Introduction to Bayesian phylogenetic inference

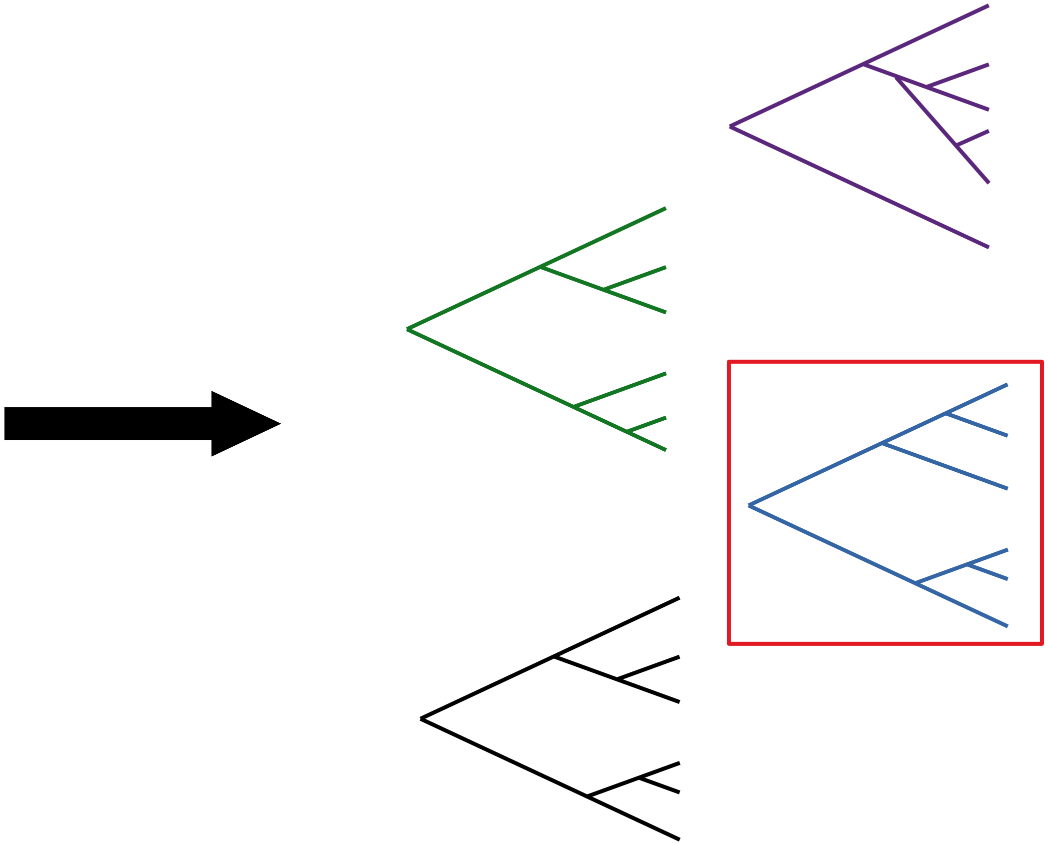
Joëlle Barido-Sottani

What is inference ?

```
ACAGACTTTCAGACTTTCAGACCC
ACACACCTACAGACTTACAGACCC
TCAGACTTTCACACCTTCAGACCT
TCACACCTACACACCCACAGACTT
TCACACCTACACACCCACAGACTT
TCAGACTTTCACACCTTCAGACCT
```

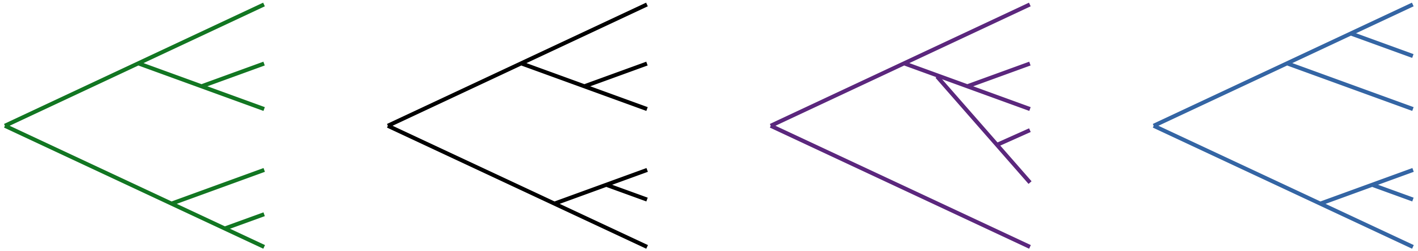
Observations

Explanations

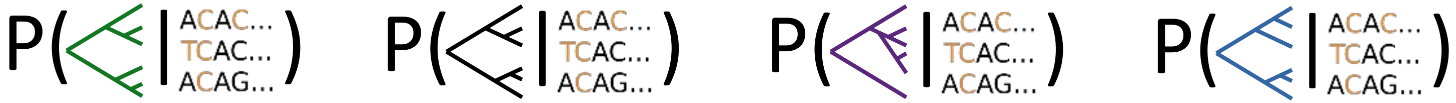


Requirements for inference

Choice of model



Ranking function



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
  => P(dice = 5) = n/N = 0.21

Issues

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

What is probability ?

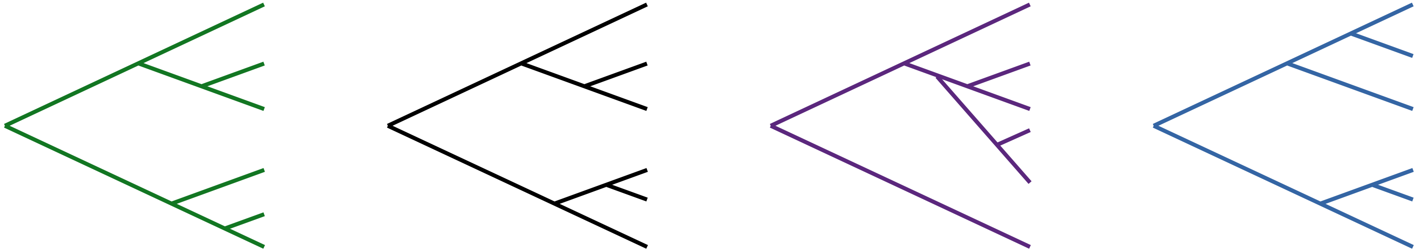
Bayesian approach

- Probability measures how plausible an outcome is based on available information
- P(dice = 5 | no information) = 1/6
  P(dice = 5 | dice is unfair) = 0.01
  P(dice = 5 | perfect information) = 1

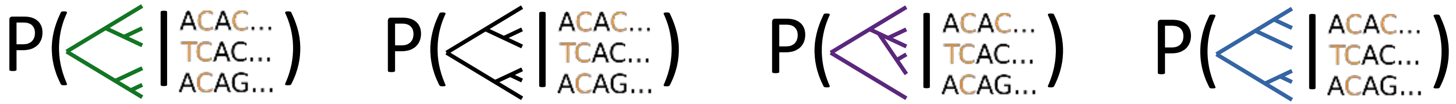
=> Probability expresses the level of certainty

Requirements for inference

Choice of model

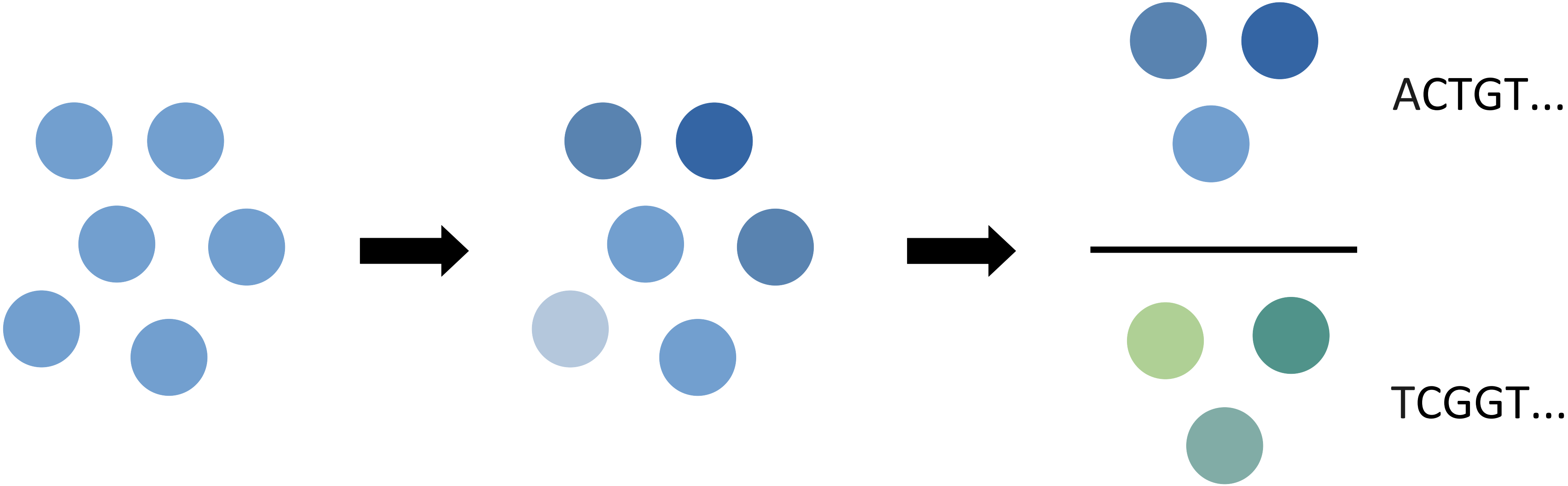


Ranking function



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

Generative models of evolution



The data is the outcome of the model

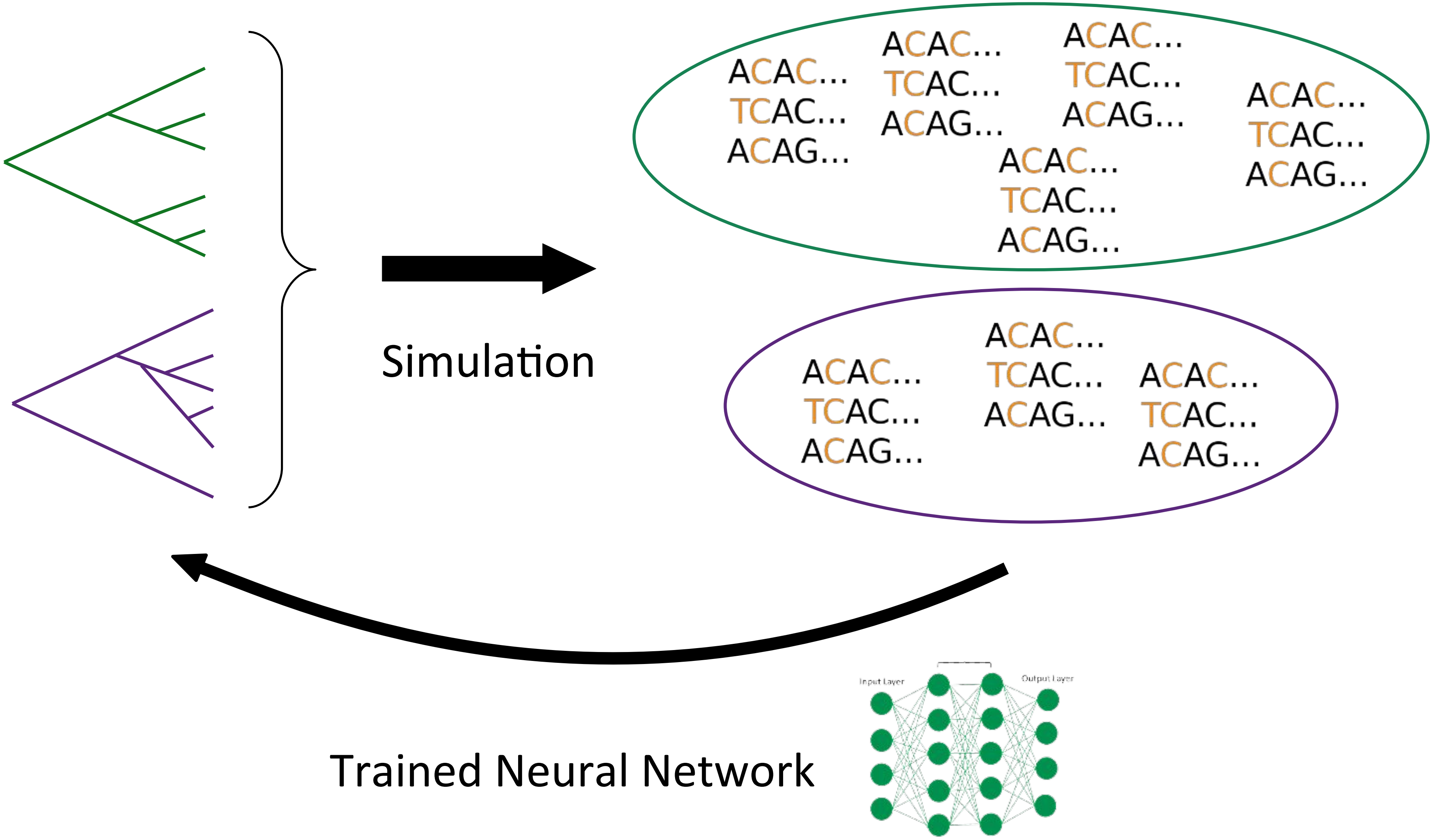
=> we can calculate P(data|parameters)

Inference based on generative models

- What we want: P(**parameters**|**data**) probability of model parameters given our observed data
- What we have: P(**data**|**parameters**) likelihood i.e. probability of generating the data given the model parameters

- Maximum likelihood approach
  => Use the likelihood P(**data**|**parameters**) as ranking function

Deep learning approach



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

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

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

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

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

Simulation

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

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

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

Trained Neural Network

Bayes’ theorem for inference

$$\underbrace{P(\text{param}|\text{data})}_{\text{Posterior}} = \frac{\overbrace{P(\text{data}|\text{param})}^{\text{Likelihood}}\ \underbrace{P(\text{param})}_{\text{Prior}}}{\underbrace{P(\text{data})}_{\text{Marginal likelihood of the data}}}$$

Bayes' theorem for inference

The data and model parameters are described by probabilities

- **Prior : P(param)** => the range of *plausible* parameter values

  NB : All model parameters have priors

- **Likelihood : P(data|param)** => the likelihood is proportional to the probability of observing the data given a hypothesis

- **Posterior : P(param|data)** => combines information from the data (likelihood) and previous knowledge (prior)

- **Marginal likelihood : P(data)** => 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

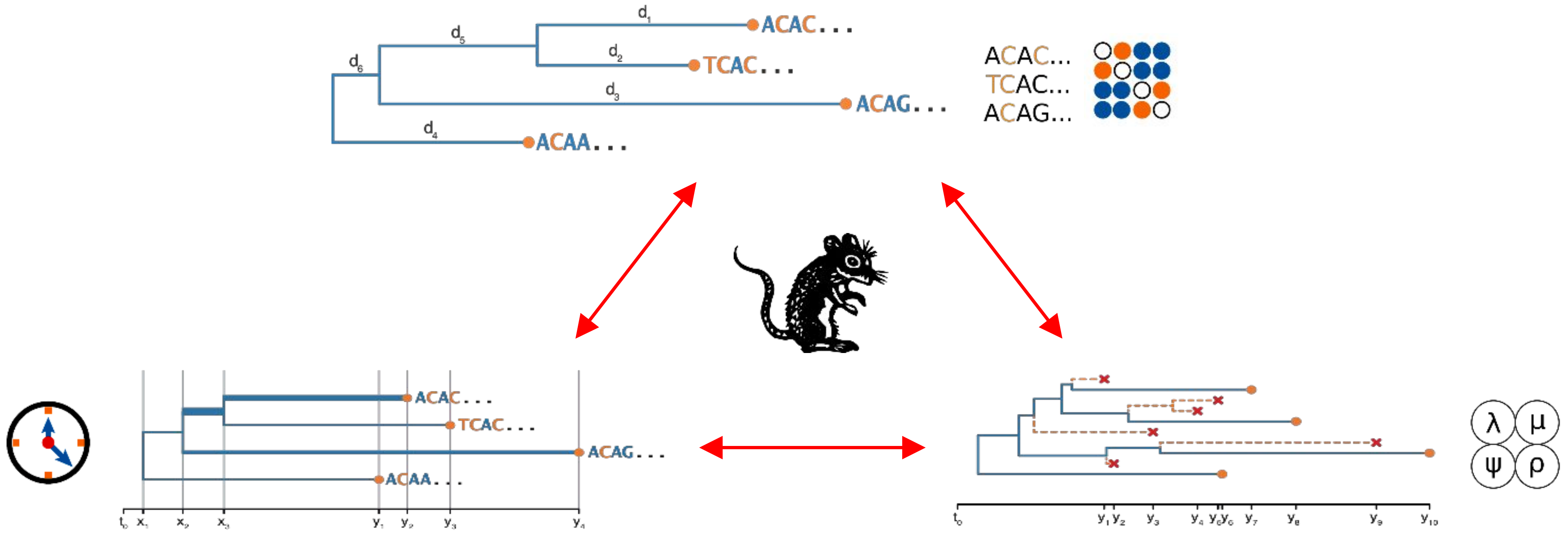
Bayesian phylogenetic and phylodynamic tools

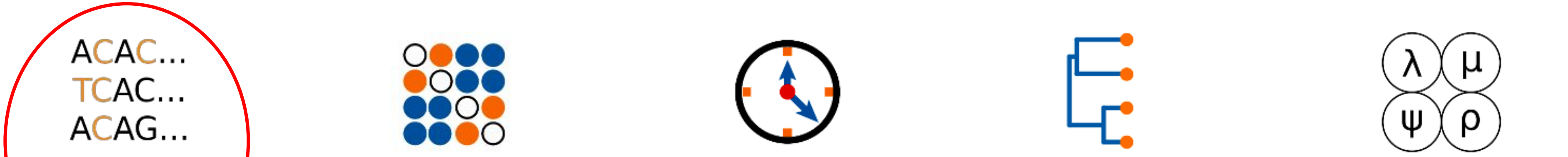
- BEAST & BEAST2
- MrBayes & RevBayes
- PhyloBayes (focus on protein alignments)
- Bali-Phy (estimating the alignment)
- SCAR (focus on recombination)
- Many more…..

Beast2

Bayesian evolutionary analysis by sampling trees

What goes into a **BEAST2** model?





Alignment data

Substitution model

Clock model

Time tree

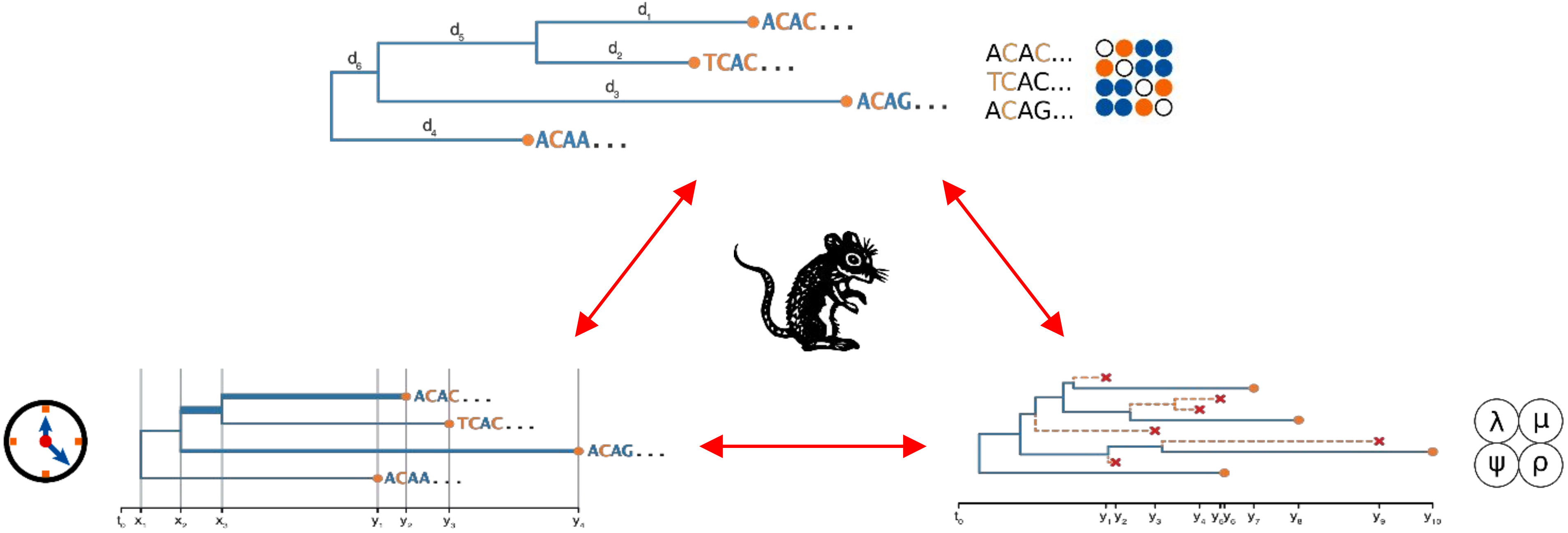
Demographic model

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?



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

Alignment data

Substitution model

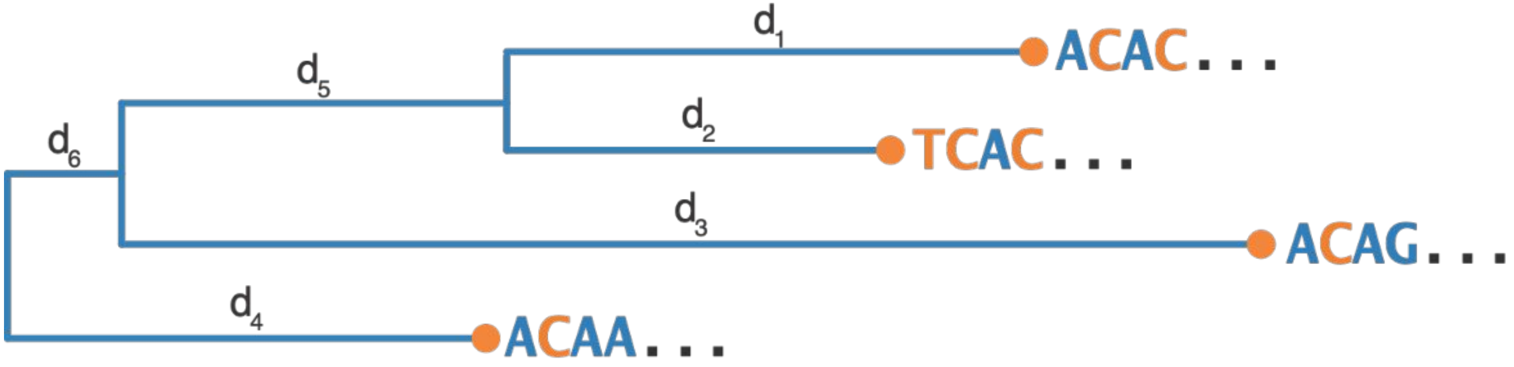
Clock model

Time tree

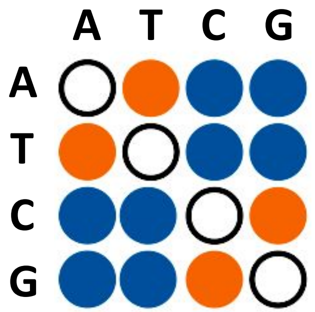
λ μ
ψ ρ

Demographic model

Substitution/site model



Genetic distance from common ancestor

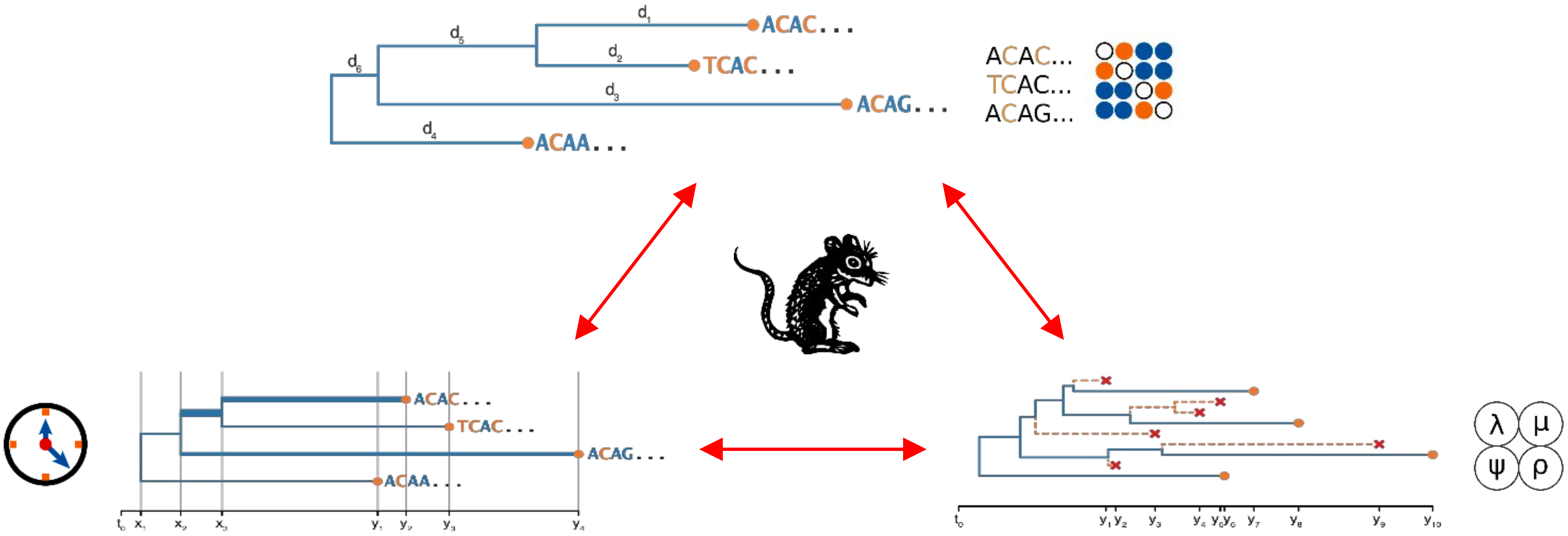


+

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

- 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?



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

Alignment data

Substitution model

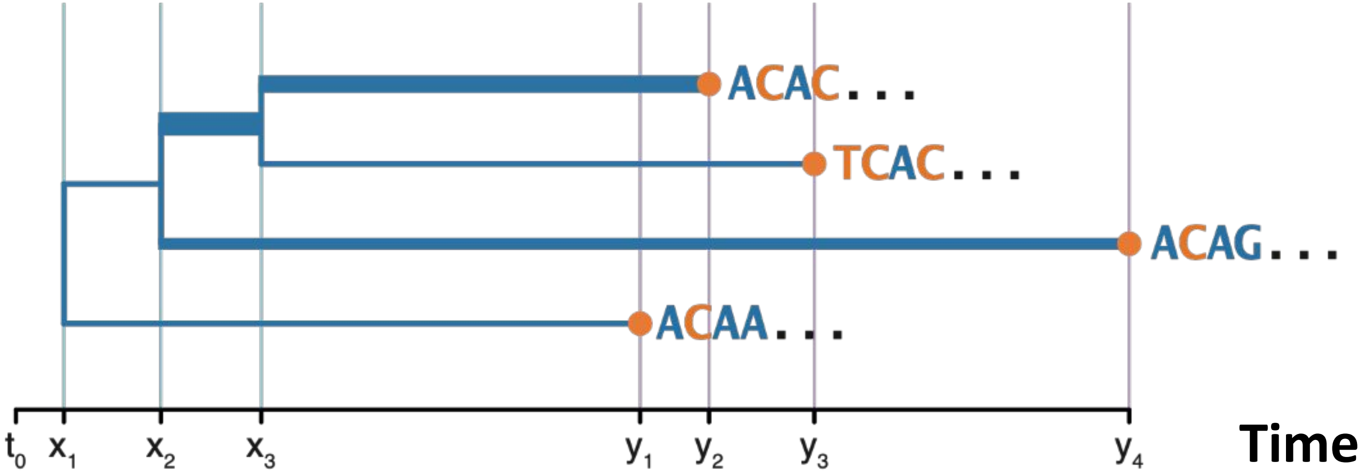
Clock model

Time tree

λ μ
ψ ρ

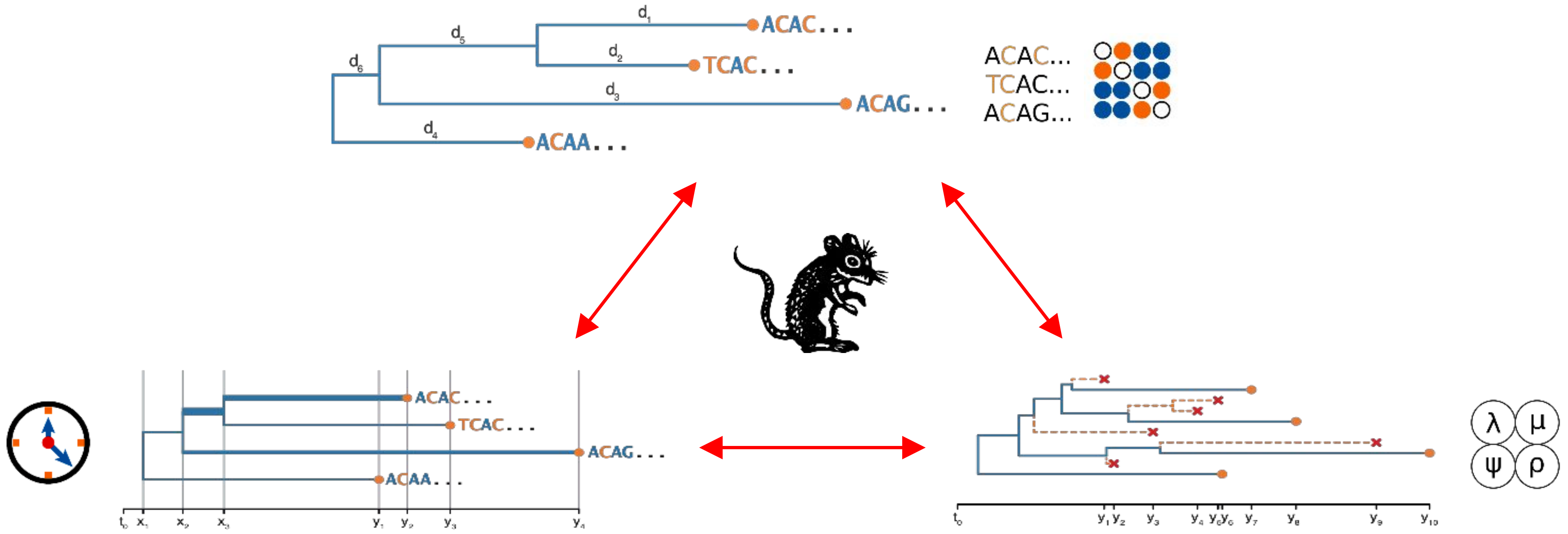
Demographic model

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?



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

Alignment data

Substitution model

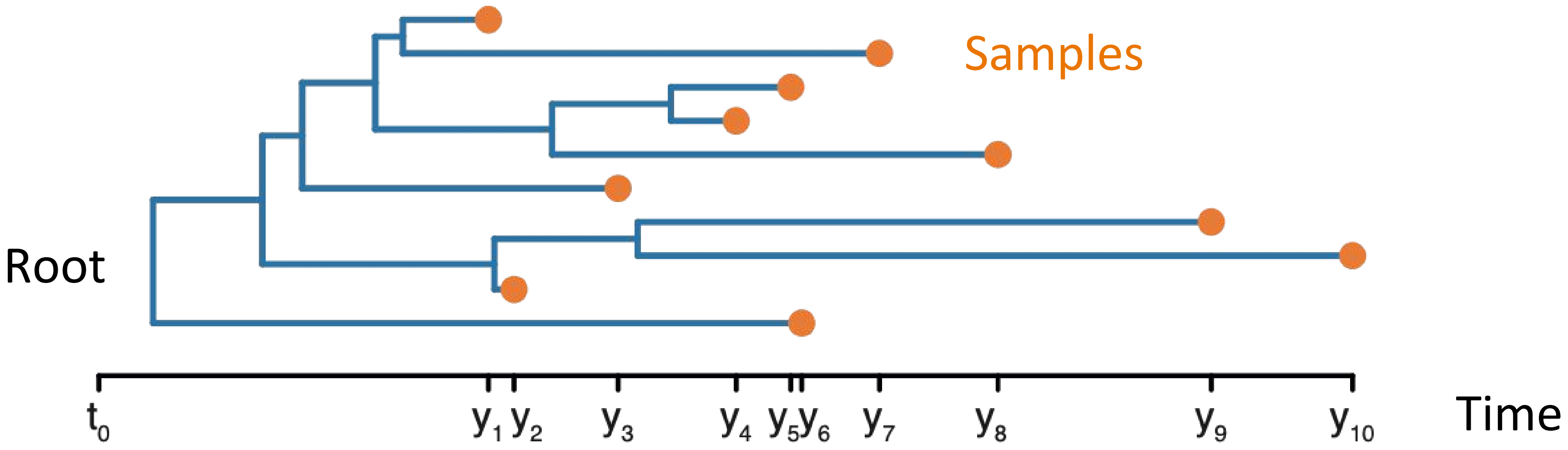
Clock model

Time tree

λ μ
ψ ρ

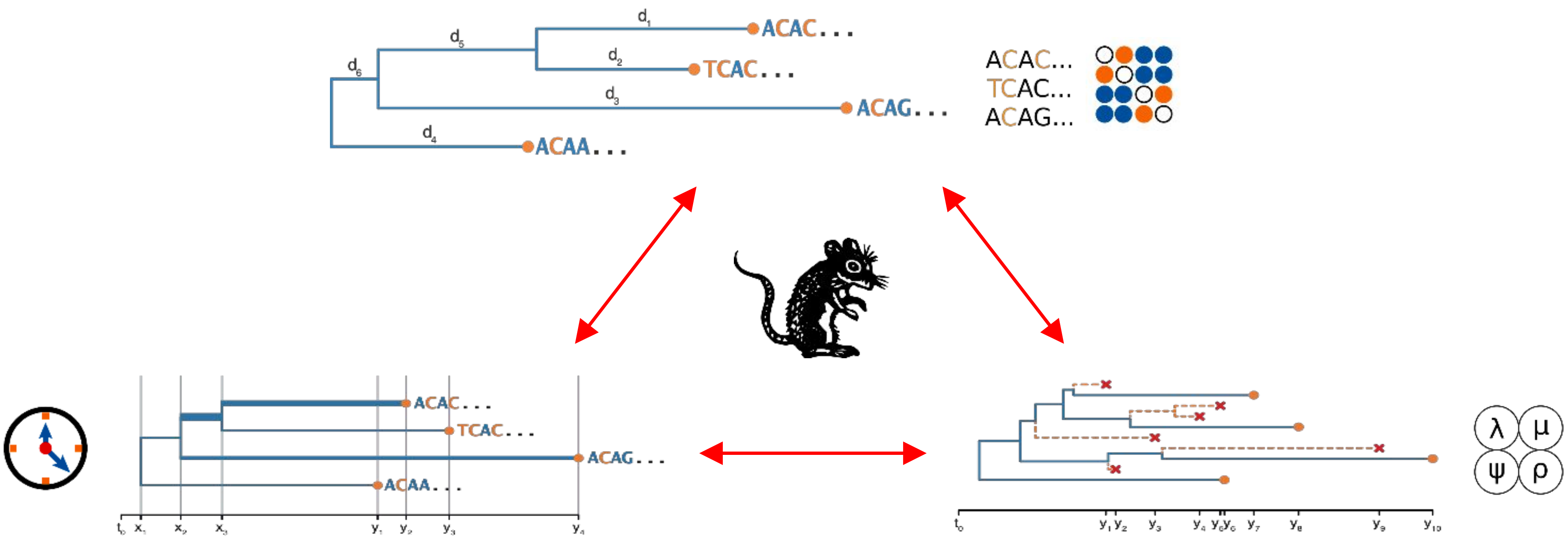
Demographic model

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?



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

Alignment data

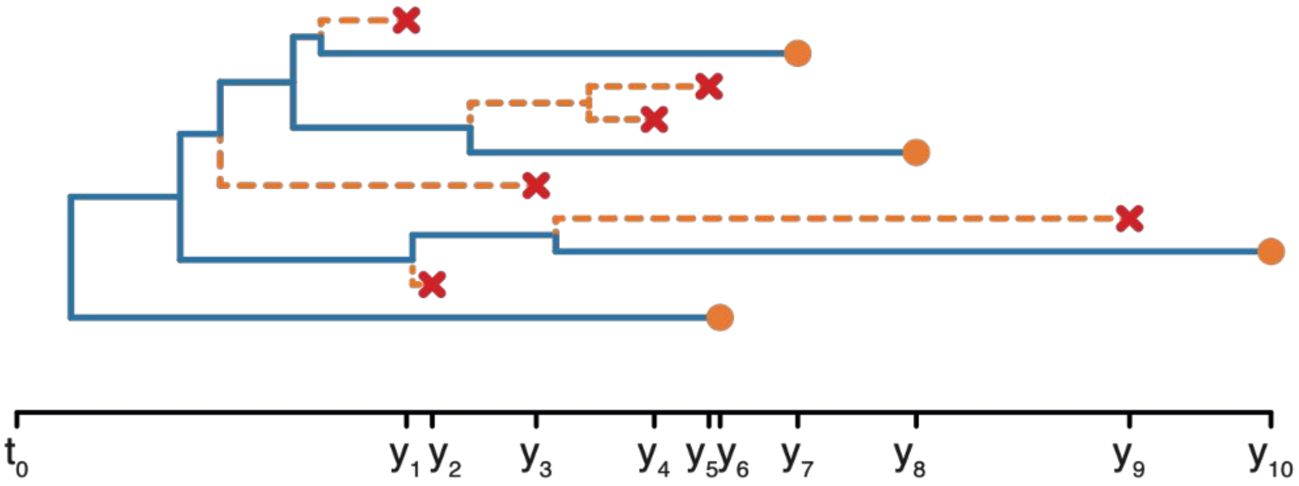
Substitution model

Clock model

Time tree

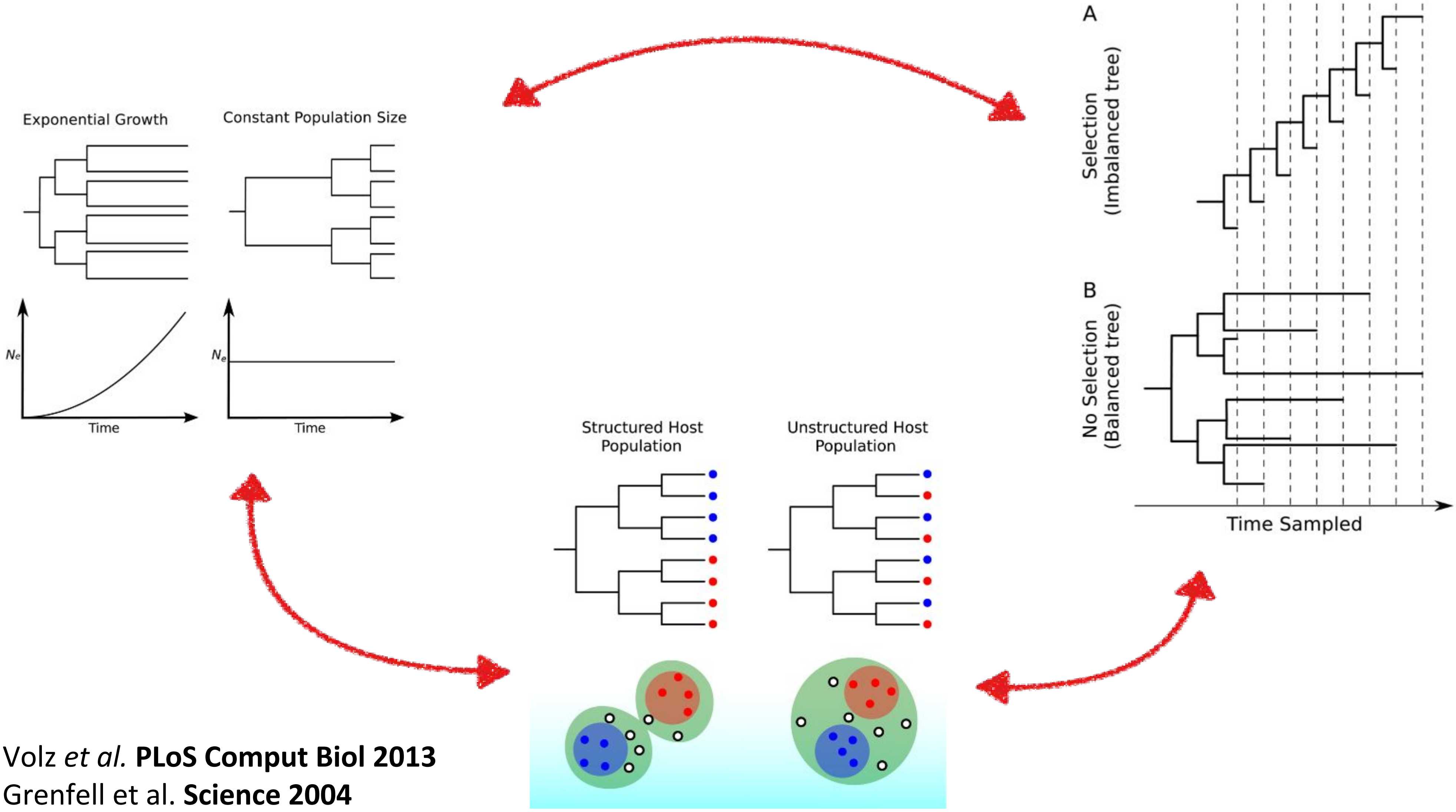
Demographic model

Demographic (tree) model



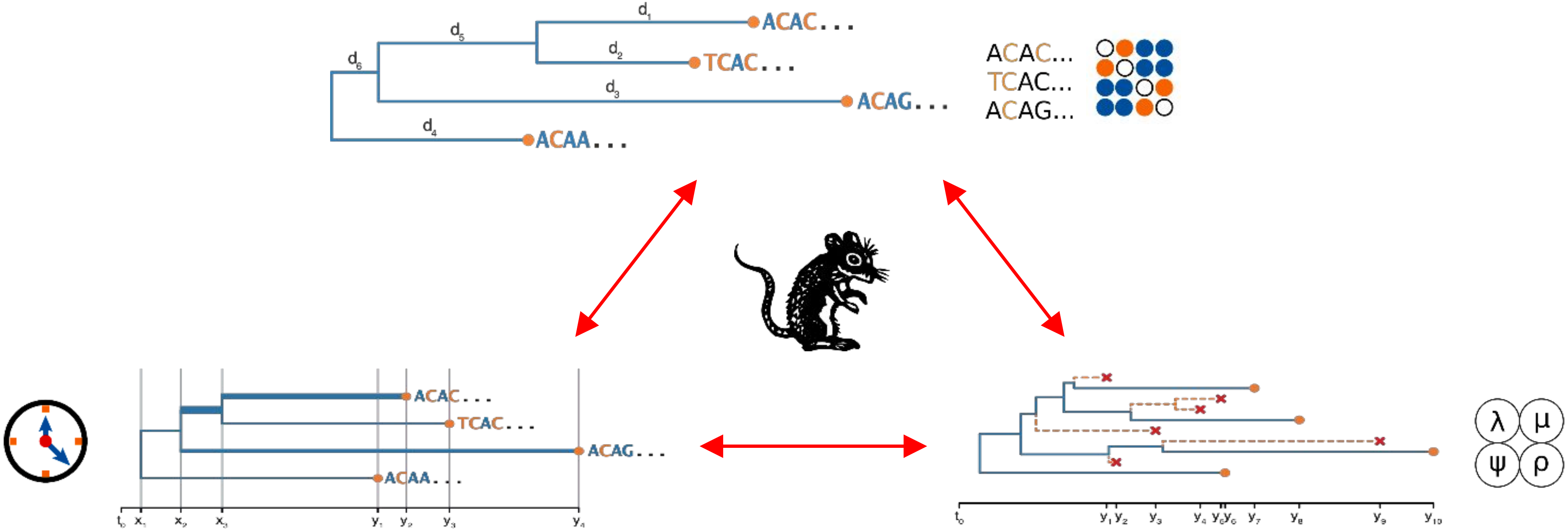
- Serves as tree prior (required since the tree is a parameter)
- Describes the population dynamics
  - How does the infected population grow over time?
  - How does the transmission rate change over time?
- Usually a birth-death or a coalescent model

Different population dynamics generate different trees



Volz *et al.* **PLoS Comput Biol 2013**
Grenfell et al. **Science 2004**

What goes into a **BEAST2** model?



ACAC...
TCAC...
ACAG...
λ
μ
ψ
ρ

Alignment data

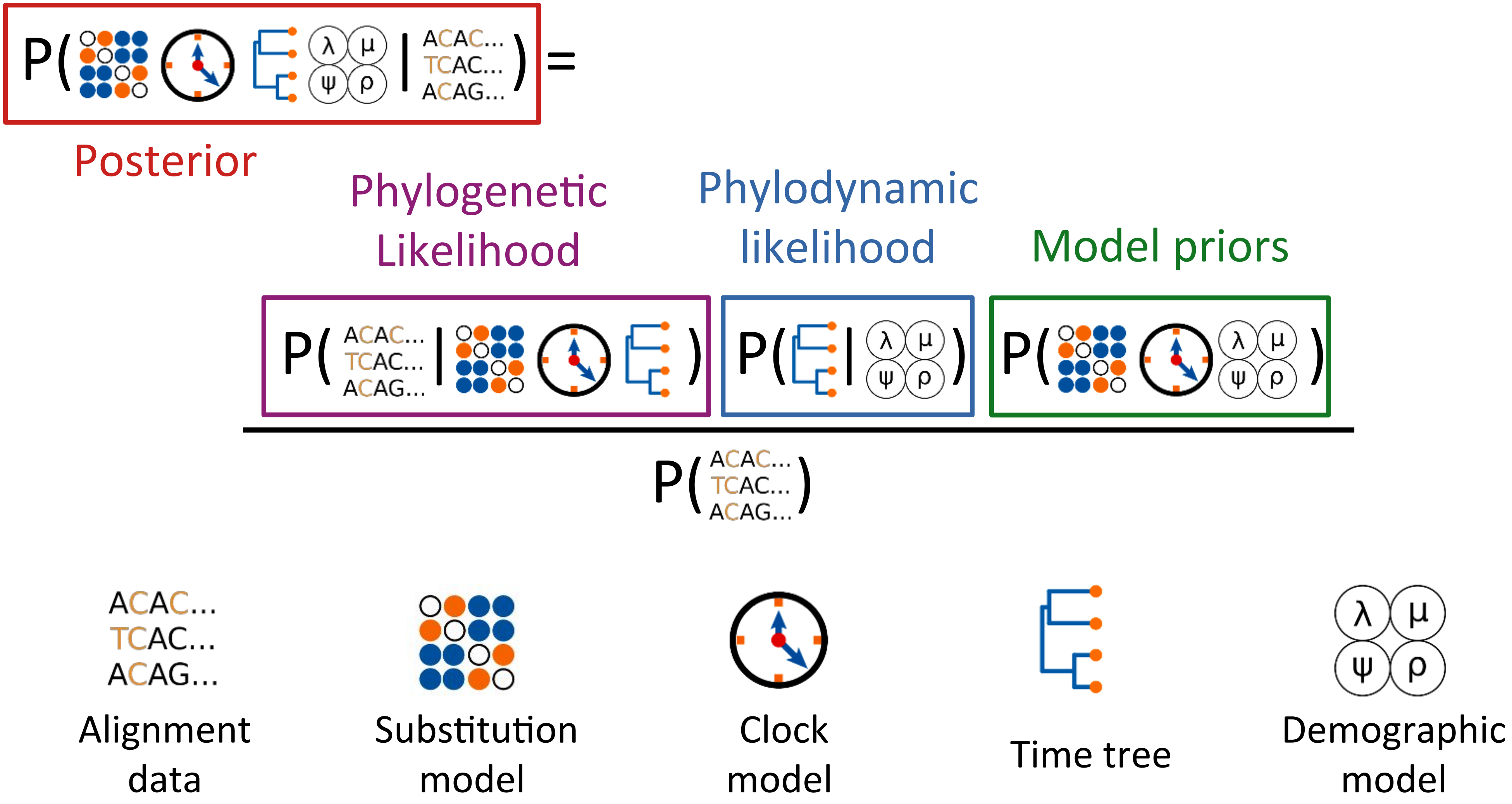
Substitution model

Clock model

Time tree

Demographic model

Final posterior distribution



Posterior

Phylogenetic Likelihood

Phylodynamic likelihood

Model priors

Alignment data

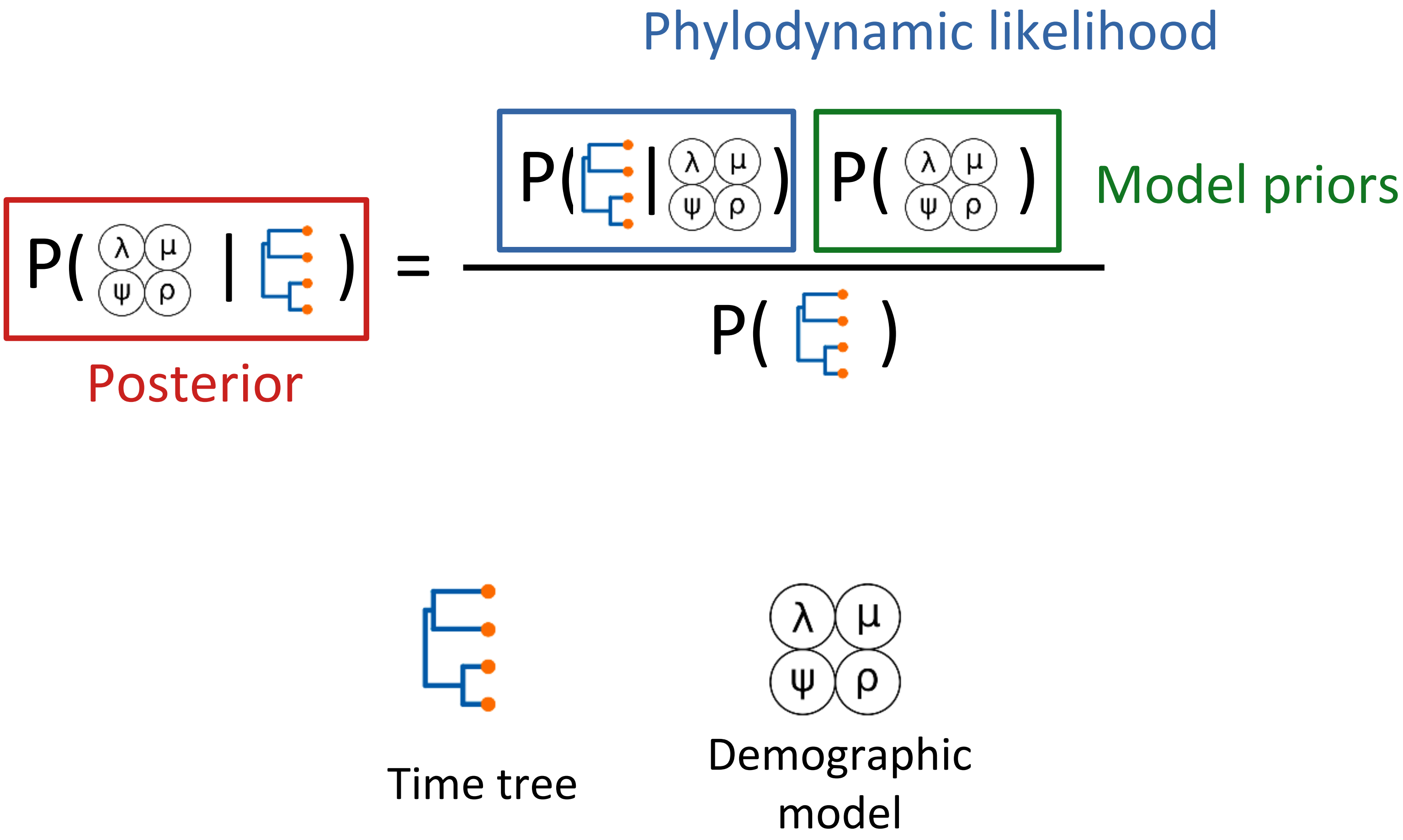
Substitution model

Clock model

Time tree

Demographic model

Final posterior distribution – fixed tree



Phylodynamic likelihood

Model priors

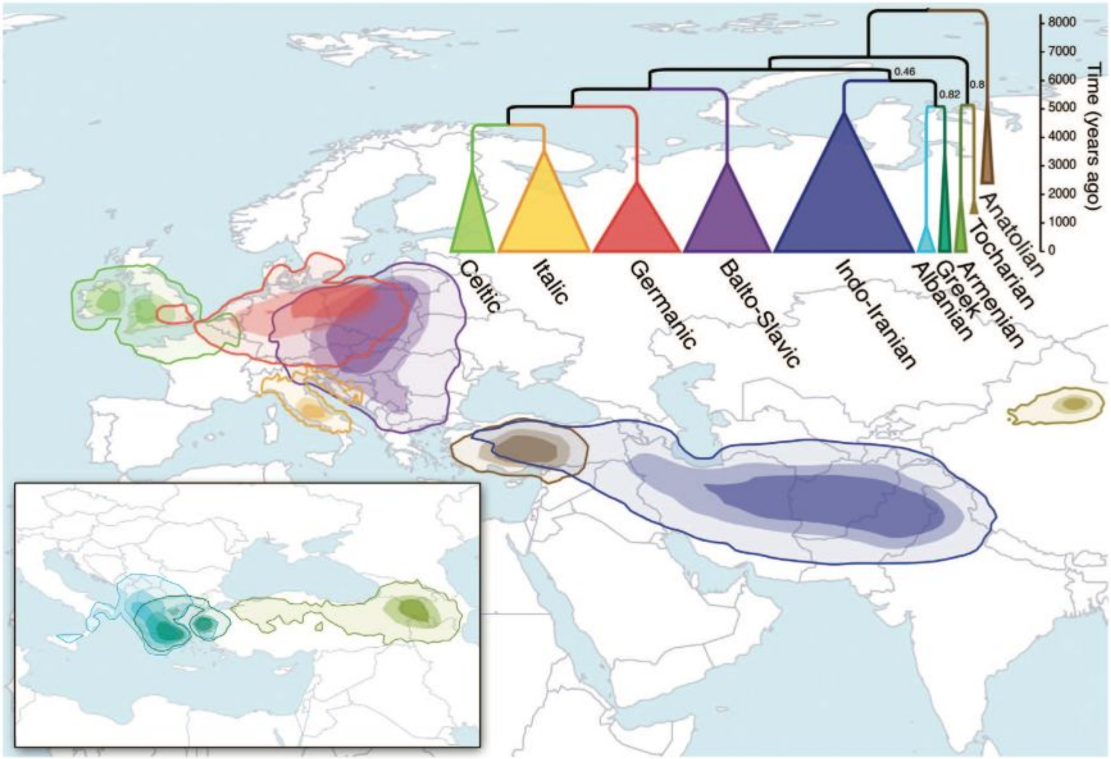
Posterior

Time tree

Demographic model

Some special cases I

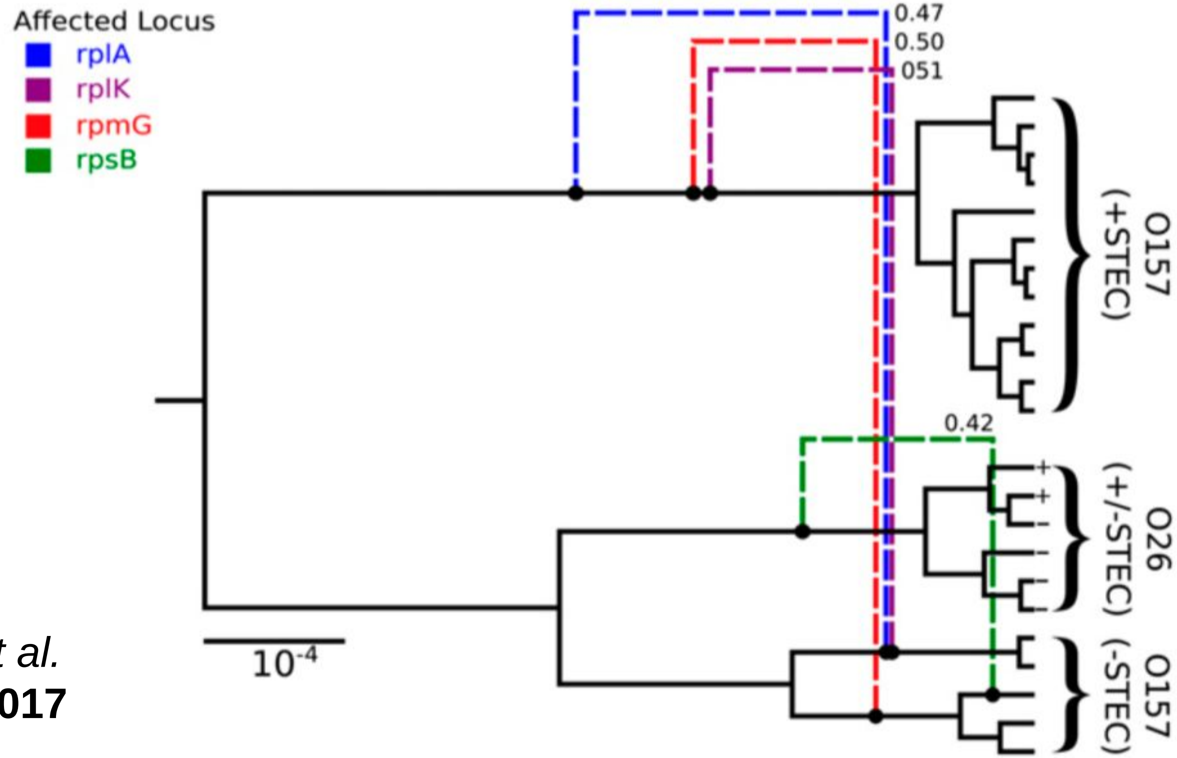
Site models don't have to be on nucleotides
=> Could be on amino acids, morphological traits, roots of words etc.



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

$$P(\text{param}\,|\,\text{data}) = \frac{P(\text{data}\,|\,\text{param})\;P(\text{param})}{P(\text{data})}$$

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

All possible param values

But the tree **is** a parameter

How many trees are there ?

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

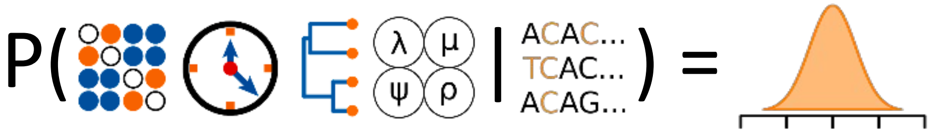
Number of tips	4	5	6	7	8	9	10	20	48
Number of trees	15	105	945	10395	135135	2.0×10^{6}	3.5×10^{7}	8.2×10^{21}	3.2×10^{70}

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



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

P(ACAC...
TCAC...
ACAG...) = ?

=> 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)

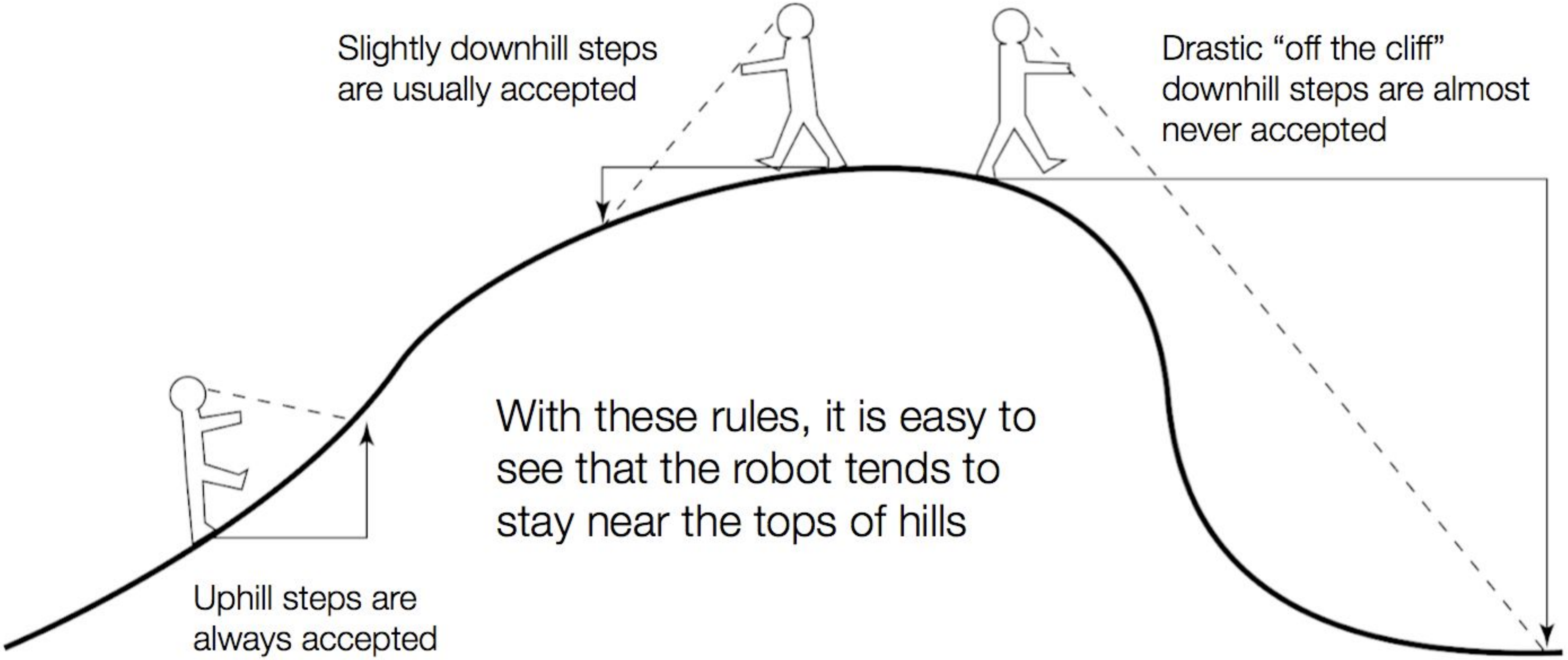
- MCMC moves through the parameter space and looks for places with high posterior

- For each step we only need to compare which posterior density is higher

  => so we only need the ratio of posteriors

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

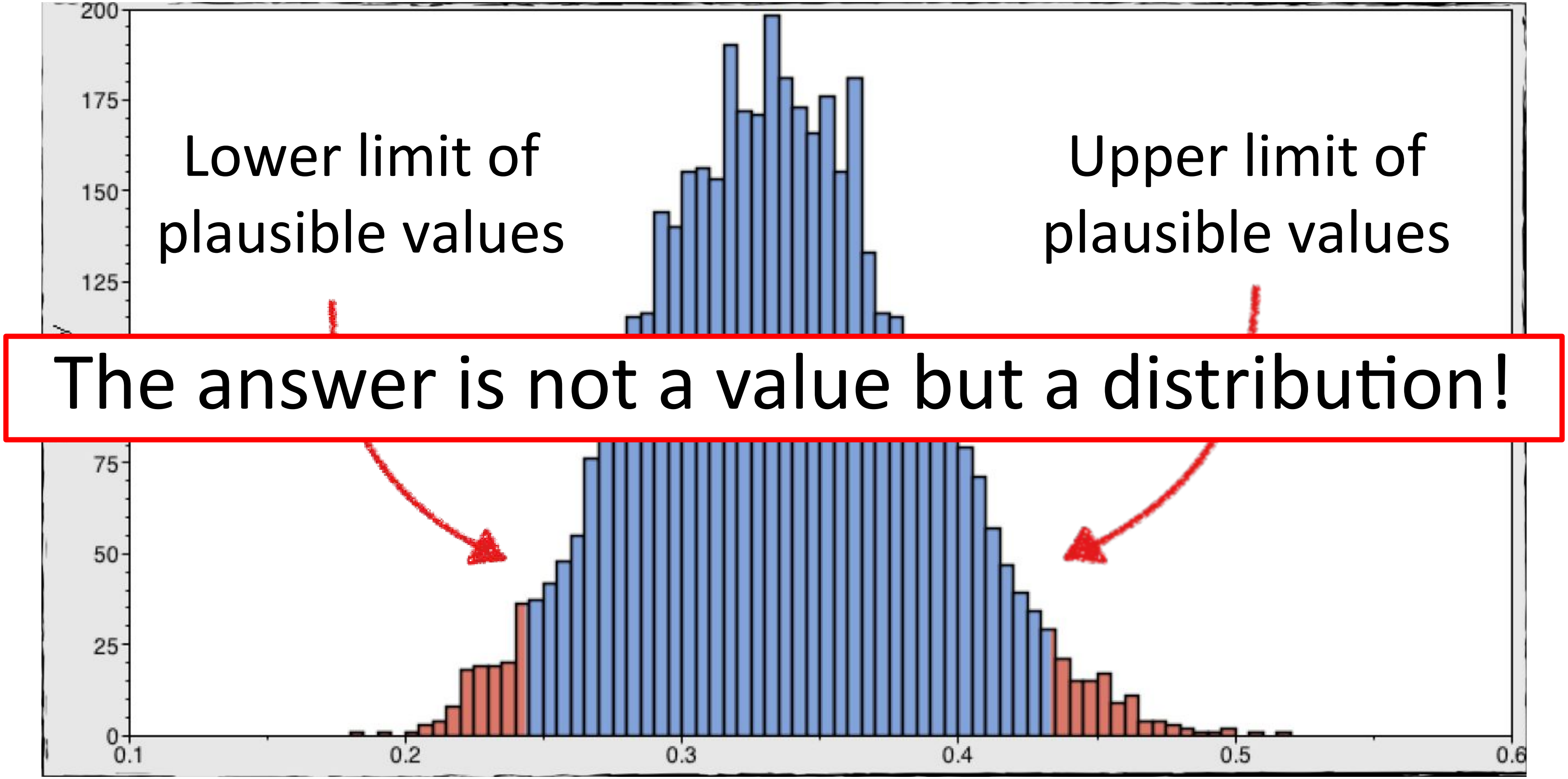
MCMC robot (courtesy of Paul Lewis)


Slightly downhill steps are usually accepted
Drastic “off the cliff” downhill steps are almost never accepted
With these rules, it is easy to see that the robot tends to stay near the tops of hills
Uphill steps are always accepted

MCMC through parameter space

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

Final posterior estimate

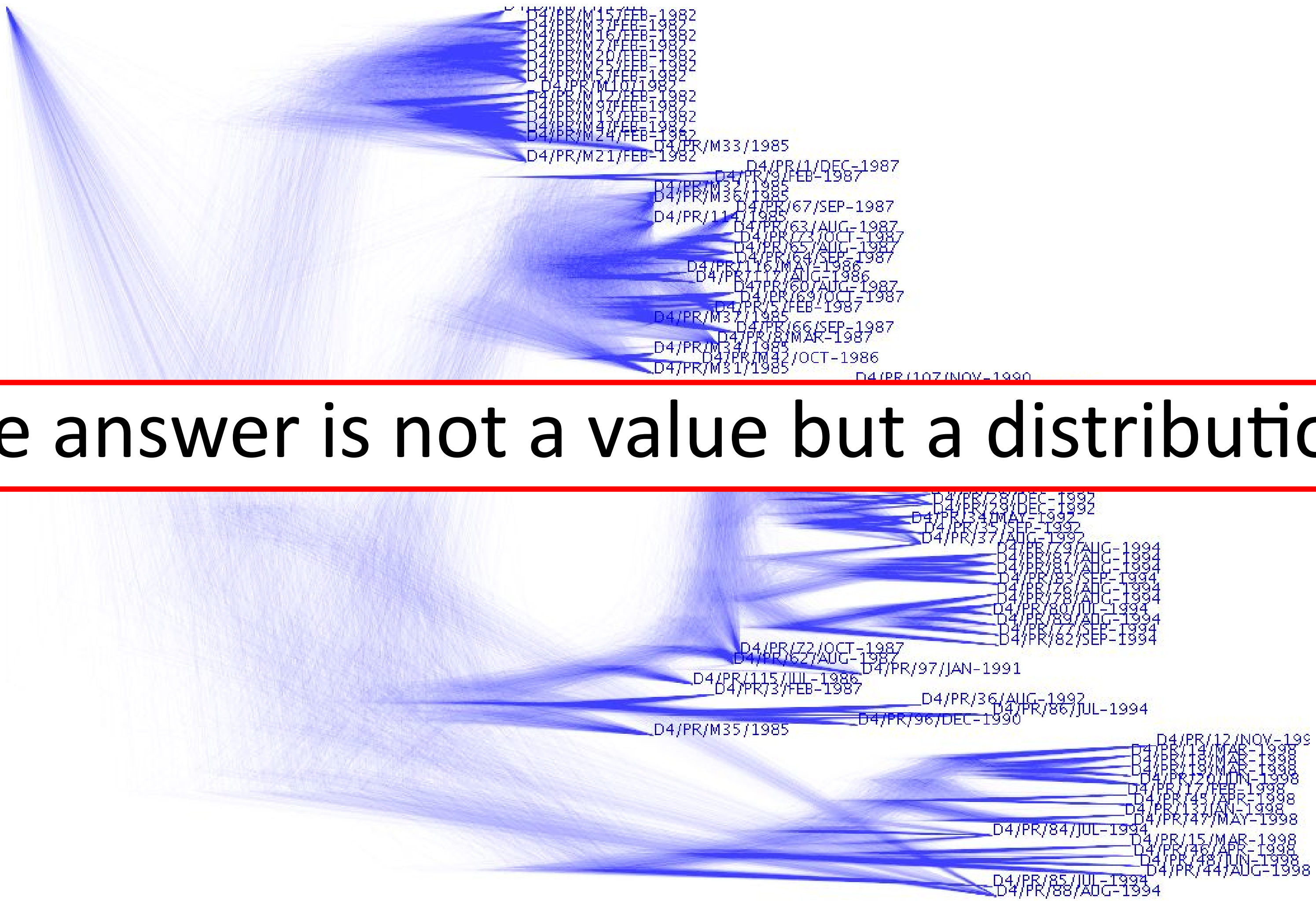


Lower limit of plausible values

Upper limit of plausible values

The answer is not a value but a distribution!

Final posterior estimate (tree edition)



The answer is not a value but a distribution!

Progress of an inference

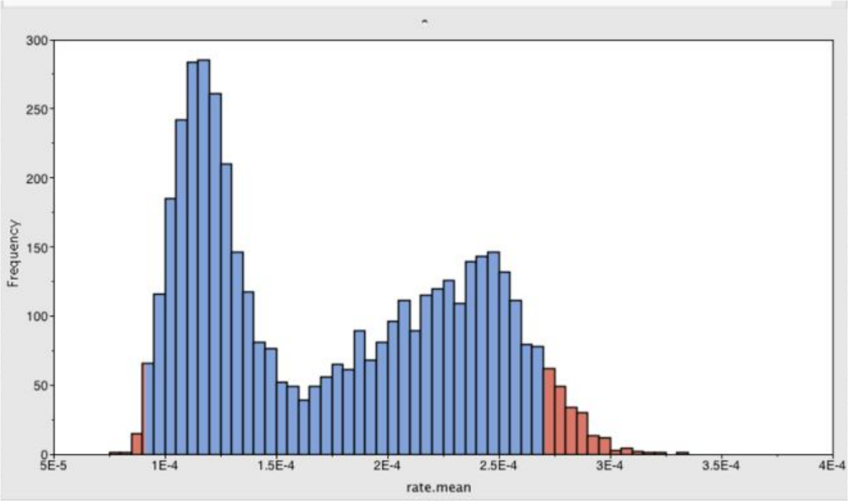
- Initial position – set by the user or by BEAST2
- Burn-in phase: moving from the initial position to the high-posterior space
- Convergence phase: the inference has reached the high-posterior space – still moving but stable
- The posterior estimates are given **only** by samples taken **after convergence**

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

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