Using BEAST2

www.github.com/richelbilderbeek/Science

GitHub

What, why, mastery

- What: Bayesian Evolutionary Analysis by Sampling Trees
- Why: Bayesian phylogenetic inference includes uncertainty
- Mastery: understand Bayesian analysis and BEAST2 priors

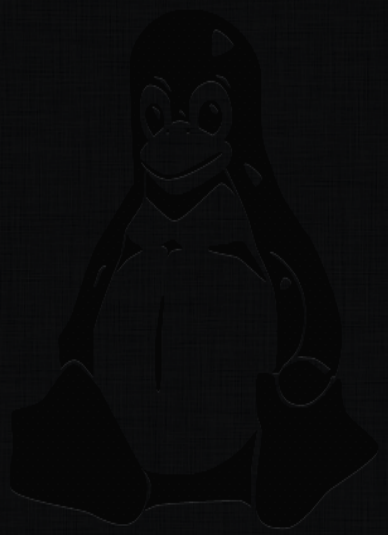
Bouckaert

Drummond

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C-H., Xie, D., Suchard, MA., Rambaut, A., & Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology, 10(4), e1003537. doi:10.1371/journal.pcbi.1003537

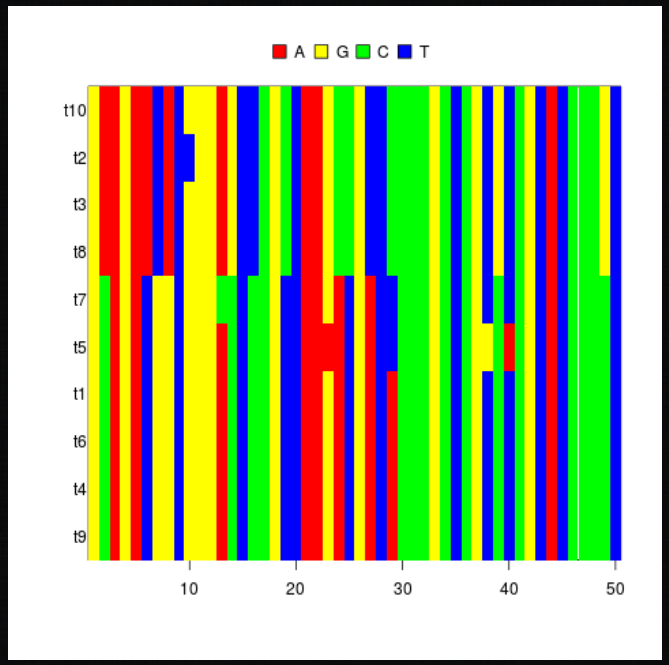
Parts

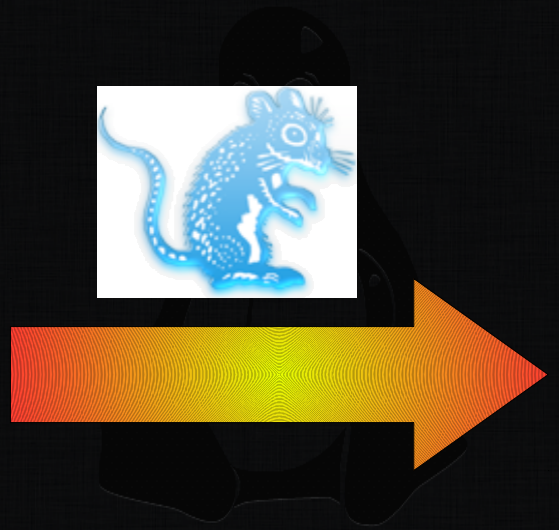
- 1-slide usage
- Workflow
- Details

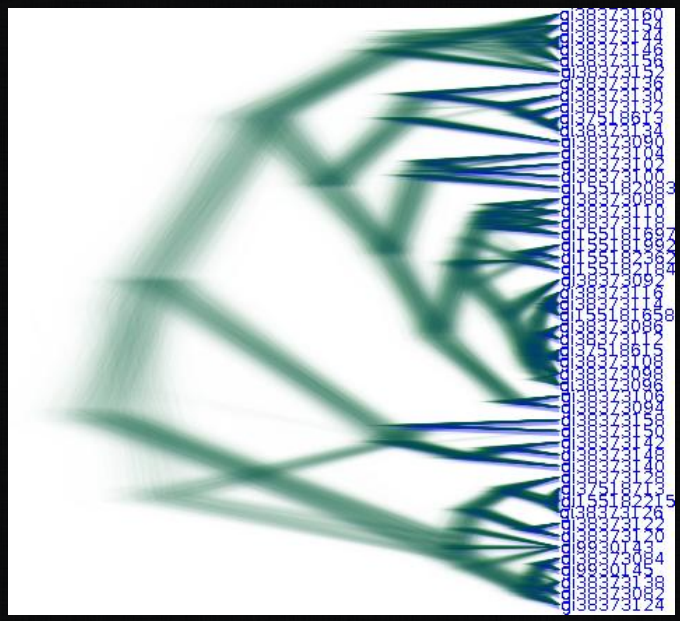


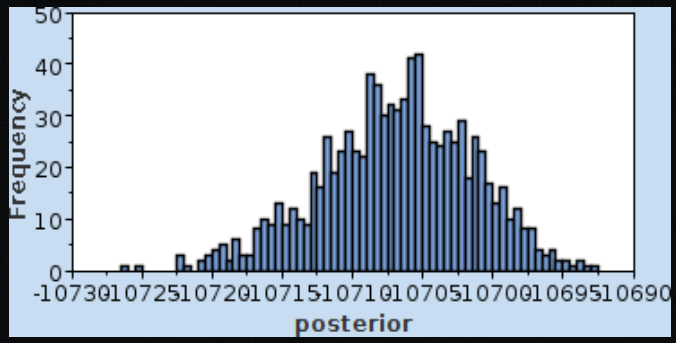
Usage

Alignment







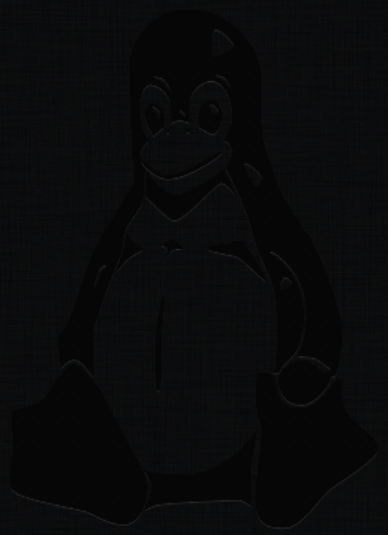


I think
it works
like this

Priors

Posterior

Demo



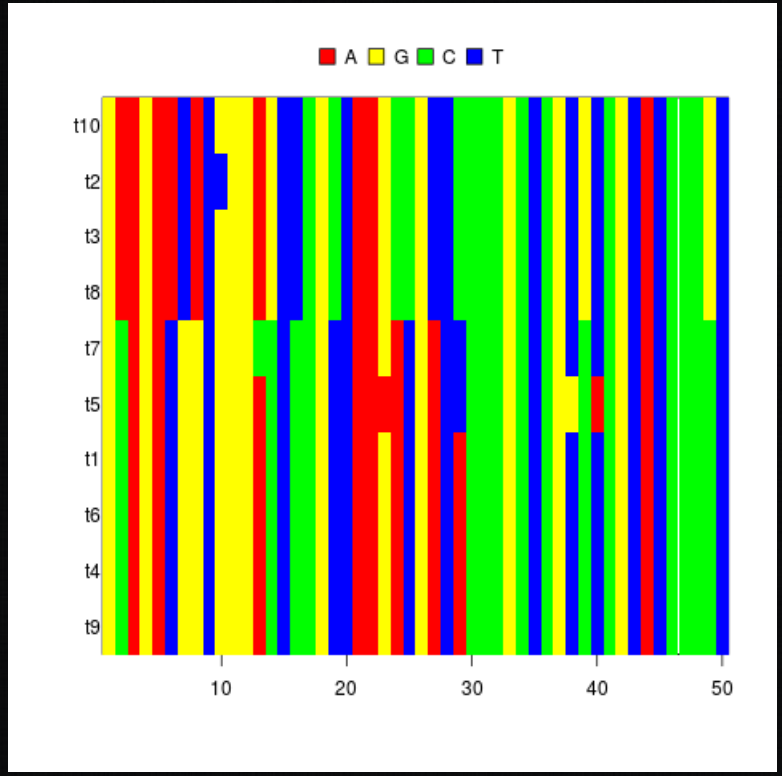
Workflow

- Create alignment
- Choose priors
- Run BEAST2
- Analyse results

Create alignments

- Many tools, e.g. ClustalW, MUSCLE, etc

```
library(ape)
library(phangorn)
alignment_phydat <- simSeq(
  rcoal(10),
  l = 50, # sequence_length,
  rate = 0.1 #mutation_rate
)
alignment_dnabin <- as.DNAbin(
  alignment_phydat
)
image(alignment_dnabin)
```

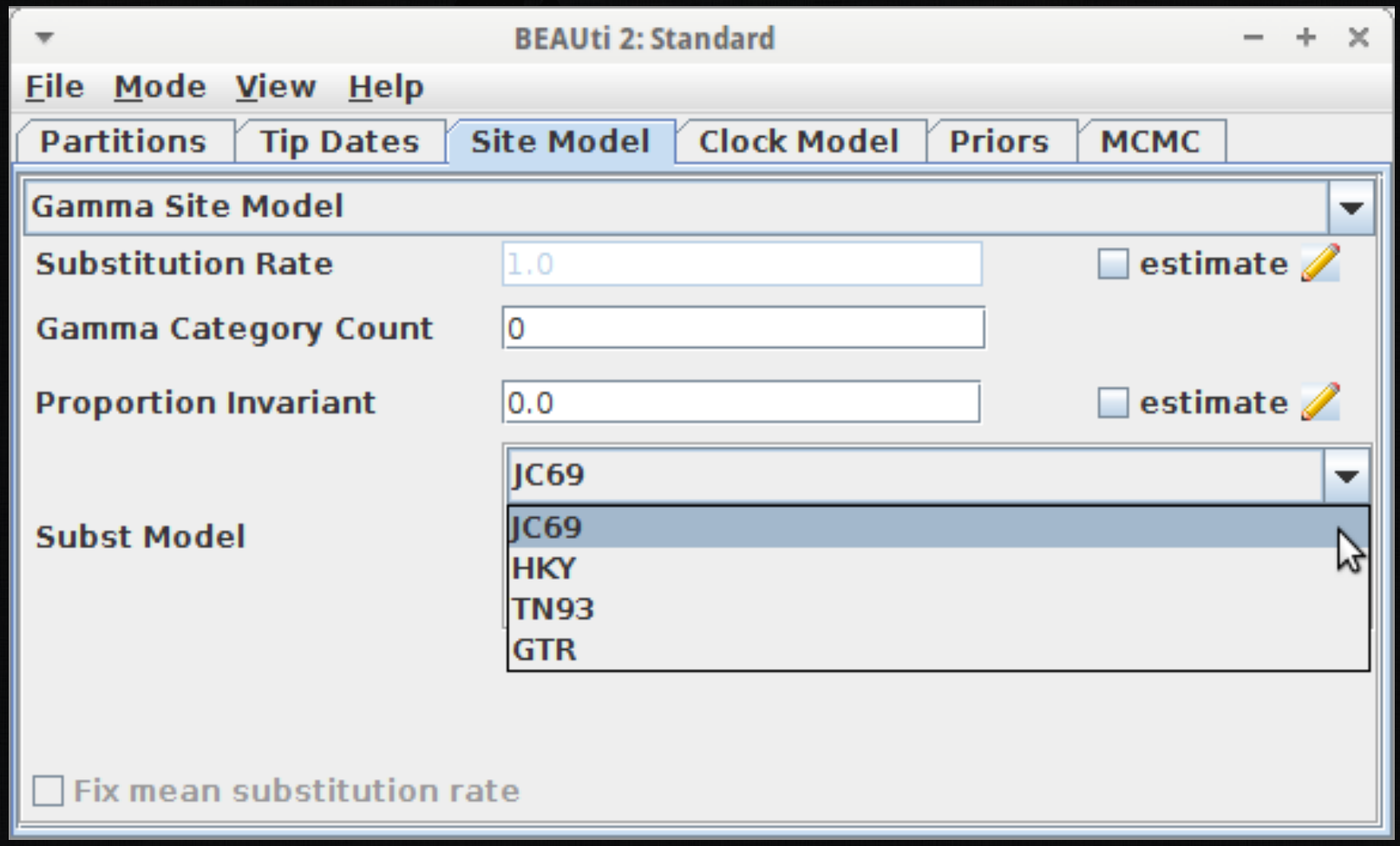


Choose priors

- Priors
  - Site model
  - Clock model
  - Speciation model
- BEAST2 can be extended with modules

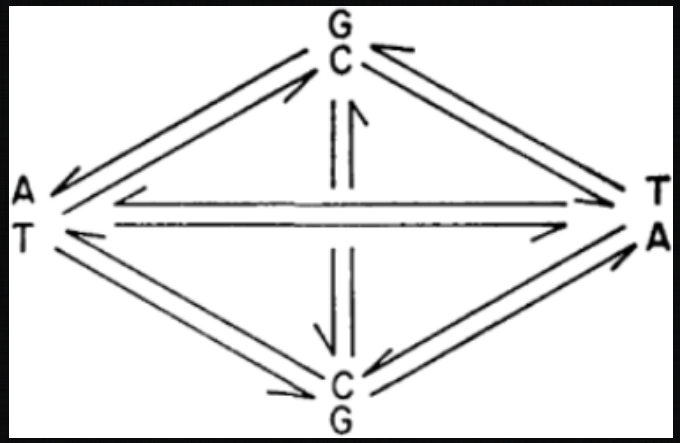
BEAST2 substitution models

- JC69
- HKY
- TN93
- GTR



JC69

- Equal base frequencies
- Equal mutation rates



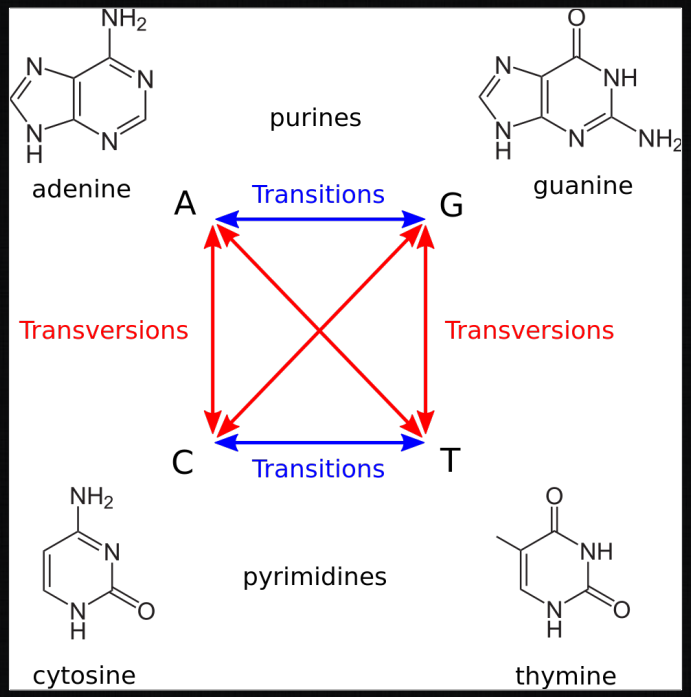
	A	C	G	T
A	*	1	1	1
C	1	*	1	1
G	1	1	*	1
T	1	1	1	*

Jukes, Thomas H., and Charles R. Cantor. "Evolution of protein molecules.", 1969

HKY

- Transitions ≠ transversions (κ)
- $\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$

"Transitions-transversions-v4"
by Petulda

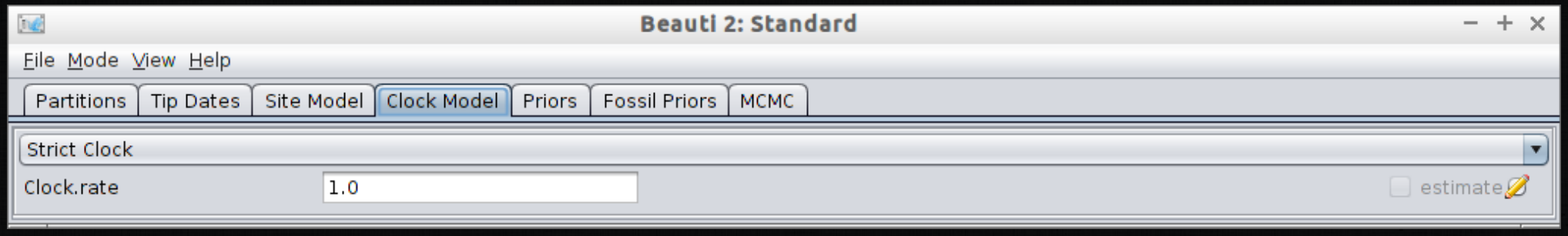


	A	C	G	T
A	*	π_C	$\pi_G K$	π_T
C	π_A	*	π_G	$\pi_T K$
G	$\pi_A K$	π_C	*	π_T
T	π_A	$\pi_C K$	π_G	*

Hasegawa M, Kishino H, Yano T (1985). "Dating of human-ape splitting by a molecular clock of mitochondrial DNA"

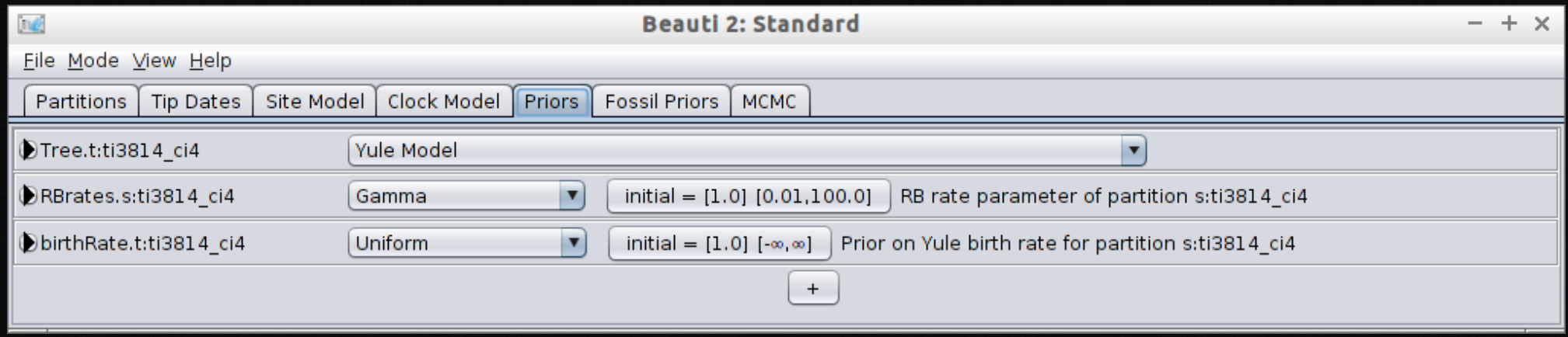
Clock model

- Strict clock
  - Intra-species data
- Relaxed clock



Species model

- Yule: constant-rate birth model
- Birth-Death model: Yule + death
- Coalescent constant population: small sample of big population



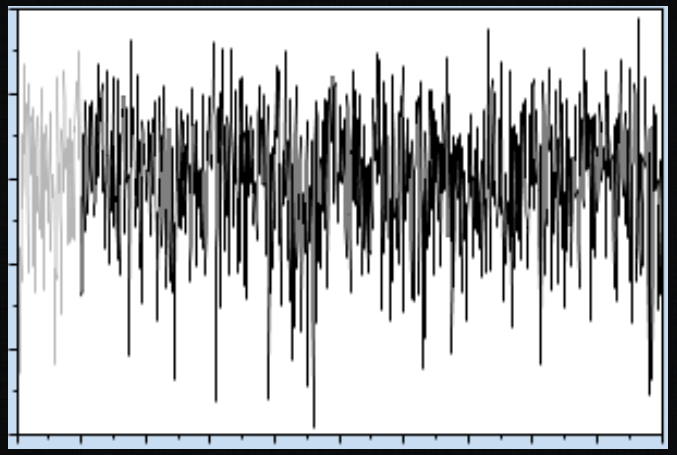
Run BEAST2

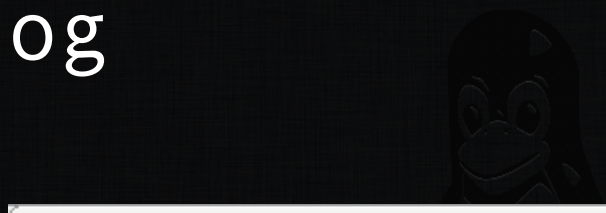
- On local computer
- Peregrine cluster

```
#!/bin/bash
#SBATCH --time=0:10:00
#SBATCH --nodes=1
#SBATCH --ntasks-per-node=1
#SBATCH --ntasks=1
#SBATCH --mem=100000
#SBATCH --job-name=example
module load Beast
beast my_parameters.xml
```

Check results

- ESS > 200
- Trace log





Tracer

File Edit Analysis Help

Trace Files:

Trace File	States	Burn-In

+ −

Traces:

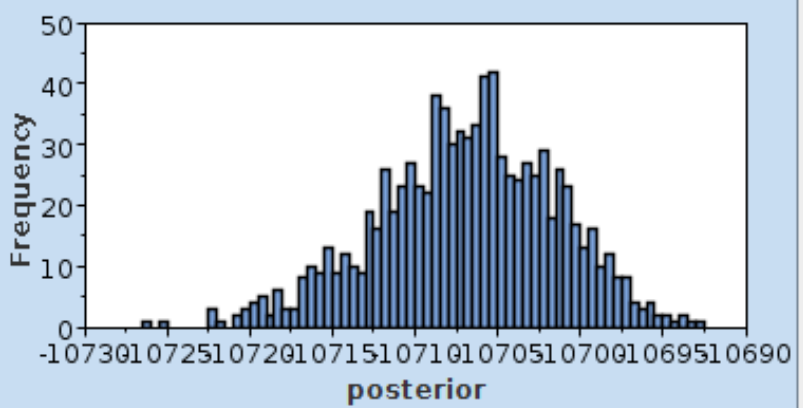
Statistic	Mean	ESS	...
posterior	-1070...	559	R
likelihood	-1089...	570	R
prior	189.438	797	R
treeLikelihood	-1089...	570	R
TreeHeight	3.768...	481	R
BirthDeath	196.346	797	R
birthRate2	98.644	199	R
relativeDeathRate2	0.378	227	R

Data type: (R)eal (I)nt (C)at

Joint-Marginal Trace

Estimates Marginal Prob Distribution

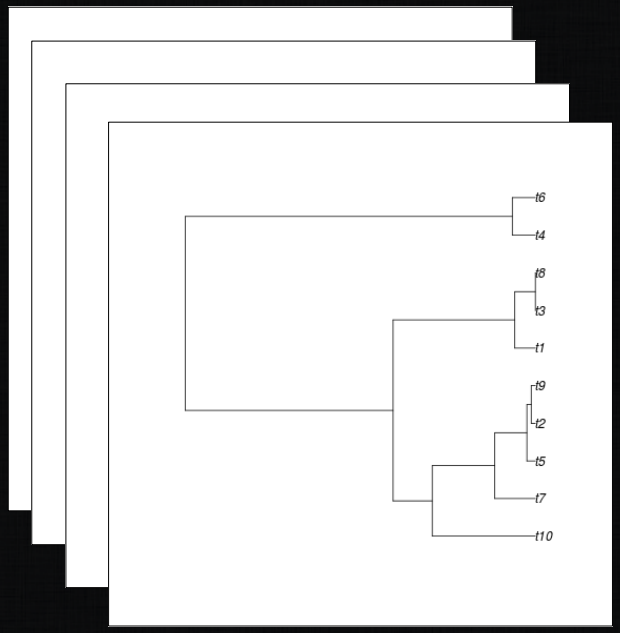
Summary Statistic



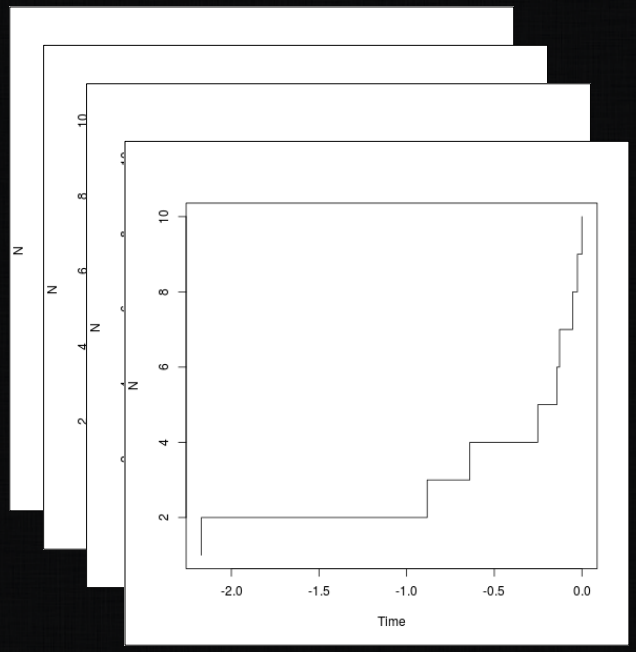
Axes... Bins: 50

Analysis

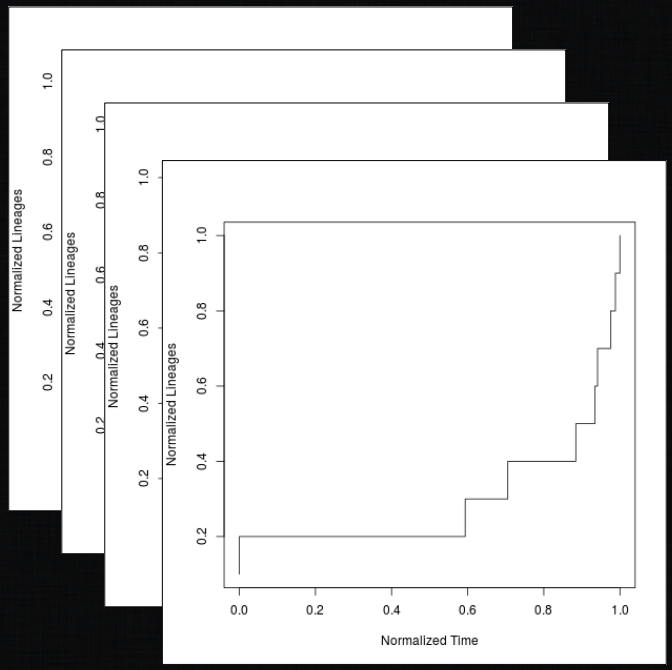
Posterior phylogenies



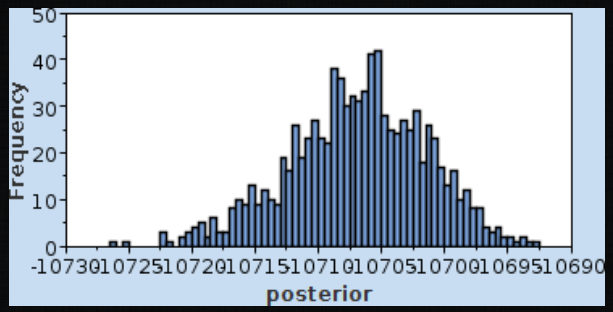
Lineages-Through-Time (LTT)



Normalized LTT (nLTT)



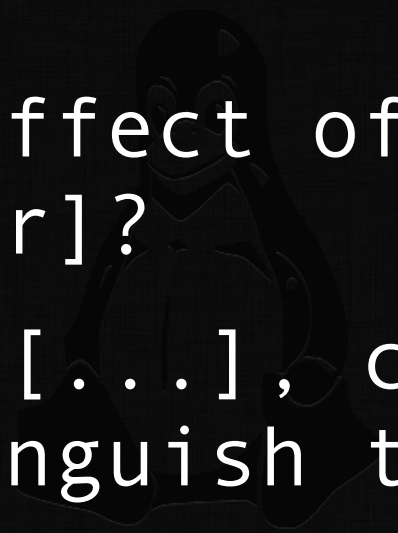
Posterior parameters



Mean
Variance

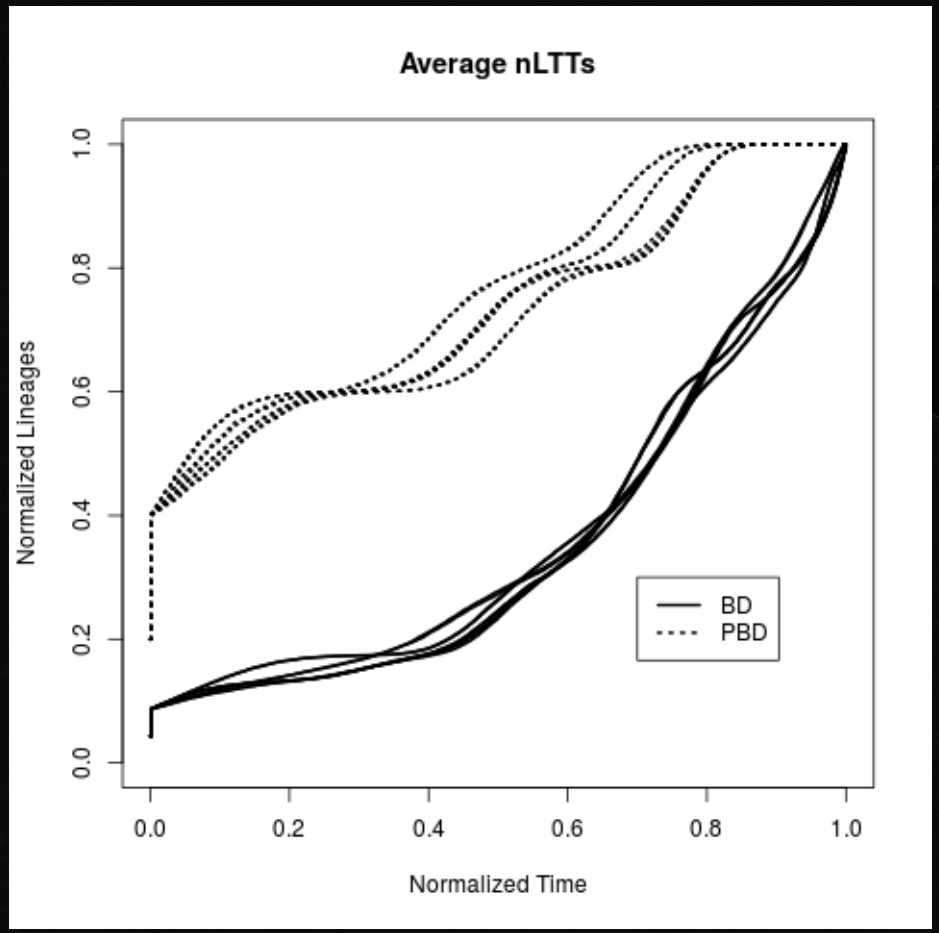
Research questions

- What is the phylogeny of this clade?
- What is the effect of [using a specific prior]?
- If nature is [...], can we recover/distinguish this?



Use cases

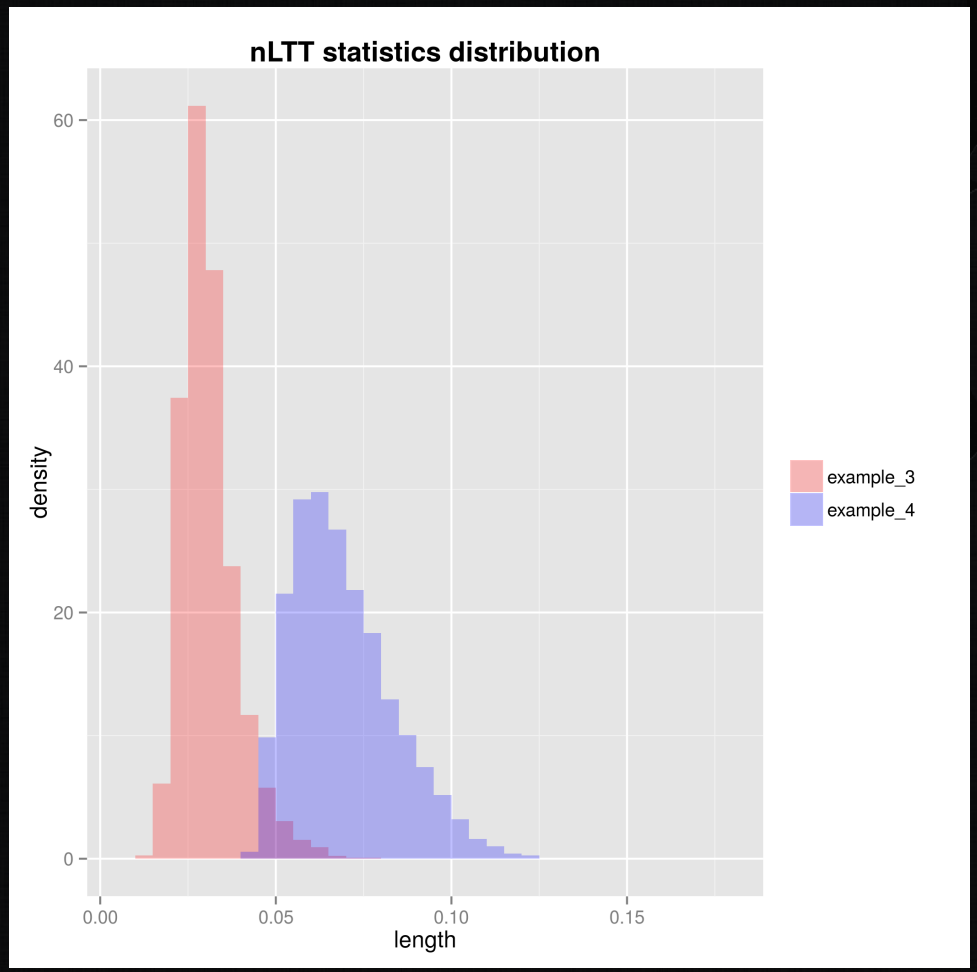
- If speciation takes time, can we observe this in BEAST2?



If speciation takes time (PBD), we can observe a difference in the shape of the phylogeny

Use cases

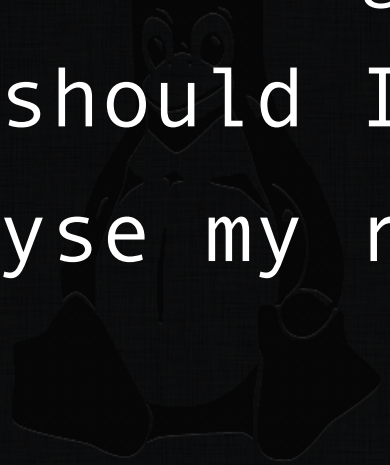
- If speciation takes time, how big will the estimation errors be?



If speciation takes time (example_4), BEAST2 will have a threefold error in recovering a known phylogeny

Details

- What is this Bayesian inference?
- How does the MCMC algorithm work?
- Which priors should I choose?
- How do I analyse my results?



Bayesian inference

$$P(H|E) = P(E|H) * P(H) / P(E)$$

posterior distribution
- How likely is the hypothesis, given the evidence?
- Our updated belief

likelihood
- How likely is the evidence, given the hypothesis?

prior distribution
- Our belief
- How likely is the prior true without evidence?

marginal likelihood
- Sum of weighted hypotheses

Example from [1]

- Elvis Presley had a twin brother who died at birth. What is the probability that Elvis had an identical twin? Assume 8% of all twins are monozygous.

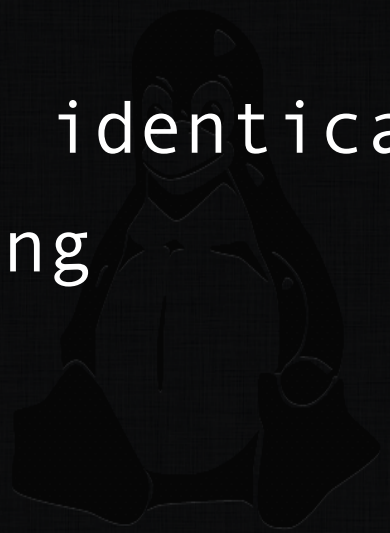
H1: Elvis had an identical twin

H2: Elvis had a fraternal twin

[1] MacKay, David JC. Information theory, inference and learning algorithms. 2003.

Bayesian inference

- P(H|E) = P(E|H) * P(H) / P(E)



- H: Elvis had an identical twin
- E: A male sibling

- P(E|H): 1.0
- P(H): 0.08
- P(E): (1.0 * 0.08) + (0.5 * 0.92) = 0.54
- P(H|E): 0.148

Bayesian inference

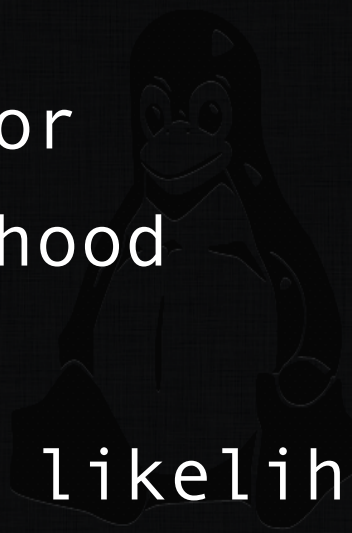
- P(H|E) = P(E|H) * P(H) / P(E)

- H: Elvis had a fraternal twin
- E: A male sibling

- P(E|H): 0.5
- P(H): 0.92
- P(E): (1.0 * 0.08) + (0.5 * 0.92) = 0.54
- P(H|E): 0.852

Bayesian inference

- f(θ|D) = Pr(D|θ) * f(θ) / Pr(D)

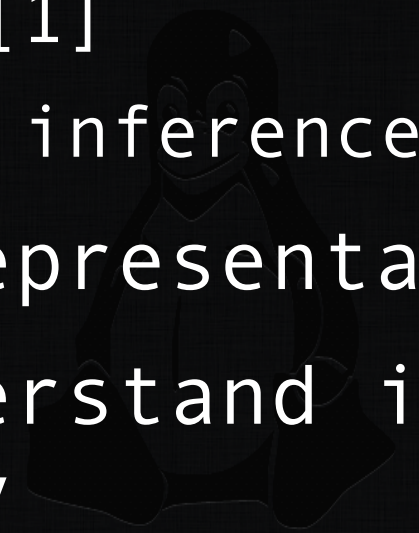


- f(θ|D): posterior
- Pr(D|θ): likelihood
- f(θ): prior
- Pr(D): marginal likelihood

- D: the sequence data
- θ: model parameters + tree
  - For example: death rate, birth rate + tree

MCMC algorithm

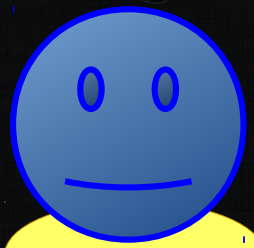
- What: Markov-Chain Monte-Carlo
  - King Markov [1]
  - Phylogenetic inference
- Why: get a representative sample
- Mastery: understand its stochasticity



[1] Example from Richard McElreath, 'Statistical Rethinking'

King Markov and the Chain Islands

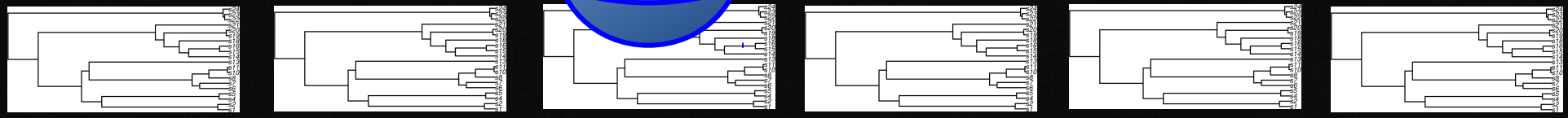
Visit your subjects representatively



N 1 2 3 4 5 6

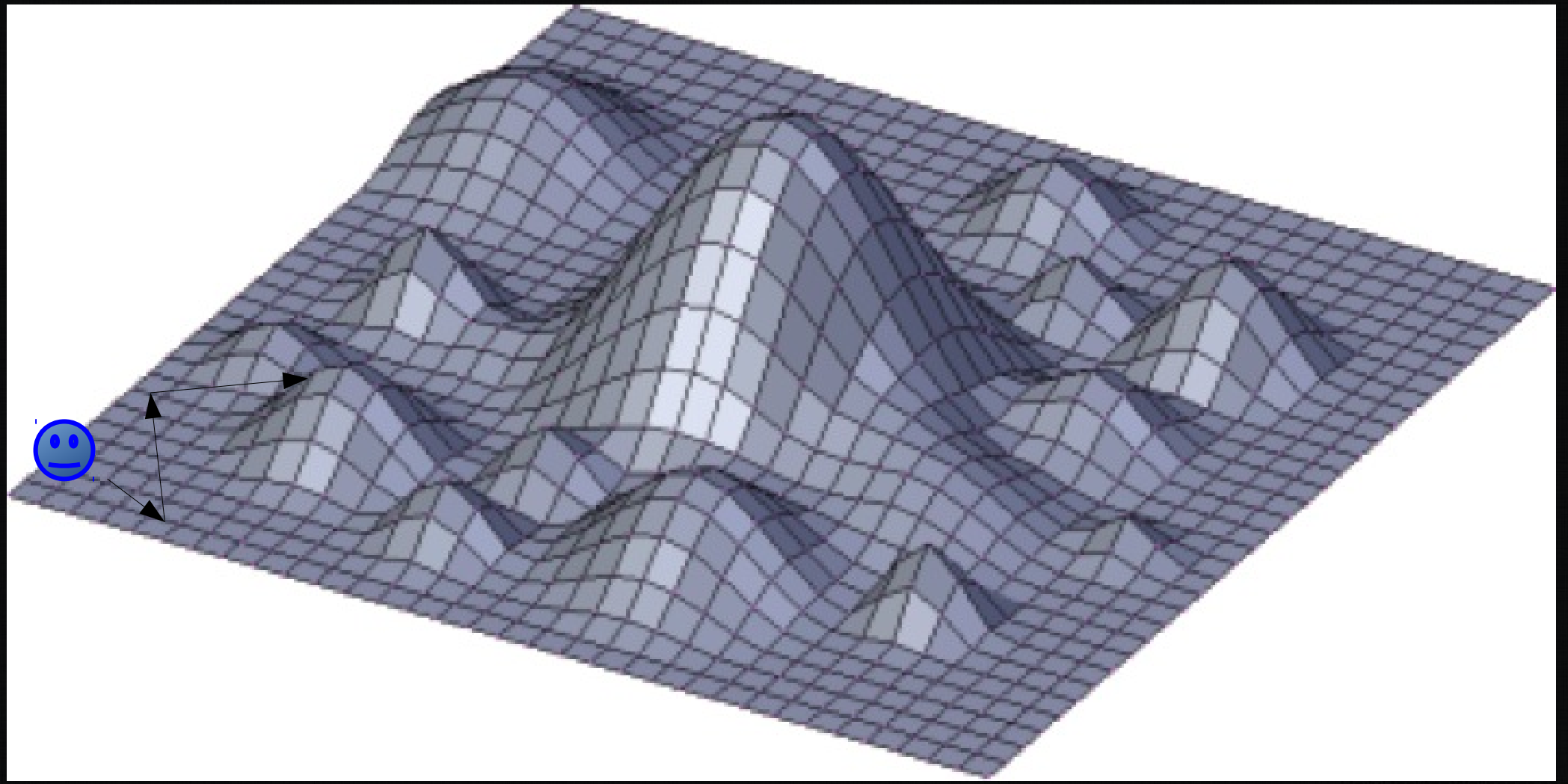
Phylogenies

Sample
the parameters
(including tree)
representatively



$Pr(D_1|\theta)$ $Pr(D_2|\theta)$ $Pr(D_3|\theta)$ $Pr(D_4|\theta)$ $Pr(D_5|\theta)$ $Pr(D_6|\theta)$

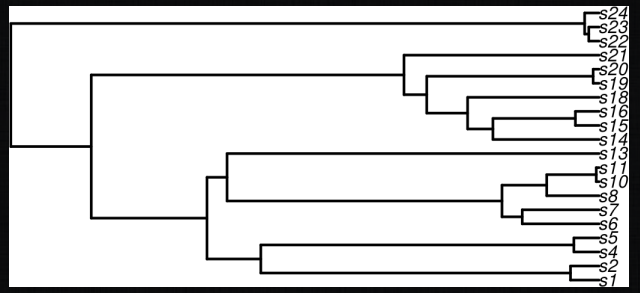
MCMC



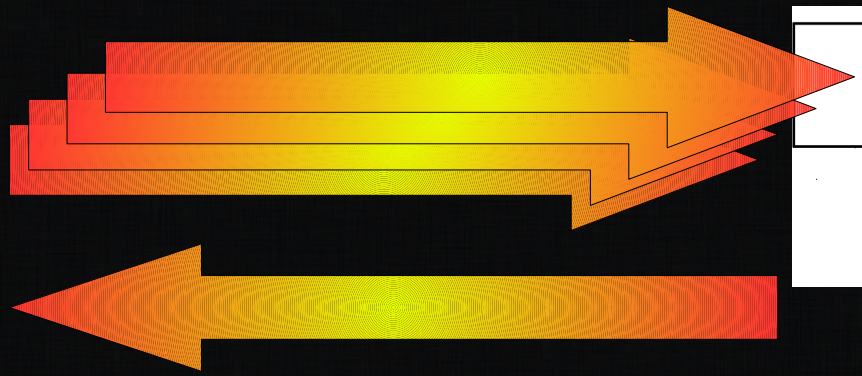
Chance of acceptance going from 1 to 2

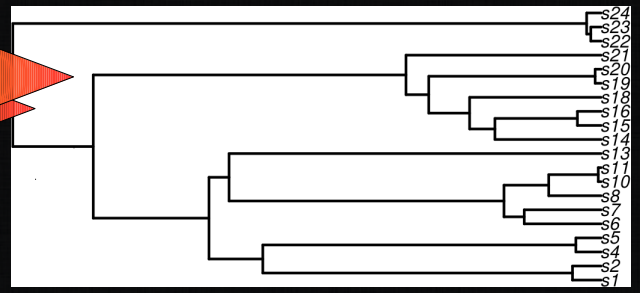
Metropolis-Hasting

$$\alpha = \min\left[1, \frac{f(\theta_2|D)}{f(\theta_1|D)} \frac{q(1\rightarrow 2)}{q(2\rightarrow 1)}\right]$$



$Pr(D_1|\theta)$

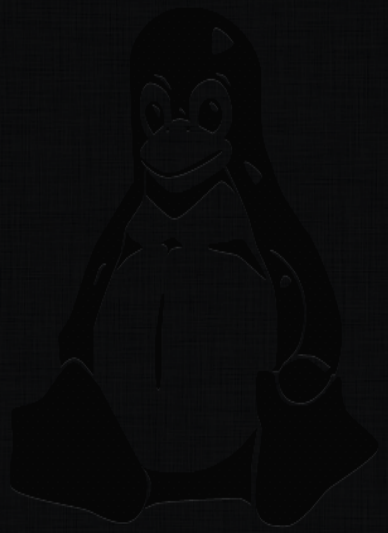




$Pr(D_2|\theta)$

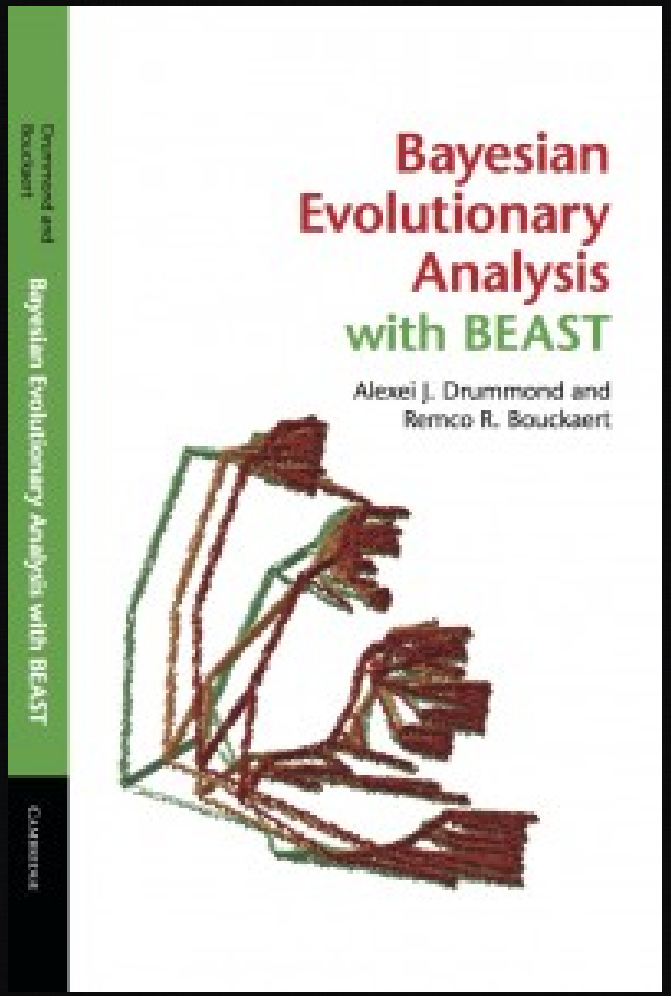
More complex things

- Operators



Which priors should I choose?

- The BEAST book has many tips
- The same as you expect/know the system to be
- When in doubt: start as simple as possible, then increase complexity



Drummond & Bouckaert, 2015

How do I analyse my results?

- All BEAST2 output is plain-text, so you can use your favorite language/tool

```
library(ape)
source("olli_rBEAST/R/fun.beast2output.R")

p <- beast2out.read.trees(
  trees_filename
)
first_phylogeny <- p[1]
plot(first_phylogeny)
```