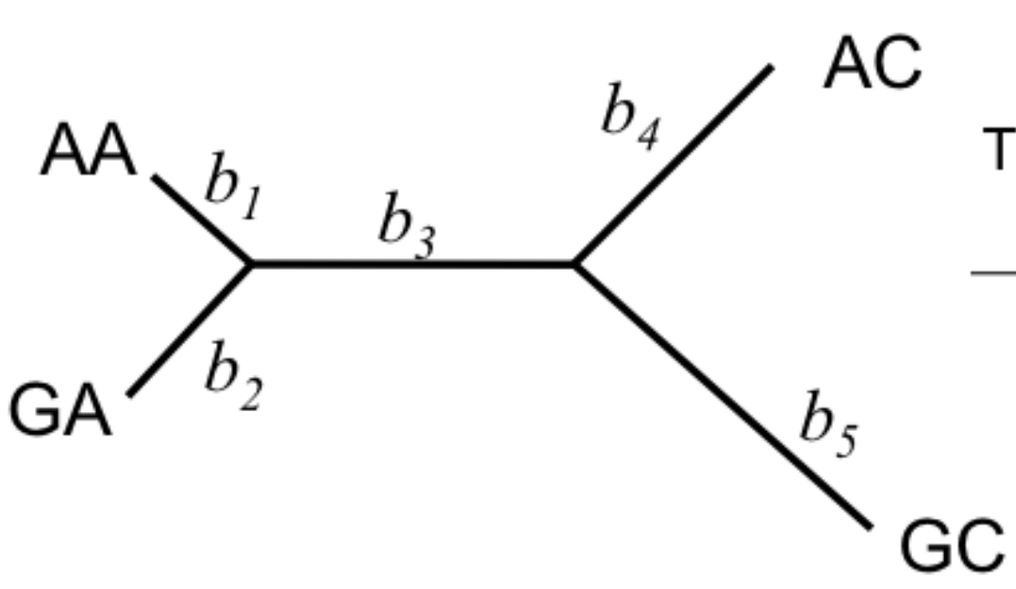
Molecular clocks, calibrations and tree ages

Slides by Alexei Drummond

Molecular clocks and calibrations

The molecular clock constraint

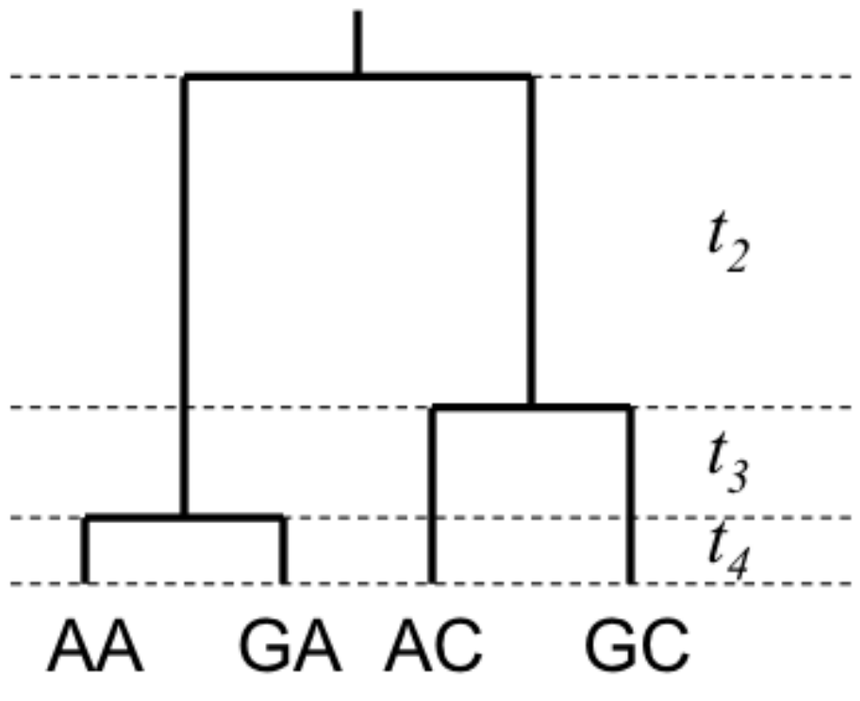
T



2n–3 branch lengths

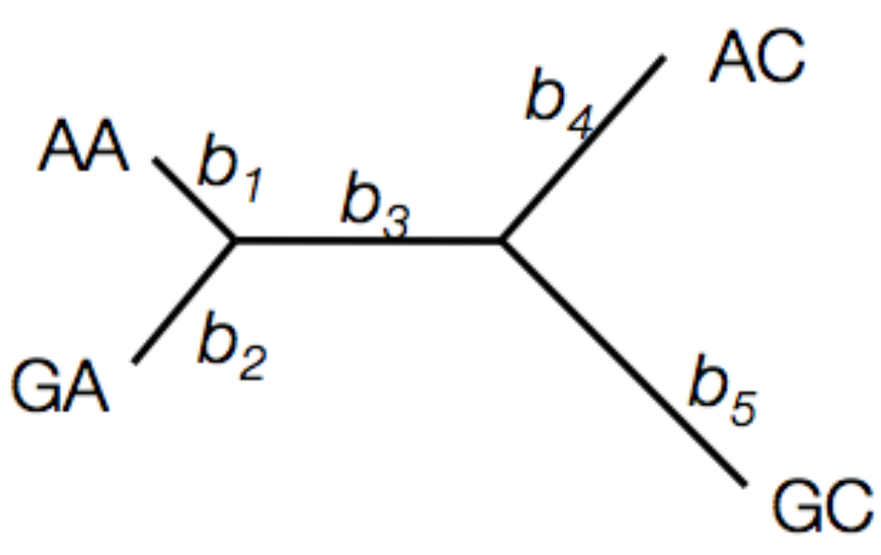
The "molecular clock" constraint

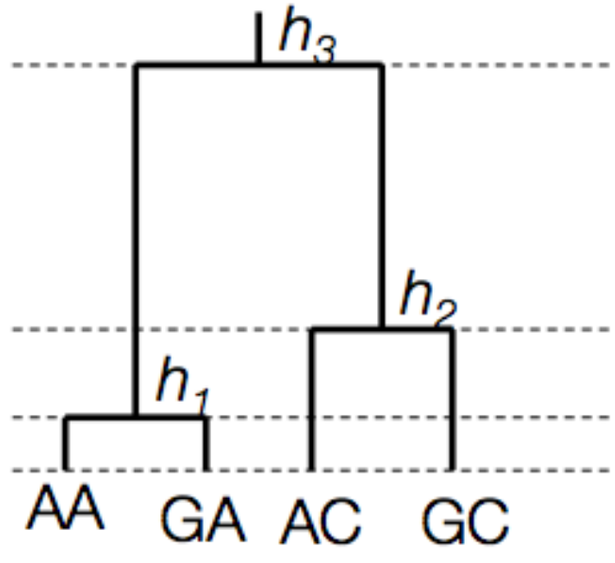
g



n–1 waiting times

Model assumptions





- Product of rate and time (branch length) is independent and identically distributed among branches.
- The root of the tree could be anywhere with equal probability.
- Topology implies nothing about individual branch lengths.

- Rate of evolution is the same on all branches.
- The root of the tree is equidistant from all tips.
- Topology constrains branch lengths (e.g. two branches in a cherry must be of equal length)

Calibration via a global molecular clock

Basic model: (Tree in expected substitutions per site)

$$p(g, \theta|D) \propto \Pr\{D|g\}p(g|\theta)p(\theta)$$

Fix (i.e. condition on) the global rate to μ:

$$p(g, \theta|D) \propto \Pr\{D|\mu \times g\}p(g|\theta)p(\theta)$$

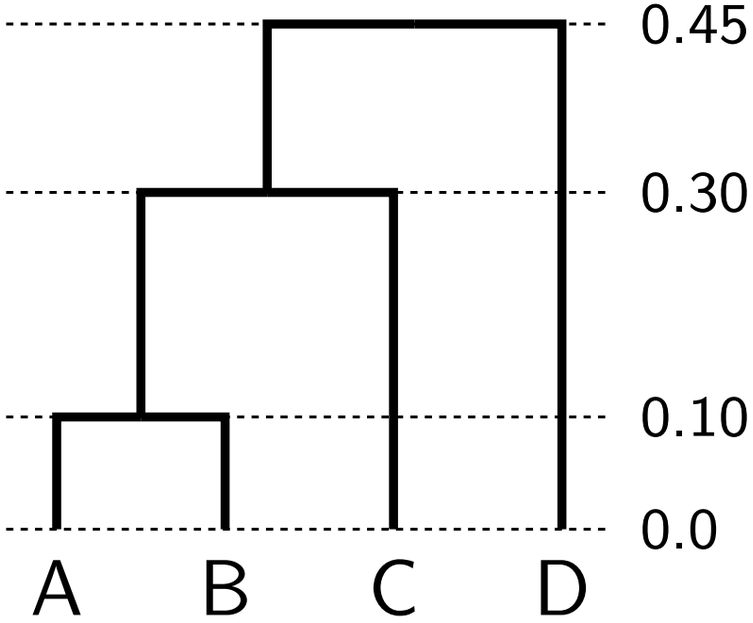
Estimate the global rate:

$$p(g, \mu, \theta|D) \propto \Pr\{D|\mu \times g\}p(g|\theta)p(\theta)p(\mu)$$

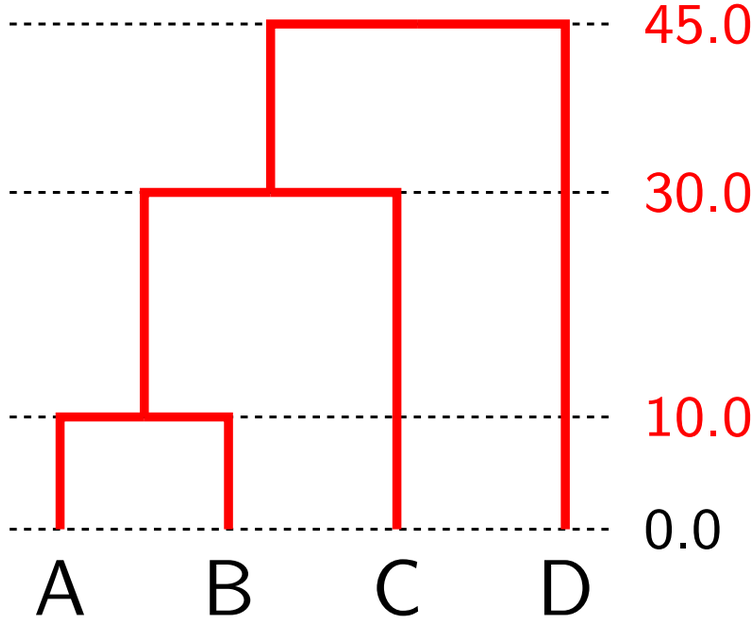
In the models above the parameters related to the details of the substitution process (Q) have been suppressed for simplicity.

Genetic distance = rate x time

$$T = \mu \times g$$



= 0.01 ×

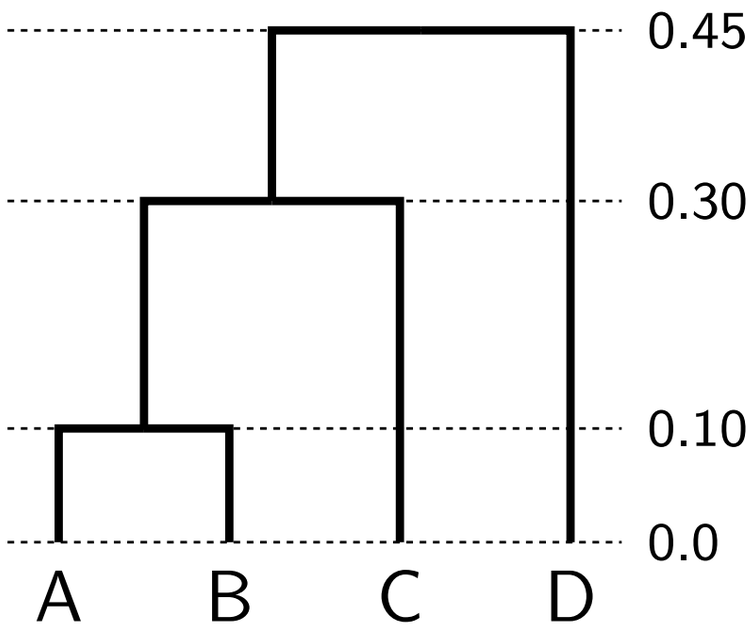


"substitution tree"

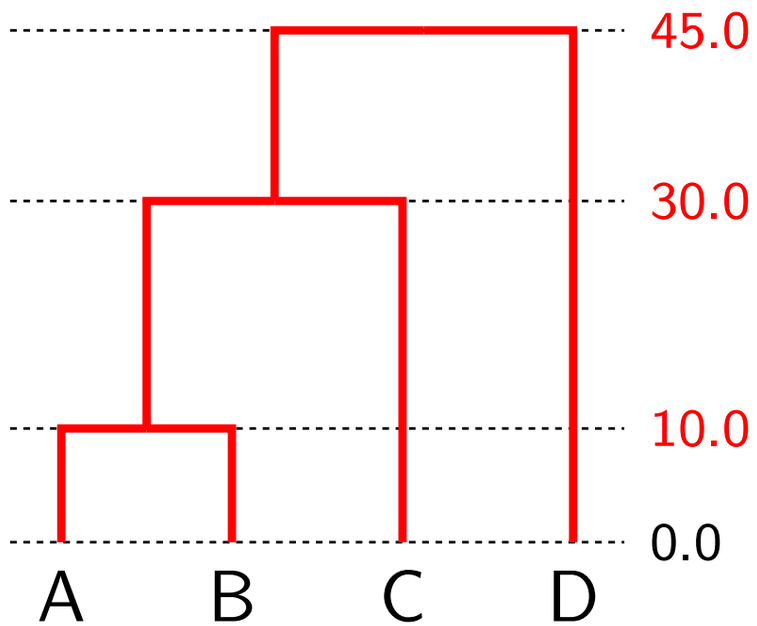
evolutionary rate
substitutions / site / unit time

time tree

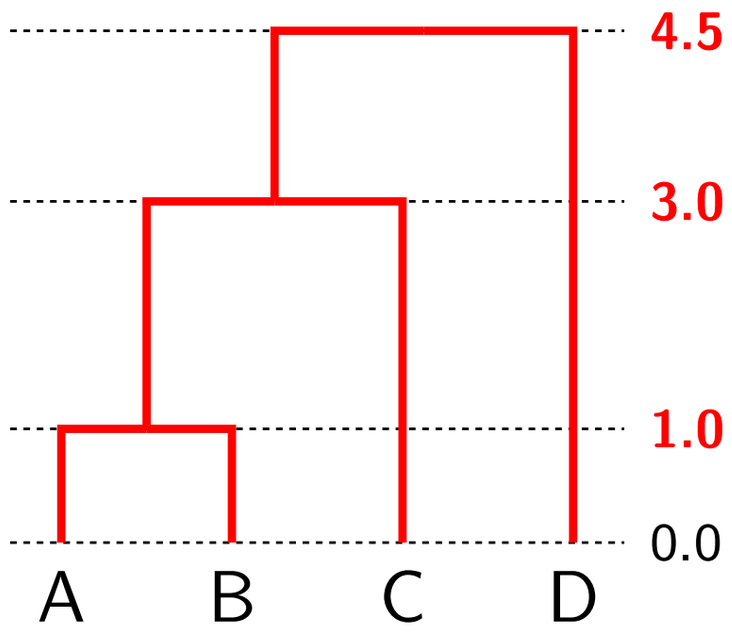
Non-identifiability of rates and times



$=$ 0.01 $\times$



$=$ **0.1** $\times$



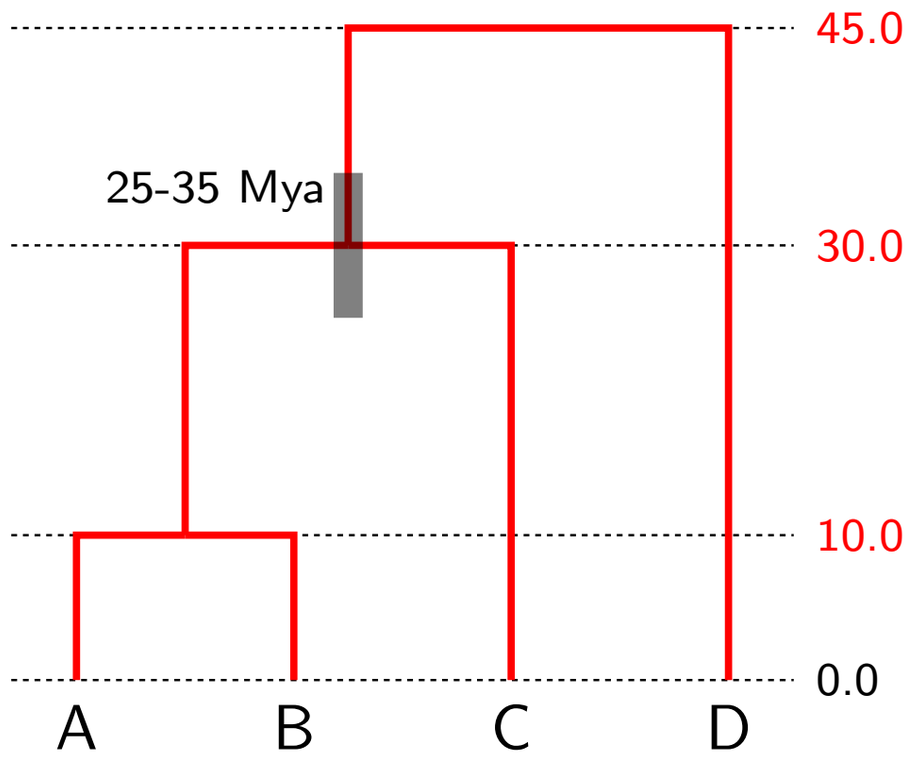
"substitution tree"

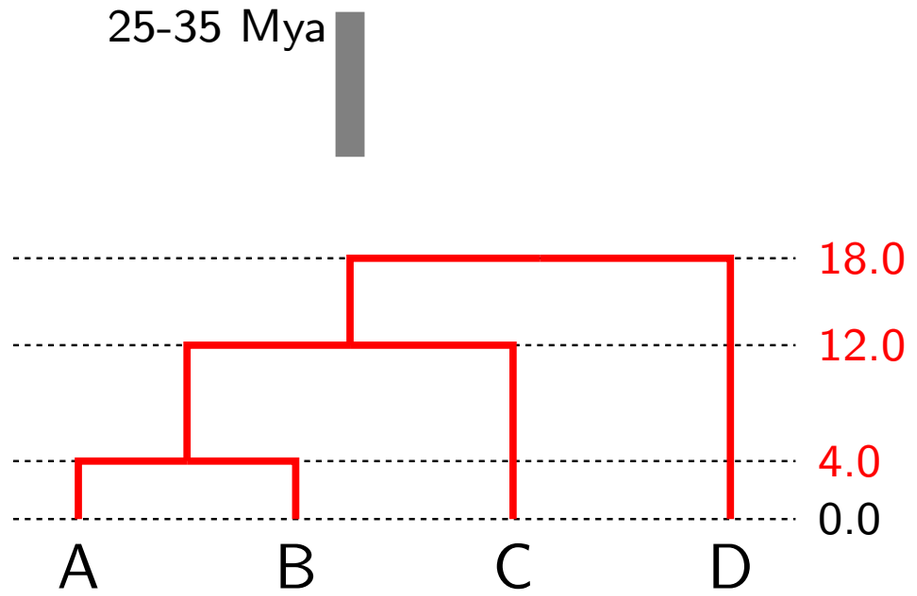
evolutionary rate
substitutions / site / unit time

time tree

Node calibration

Suppose fossil evidence shows the common ancestor of species A, B and C lived 25-35 Mya.





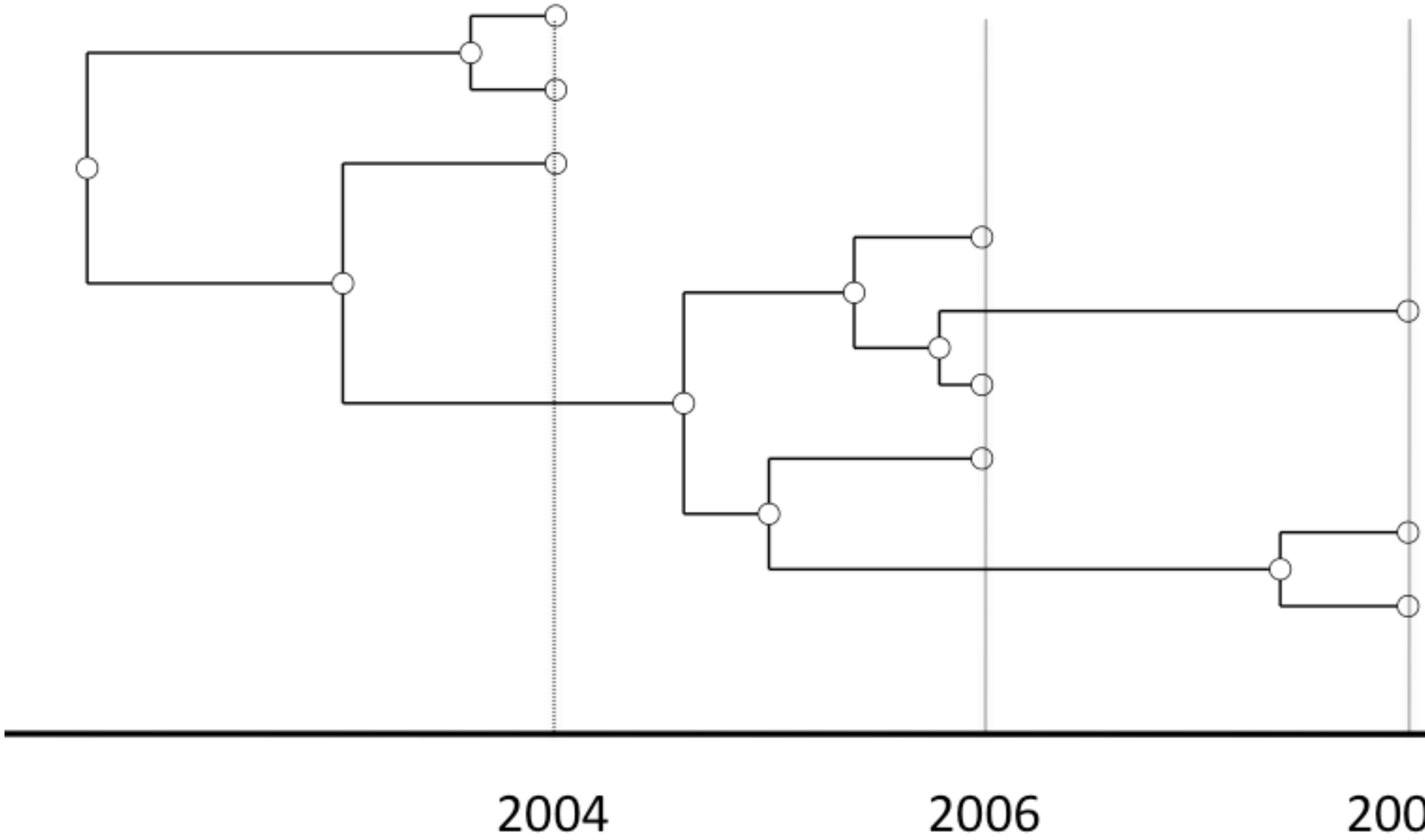
With a strict molecular clock, only the age (range) of a single node in the tree is needed in order to interpolate and extrapolate the ages of all other divergence times.

Once a known node age like this "calibrates" the tree, the genetic distances can be separated into an absolute rate and divergence times.

Tip calibration

Modelling phylogenetic data sampled through time

Drummond *et al* (2002)

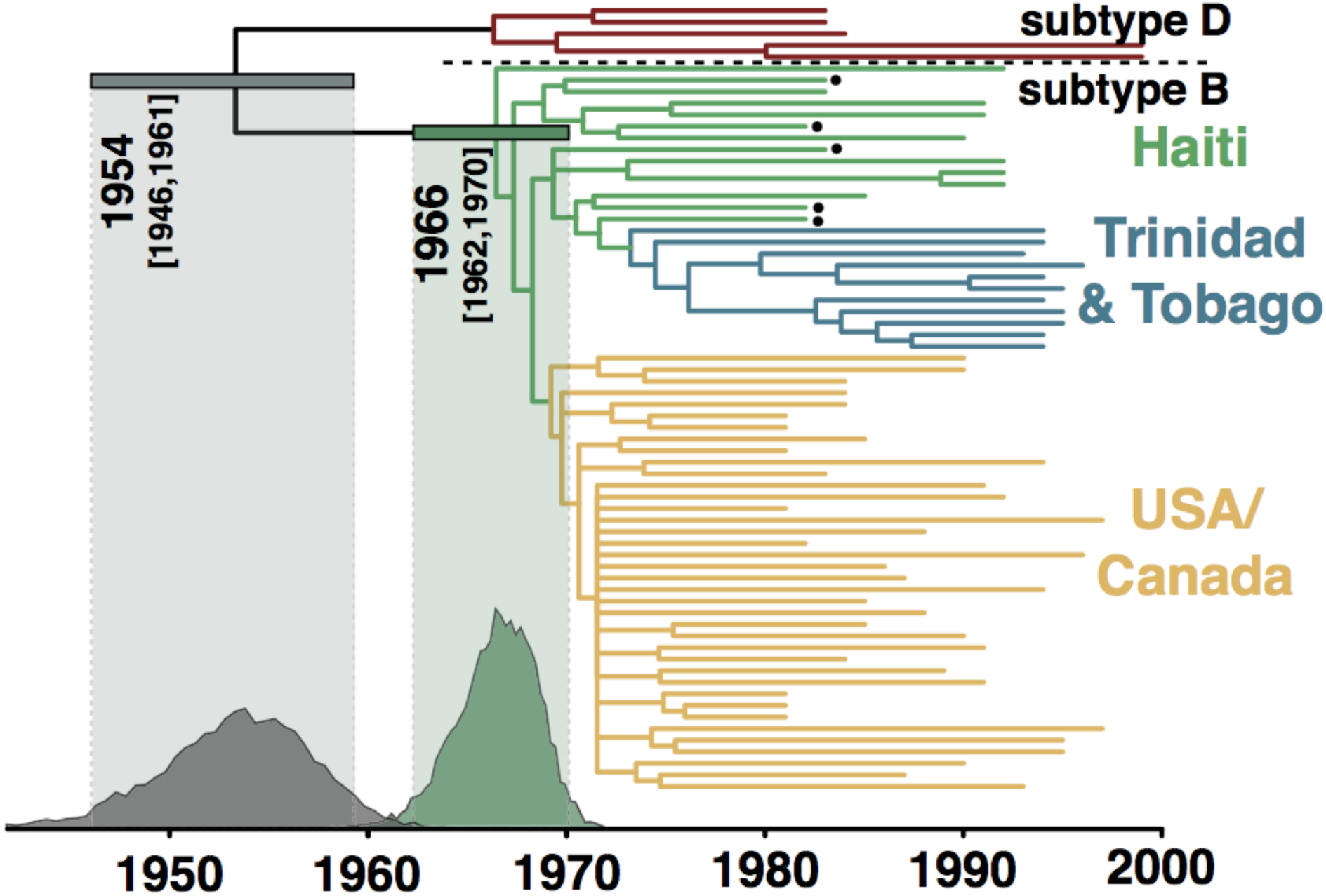


- Rapidly evolving microbes
- Ancient DNA
- Cancer
- Somatic evolution
- Languages
- *et cetera*

$$P(g, \mu, Q, \theta | D) \propto \Pr(D | g \times \mu, Q) P(g | \theta) P(\theta) P(Q) p(\mu)$$

A calibrated phylogenetic inference

Origin of the HIV epidemic in the Americas, Gilbert *et al* (2007)



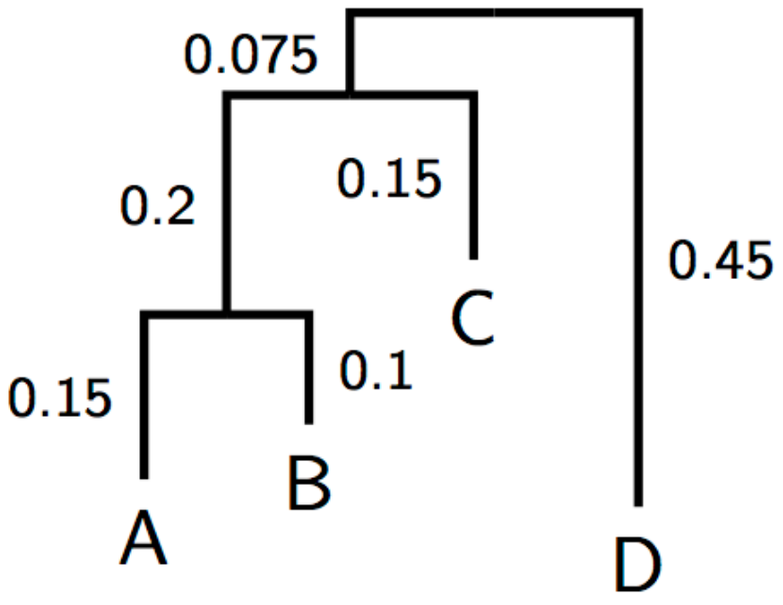
A phylogenetic reconstruction of samples of HIV-1 virus. Each tip represents a single infected individual from whom a blood sample has been taken.

Relaxed phylogenetics

Genetic distance = rate x time

Relaxed molecular clock

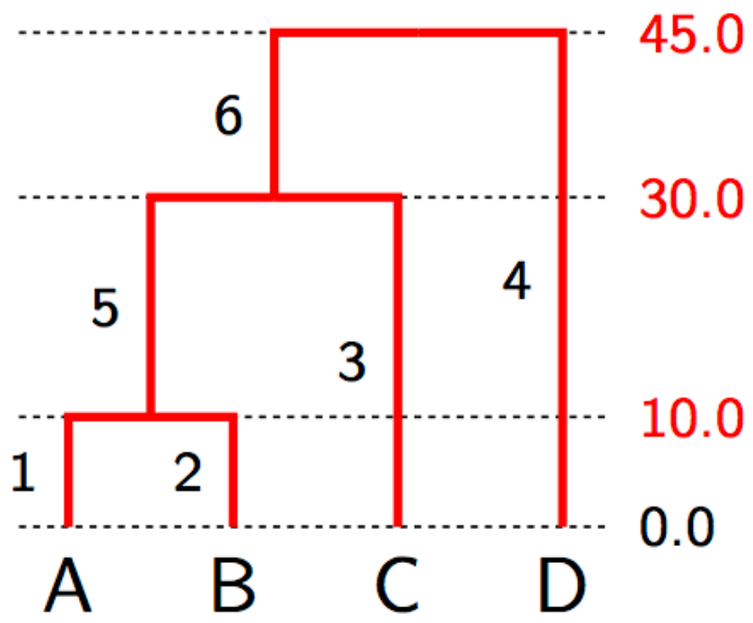
$$T = \vec{\mu} \star g$$



"substitution tree"

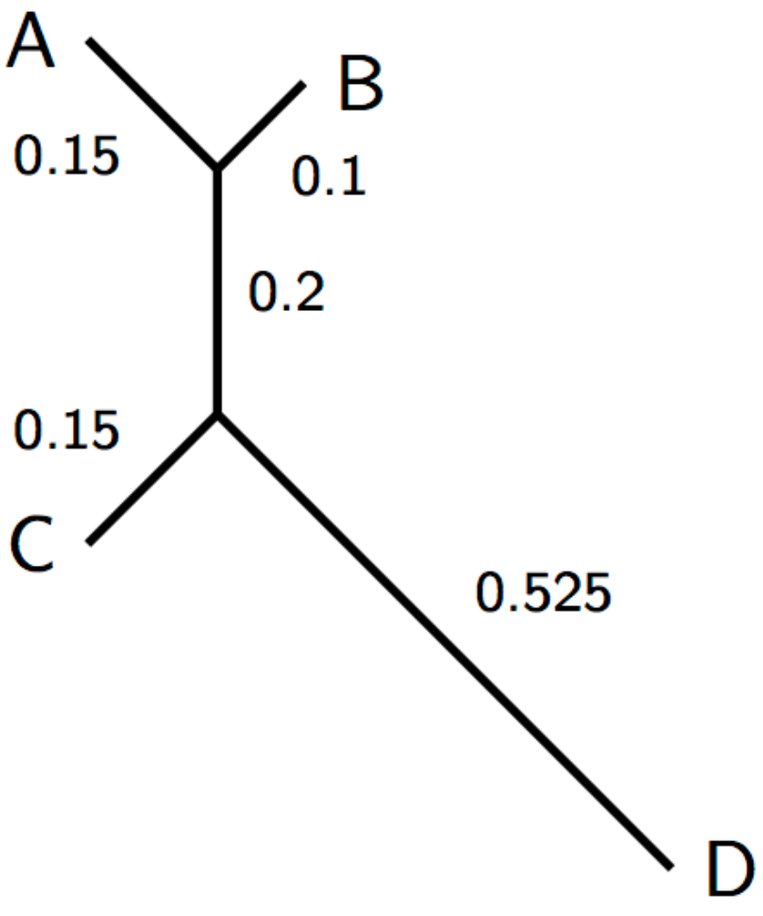
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time



time tree

Genetic distance = rate x time

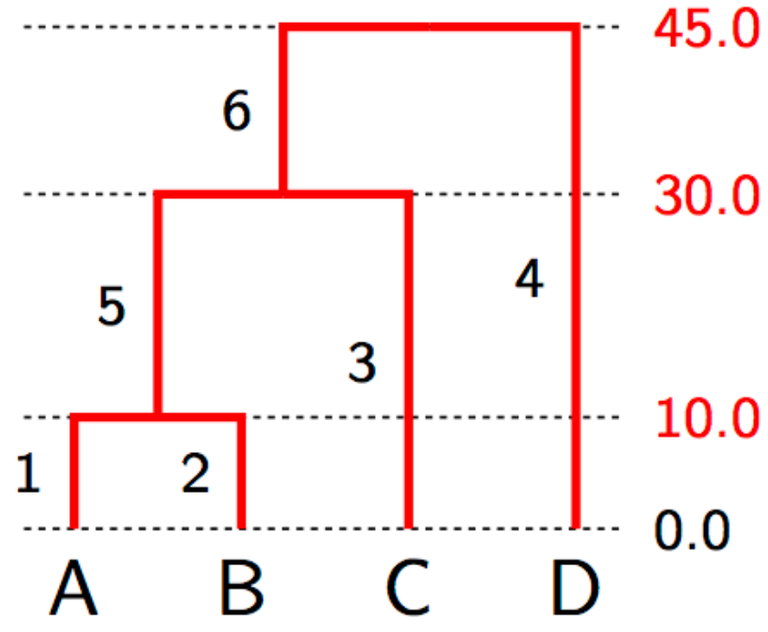


"substitution tree"

$$T = \vec{\mu} \star g$$

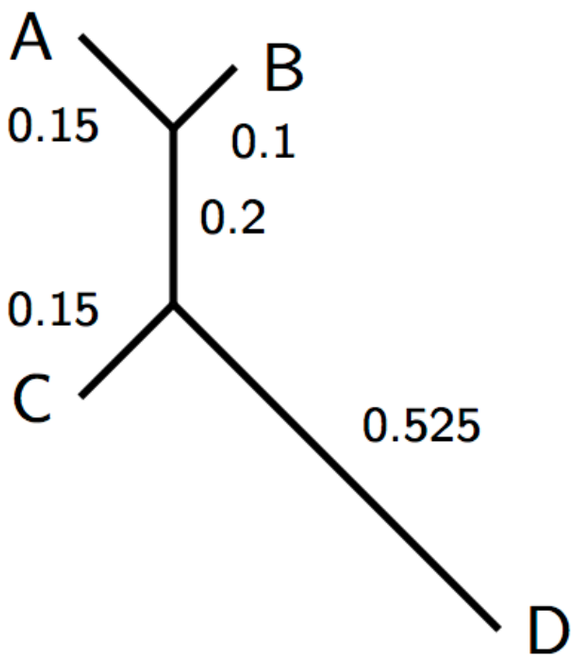
$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$

evolutionary rates
substitutions / site / unit time

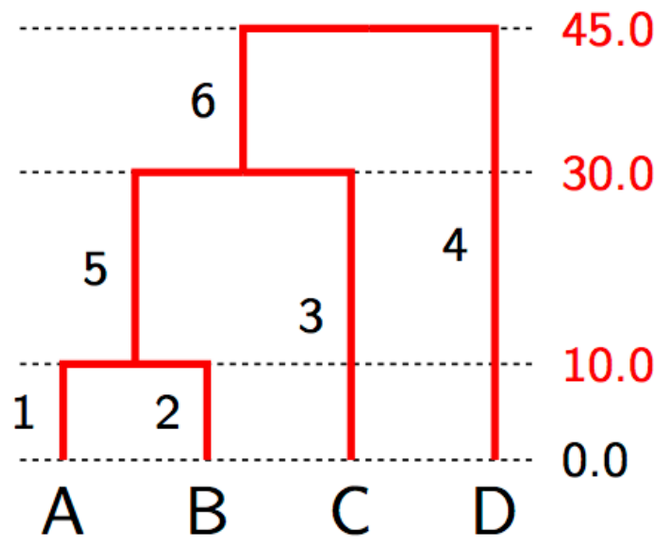


time tree

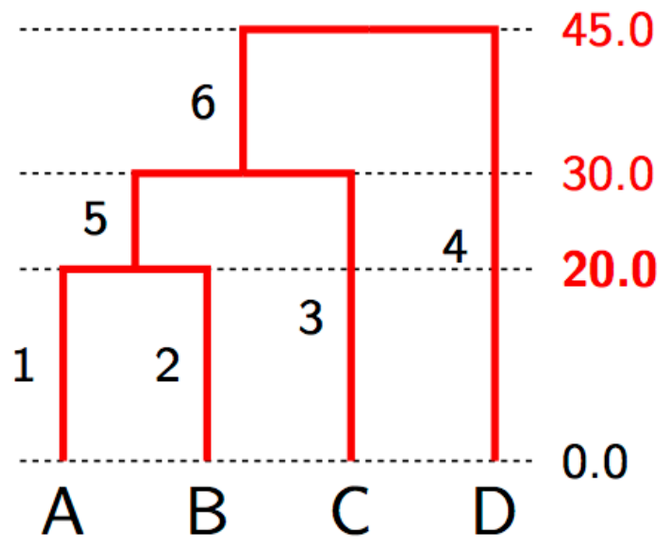
Non-identifiability of rates and times



$$= \begin{pmatrix} 0.015 \\ 0.01 \\ 0.005 \\ 0.01 \\ 0.01 \\ 0.005 \end{pmatrix} \star$$



$$= \begin{pmatrix} \mathbf{0.0075} \\ \mathbf{0.005} \\ 0.005 \\ 0.01 \\ \mathbf{0.02} \\ 0.005 \end{pmatrix} \star$$

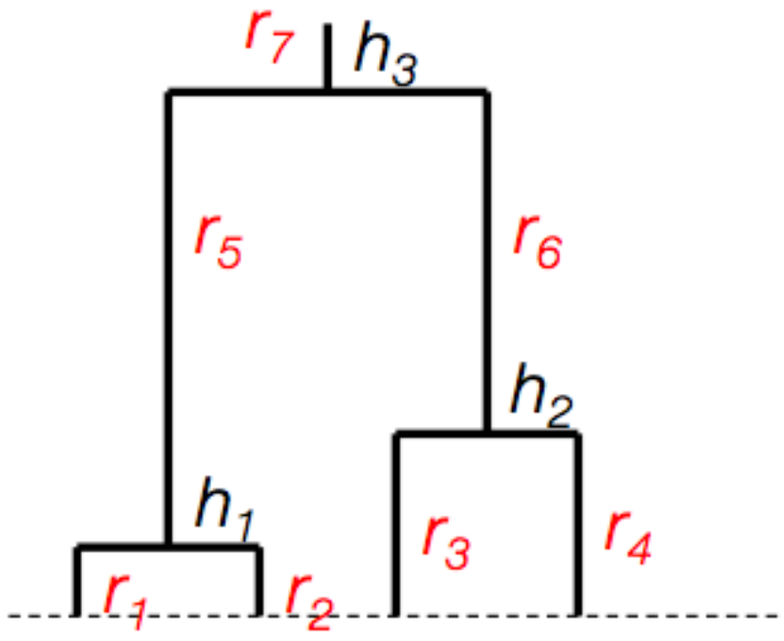


"substitution tree"

evolutionary rates
substitutions / site / unit time

time tree

Relaxing the molecular clock



In the field of divergence time estimation auto-correlated relaxed clocks have been considered.

e.g. Thorne et al, 1998:

$$r_i \sim LogNormal(r_{A(i)}, \sigma^2 \Delta t_i)$$

AC

$r \sim Exp(\lambda)$

$r \sim LogNormal(\mu, \sigma^2)$

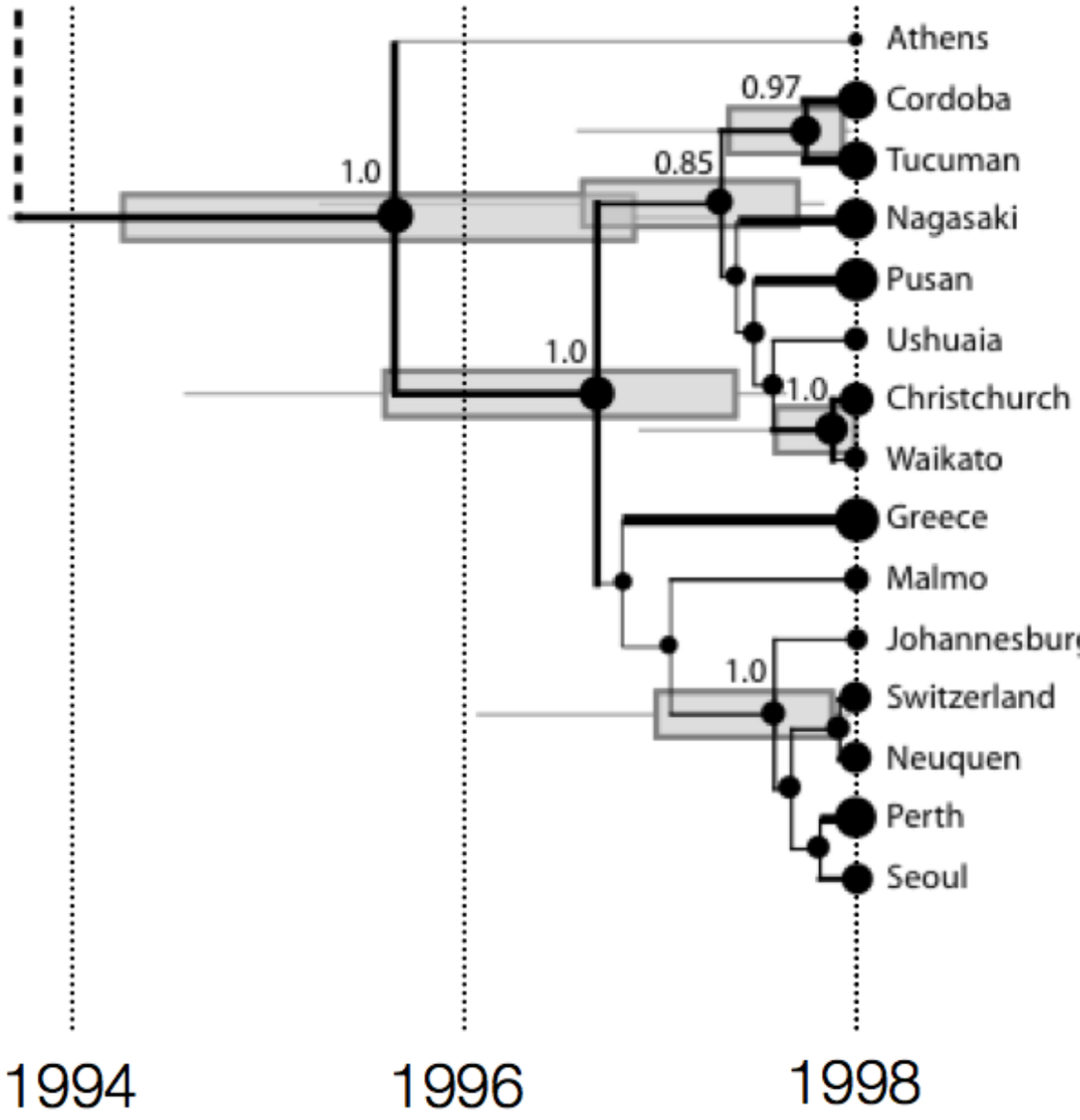
$r \sim Gamma(\alpha, \beta)$

We introduce a relaxed clock model in which there is no prior correlation between child and parent rates

"Un-correlated" or "memory-less" relaxed clocks

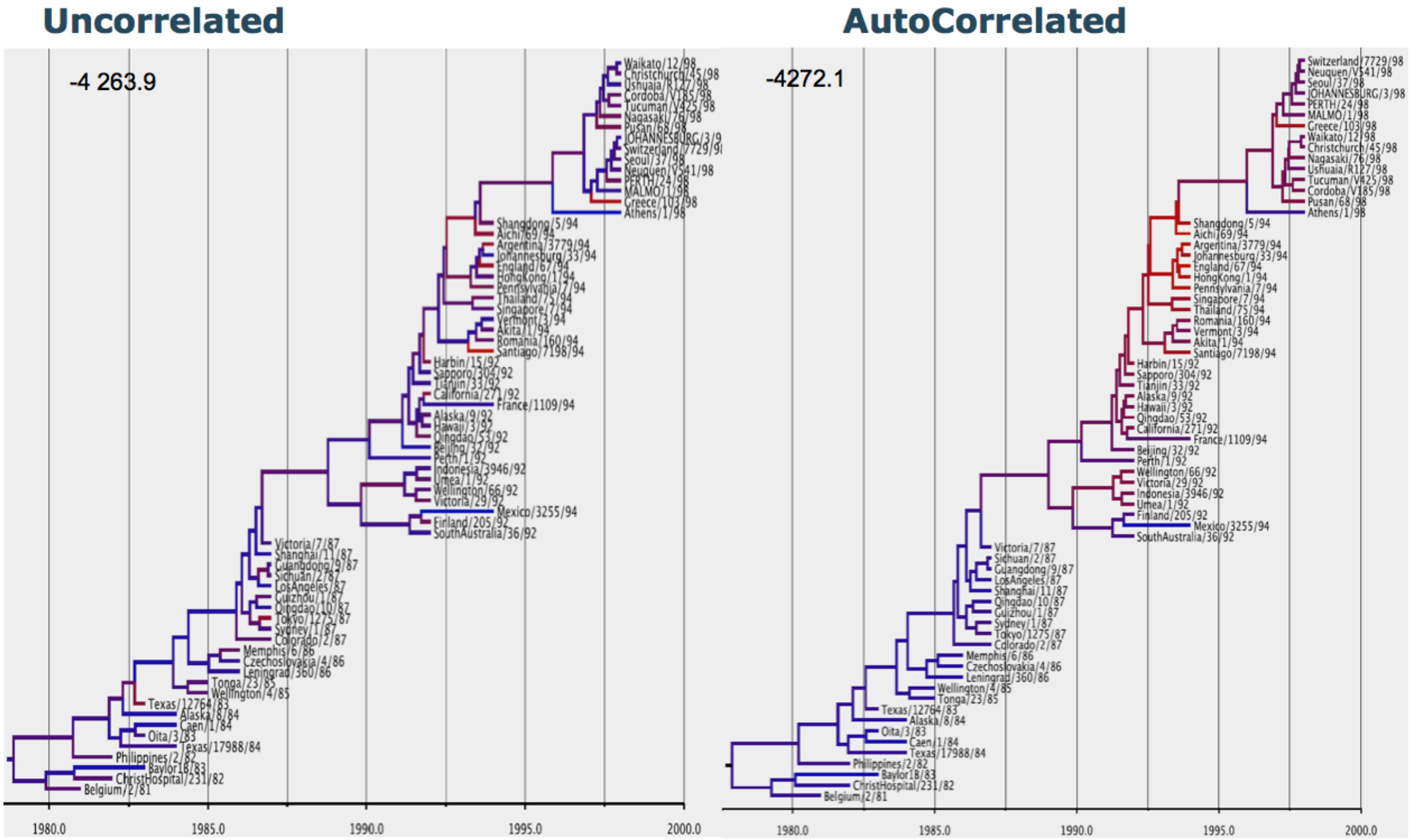
ML

Influenza A gene tree estimating using relaxed molecular clock

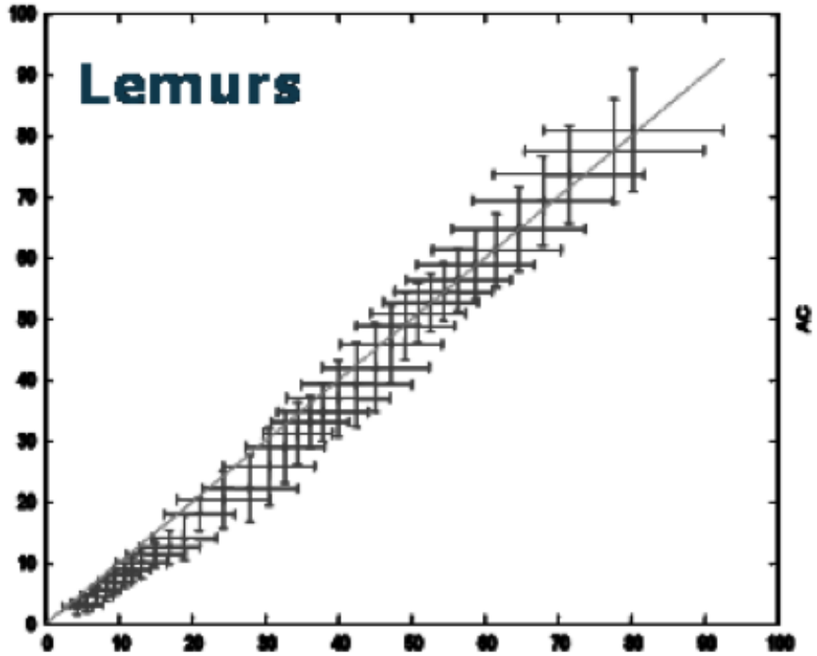


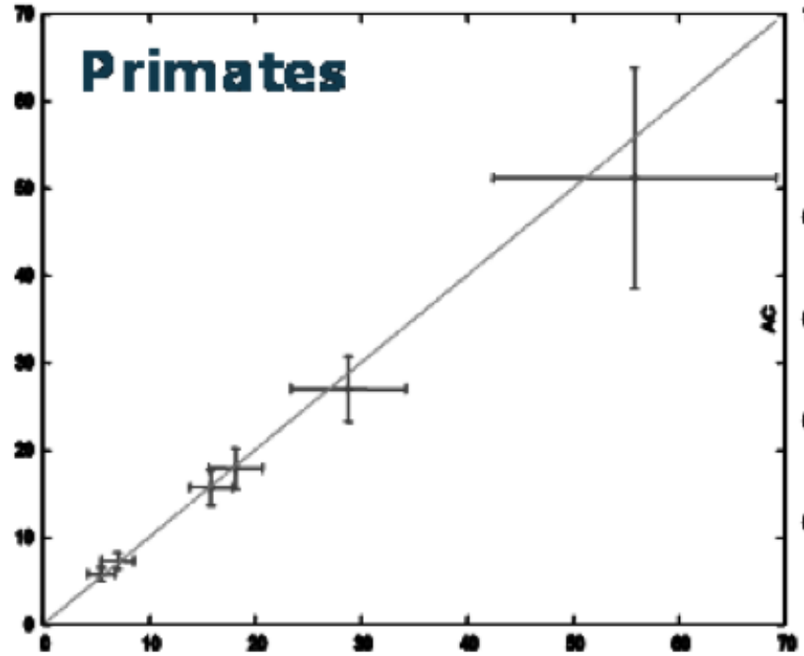
- Box-and-whisker plots show uncertainty in divergence times (only for splits with posterior probability > 0.5)
- Node size and branch thickness proportional to evolutionary rate.

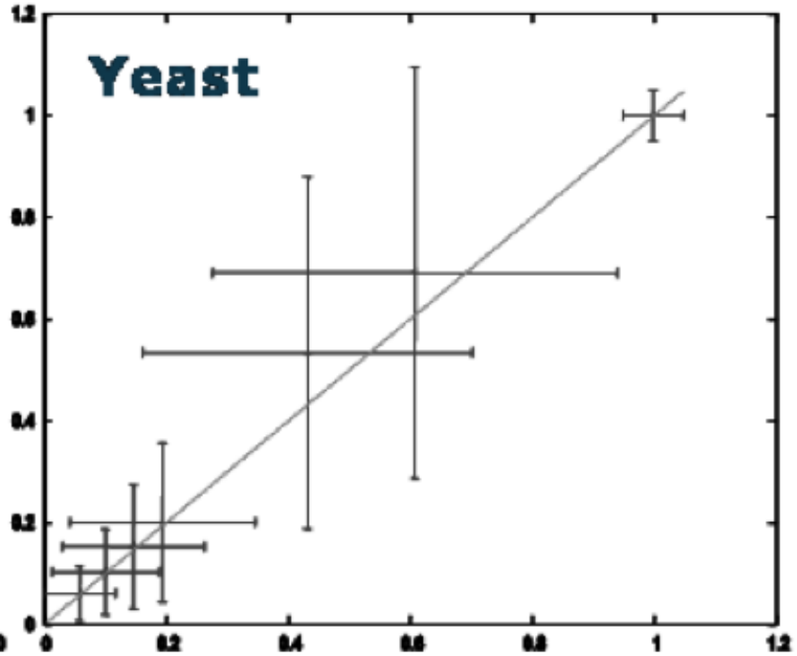
Influenza trees under different relaxed molecular clocks

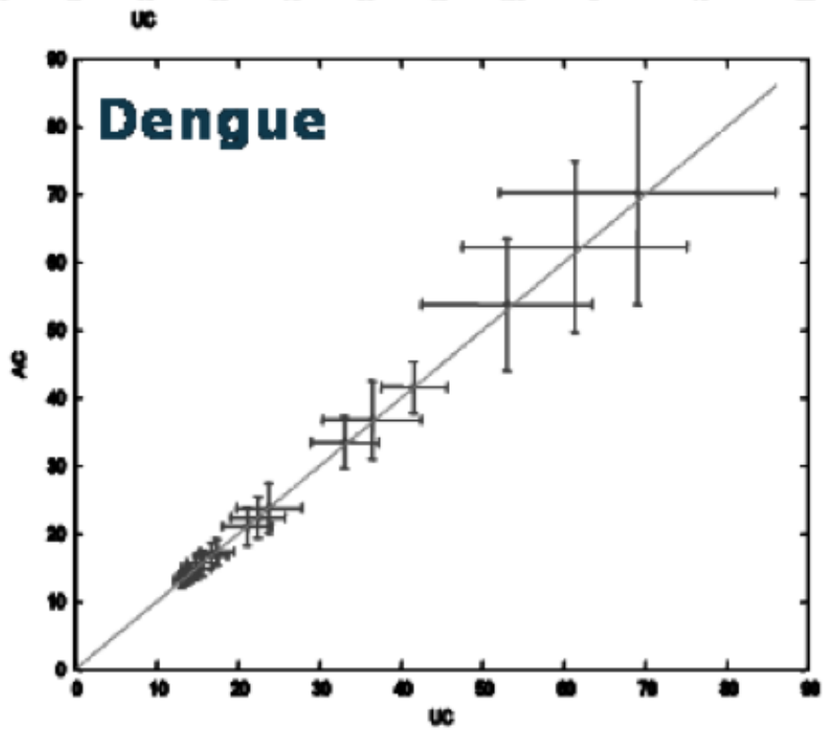


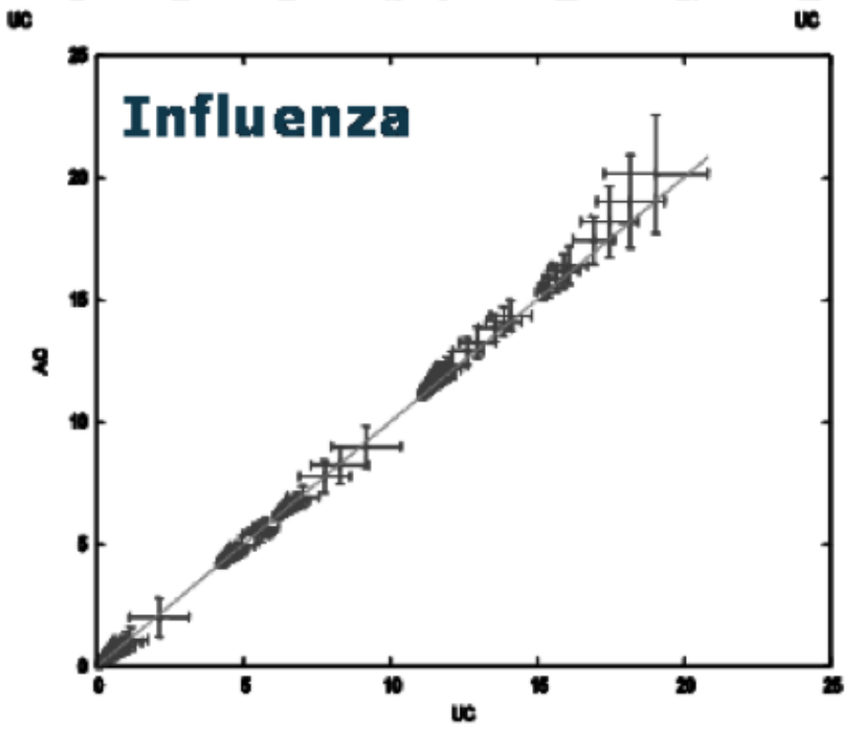
UC versus AC on five data sets



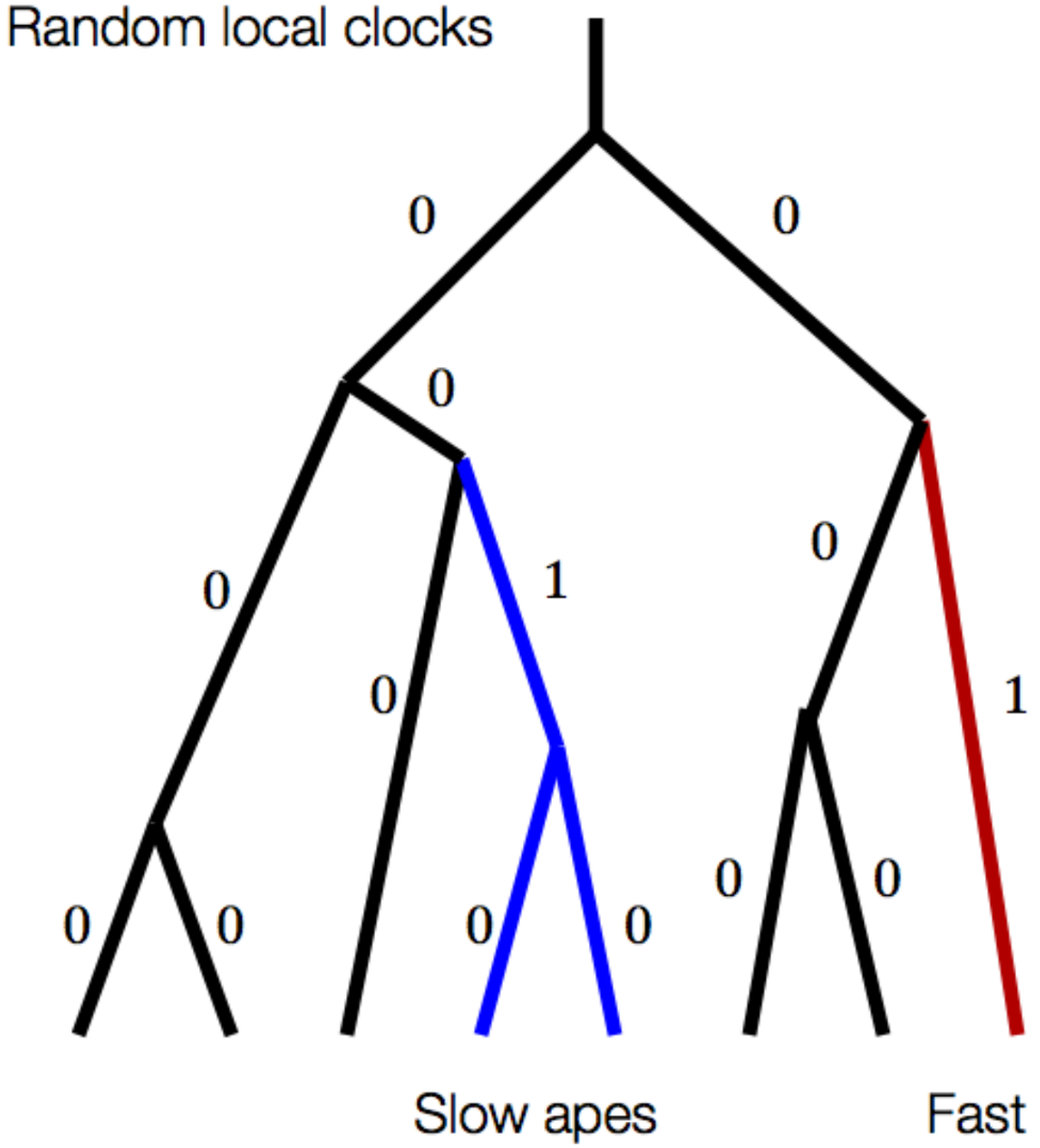






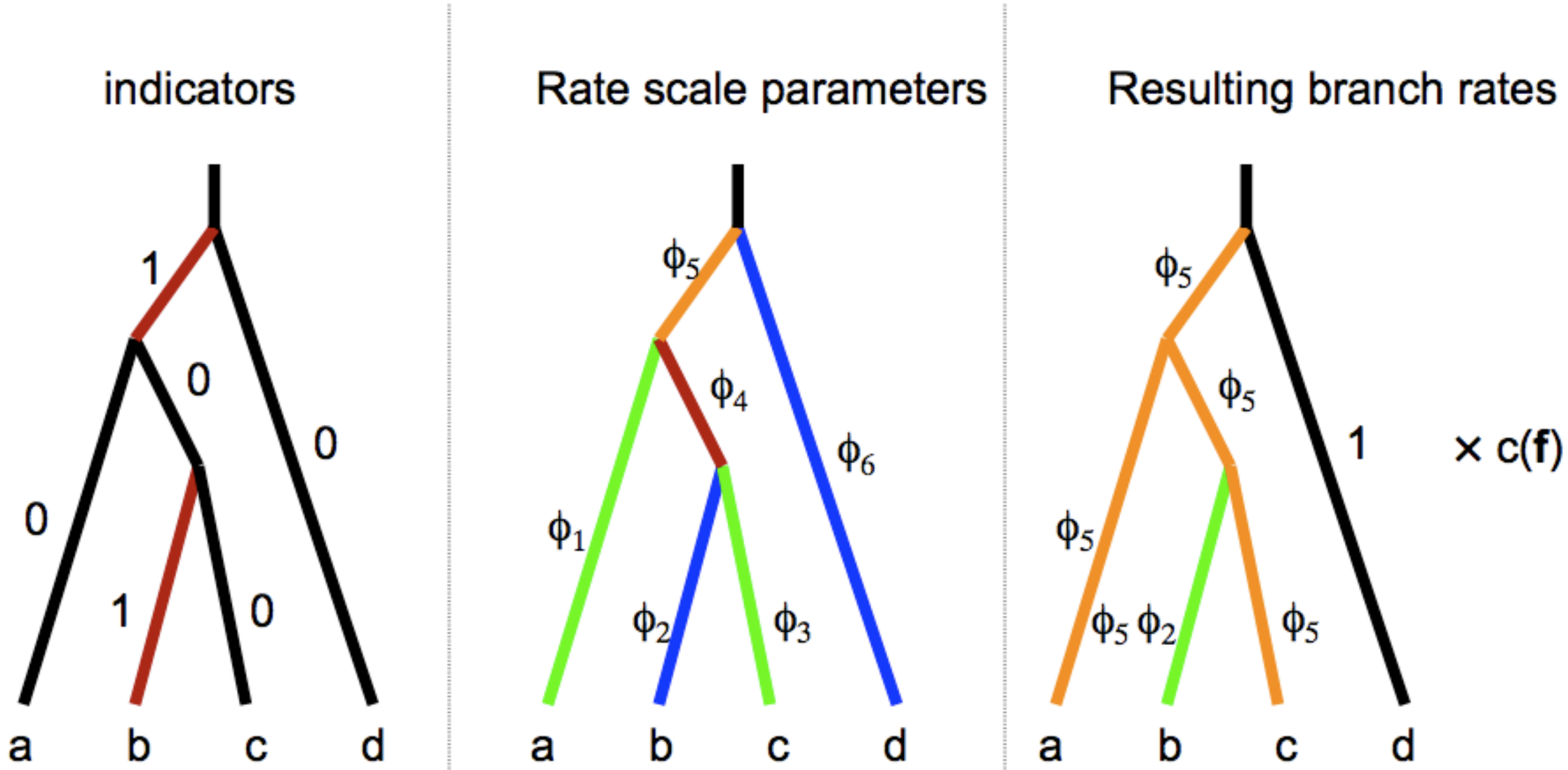


Random local molecular clocks



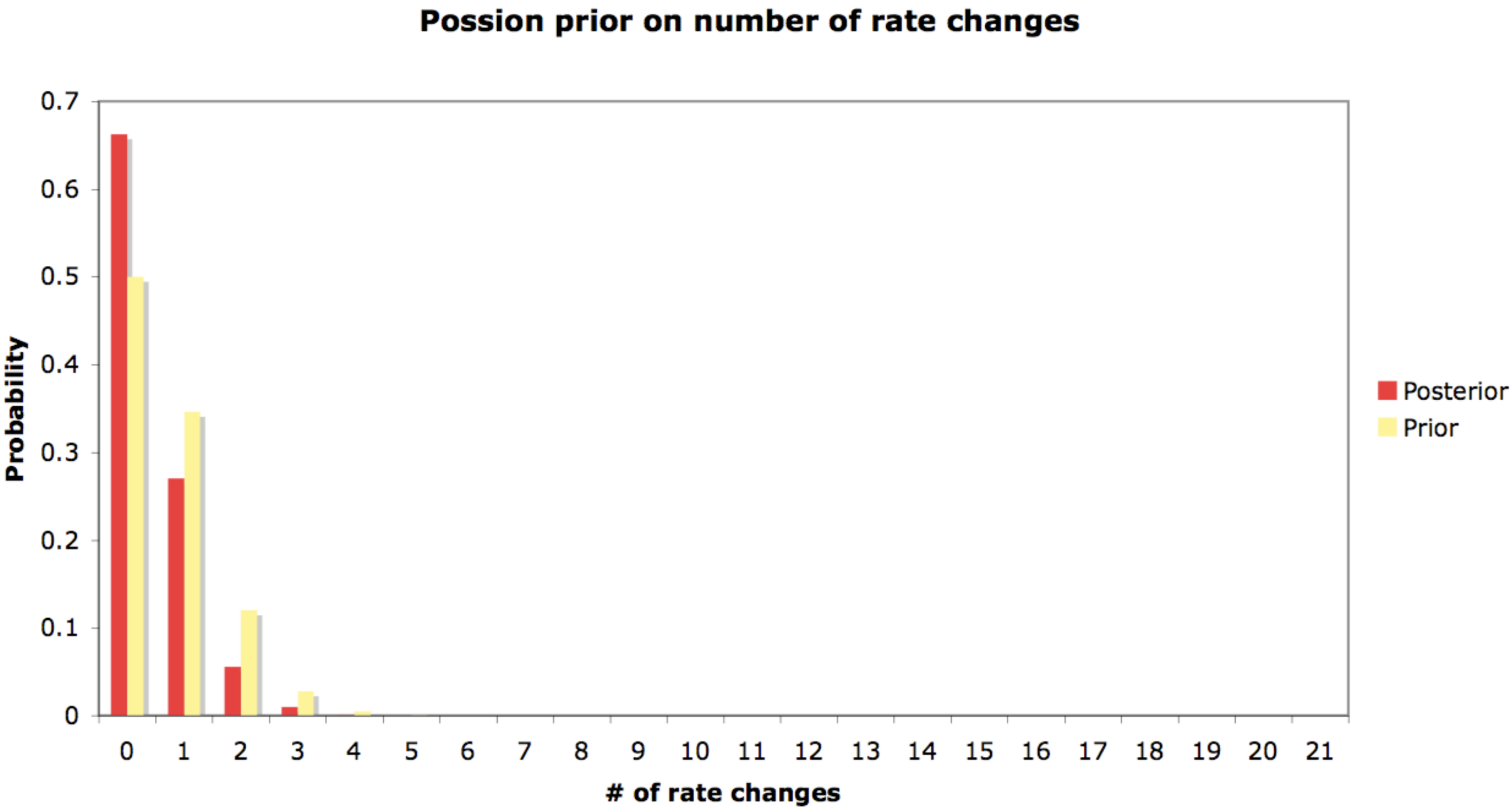
2^{2n-2} local molecular clock models

Random local molecular clocks

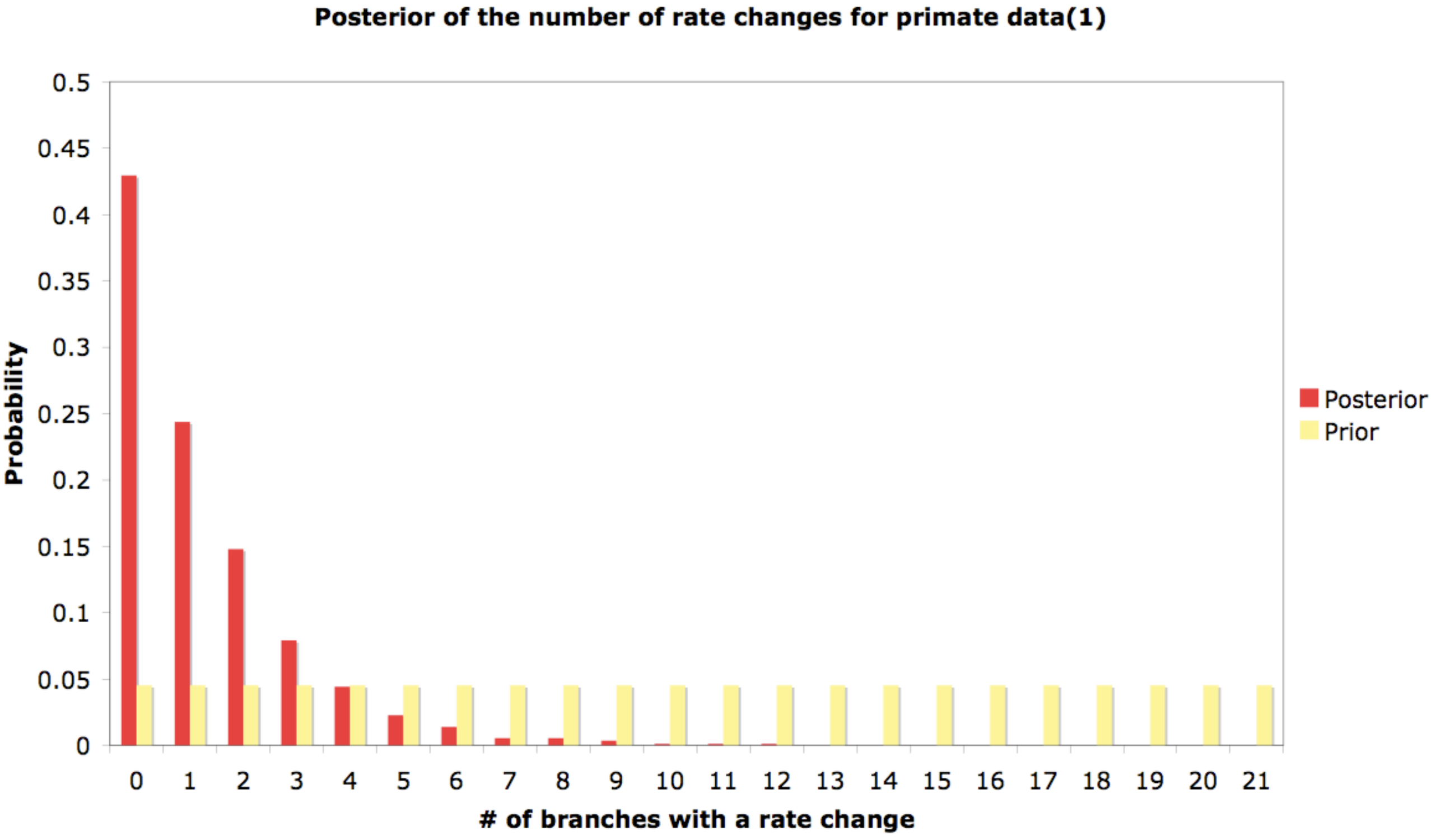


Red/Orange fast, Green/Blue slow

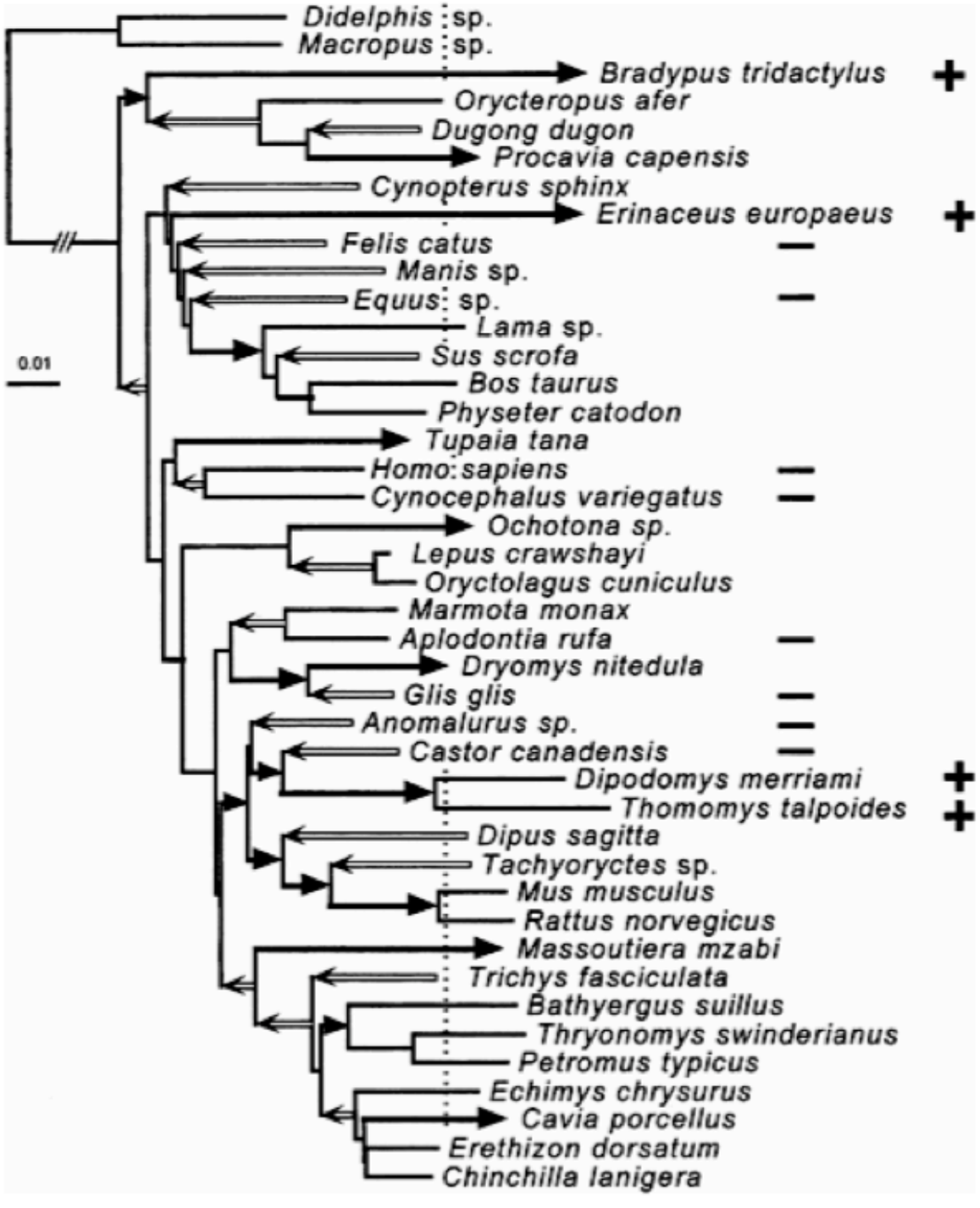
Primate data set Poisson prior



Primate data set - uniform prior



Rodents

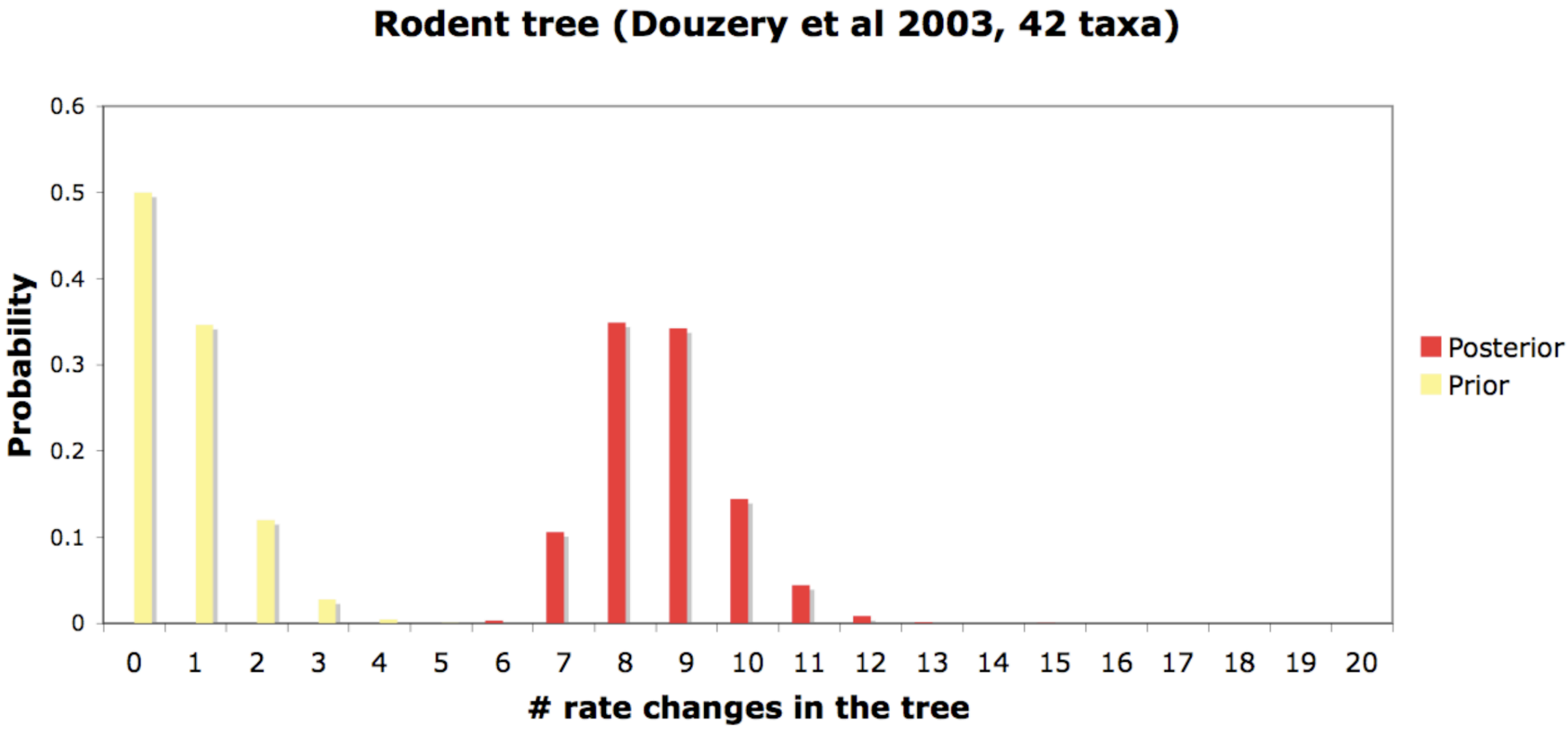


82 branches

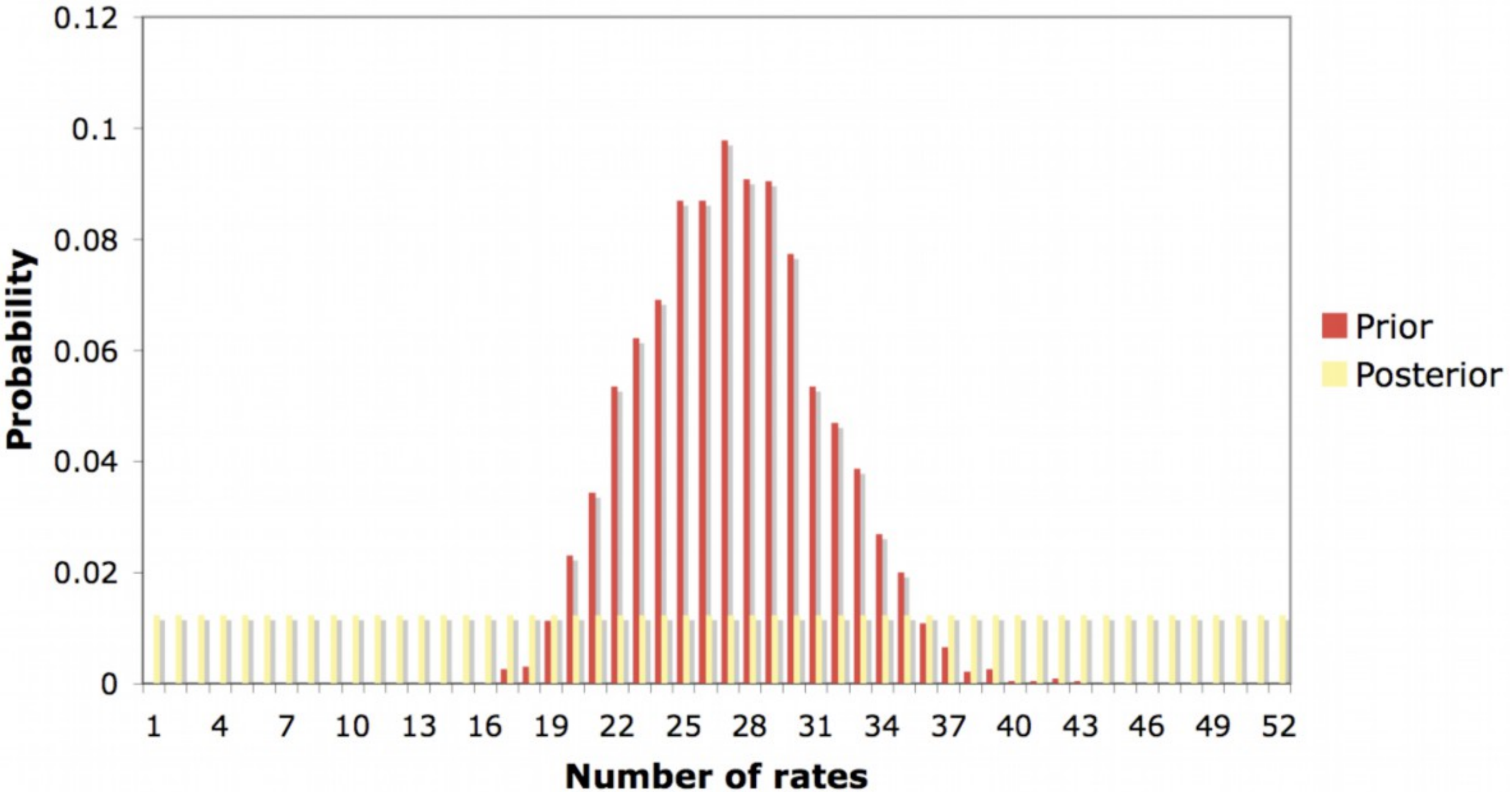
38 rate changes according to Douzery et al 2003

Fig. 1. Extensive nucleotide substitution rate variations in the first two codon positions of the ADRA2B + IRBP + vWF nuclear genes between placental mammals. The *vertical dashed line* indicates the mean value of the root-to-tip distance of the 40 placental taxa. Significantly faster- or slower-evolving species are indicated, respectively, by a + or a – as evidenced by the branch-length test. Significantly faster- and slower-evolving branches as evidenced by the two-cluster test are indicated, respectively, by *filled arrows* pointing right and *open arrows* pointing left. The scale unit corresponds to the expected number of nucleotide substitutions per site. The log-likelihood of this tree is ln$L$ = –26,054.36, and its AIC is 52282.78. In the clock-like constrained model—with a single global clock—a significant loss of log-likelihood is observed (ln$L$ = –26,222.37, AIC = 52,538.74).

Rodent data set (Poisson prior on # changes)



Rodent data set (uniform prior on # changes)



Final Perspectives

- The **molecular clock** transforms genetic distances into divergence times through substitution rates

- The **strict clock model** is the simplest, but inappropriate for most empirical datasets.

- **Relaxed clock model** model rate variations between lineages. Many different relaxed models exist and can lead to different results.

- A clock model always requires **calibration**, i.e. the use of external time information (**sample ages** or **node calibration densities**) to be identifiable.