
A phylogeny captures part of evolutionary history that is otherwise not directly observable

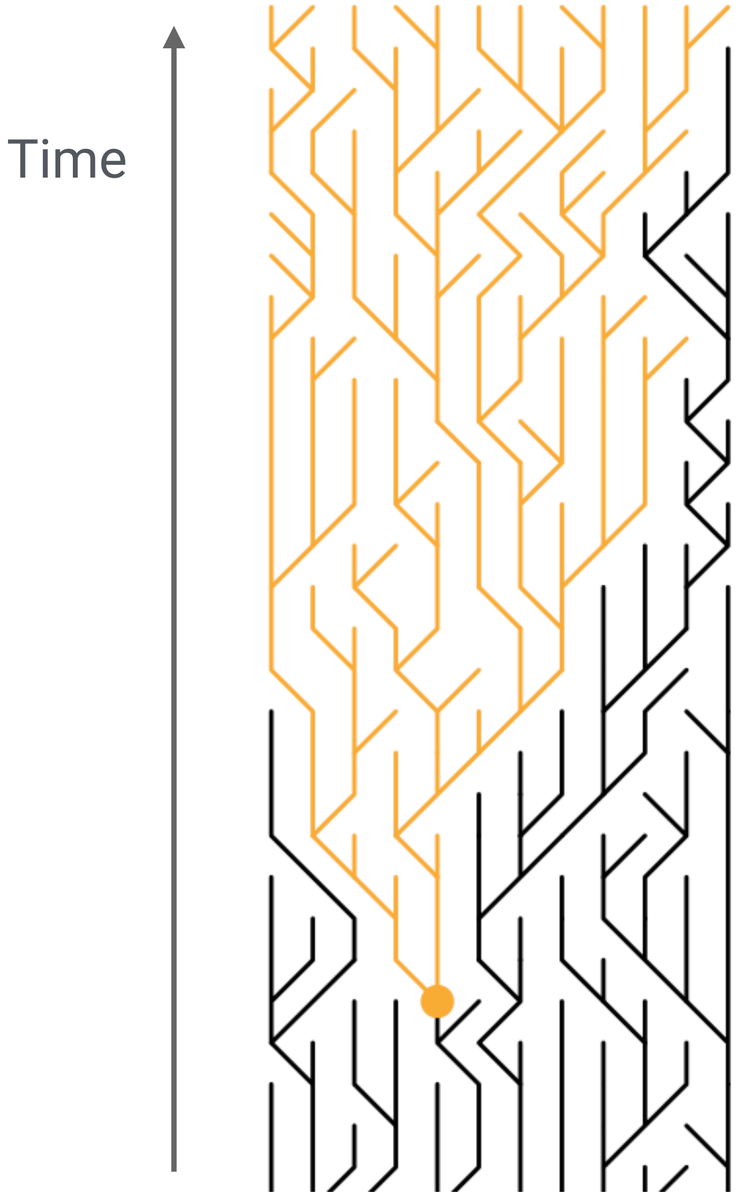
Phylodynamics aims to quantify the **branching processes** that gave rise to the tree, e.g., speciation, extinction

Phylogenetic comparative methods are used to study trait evolution, taking into account phylogenetic non-independence

Explore the tree of life using the Open Tree of Life tool, currently inc. 2,385,875 tips

Research topics in phylogenetics





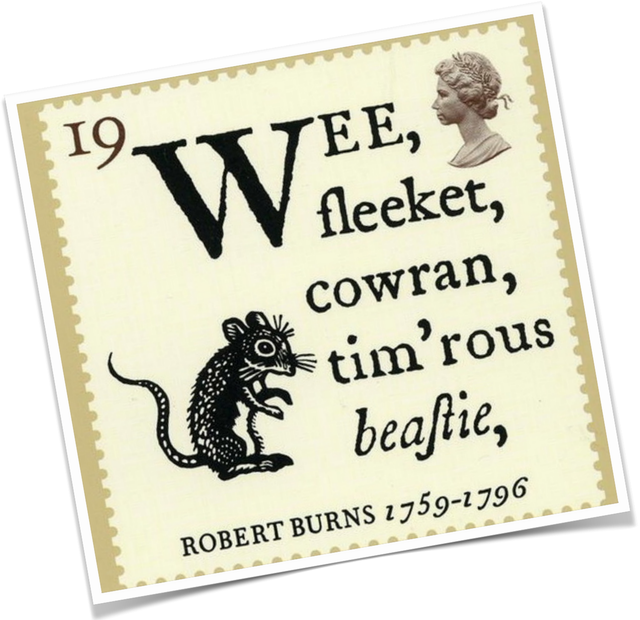
Biological unit

- populations
- species
- viruses
- cells
- languages

Data

- DNA
- morphology
- words

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

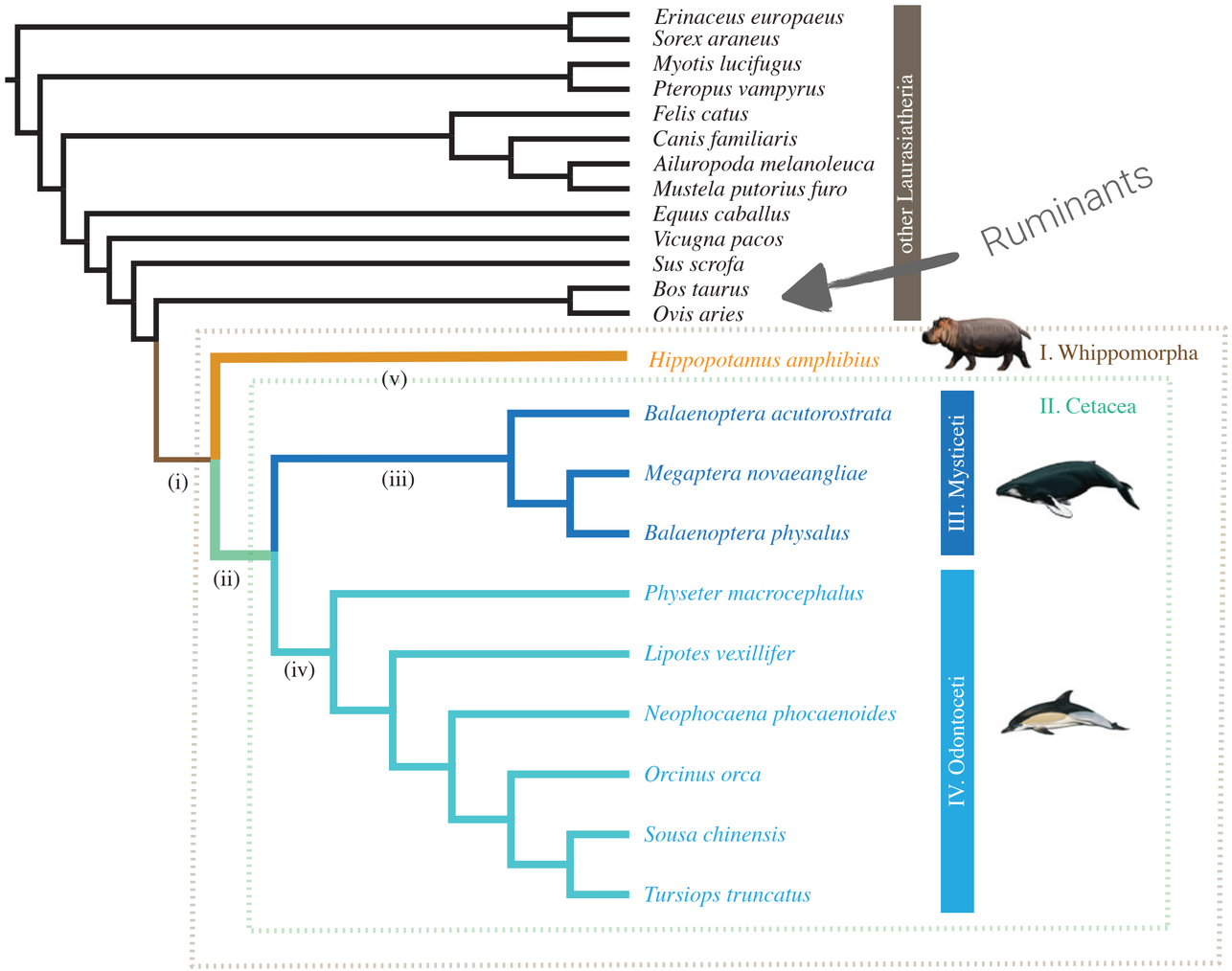


Scots poem - also the BEAST2 logo!

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 Savage* (1997)

What can we learn from trees?



How are our favourite species related?

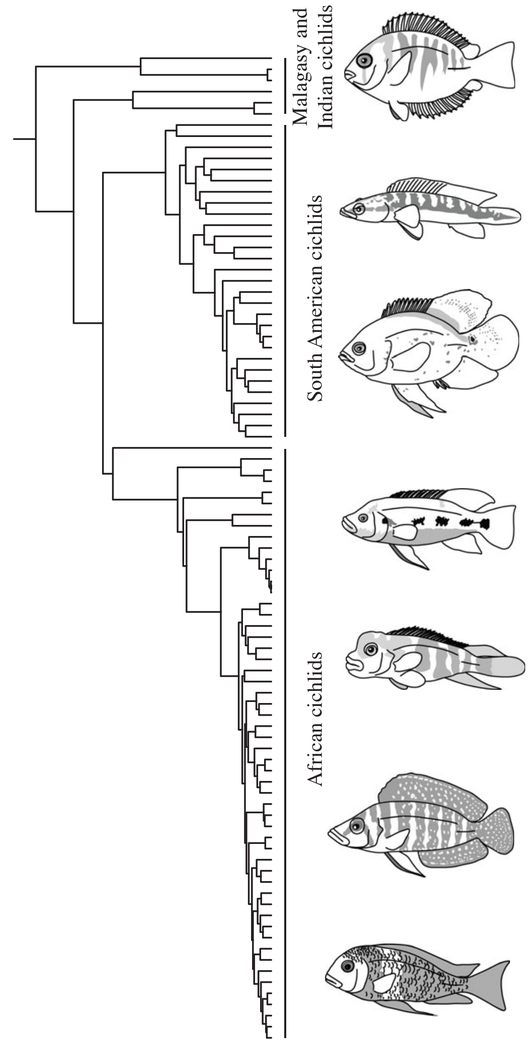
Does the phylogeny support the taxonomy?

What was the sequence of character evolution?

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

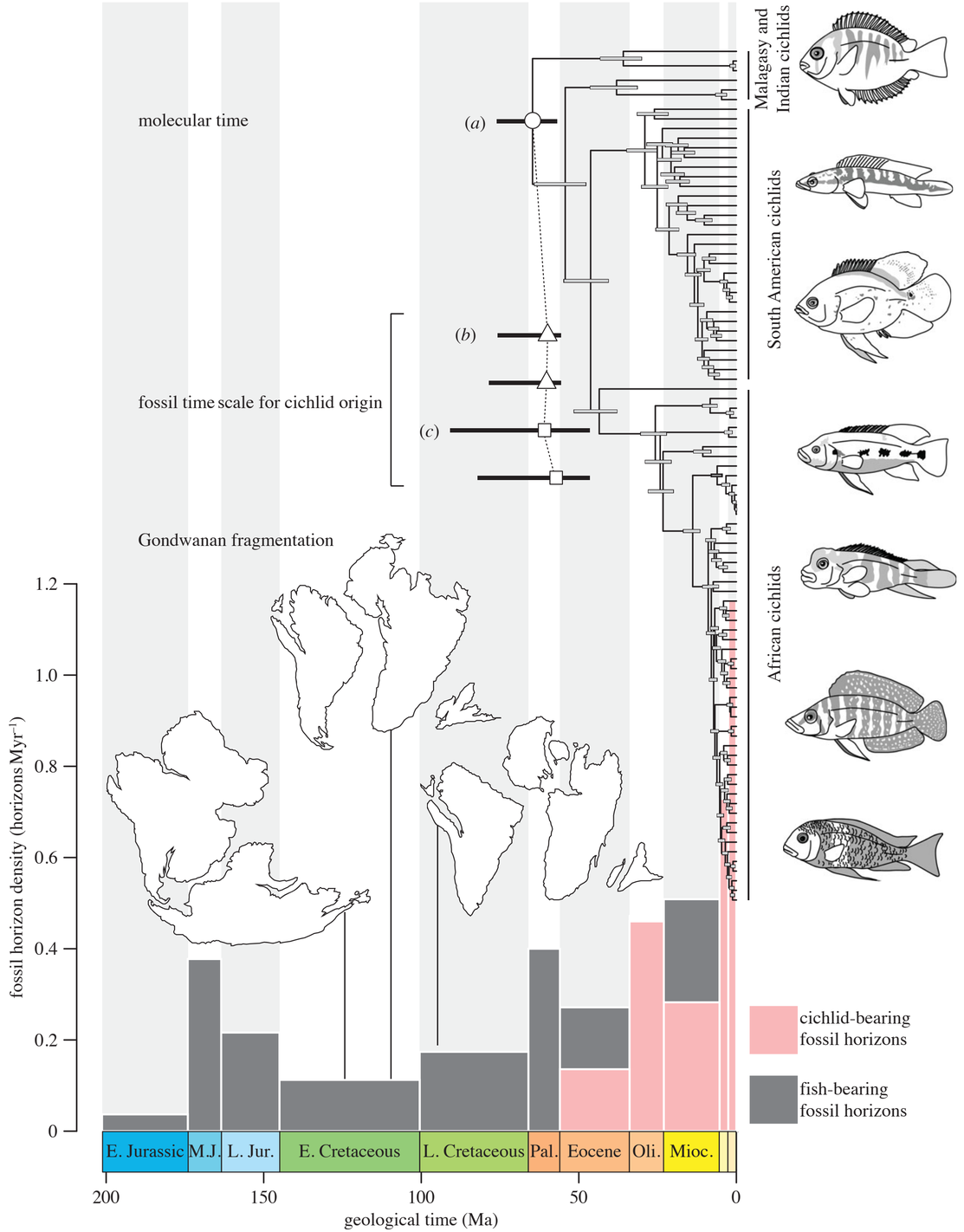
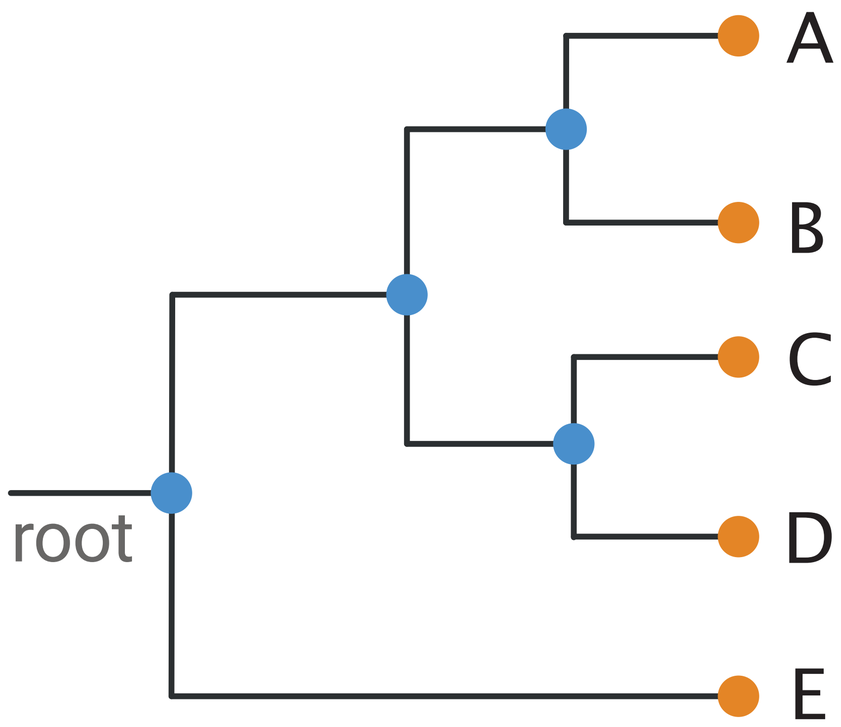


Image adapted from *Friedmann et al.* (2013)

Some basic terms

MRCA = most recent common ancestor



internal nodes or MRCAs

tips or leaves

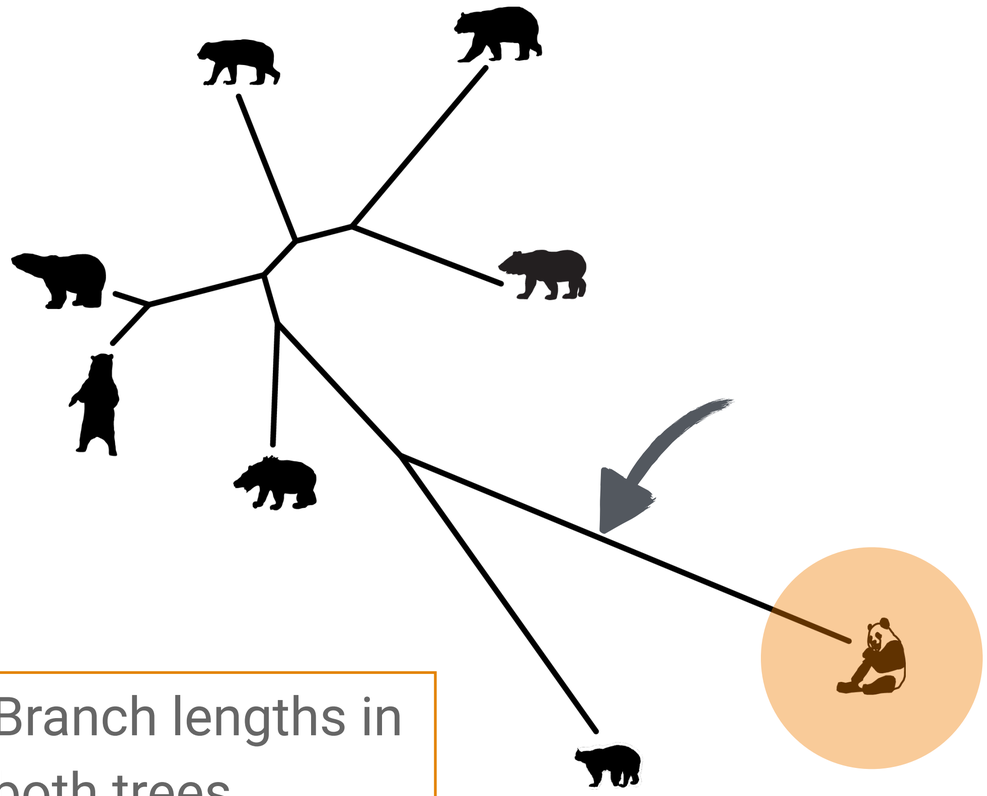
branches or edges

branch lengths = genetic distance *or* time

How to read a phylogenetic tree

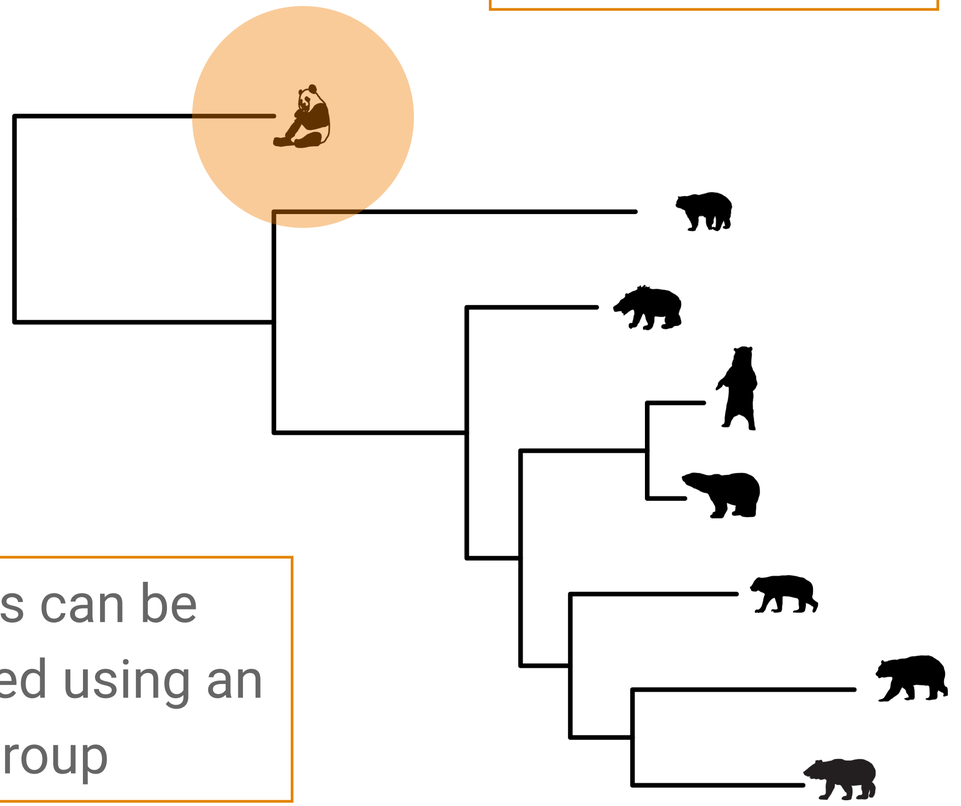
Two types of trees

Unrooted vs. rooted trees



Branch lengths in both trees represents genetic distance

A rooted tree shows the direction of time



Trees can be rooted using an outgroup

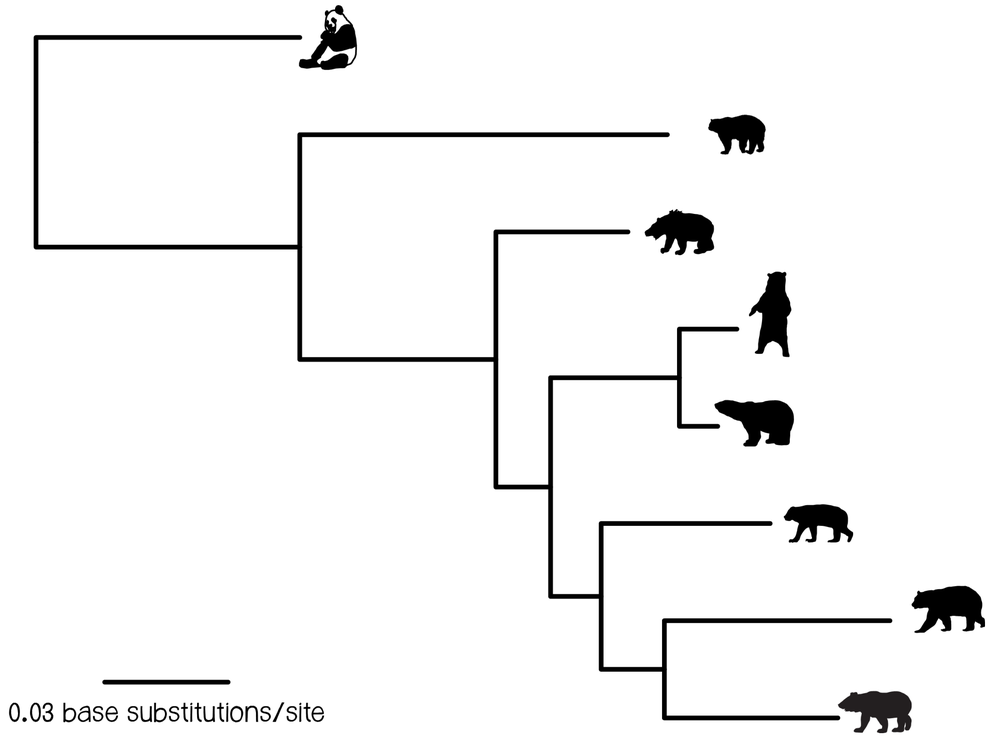
Image source Tracy Heath

[Explore tree rooting]

Branch lengths = genetic *or* time

Tip look at the axis

Trees show the relationships among **extant** (living) bear species



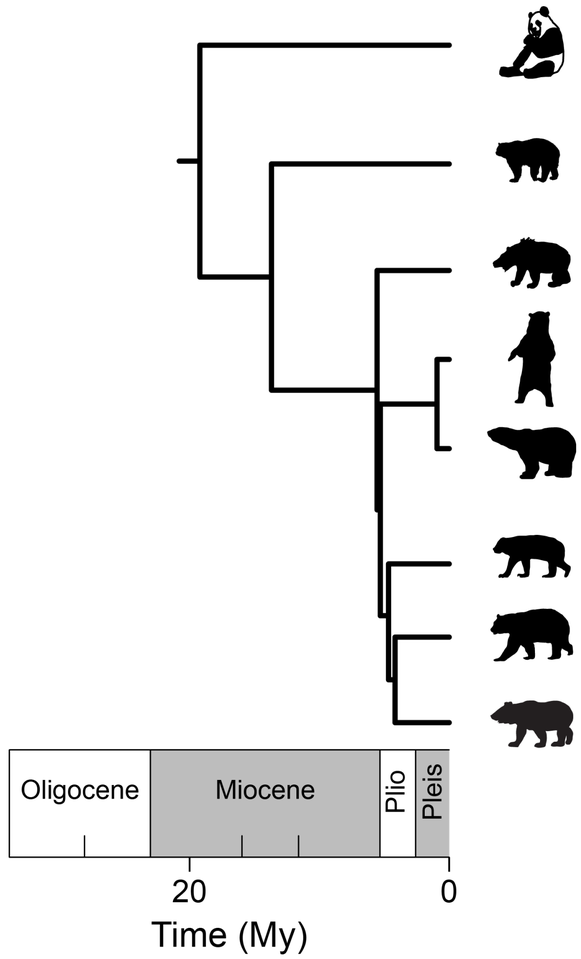


Image source Tracy Heath

Dated tree showing the relationship of extant *and* fossil bear species

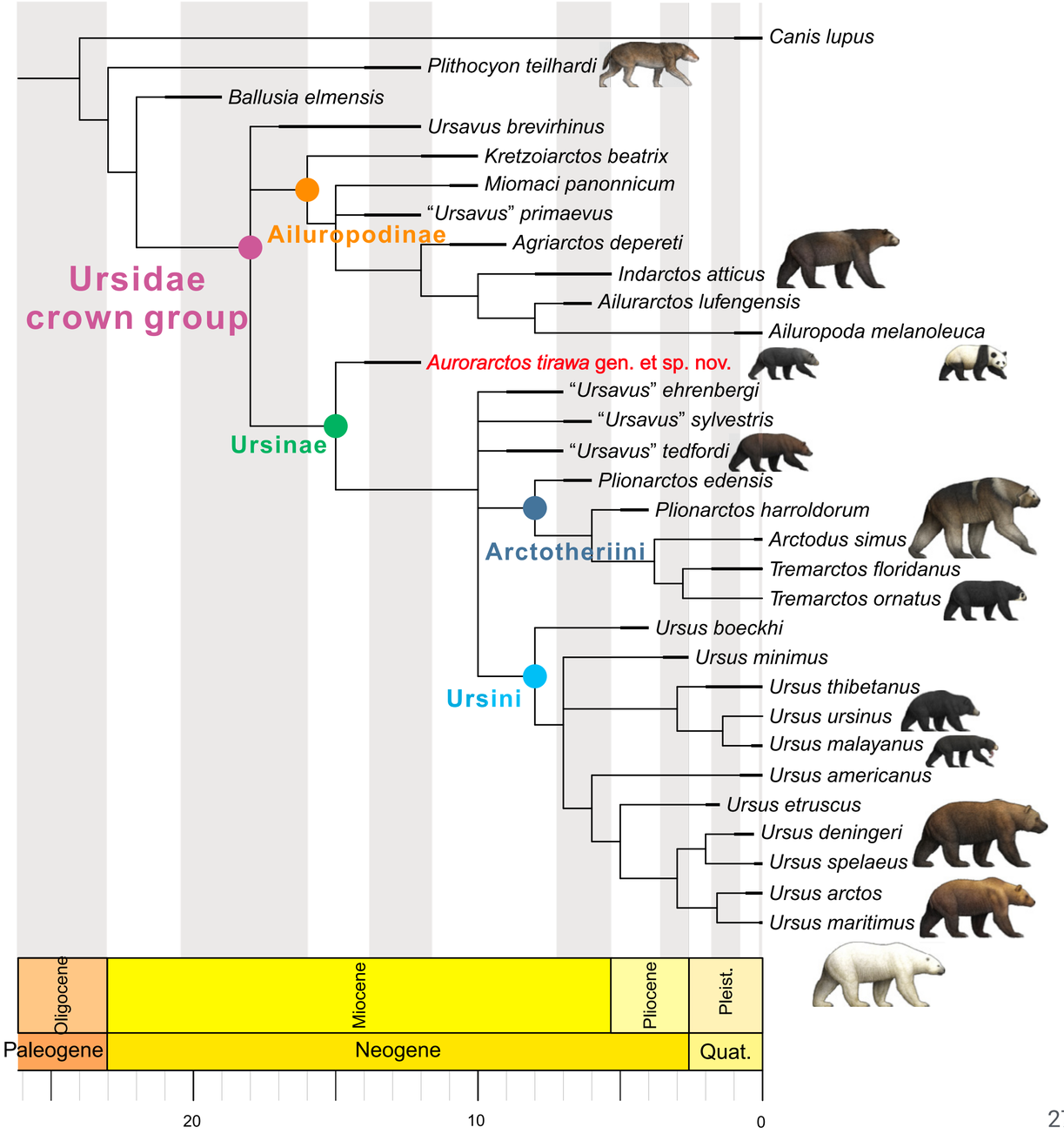


Image adapted from *Jiangzuo and Flynn* (2020)

How do we find the ‘best’ tree?

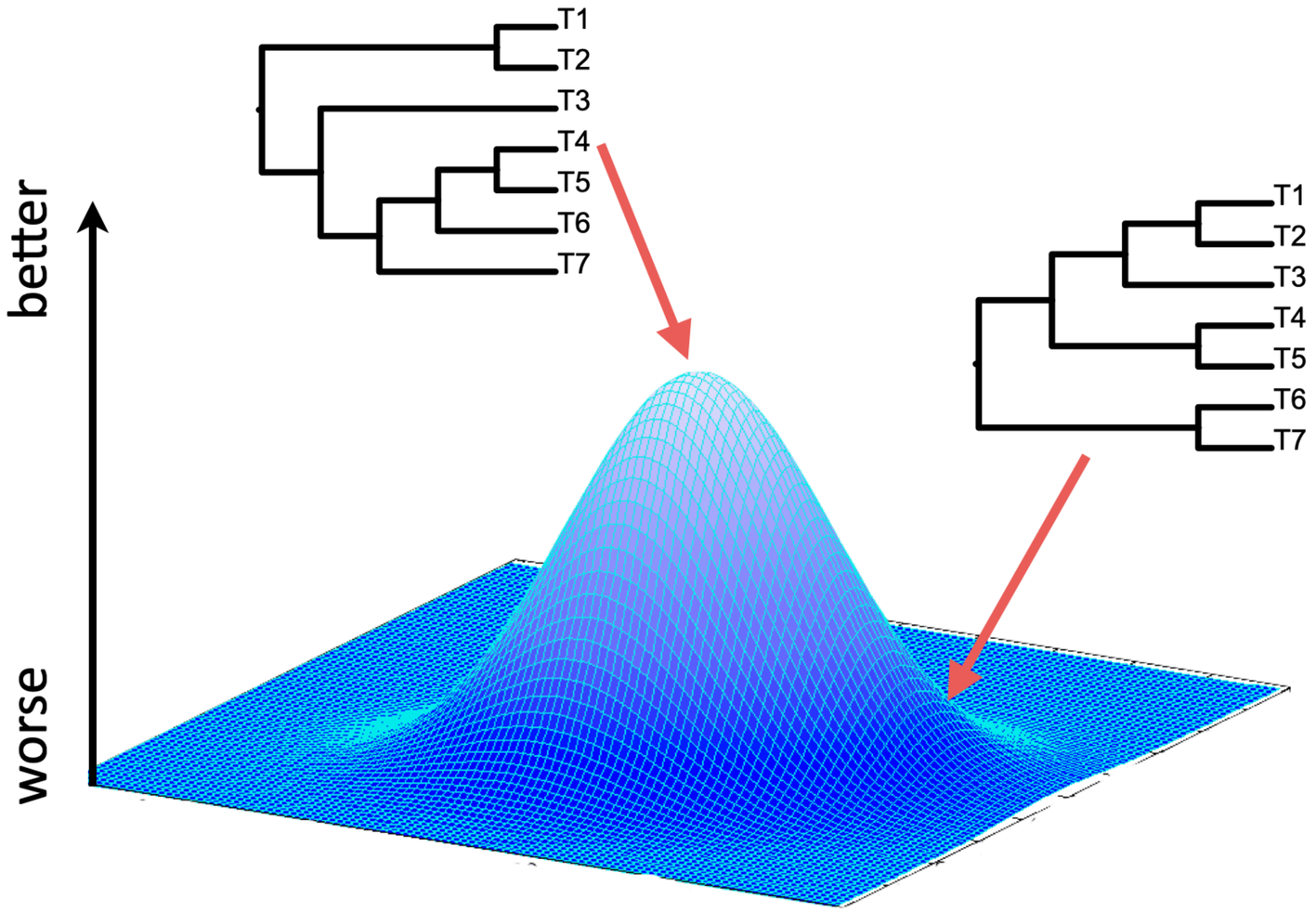


Image source Tracy Heath

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 (undated)

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

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

Bayes' theorem

$$\Pr(\textbf{model} \mid \textbf{data}) \propto \Pr(\textbf{data} \mid \textbf{model})\Pr(\textbf{model})$$

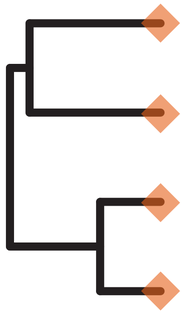
The posterior is proportional to the product of the prior and the likelihood

Components used to infer trees

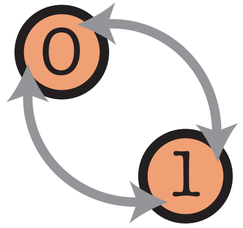
without yet considering time

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

data
sequences or characters

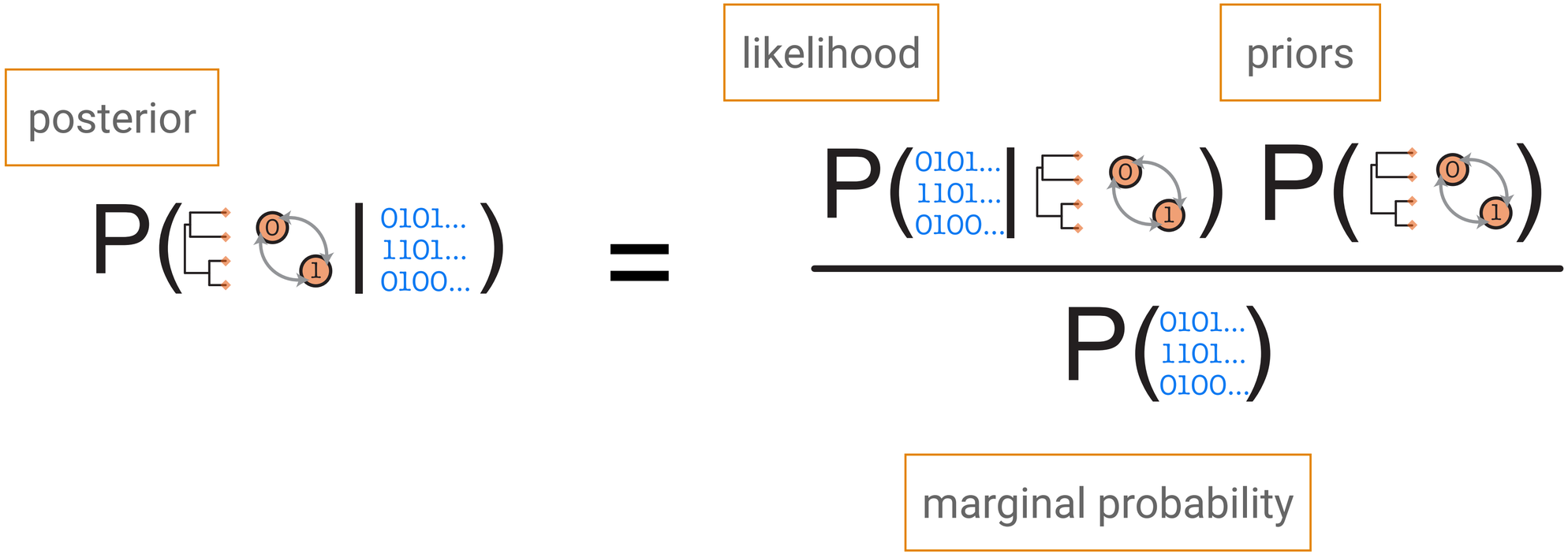


tree
topology and branch lengths



substitution model

Bayesian tree inference



Molecular evolution

Character evolution

along species trees

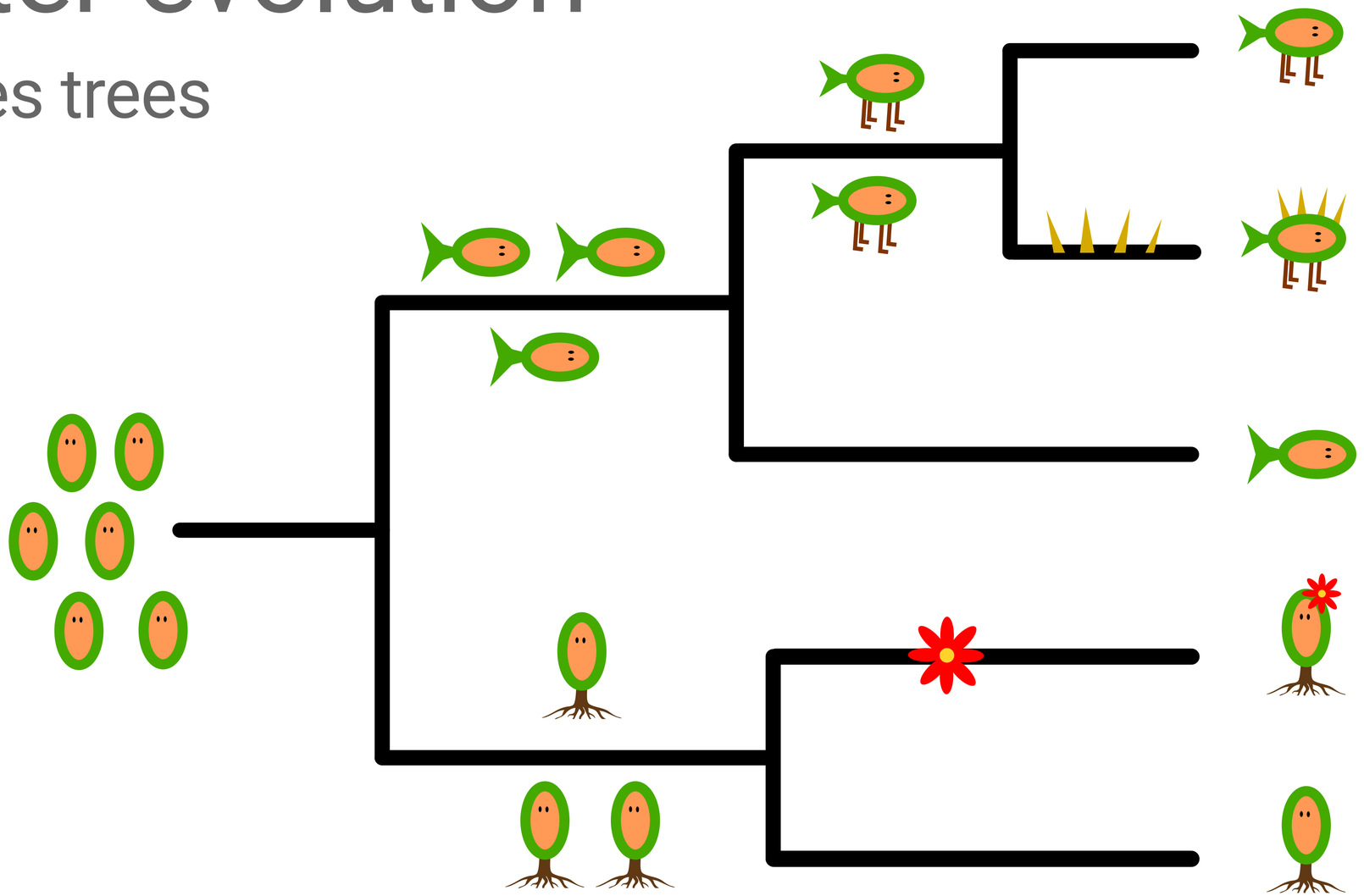
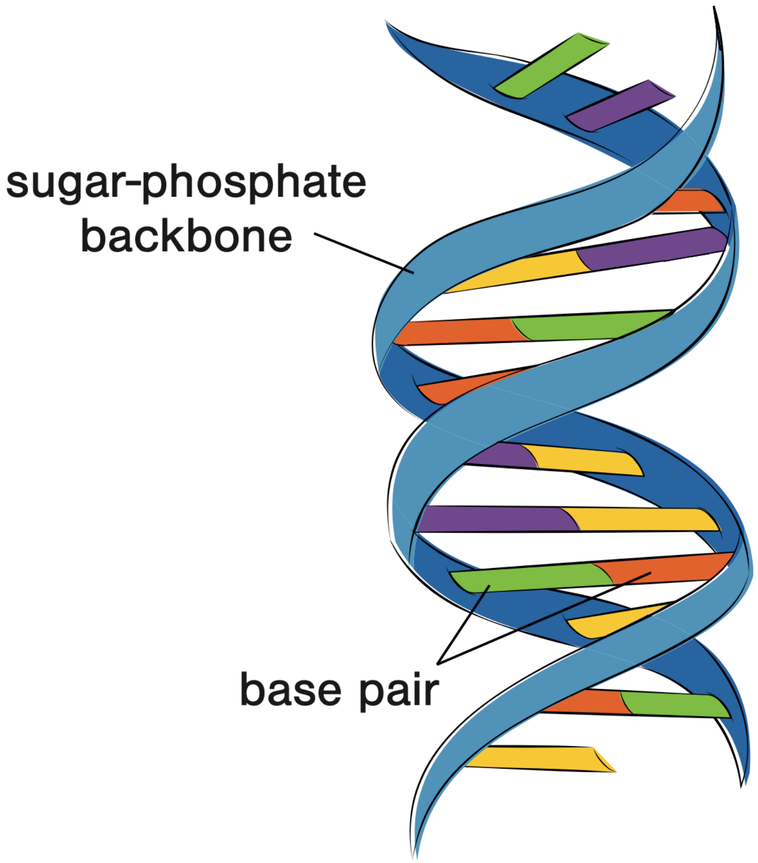
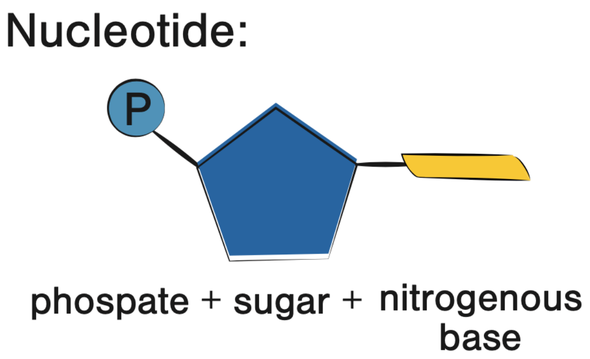
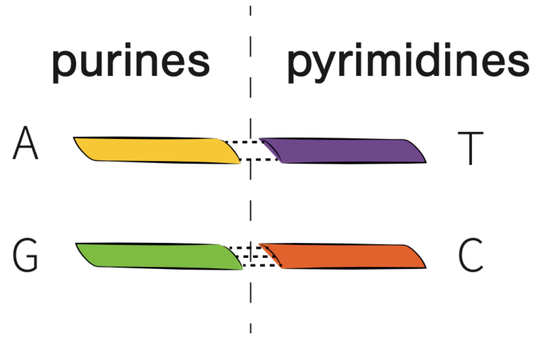


Image: Joëlle Barido-Sottani

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

Mutation vs. substitution

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

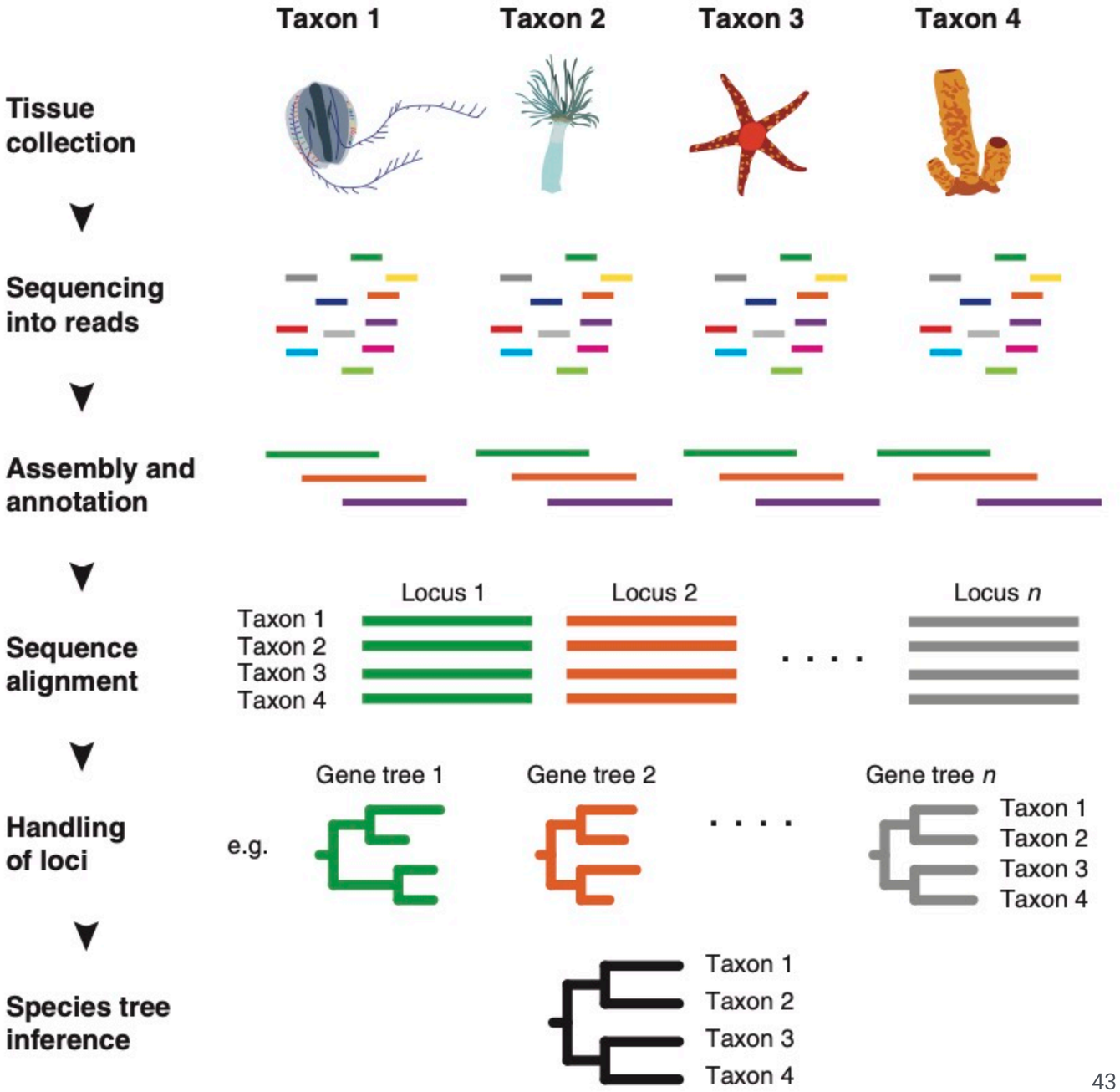
Individuals of the same species have identical characters at *most* positions in their genome (only 0.1% vary among humans)

Most mutations are repaired but can persist across generations

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

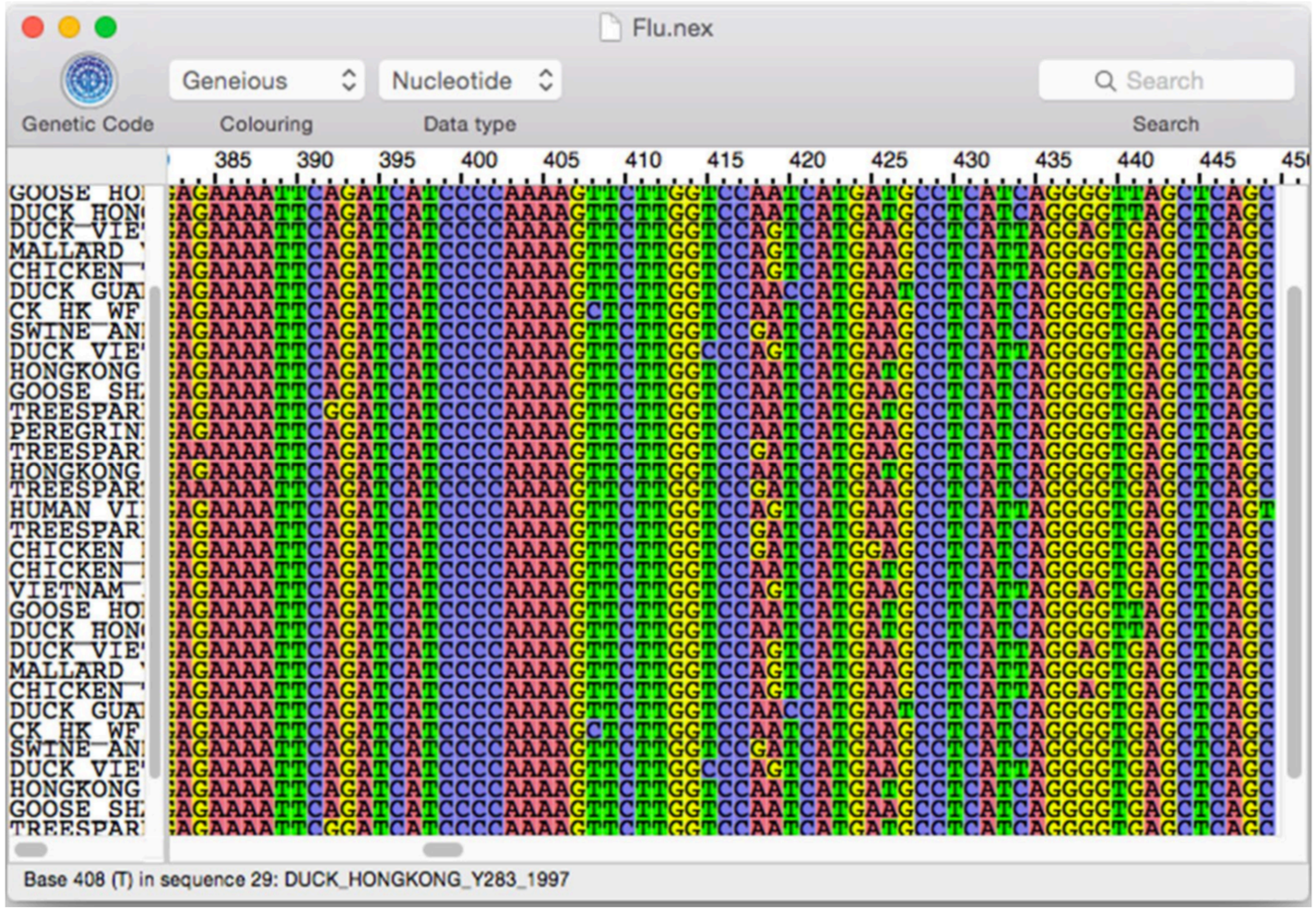
DNA sequencing

Multiple sequence alignment software establishes homology across sites from different species



Duchêne (2021) Phylogenomics Primer

Multiple sequence alignment



See extra slides for details

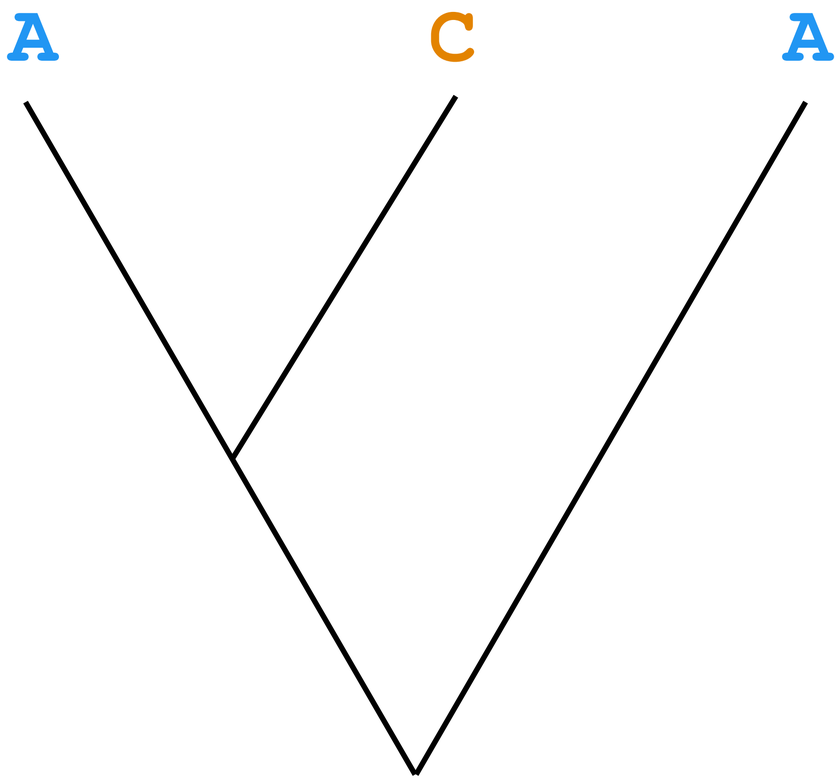
```
#NEXUS

[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;
```



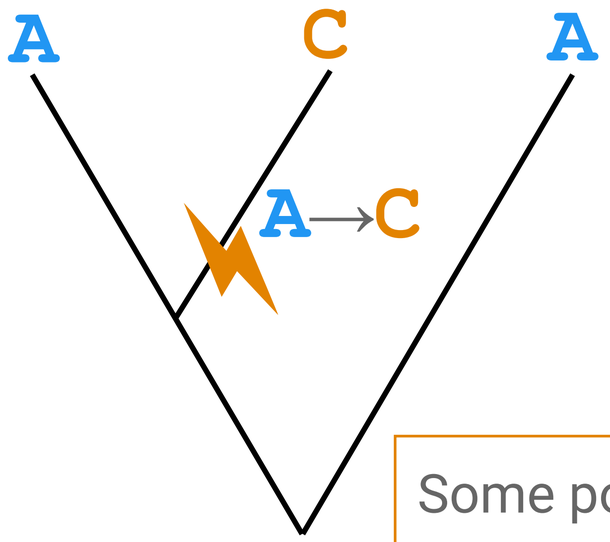
Substitution models

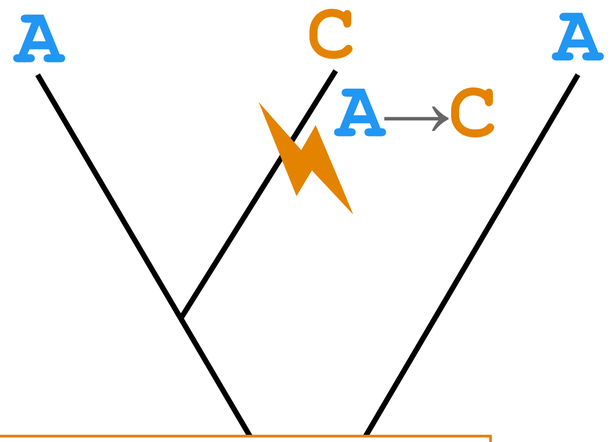
The data are the observed states at the tips

How probable is our data, given my tree?

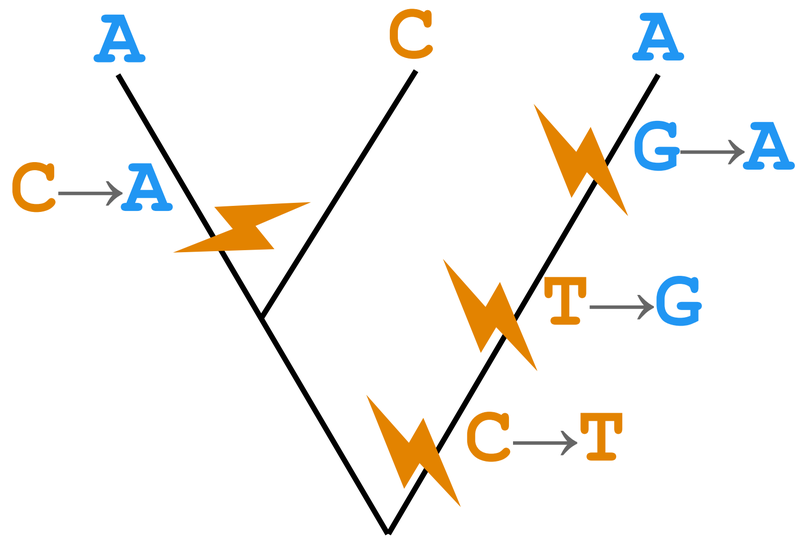


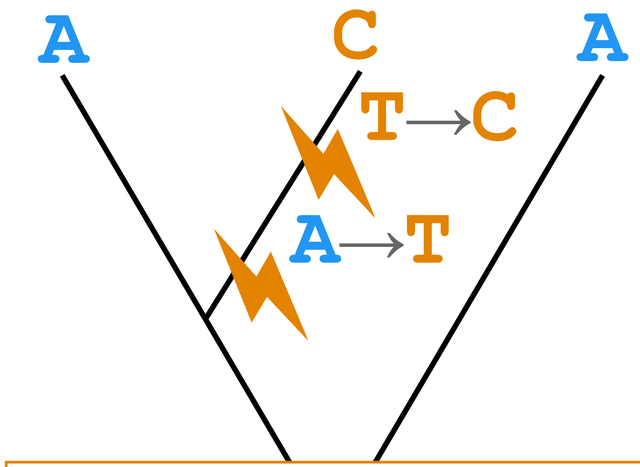
To apply a model based approach we need to be able to compute the probability of our sequence alignment (or character matrix)

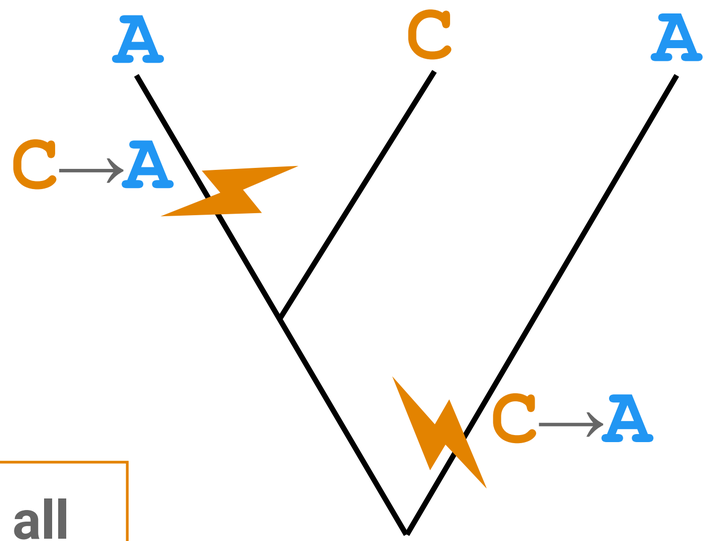




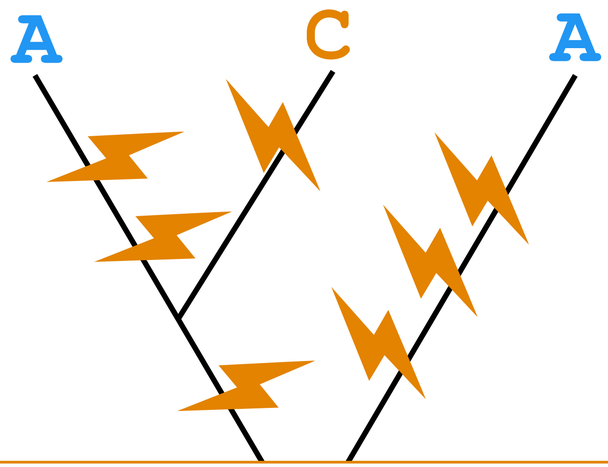
Some possible character histories...







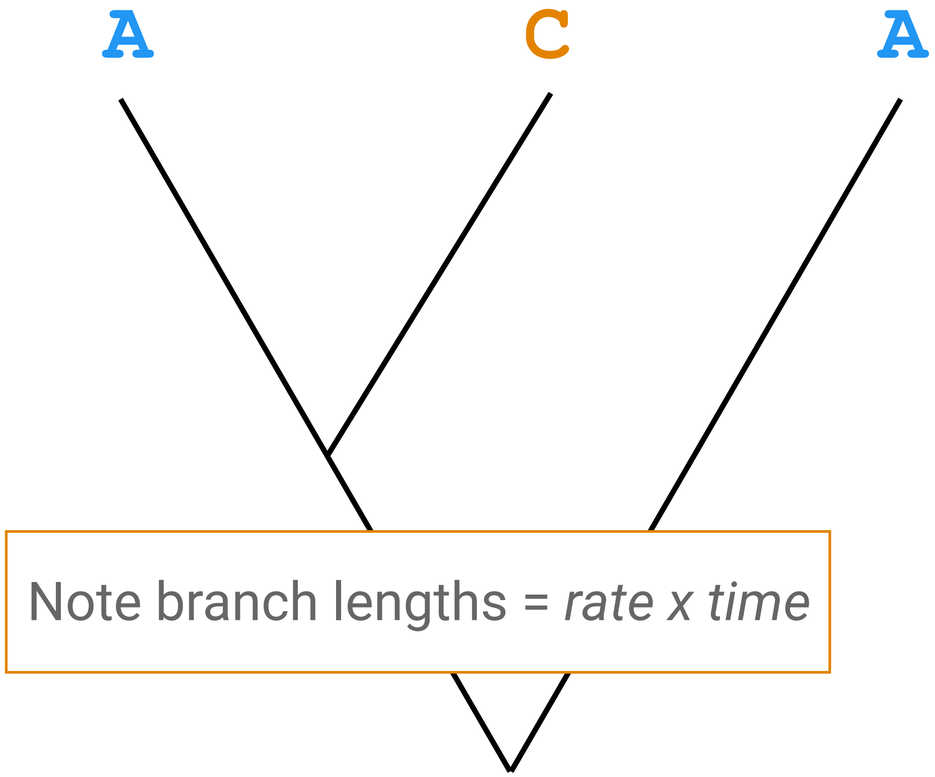
Model based approaches take **all possible histories** into account!



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)

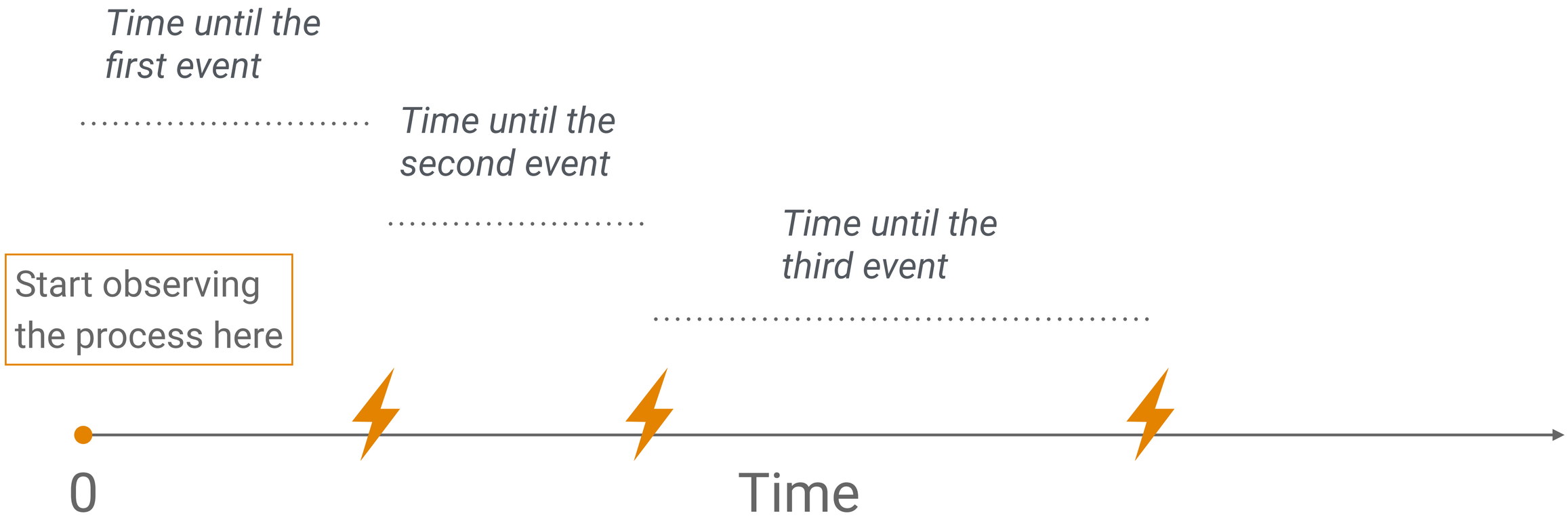
Models of molecular sequence evolution

Also known as substitution / site / character models

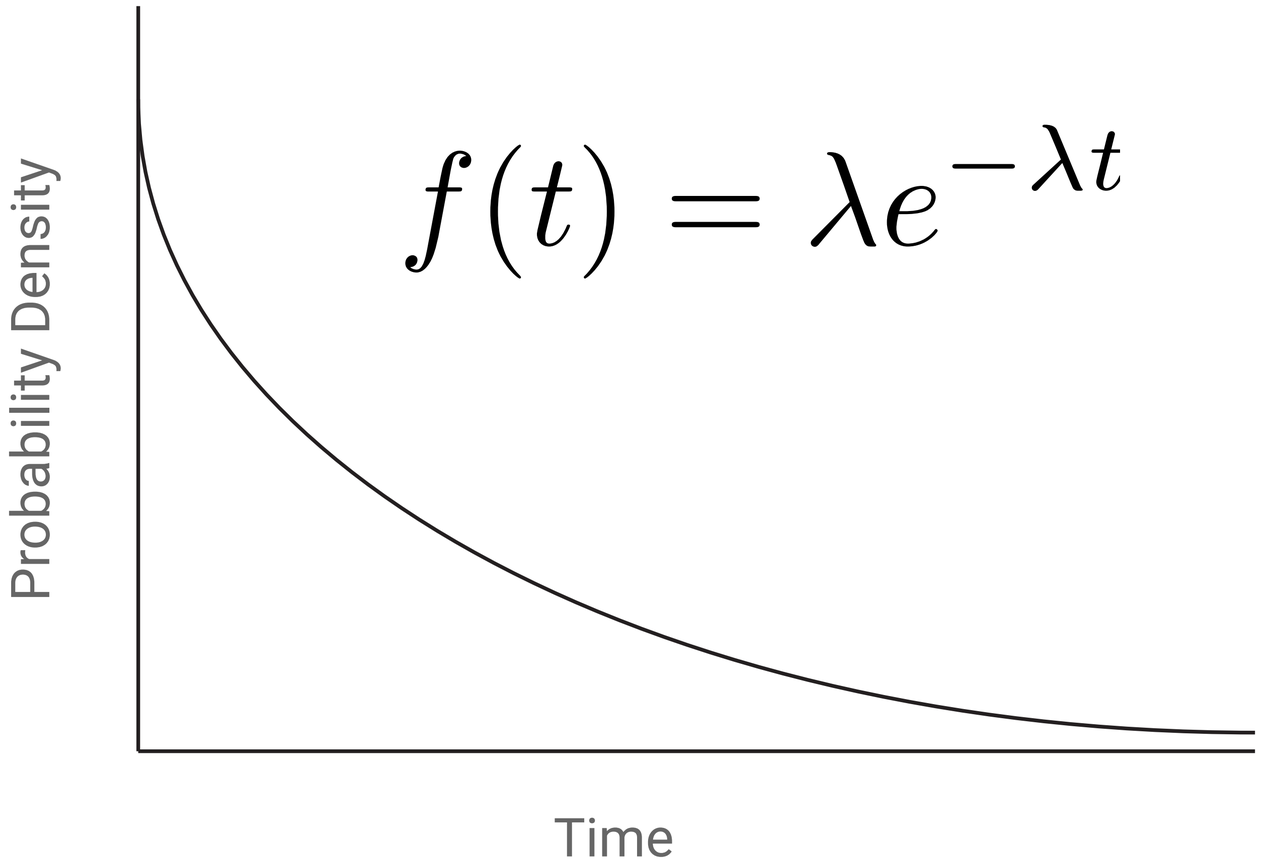
They allow us to compute the probability of changing from one state to another over branch length *v*

Continuous time Markov chain

Nucleotide substitutions (events) occur at a constant rate (λ)



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

Rate matrix

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

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

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

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

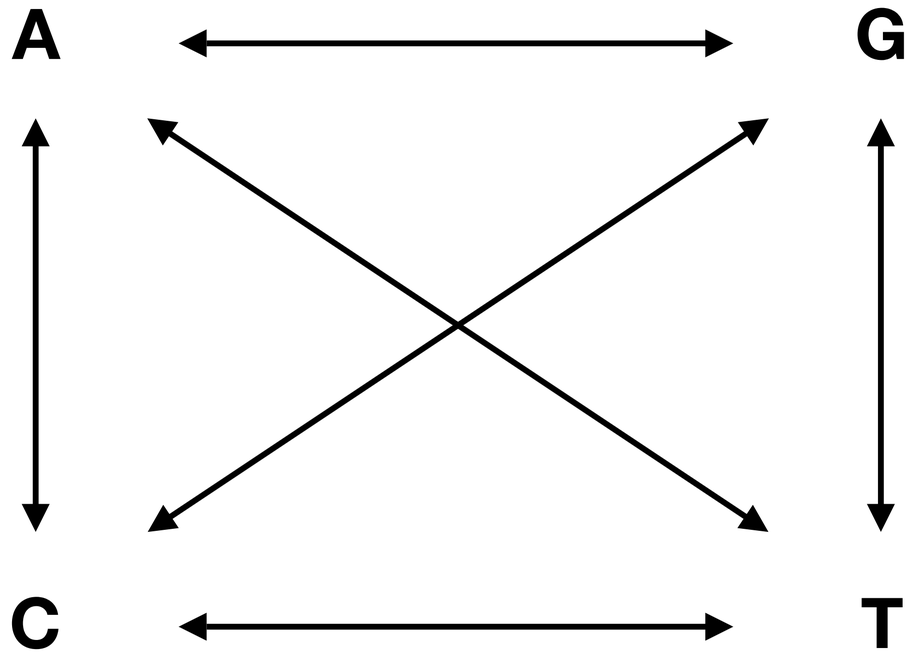
This is the Jukes-Cantor (1969) or JC69 model

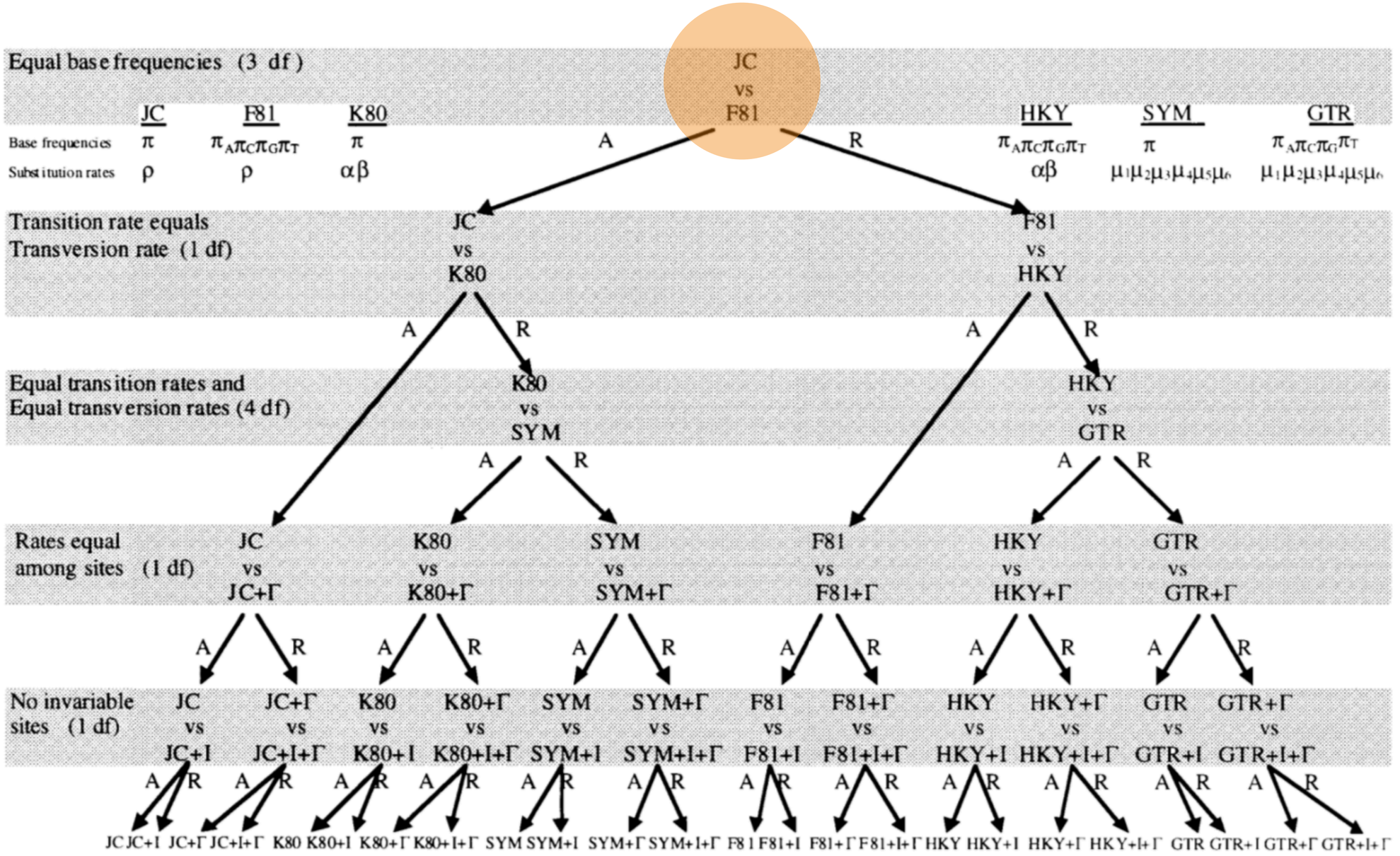
Rate matrix

An alternative representation of the JC model

Arrows represent (symmetric) rates of change between states

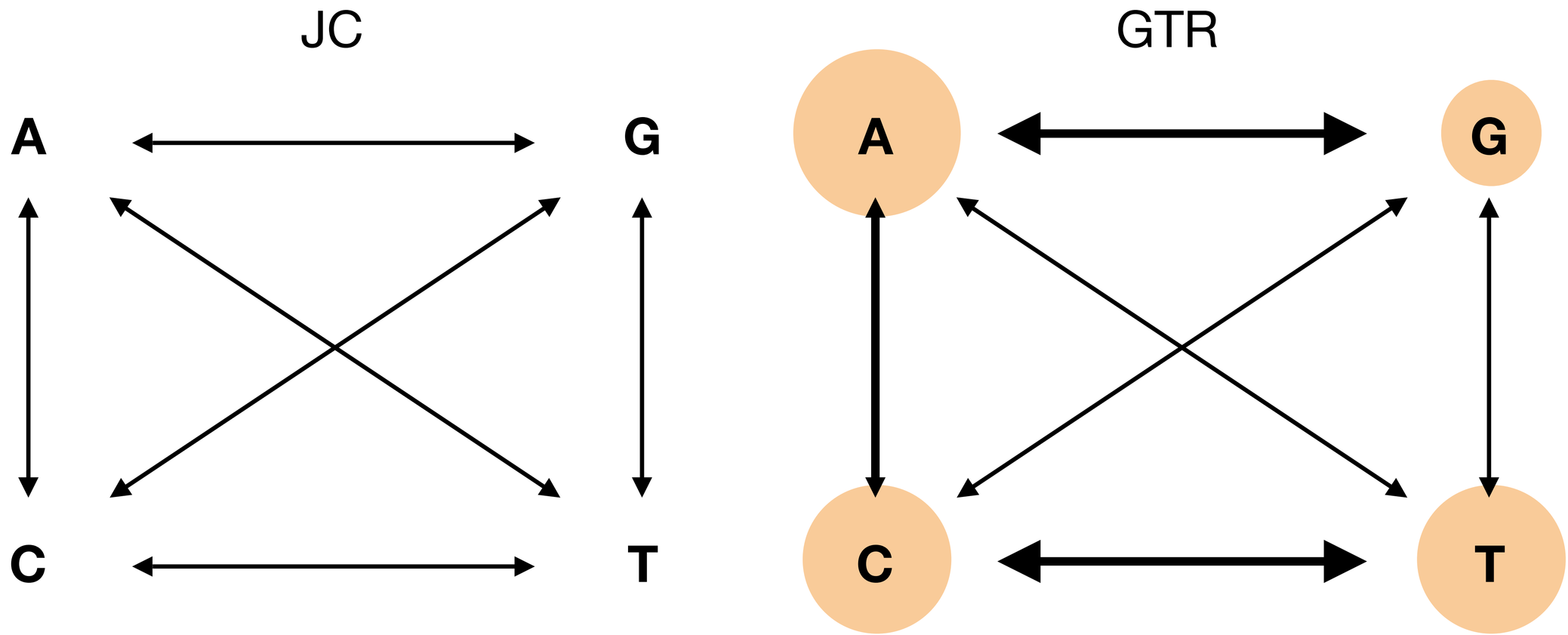
The model assumes base frequencies are equal



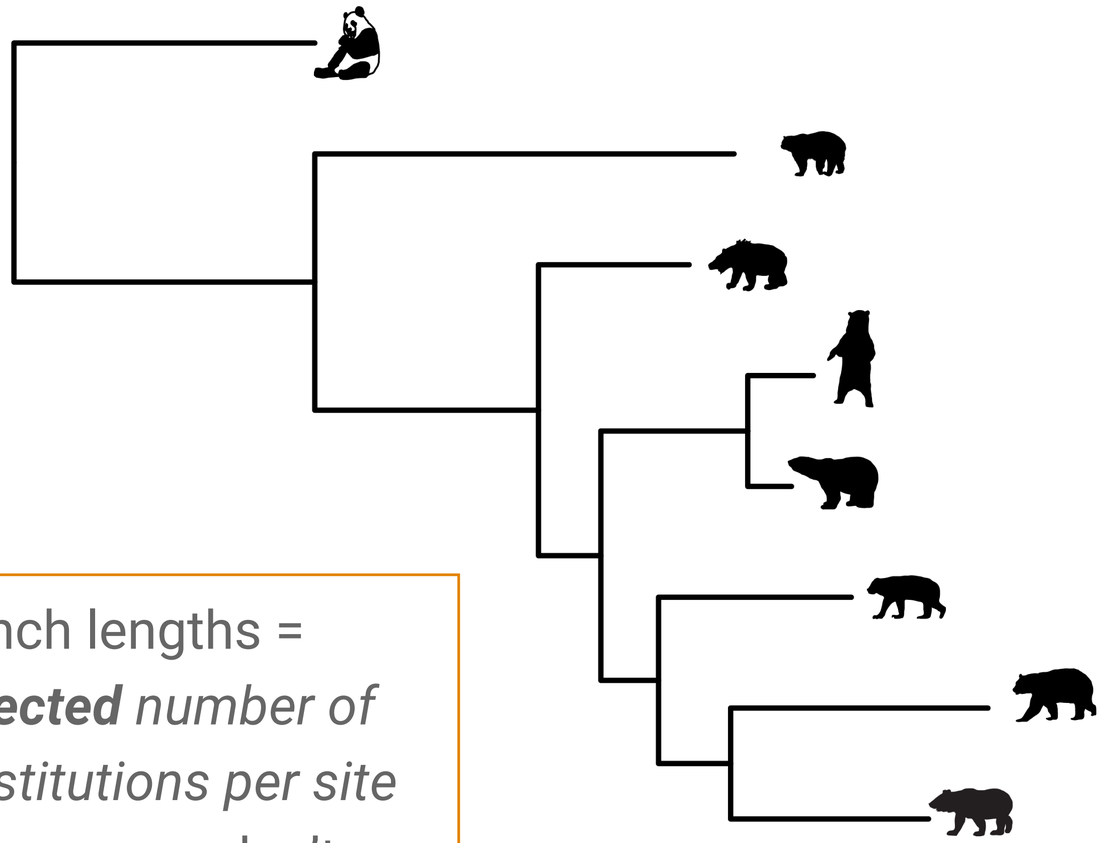


Posada & Crandall (1998)

The JC vs. GTR models



Branch lengths



Branch lengths = ***expected*** *number of substitutions per site* (because we don't observe every change)

Branch lengths are a product of rate and time

Without temporal information we can only measure relative genetic distance

Exercise

Jukes-Cantor model transition probability applet

(skip for now)

Written by Paul Lewis

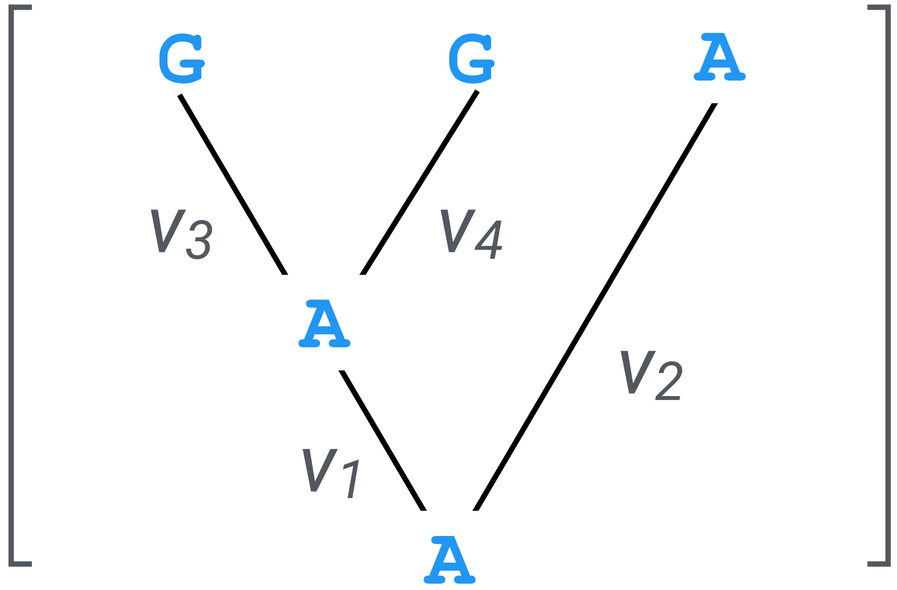
Felsenstein’s pruning algorithm

(skip for now)

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

Computing the probability of the observed data

$Pr =$

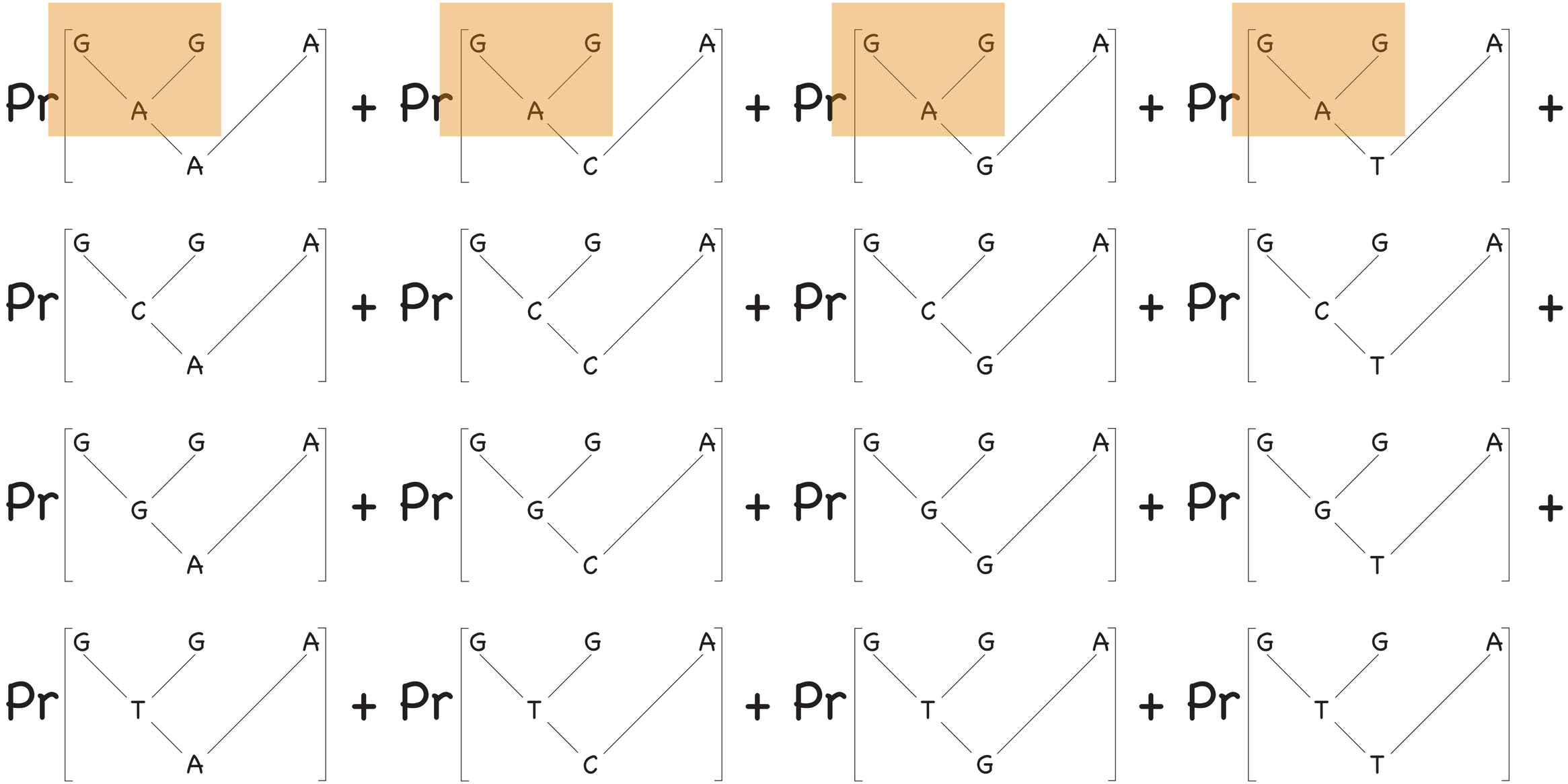


Just suppose for now we know the ancestral states at internal nodes

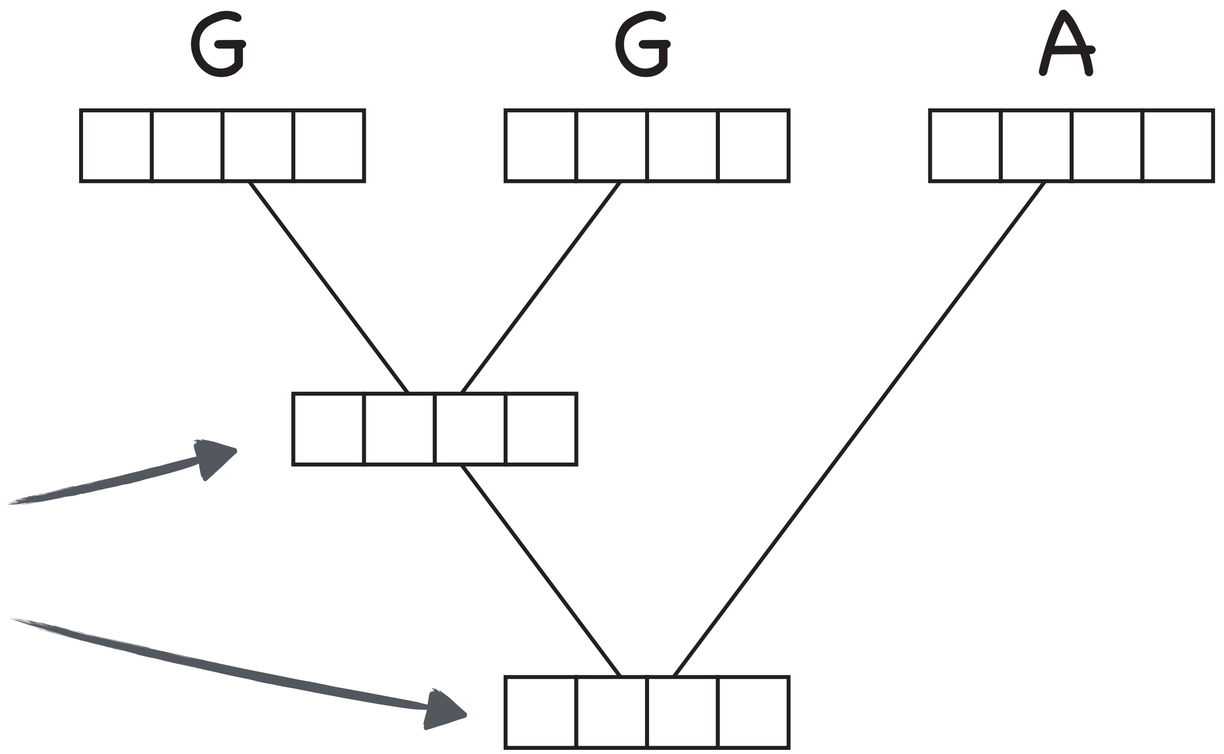
$\pi_A \times P_{AA}(v_1) \times P_{AA}(v_2) \times P_{AG}(v_3) \times P_{AG}(v_4)$

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

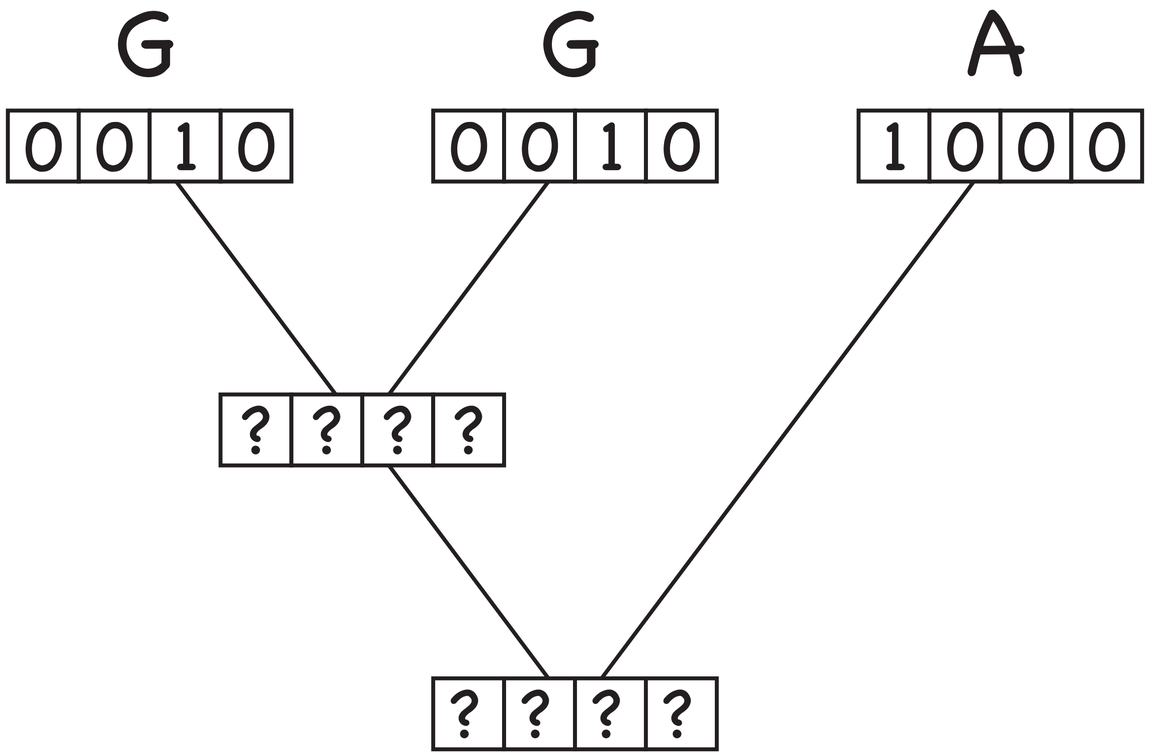
π_i — stationary frequencies


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]

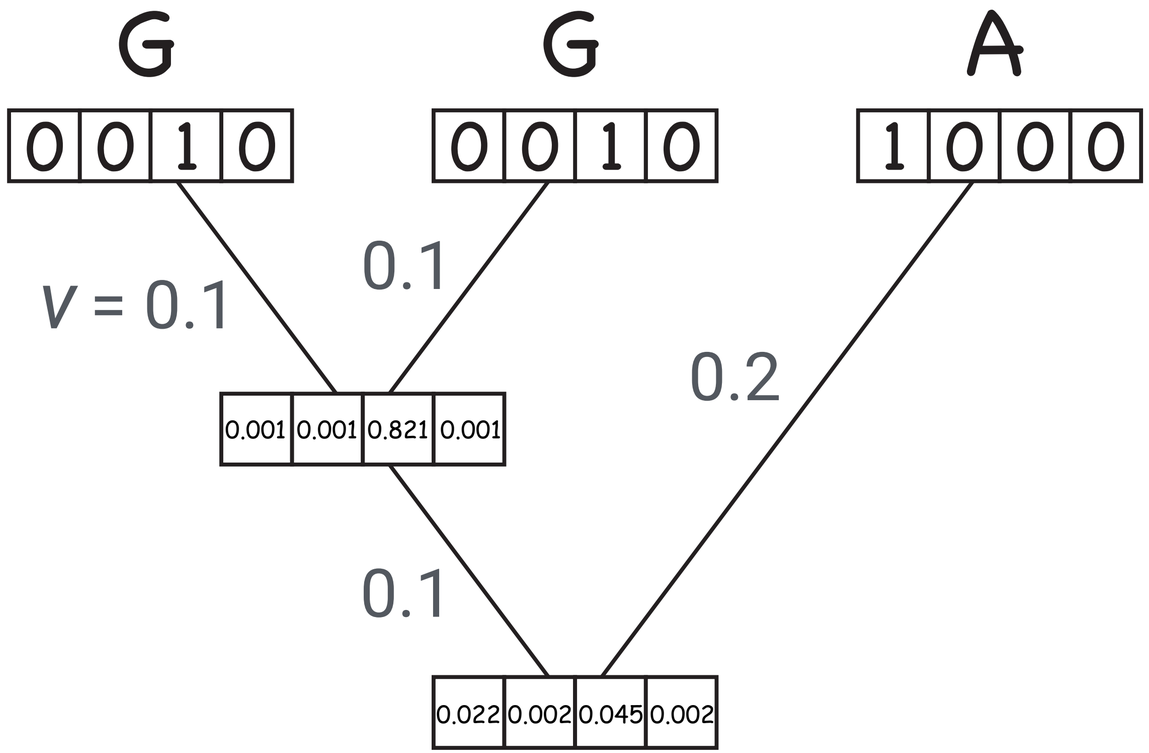
We’re going to calculate these probabilities based on the Pr calculated for the tips / nodes below



Felsenstein (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach.

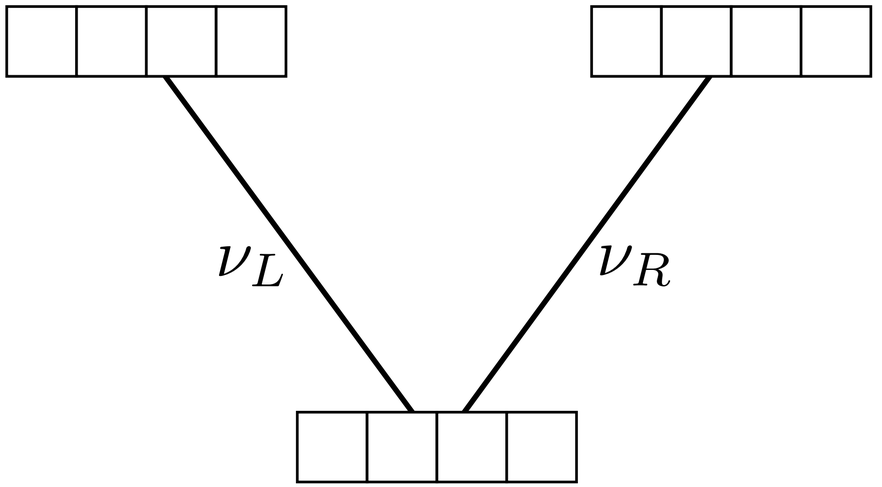


Next, we calculate the probabilities at each internal node



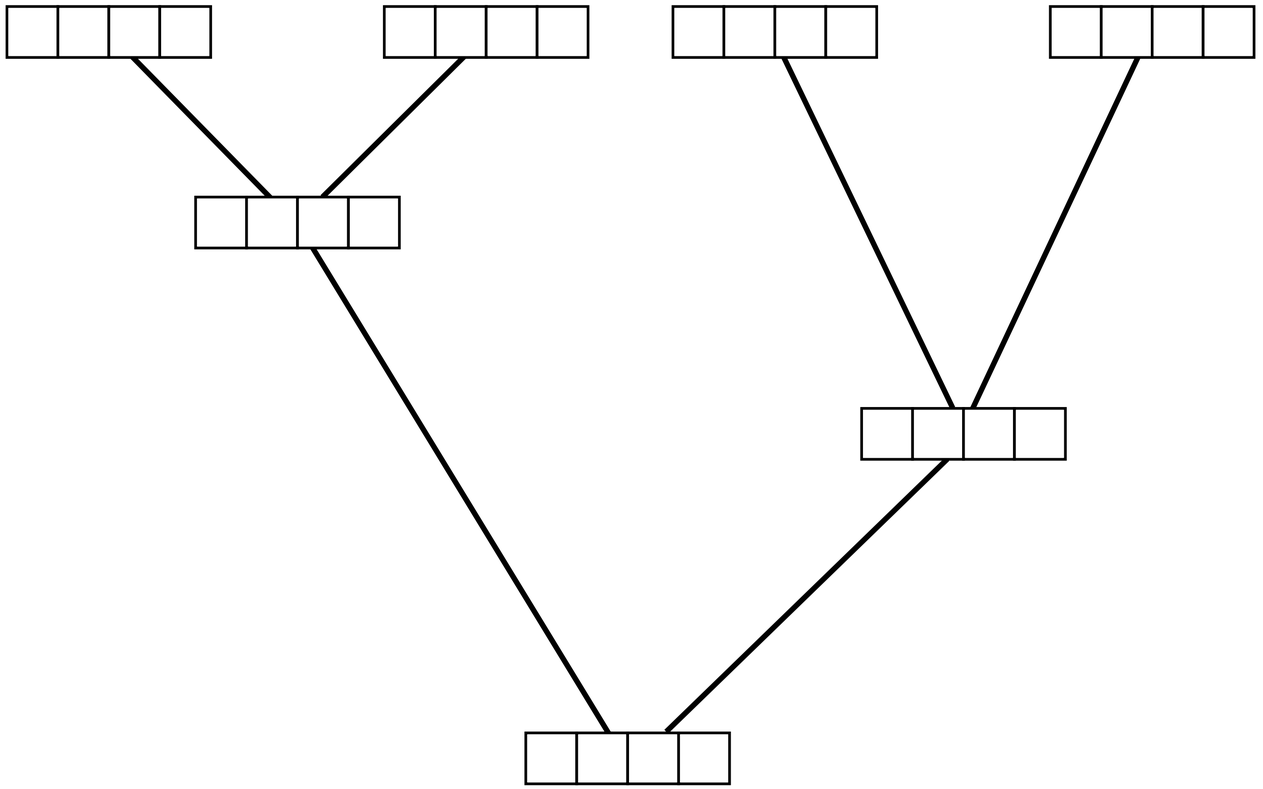
The values at internal nodes are *likelihoods*

Likelihood = Pr(sequence data | (sub)tree, model)



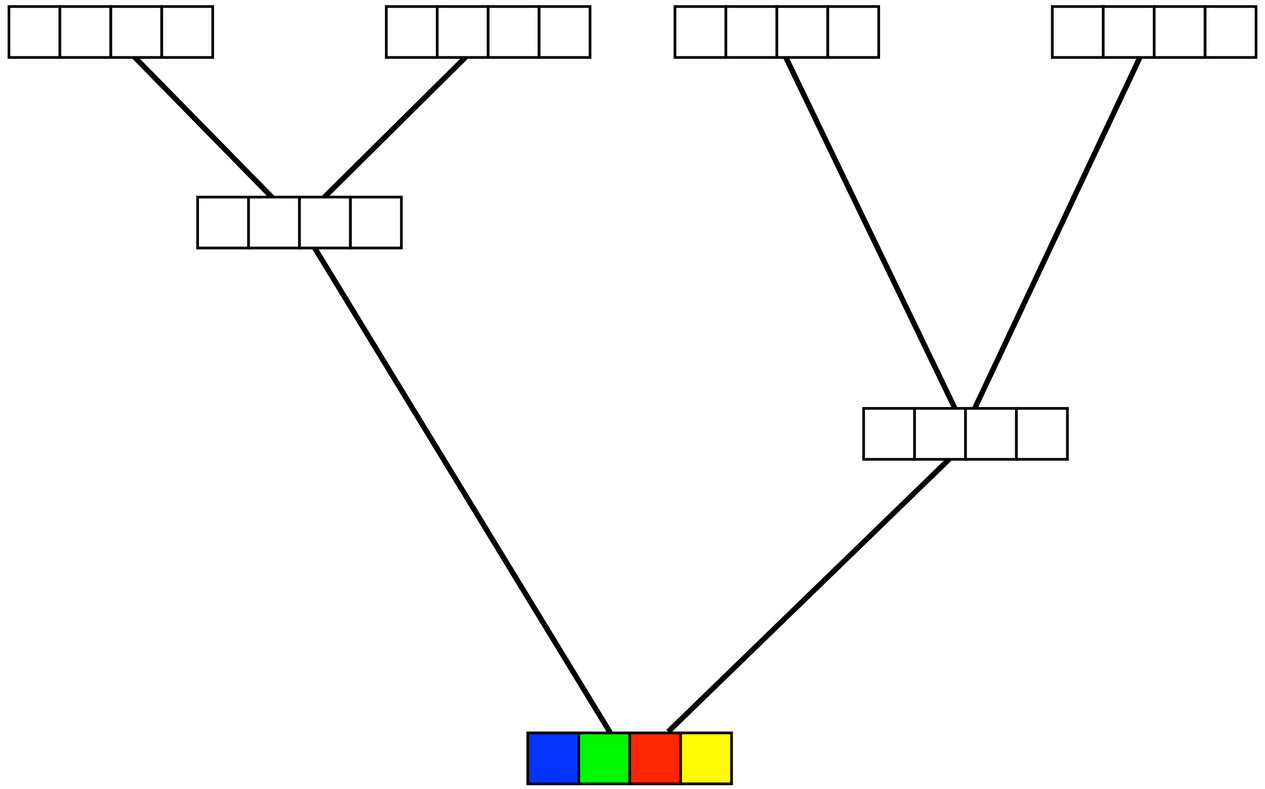
For each node, we take into account all possible changes, along both branches

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



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

n.b. This is the probability of a **single site** in your alignment!



Finally, we need to calculate & multiply this Pr across all sites in our alignment

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

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

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

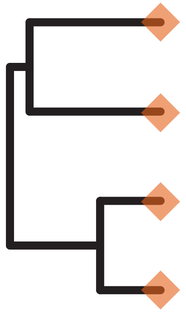
Another nice description of the pruning algorithm: *Harmon* (2019) Phylogenetic Comparative Methods, Chapter 8

Components used to infer trees

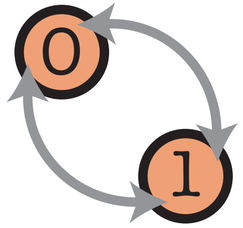
without yet considering time

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

data
sequences or characters

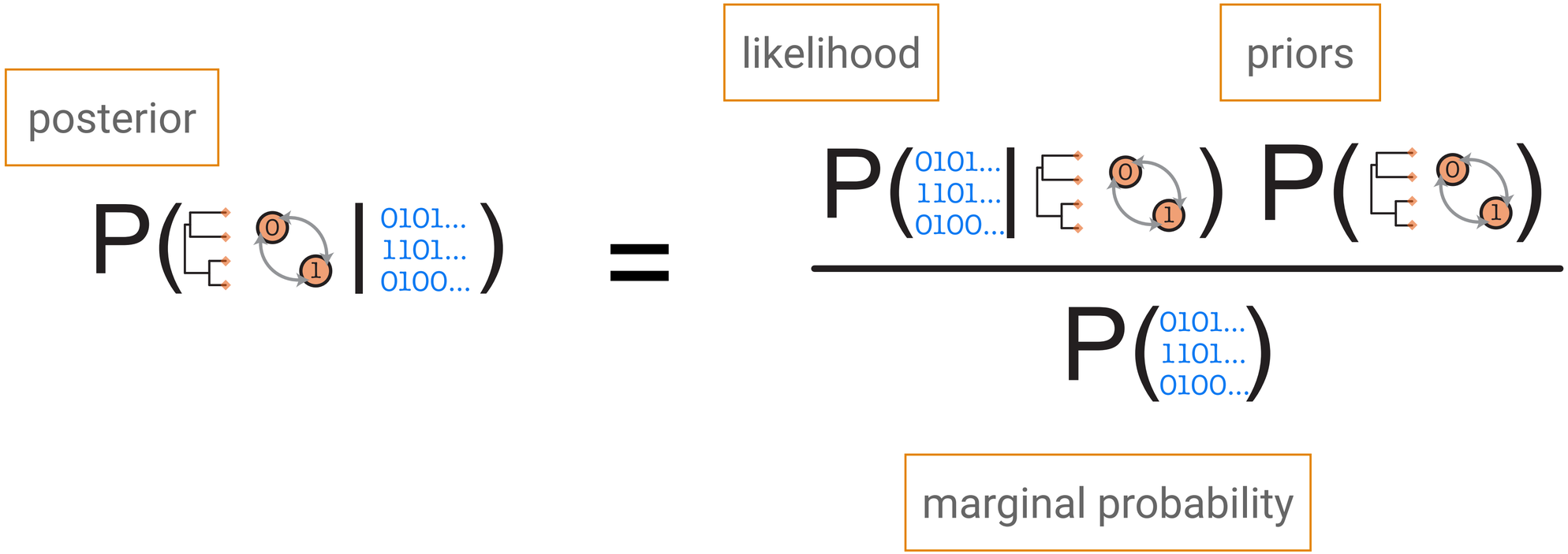


tree
topology and branch lengths



substitution model

Bayesian tree inference

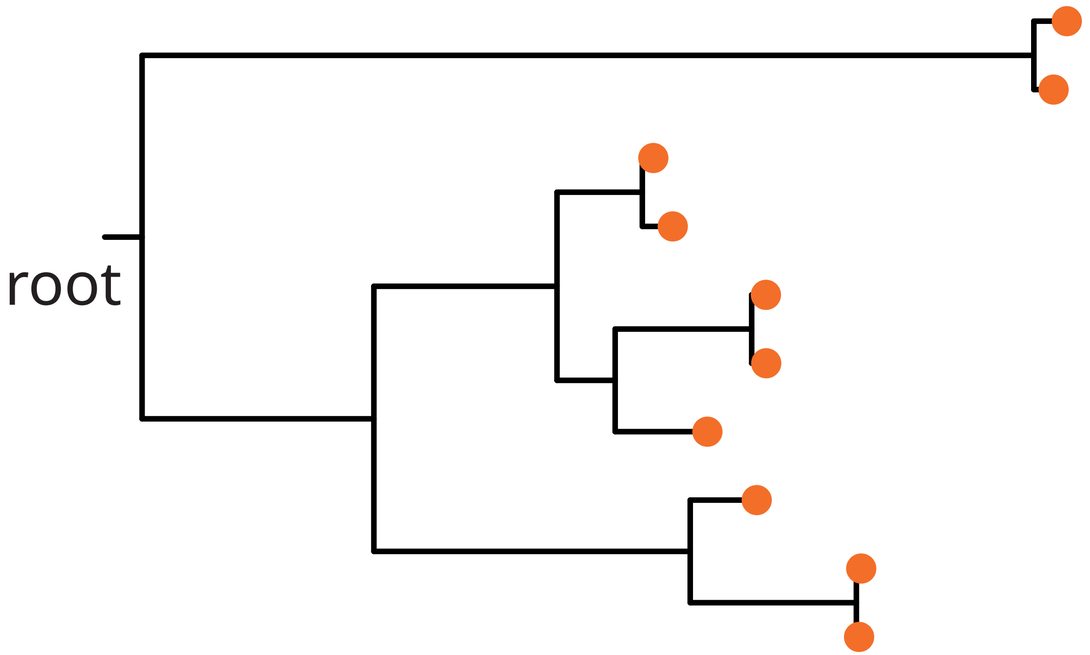


Short break



Introduction to phylogenetic dating

Molecular (or morphological) characters are not independently informative about time





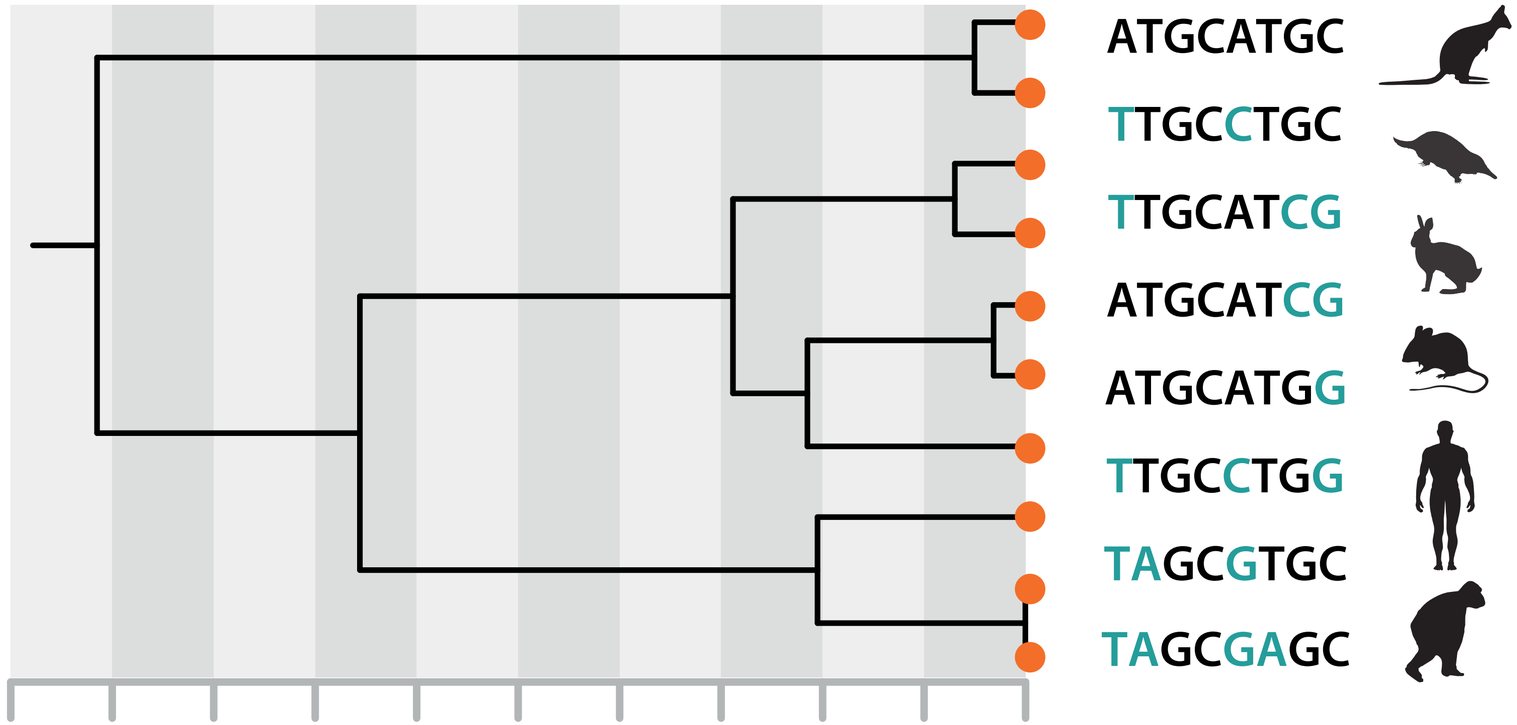
Slow rate, long interval **or** fast rate, short interval?

Goal: to disentangle evolutionary rate and time

branch lengths = genetic distance

$v = rt$

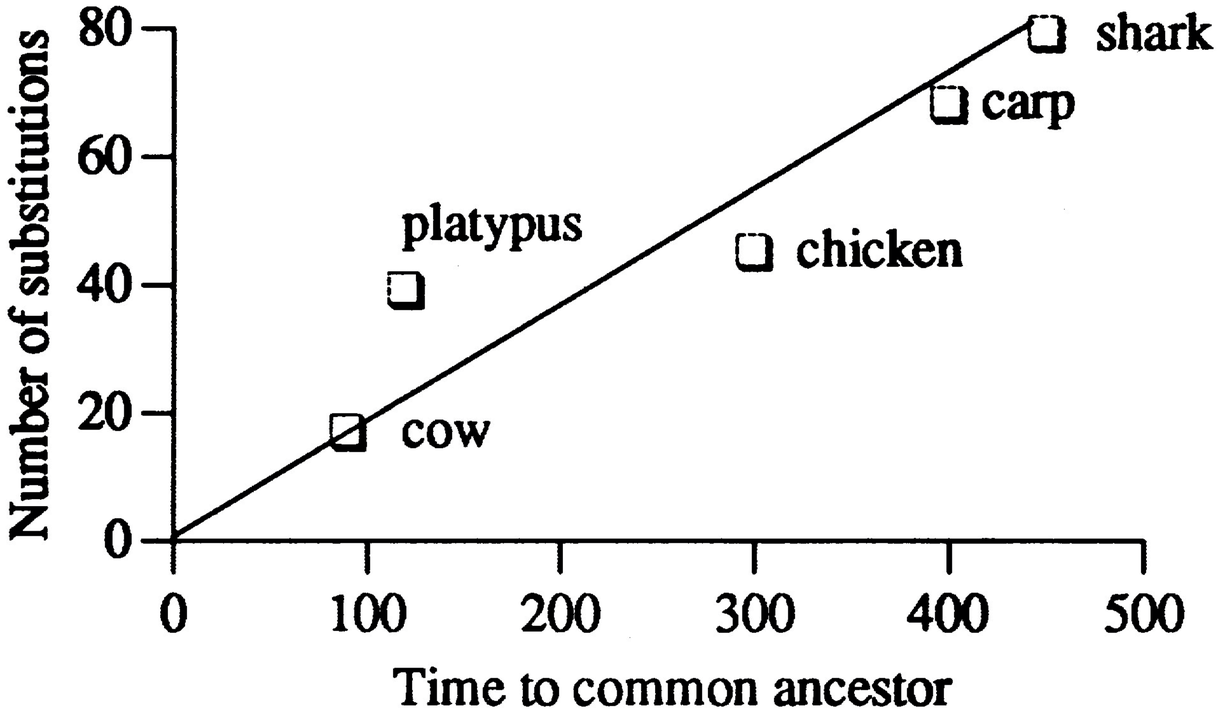
Molecular (or morphological) characters are not independently informative about time



branch lengths = time

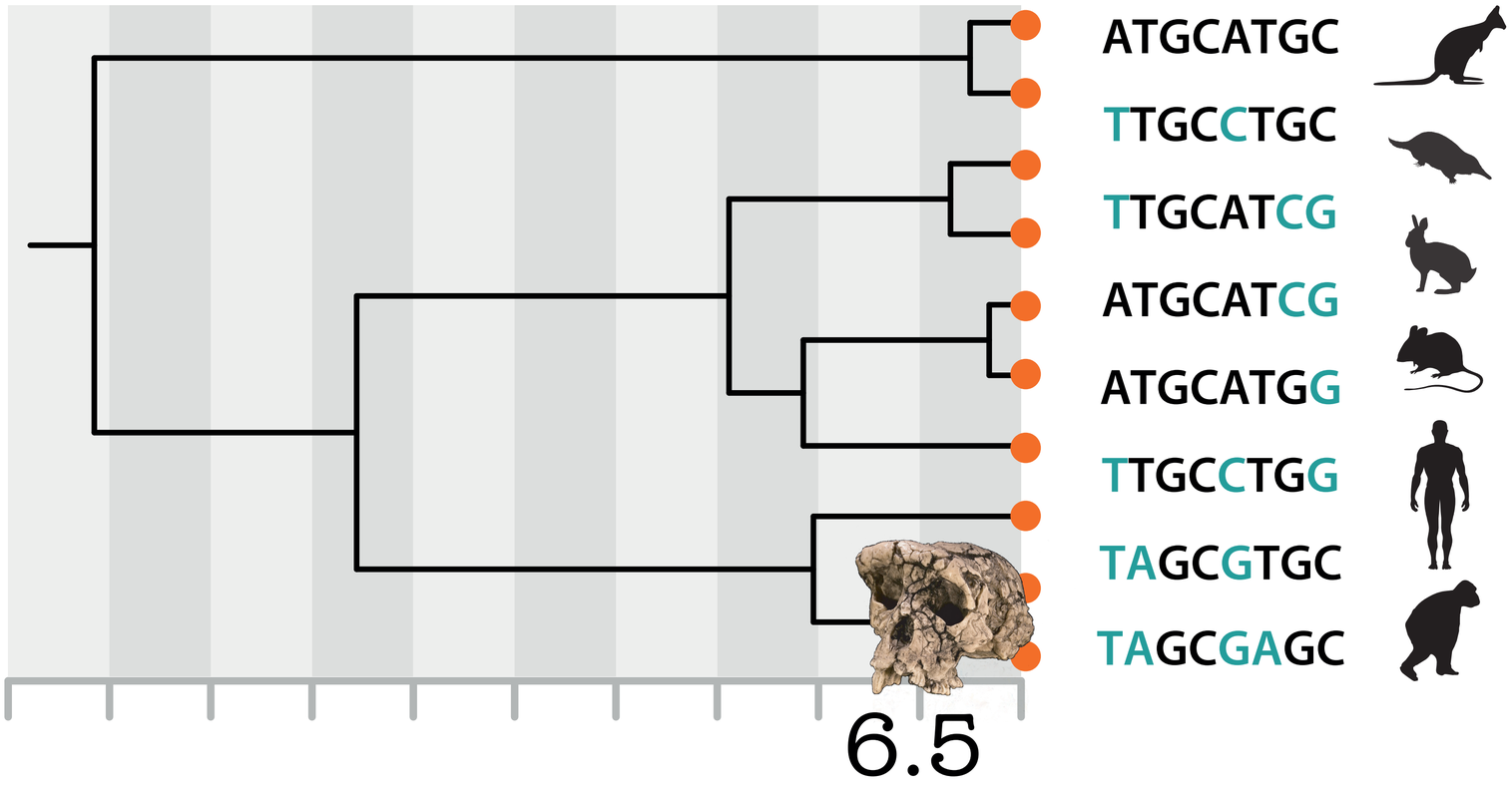
Goal: to disentangle evolutionary rate and time

The molecular clock hypothesis



Molecules as documents of evolutionary history *Zuckerkandl & Pauling* (1965)
A history of the molecular clock *Morgan* (1998)

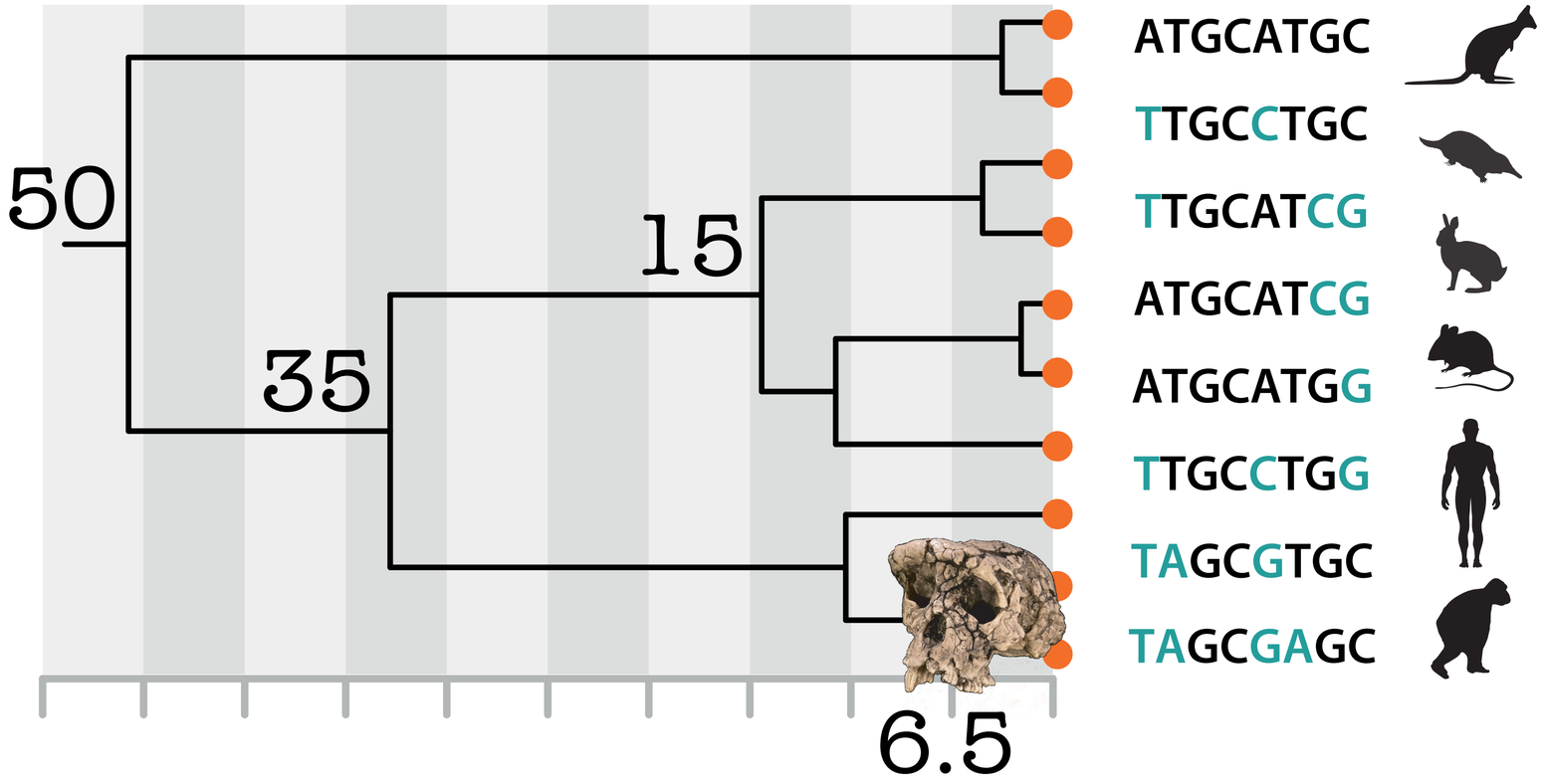
Calibrating the substitution rate



branch lengths = time

Temporal evidence of divergence for one species pair let's us calibrate the average rate of molecular evolution

Calibrating the substitution rate

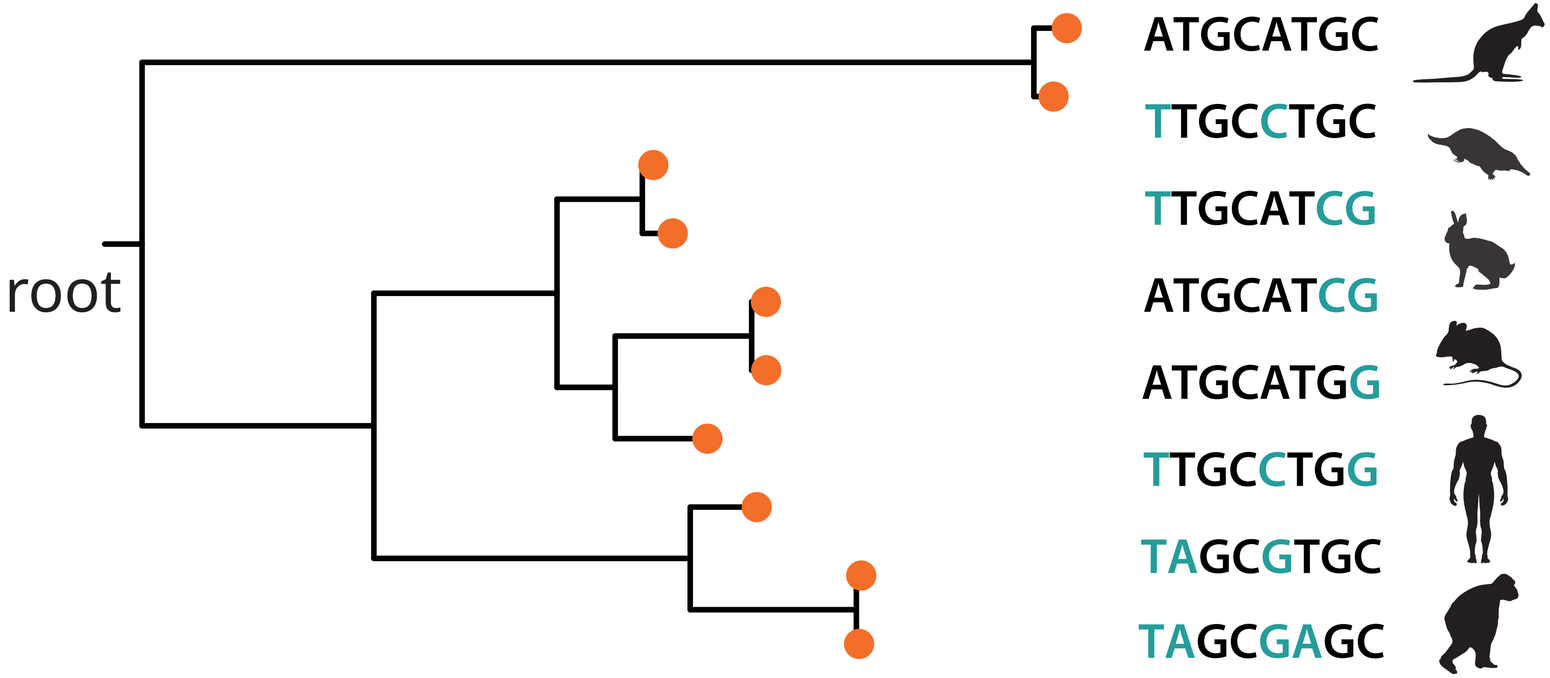


branch lengths = time

We can use this rate to extrapolate the divergence times for other species pairs

Molecular dating: challenges

Rate and time are not **fully identifiable**!

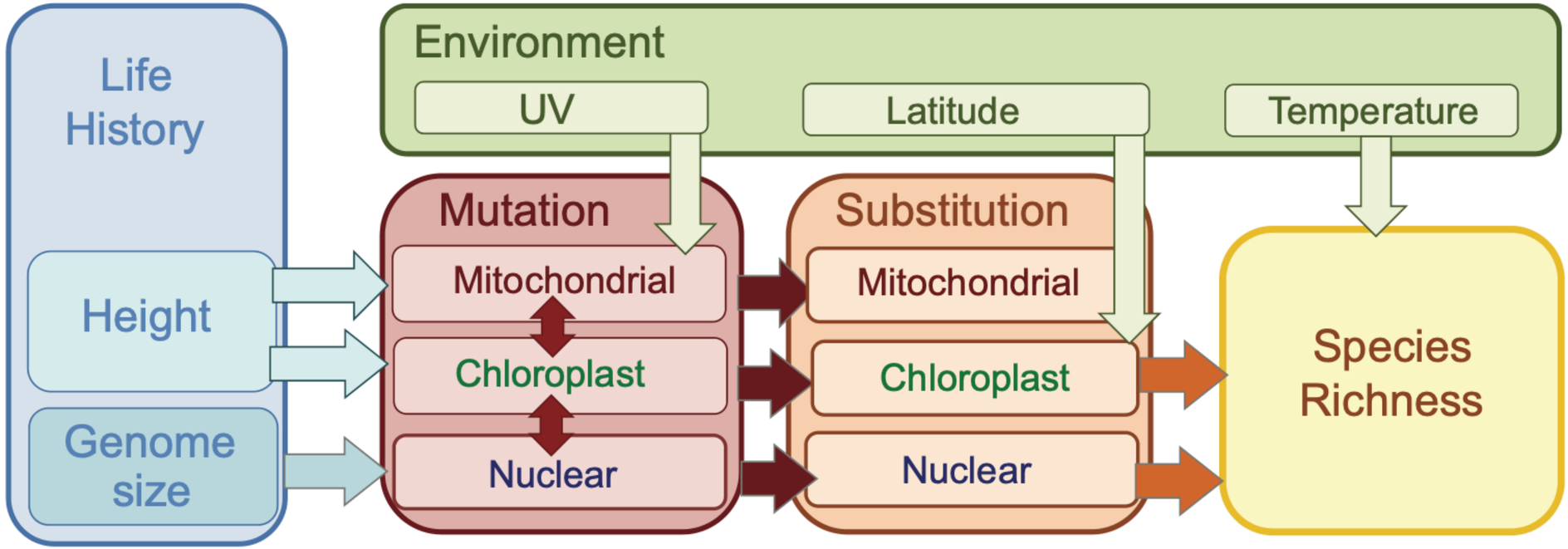


branch lengths = genetic distance

$v = rt$

Molecular dating: challenges

Many variables contribute to **variation in the substitution rate**



Bromham et al. (2015)

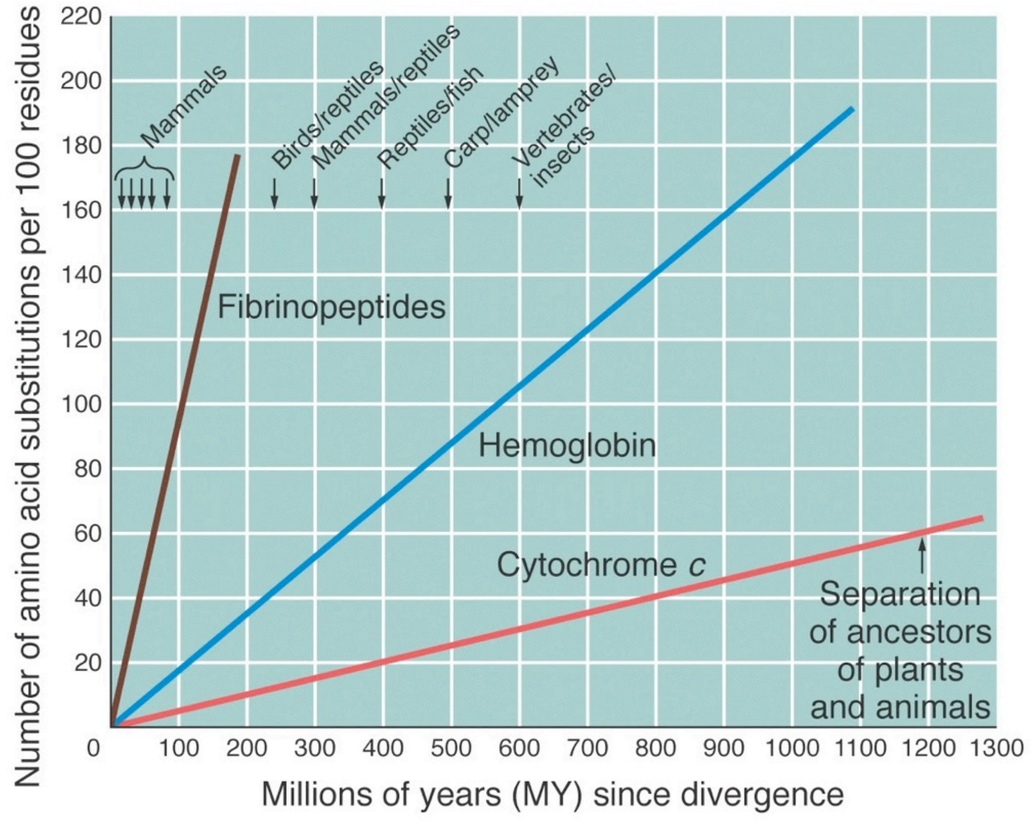
Molecular dating: challenges

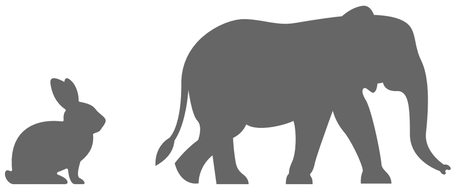
Many variables contribute to **variation in the substitution rate**

The molecular clock is not constant

Rates vary across:

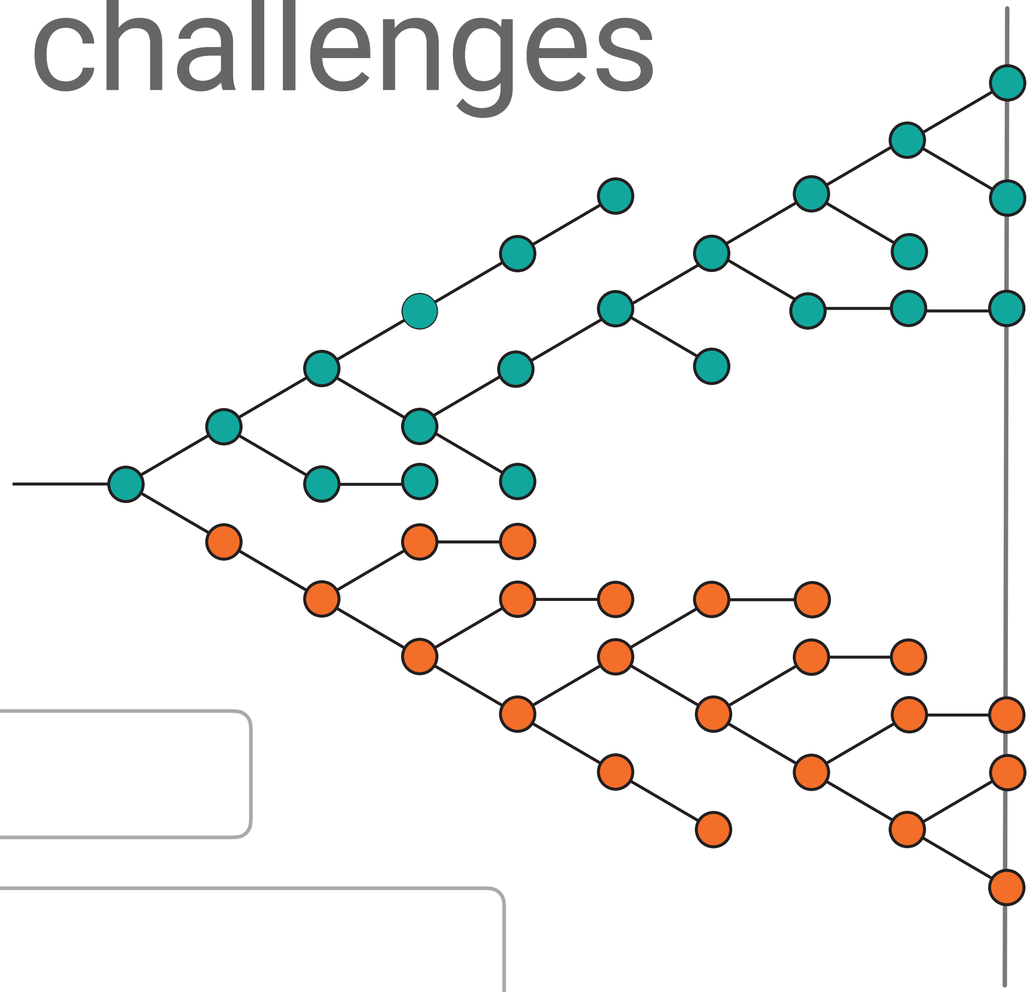
- taxa
- time
- genes
- sites within the same gene





Molecular dating: challenges

Calibrations are imprecise



Molecular evolution:

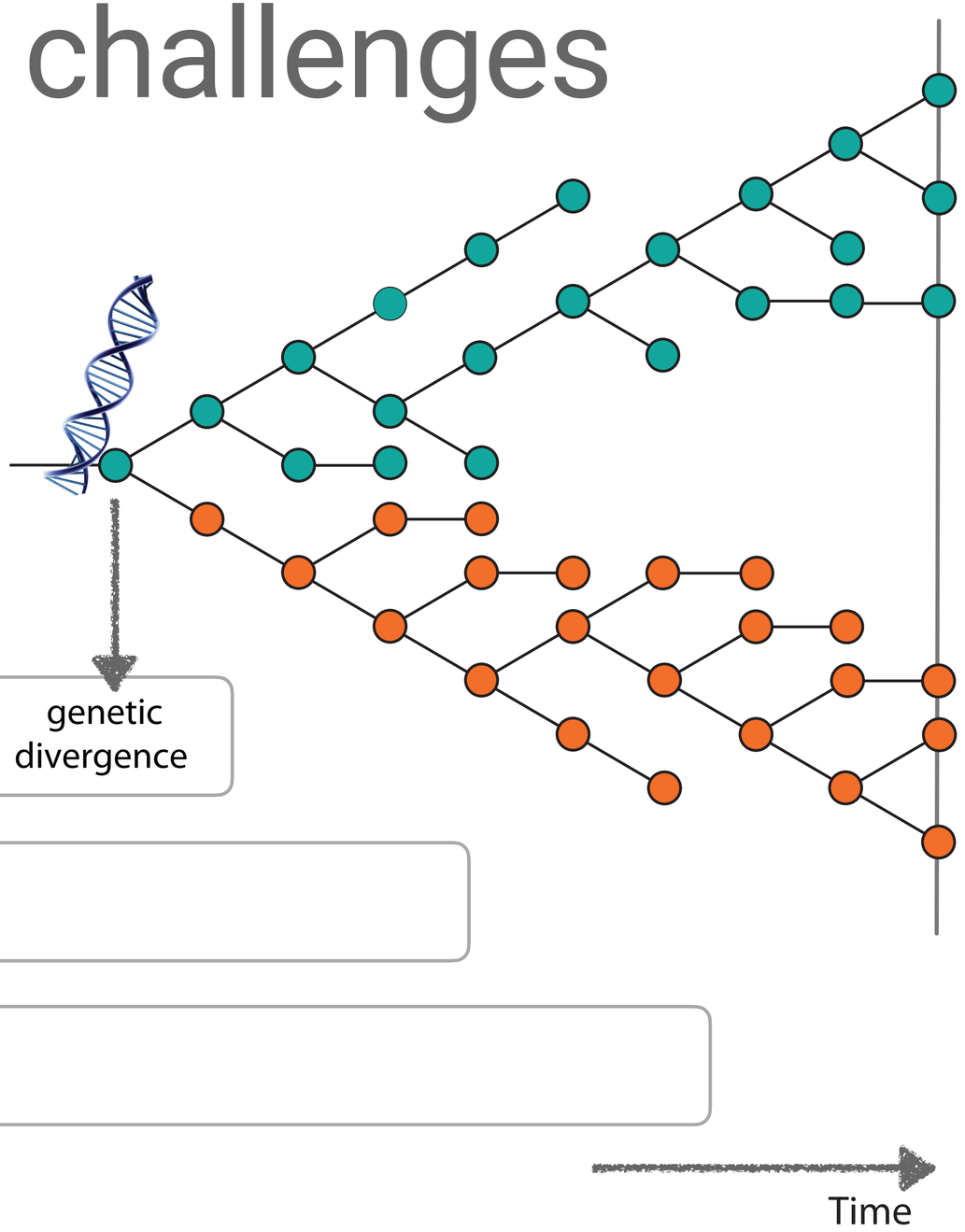
Morphological evolution:

Fossil preservation:

Time

Molecular dating: challenges

Calibrations are imprecise



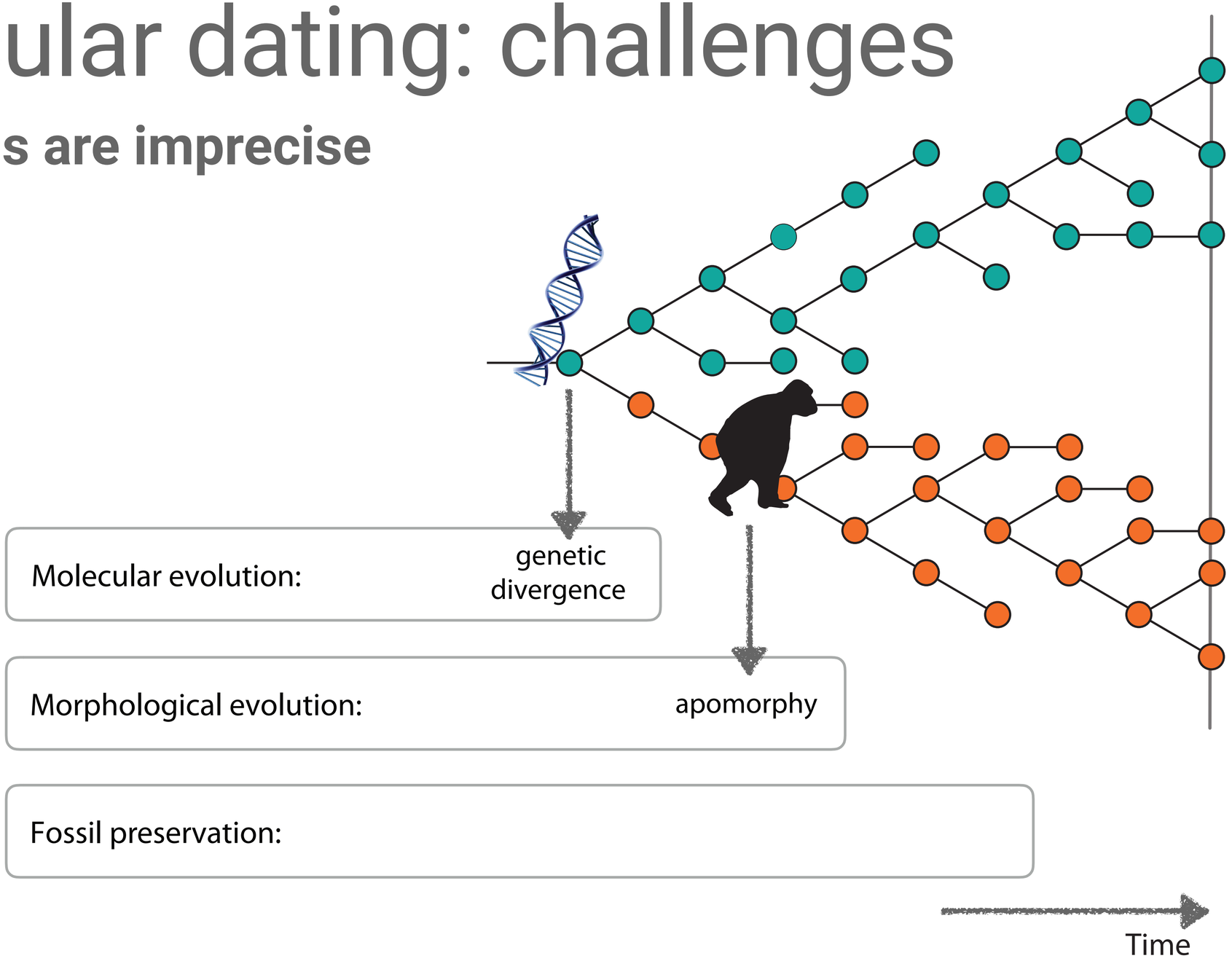
Molecular evolution:

Morphological evolution:

Fossil preservation:

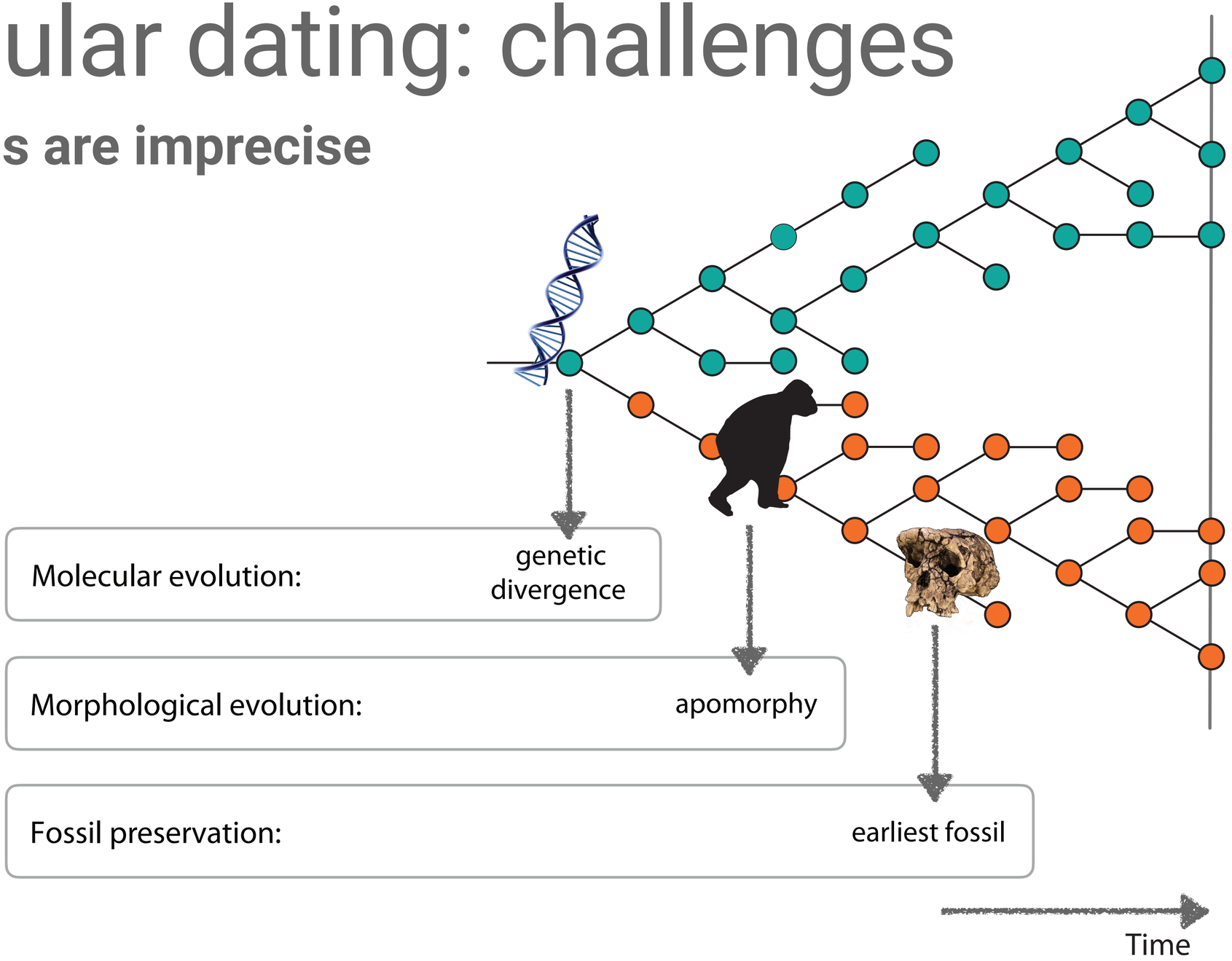
Molecular dating: challenges

Calibrations are imprecise



Molecular dating: challenges

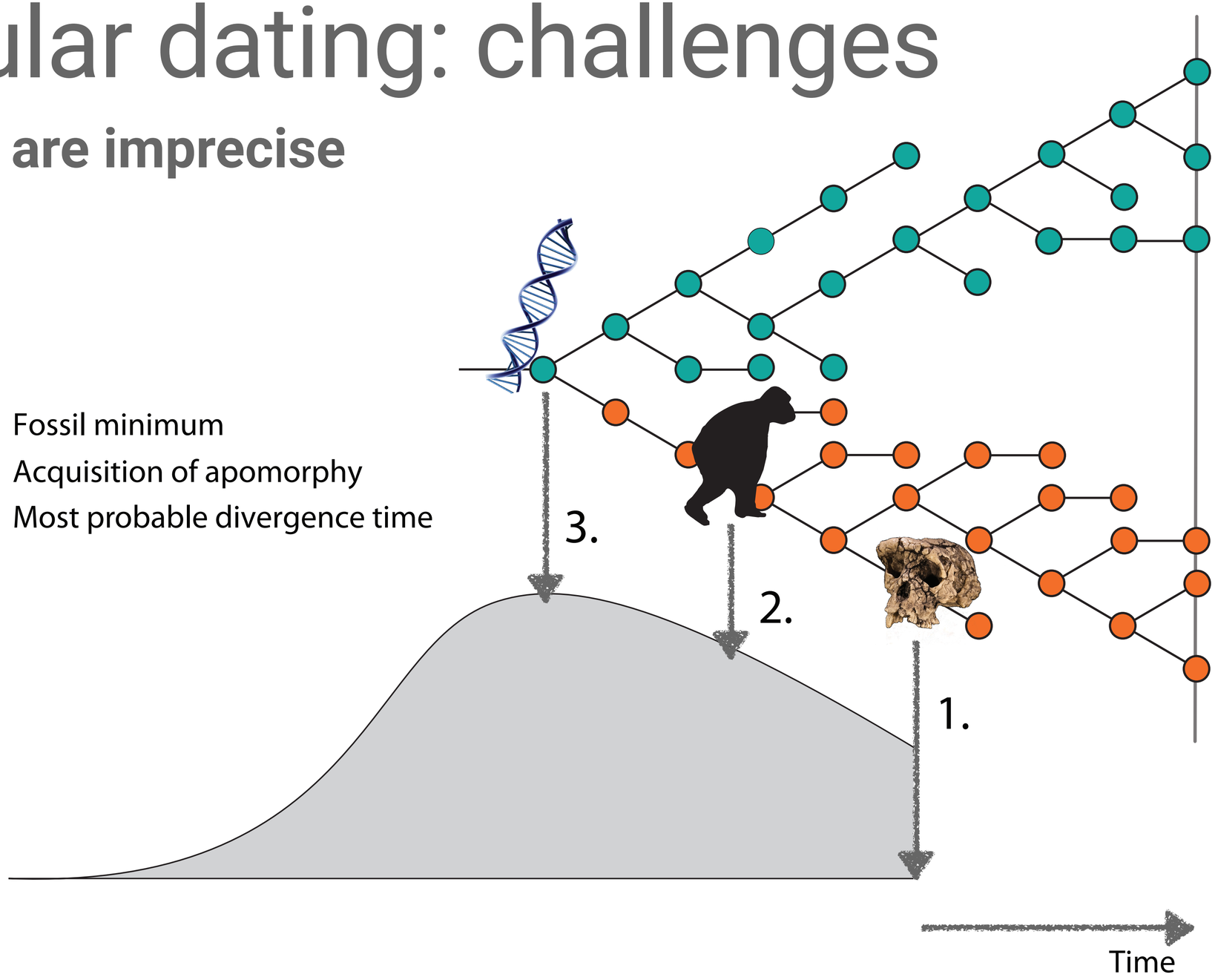
Calibrations are imprecise



Molecular dating: challenges

Calibrations are imprecise

1. Fossil minimum
2. Acquisition of apomorphy
3. Most probable divergence time



Molecular dating: challenges

Summary

1. Rate and time are not fully identifiable

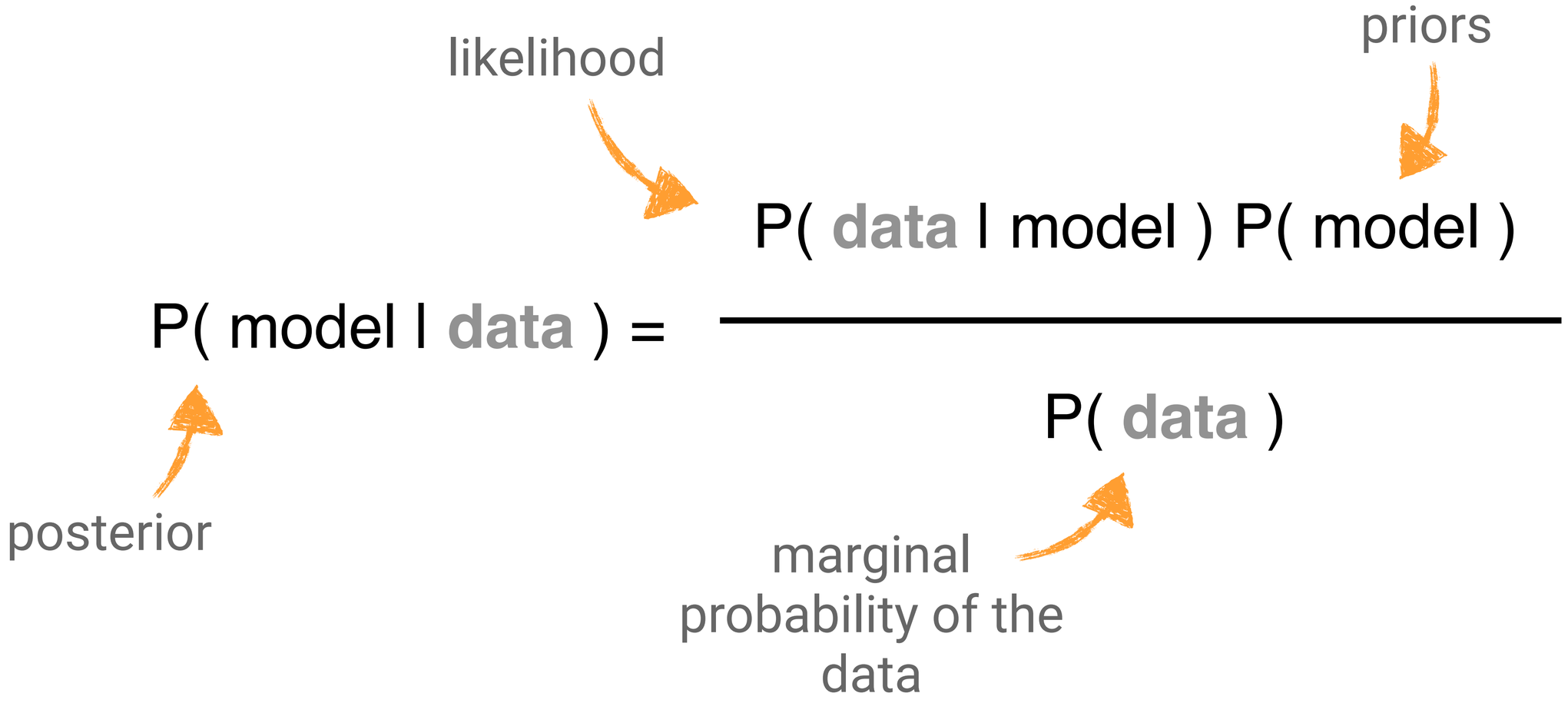
2. The substitution rate varies

3. Calibrations are imprecise

→ we need a flexible statistical framework that deals well with uncertainty!

Bayesian **time tree** inference

We use a Bayesian framework



Bayesian divergence time estimation

The data

and / or

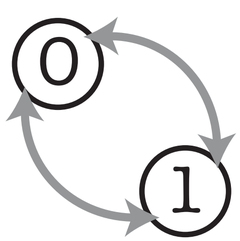
```
0101... ATTG...
1101... TTGC...
0100... ATTC...
```

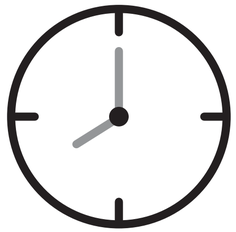


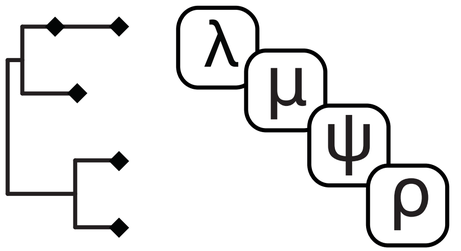
phylogenetics characters

sample ages

3 model components







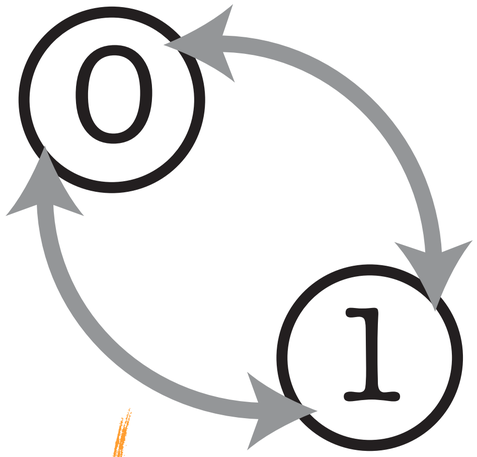
substitution model

clock model

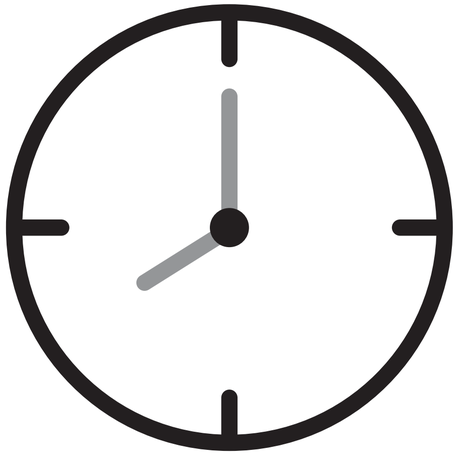
tree and tree model

Understanding the tripartite approach to Bayesian divergence time estimation
Warnock, Wright. (2020)

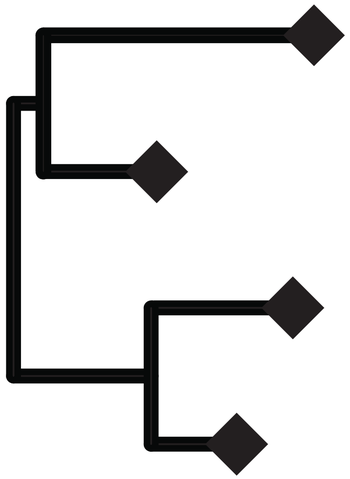
substitution model



clock model

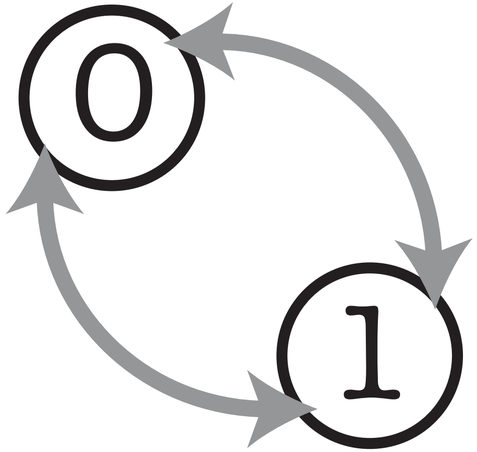


tree model

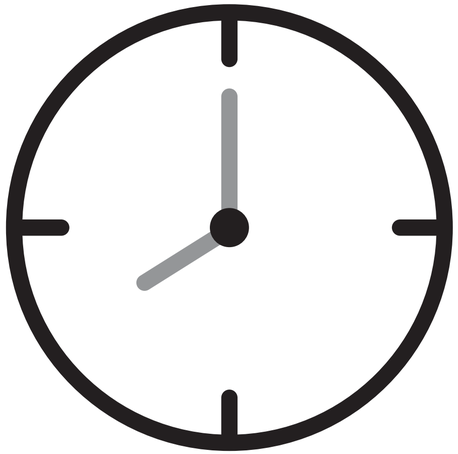


How likely are we to observe a change between character states? e.g., A → T

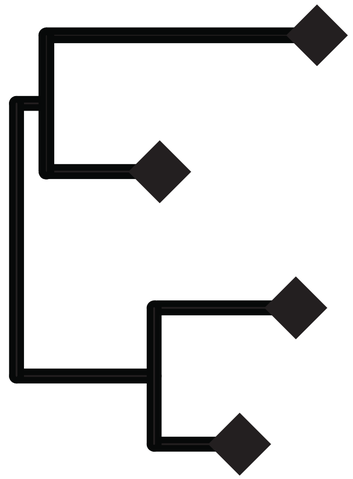
substitution model



clock model

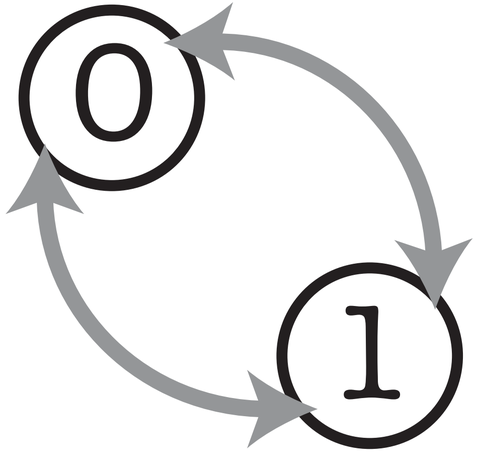


tree model

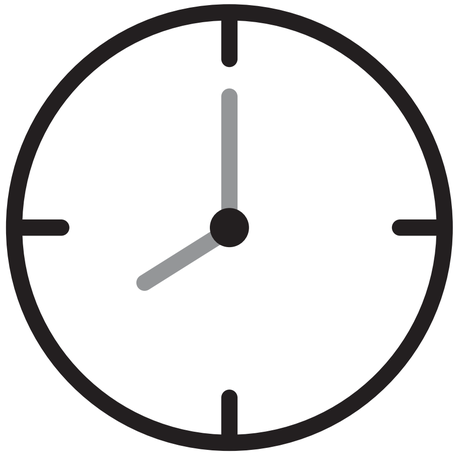


How have rates of evolution varied (or not) across the tree?

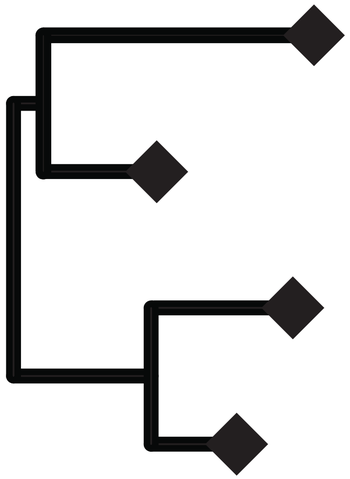
substitution model



clock model



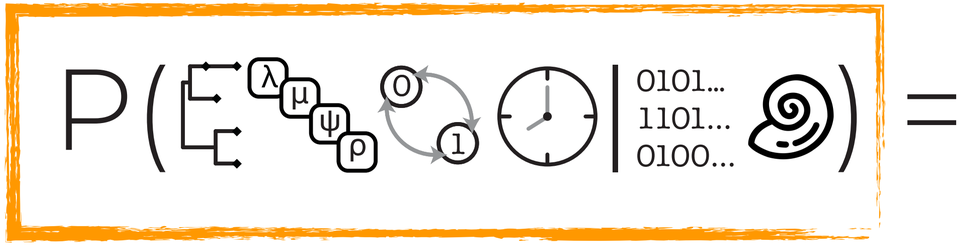
tree model



How have species originated, gone extinct and been sampled through time?

Bayesian divergence time estimation

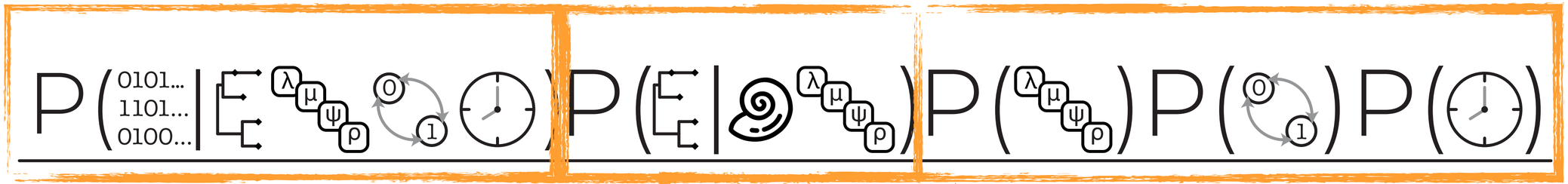
posterior

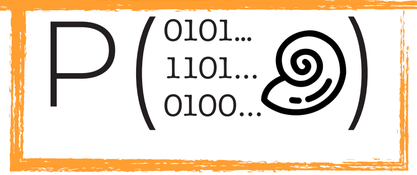


likelihood

probability of the time tree

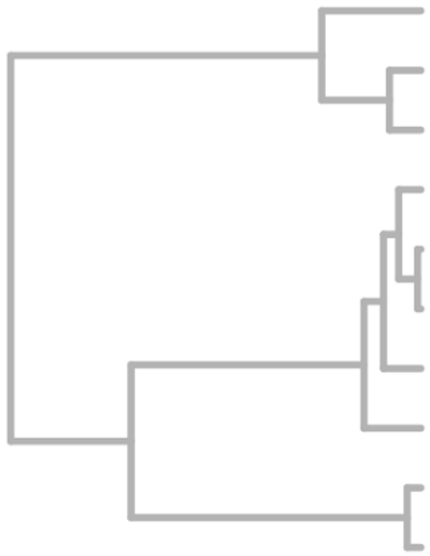
priors





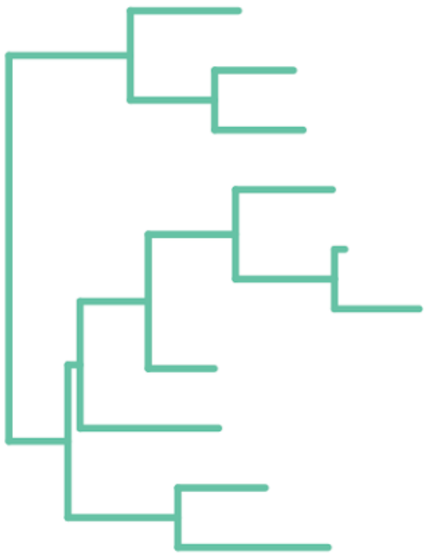
marginal pr of the data

Calculating the likelihood



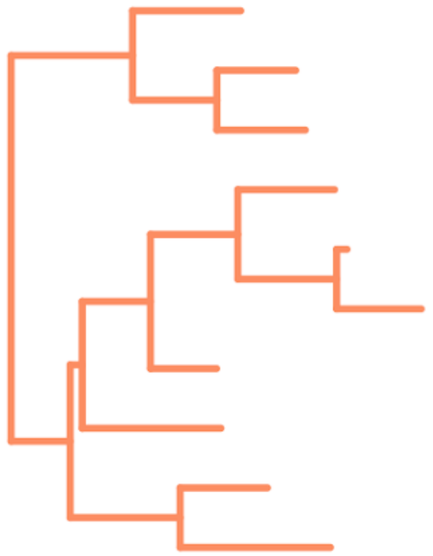
Time (year)

X



Clock rate (subs/site/year)

=



Genetic distance (subs/site)

Prior

Prior

Likelihood

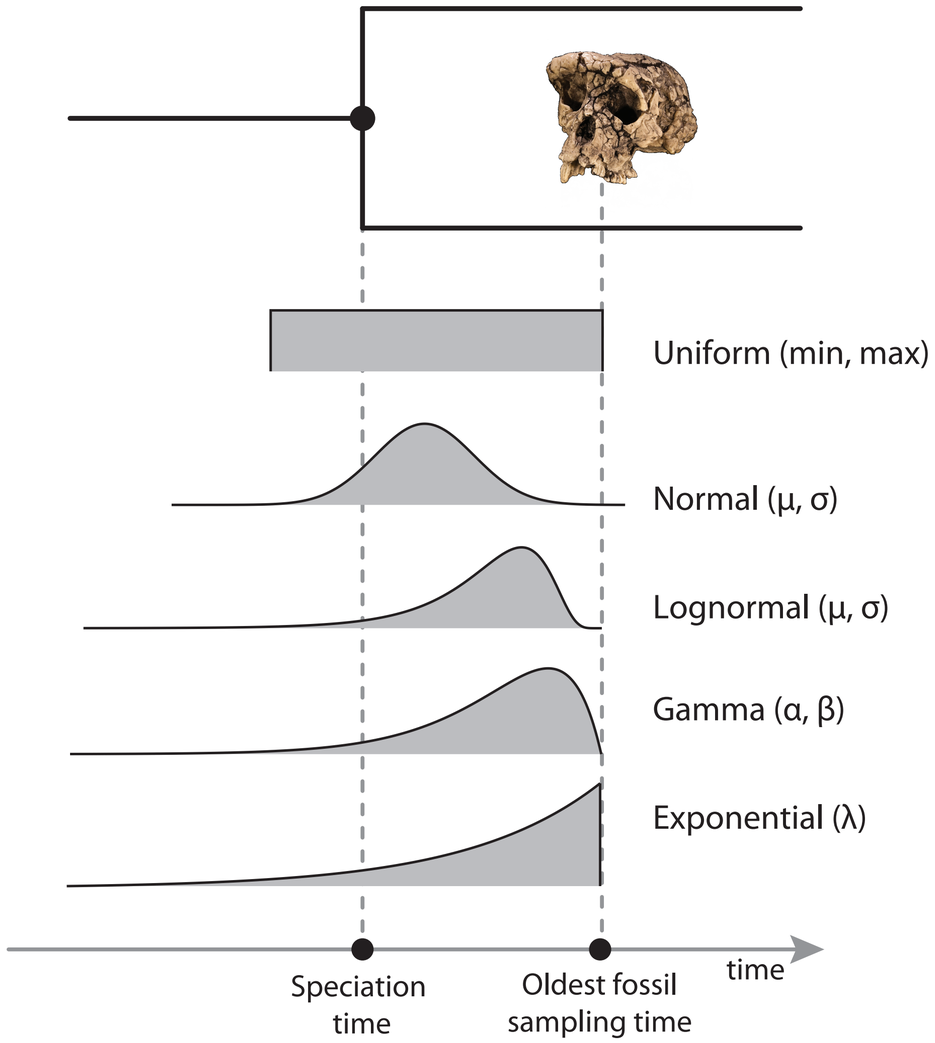


Based on the calibration times we can estimate the rate over time



Once we have the rate we can transform evolutionary rates into genetic distance

Node dating



The tree includes extant (living) species only

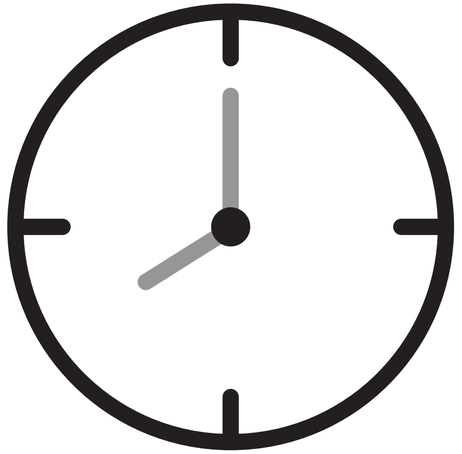
We can use a calibration density to constrain internal node ages

We typically use a birth-death process model to describe the tree generating process

Adapted from *Heath* (2012) Sys Bio

clock model

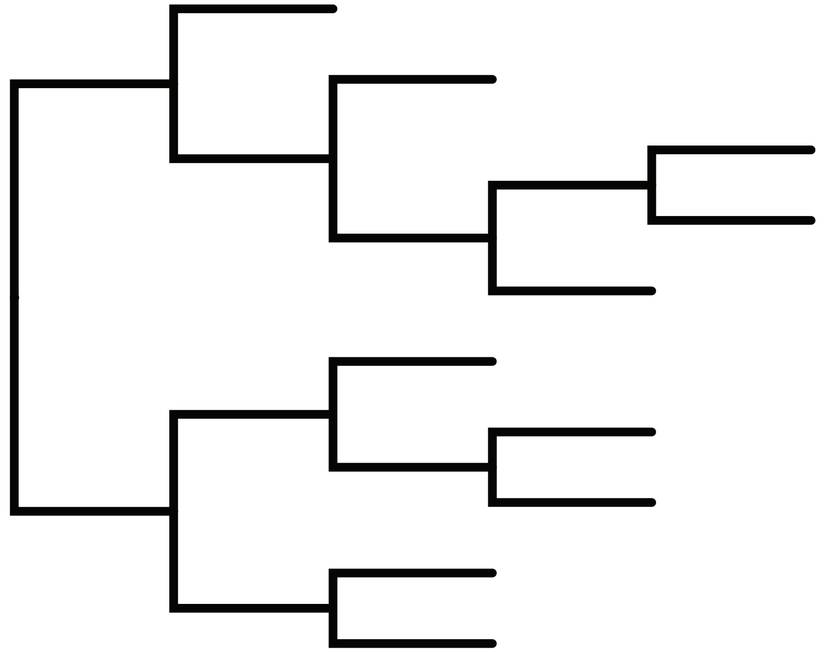
The clock model describes how evolutionary rates vary (or not) across the tree



Strict / constant molecular clock model

Assumptions

- The substitution rate is constant over time
- All lineages share the same rate



branch length = substitution rate

low high

Image source Tracy Heath

Relaxed clock models

Assumptions

- Lineage-specific rates

- The rate assigned to each branch is drawn from some underlying distribution

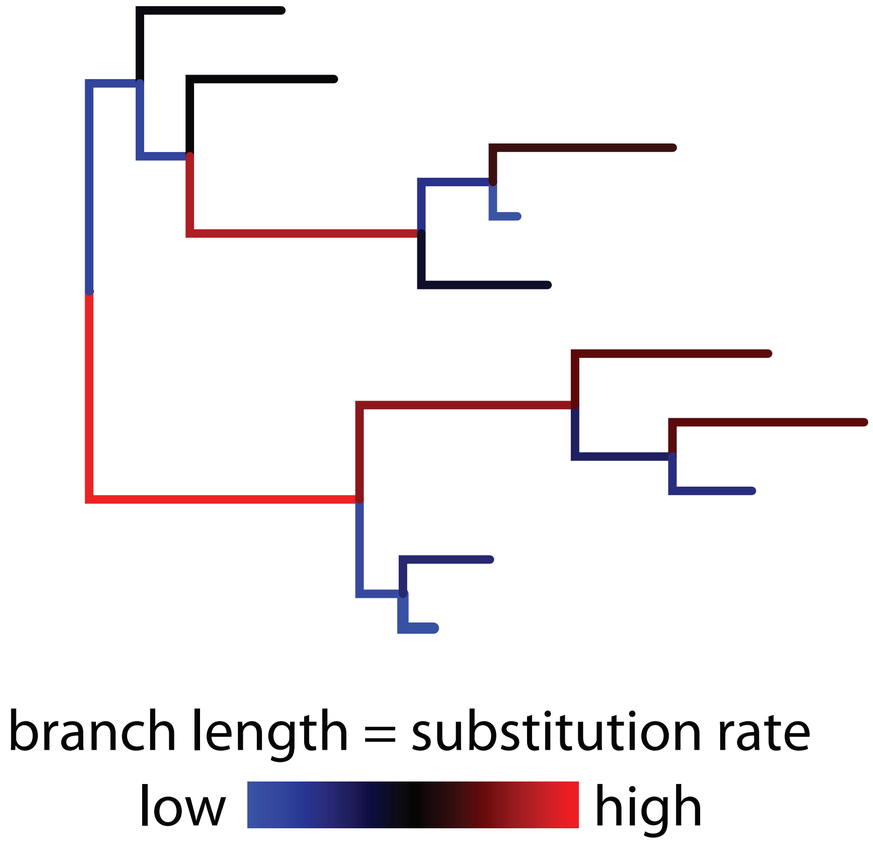


Image source Tracy Heath

Many different clock models

- Strict clock
- Uncorrelated or independent clock (= the favourite)
- Autocorrelated clock
- Local clocks
- Mixture models

The changing face of the molecular evolutionary clock. *Ho*. 2014
See also: *Warnock, Wright*. 2020. Elements of Paleontology

Applications of node dating

Hypothesis testing: did the break up of Gondwana drive the radiation of cichlids?

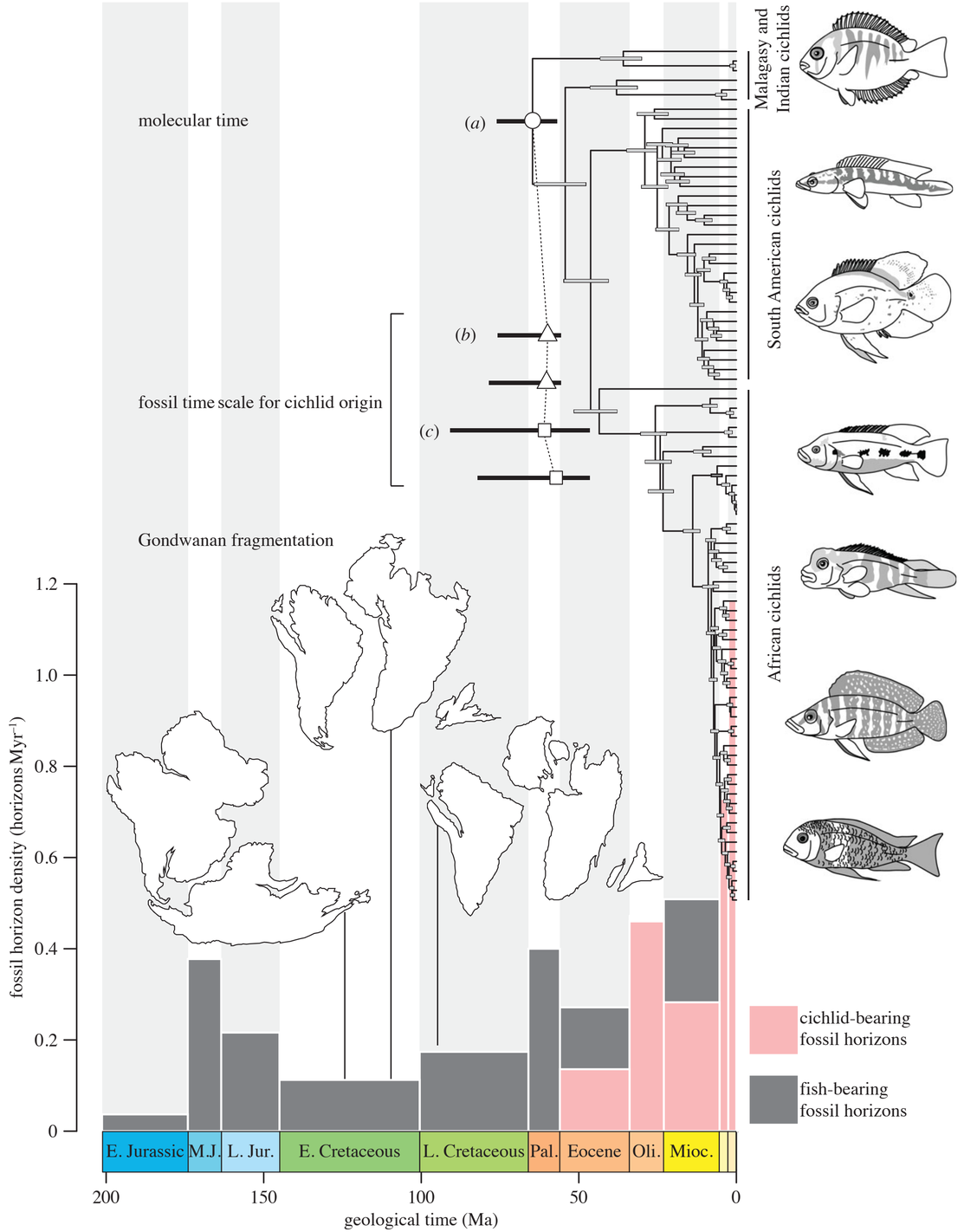
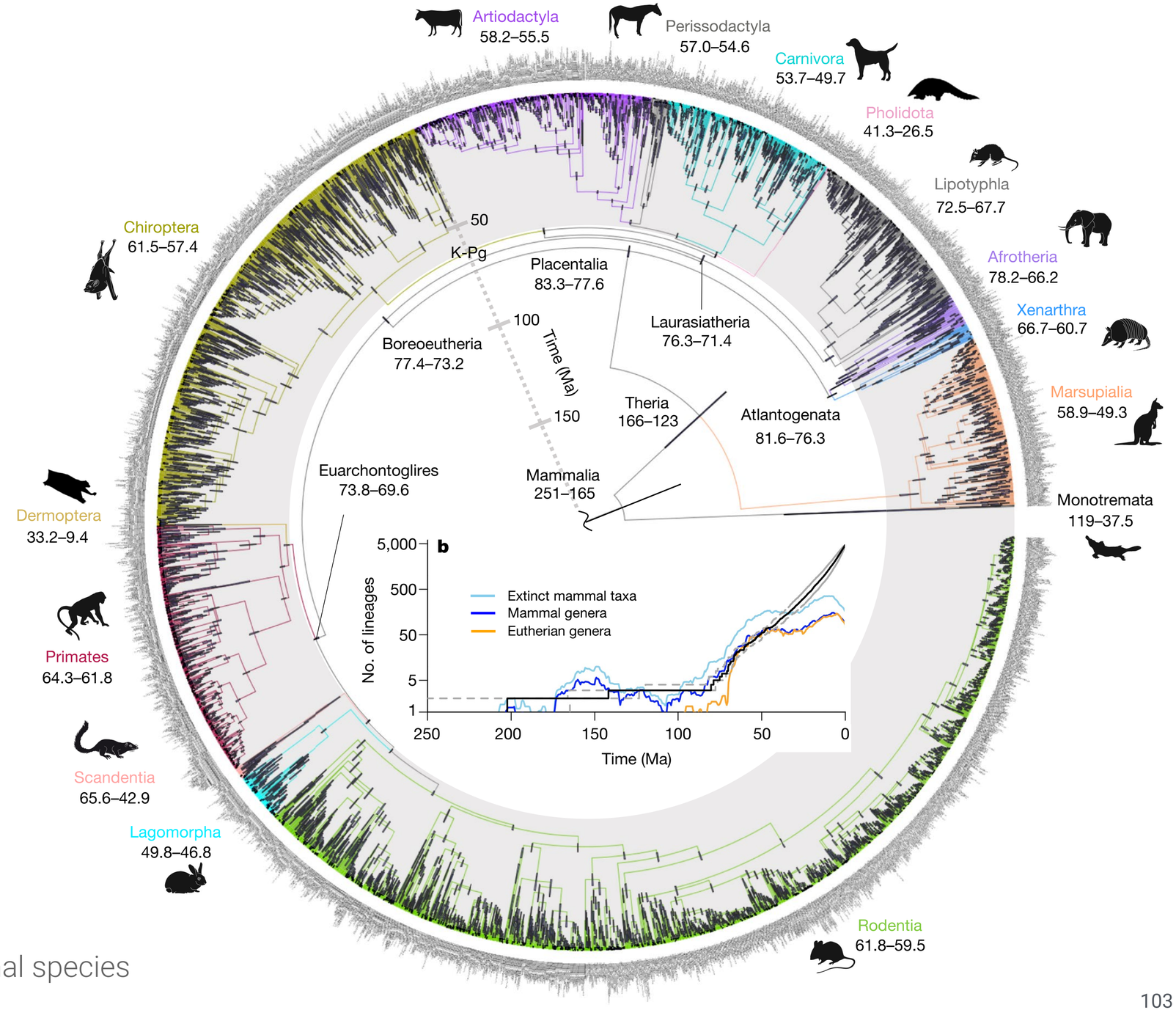


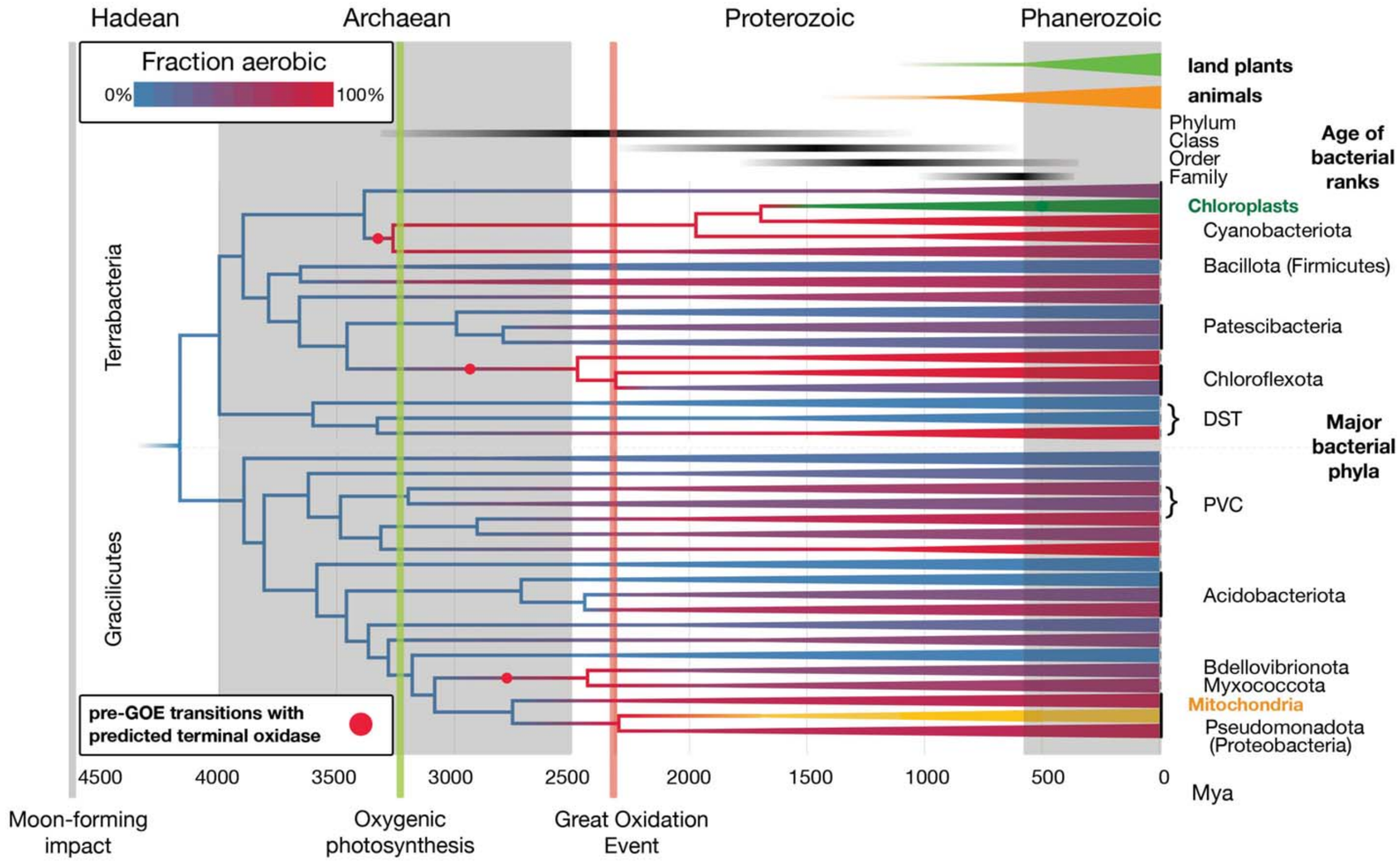
Image adapted from *Friedmann et al.* (2013)

Very large trees can only be time-calibrated using a node dating approach



Álvarez-Carretero et al. (2021) — 4,705 mammal species

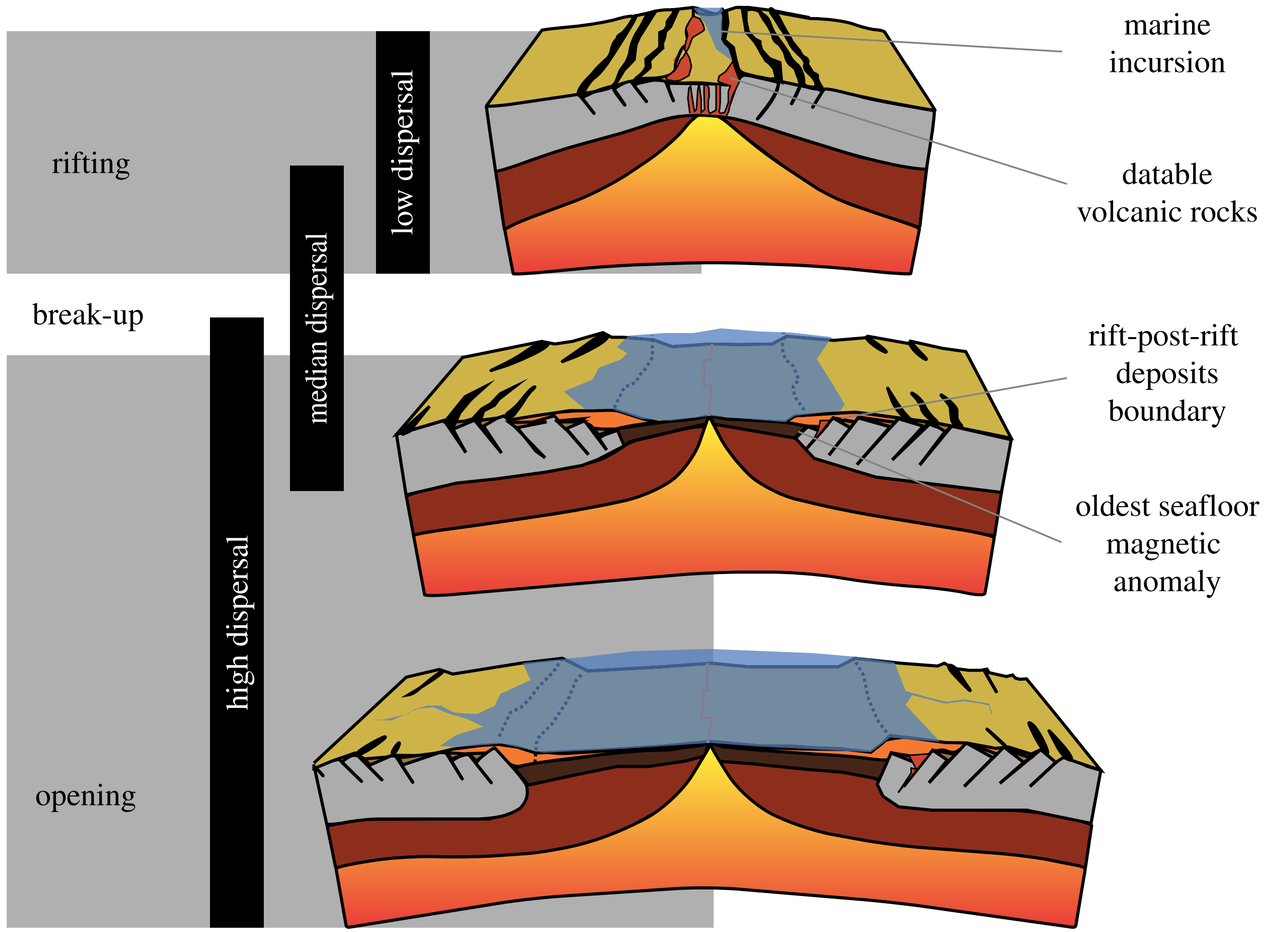
Dating with sparse calibration information



Davín et al. (2025)

A geological timescale for bacterial evolution, calibrated using atmospheric oxygenation and the spread of aerobic metabolism

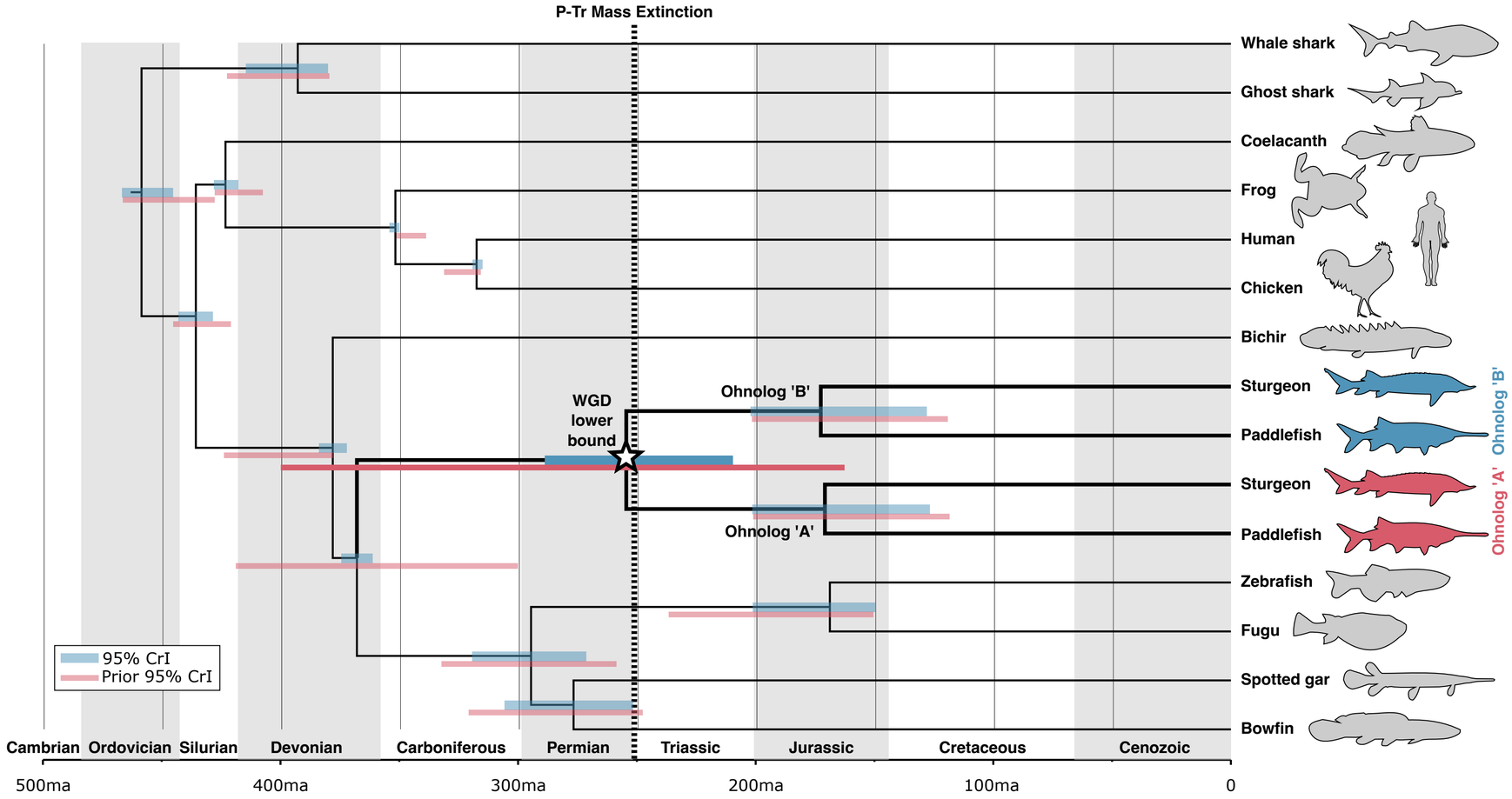
Biogeographic calibrations



De Baets et al. (2016)

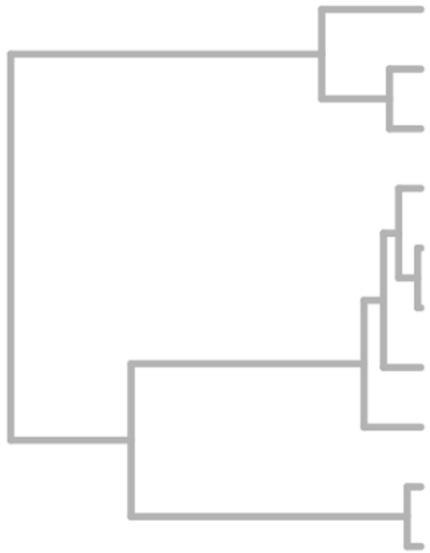
See also: Ho et al. (2015), *Landis* (2017, 2021), *Papadopoulou et al.* (2010)

Dating gene or genome duplications



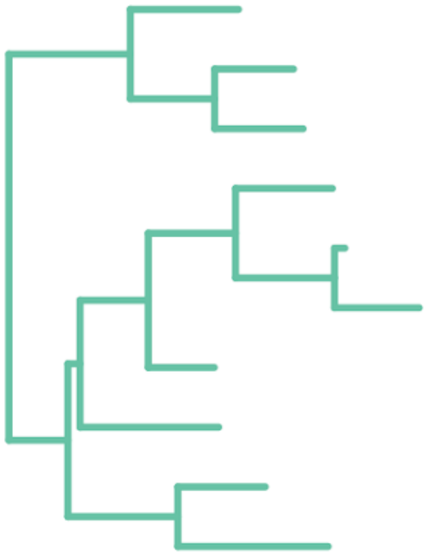
Redmond et al. (2023)

⚠️ Times and rates are not fully identifiable!



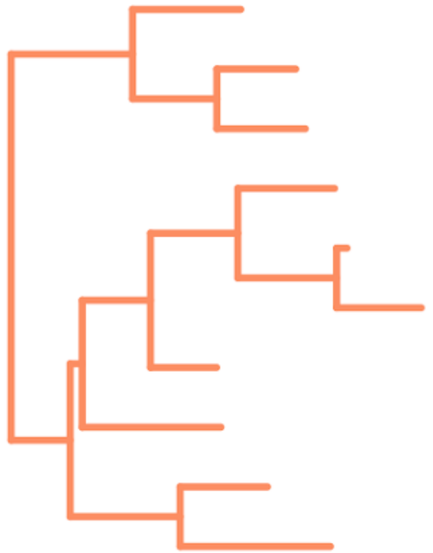
Time (year)

X



Clock rate (subs/site/year)

=



Genetic distance (subs/site)

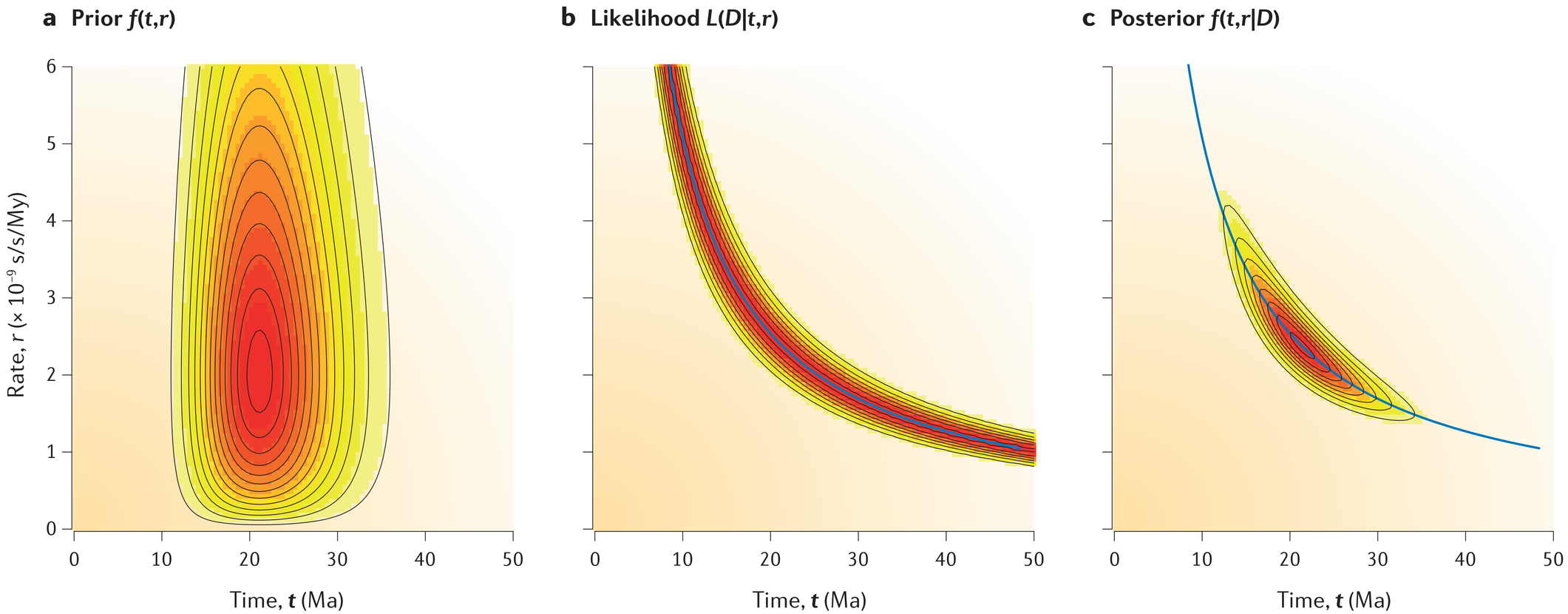
Prior

Prior

Likelihood

Slide adapted from Sebastian Duchene

⚠ The priors will **always** influence the results



dos Reis et al. 2015. Nature Reviews Genetics

Short break?



Introduction to MCMC

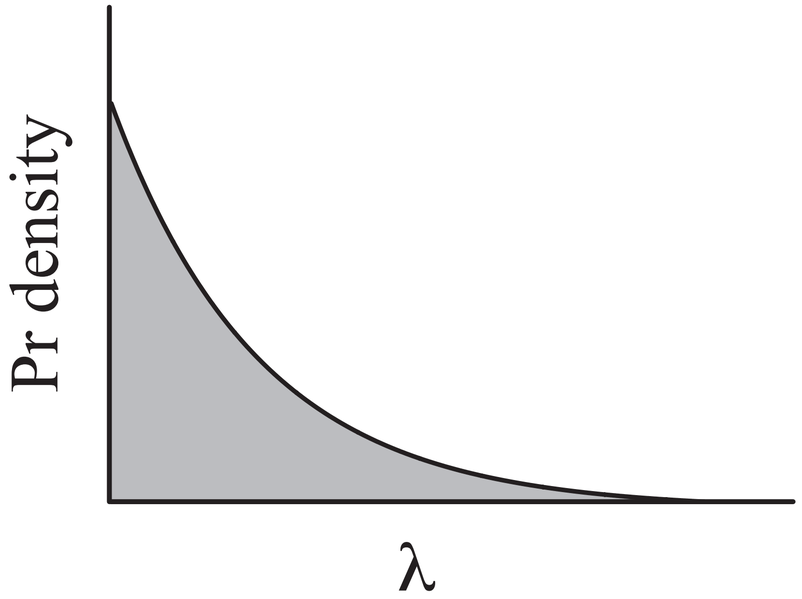
Probabilities vs. probability densities

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



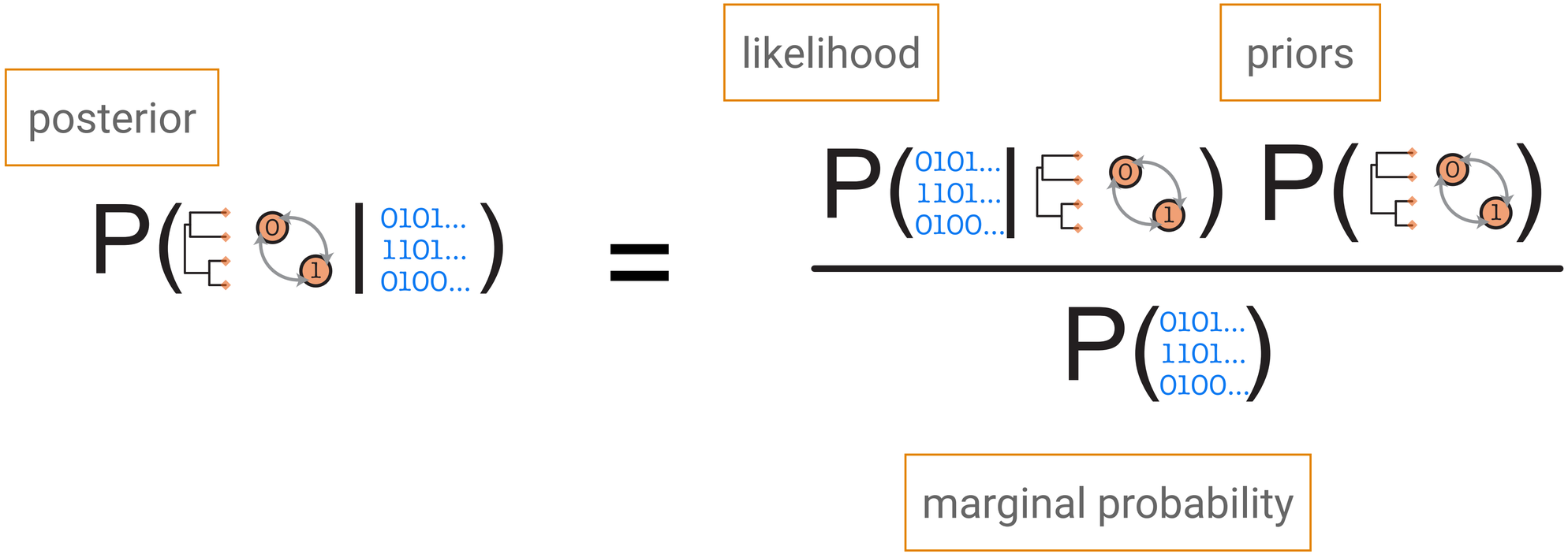
λ is drawn from an exponential distribution

The x-axis represents the value of our parameter λ

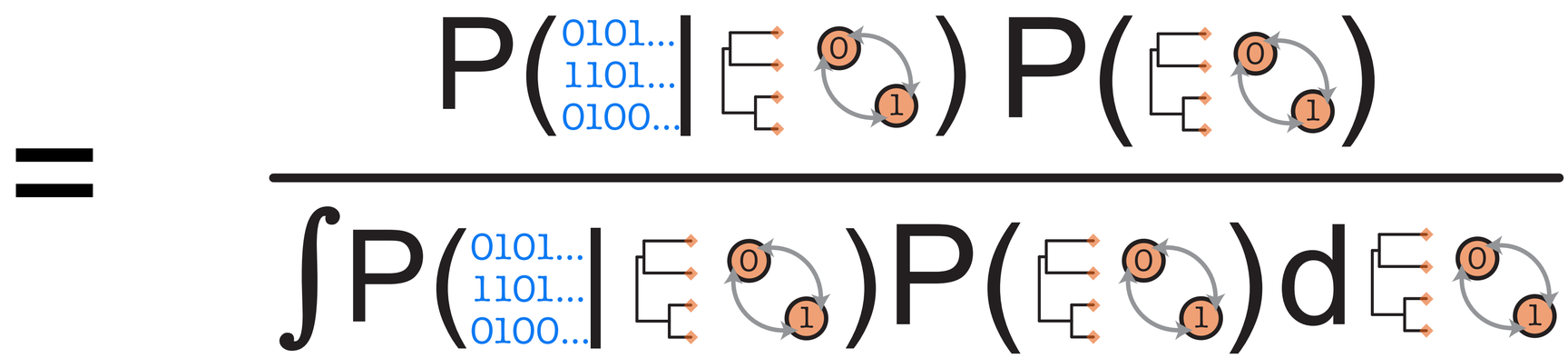
The y-axis does have a value but it is not so easily interpretable

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)?

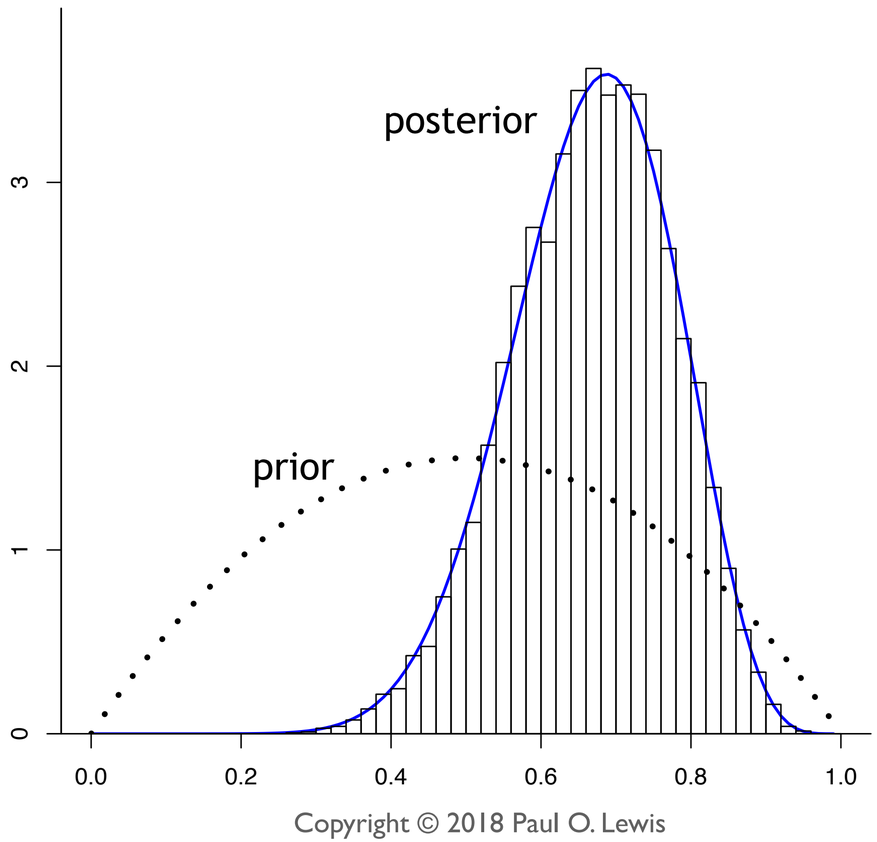
A group of algorithms for approximating the posterior distribution (also known as samplers)

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

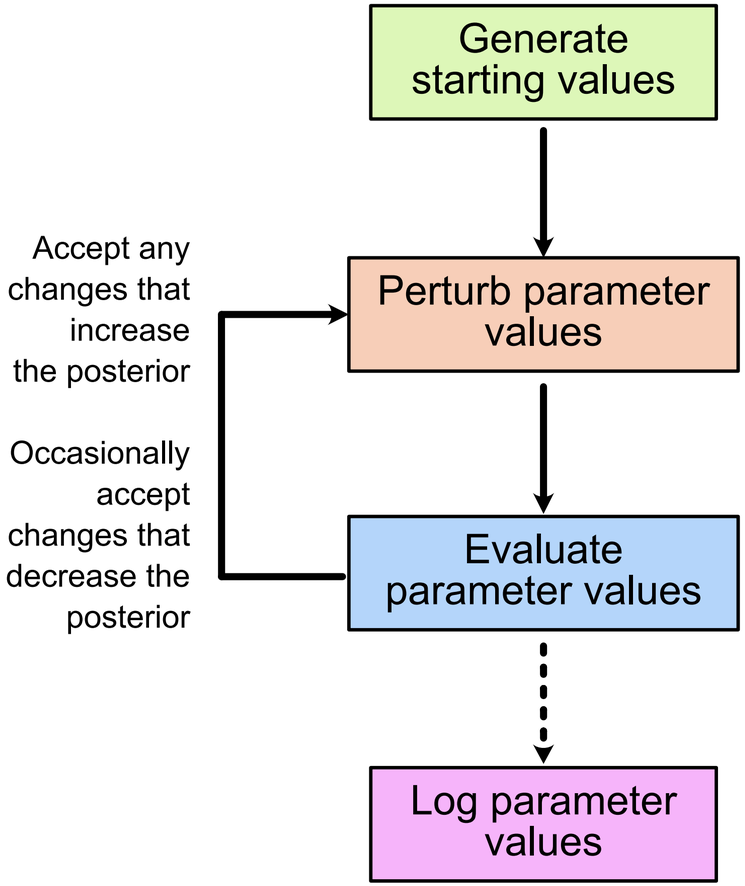
What is Markov chain Monte Carlo (MCMC)?



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

The Metropolis-Hastings algorithm

Flowchart

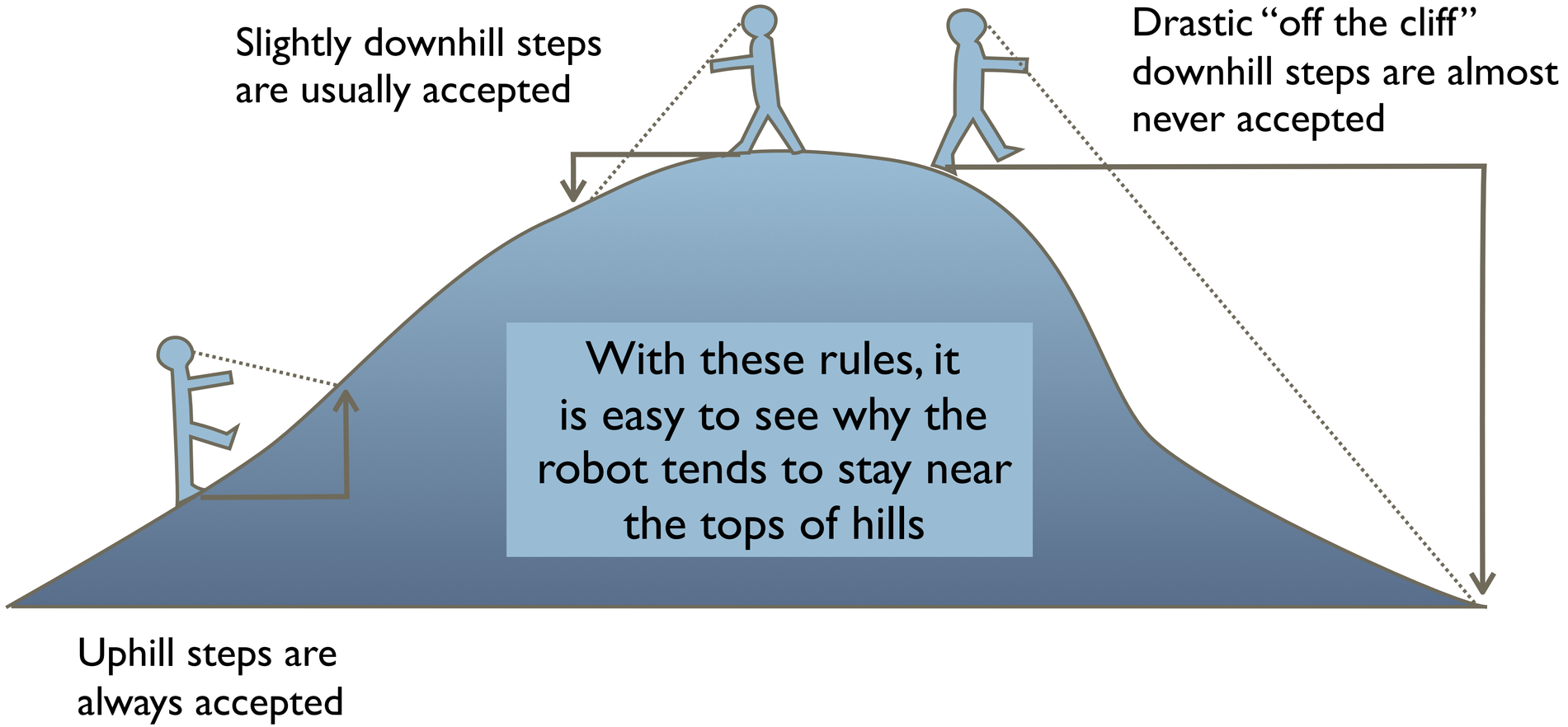


Pseudocode

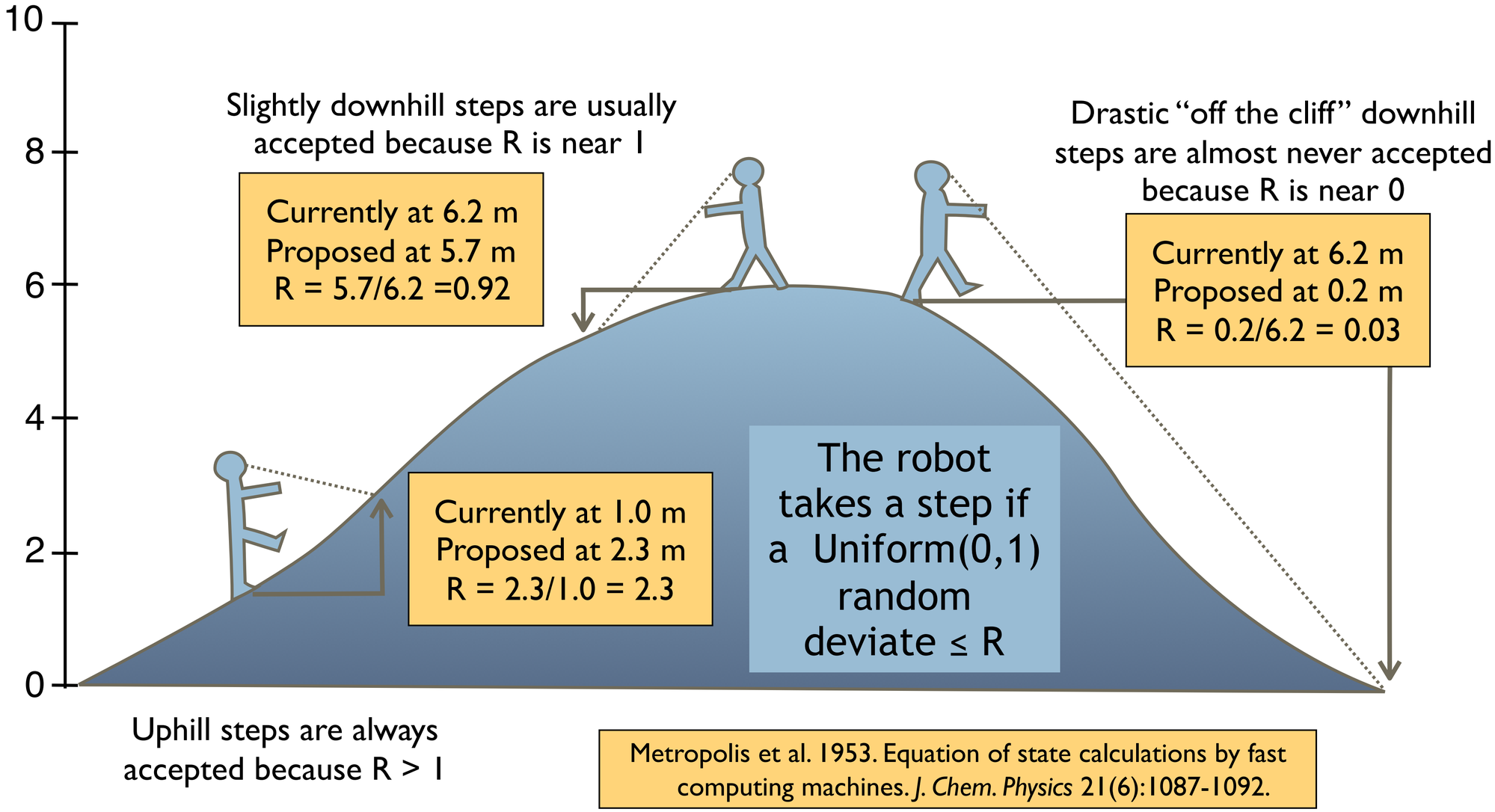
```
initialize starting values;

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
```

Barido-Sottani et al. (2024)

MCMC robot's rules



Copyright © 2018 Paul O. Lewis



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

Copyright © 2018 Paul O. Lewis

Hastings ratio

new parameter values

$$R = \frac{P(\text{tree}^*, \text{rates}^* \mid \text{data})}{P(\text{tree}, \text{rates} \mid \text{data})}$$

$$= \frac{P(\text{data} \mid \text{tree}^*, \text{rates}^*)\,P(\text{tree}^*, \text{rates}^*)}{P(\text{data})}$$

$$= \frac{P(\text{data} \mid \text{tree}^*, \text{rates}^*)\,P(\text{tree}^*, \text{rates}^*)}{P(\text{data} \mid \text{tree}, \text{rates})\,P(\text{tree}, \text{rates})}$$

The marginal probability of the data cancels out

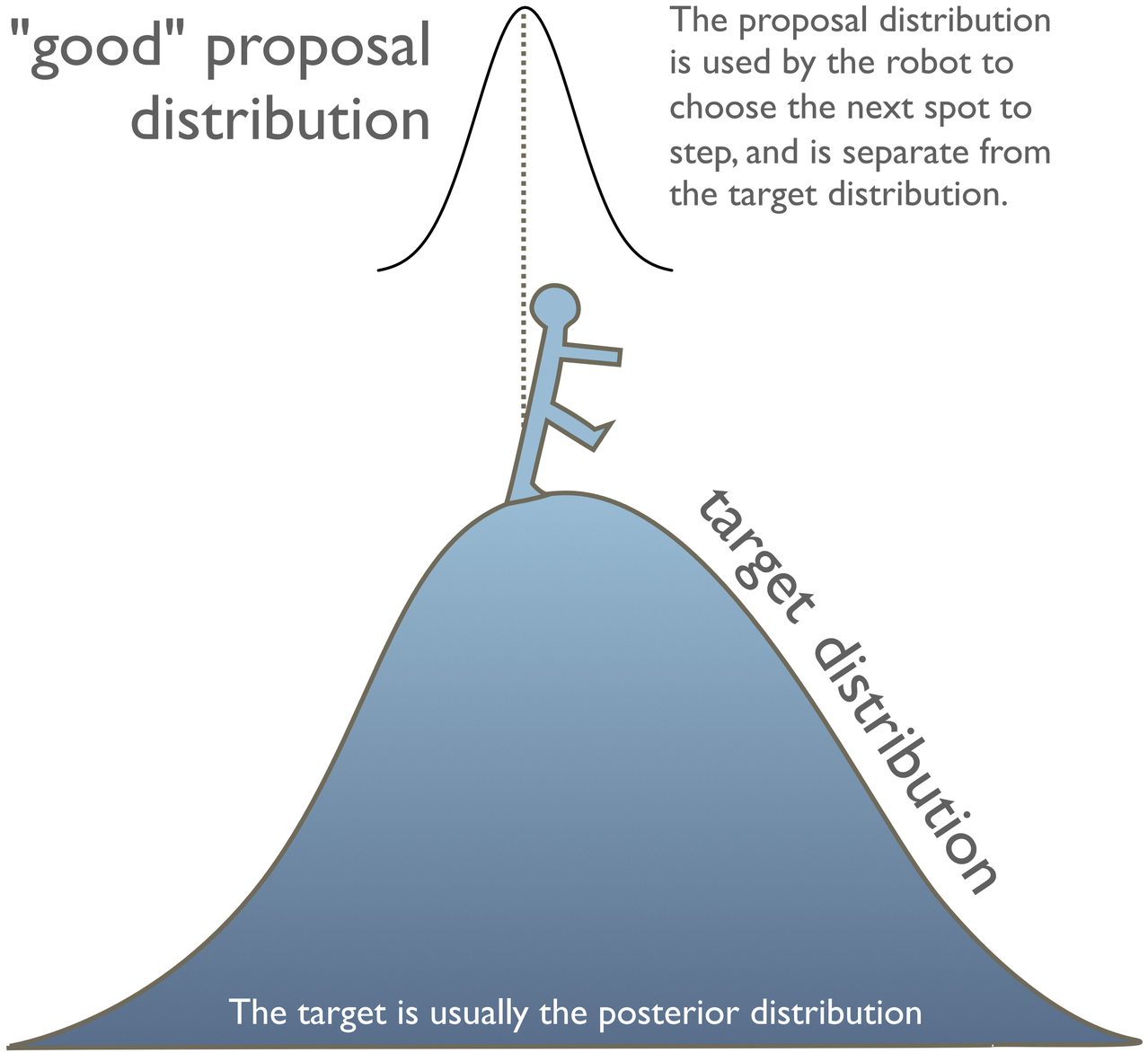
All we're left to calculate is the likelihood ratio and the prior odds ratio

Barido-Sottani et al. (2024)

Moves

Known as moves, proposals or operators — *how* we move across the parameter space

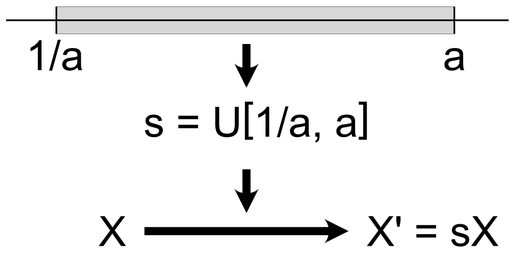
i.e., how do we select the next set of parameters to evaluate



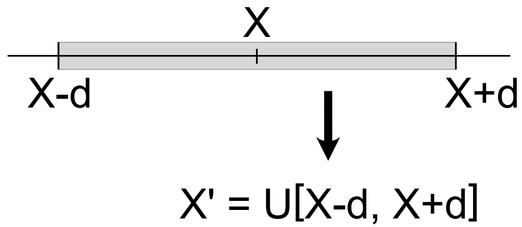
Copyright © 2018 Paul O. Lewis

Moves

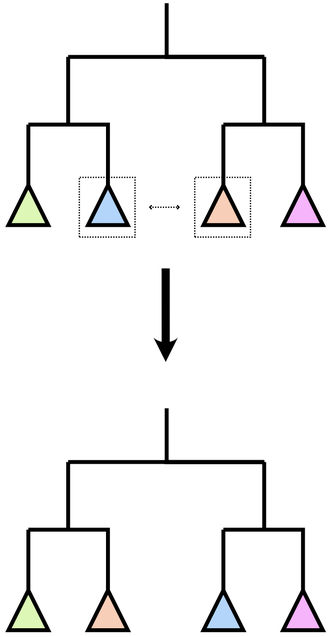
Scaling move



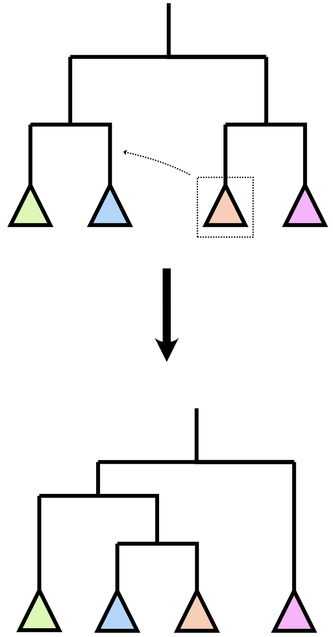
Sliding move



SubTree eXchange (STX) move



Subtree Pruning and Regrafting (SPR) move



Barido-Sottani et al. (2024)

MCMC Robot demo

Convergence

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

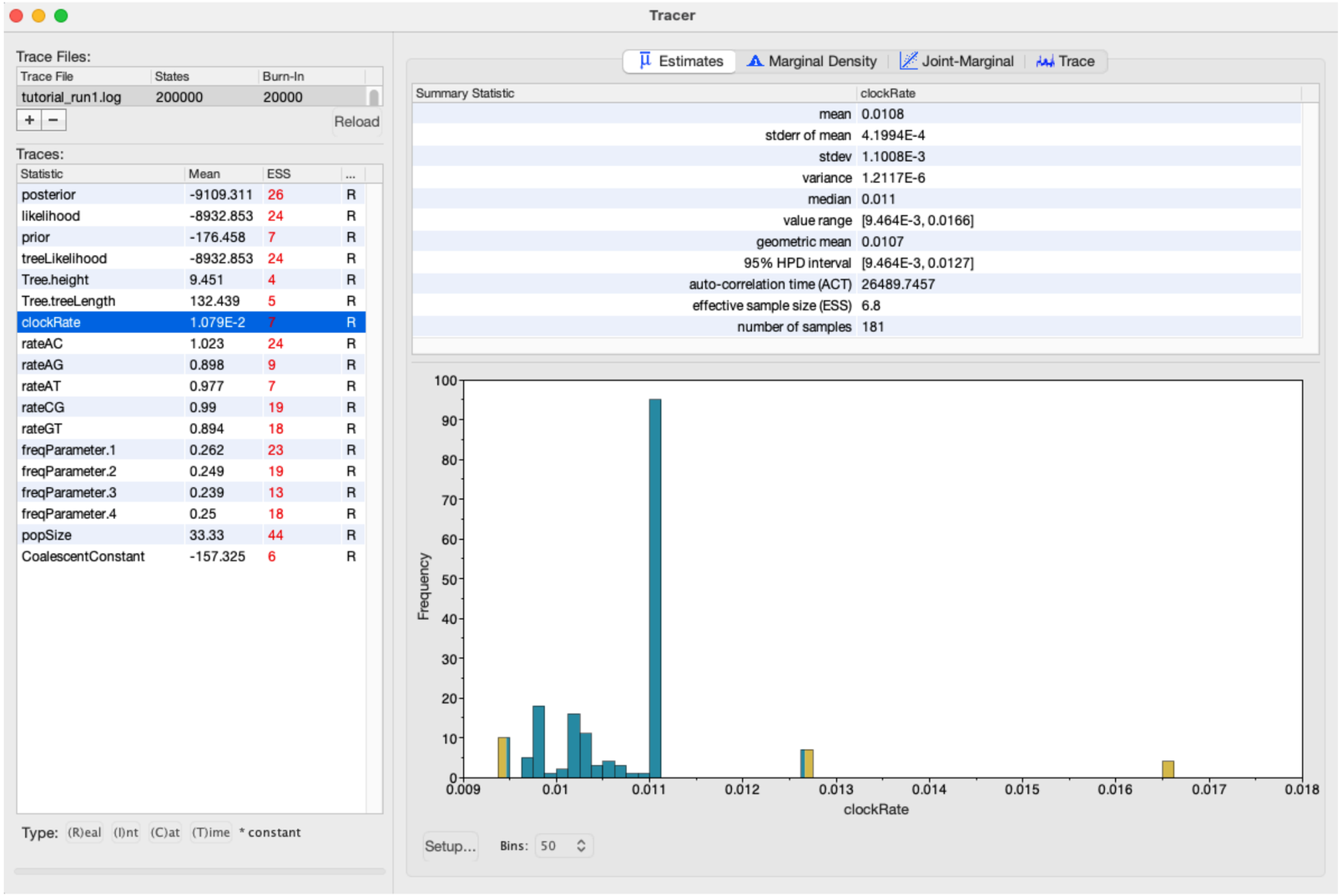
Key indicators based on trace (log) files:

- The effective sample sizes (ESS) - the effective number of independent draws from the target distribution, taking into account autocorrelation
- Does the histogram resemble a smooth distribution?
- Does the trace resemble a hairy caterpillar?
- Compare independent runs - they should overlap!

If we are confident the MCMC output provides a good representation of the posterior, we say the chain has converged or reached convergence

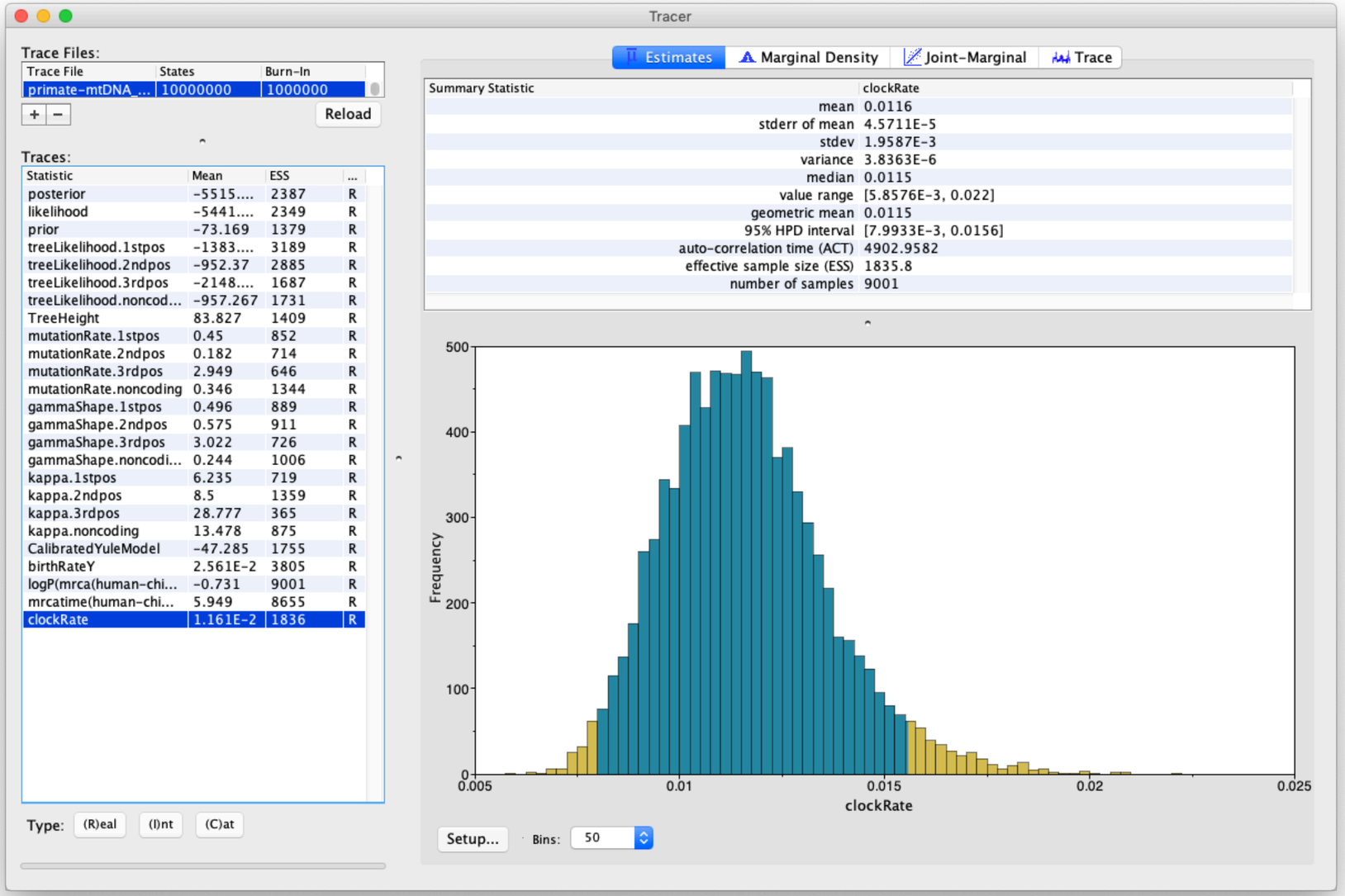
Exploring the posterior

Tracer is an amazing program for exploring MCMC output

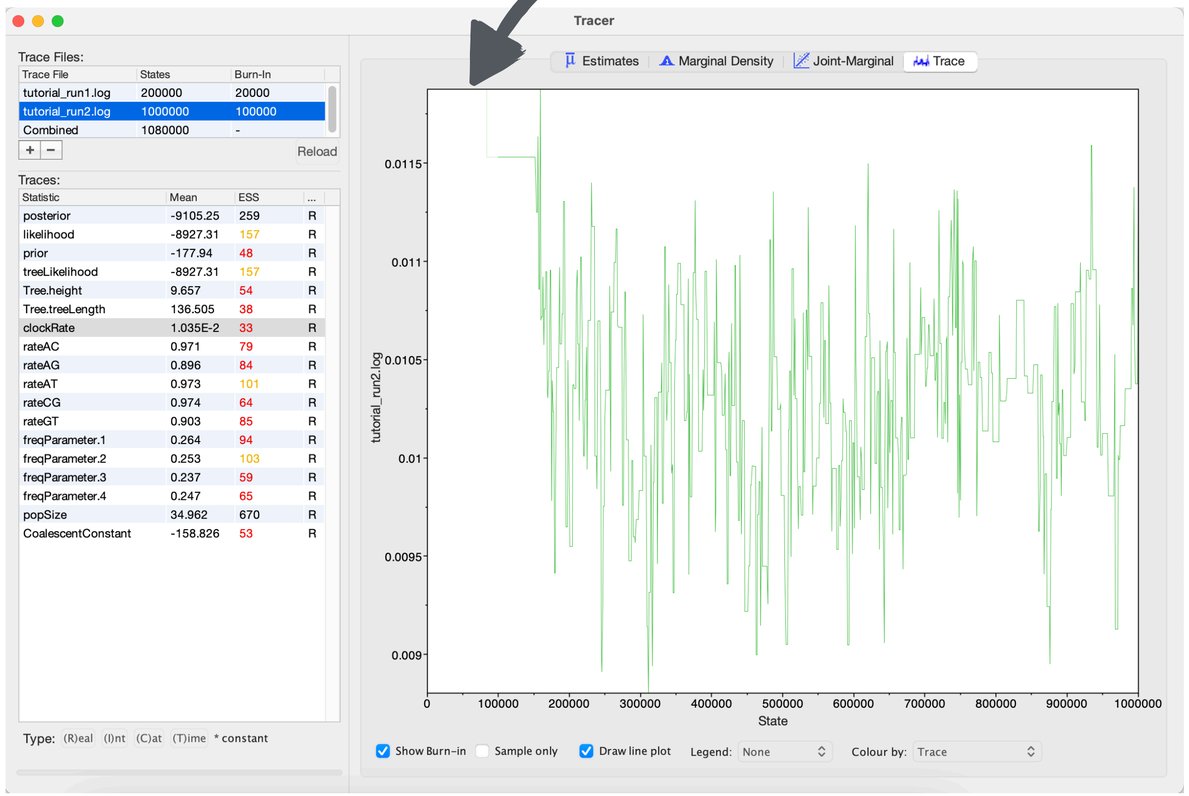


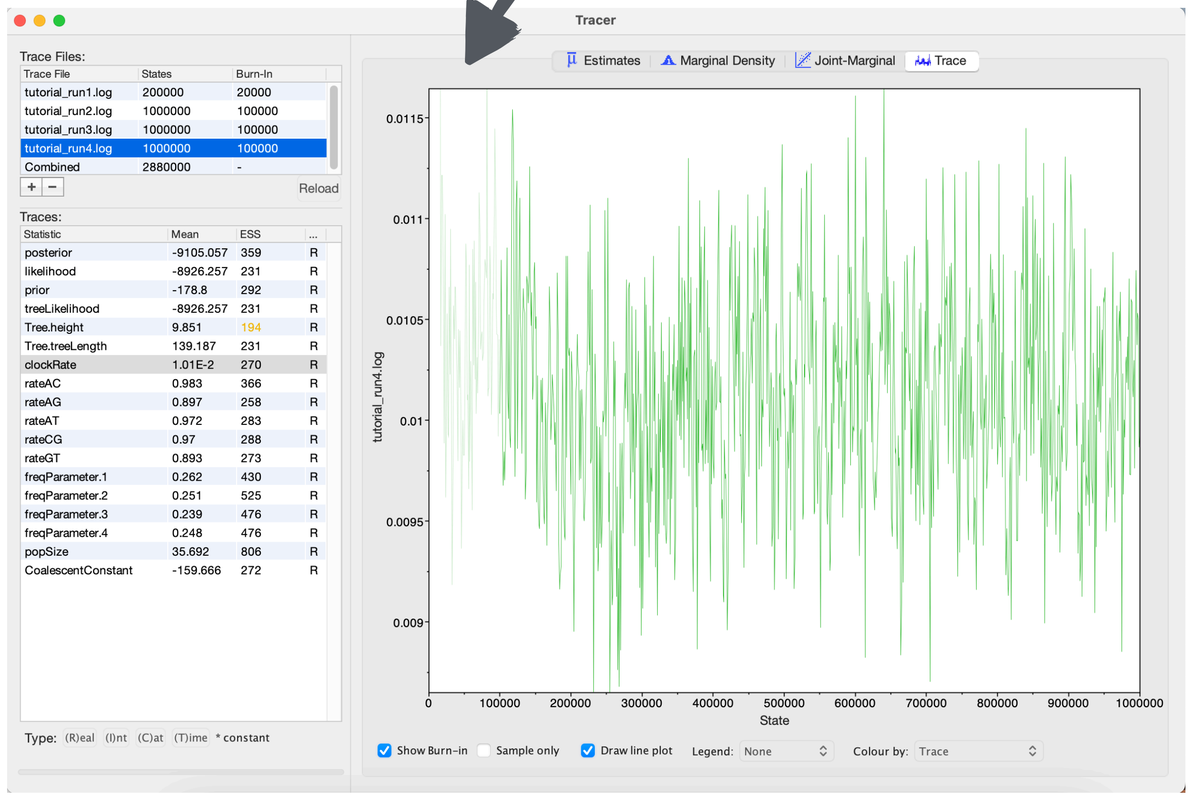
Exploring the posterior

Tracer is an amazing program for exploring MCMC output

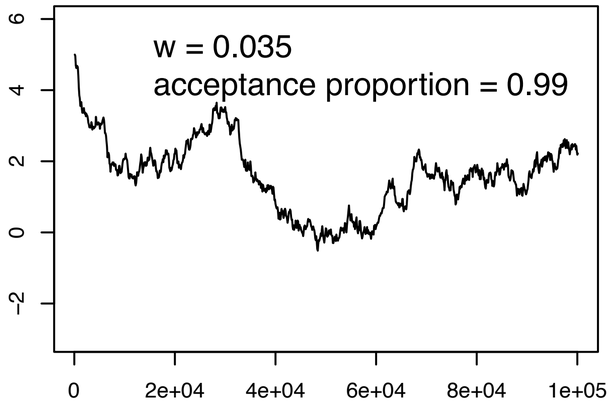


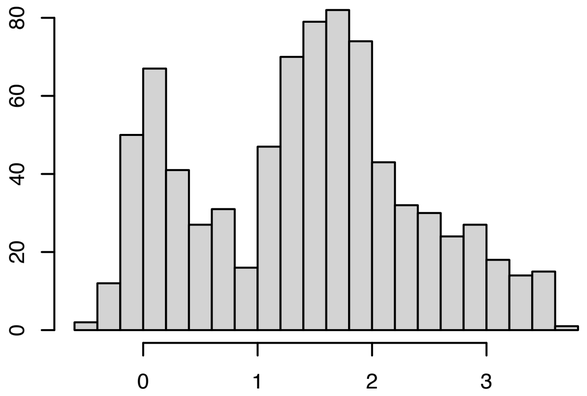
“Burn-in” - usually we discard the first 10-20% of steps

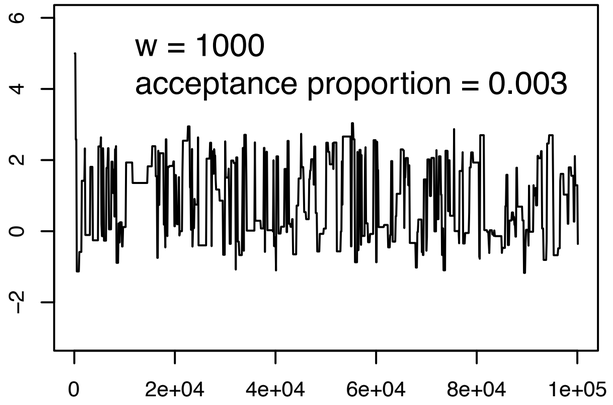


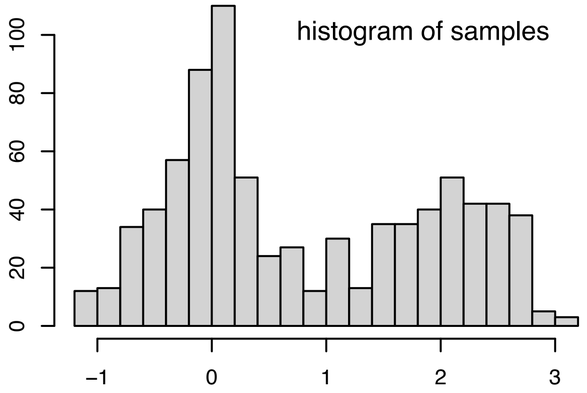




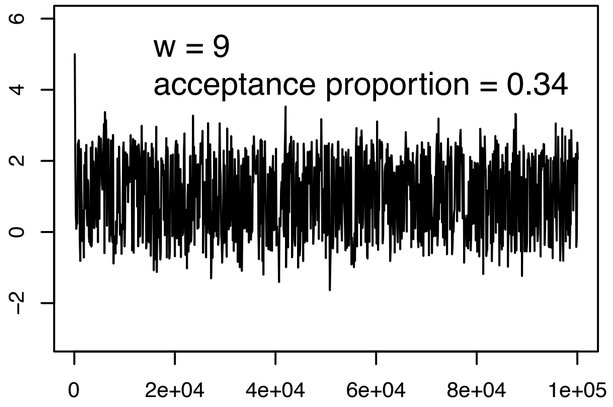


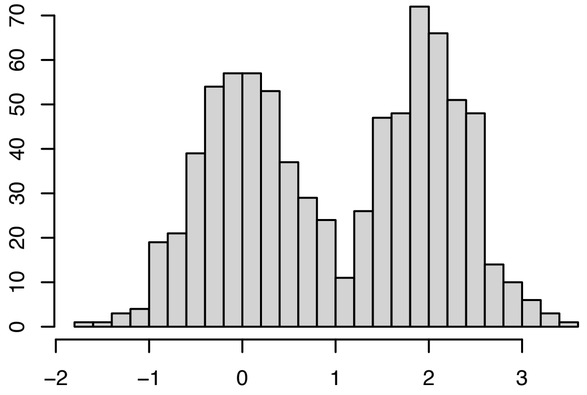






Better!



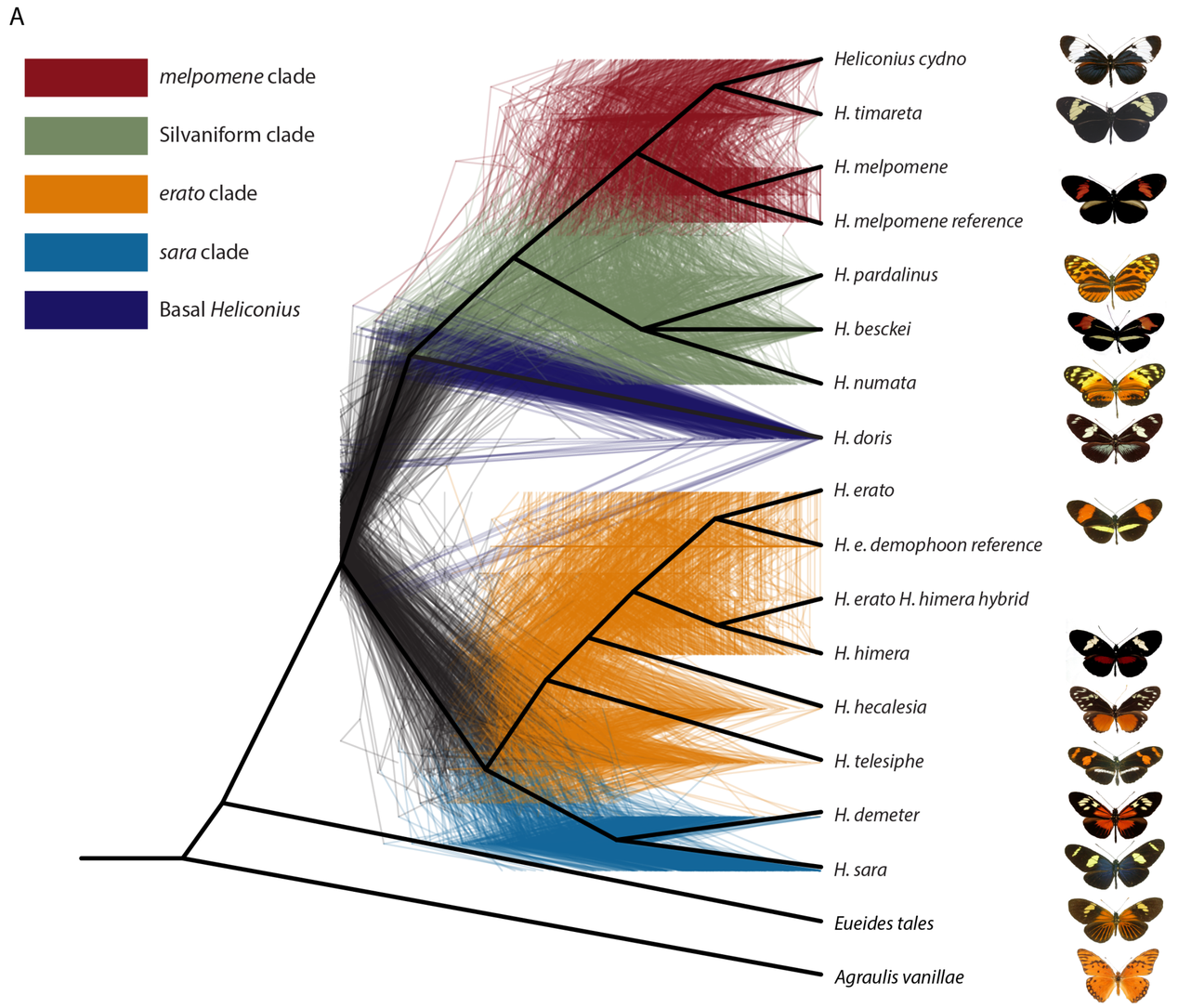


Barido-Sottani et al. (2024)

Summarising the posterior

Summarising trees is much more challenging

Presenting a single summary tree can be misleading



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

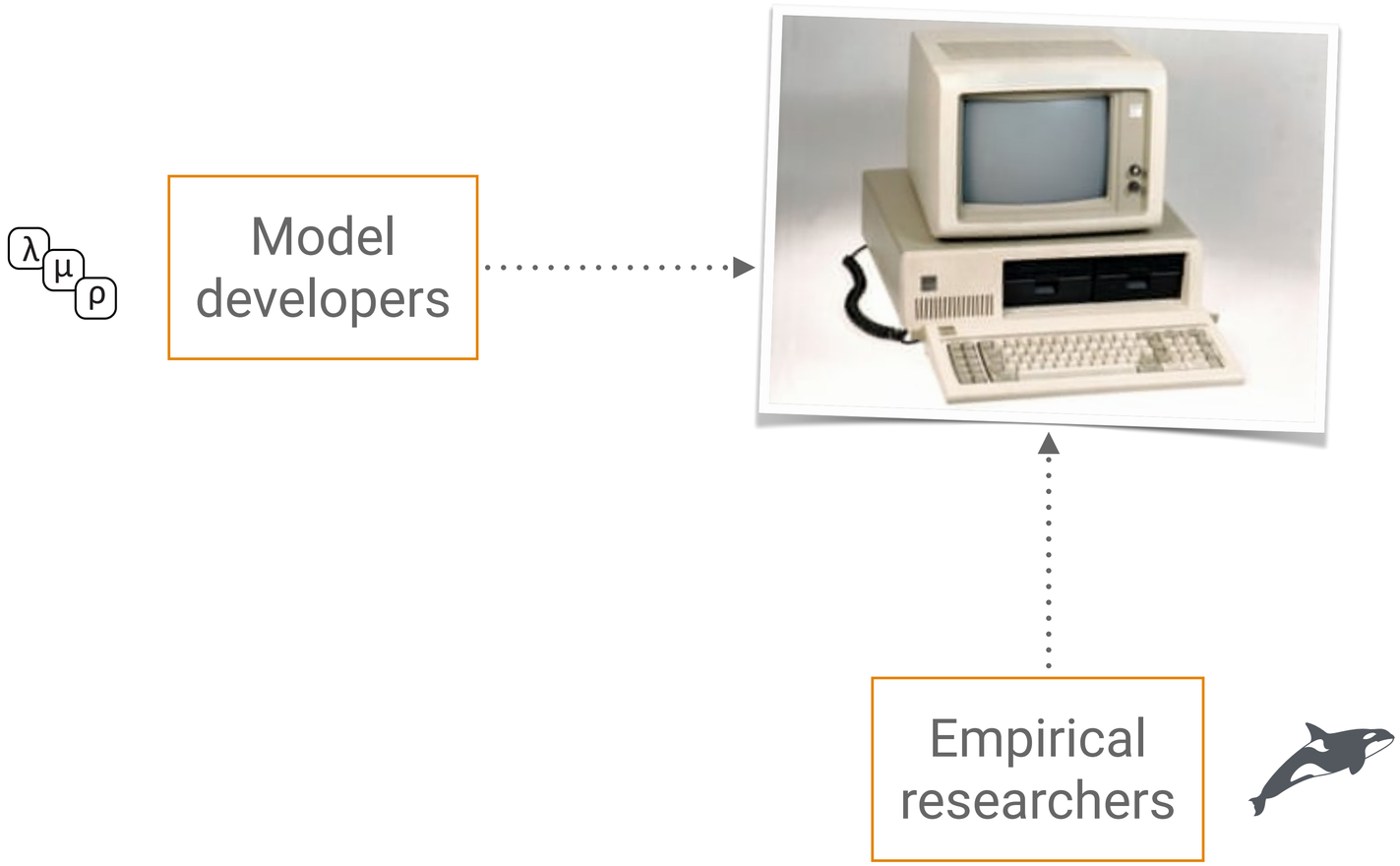
Maximum clade credibility (MCC) tree — “the tree with the maximum product of the posterior clade probabilities”

Maximum a posteriori (MAP) tree — “the tree topology that has the greatest posterior probability, averaged over all branch lengths and substitution parameter values”

⚠️ Fossil only trees are usually highly uncertain! *Capobianco* (2025)

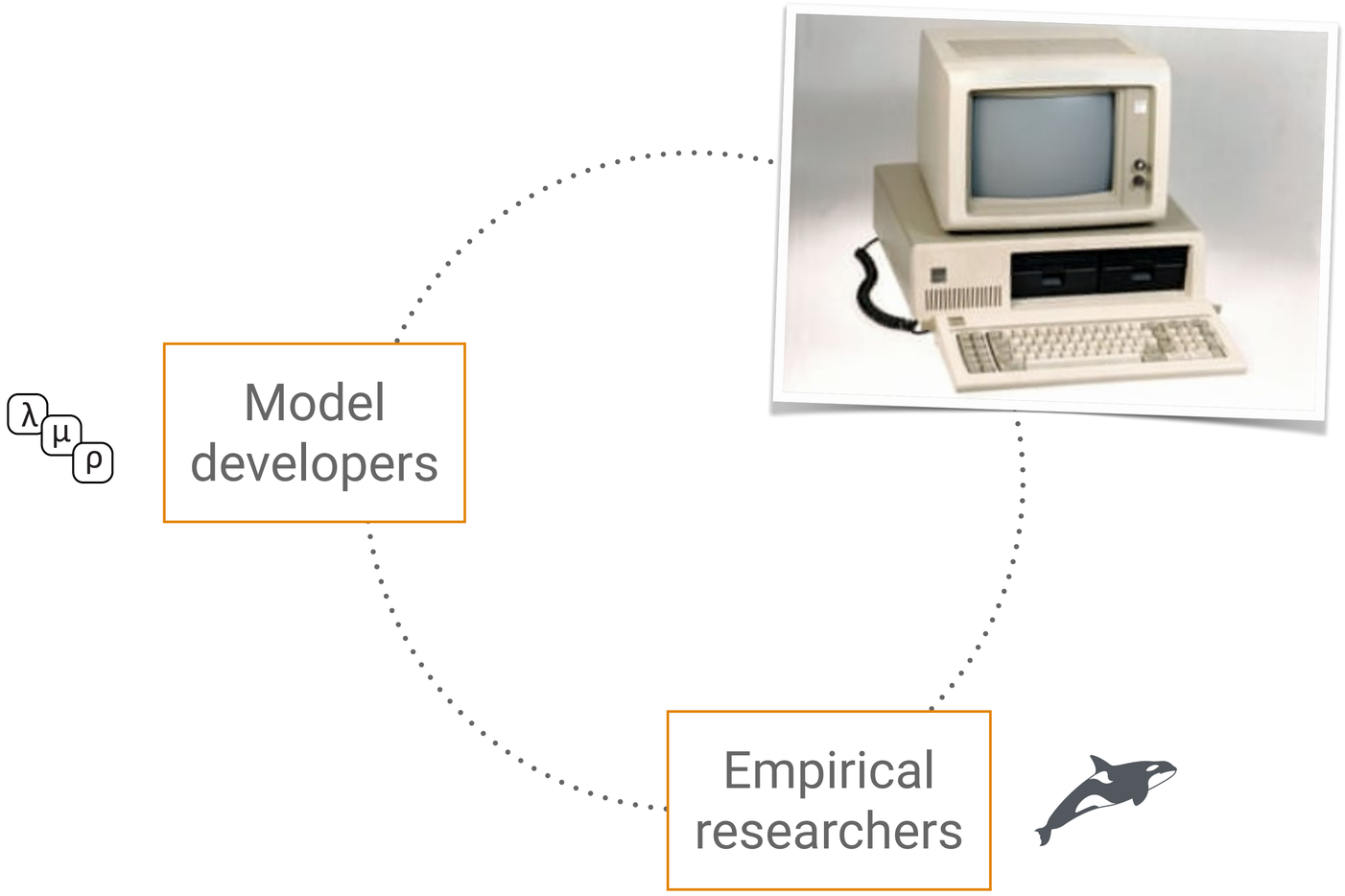
Intro to BEAST2

Phylogenetic inference — the old way

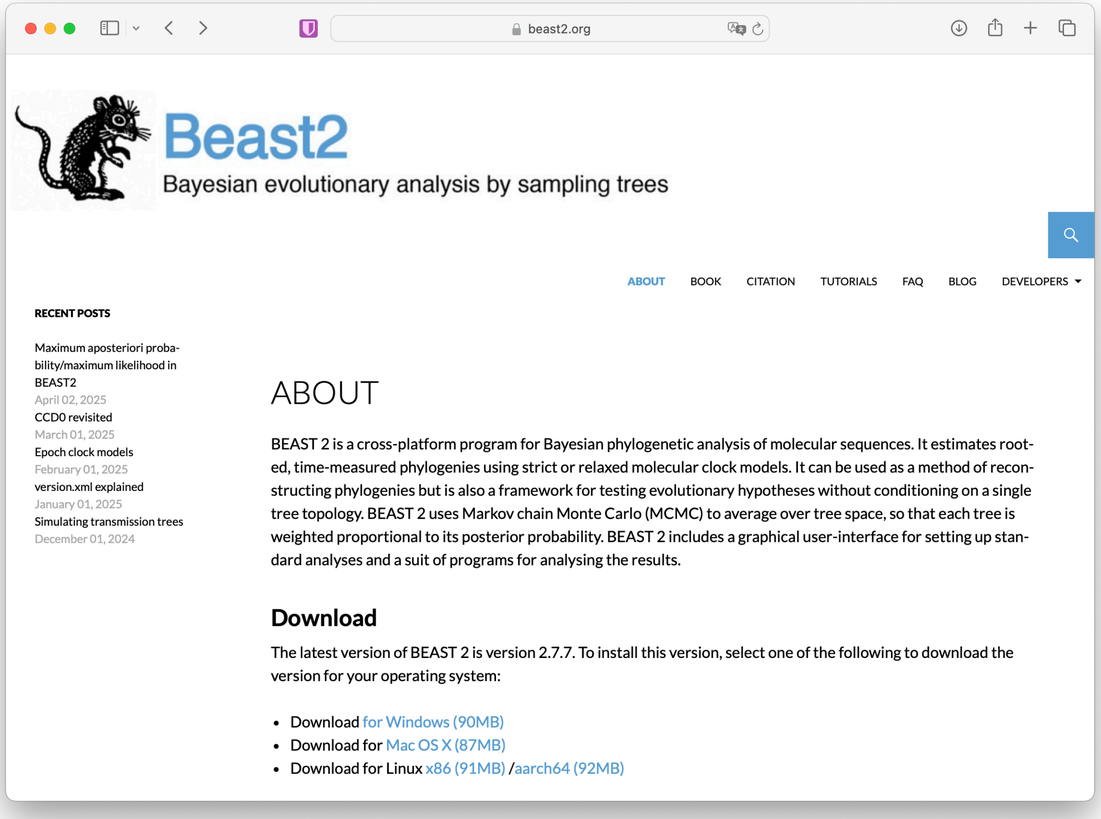


What we might call a **“black box”** approach

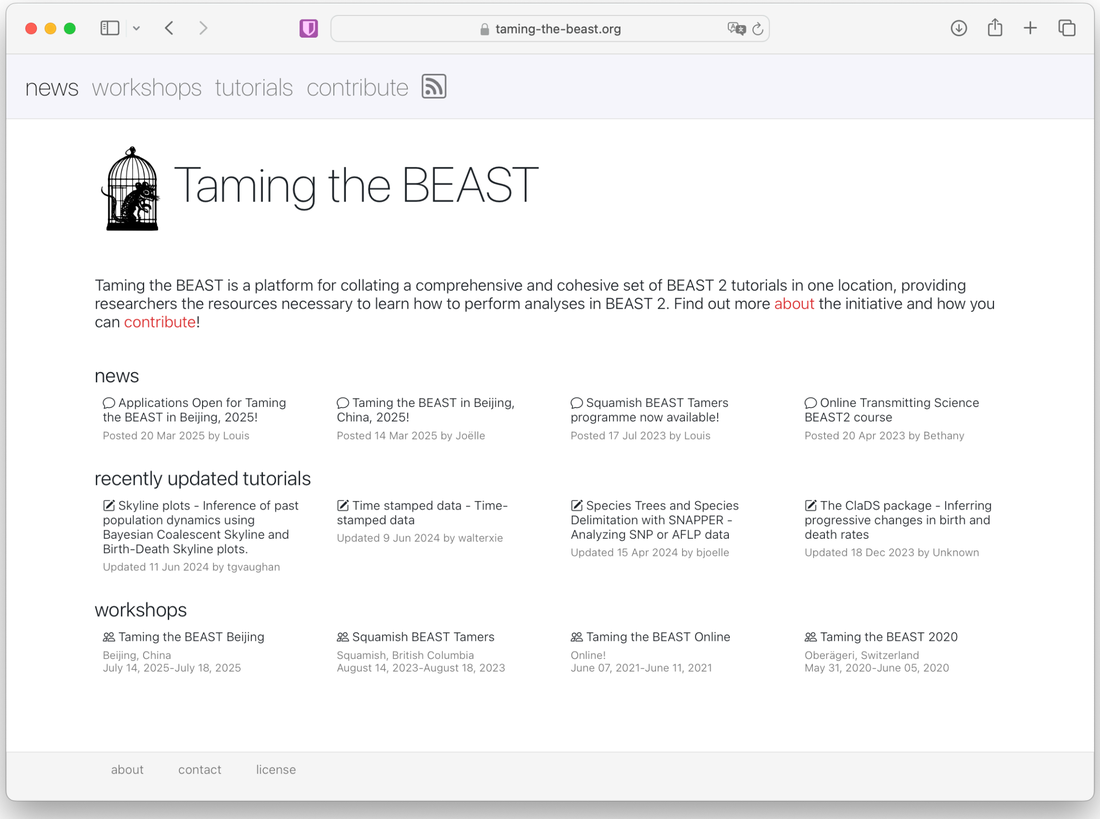
Phylogenetic inference — a better way?



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



www.beast2.org



taming-the-beast.org

BEAST2

Designed with extendability and flexibility in mind

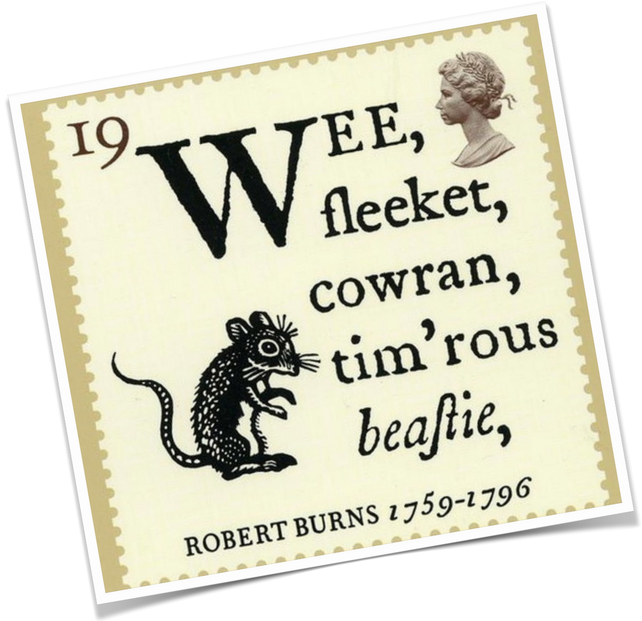
Developed and supported by a large international team of developers

Has a suite of apps that can used to generate input files and parse the output

www.beast2.org

Beast2

Bayesian evolutionary analysis by sampling trees



Scots poem - also the BEAST2 logo!

BEAST2 toolkit and work flow



Step 1. generate the xml file in **BEAUti**



Step 2. run your analysis in **BEAST**



Step 3a. Examine you log files using **Tracer**



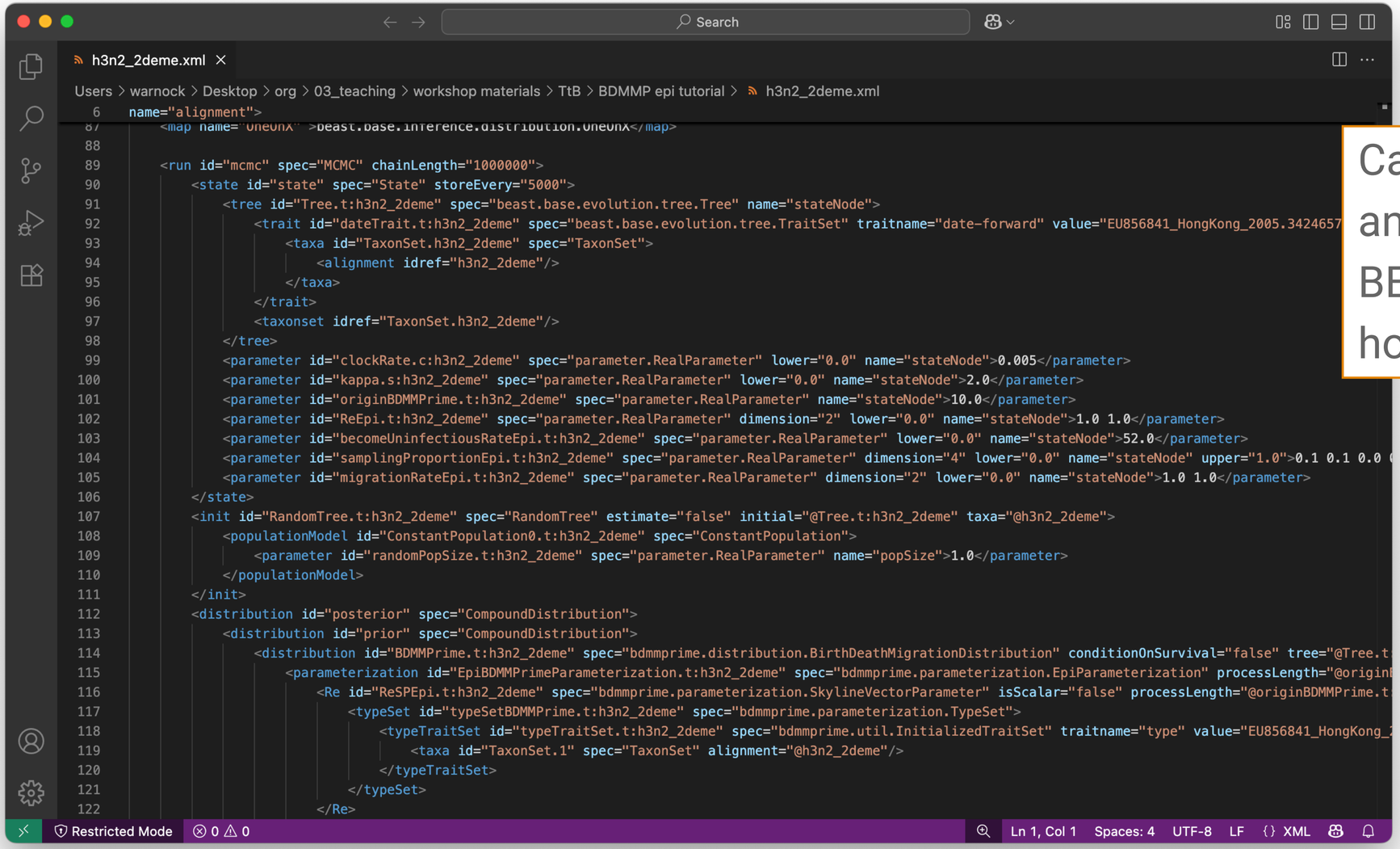
Step 3b. Generate a summary tree using **TreeAnnotator**



Step 4. Examine your summary tree in **FigTree**

Step… any other downstream analysis

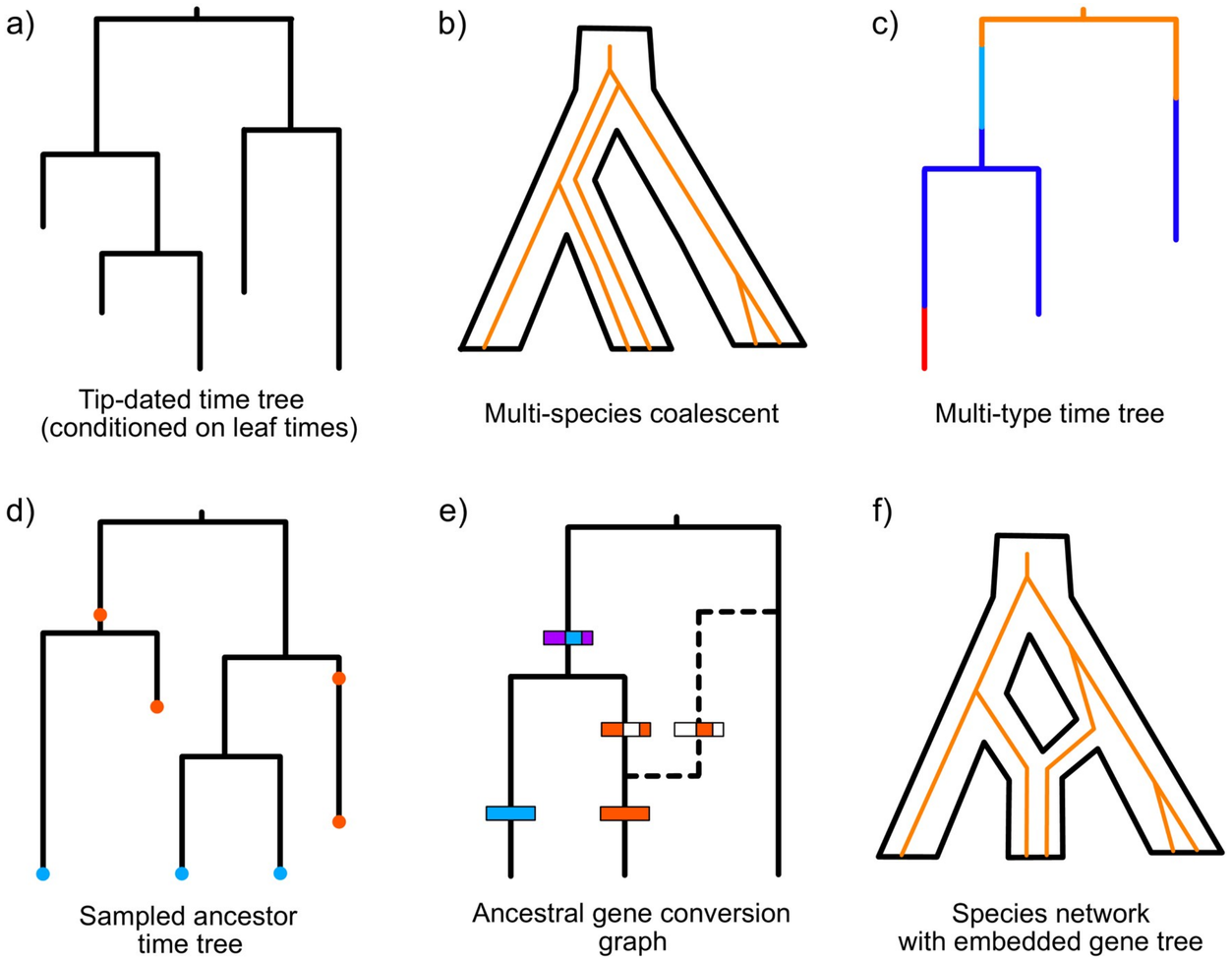
BEAST2 input: the XML file



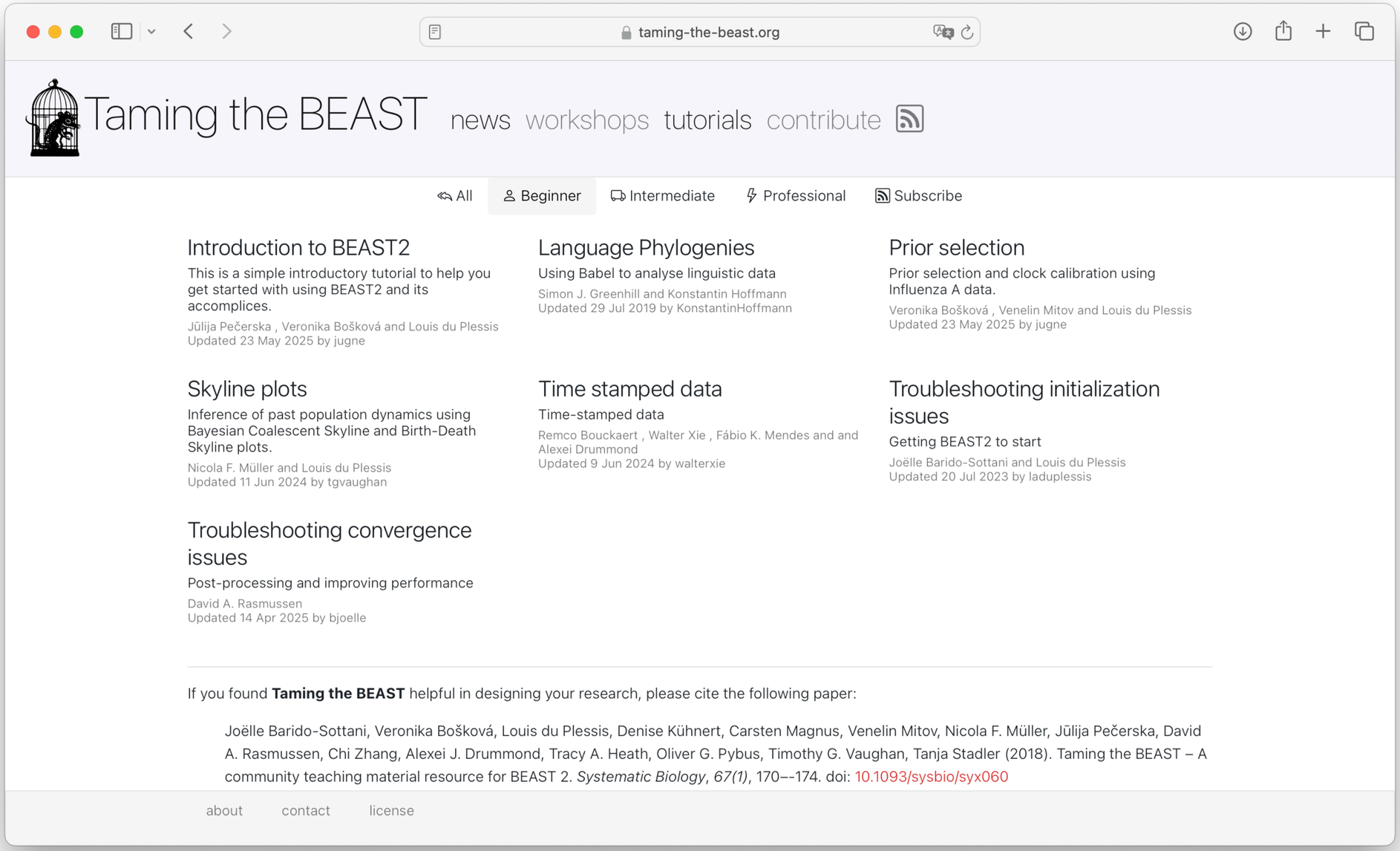
Caveat: if you want to use an option not available in BEAUti you have to learn how to edit the XML

A wide range of models and tree structures

Note all tree models in BEAST2 incorporate a temporal component



Bouckaert et al. (2019) *PLOS Comp Bio*



https://taming-the-beast.org/

Exercise

Further reading



Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology *Huelsenbeck* et al. (2001)



MCMC Robot App by Paul Lewis



Practical guidelines for Bayesian phylogenetic inference using Markov chain Monte Carlo (MCMC) *Barido-Sottani et al.* (2024)



decodinggenomes.org (free pdf available) *Stadler et al.* (2024) Chap 9, Phylodynamics

Extra slides

#NEXUS

Initiates the nexus block

```
[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia         00?1??1???????????0???1?211201??????????0-?-????????
Laccaria        110?001000000000000-000--00--00---00--0--------000-
Capsaspora      001011-00------000-00---00--------0----------0-0-
Monosiga        001000-11------000-00---00--------0----------0-0-
Sycon           00110011111111111110100110--0000000000----0---000-
Amphimedon      00110011111111111121100110--00000000000---0---000-
Trichoplax      0011001111111100000-0000210--01111110000--00---000-
Mnemiopsis      0011001111111100000-00112112001100101111121110011110
Nematostella    00110011111110000000112111001100101111111101111111
Hydra           0011001111111100000-00112111001100101111111110111111

;

END;
```

```
#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia        00?1??1???????????0???1?211201??????????0-?-????????
Laccaria       110?001000000000000-000--00--00---00--0--------000-
Capsaspora     001011-00------000-00---00--------0----------0-0-
Monosiga       001000-11------000-00---00--------0----------0-0-
Sycon          00110011111111111110100110--0000000000----0---000-
Amphimedon     00110011111111111121100110--00000000000---0---000-
Trichoplax     0011001111111100000-0000210--01111110000--00---000-
Mnemiopsis     0011001111111100000-00112112001100101111211100111110
Nematostella   00110011111111000000011211100110010111111110111111
Hydra          0011001111111100000-00112111001100101111111110111111

;

END;
```

Comments go in square brackets

Great for keeping track of data sources

```
#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia         00?1??1???????????0???1?211201??????????0-?-????????
Laccaria        110?001000000000-000--00--00---00--0--------000-
Capsaspora      001011-00------000-00---00--------0----------0-0-
Monosiga        001000-11------000-00---00--------0----------0-0-
Sycon           00110011111111111110100110--0000000000----0---000-
Amphimedon      00110011111111111121100110--00000000000---0---000-
Trichoplax      0011001111111100000-0000210--01111110000--00---000-
Mnemiopsis      0011001111111100000-00112112001100101111211100111110
Nematostella    00110011111110000000112111001100101111111101111111
Hydra           0011001111111100000-00112111001100101111111110111111

;

END;
```

Initiates the data block

```
#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia         00?1??1??????????0???1?211201??????????0-?-????????
Laccaria        110?001000000000-000--00--00---00--0--------000-
Capsaspora      001011-00------000-00---00--------0----------0-0-
Monosiga        001000-11------000-00---00--------0----------0-0-
Sycon           00110011111111111110100110--0000000000----0---000-
Amphimedon      00110011111111111121100110--00000000000---0---000-
Trichoplax      0011001111111100000-0000210--01111110000--00---000-
Mnemiopsis      0011001111111100000-00112112001100101111211100111110
Nematostella    00110011111110000000112111001100101111111101111111
Hydra           0011001111111100000-00112111001100101111111110111111

;

END;
```

Number of taxa and characters in the matrix

```
#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia         00?1??1???????????0???1?211201??????????0-?-????????
Laccaria        110?001000000000000-000--00--00---00--0--------000-
Capsaspora      001011-00------000-00---00--------0----------0-0-
Monosiga        001000-11------000-00---00--------0----------0-0-
Sycon           00110011111111111110100110--0000000000----0---000-
Amphimedon      00110011111111111121100110--00000000000---0---000-
Trichoplax      0011001111111100000-0000210--01111110000--00---000-
Mnemiopsis      0011001111111100000-00112112001100101111121110011110
Nematostella    00110011111110000000112111001100101111111101111111
Hydra           0011001111111100000-00112111001100101111111110111111

;

END;
```

Details about the data

```
#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia         00?1??1???????????0???1?211201??????????0-?-????????
Laccaria        110?001000000000000-000--00--00---00--0--------000-
Capsaspora      001011-00------000-00---00--------0----------0-0-
Monosiga        001000-11------000-00---00--------0----------0-0-
Sycon           00110011111111111110100110--0000000000----0---000-
Amphimedon      00110011111111111121100110--00000000000---0---000-
Trichoplax      0011001111111100000-0000210--01111110000--00---000-
Mnemiopsis      0011001111111100000-00112112001100101111211100111110
Nematostella    00110011111110000000112111001100101111111101111111
Hydra           0011001111111100000-00112111001100101111111110111111

;

END;
```

Initiates the data matrix

```
#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia         00?1??1??????????0???1?211201??????????0-?-????????
Laccaria        110?00100000000000-000--00--00---00--0--------000-
Capsaspora      001011-00------000-00---00--------0----------0-0-
Monosiga        001000-11------000-00---00--------0----------0-0-
Sycon           00110011111111111110100110--0000000000----0---000-
Amphimedon      00110011111111111121100110--00000000000---0---000-
Trichoplax      00110011111100000-0000210--01111110000--00---000-
Mnemiopsis      00110011111100000-00112112001100101111211100111110
Nematostella    00110011111110000000112111001100101111111101111111
Hydra           00110011111100000-00112111001100101111111110111111

;

END;
```

Phylogenetic data matrix

Taxon names left, characters right after tabs or spaces

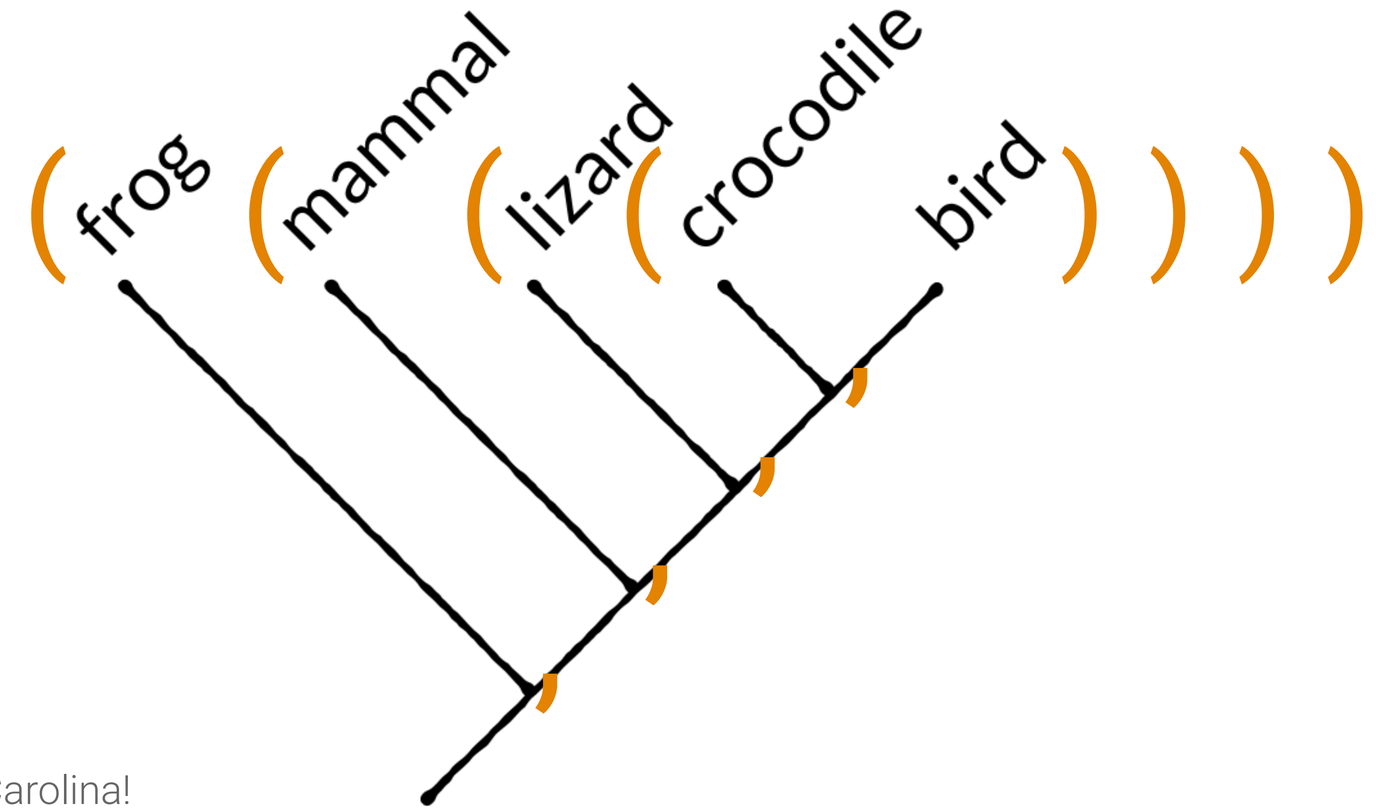
```
#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia         00?1??1???????????0???1?211201??????????0-?-????????
Laccaria        110?001000000000000-000--00--00---00--0--------000-
Capsaspora      001011-00------000-00---00--------0----------0-0-
Monosiga        001000-11------000-00---00--------0----------0-0-
Sycon           00110011111111111110100110--0000000000----0---000-
Amphimedon      00110011111111111121100110--00000000000---0---000-
Trichoplax      0011001111111100000-0000210--01111110000--00---000-
Mnemiopsis      0011001111111100000-00112112001100101111211100111110
Nematostella    00110011111110000000112111001100101111111101111111
Hydra           0011001111111100000-00112111001100101111111110111111

;

END;
```

Initiates the end of the data matrix

```
#NEXUS

[Matrix modified from Dunn (2016)]

BEGIN DATA;

DIMENSIONS  NTAX=10 NCHAR=117;
    FORMAT DATATYPE = STANDARD SYMBOLS= " 0 1 2 3" MISSING=? GAP=- ;

MATRIX

Charnia         00?1??1???????????0???1?211201??????????0-?-????????
Laccaria        110?001000000000000-000--00--00---00--0--------000-
Capsaspora      001011-00------000-00---00--------0----------0-0-
Monosiga        001000-11------000-00---00--------0----------0-0-
Sycon           00110011111111111110100110--0000000000----0---000-
Amphimedon      00110011111111111121100110--00000000000---0---000-
Trichoplax      0011001111111100000-0000210--01111110000--00---000-
Mnemiopsis      0011001111111100000-00112112001100101111211100111110
Nematostella    00110011111110000000112111001100101111111101111111
Hydra           0011001111111100000-00112111001100101111111110111111

;

END;
```

Initiates the end of the nexus block

Newick format



Named after Newick's Lobster House, North Carolina!

```
(frog, (mammal, (lizard, (crocodile, bird))));
```