Phylolinguistics part 1

Benedict King, MPI EVA

PREPARE FOR TUTORIAL

https://github.com/king-ben/winterschool

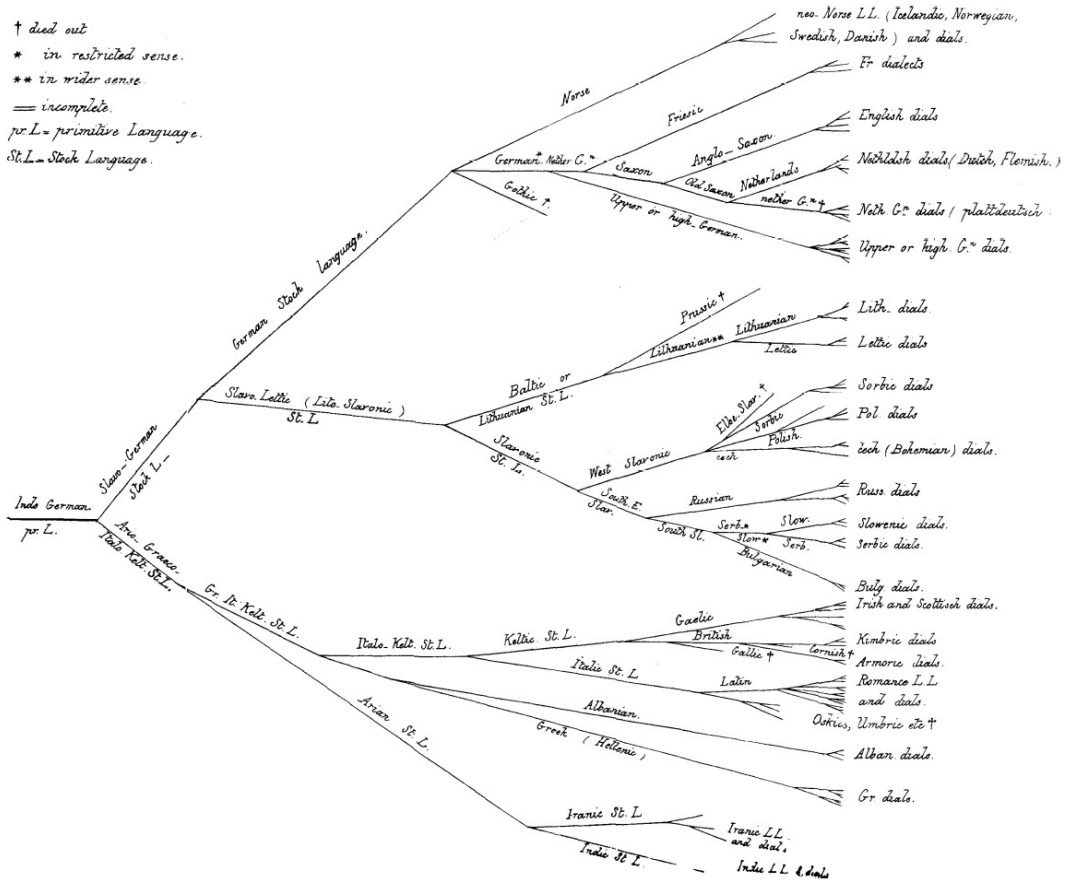
Click on required_software.md

Follow instructions

Learning Goals

- Understand the motivation and philosophy behind Phylogenetics
- What kind of data is required for input into a phylogenetic analysis of languages
- Understand the major components of Bayesian phylogenetic methods
- How to interpret the output of a phylogenetic analysis

Why phylogenetics for linguistics?



Schleicher 1863

- Variation in traits
- Inheritance of traits
- Lineage splitting

What is phylogenetics?

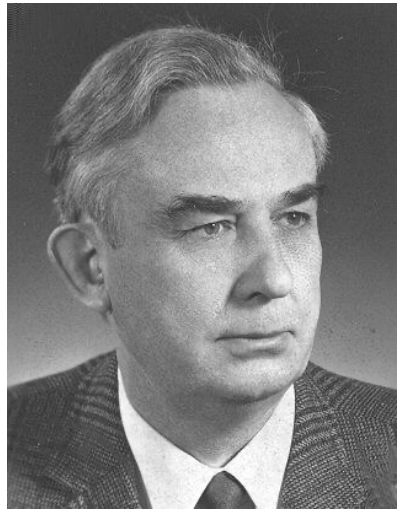
It does NOT involve overall similarity measures

Biology	Linguistics
Phenetics	Lexicostatistics

What is phylogenetics?

By making assumptions about the underlying PROCESS, we can reconstruct more informative relationship hypotheses

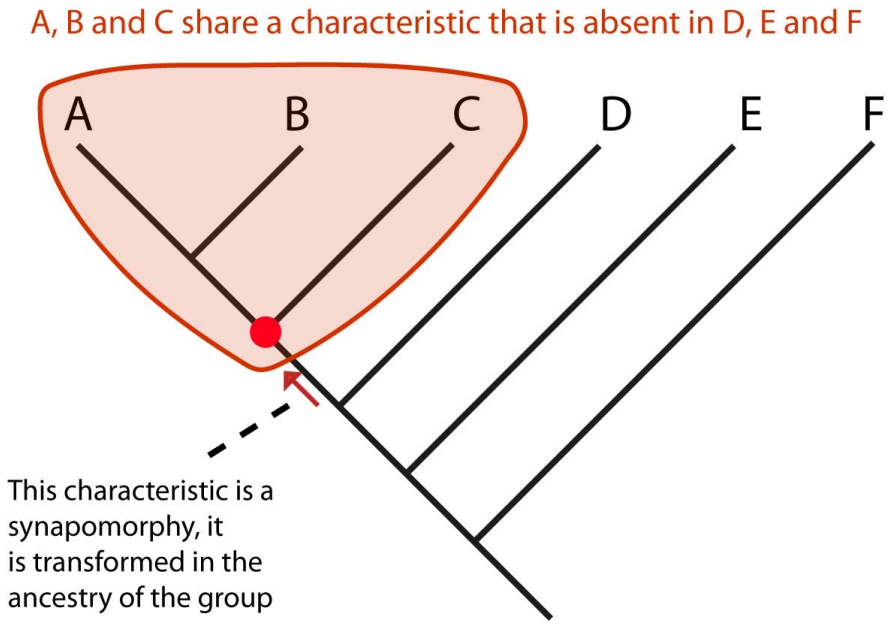
- Lineages split
- Innovations appear and are transmitted to descendants



Willi Hennig

What is phylogenetics?

It is forming groups based only on INNOVATIONS



Biology	Linguistics
Cladistics, Phylogenetic systematics	Historical Linguistics, Linguistics Phylogenetics
Synapomorphy	Innovation
Symplesiomorphy	Retention

Data for linguistic phylogeny

1. Swadesh List of basic vocabulary **concepts/meanings**

2. Record **lexeme** for each language

3. Classify lexemes into **cognate sets**

woman	Italic	Latin	mulier		ˈmuli.er	ˈmuli.er	mulier [Latin]
woman	Italic	Romanian	femeie		feˈmeje	feˈmeje	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Megleno-Romanian	mul'ári				mulier [Latin]
woman	Italic	Dalmatian: Vegliote	mul´ér		muˈʎer		mulier [Latin]
woman	Italic	Neapolitan	femmena		ˈfemmənə	ˈfemmənə	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Italian	donna		ˈdɔnna	ˈdɔnna	*dem- [Proto-Indo-European]
woman	Italic	Friulian	femine		ˈfɛmine	ˈfɛminɛ	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Ladin	ela		ˈela		*ela [Ladin]*
woman	Italic	Milanese	dòna		ˈdɔna		*dem- [Proto-Indo-European]
woman	Italic	Sardinian: Nuoro	femmina		ˈfɛmmina	ˈfɛmmina	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Sardinian: Logudoro	fémina		ˈfɛmina	ˈfɛmina	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Anglo-Norman	femme				*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	French	femme		fam	fam	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Walloon	feume		fœm	fœm	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Old Occitan	femna				*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Old Occitan	moulher, molher				mulier [Latin]
woman	Italic	Franco-Provençal	fna		fˈna	fˈna	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Old Catalan	fembra		fembrə	fembrə	*dʰeh₁(i)- [Proto-Indo-European]
woman	Italic	Catalan	dona		ˈdɔnə	ˈdɔnə	*dem- [Proto-Indo-European]
woman	Italic	Old Spanish	mujer		muˈʒer		mulier [Latin]
woman	Italic	Spanish	mujer		muˈxer	muˈxɛr	mulier [Latin]
woman	Italic	Portuguese	mulher		muˈʎɛr	muˈʎɛr	mulier [Latin]
woman	Italic	Portuguese: Brazilian	mulher				mulier [Latin]

Data is binarised for analysis

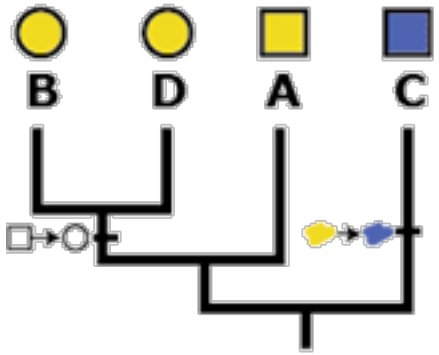
	mulier [Latin]	*dem- [Proto-Indo-European]	*dʰeh₁(i)- [Proto-Indo-European]	*ela [Ladin]*
Portuguese: Brazilian	1	0	0	0
Portuguese	1	0	0	0
Spanish	1	0	0	0
Old Spanish	1	0	0	0
Catalan	0	1	0	0
Old Catalan	0	0	1	0
Franco-Provençal	0	0	1	0
Old Occitan	1	0	0	0
Old Occitan	0	0	1	0
Walloon	0	0	1	0
French	0	0	1	0
Anglo-Norman	0	0	1	0
Sardinian: Logudoro	0	0	1	0
Sardinian: Nuoro	0	0	1	0
Milanese	0	1	0	0
Ladin	0	0	0	1
Friulian	0	0	1	0
Italian	0	1	0	0
Neapolitan	0	0	1	0
Dalmatian: Vegliote	1	0	0	0
Megleno-Romanian	1	0	0	0
Romanian	0	0	1	0
Latin	1	0	0	0

Parsimony

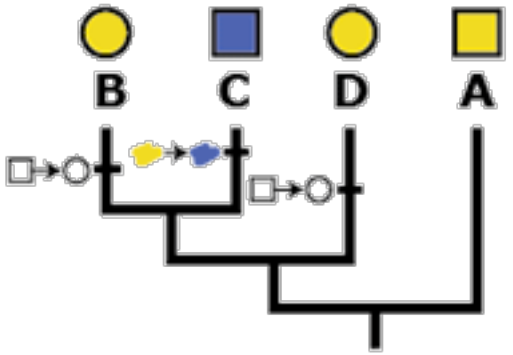
- For a particular tree, the total number of steps required across all characters is known as the **tree length**

taxon	characters	
	shape	color
A	□	yellow
B	○	yellow
C	□	blue
D	○	yellow

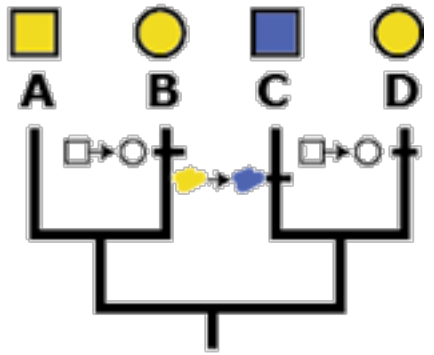
- The tree with the shortest length is known as the **most parsimonious tree**
- There may be more than one



tree length:
2 steps



tree length:
3 steps

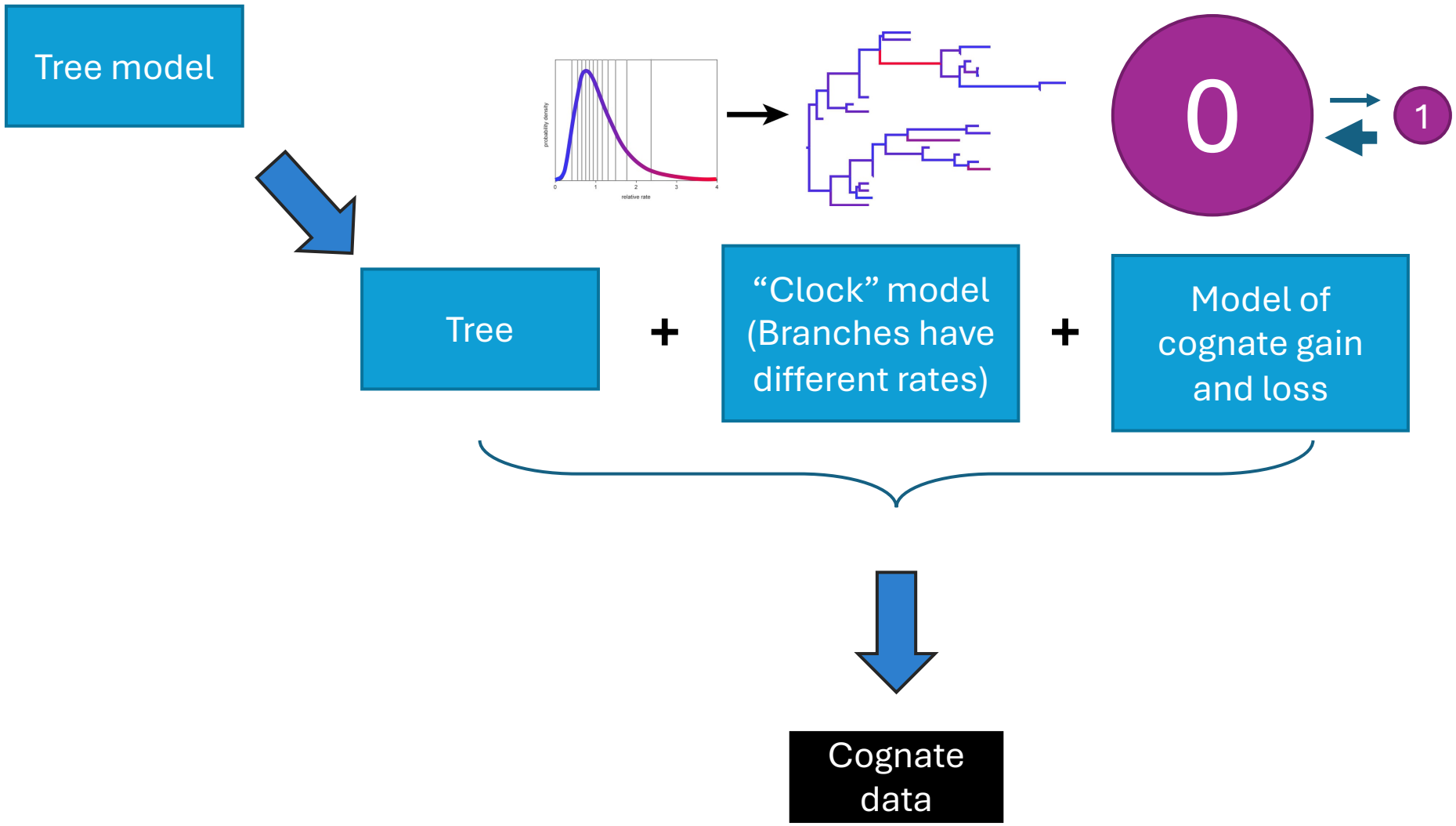


tree length:
3 steps

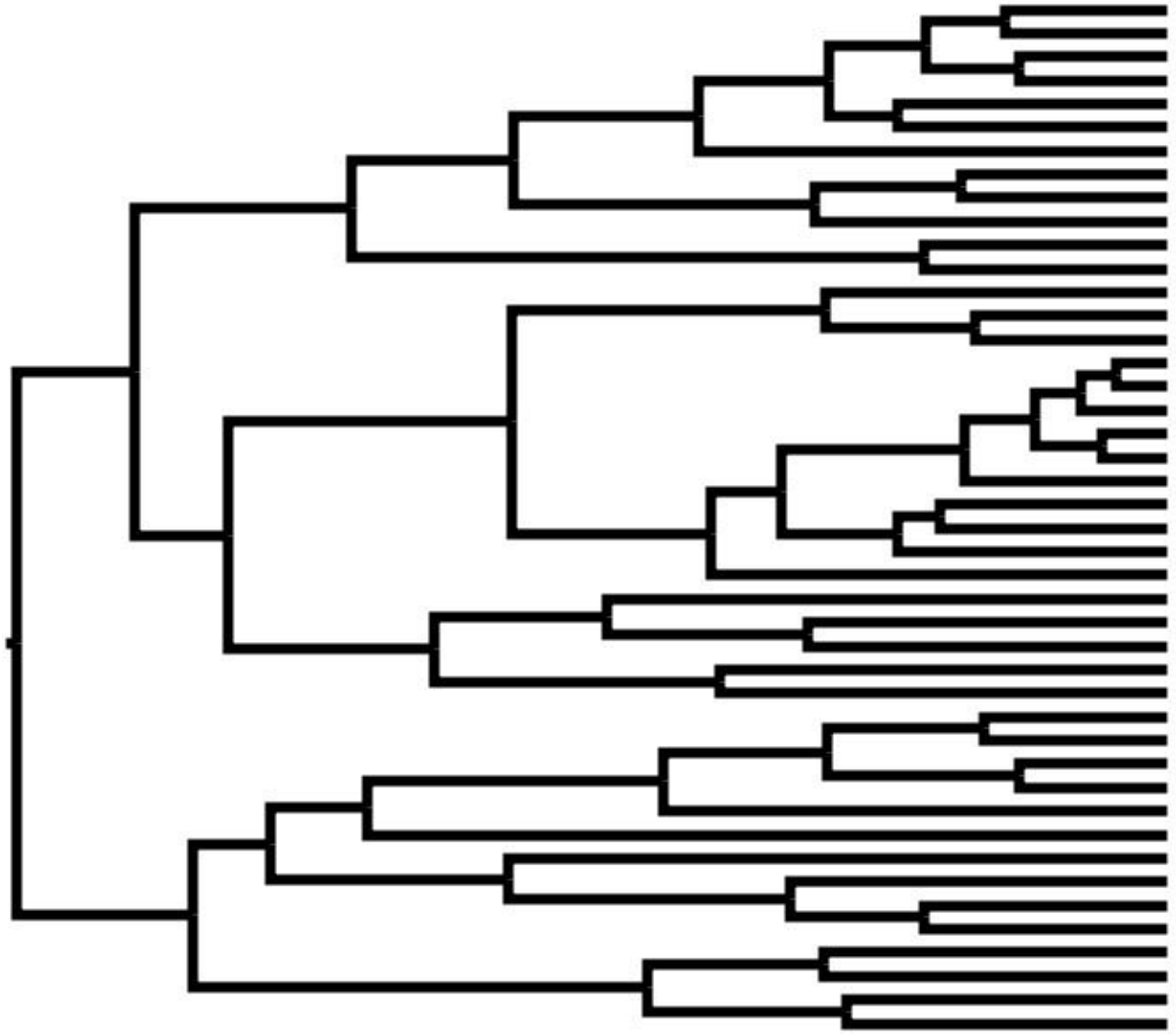
Model-based Phylogenetics – why?

- Rates of change across branches of the tree
- Allow varying rates between different meanings
- Estimate the age of divergences in the phylogeny
- Estimate meaningful branch lengths – either in units of time or amount of linguistic change
- Account for and quantify uncertainty

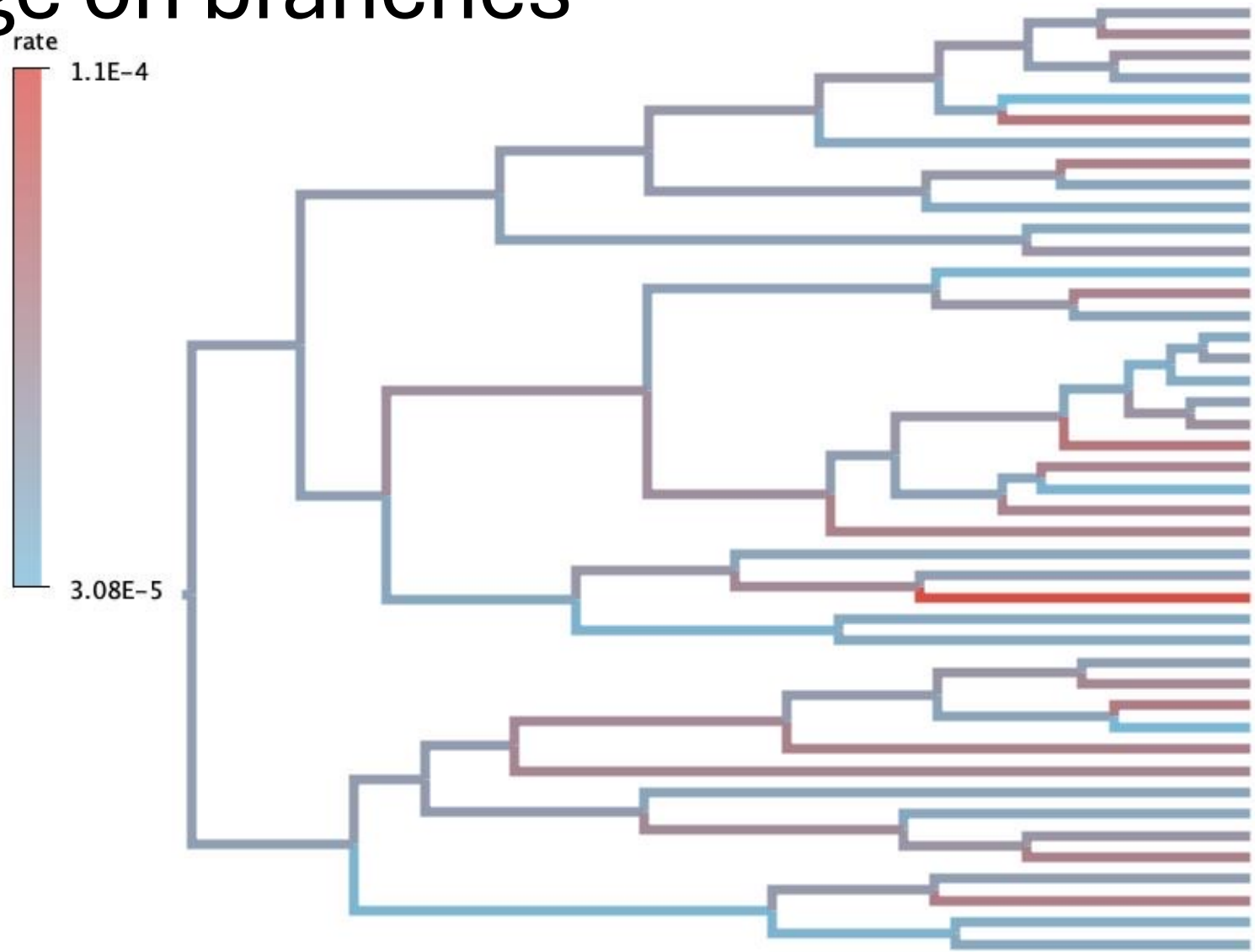
Overview of a generative phylogenetic model



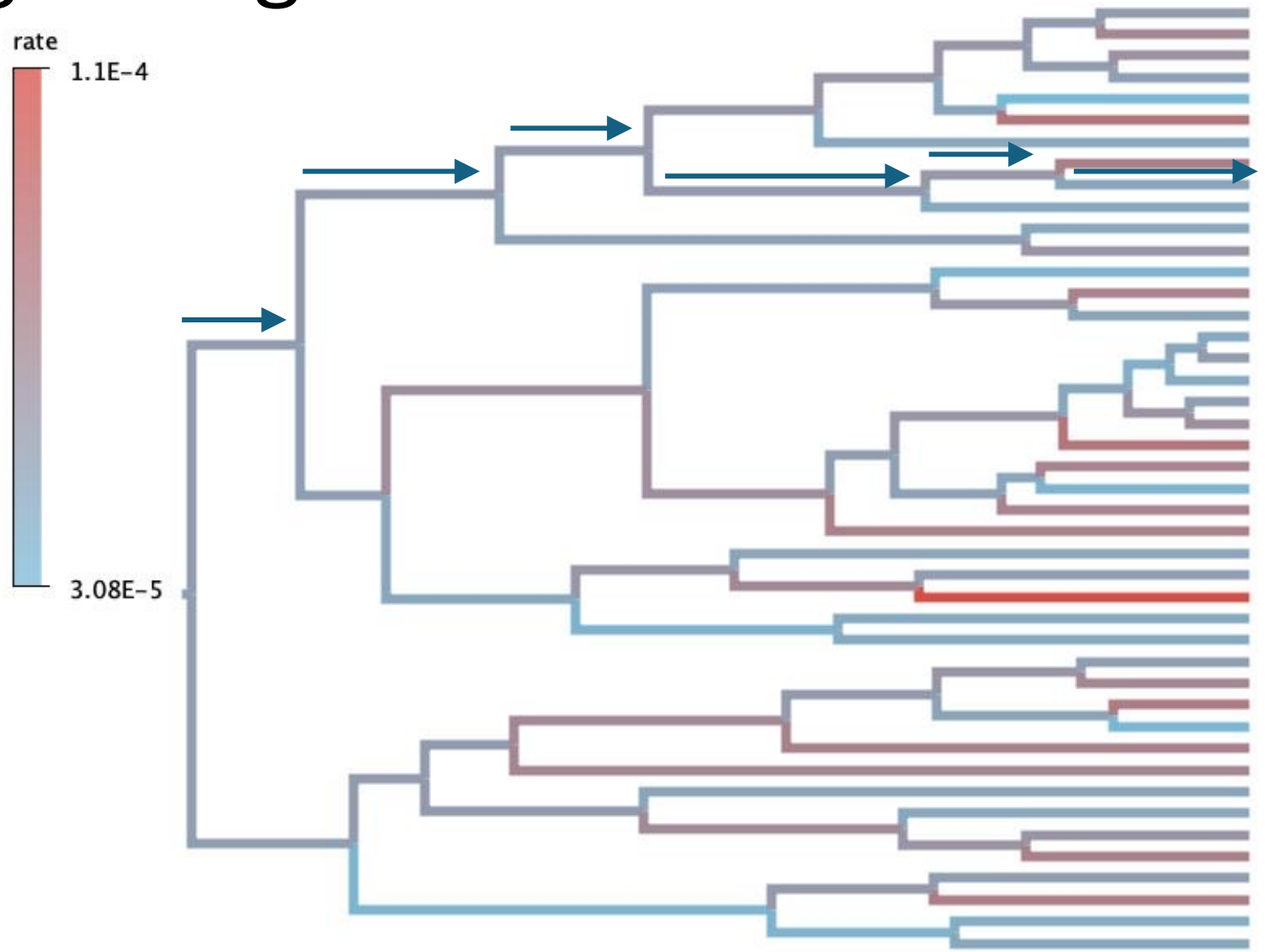
Tree model generates the branching process



Rate model (“clock”) determines speed of change on branches



Substitution model describes how traits change along the tree

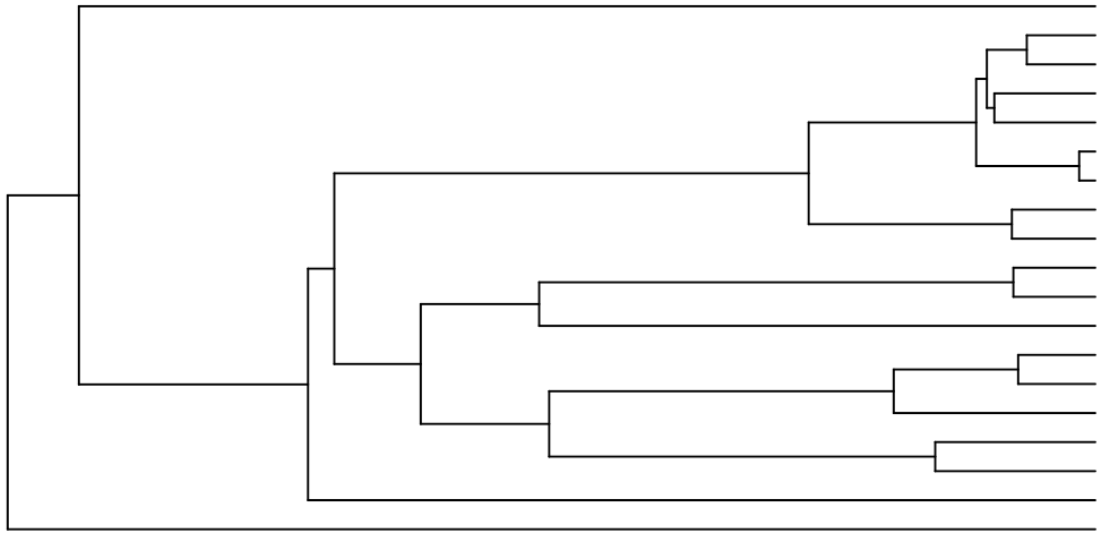


1. Tree models

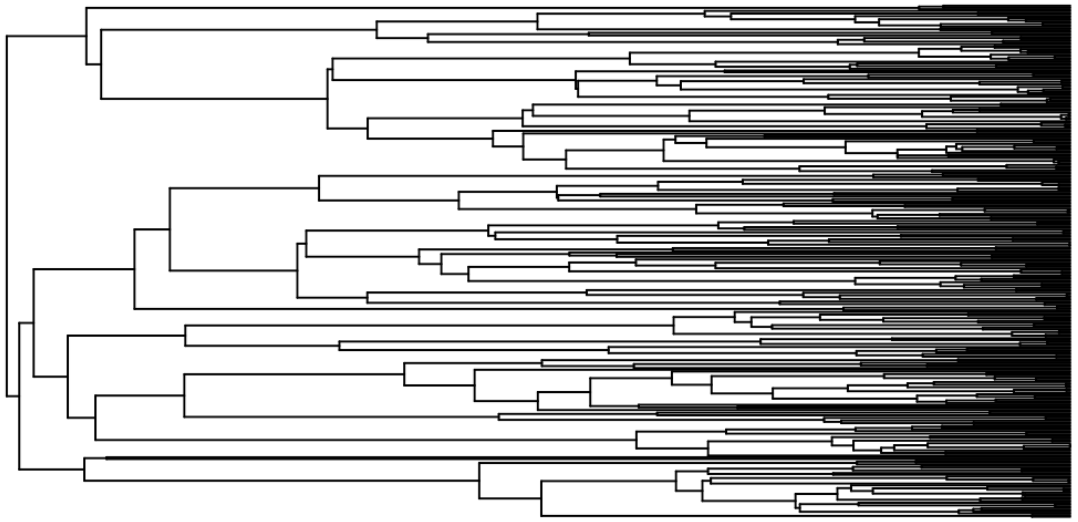
- Describe the branching of the tree
- Birth rate – how often languages branch in two
- Death rate – how often languages go extinct
- Any variation in the birth rate and death rate
- Sampling rate: how often through time languages are sampled, e.g. as written records

Yule model: pure birth

- Age=10
- Birth rate = 0.3
- Death rate = 0.0

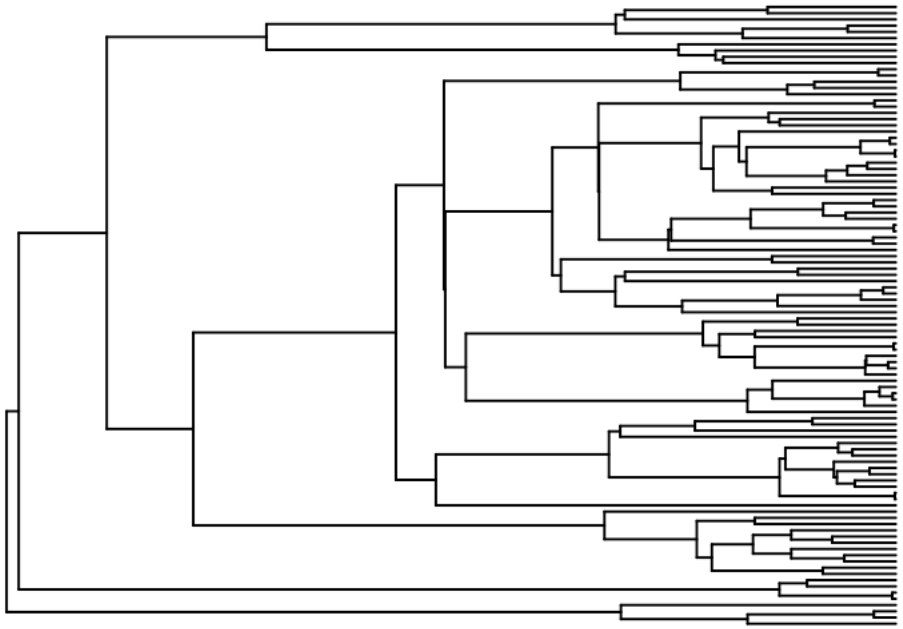


- Age=10
- Birth rate = 0.5
- Death rate = 0.0

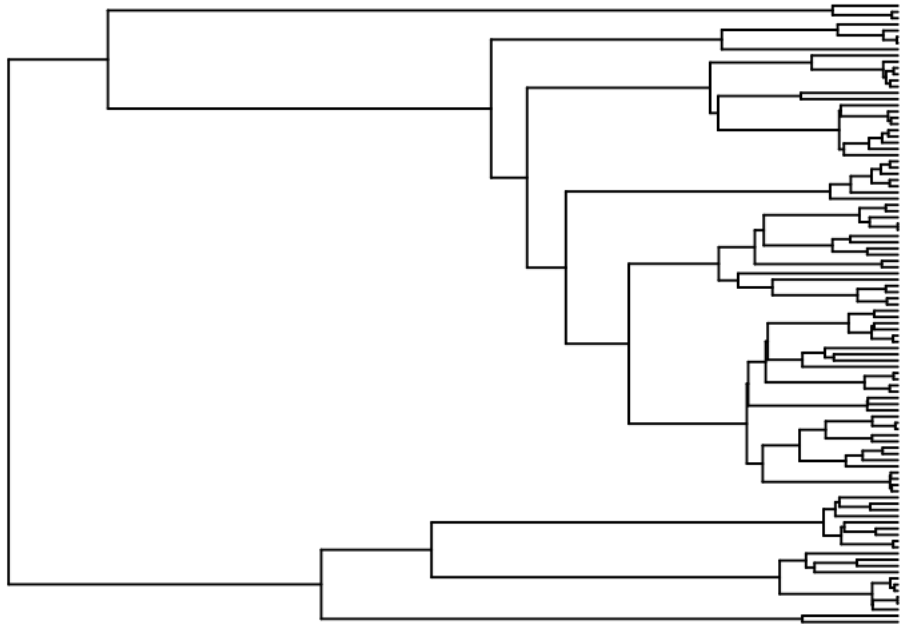


Birth-death model

- 100 taxa
- Birth rate = 0.3
- Death rate = 0.0

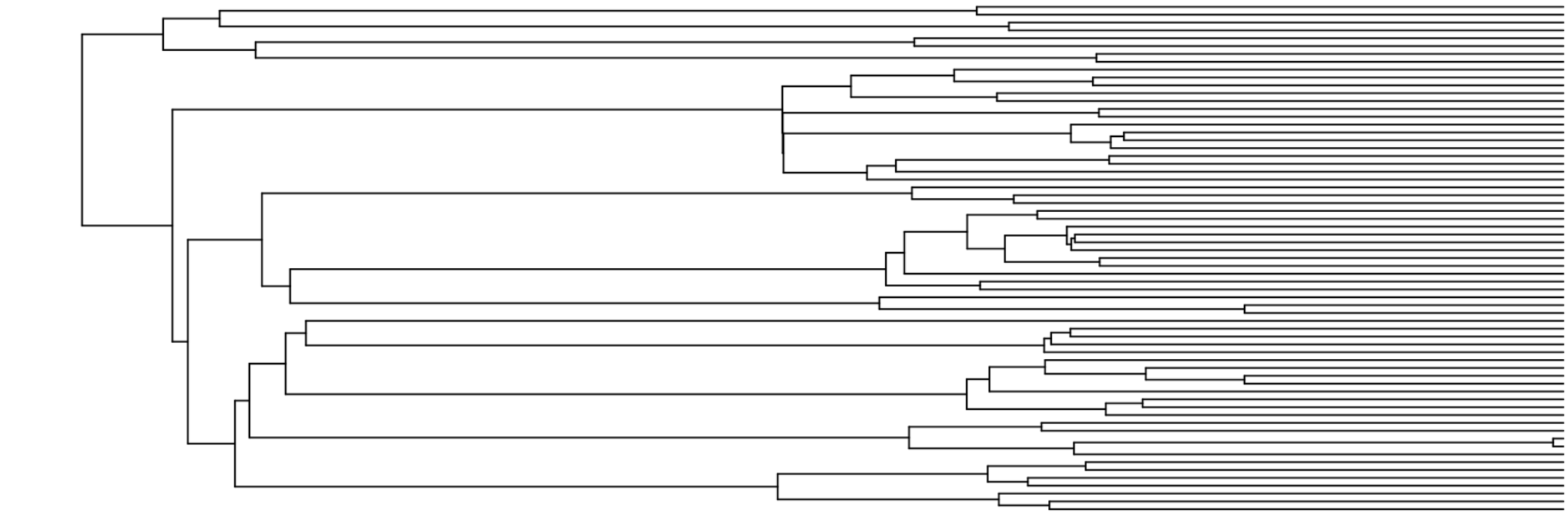


- 100 taxa
- Birth rate = 3.0
- Death rate = 2.9



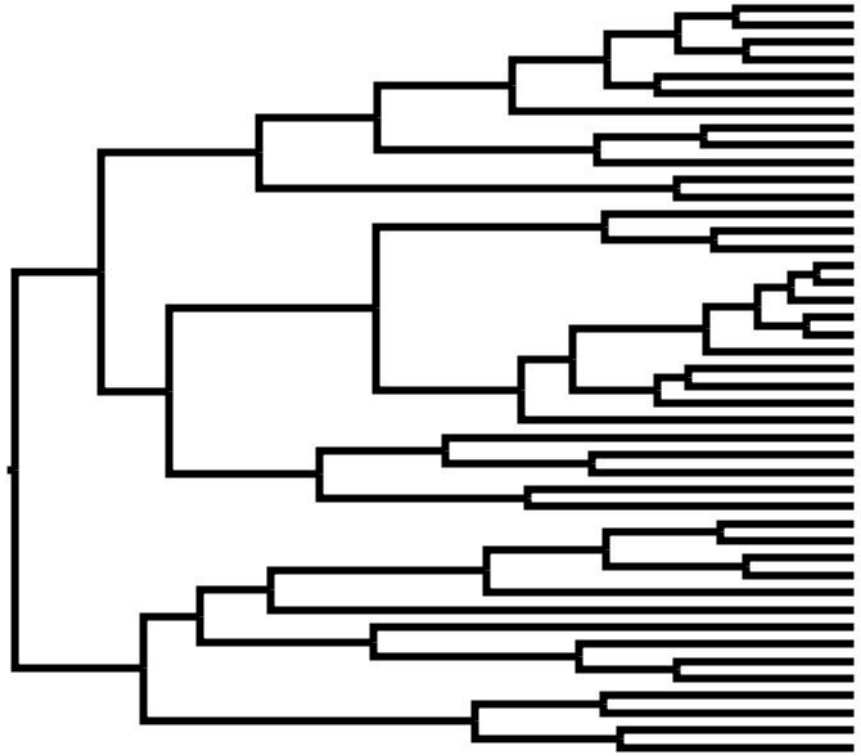
Birth-death skyline model

- Birth rate = 0.3
- Birth rate = 0.01
- Birth rate = 0.4
- Birth rate = 0.01

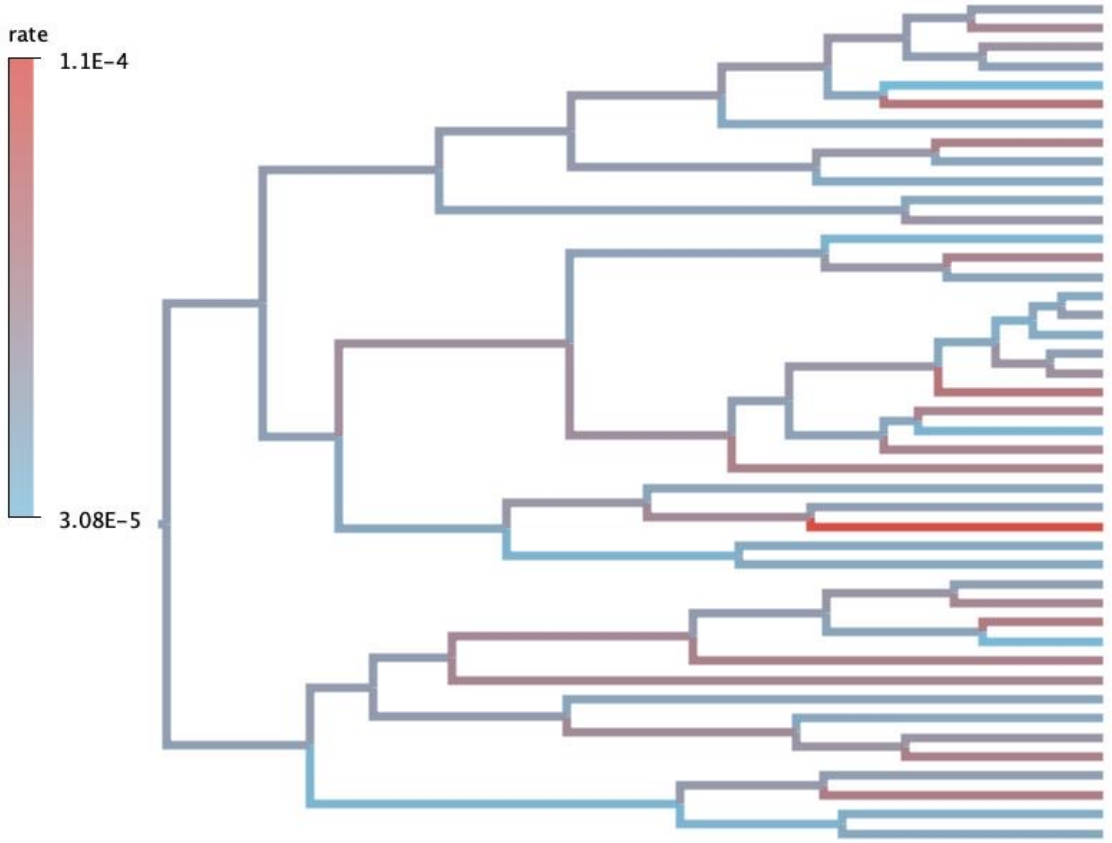


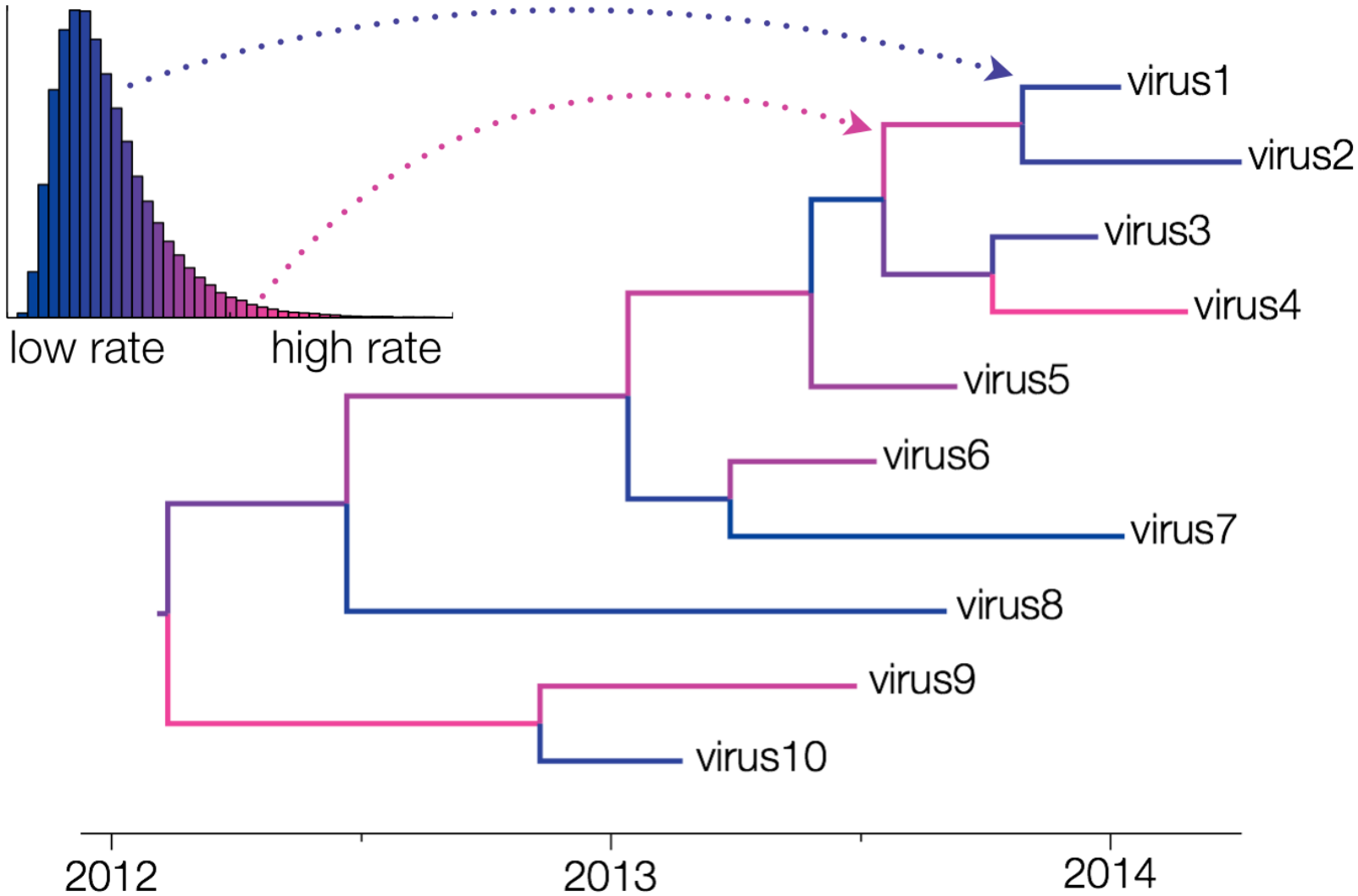
Rates model: the “linguistic clock”

Strict clock: all branches have a single rate



Relaxed clock: branch rates are different




low rate
high rate
virus1
virus2
virus3
virus4
virus5
virus6
virus7
virus8
virus9
virus10
2012
2013
2014

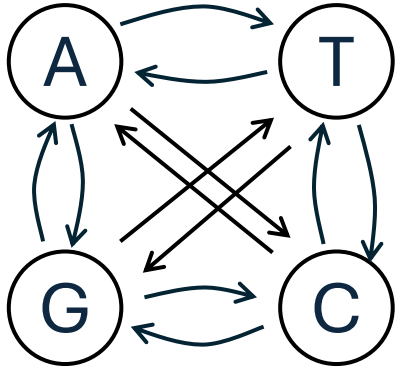
Clock model summary

2 model parameters:

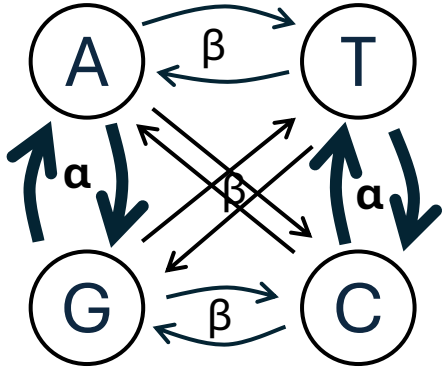
1. **Clock rate mean**
   - Average rate of change across all branches of the tree
2. **Clock rate standard devation**
   - How much variation between branches of the tree

Part 3: the substitution model

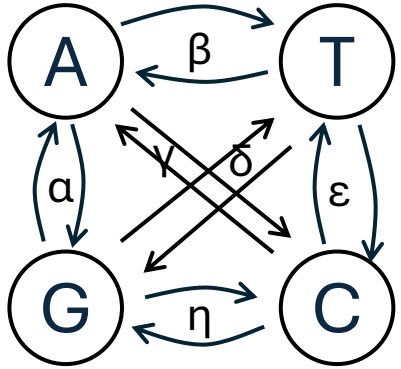
Jukes-Cantor model



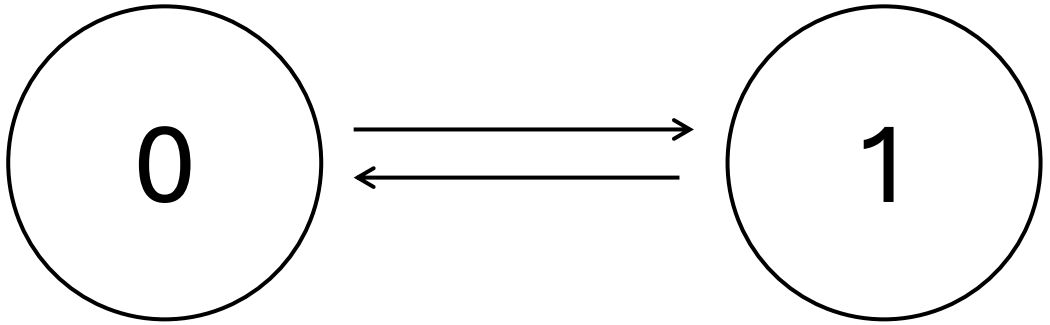
Kimura-80 model



General Time-Reversible model



For linguistic models we have 2 states: absent (0) and present (1)

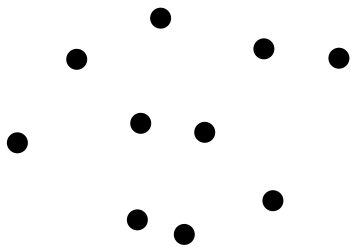


Rate of gain of a cognate and rate of loss of a cognate is determined by the **stable frequency** parameters

Stable frequencies are driven by rates

Chance of moving to city in the next year = 10%

Population = 10000

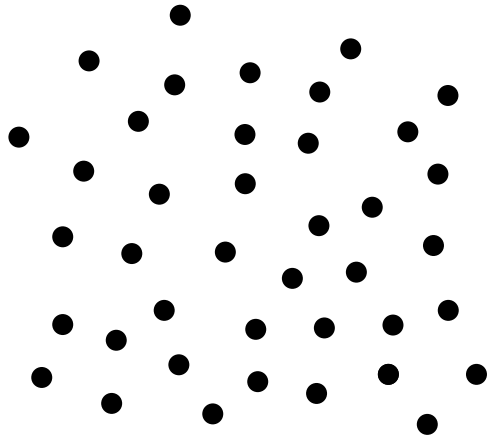


1000 people move

1000 people move

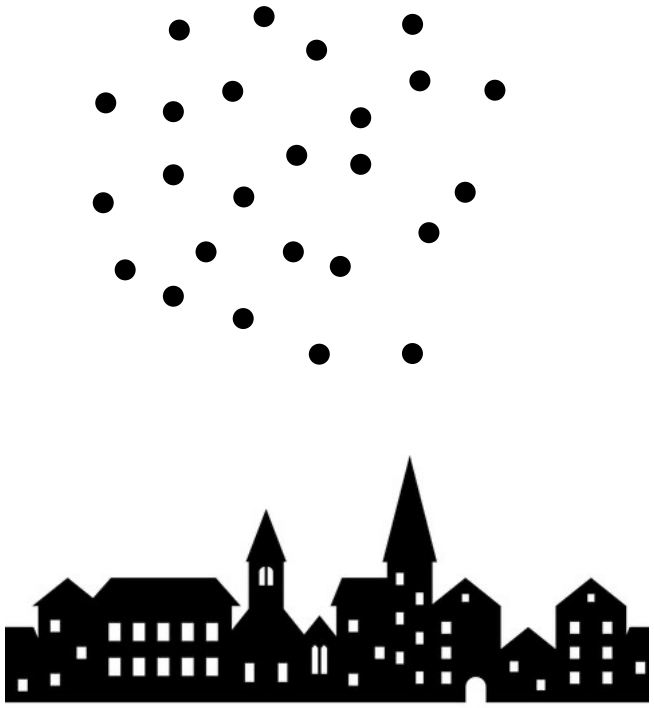
Chance of moving to town in the next year = 2.5%

Population = 40000



Chance of moving to city in the next year = 10%

Population = 25000



2500 people move

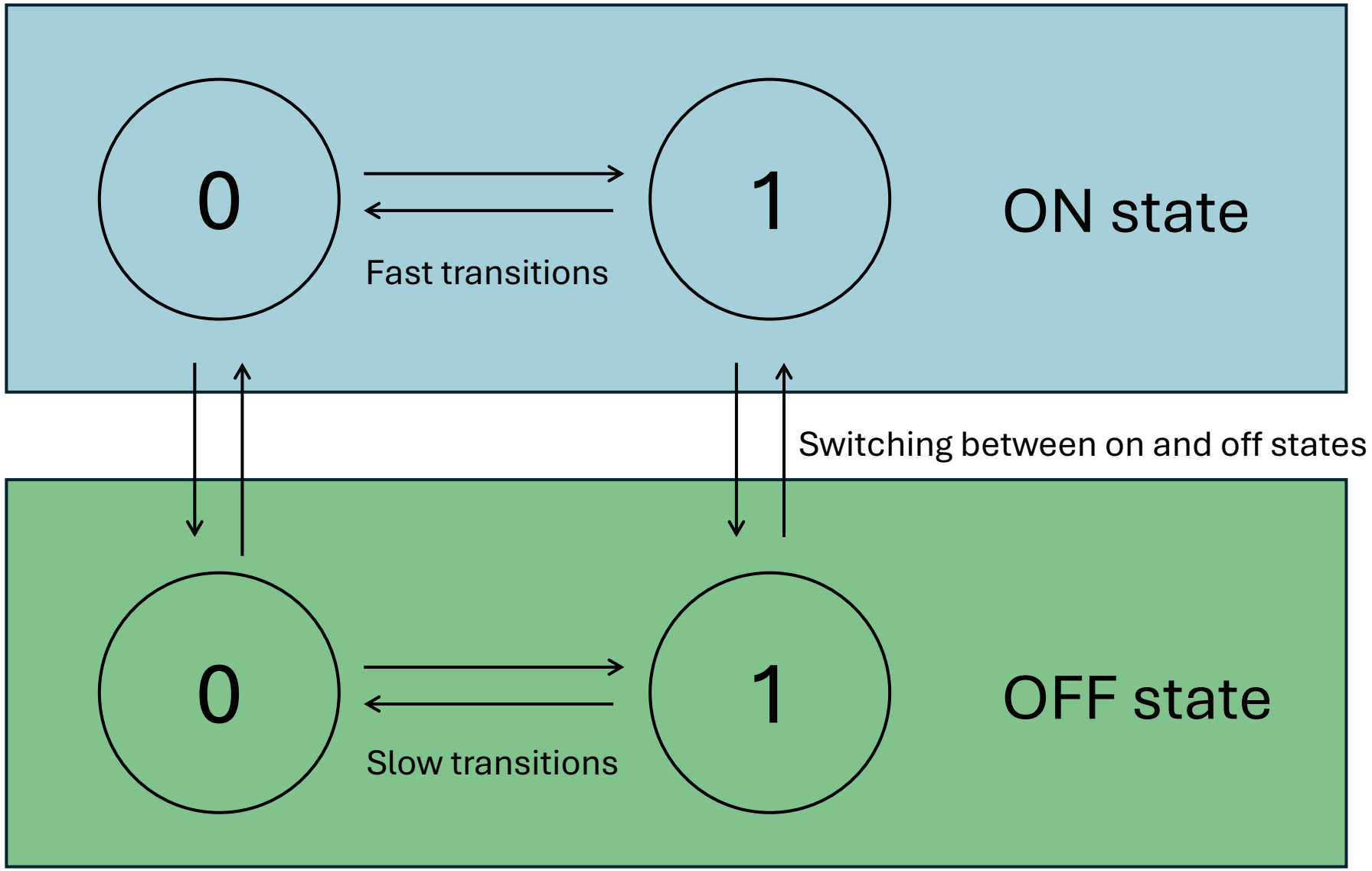
625 people move

Chance of moving to town in the next year = 2.5%

Population = 25000

Stable frequencies of the present and absent states

The binary covarion model

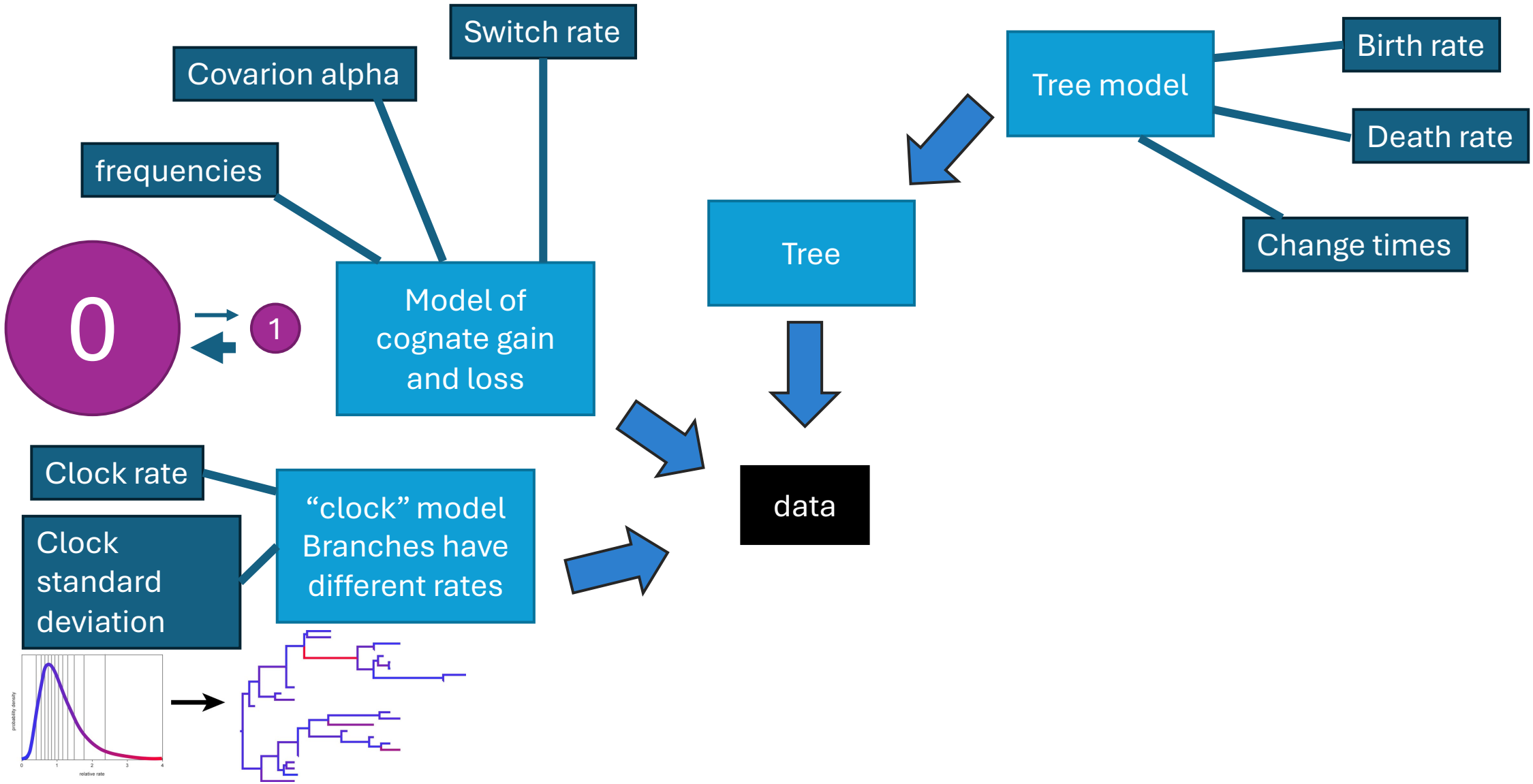


Substitution model summary

3 model parameters:

1. **Stable frequencies** of present and absent states
   - Determines how common gains and losses of a cognate are
2. **Switch rate** between the on and off states
3. **Relative rate of the on and off states** (covarion alpha parameter)

Overview of a generative phylogenetic model

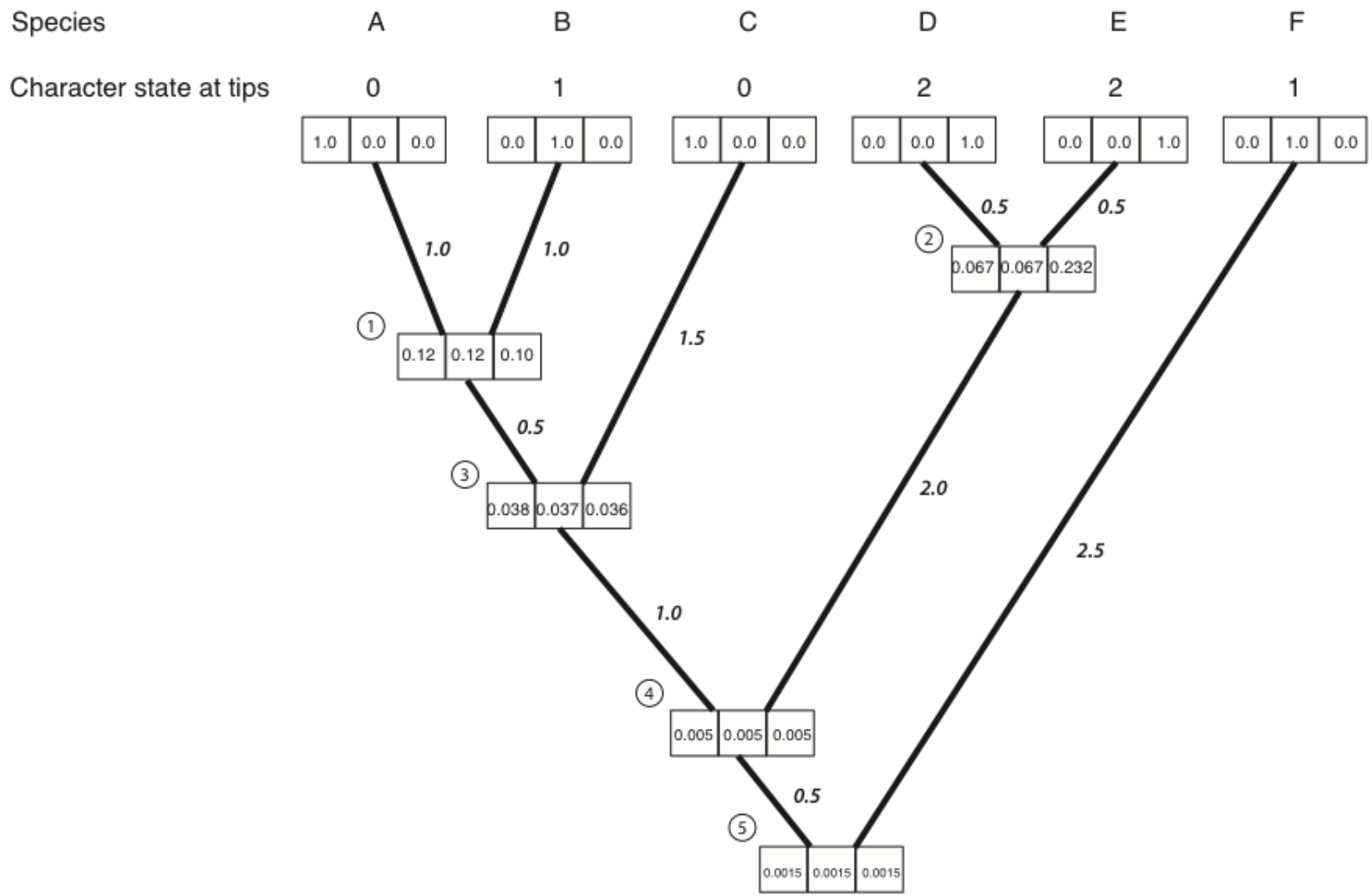


Phylogenetic inference

- A phylogenetic generative model describes how we go from the model to generating data
- We hope that this generative model is a good approximation of the real process that produced the real data
- In phylogenetic inference we must work backwards to estimate the model parameters from the data

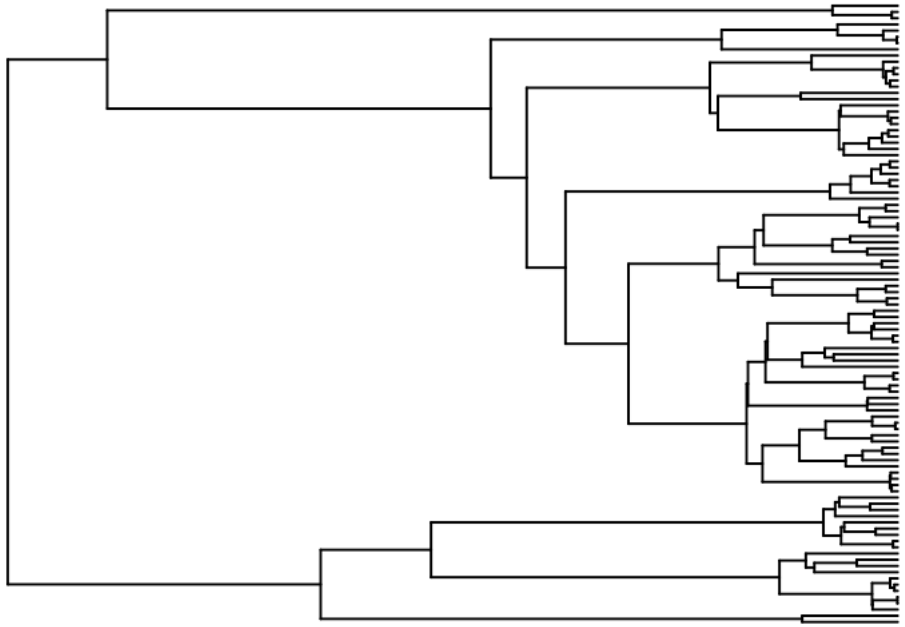
Step 1: probability of the data (the likelihood)

- Felsenstein's pruning algorithm
- Given a tree, branch rates and a substitution model, we can calculate the probability of observing the data



Step 2: probability of the tree

- 100 taxa
- Birth rate = 3.0
- Death rate = 2.9



$$f[\mathcal{T} \mid \lambda, \mu, \psi, \rho, t, S] = \frac{q_1(0)}{1 - p_1(0)} \prod_{i=1}^{N+n-1} \lambda_{l(x_i)} q_{l(x_i)}(x_i)$$

$$\cdot \prod_{i=1}^{n} \frac{\psi_{l(y_i)}}{q_{l(y_i)}(y_i)} \prod_{i=1}^{m} \rho_i^{N_i} ((1 - \rho_i) q_{i+1}(t_i))^{n_i},$$

with $l(t) = i$ iff $t_{i-1} \leq t < t_i$, and with $(t_{i-1} \leq t < t_i,\ i = 1, \ldots, m)$,

$$A_i = \sqrt{(\lambda_i - \mu_i - \psi_i)^2 + 4\lambda_i \psi_i},$$

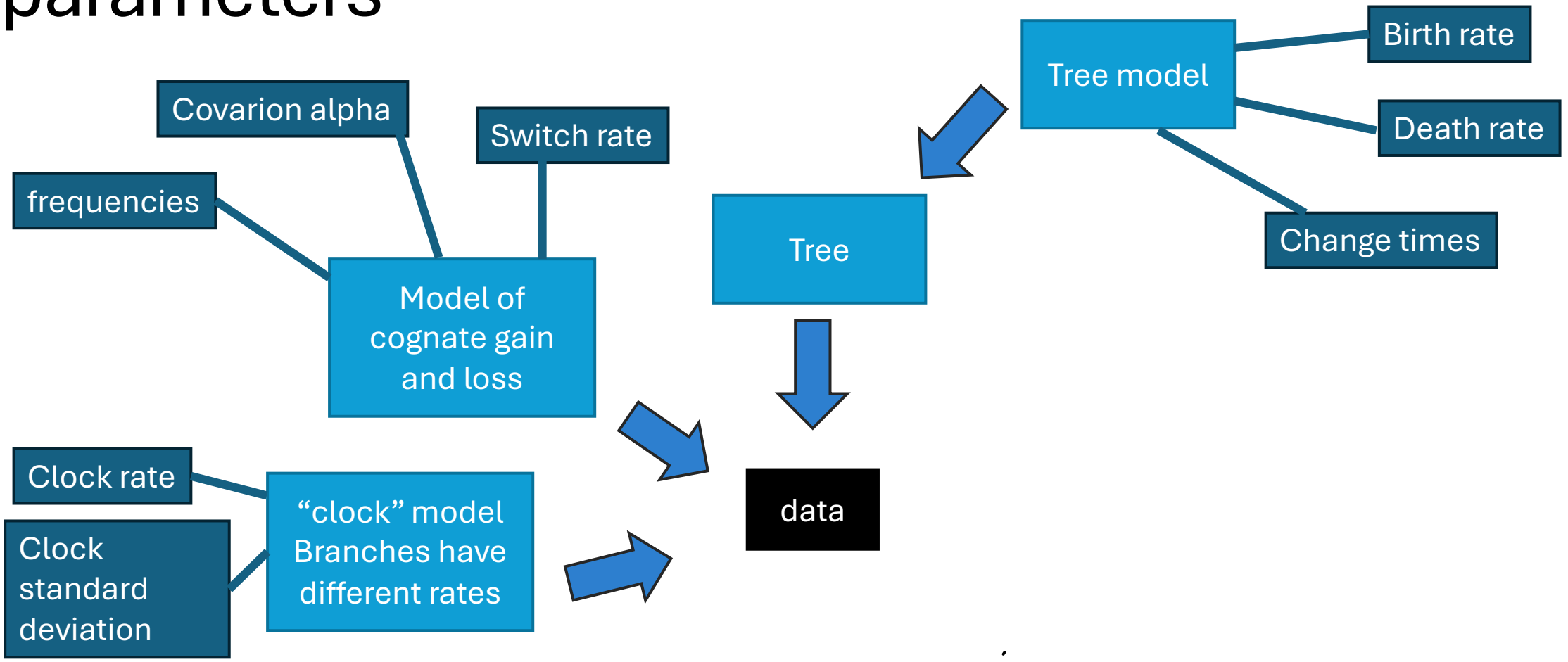
$$B_i = \frac{(1 - 2(1 - \rho_i) p_{i+1}(t_i))\lambda_i + \mu_i + \psi_i}{A_i},$$

$$p_i(t) = \frac{\lambda_i + \mu_i + \psi_i - A_i \dfrac{e^{A_i(t - t_i)}(1 + B_i) - (1 - B_i)}{e^{A_i(t - t_i)}(1 + B_i) + (1 - B_i)}}{2\lambda_i},$$

$$q_i(t) = \frac{4e^{-A_i(t - t_i)}}{(e^{-A_i(t - t_i)}(1 + B_i) + (1 - B_i))^2},$$

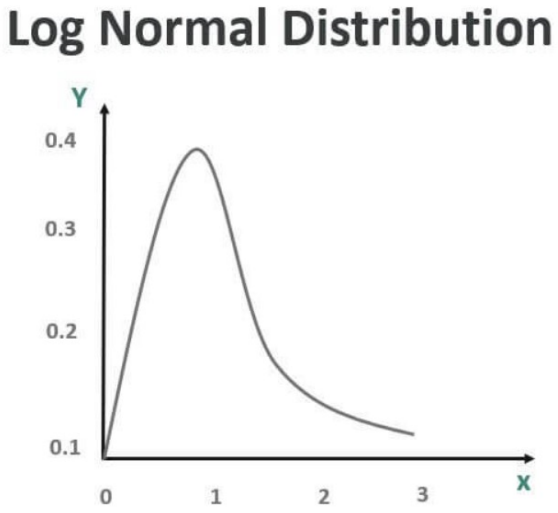
It's complicated…

Step 3: prior probabilities for model parameters

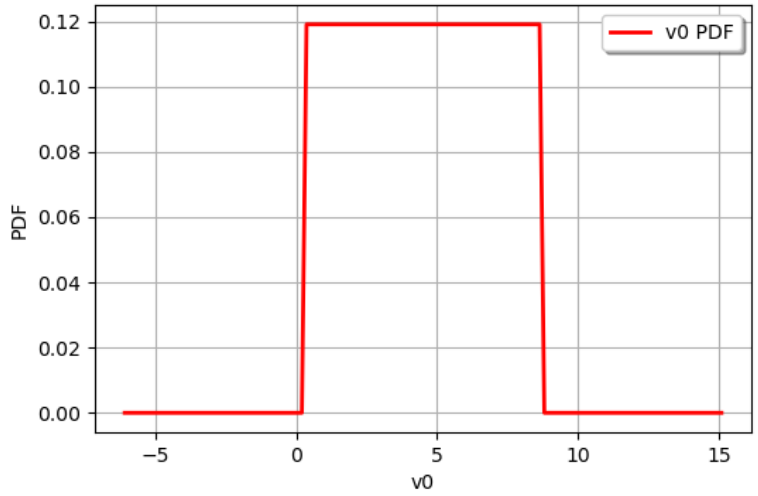


Prior probabilities

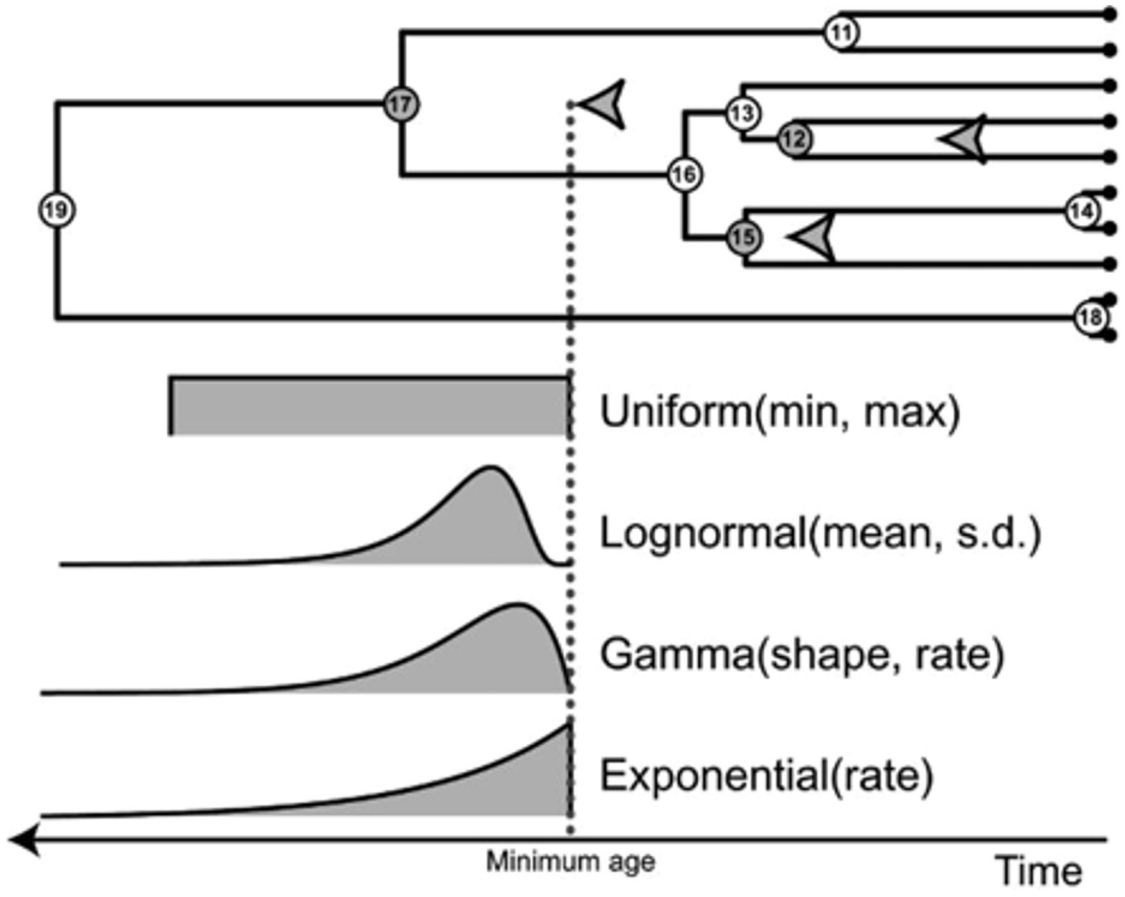
- Describe knowledge about what sensible parameter values are
- This knowledge should not come from looking at the data
- Often we have no other knowledge, and use uninformative priors (high standard deviation on a normal or lognormal distribution), or a uniform distribution



Uniform distribution



- The clock model can be **calibrated** using a **Node age prior**

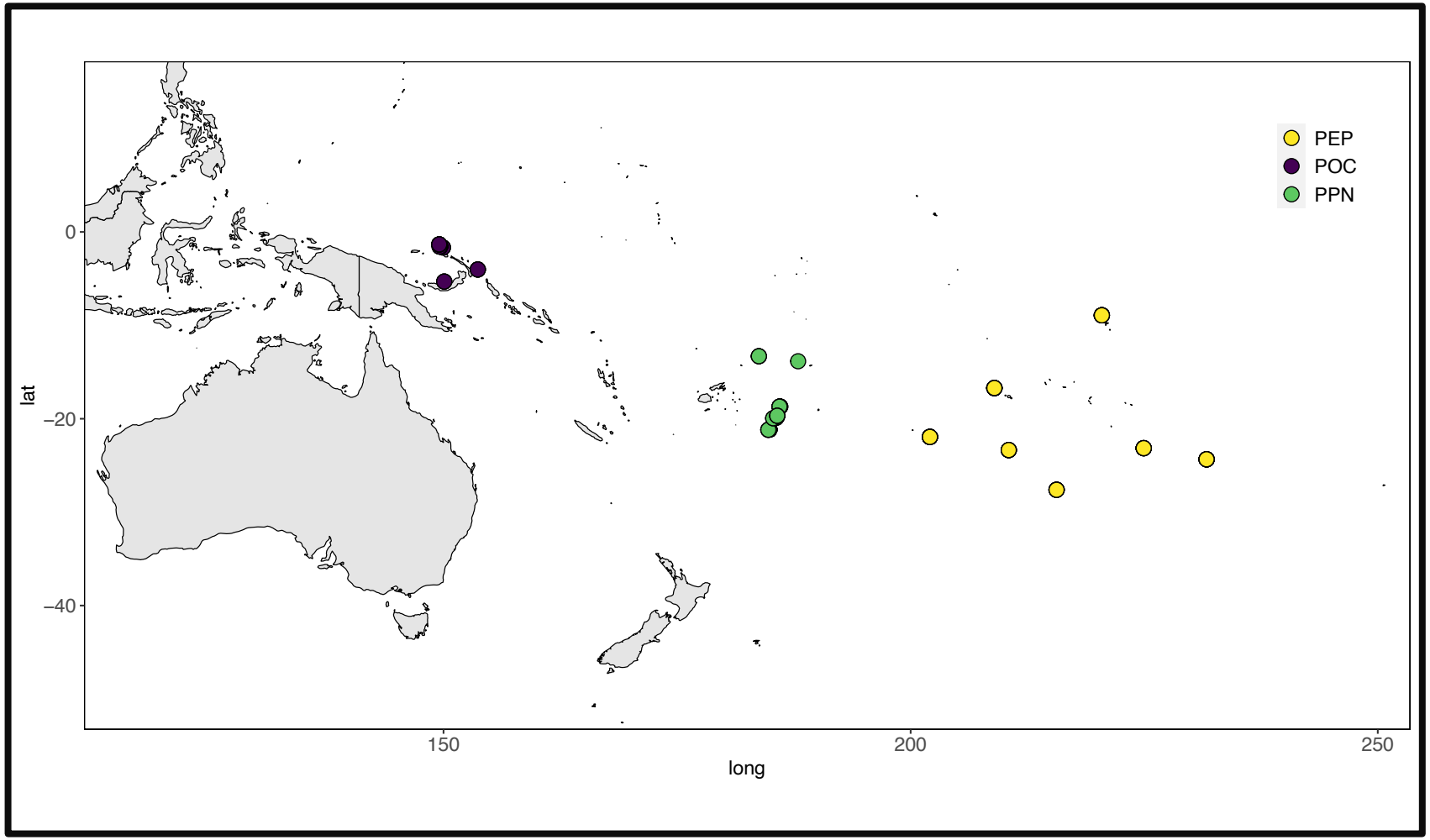


Node calibrations are set as **prior probability** distributions on node age in a Bayesian analysis.

These are usually in the form a probability distribution, with a lower bound set at the age of the calibration fossil

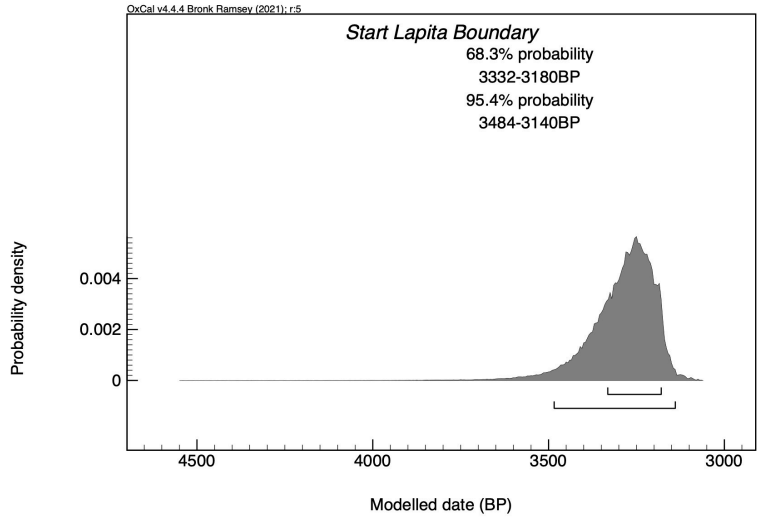
Node calibration example: Pacific islands

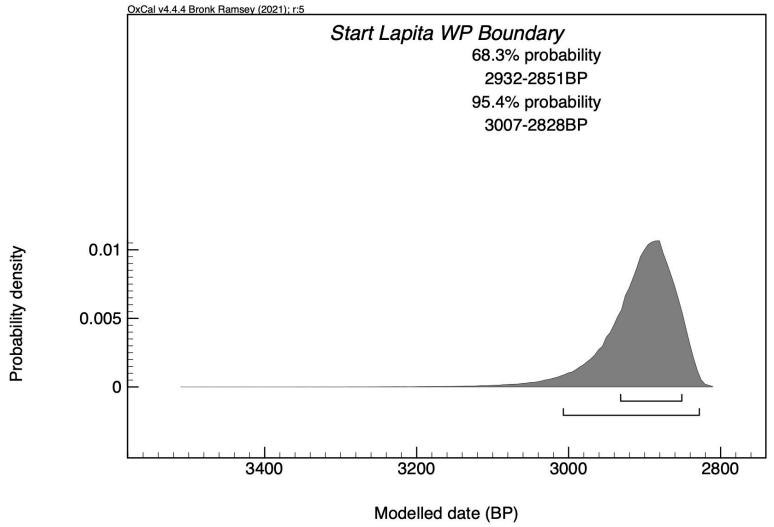
- **Radiocarbon dates for Northern Melanesia, Western Polynesia and Eastern Polynesia**

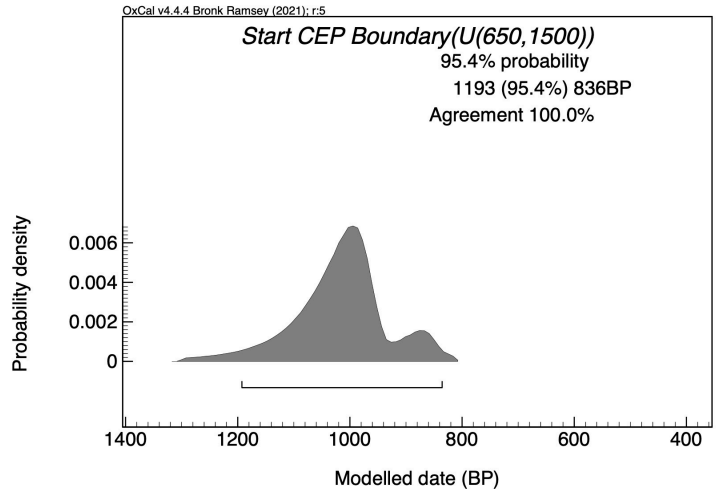


Radiocarbon dates

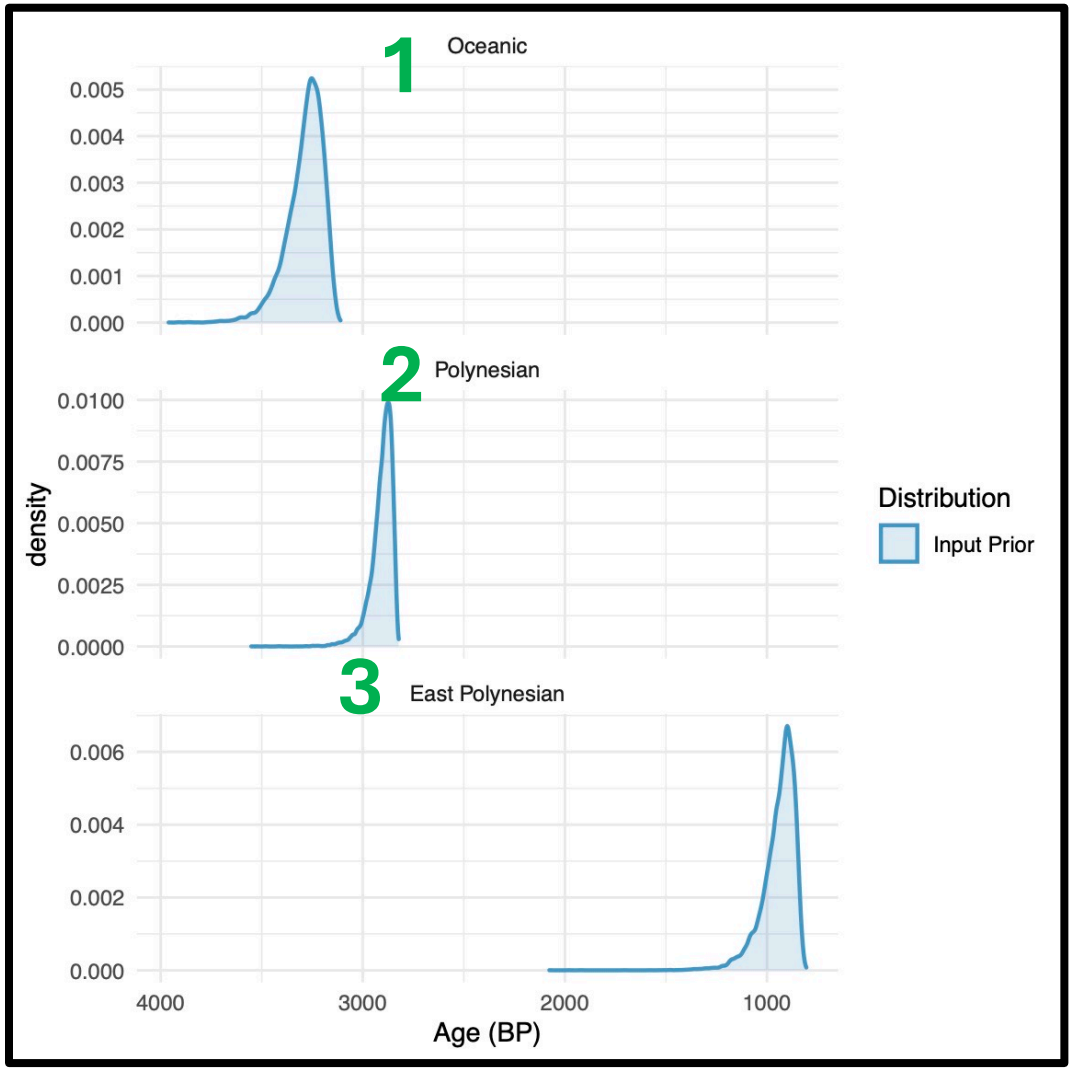
- Approximate first arrival in Melanesia, Polynesia and Eastern Polynesia
- If we assume this marks the beginning of independent language evolution, this information can be used for node calibration

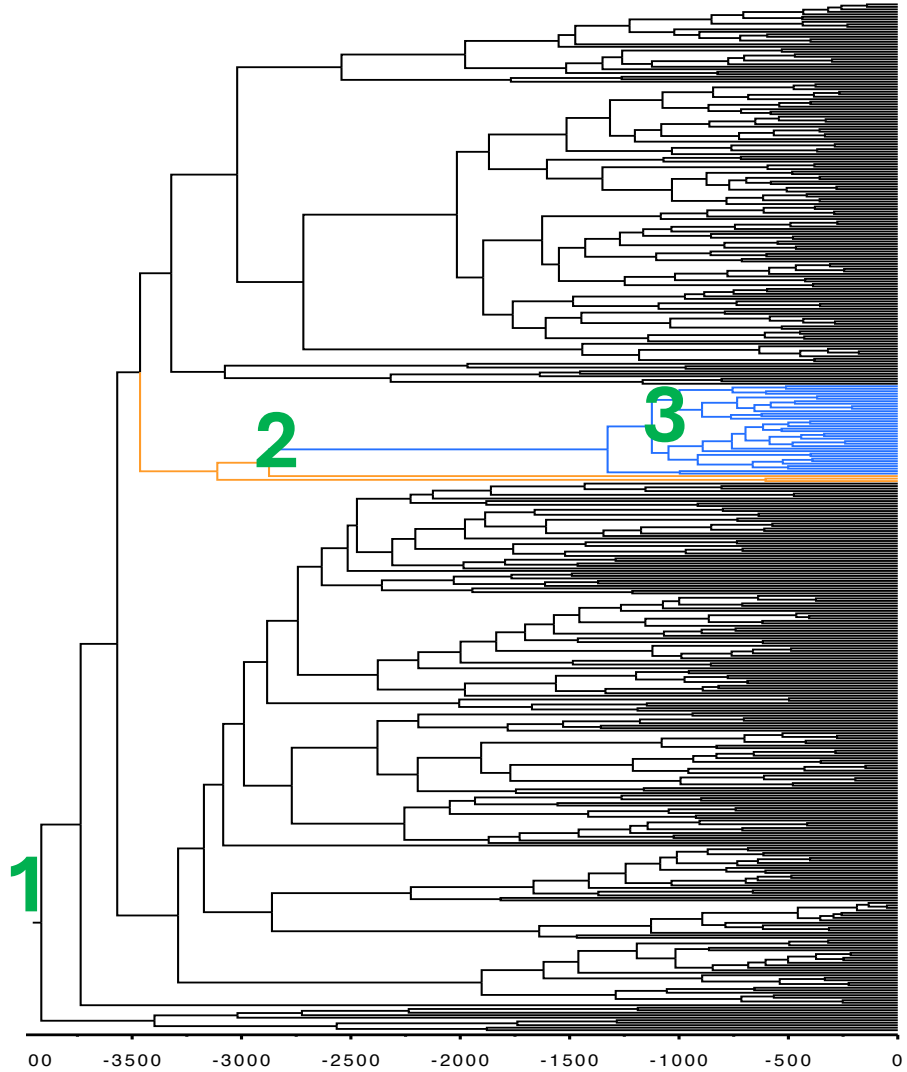




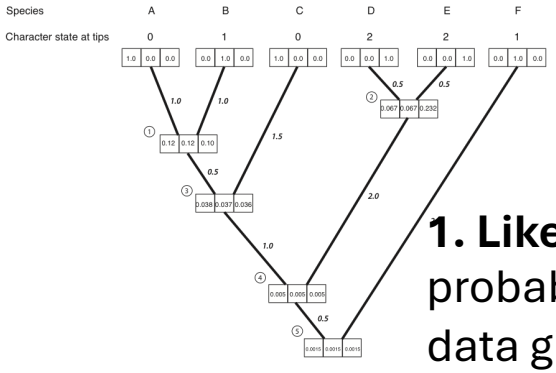


Implementing node age calibrations



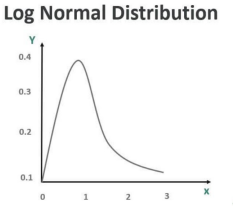


Putting it all together: Bayes theorum



1. Likelihood: probability of the data given the tree and parameter values

2. Probability of the tree



3. Prior probabilities of model parameters

likelihood

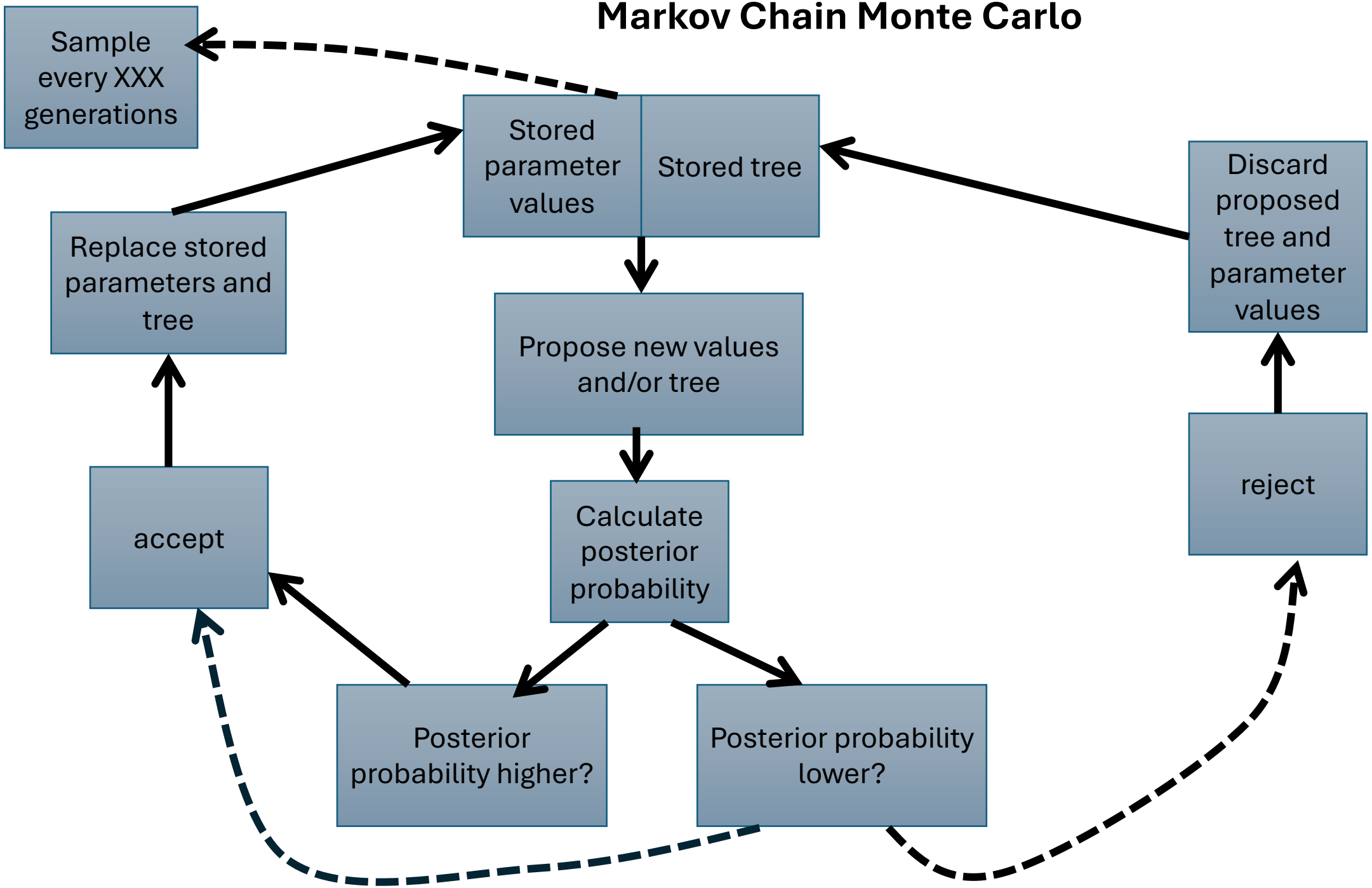
prior

Posterior probability

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)}$$

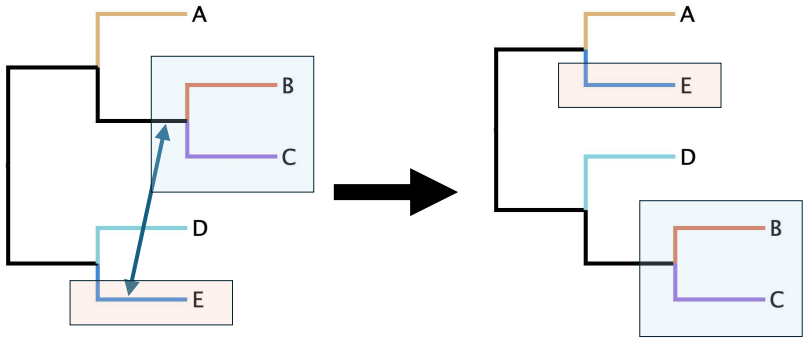
Model evidence

Markov Chain Monte Carlo

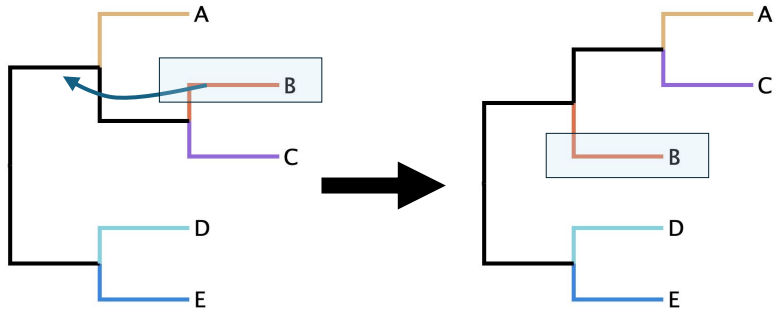


Operators

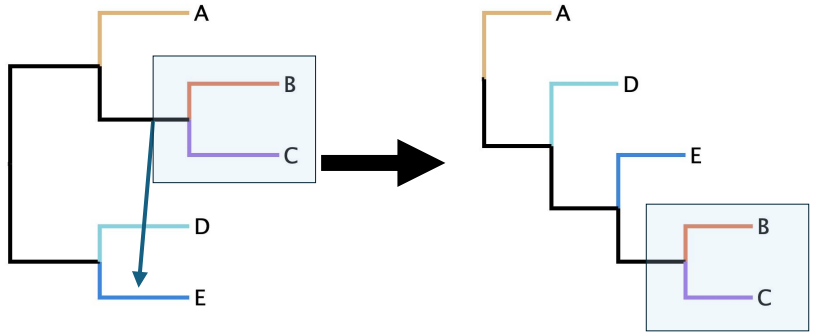
Exchange operator

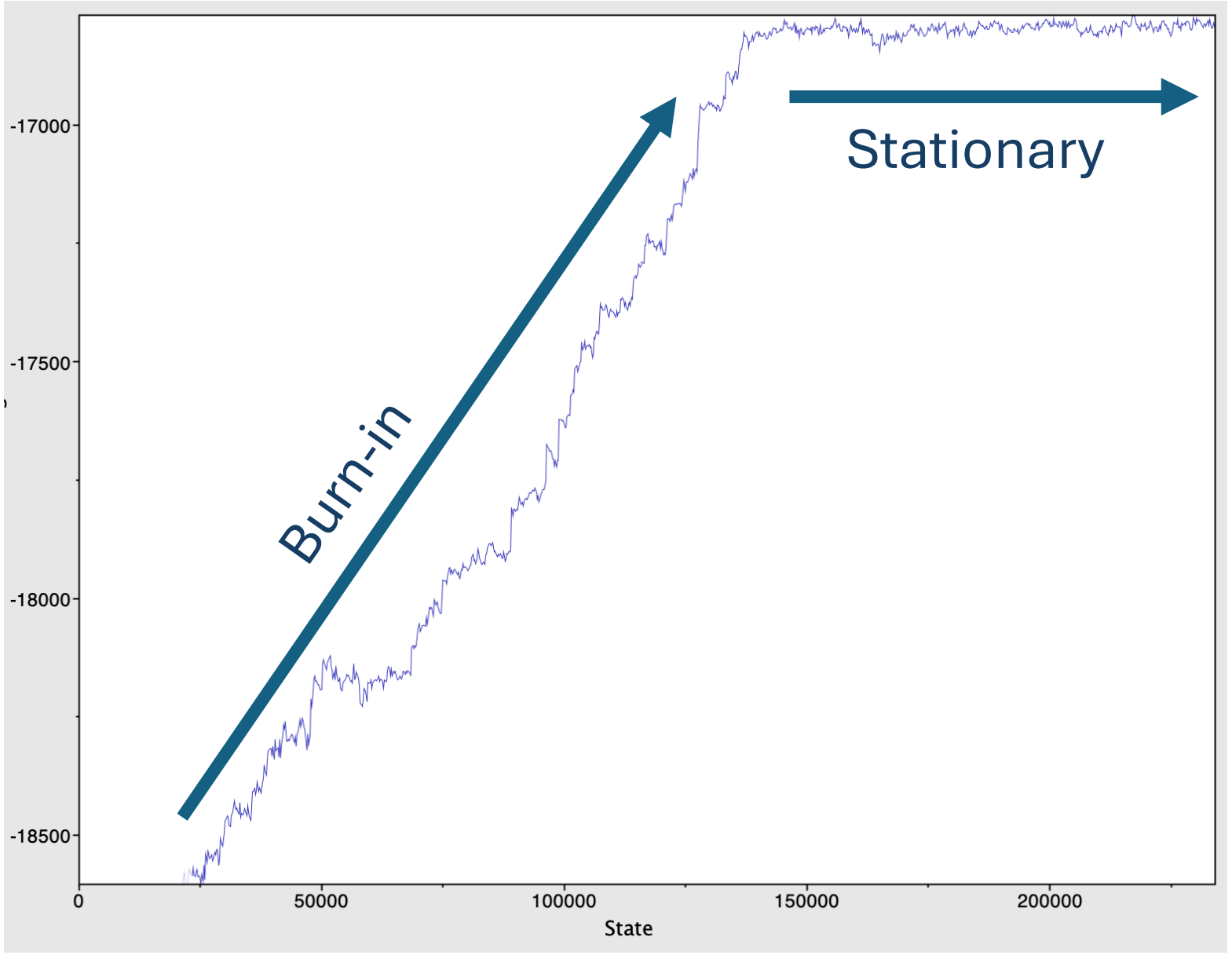


Subtree slide operator

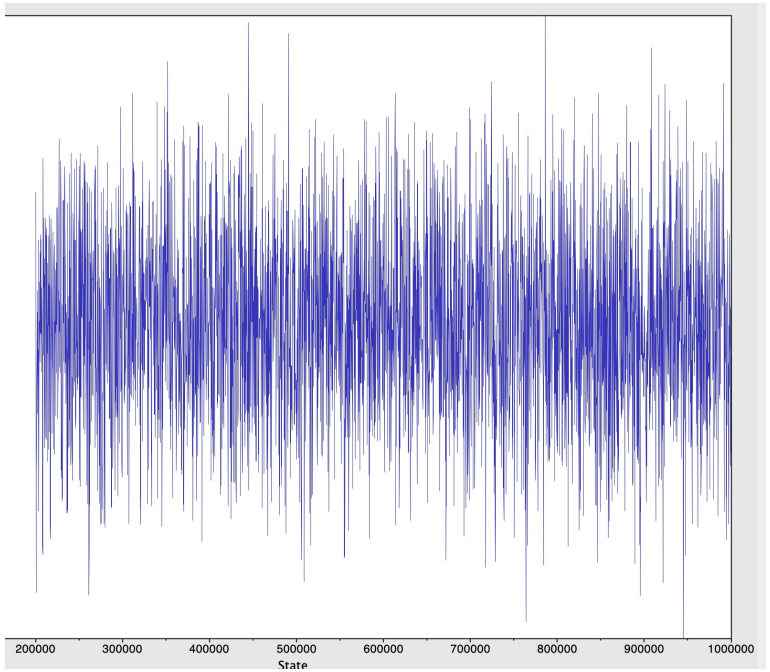


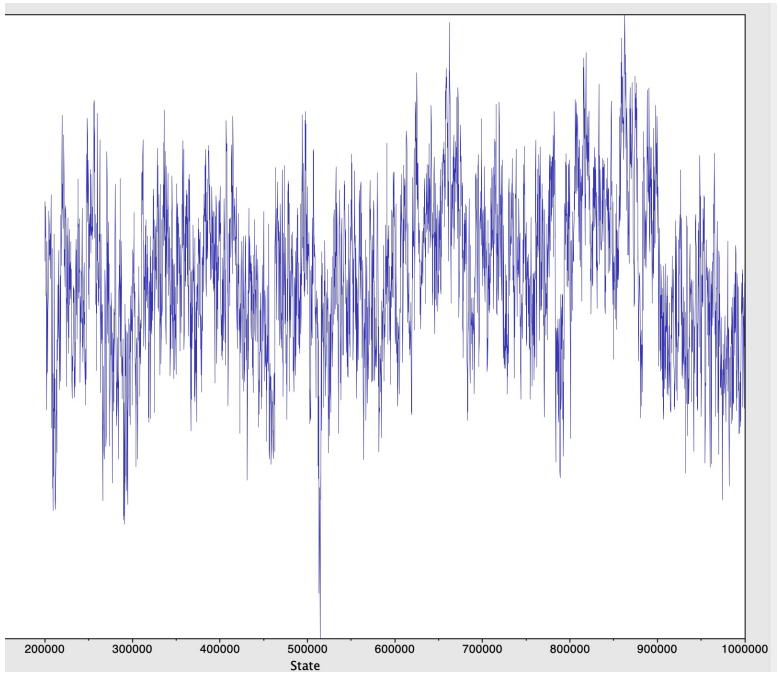
Wilson-Balding operator

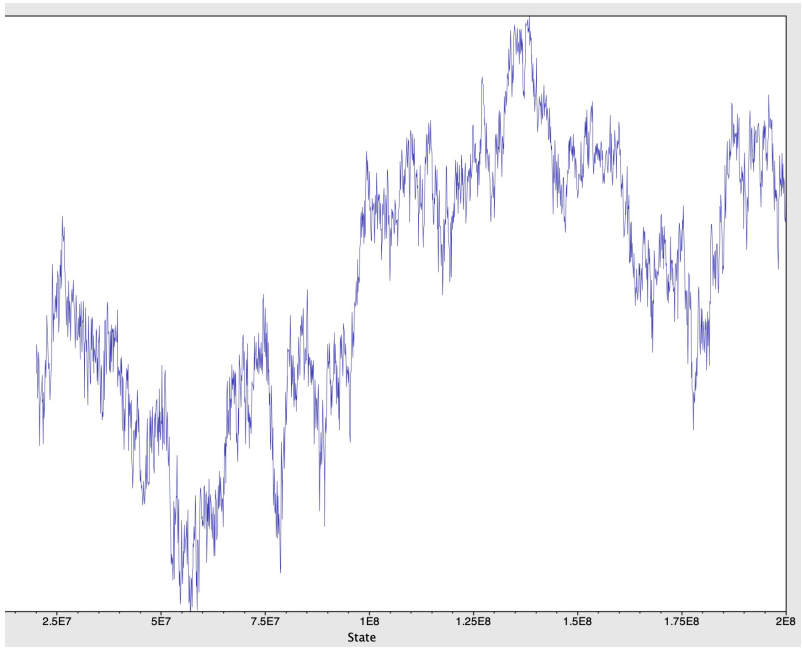



-17000
-17500
-18000
-18500
0
50000
100000
150000
200000
State
Burn-in
Stationary

Good and bad chains



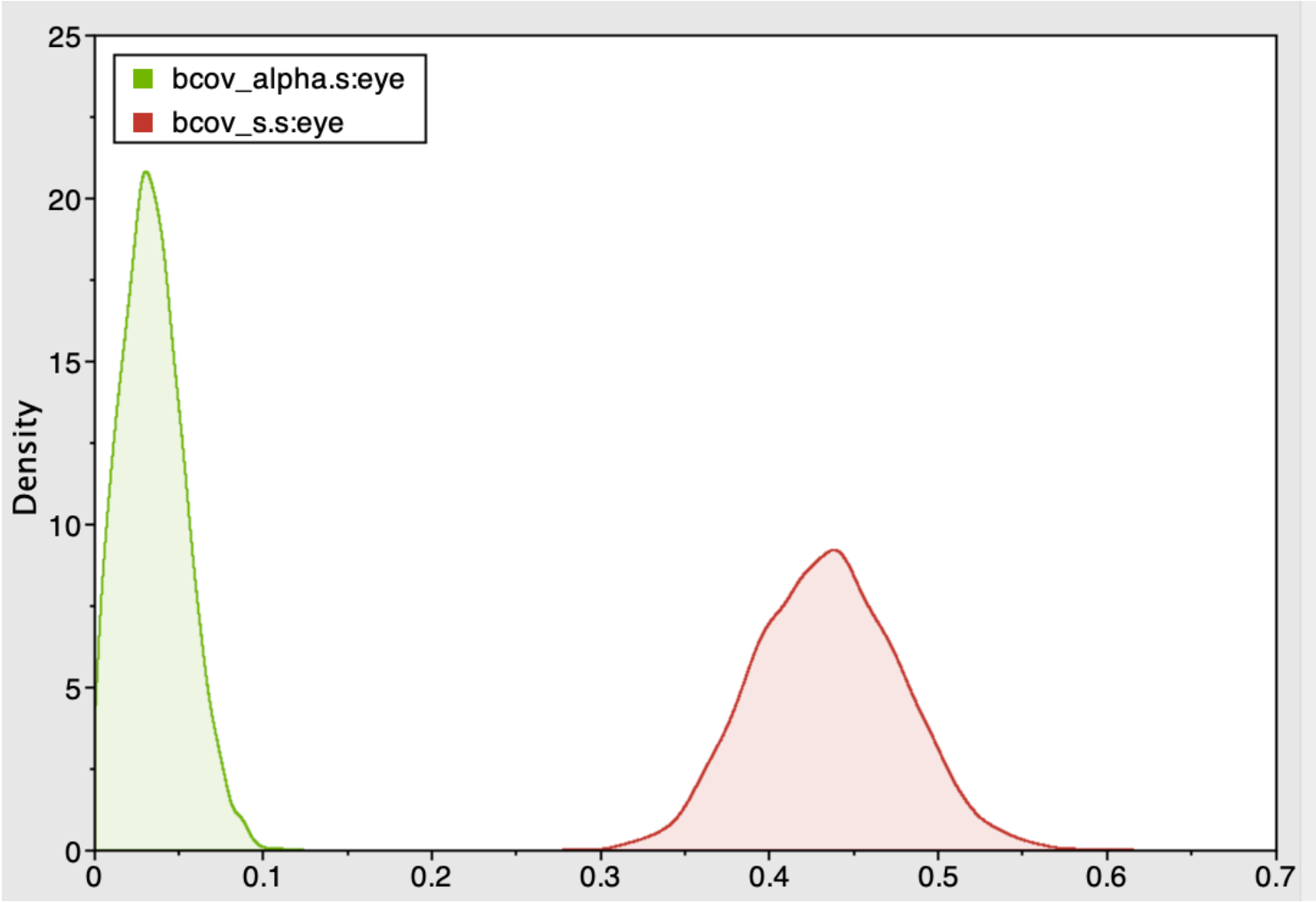




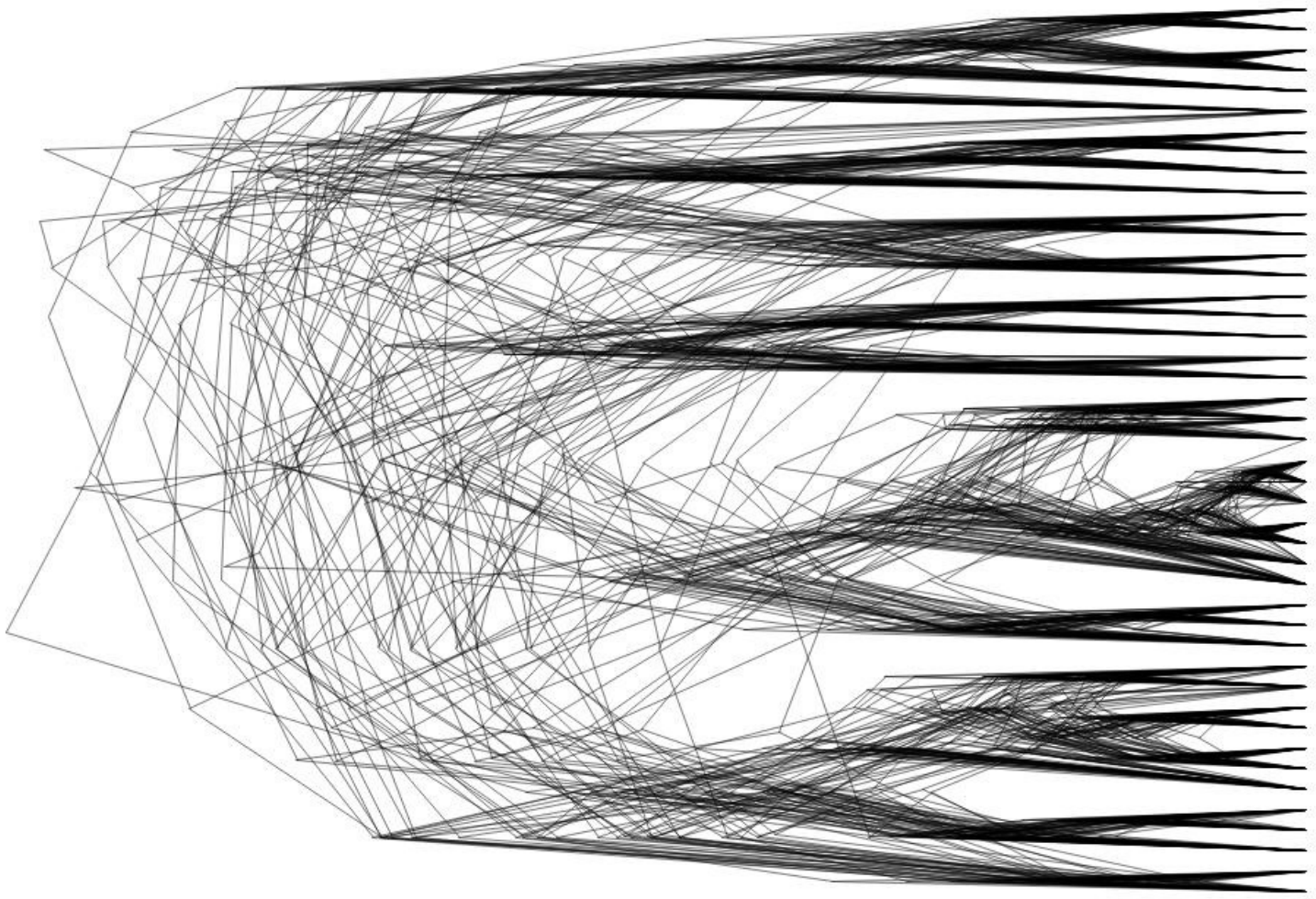
The end result is a sample

Sample	posterior	likelihood	prior	Tree model	birth rate	covarion alpha	covarion switch	frequency absent	frequency present	clock mean	clock standard dev
10000	-16826.3	-16430.0	-396.4	-339.8	9.86E-04	0.0394	0.4450	0.9535	0.0465	5.14E-05	0.4874
20000	-16820.6	-16435.6	-385.0	-318.4	1.50E-03	0.0155	0.4249	0.9502	0.0498	9.48E-05	0.5483
30000	-16841.3	-16429.1	-412.2	-336.1	1.02E-03	0.0310	0.5375	0.9528	0.0472	6.18E-05	0.7948
40000	-16827.4	-16441.0	-386.4	-336.6	9.03E-04	0.0232	0.4400	0.9526	0.0474	6.30E-05	0.5497
50000	-16821.1	-16430.1	-391.0	-341.7	8.66E-04	0.0590	0.3955	0.9526	0.0474	5.43E-05	0.4969
60000	-16839.0	-16447.7	-391.4	-330.5	1.26E-03	0.0651	0.3525	0.9552	0.0448	6.40E-05	0.5866
70000	-16813.5	-16434.5	-378.9	-327.5	9.98E-04	0.0562	0.3988	0.9540	0.0460	7.90E-05	0.5809
80000	-16810.5	-16426.9	-383.6	-328.2	1.18E-03	0.0562	0.4185	0.9541	0.0459	7.25E-05	0.5421
90000	-16843.1	-16443.6	-399.5	-341.8	8.29E-04	0.0562	0.4501	0.9541	0.0459	5.31E-05	0.7221
100000	-16822.7	-16430.8	-391.9	-346.0	8.70E-04	0.0313	0.3919	0.9541	0.0459	4.37E-05	0.4759

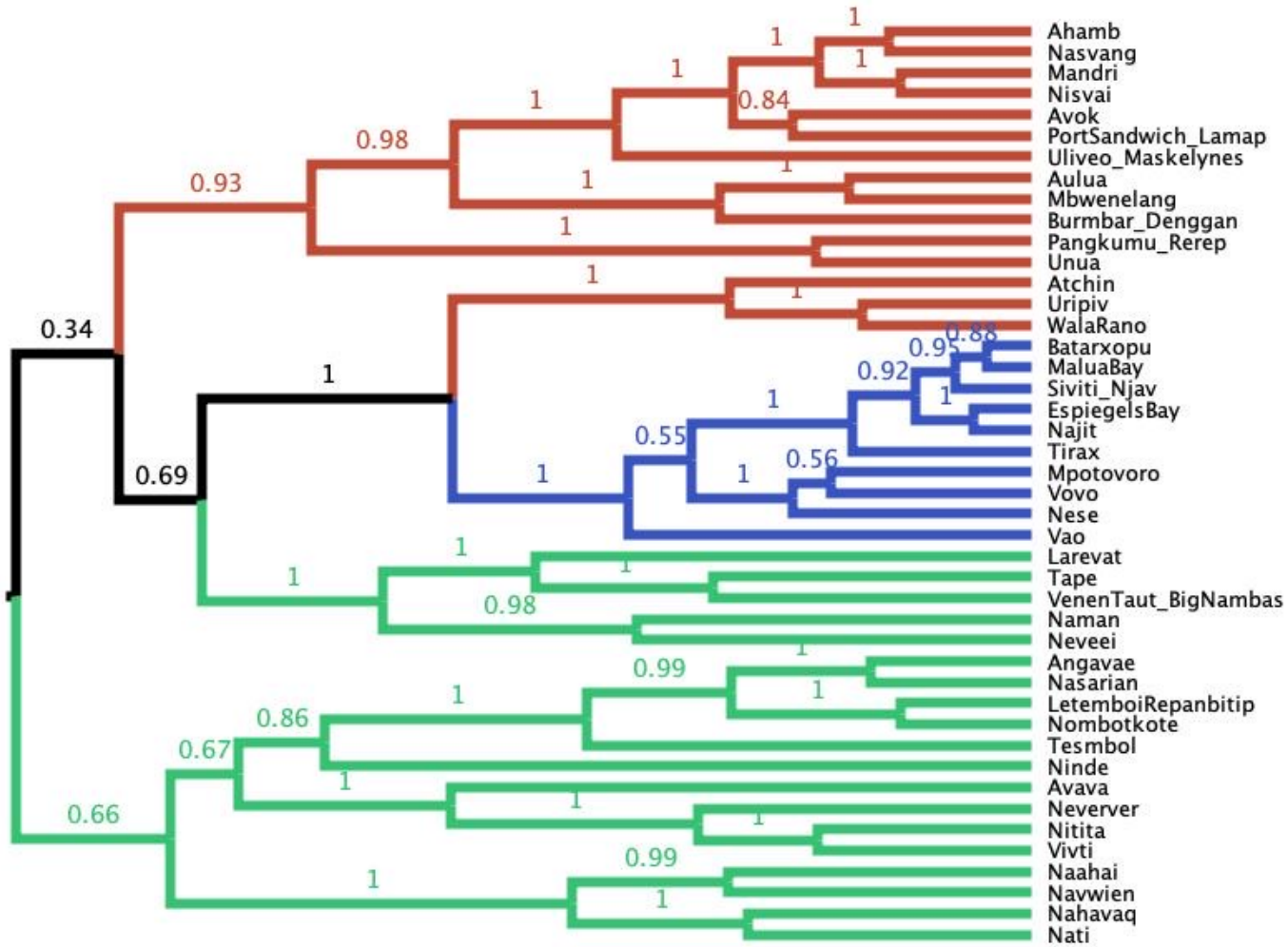
Summarising output - parameters



Tree samples



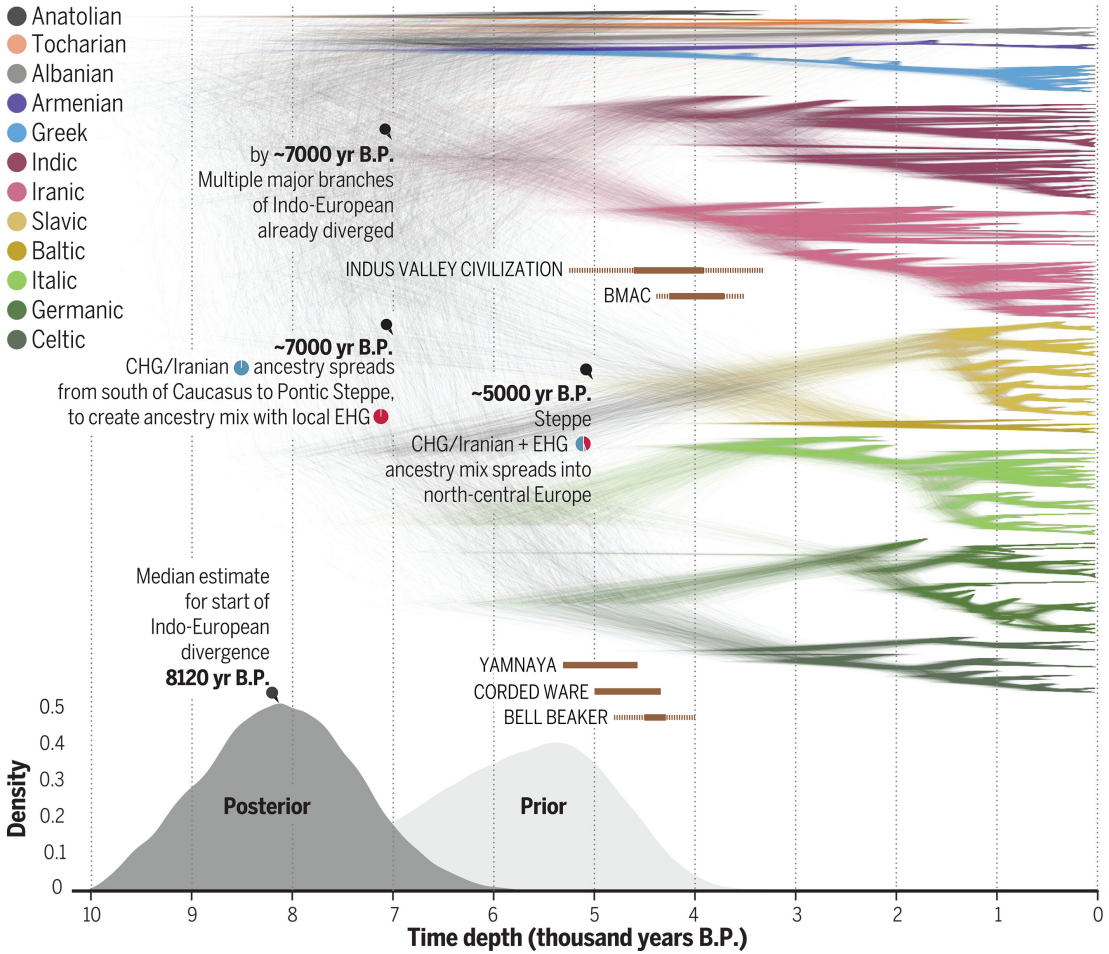
Summarising output – consensus trees

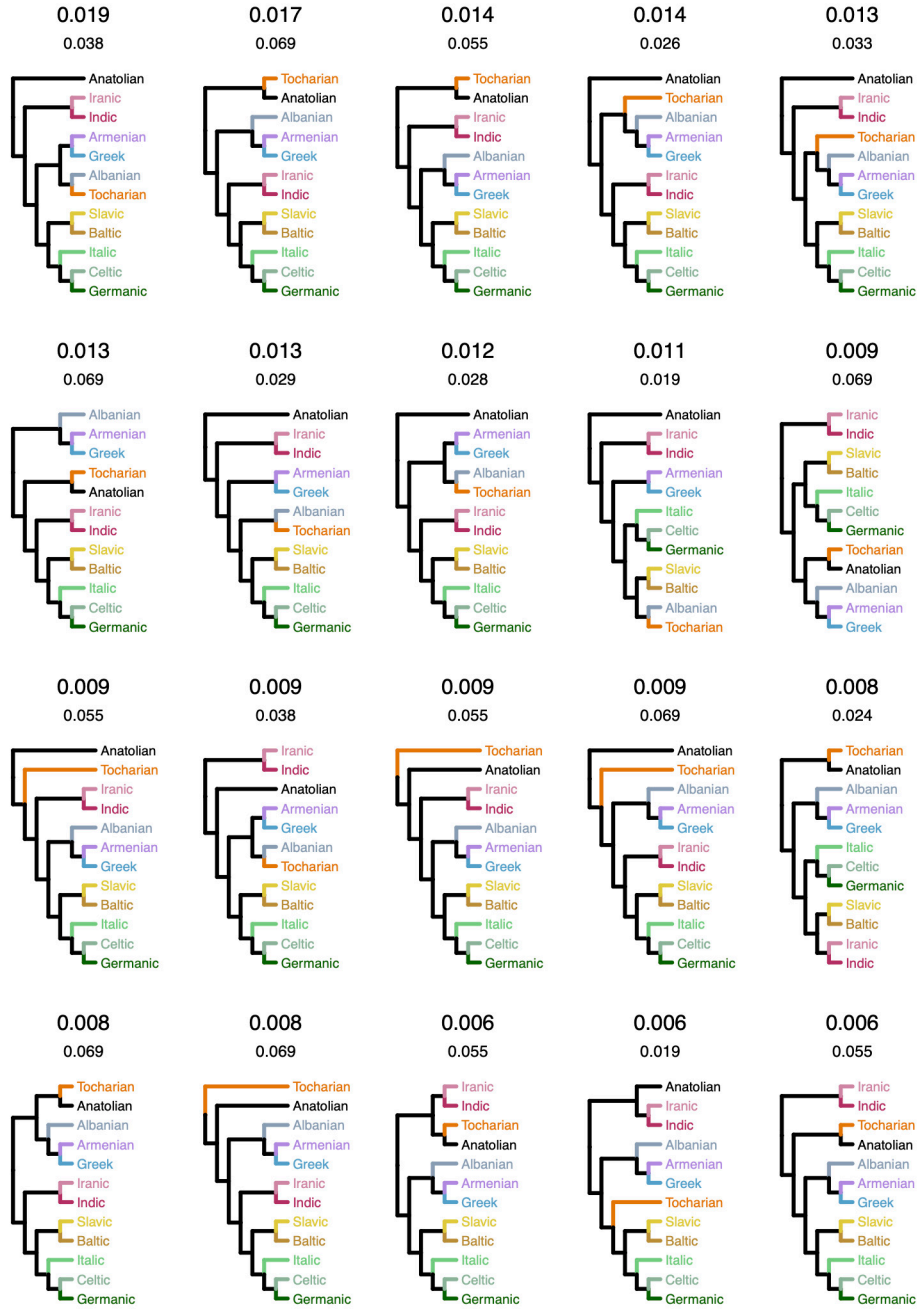


DANGER!
LINGUISTS!

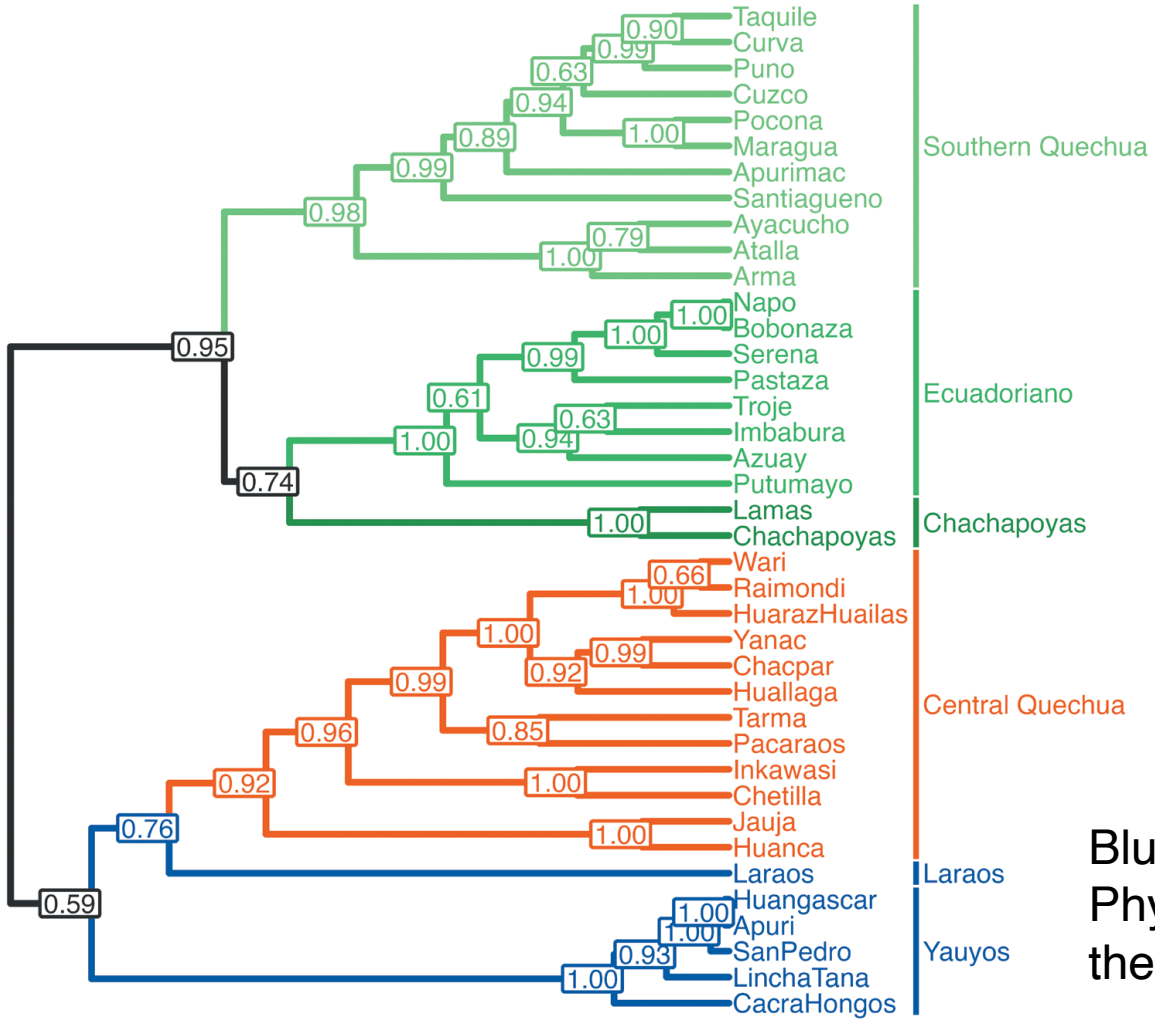
Summarising output trees

Densitree



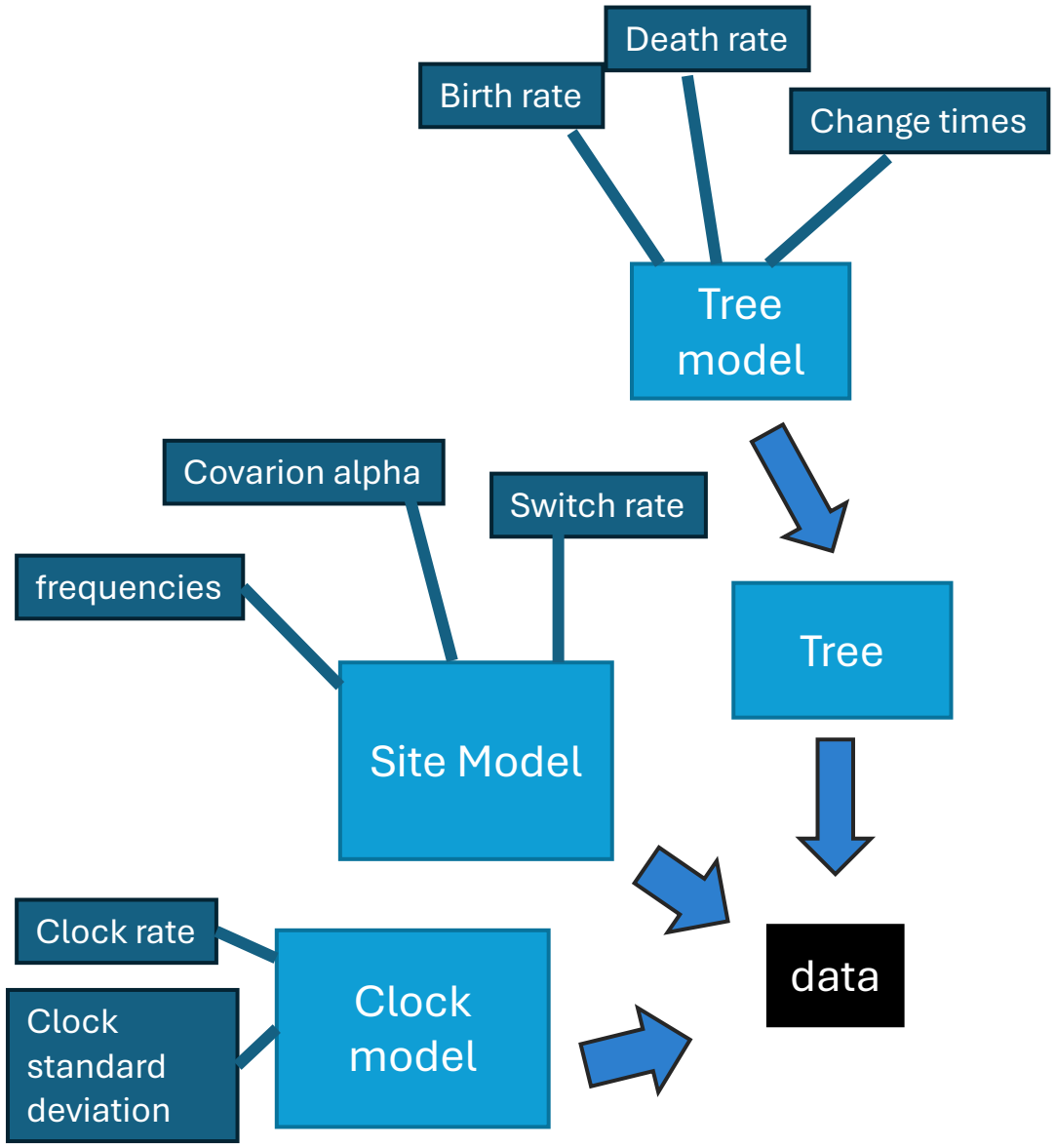


Morning tutorial: Phylogenetic analysis of the Quechua Language Family



Blum *et al.* (2023) A Phylolinguistic Classification of the Quechua Language Family

Model



Analysis

1. Likelihood: probability of the data given the tree and parameter values

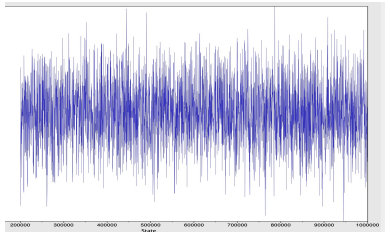
2. Probability of the tree

3. Prior probabilities of model parameters

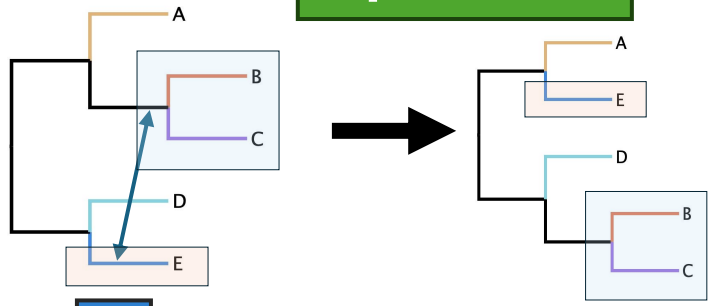
Posterior probability

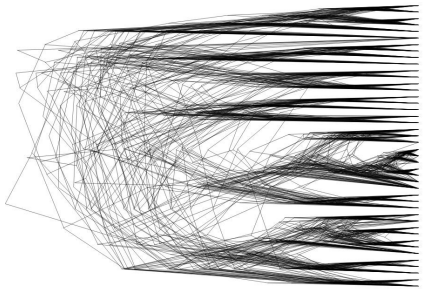
$$P(H|D) = \frac{\overbrace{P(D|H)}^{\text{likelihood}} * \overbrace{P(H)}^{\text{prior}}}{\underbrace{P(D)}_{\text{Model evidence}}}$$

Markov Chain Monte Carlo



Operators





Sample of trees and parameter values

Workflow

CLDF

Data
.nex file

BEAUti

Analysis
file
(.xml)

BEAST2

Trace file
(.log)

Tree file
(.trees)

TreeAnnotator

Summary
tree
(.tree)

Figtree/
Icytree

Figure
(.pdf)