



Introduction to Bayesian phylogenetic inference

Joëlle Barido-Sottani

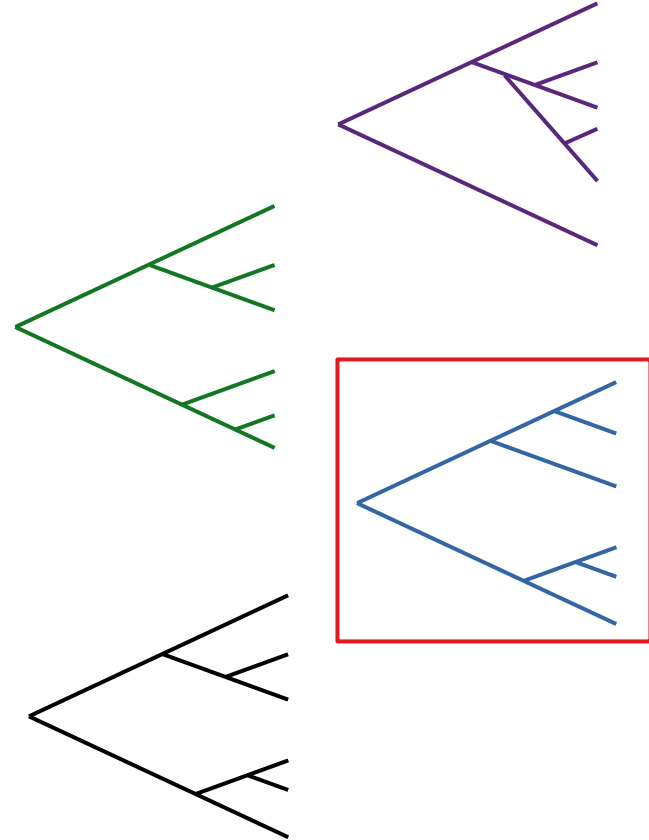
What is inference ?

ACAGACTTTTCAGACTTTTCAGACCC
ACACACCTACAGACTTACAGACCC
TCAGACTTTTCACACCTTCAGACCT
TCACACCTACACACCCACAGACTT
TCACACCTACACACCCACAGACTT
TCAGACTTTTCACACCTTCAGACCT

Observations

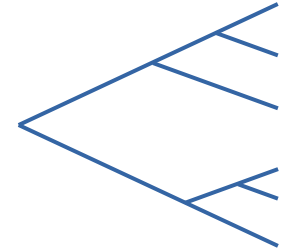
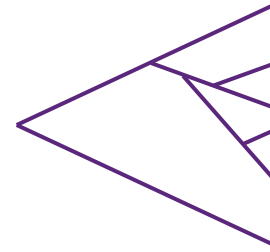
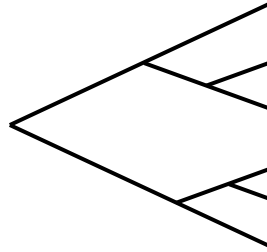
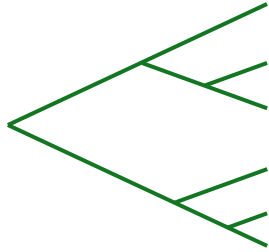


Explanations



Requirements for inference

Choice of
model



Ranking
function

$$P(\text{Green Tree} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix})$$

$$P(\text{Black Tree} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix})$$

$$P(\text{Purple Tree} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix})$$

$$P(\text{Blue Tree} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix})$$

Inference = optimizing **parameters** within a **model**
to fit **observations**

What is probability ?



Frequentist approach

- Based on repeated experiments
- $N = 1000$ dice rolls, $n = 210$ rolls with value 5
 $\Rightarrow P(\text{dice} = 5) = n/N = 0.21$

Issues

- Assumes that experiments can be repeated
- Assumes that the underlying system is random

What is probability ?



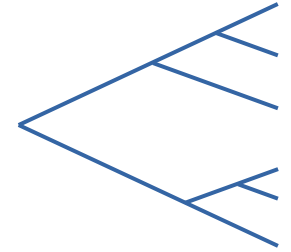
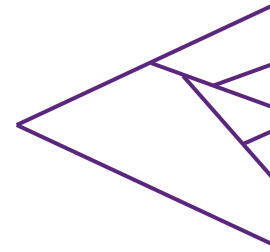
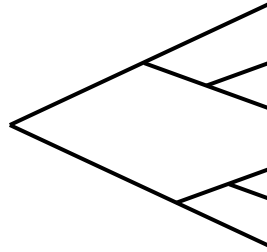
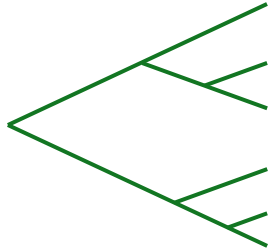
Bayesian approach

- Probability measures how plausible an outcome is based on available information
- $P(\text{dice} = 5 \mid \text{no information}) = 1/6$
 $P(\text{dice} = 5 \mid \text{dice is unfair}) = 0.01$
 $P(\text{dice} = 5 \mid \text{perfect information}) = 1$

=> Probability expresses the level of certainty

Requirements for inference

Choice of
model



Ranking
function

$$P(\text{Green Tree} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix})$$

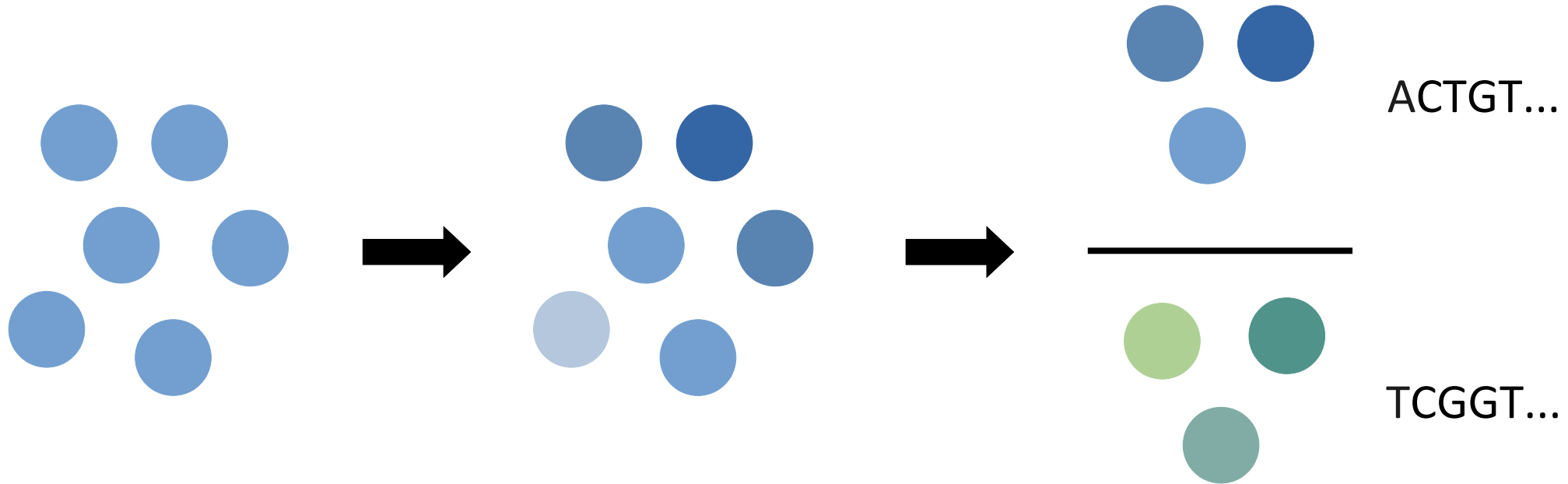
$$P(\text{Black Tree} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix})$$

$$P(\text{Purple Tree} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix})$$

$$P(\text{Blue Tree} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix})$$

Inference = optimizing **parameters** within a **model**
to fit **observations**

Generative models of evolution

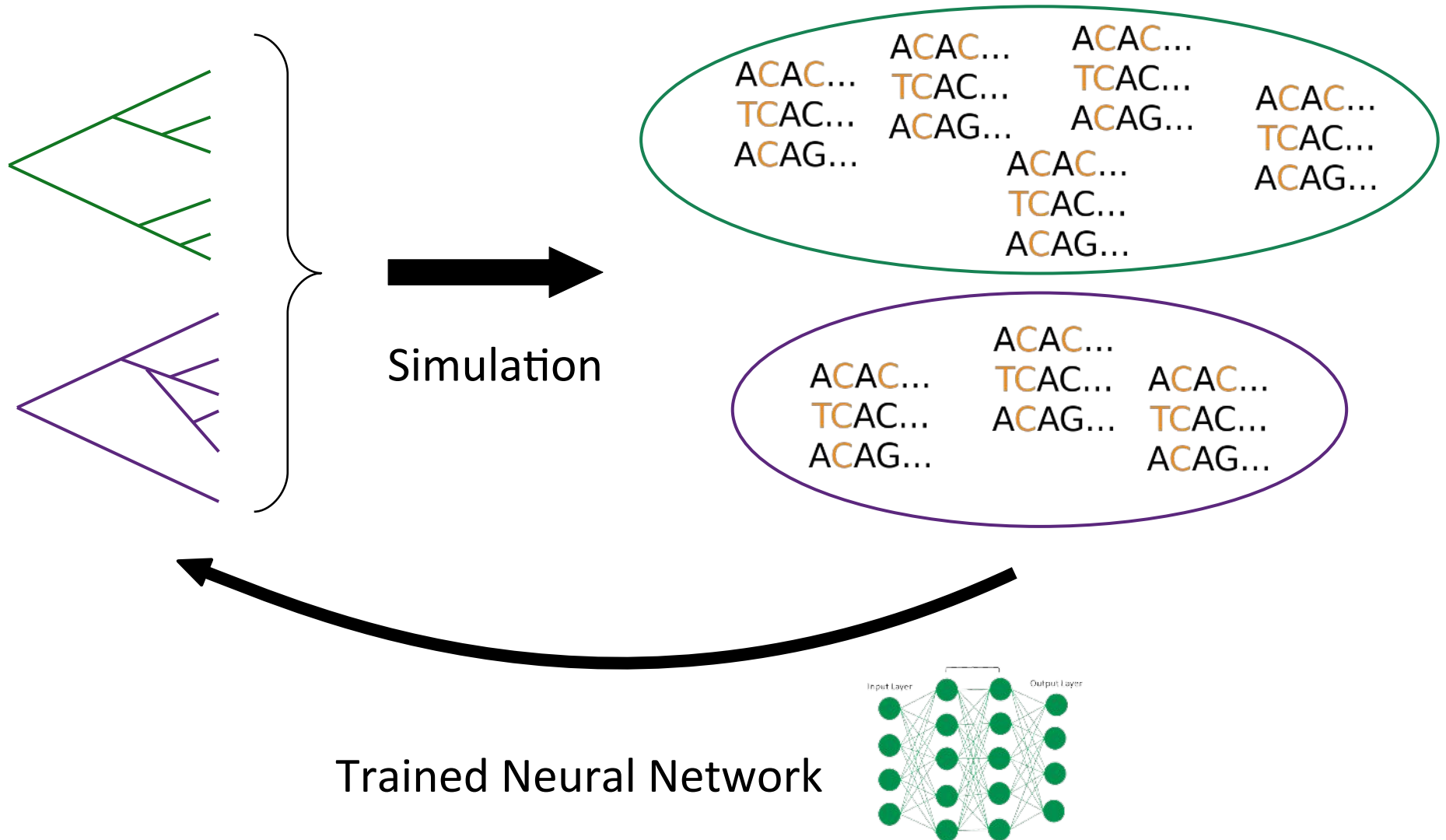


The data is the outcome of the model
=> we can calculate $P(\text{data} | \text{parameters})$

Inference based on generative models

- What we want: $P(\text{parameters} \mid \text{data})$ probability of model parameters given our observed data
- What we have: $P(\text{data} \mid \text{parameters})$ likelihood i.e. probability of generating the data given the model parameters
- Maximum likelihood approach
=> Use the likelihood $P(\text{data} \mid \text{parameters})$ as ranking function

Deep learning approach



Bayes' theorem for inference

$$P(\text{param} \mid \text{data}) = \frac{P(\text{data} \mid \text{param}) P(\text{param})}{P(\text{data})}$$

The diagram illustrates the components of Bayes' theorem for inference. The equation is written as $P(\text{param} \mid \text{data}) = \frac{P(\text{data} \mid \text{param}) P(\text{param})}{P(\text{data})}$. Red arrows point from labels to specific parts of the equation: 'Likelihood' points to $P(\text{data} \mid \text{param})$, 'Prior' points to $P(\text{param})$, 'Marginal likelihood of the data' points to $P(\text{data})$, and 'Posterior' points to $P(\text{param} \mid \text{data})$. The terms 'param' and 'data' are color-coded: 'param' is blue and 'data' is orange.

Likelihood

Posterior

Marginal likelihood of the data

Prior

Bayes' theorem for inference

The data and model parameters are described by probabilities

- **Prior** : $P(\text{param}) \Rightarrow$ the range of *plausible* parameter values
NB : All model parameters have priors
- **Likelihood** : $P(\text{data} | \text{param}) \Rightarrow$ the likelihood is proportional to the probability of observing the data given a hypothesis
- **Posterior** : $P(\text{param} | \text{data}) \Rightarrow$ combines information from the data (likelihood) and previous knowledge (prior)
- **Marginal likelihood** : $P(\text{data}) \Rightarrow$ probability of the data given the chosen model(s) over all possible parameter values

A note on priors

- Priors should be **distinct** from the data
 - Previous literature (on a different dataset)
 - Knowledge of biological processes
- Estimates are influenced by both priors **and** data
- Are other types of analyses free of priors?
 - ML inference : all values are equally likely – implicit *uniform* prior
 - DL inference : priors given by the training dataset
 - More generally : post-processing choices **are priors**
e.g. investigating further a value which seems absurd

Bayesian phylogenetic and phylodynamic tools

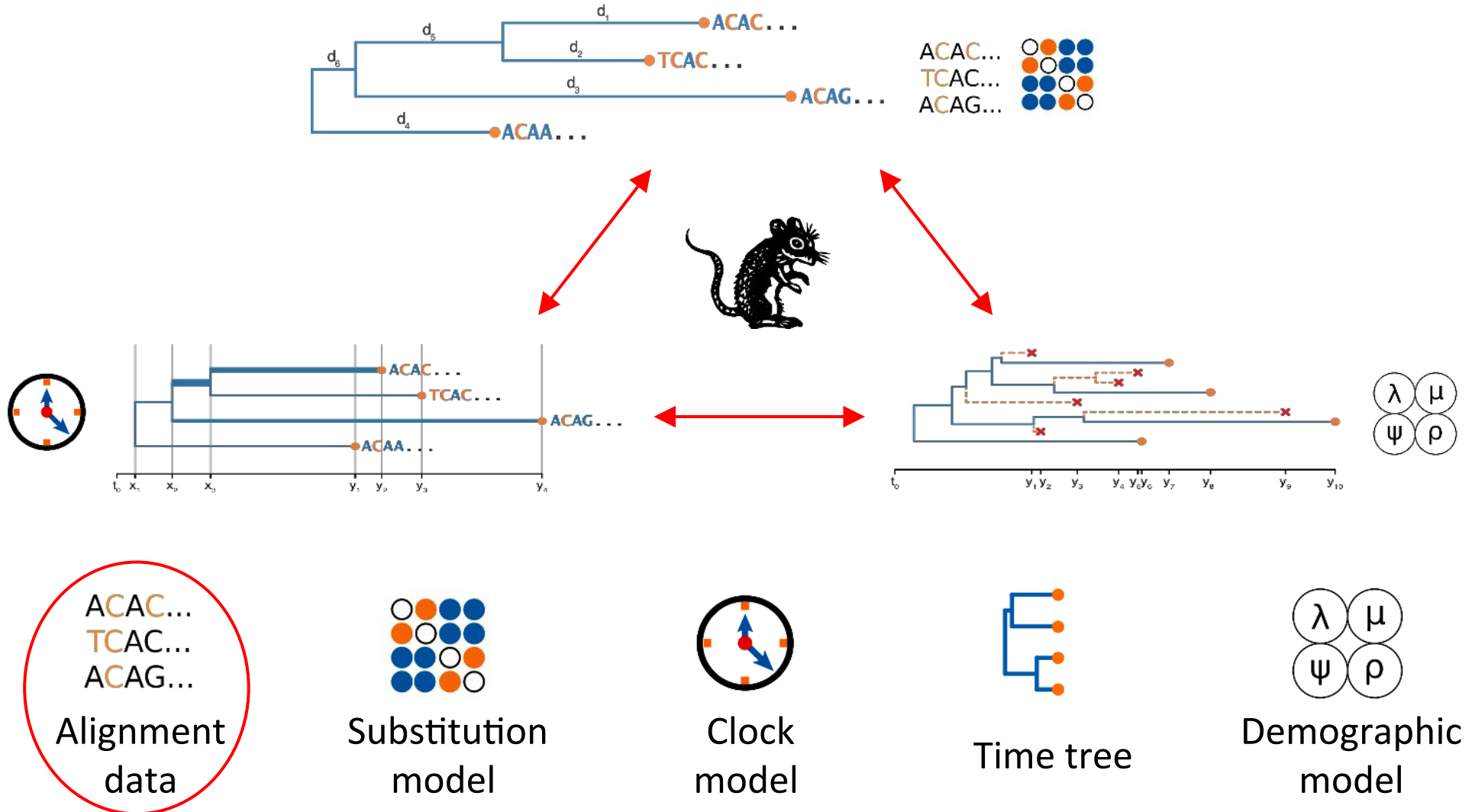
- BEAST & BEAST2
- MrBayes & RevBayes
- PhyloBayes (focus on protein alignments)
- Bali-Phy (estimating the alignment)
- SCAR (focus on recombination)
- Many more.....



Beast2

Bayesian evolutionary analysis by sampling trees

What goes into a **BEAST2** model?

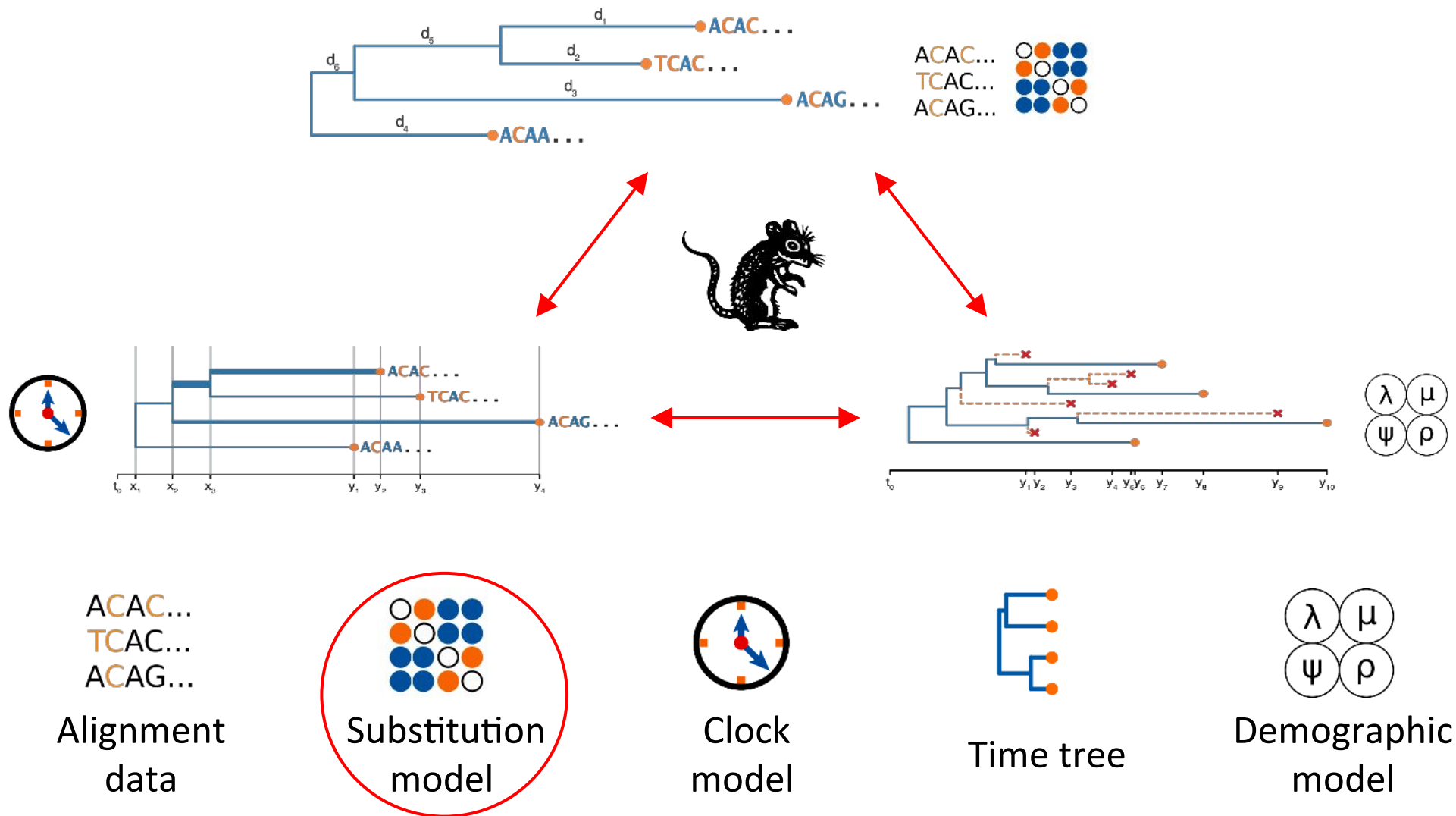


The alignment data

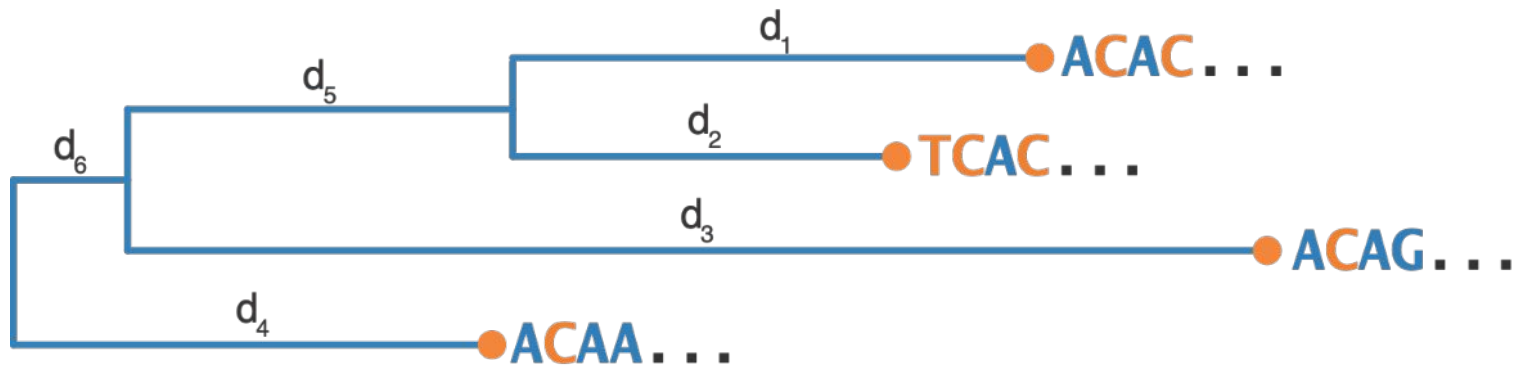
ACAC...
TCAC...
ACAG...

- Typically an alignment of DNA or RNA sequences
- Can also be amino acids or codons
- Sampled at one point in time or several
- Is often split into multiple partitions
 - Multiple genes
 - 1st, 2nd and 3rd codon positions

What goes into a **BEAST2** model?



Substitution/site model



Genetic distance from common ancestor

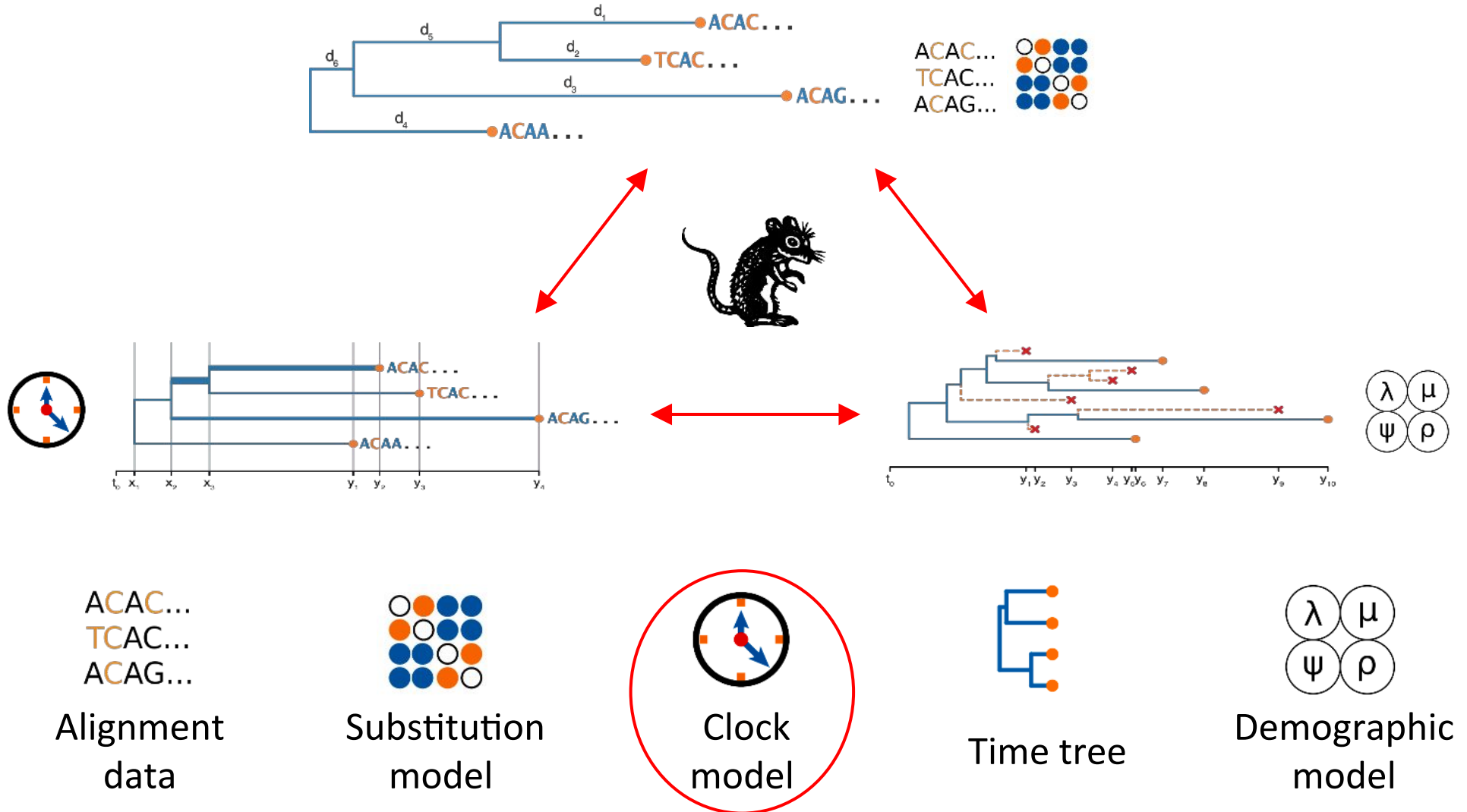
	A	T	C	G
A	○	●	●	●
T	●	○	●	●
C	●	●	○	●
G	●	●	●	○

+

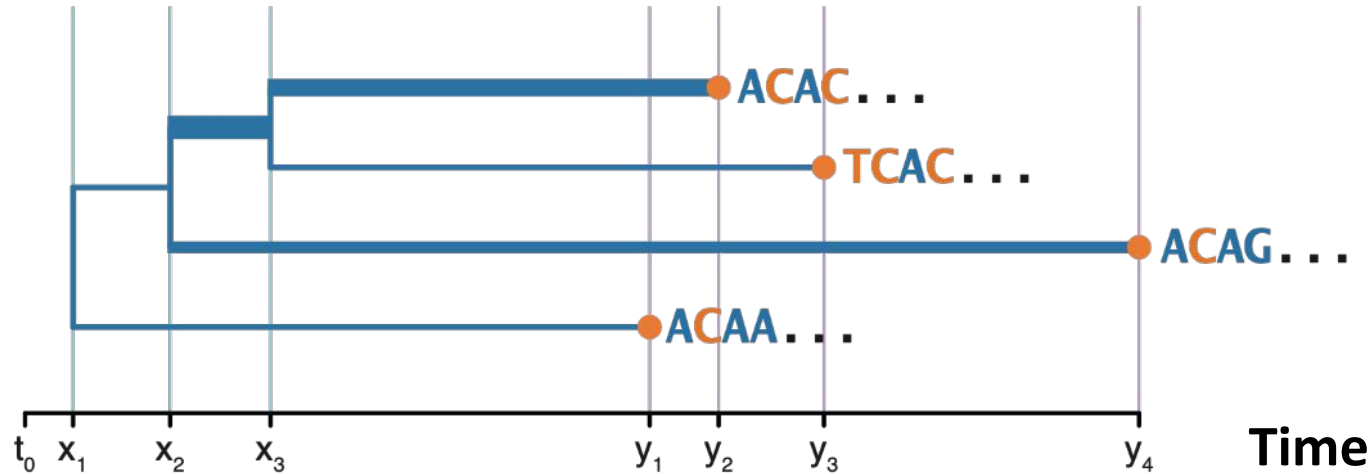
$(\pi_T, \pi_C, \pi_A, \pi_G)$

- Links the genome sequences to the genealogy
- We observe sequences at the tips, not their histories
- Not all substitutions are observed (multiple substitutions at the same site, reverse substitutions)

What goes into a **BEAST2** model?

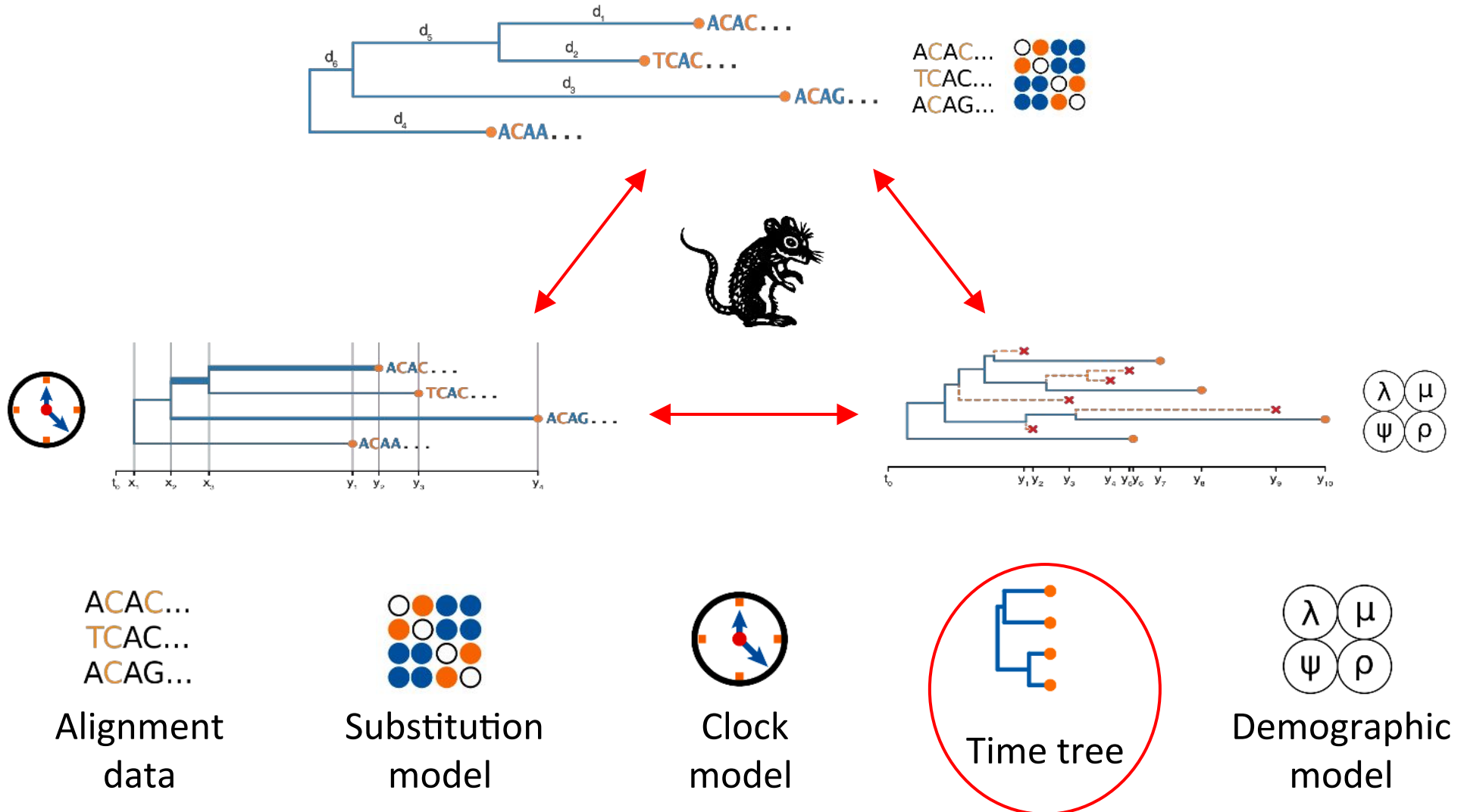


Molecular clock model

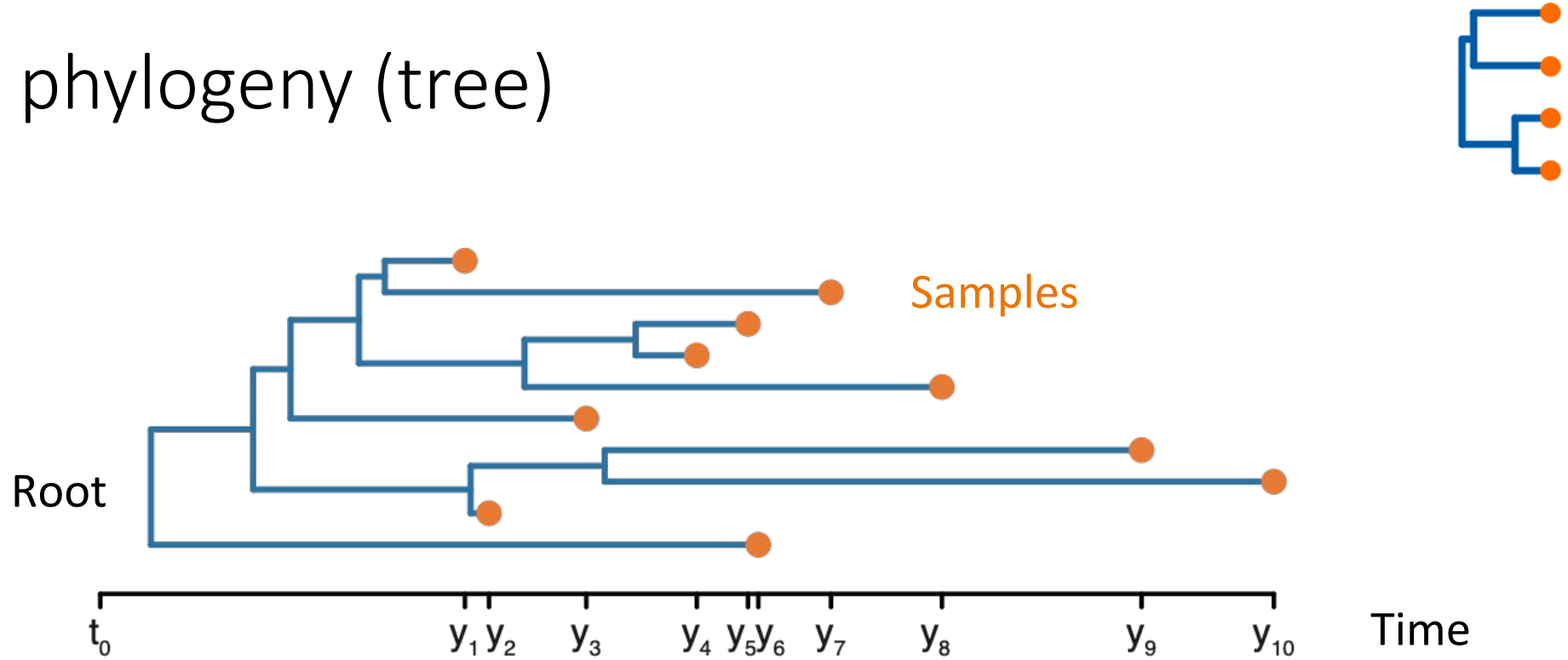


- Scales branch lengths to calendar time => how long does it take for substitutions to appear?
- Different branches may have different clock rates
- Time information is needed to calibrate the clock

What goes into a **BEAST2** model?

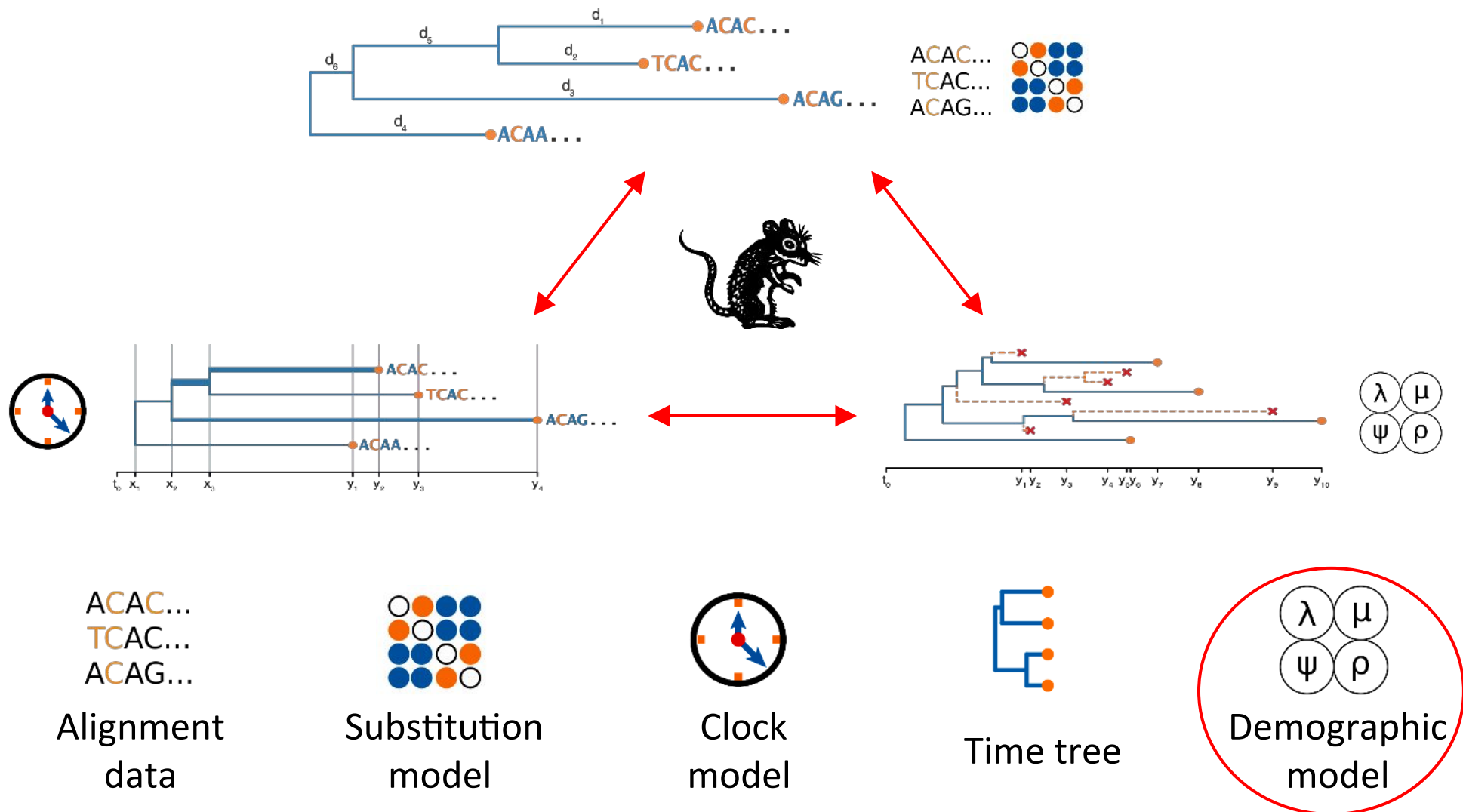


The phylogeny (tree)

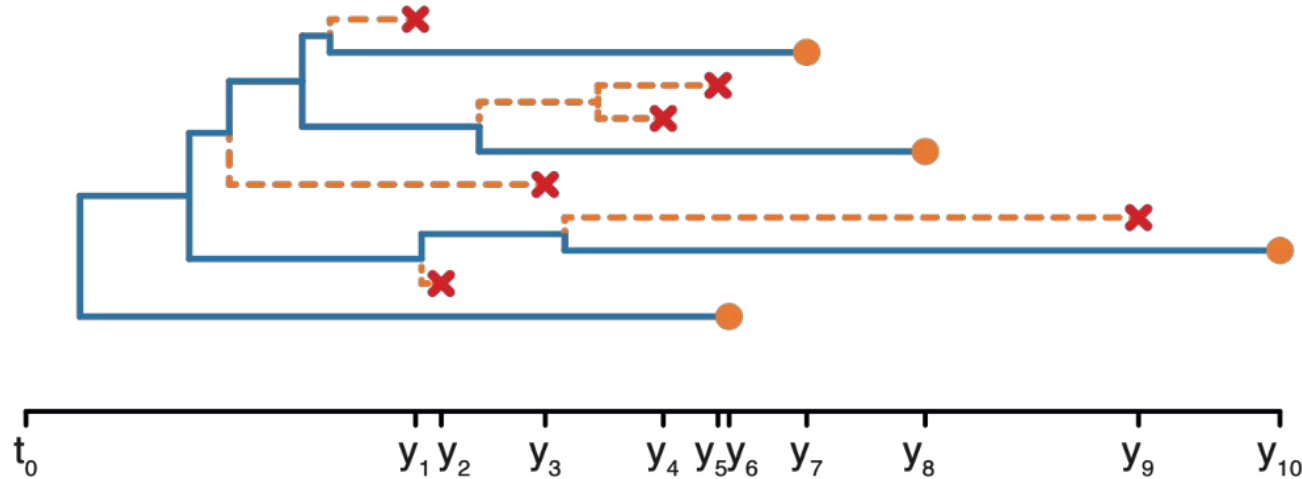


- Phylogenies in phylodynamics are **rooted, time trees**
- Displays the ancestral relationships between the **sampled** sequences and the divergence times

What goes into a **BEAST2** model?

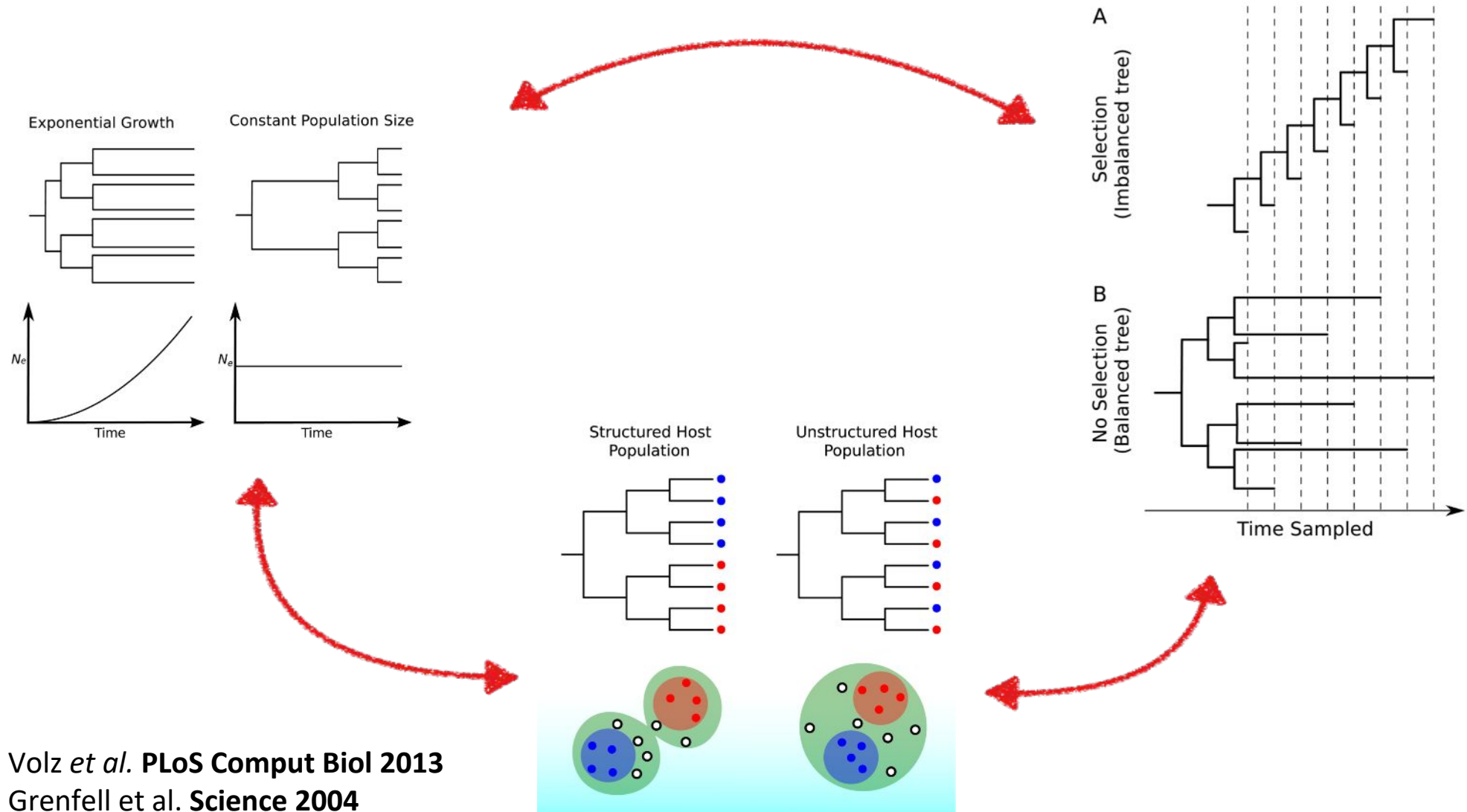


Demographic (tree) model

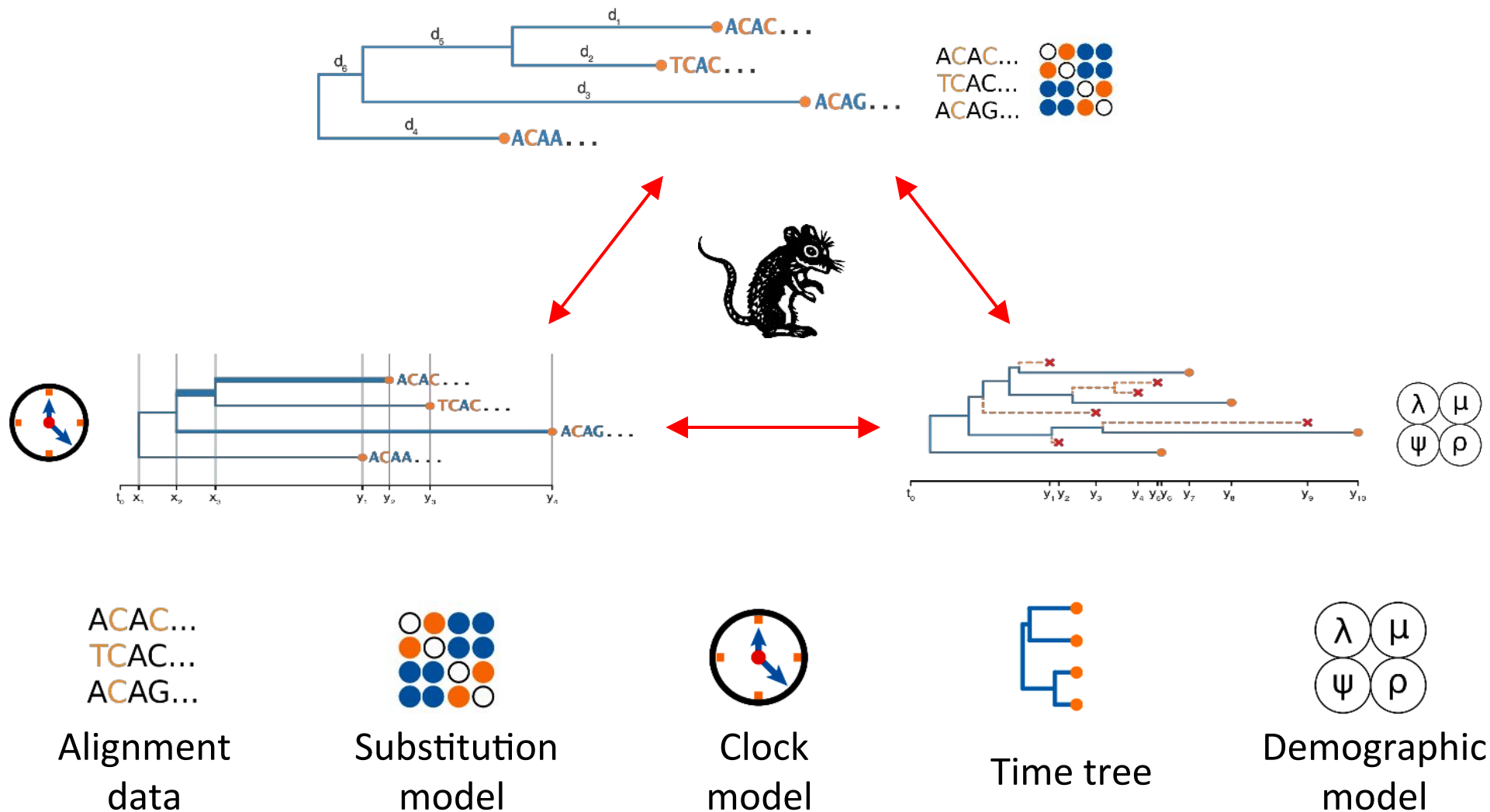


- Serves as tree prior (required since the tree is a parameter)
- Describes the population dynamics
 - How does the infected population grow over time?
 - How does the transmission rate change over time?
- Usually a birth-death or a coalescent model

Different population dynamics generate different trees



What goes into a **BEAST2** model?



Final posterior distribution

$$P(\text{Posterior}) =$$

Posterior

Phylogenetic
Likelihood

Phylodynamic
likelihood

Model priors

$$P(\text{ACAC... TCAC... ACAG...} \mid \text{Substitution model, Clock model, Time tree, Demographic model})$$

$$P(\text{ACAC... TCAC... ACAG...})$$

ACAC...
TCAC...
ACAG...

Alignment
data



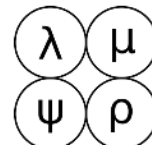
Substitution
model



Clock
model



Time tree



Demographic
model

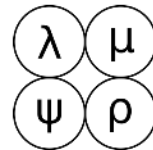
Final posterior distribution – fixed tree

Phylogenetic likelihood

$$\underbrace{P\left(\begin{array}{cc} \lambda & \mu \\ \psi & \rho \end{array} \middle| \begin{array}{c} \text{Tree} \end{array}\right)}_{\text{Posterior}} = \frac{\underbrace{P\left(\begin{array}{cc} \lambda & \mu \\ \psi & \rho \end{array} \middle| \begin{array}{c} \text{Tree} \end{array}\right)}_{\text{Phylogenetic likelihood}} \underbrace{P\left(\begin{array}{cc} \lambda & \mu \\ \psi & \rho \end{array}\right)}_{\text{Model priors}}}{P\left(\begin{array}{c} \text{Tree} \end{array}\right)}$$



Time tree

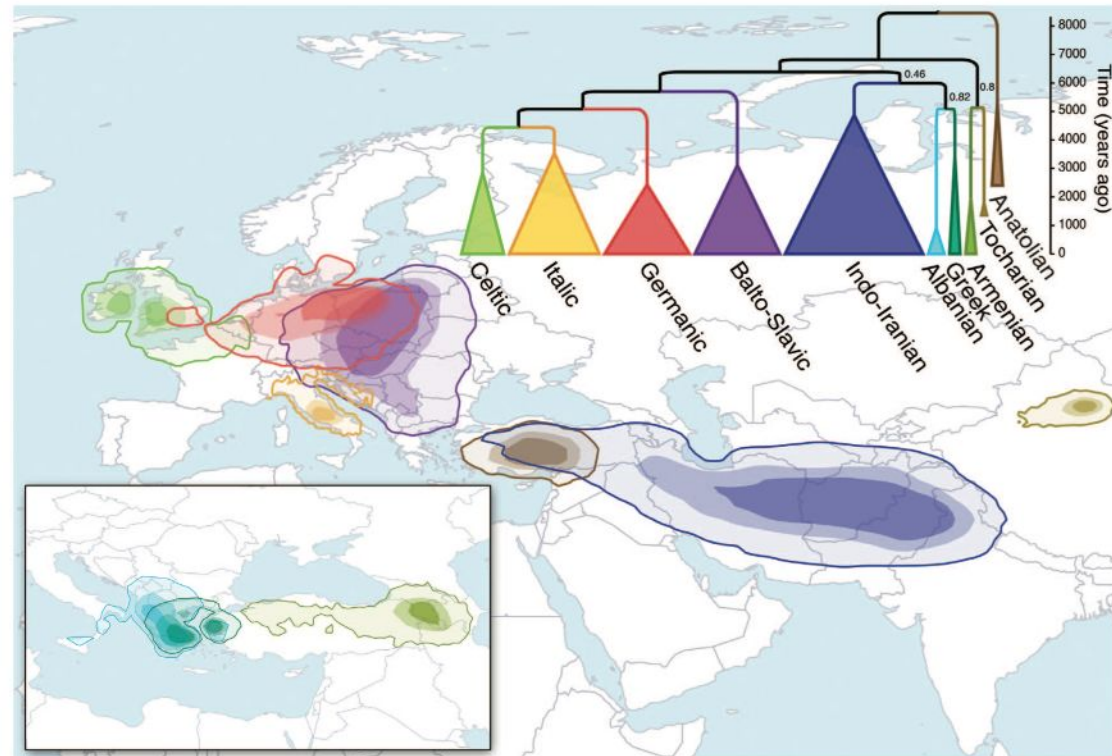


Demographic
model

Some special cases I

Site models don't have to be on nucleotides

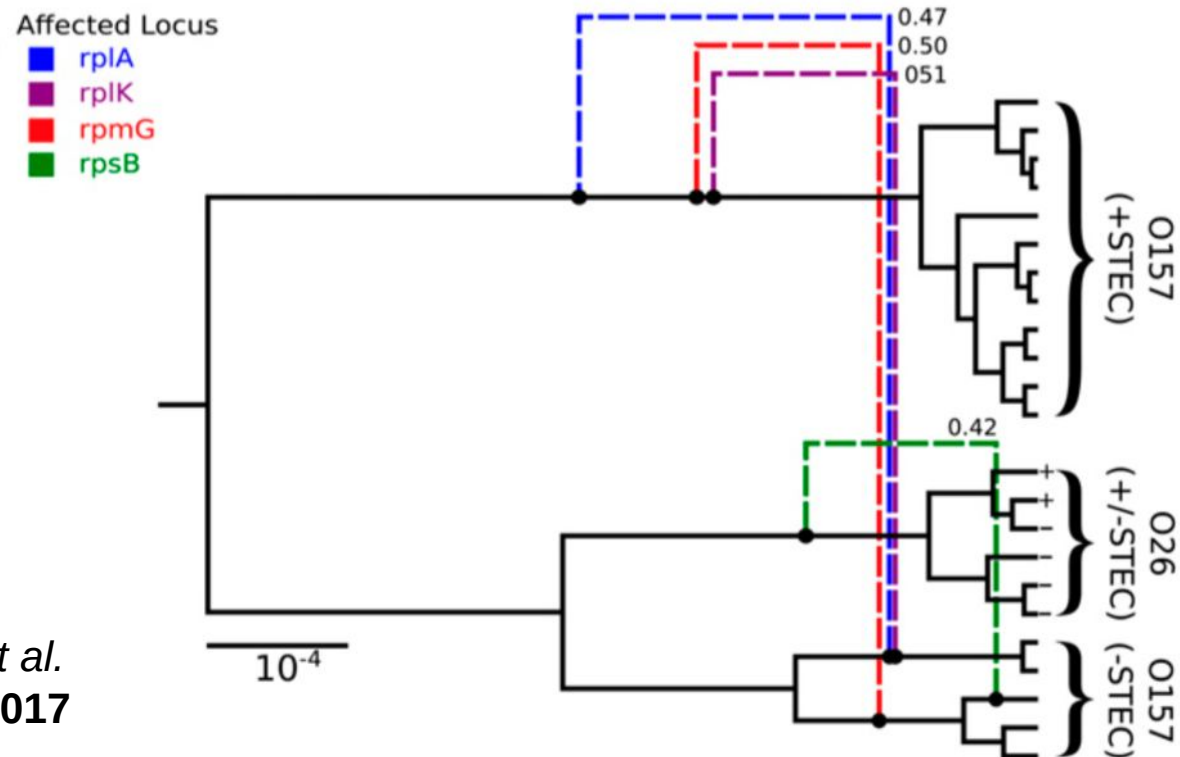
=> Could be on amino acids, morphological traits, roots of words etc.



Bouckaert *et al.*
Science 2012


Special cases II

BEAST2 doesn't always use trees!



Vaughan *et al.*
Genetics 2017

Inference in practice – calculating the posterior

$$P(\text{param} \mid \text{data}) = \frac{P(\text{data} \mid \text{param}) P(\text{param})}{P(\text{data})}$$


$$P(\text{data}) = \int P(\text{data} \mid \text{param})$$

All possible **param** values

But the tree is a parameter

How many trees are there ?

$$T_n = (2n - 3)!! = 1 \times 3 \times 5 \times \dots \times 2n - 5 \times 2n - 3$$

Number of tips	4	5	6	7	8	9	10	20	48
Number of trees	15	105	945	10395	135135	2.0×10^6	3.5×10^7	8.2×10^{21}	3.2×10^{70}

For realistic tree size ($n = 136$): $T_n = 2.1 \times 10^{267}$

=> There are too many trees

Calculating the posterior

- We want to calculate the posterior distribution

$$P(\text{grid, clock, tree} \mid \begin{matrix} \lambda & \mu \\ \psi & \rho \end{matrix} \mid \begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix}) = \text{bell curve}$$

- **But** we cannot easily calculate the marginal likelihood

$$P(\begin{matrix} \text{ACAC...} \\ \text{TCAC...} \\ \text{ACAG...} \end{matrix}) = ?$$

=> use **MCMC** (Markov-chain Monte Carlo)

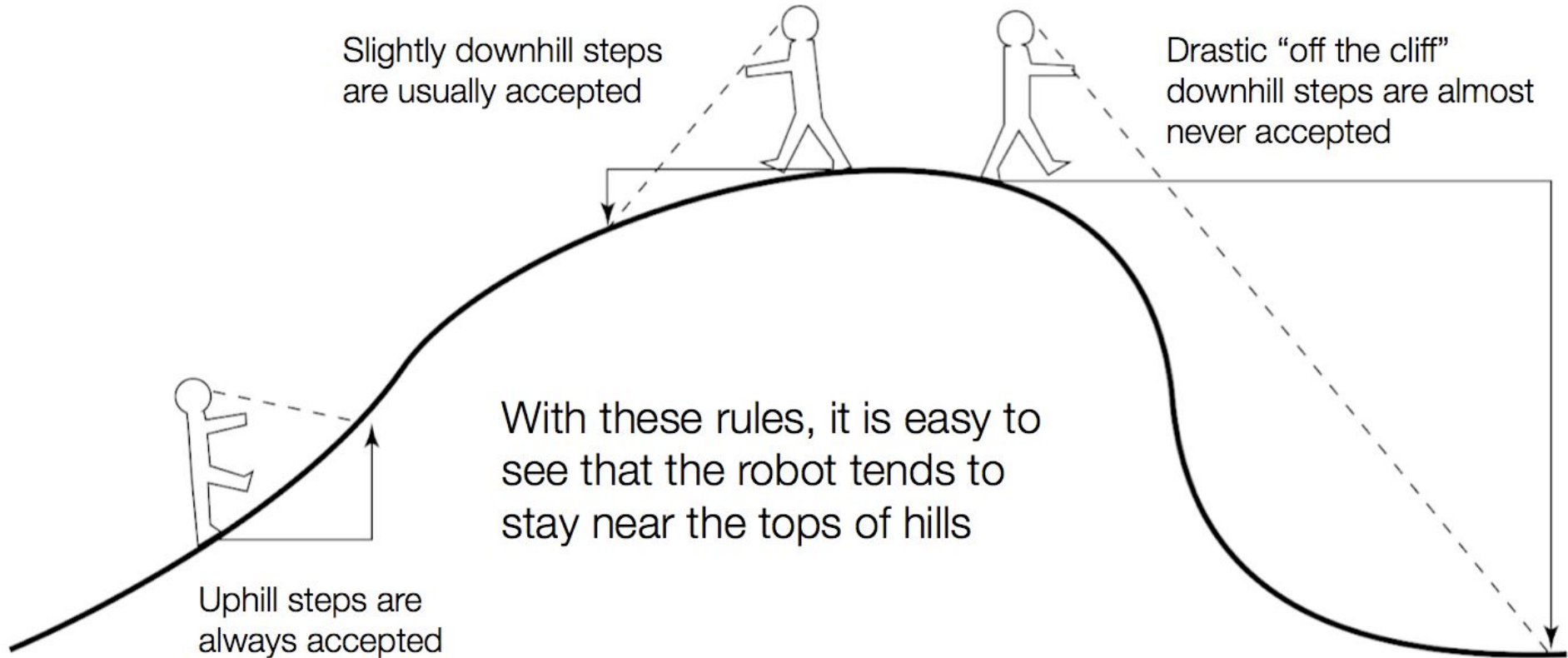
- MCMC performs a random walk in the parameter space, sampling areas based on their posterior value

MCMC (Markov-chain Monte-Carlo)

- MCMC moves through the parameter space and looks for places with high posterior
- For each step we only need to compare which posterior density is higher
=> so we only need the ratio of posteriors

$$\frac{P(\text{model}_1 \mid \text{data})}{P(\text{model}_2 \mid \text{data})} = \frac{\frac{P(\text{data} \mid \text{model}_1) P(\text{model}_1)}{P(\text{data})}}{\frac{P(\text{data} \mid \text{model}_2) P(\text{model}_2)}{P(\text{data})}}$$

MCMC robot (courtesy of Paul Lewis)



MCMC through parameter space

<https://chi-feng.github.io/mcmc-demo/app.html?algorithm=RandomWalkMH>

Operators

- MCMC steps through the state space and samples the posterior
- **Operators/proposals** are used to decide where to step to next
=> a parameter (or multiple parameters) are selected and modified to propose a step
- Operators are part of the **algorithm**, not the model
=> the choice and configuration of operators only affects the **efficiency** of the algorithm
- How to configure operators?
 - Most operators can be **tuned** to change the size of proposed steps
 - Default operators by BEAUti + **auto-tuning** by BEAST2
 - Additional performance **suggestions** at the end of a run

Progress of an inference

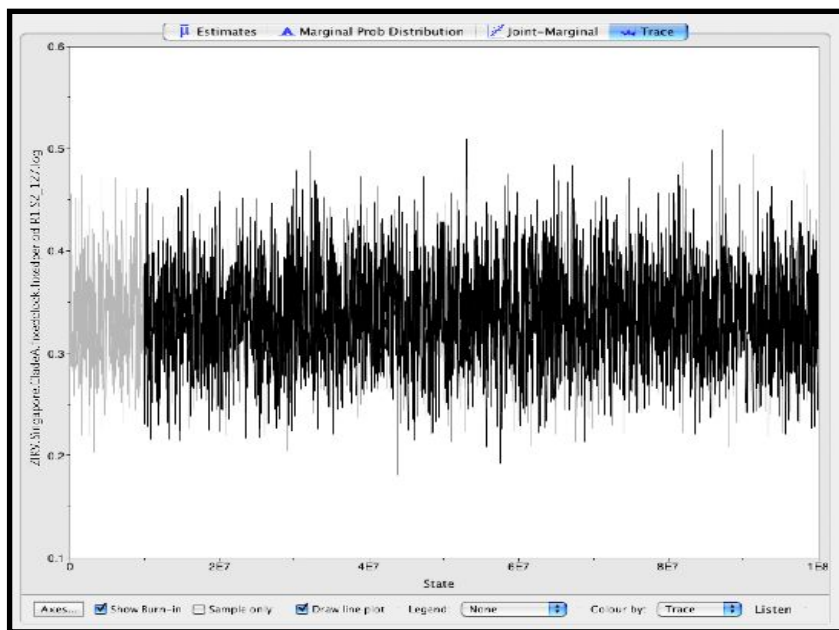
- Initial position – set by the user or by BEAST2
- Burn-in phase: moving from the initial position to the high-posterior space
- Convergence phase: the inference has reached the high-posterior space – still moving but stable
- The posterior estimates are given **only** by samples taken **after convergence**

MCMC inference – when is it done?

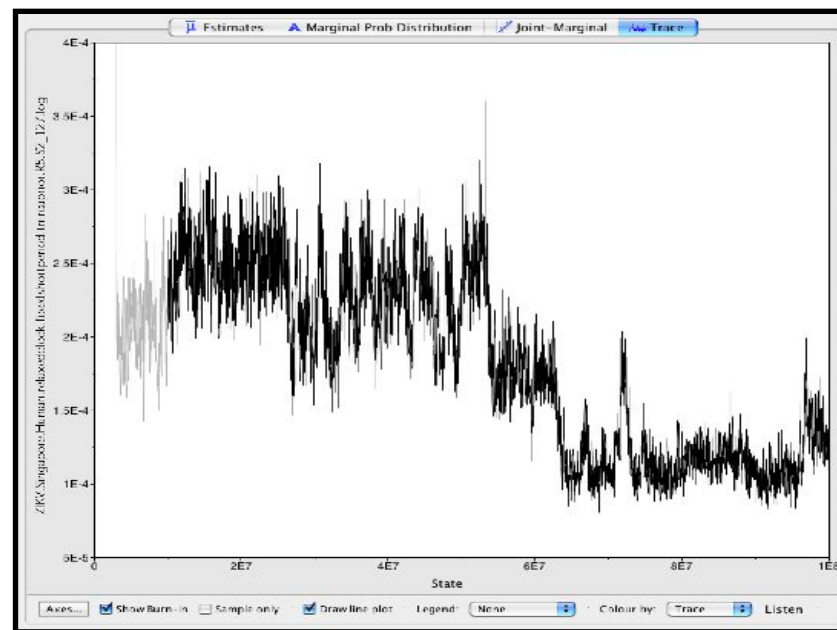
- A proper MCMC inference is guaranteed to converge – but not when!!
- Results obtained before convergence are not reliable
- The number of steps needed depends on many factors
 - Complexity of the analysis (partitions, models, etc...)
 - Size of the dataset
 - Starting values
 - Efficiency of the implementation / operators

Convergence assessment

Looking at the traces



Mixing well! 😊



Not mixing! 😞

The Effective Sample Size (ESS)

- In an MCMC, samples are correlated
=> number of samples in the chain (or the log) \neq number of independent samples
- How do we estimate the number of independent samples?
=> the **effective sample size** (ESS)
- ESS is specific to a particular inference **and** a particular parameter
- ESS > 200 is usually considered ok, more recently higher thresholds (500-600) have been proposed

Trace Files:

Trace File	States	Burn-In
primate-mtDNA....	1000000	100000

+ - Reload

Traces:

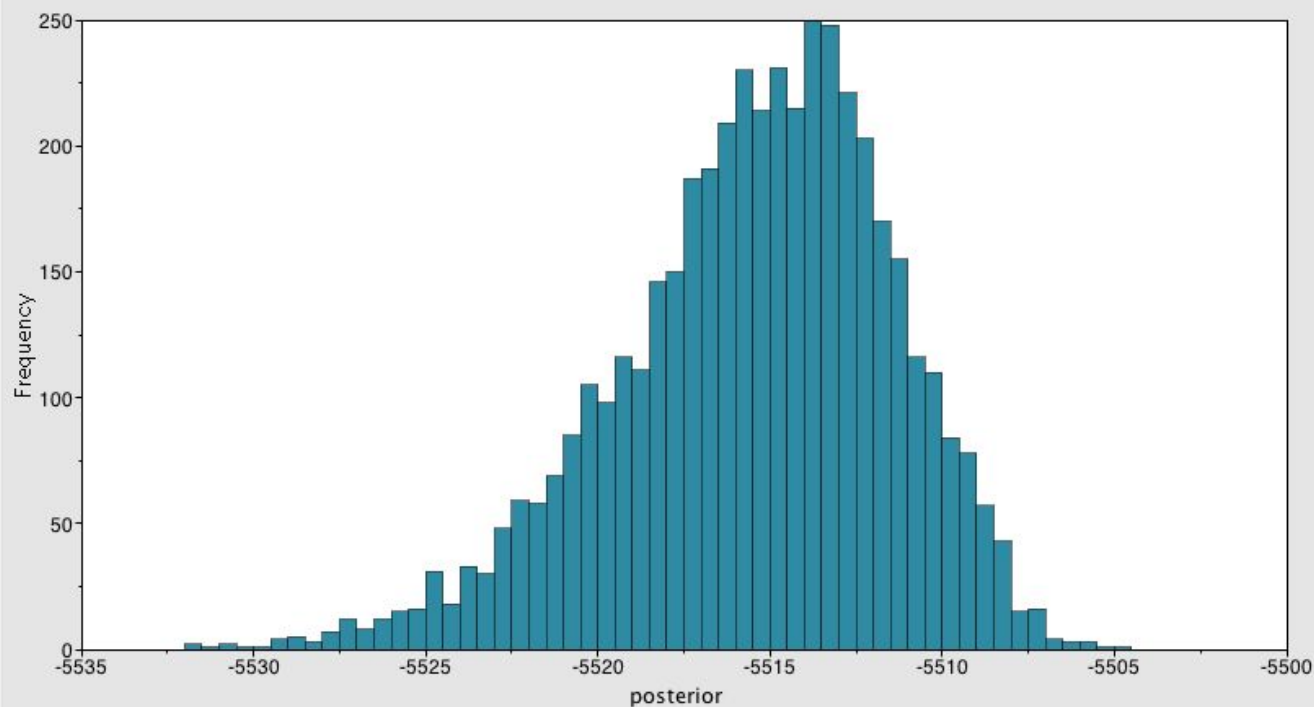
Statistic	Mean	ESS	...
posterior	-5515....	261	R
likelihood	-5442....	335	R
prior	-73.432	113	R
treeLikelihood.1stpos	-1383....	502	R
treeLikelihood.2ndpos	-952.555	321	R
treeLikelihood.3rdpos	-2148....	202	R
treeLikelihood.noncod...	-957.464	192	R
TreeHeight	85.181	235	R
mutationRate.1stpos	0.451	120	R
mutationRate.2ndpos	0.179	124	R
mutationRate.3rdpos	2.955	96	R
mutationRate.noncoding	0.34	180	R
gammaShape.1stpos	0.477	105	R
gammaShape.2ndpos	0.553	76	R
gammaShape.3rdpos	2.998	98	R
gammaShape.noncodi...	0.249	88	R
kappa.1stpos	6.424	86	R
kappa.2ndpos	8.681	79	R
kappa.3rdpos	29.35	42	R
kappa.noncoding	13.619	69	R
CalibratedYuleModel	-47.452	320	R
birthRateY	2.547E-2	731	R
logP(mrca(human-chi...	-0.744	4203	R
mrctime(human-chi...	5.95	2567	R
clockRate	1.165E-2	391	R

Type: (R)real (I)nt (C)at

Estimates Marginal Density Joint-Marginal Trace

Summary Statistic

	posterior
mean	-5515.4884
stderr of mean	0.2487
stdev	4.0176
variance	16.141
median	-5515.0556
value range	[-5531.8494, -5504.8478]
geometric mean	n/a
95% HPD interval	[-5523.3461, -5508.1268]
auto-correlation time (ACT)	3449.4716
effective sample size (ESS)	261
number of samples	4501



Setup...

Bins: 50

Trace Files:

Trace File	States	Burn-In
primate-mtDNA...	1000000	100000
+ -		
Reload		

Traces:

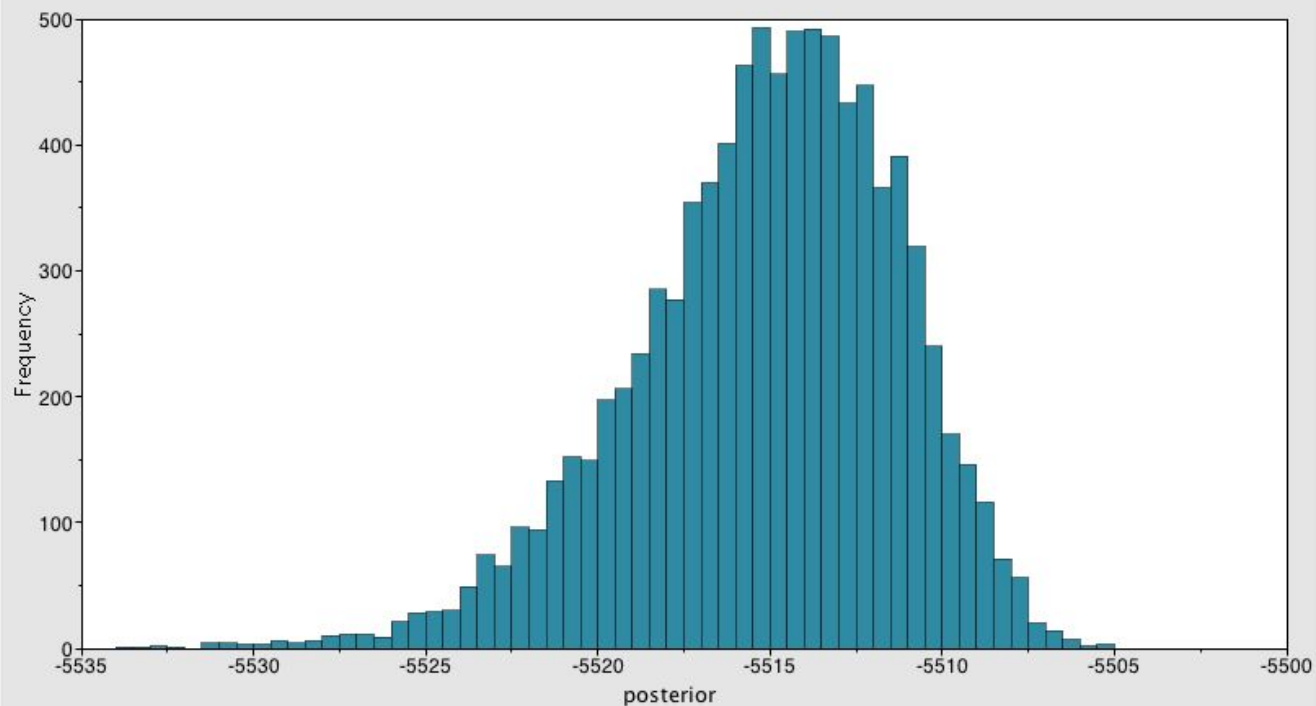
Statistic	Mean	ESS	...
posterior	-5515....	2387	R
likelihood	-5441....	2349	R
prior	-73.169	1379	R
treeLikelihood.1stpos	-1383....	3189	R
treeLikelihood.2ndpos	-952.37	2885	R
treeLikelihood.3rdpos	-2148....	1687	R
treeLikelihood.noncod...	-957.267	1731	R
TreeHeight	83.827	1409	R
mutationRate.1stpos	0.45	852	R
mutationRate.2ndpos	0.182	714	R
mutationRate.3rdpos	2.949	646	R
mutationRate.noncodi...	0.346	1344	R
gammaShape.1stpos	0.496	889	R
gammaShape.2ndpos	0.575	911	R
gammaShape.3rdpos	3.022	726	R
gammaShape.noncodi...	0.244	1006	R
kappa.1stpos	6.235	719	R
kappa.2ndpos	8.5	1359	R
kappa.3rdpos	28.777	365	R
kappa.noncoding	13.478	875	R
CalibratedYuleModel	-47.285	1755	R
birthRateY	2.561E-2	3805	R
logP(mrca(human-chi...	-0.731	9001	R
mrctime(human-chi...	5.949	8655	R
clockRate	1.161E-2	1836	R

Type: (R)eal (I)nt (C)at

[Estimates](#)
[Marginal Density](#)
[Joint-Marginal](#)
[Trace](#)

Summary Statistic

	posterior
mean	-5515.1348
stderr of mean	0.0786
stdev	3.8428
variance	14.7672
median	-5514.7368
value range	[-5533.9705, -5505.0922]
geometric mean	n/a
95% HPD interval	[-5523.187, -5508.5802]
auto-correlation time (ACT)	3770.478
effective sample size (ESS)	2387.2
number of samples	9001



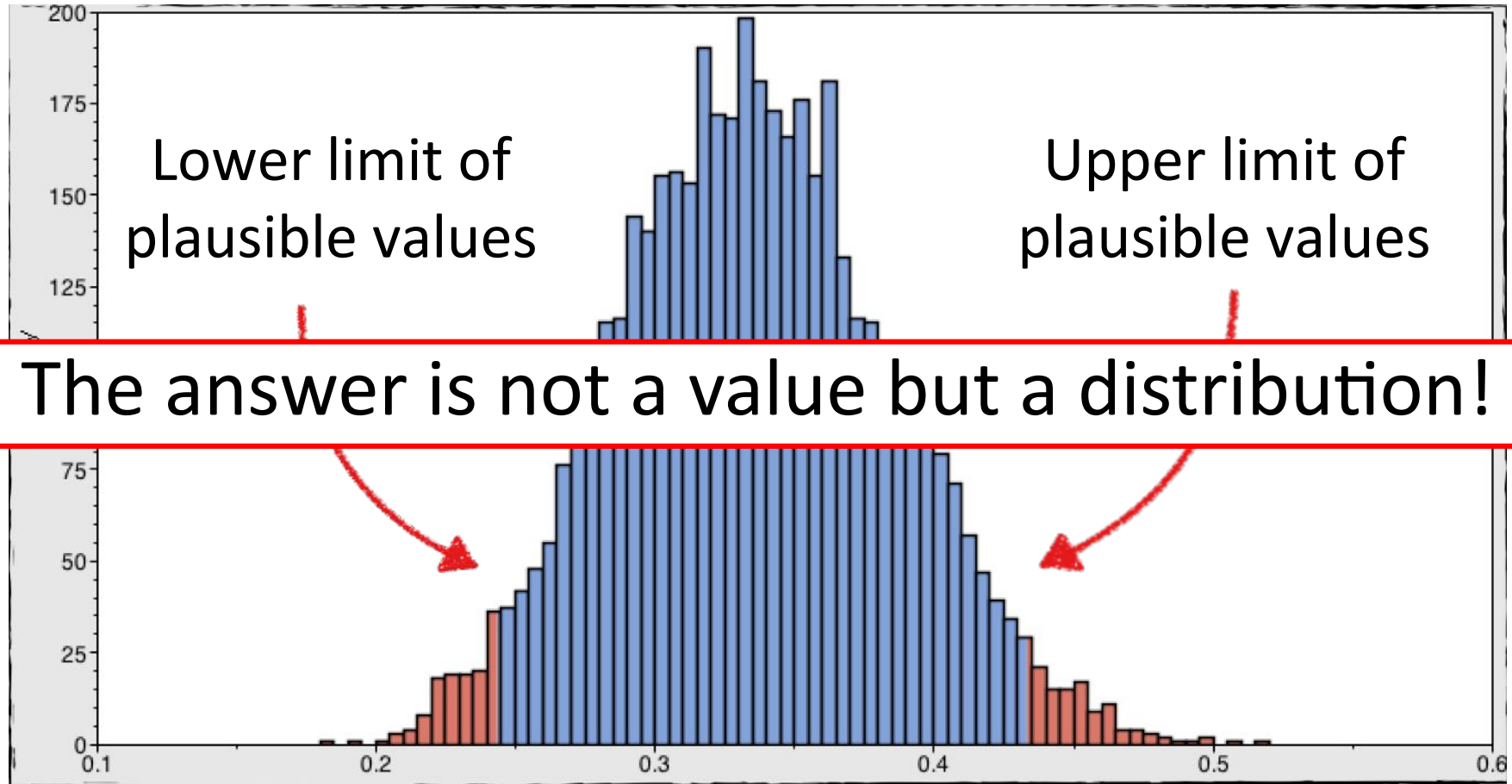
Setup...

Bins: 50

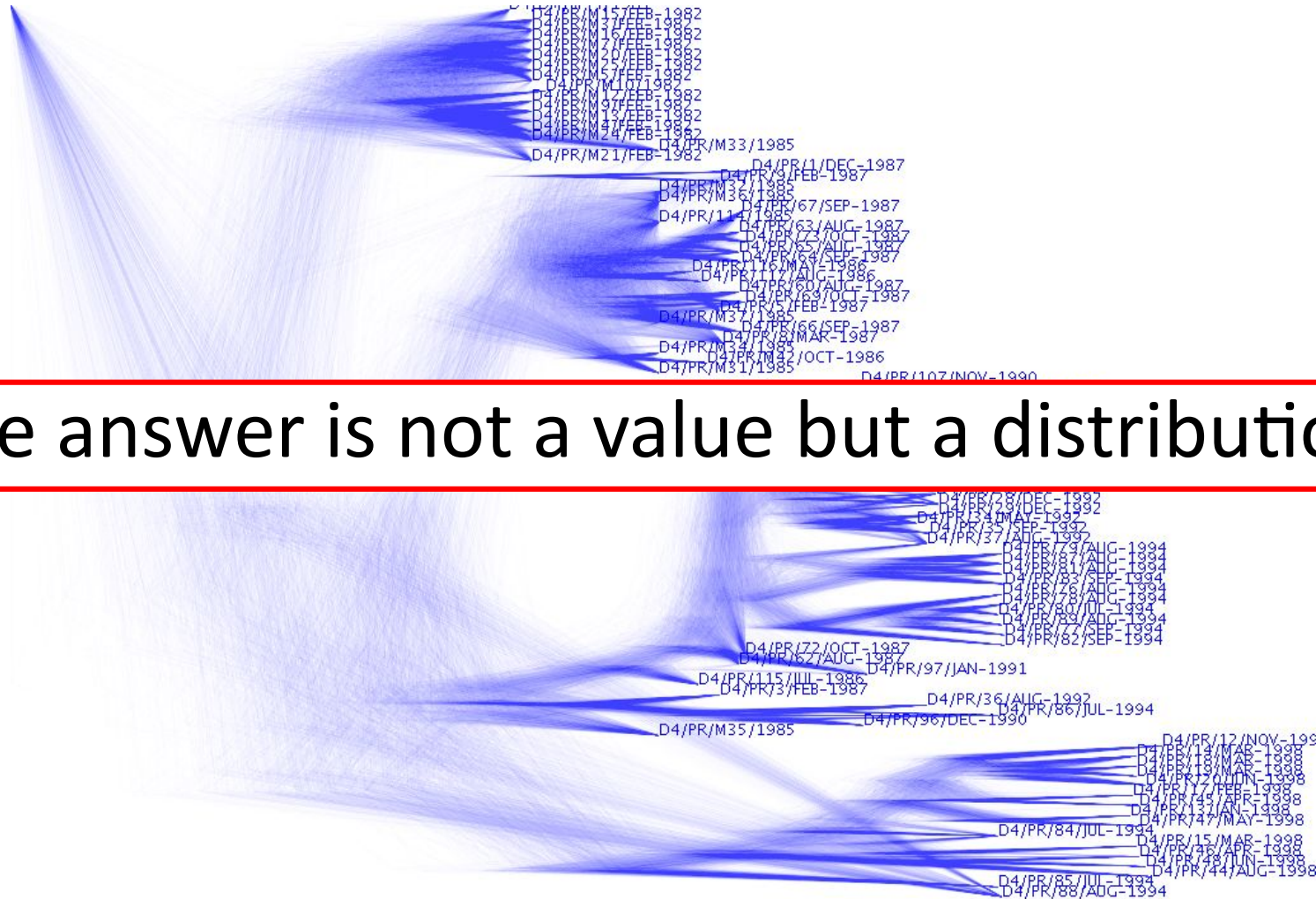
What about tree convergence?

- Does not appear into Tracer **but** can be somewhat estimated from the other parameters
- RWTY (R package): estimates the ESS of the overall tree topology
<https://cran.r-project.org/package=rwty>
- Convenience (R package): estimates the ESS of splits in the tree
<https://github.com/lfabreti/convenience>

Final posterior estimate



Final posterior estimate (tree edition)

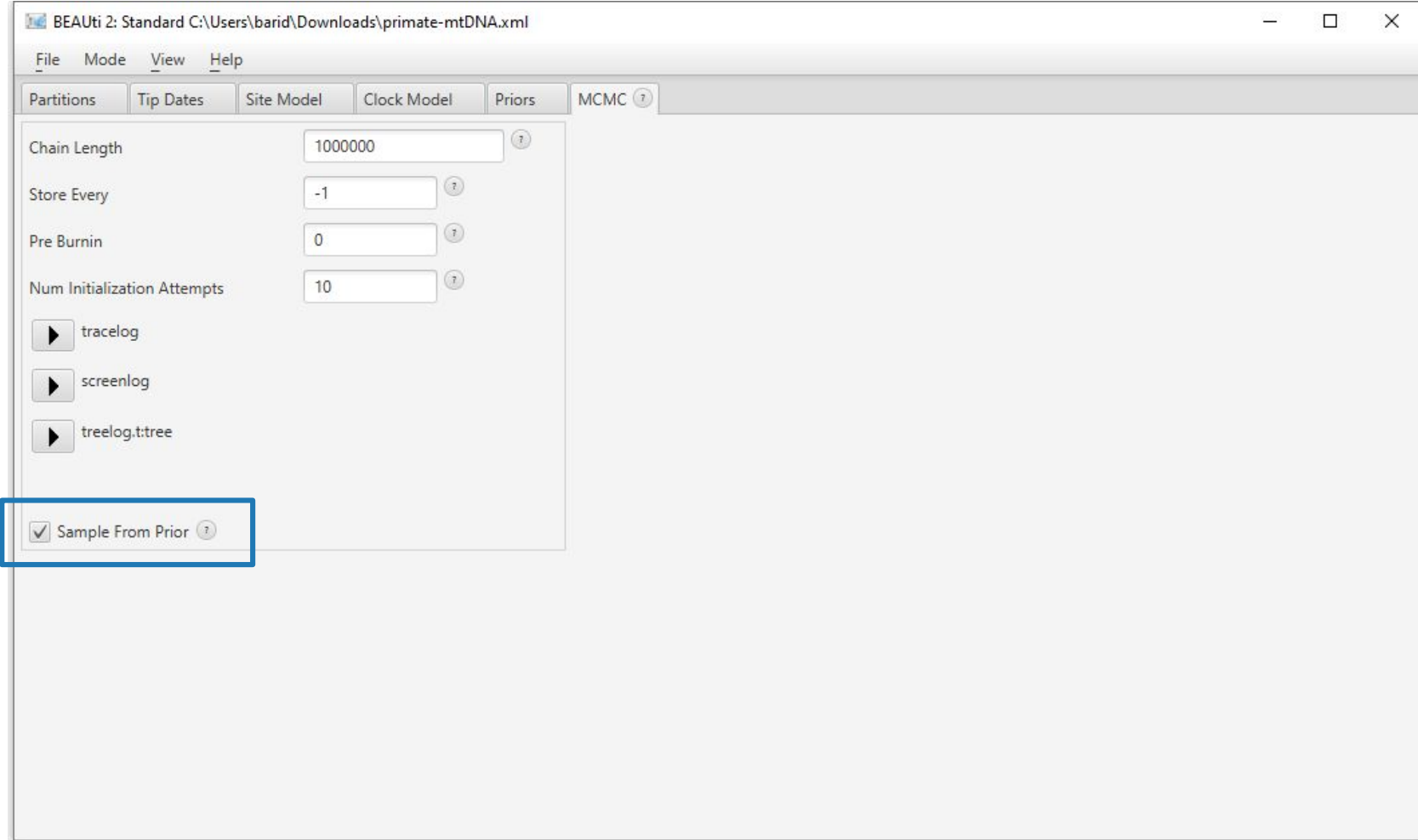


The answer is not a value but a distribution!

Sampling from the prior

- Bayesian analyses include priors on all parameters
 - the chosen priors affect the results
 - if priors interact, the effective prior can be different from the intended prior
 - How do we know whether the results come from the data **or** from the priors?
- => **Solution:** evaluate the results while removing the influence of the data

Sampling from the prior





Tracer

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In
primate-mtDNA_long_sfp.log	10000000	1000000
primate-mtDNA_long.log	10000000	1000000
Combined	18000000	-

+

-

Reload

Traces:

Statistic	Mean	ESS	Type
posterior	-381.17	2105	R
likelihood	◆	-	R
prior	-381.17	2105	R
treeLikelihood.1stpos	◆	-	R
treeLikelihood.2ndpos	◆	-	R
treeLikelihood.3rdpos	◆	-	R
treeLikelihood.noncoding	◆	-	R
Tree.height	1.448E17	4179	R
Tree.treeLength	7.082E17	3980	R
mutationRate.1stpos	0.909	81	R
mutationRate.2ndpos	1.003	122	R
mutationRate.3rdpos	0.843	78	R
mutationRate.noncoding	1.276	48	R
gammaShape.1stpos	1.106	689	R
gammaShape.2ndpos	1.061	1243	R
gammaShape.3rdpos	1.097	1122	R
gammaShape.noncoding	1.164	882	R
kappa.1stpos	5.753	1278	R
kappa.2ndpos	5.49	992	R
kappa.3rdpos	6.034	642	R
kappa.noncoding	5.493	1265	R
CalibratedYuleModel	-396.069	2300	R
birthRateY	3.764E-17	2449	R
logP(mrca(human-chimp))	-0.73	8523	R
mrca.age(human-chimp)	6.004	8930	R
clockRate	0.988	9001	R

Type: (R)eal (I)nt (C)at (T)ime * constant

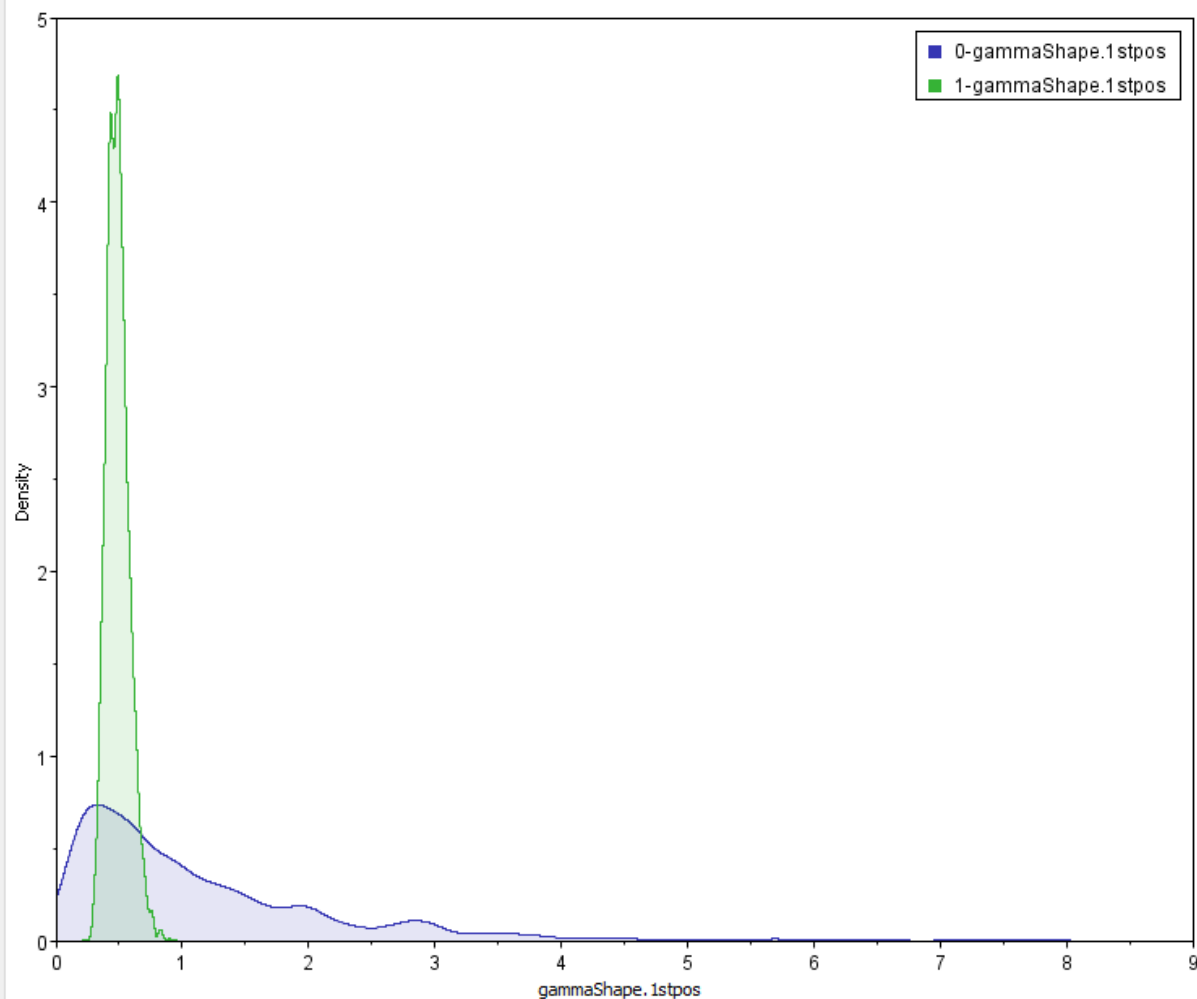
Estimates

Marginal Density

Joint-Marginal

Trace

Display: KDE



Setup... Legend: Top-Right

Colour by: Trace file

Summary trees



TreeAnnotator v2.4.6

Burnin percentage:

Posterior probability limit:

Target tree type:

Node heights:

Target Tree File:

Input Tree File:

Output File:

Low memory: ☐

Summary trees

- Maximum A Posteriori (MAP) tree : sample with the highest posterior
- Maximum clade credibility (MCC) tree: sample with the highest clade score
- **New method:** Conditional Clade Distributions (CCD) trees
 - Based on the full distribution including unsampled trees
 - More accurate than MCC trees
 - Currently **not available** for FBD trees

Summary trees

NB: All summary methods produce a **filtered representation** of the full result

In all analyses:

- Check the **posterior** values at the splits
- Check the uncertainty **around** the node ages
- If the posterior is diffuse, check for **alternative** configurations in the distribution

BEAST2 best practices

Before you begin

- Decide on plausible parameter values
- Plan for necessary vs unnecessary complexity
- Check potential sources of error (e.g. sampling biases)

Before you run the analysis

- Ask someone else to look at your XML file
- Decide on the length of the chain & sampling frequency
=> Aim for **1,000 – 10,000** samples

Actually running the analysis

- Sample from the prior (run without data)
- Run the analysis with multiple chains
- (If necessary) re-evaluate the complexity

BEAST2 best practices

After the analysis

- Assess convergence and mixing
- Combine the chains
- Examine the full posterior distribution

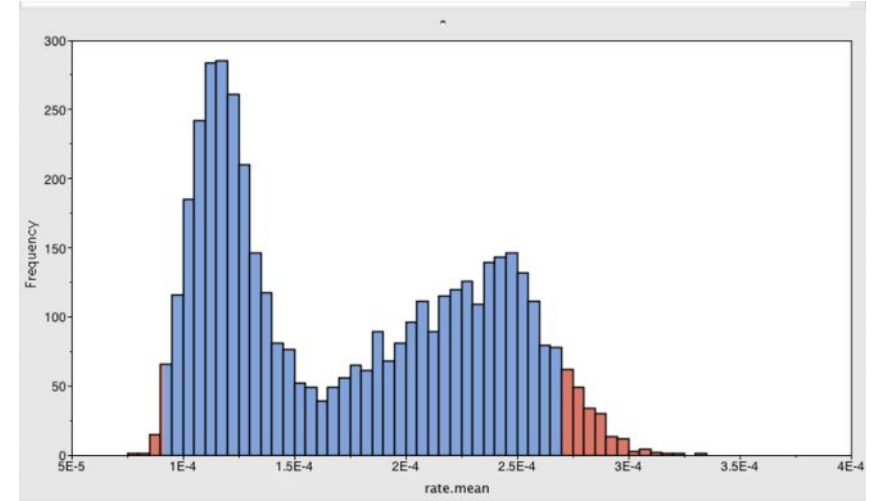
For future researchers

- Keep all input files (*Sequences were aligned manually...*)
- Keep the final result files (PDF is **not** a tree format)
- Keep the pre and post-processing code

Bayesian inference: pros and cons

- Pros

- Complete posterior distribution
=> good with uncertain and complex scenarios
- Use of priors => uses results from previous studies and biological knowledge



- Cons

- (Very) computationally expensive
- Use of priors => more complex analysis setup
- Convergence can be a major issue