Substitution models

Sebastian Duchene

University of Melbourne, Australia

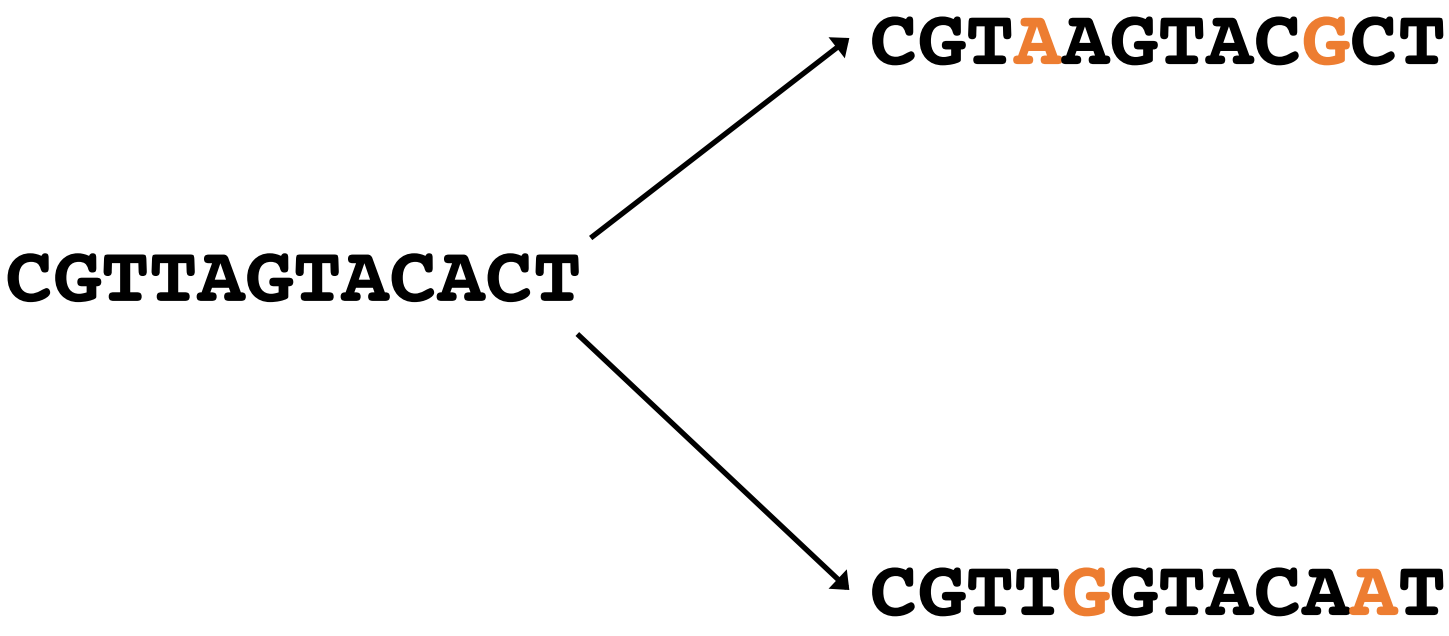
- Substitution models
- Evolutionary rates and timescales
- Calibrating the Molecular clock
- Model selection

Some of the figures in this presentation were adapted from Simon Ho's slides

Phylogenetic methods

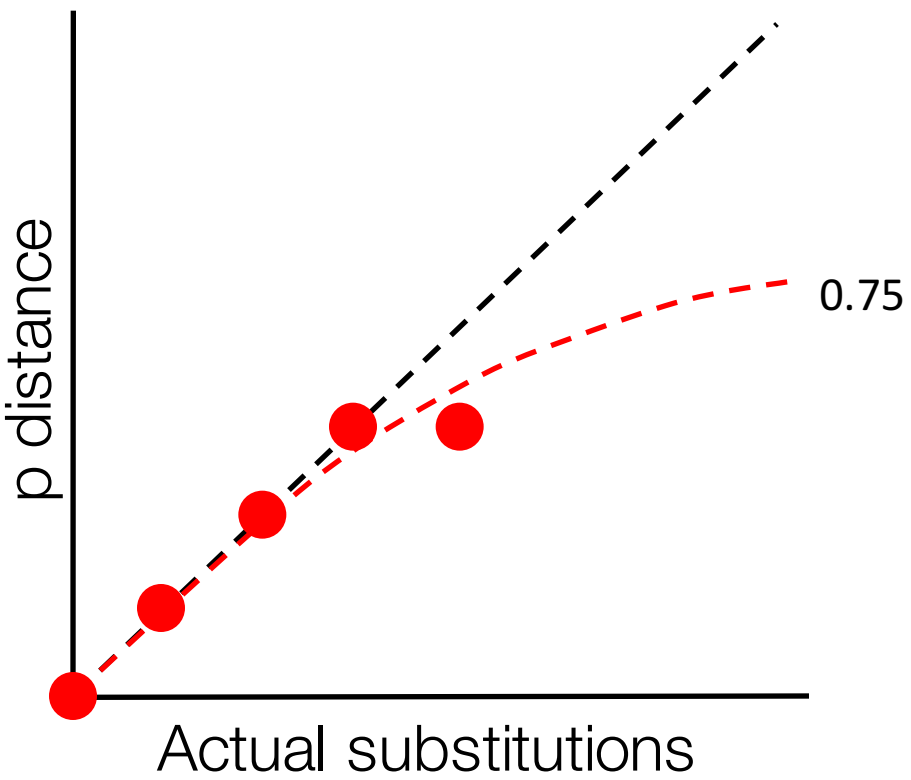
1. Maximum parsimony
2. Distance-based methods
3. Maximum likelihood
4. Bayesian inference

Model-based methods



Seq1-CGTAAGTACGCT
Seq2-CGTTGGTACAAT

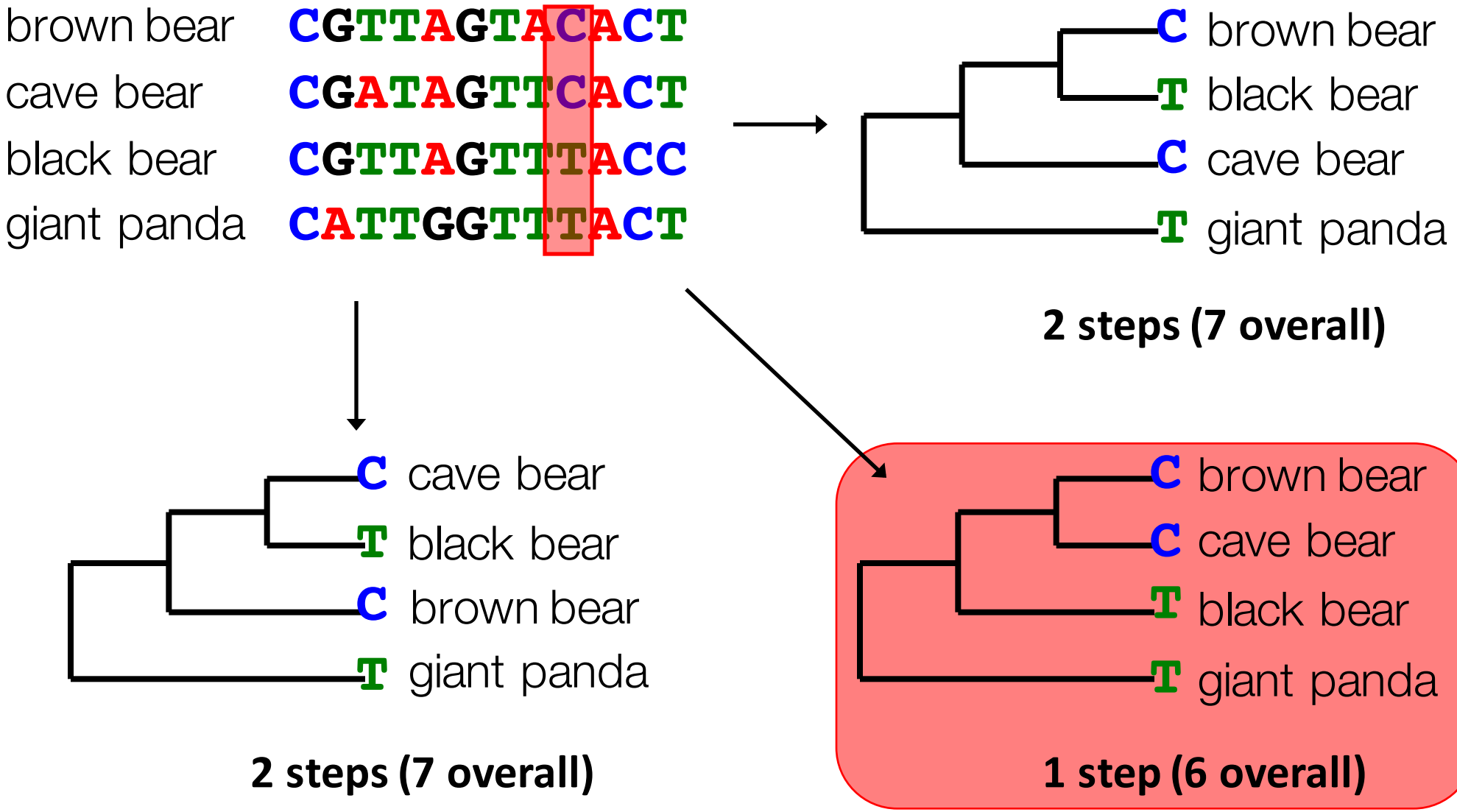
p distance = 4/12



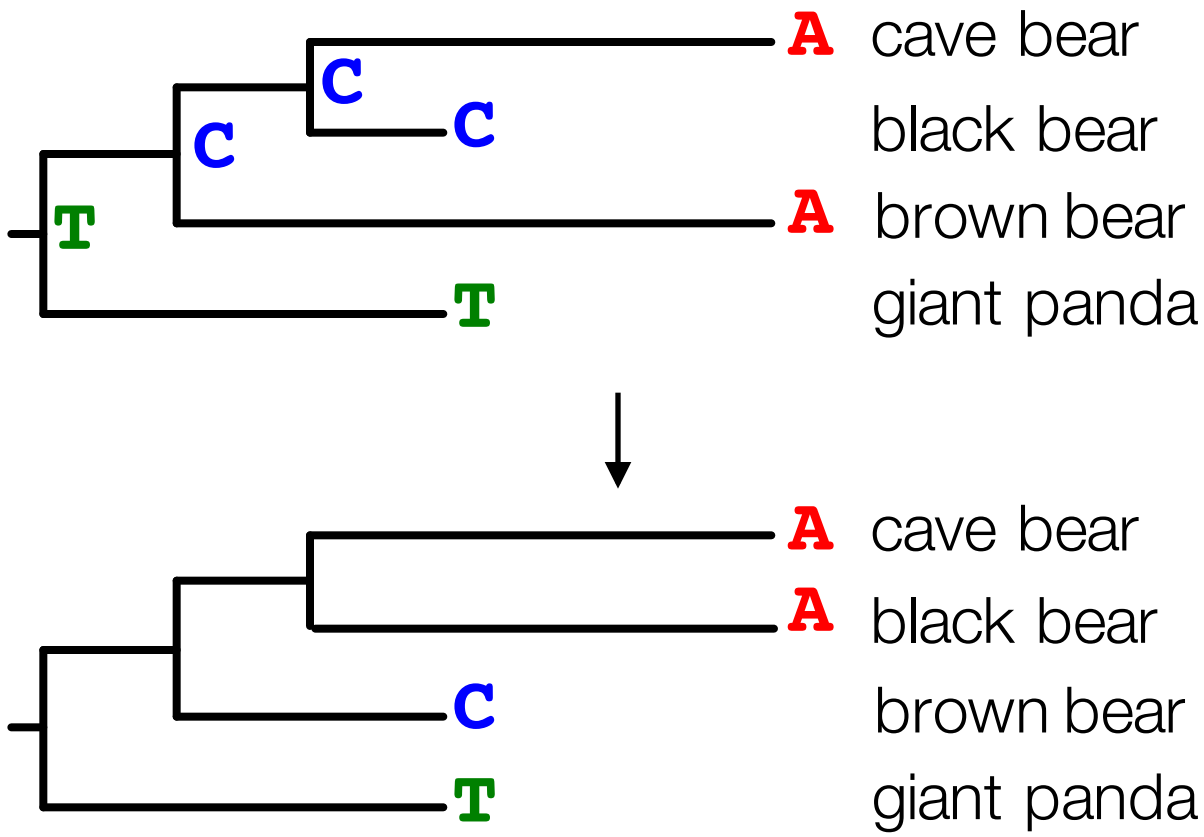
A	A	A	A	A
A	T	T	T	T
C	C	G	G	G
A	A	A	A	A
T	T	T	T	T
T	T	T	T	T
A	A	A	A	A
G	G	G	G	G
T	T	T	A	C

- This leads to a problem known as ‘long-branch attraction’
  - Long branch = many substitutions
  - Similarities arise by chance
  - Long branches cluster together

Maximum parsimony



Long-branch attraction



Maximum parsimony

- Identifies the tree topology that can explain the sequence data, using the smallest number of inferred substitution events
- Commonly used for morphological data
- Now rarely used for analysing genetic data
  - Cannot estimate evolutionary rates or timescales
  - Effects of multiple substitutions

Weaknesses

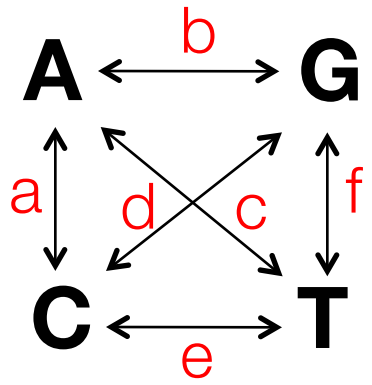
- Maximum parsimony does not correct for multiple substitutions at the same site
- This leads to a problem known as ‘long-branch attraction’
  - Long branches in the tree tend to group together

We can correct for multiple substitutions using **models** of the molecular evolutionary process

Substitution Models

Nucleotide substitution models

Rate Matrix



Base Frequencies

$\pi_A + \pi_C + \pi_G + \pi_T = 1$

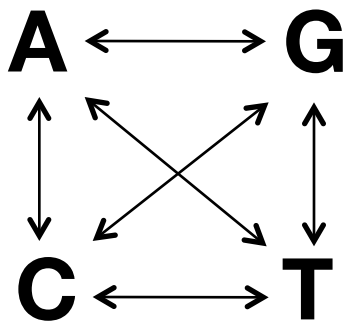
Site Rates

+ I + G

JC	**HKY**	**GTR**	**GTR+I+G**
a=b=c=d=e=f	a=c=d=f, b=e	a, b, c, d, e, f	a, b, c, d, e, f
$\pi_A=\pi_C=\pi_G=\pi_T$	π_A, π_C, π_G, π_T	π_A, π_C, π_G, π_T	π_A, π_C, π_G, π_T
No I or G	No I or G	No I or G	I, G
0 free parameters	4 free parameters	8 free parameters	10 free parameters

Nucleotide substitution models

Rate matrix (rate)



x

Base frequency Vector`

π

=

Q matrix

$$\begin{pmatrix} - & \pi_C r_{ac} & \pi_G r_{ag} & \pi_T r_{at} \\ \pi_A r_{ac} & - & \pi_G r_{cg} & \pi_T r_{ct} \\ \pi_A r_{ag} & \pi_C r_{cg} & - & \pi_T r_{gt} \\ \pi_A r_{at} & \pi_C r_{ct} & \pi_G r_{gt} & - \end{pmatrix}$$

Continuous time Markov Chain

- Consider a CTMC with two states (0 and 1).

$$Q = \begin{bmatrix} - & 1 \\ 2 & - \end{bmatrix}$$

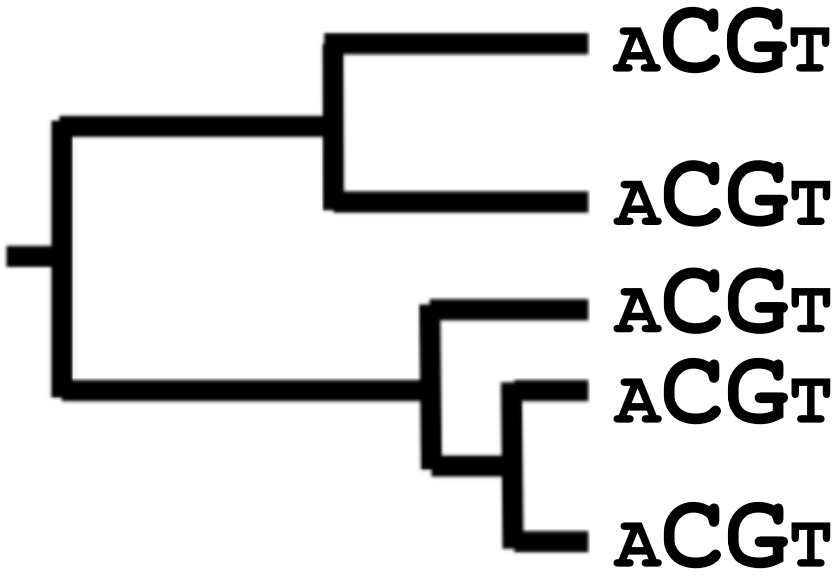
- Expected number of substitutions.
- Time at each state.

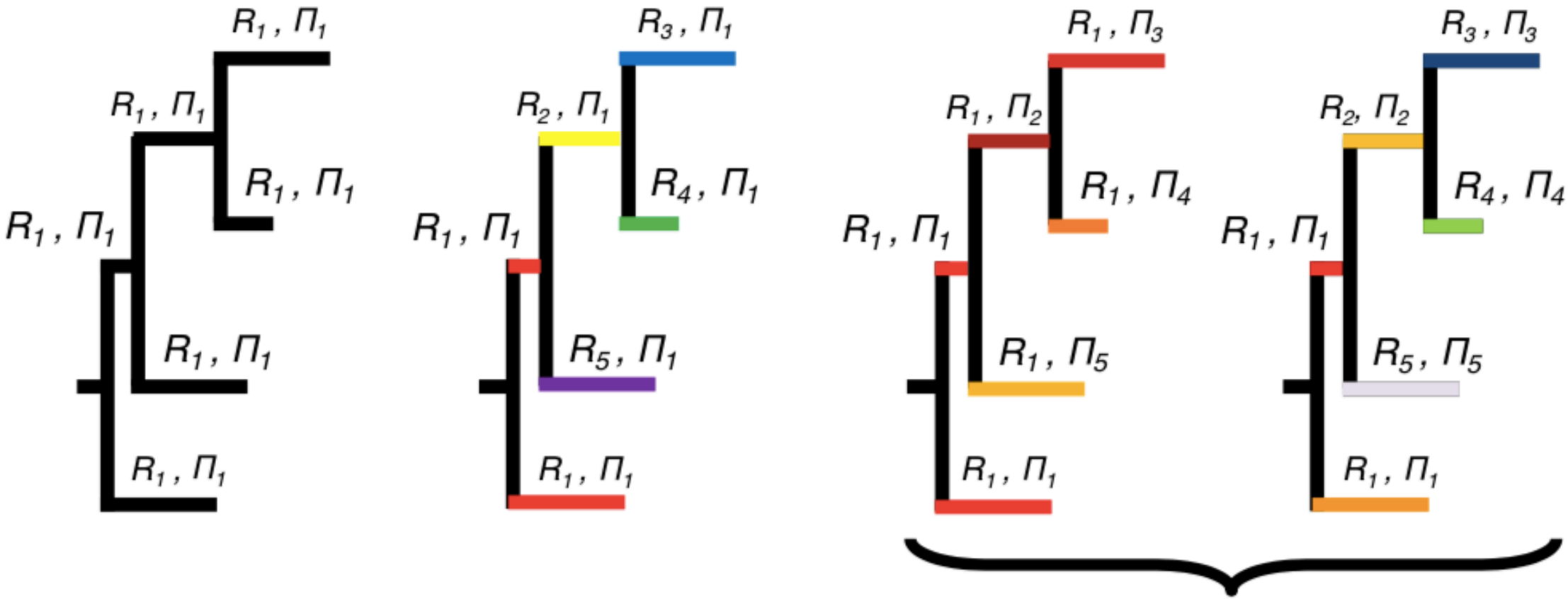
CTMC example

- Mean substitutions = 20
- Mean time at 0 = 6.33
- Mean time at 1 = 13.33

Fundamental assumptions

- Stationary.
- Reversible. $\pi_A Q_{AT} = \pi_T Q_{TA}$
- Homogeneous.





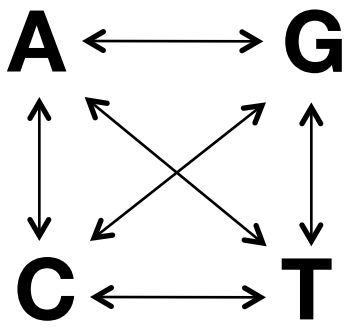
Time-reversible,
Homogeneous,
Stationary

Non-reversible,
Non-homogeneous,
Stationary

Non-reversible,
Non-homogeneous,
Non-stationary

Nucleotide substitution models

Rate Matrix Base Frequencies Site Rates

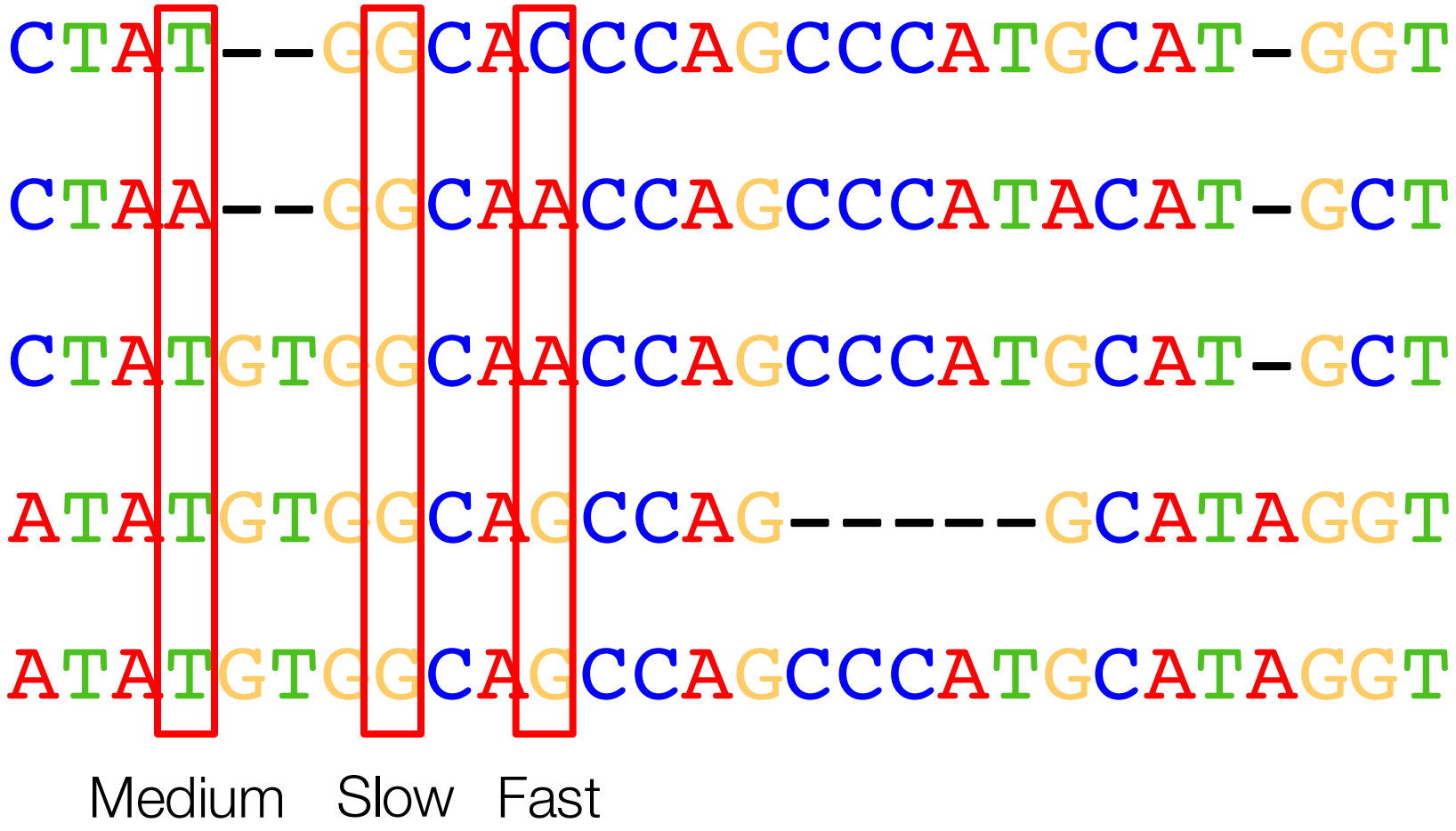


$\pi_A + \pi_C + \pi_G + \pi_T = 1$ **+ I + G**

Instantaneous rate matrix

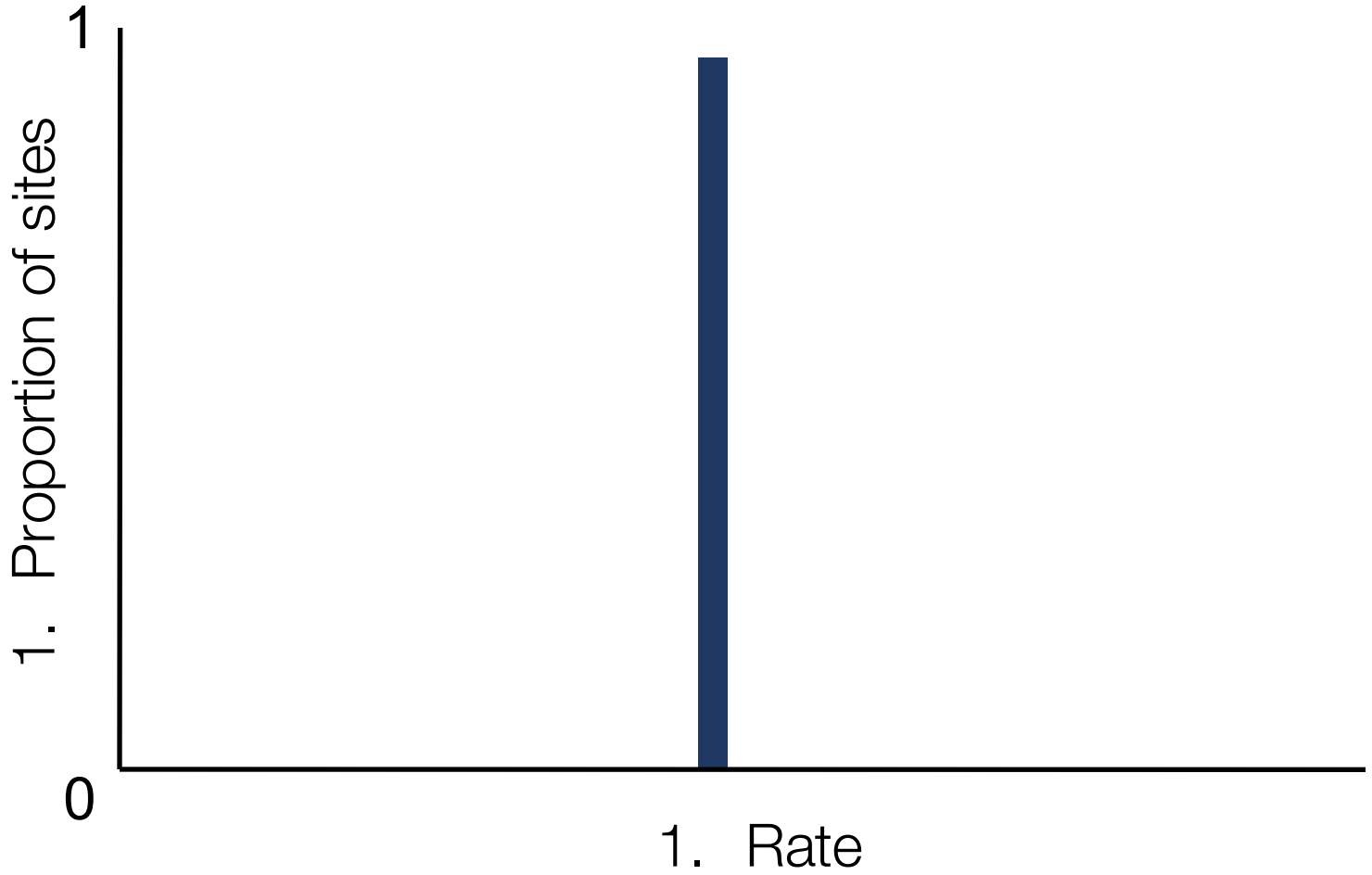
$$Q = \begin{pmatrix} - & \pi_C r_{ac} & \pi_G r_{ag} & \pi_T r_{at} \\ \pi_A r_{ac} & - & \pi_G r_{cg} & \pi_T r_{ct} \\ \pi_A r_{ag} & \pi_C r_{cg} & - & \pi_T r_{gt} \\ \pi_A r_{at} & \pi_C r_{ct} & \pi_G r_{gt} & - \end{pmatrix}$$

Rate variation across sites



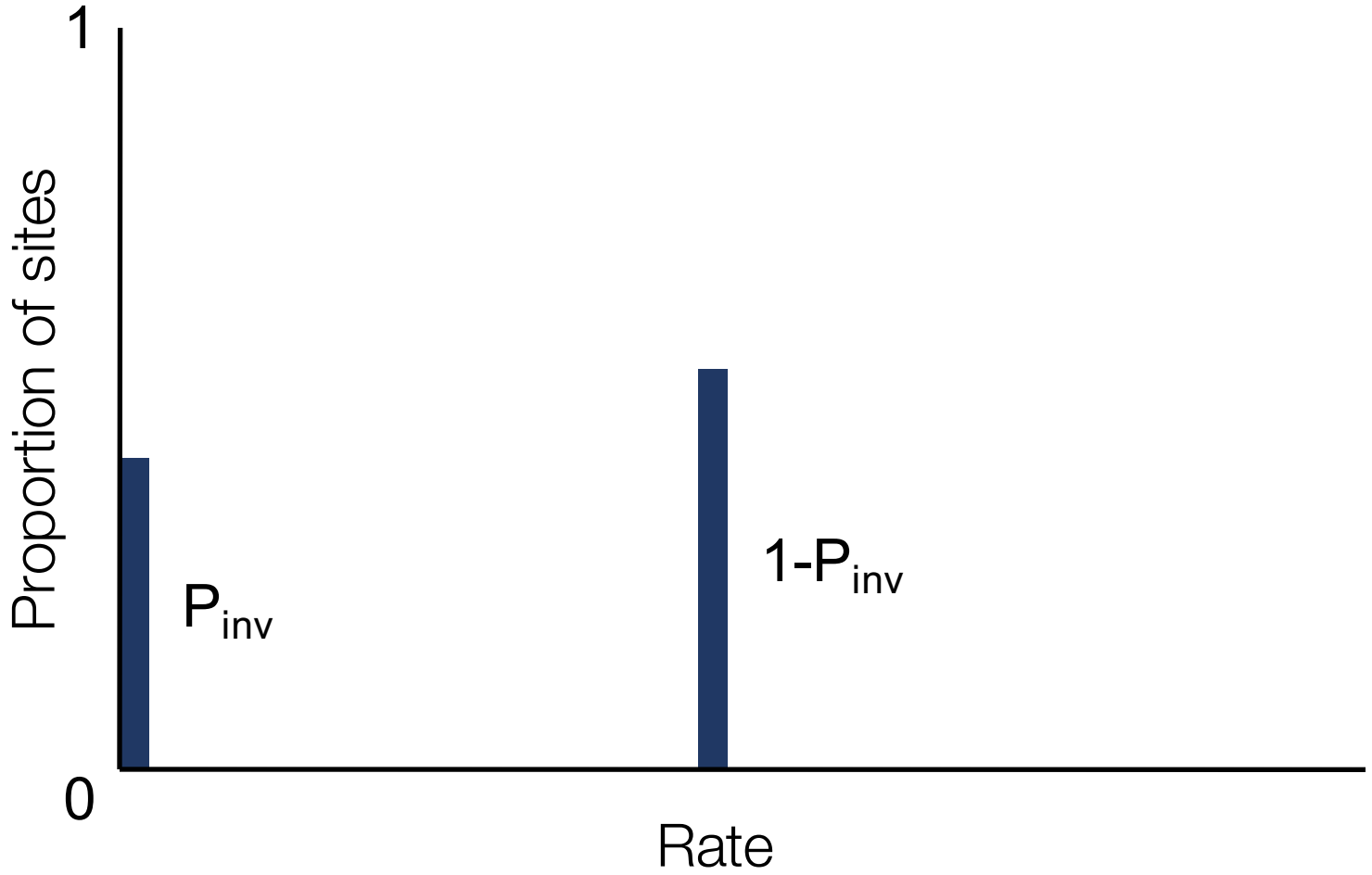
1. Rate variation among sites

1. Equal rates among sites (e.g., **JC**, **GTR**, **HKY** models)



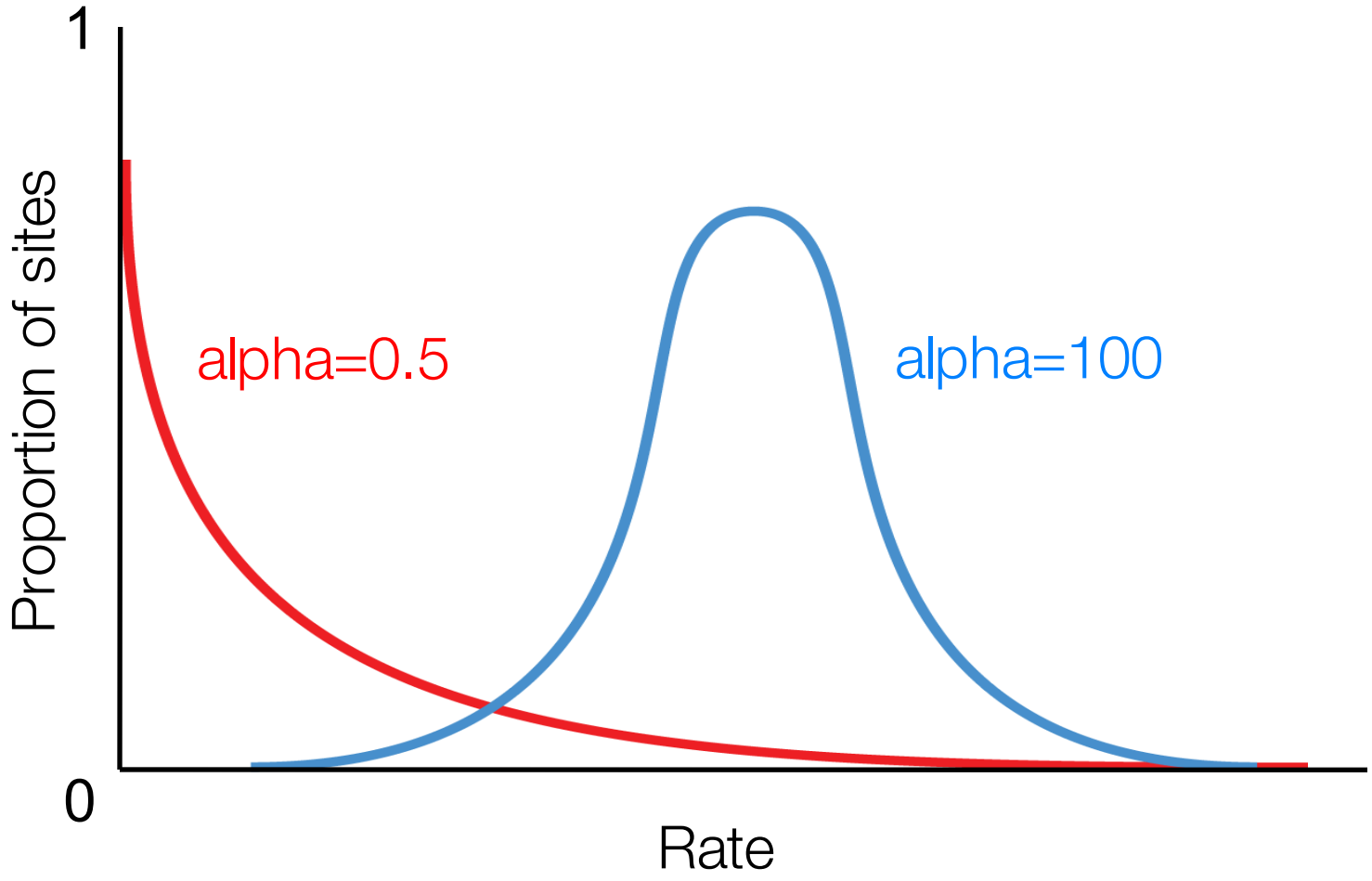
Rate variation among sites

- Proportion of invariable sites (e.g., **JC+I**, **GTR+I**, **HKY+I** models)



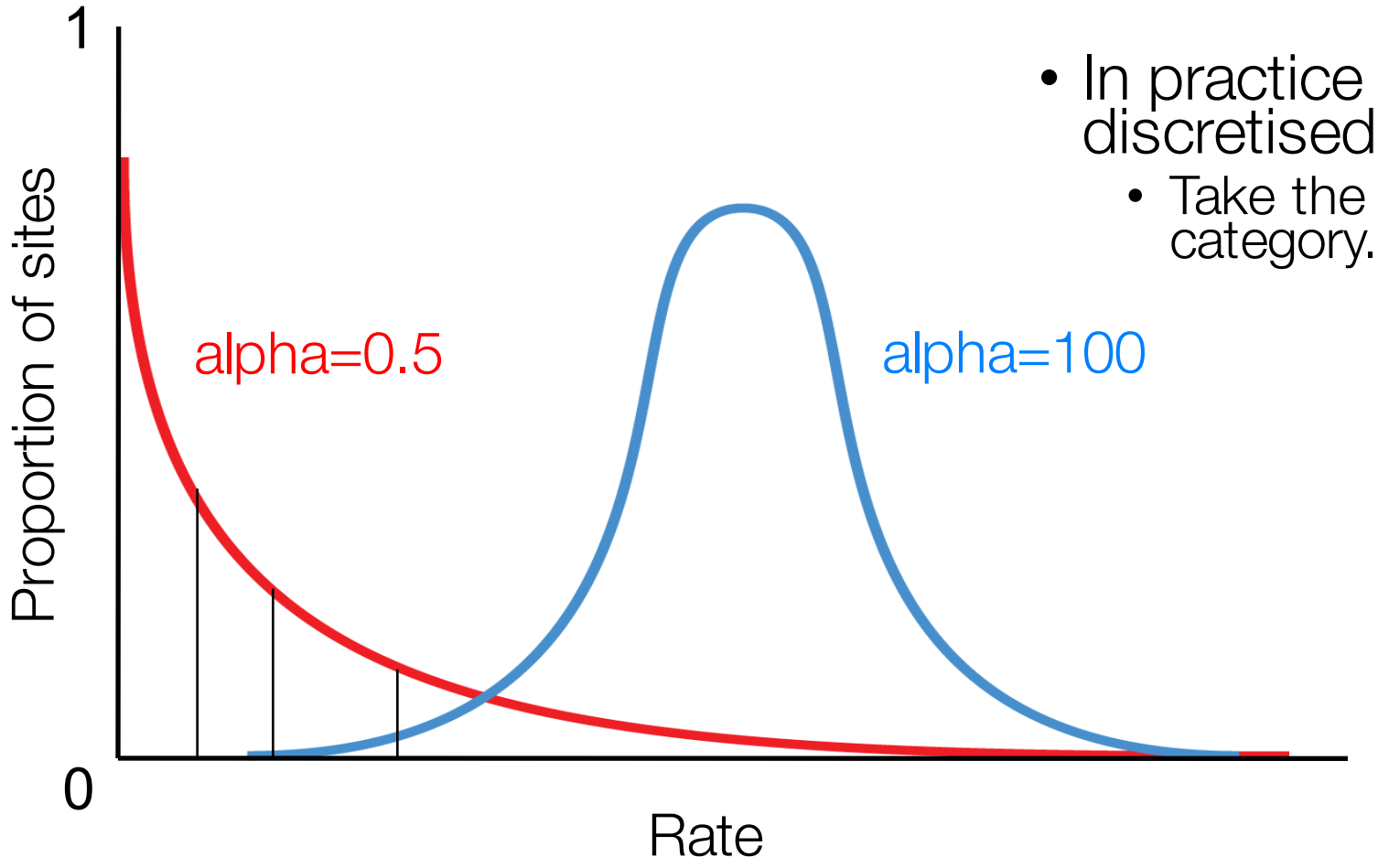
Rate variation among sites

- Gamma-distributed rate variation among sites (e.g., **JC+G**, **GTR+G**, **HKY+G** models)



Rate variation among sites

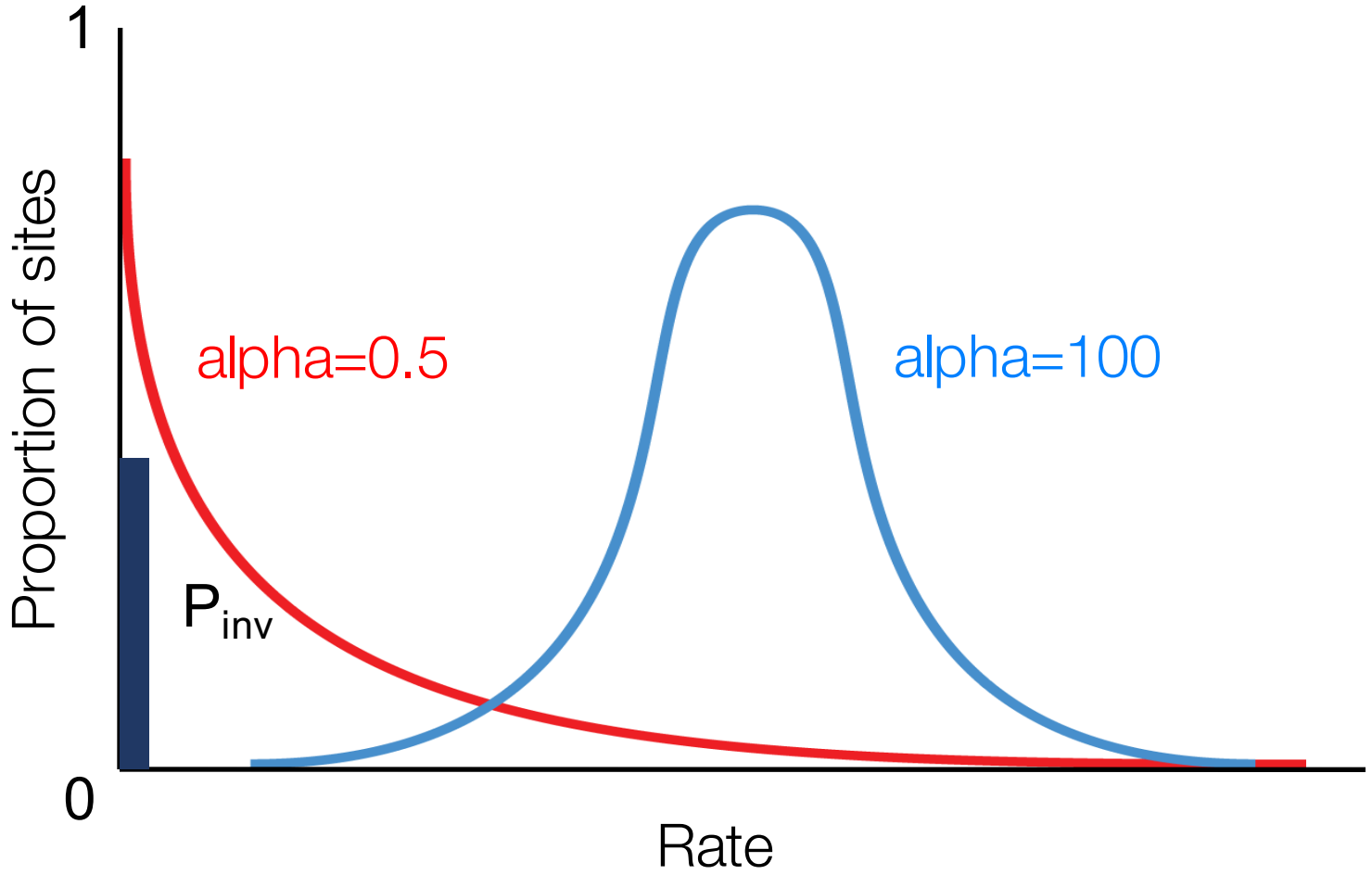
- Gamma-distributed rate variation among sites (e.g., **JC+G**, **GTR+G**, **HKY+G** models)



- In practice we use a discretised distribution.
  - Take the mean of each category.

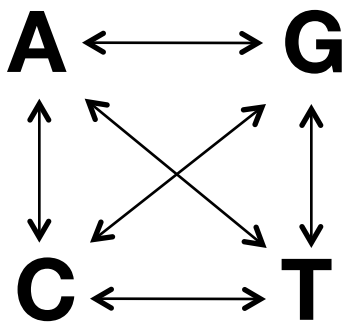
Rate variation among sites

- Gamma-distributed rate variation among sites and a proportion of invariable sites (e.g., **JC+G+I**, **GTR+G+I**, **HKY+G+I** models)



Nucleotide substitution models

Rate Matrix Base Frequencies Site Rates



$\pi_A + \pi_C + \pi_G + \pi_T = 1$ **+ I + G**

#Models

203 x 15 x 4 = 12,180

In phylogenetics, we typically consider a small subset of these

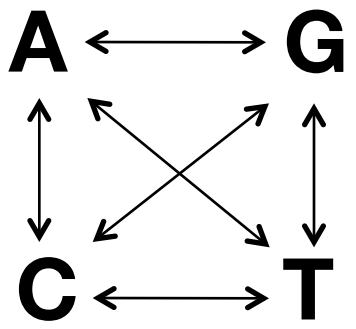
Amino acid substitution matrices

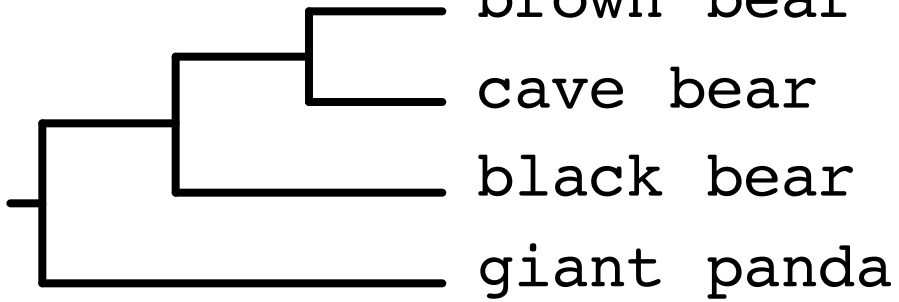
- 20x20 matrix of substitution probabilities
- Too many parameters to estimate
  - GTR model for DNA: 6 parameters
  - GTR model for proteins: 190 parameters
- Estimate substitution probabilities using a large data set
- Standard matrices:
  - PAM, BLOSUM, etc.

Fundamental assumptions

- Stationary
- Reversible
- Homogeneous
- Independent across sites

π_A π_C π_G π_T



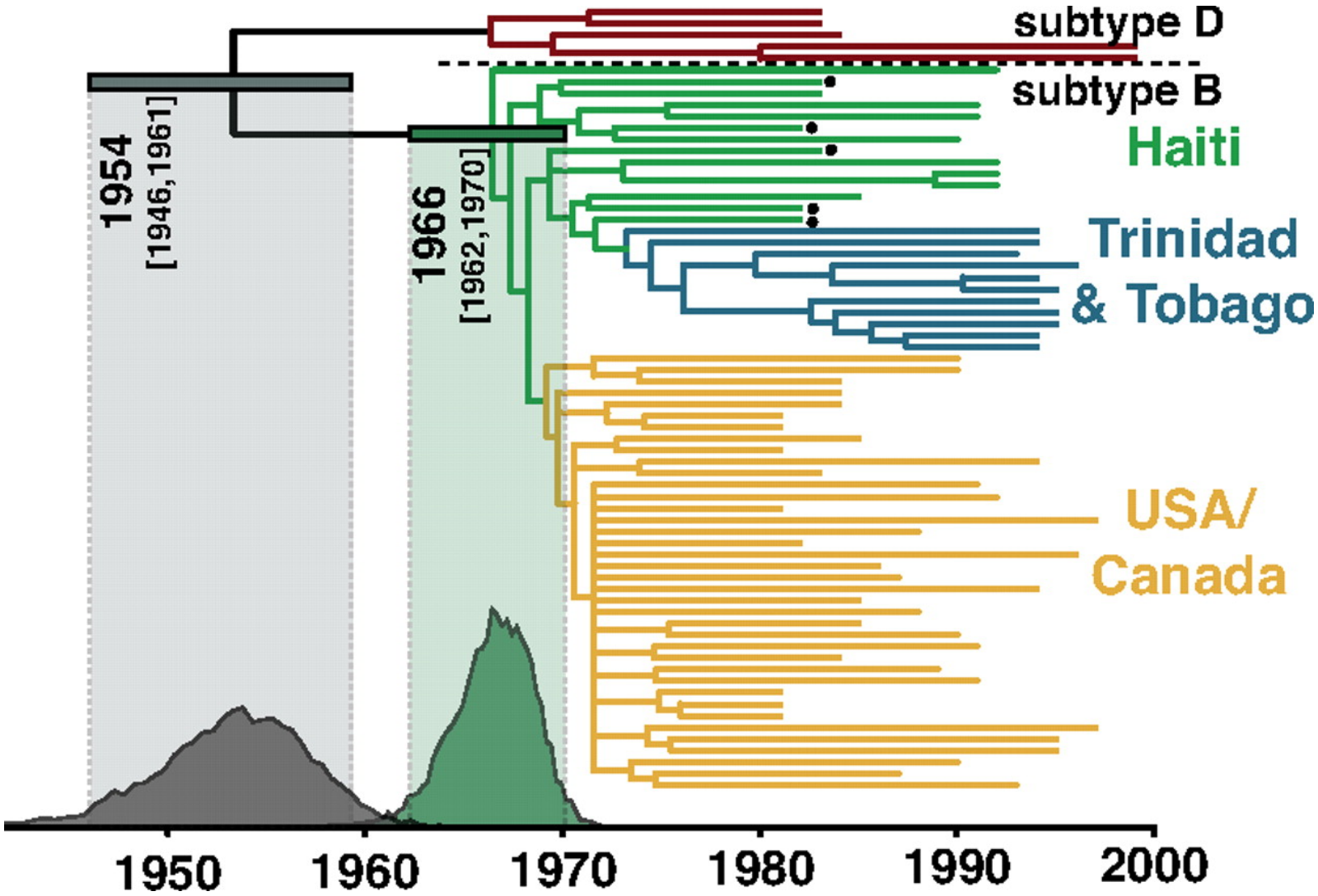


CGTTAGTACACT
CGATAGTTCACT
CGTTAGTTTACC
CATTGGTTTACT

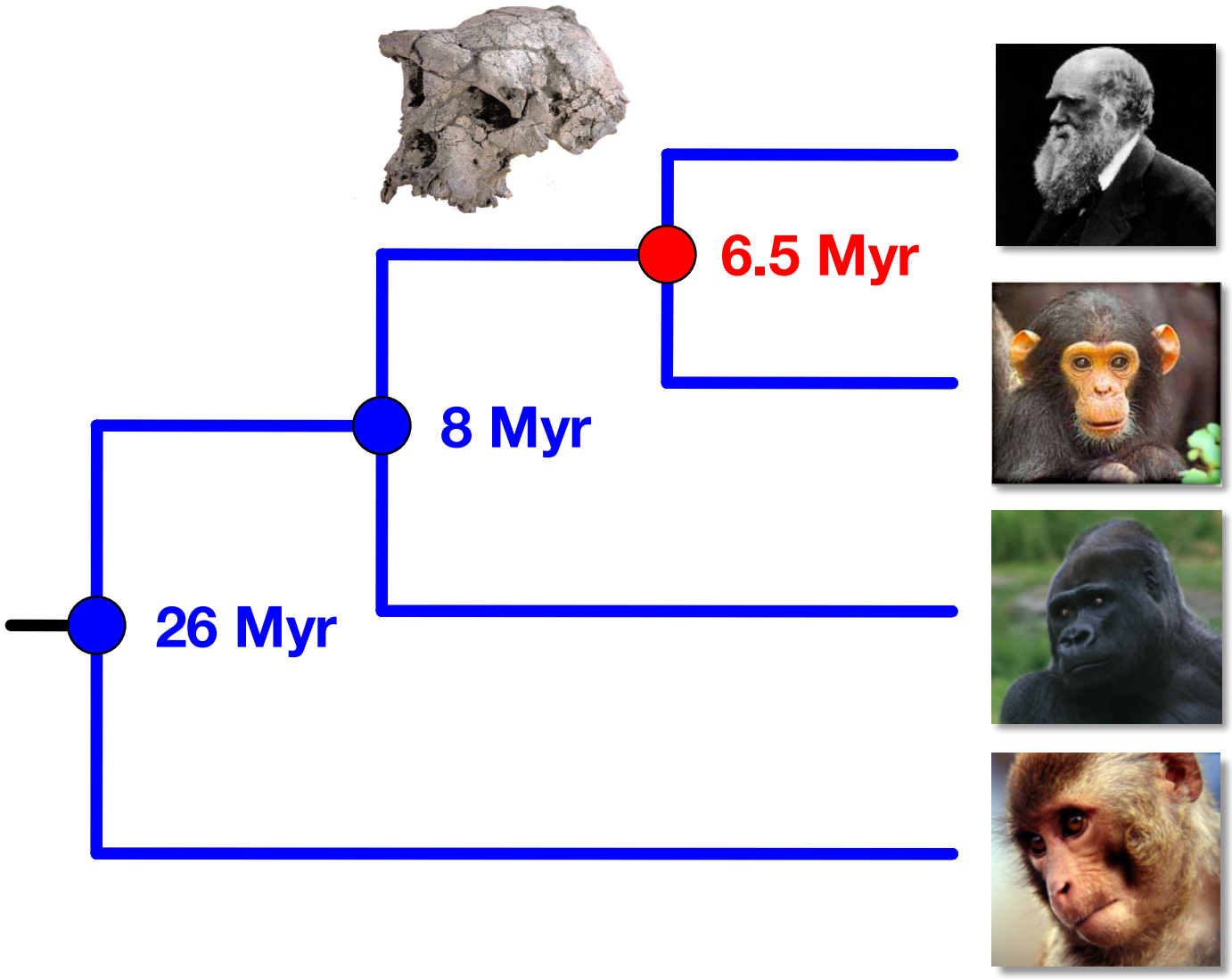
Evolutionary rates and timescales

Diversification of mammals

Emergence of HIV/AIDS in the Americas

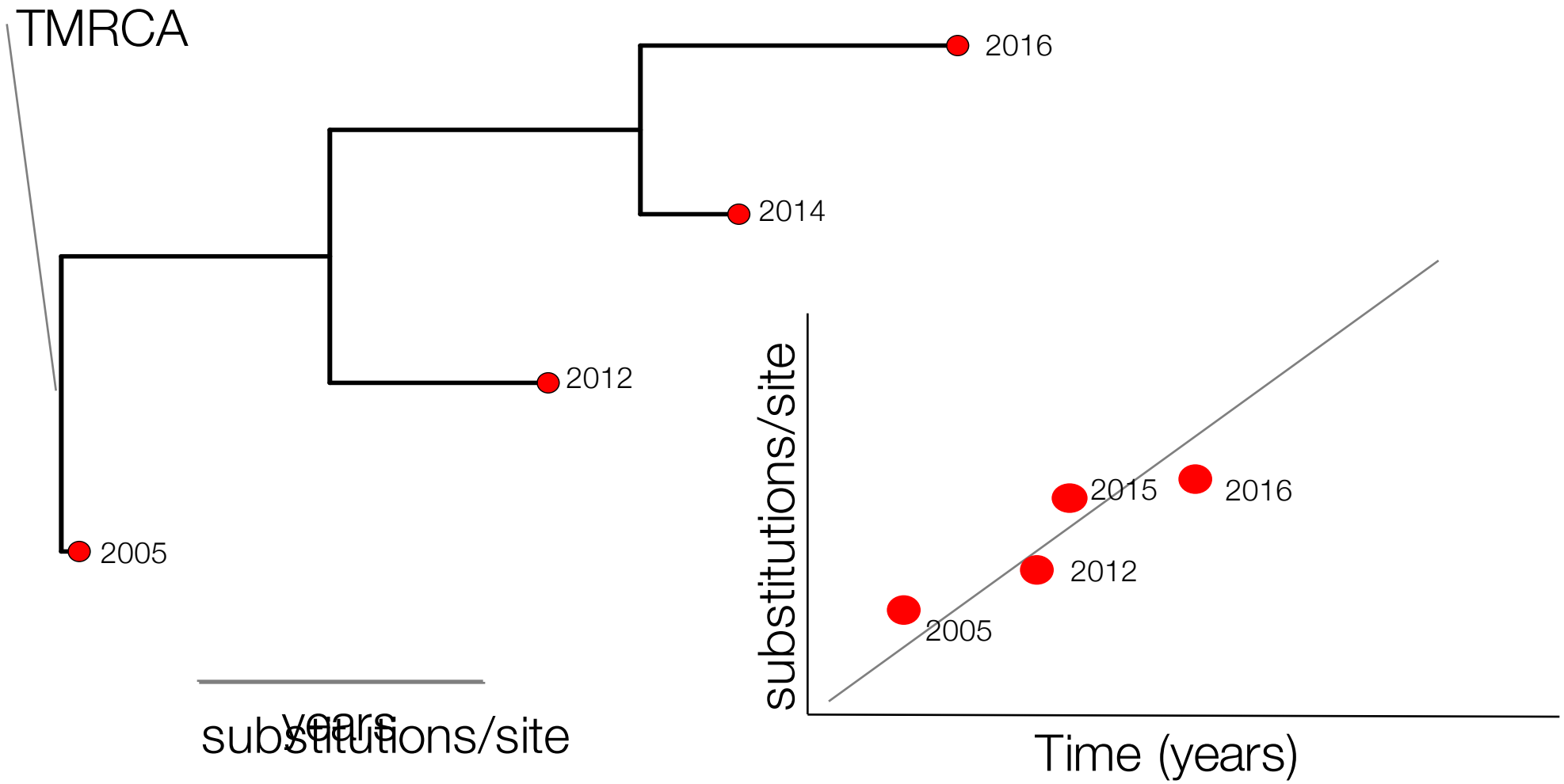


The Molecular Clock



5% genetic difference

The molecular clock



The molecular clock

- Zuckerkandl & Pauling (1962)
- Margoliash (1963)
- Doolittle & Blomback (1964)
- **Zuckerkandl & Pauling (1965)**

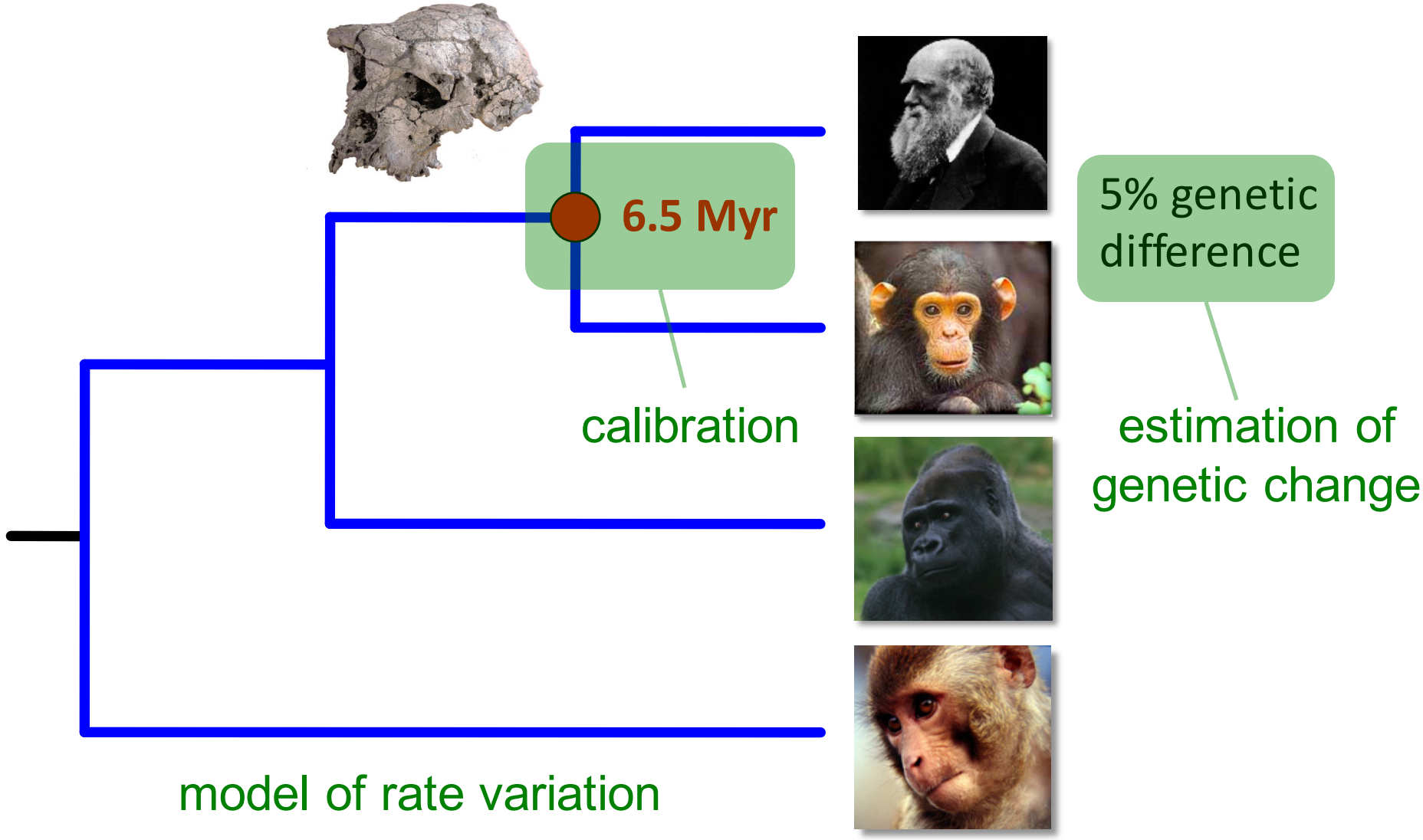
Assumed constant rate among species to estimate timing of globin gene duplications

Proportional relationship between genetic distance and time since divergence

Examined correlation between time and genetic divergence in mammalian fibrinopeptides

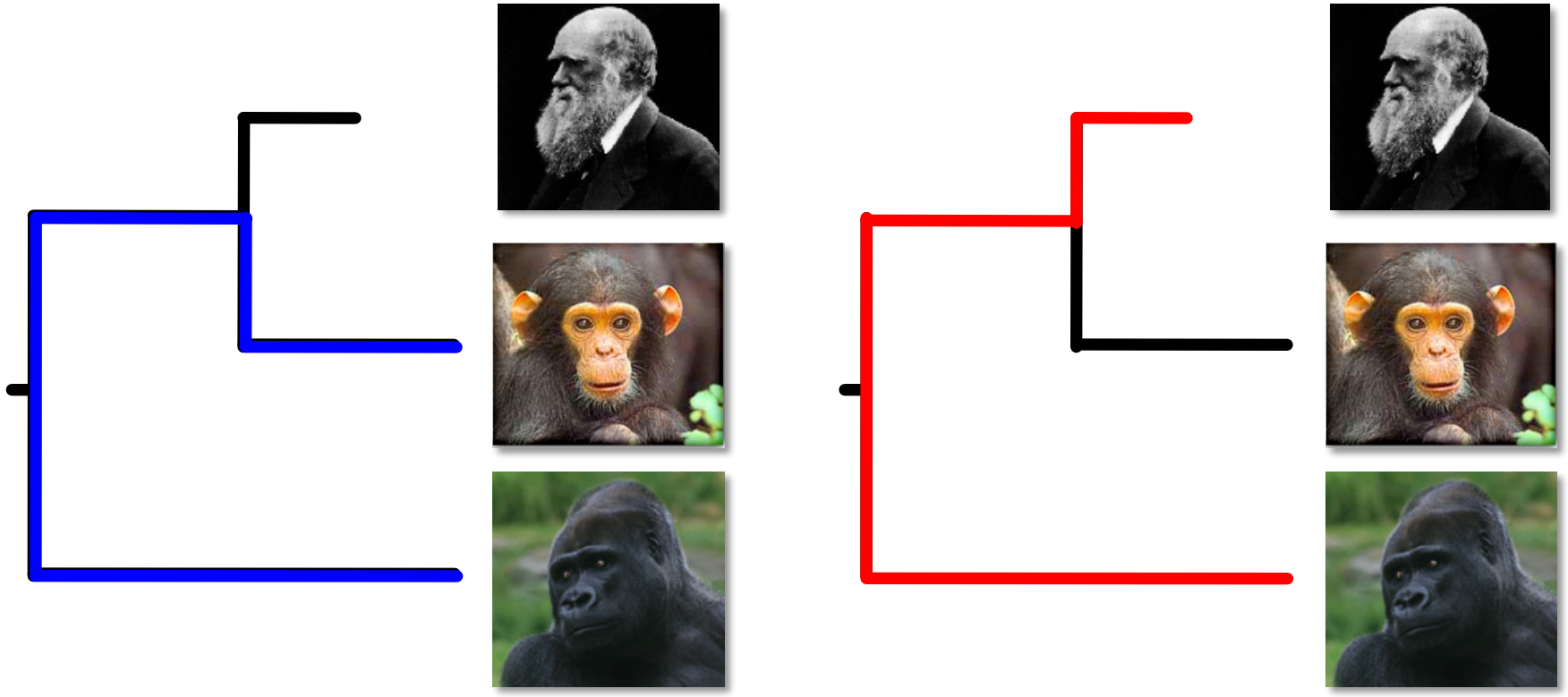
Introduced the term ‘molecular evolutionary clock’

Sources of uncertainty



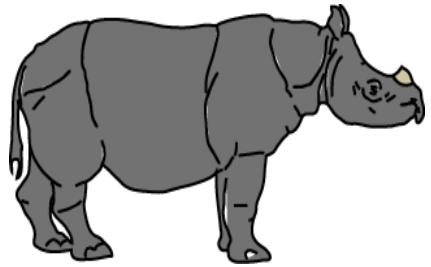
Testing for clocklike evolution

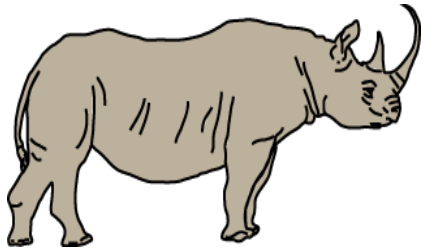
- Relatives-rates test (Fitch 1976)



Departures from the clock

- Rates vary among lineages
  - Differences in mutation rates
  - Differences in strength and direction of selection
  - Differences in population size
    - Predictors of rates:
      - Longevity (mammals)
      - Height (flowering plants)





Why keep the molecular clock?

- The behaviour of most real sequences does not satisfy the assumption of a strict molecular clock

Bromham & Penny (2003):

The molecular clock is an irreplaceable source of information in evolutionary biology and it would be foolish to abandon it altogether

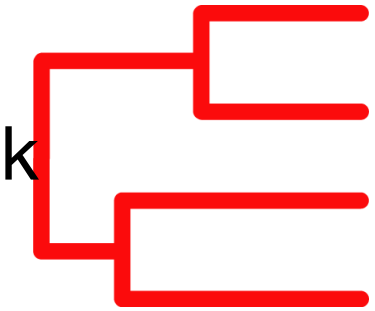
Testing for clocklike evolution

- Likelihood-ratio test
  - Strict clock vs unconstrained model
- Bayes factors
  - Comparing strict clock against other models
  - Assess rate variation in relaxed clock models

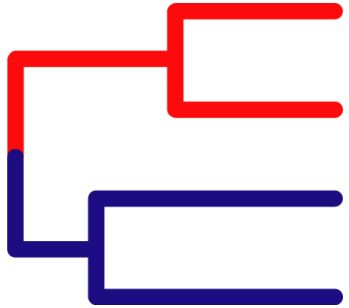
Relaxed Molecular Clocks

Molecular Clock Models

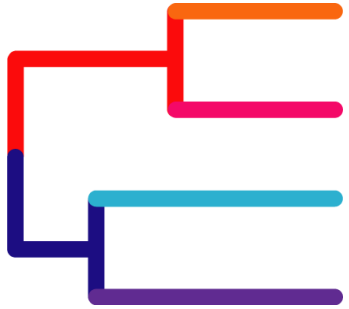
- Strict or ‘global’ molecular clock



- Multi-rate clocks



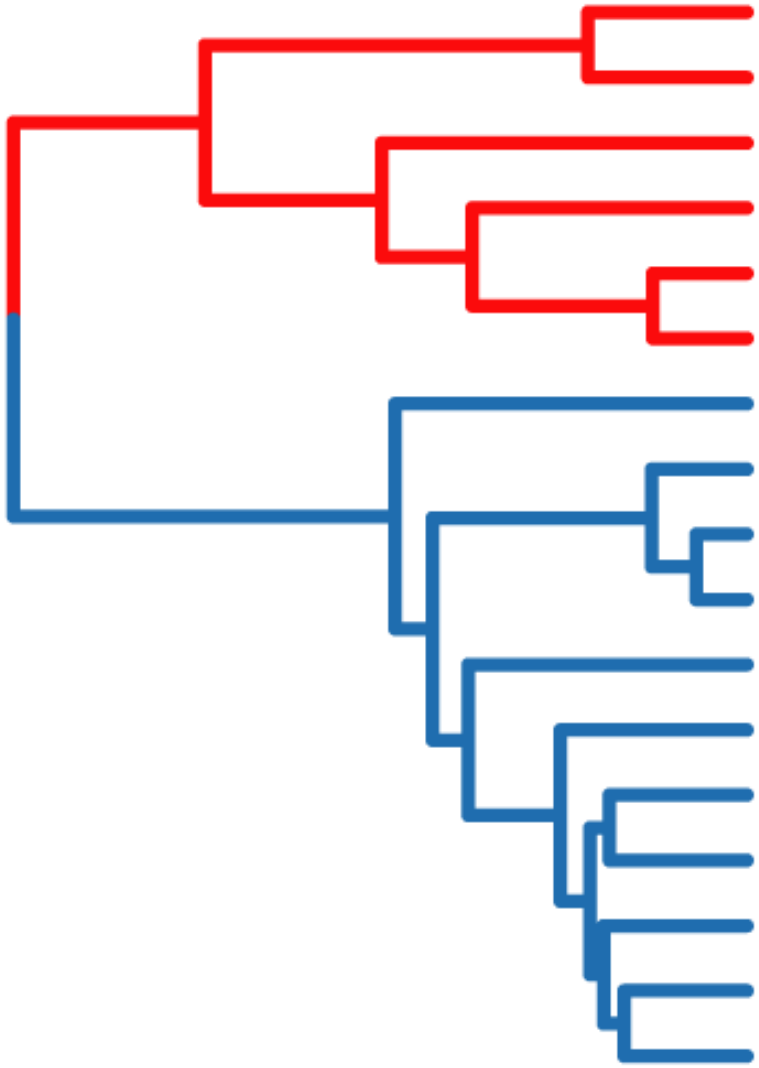
- Relaxed clocks



Multi-rate clocks

- Small number of rates
  - More than 1 rate (i.e., not a strict clock)
  - Fewer than number of branches (i.e., not a relaxed clock)
- Local clock
  - Same rate shared by neighbouring branches
- Discrete clock
  - Small number of branch rates, distributed across tree

Local clocks



Local clocks

- User-defined local clock
  - Fixed tree topology
- Random local clock
  - Each branch has a probability of inheriting rate from ancestor
  - Tree estimated

Discrete clocks

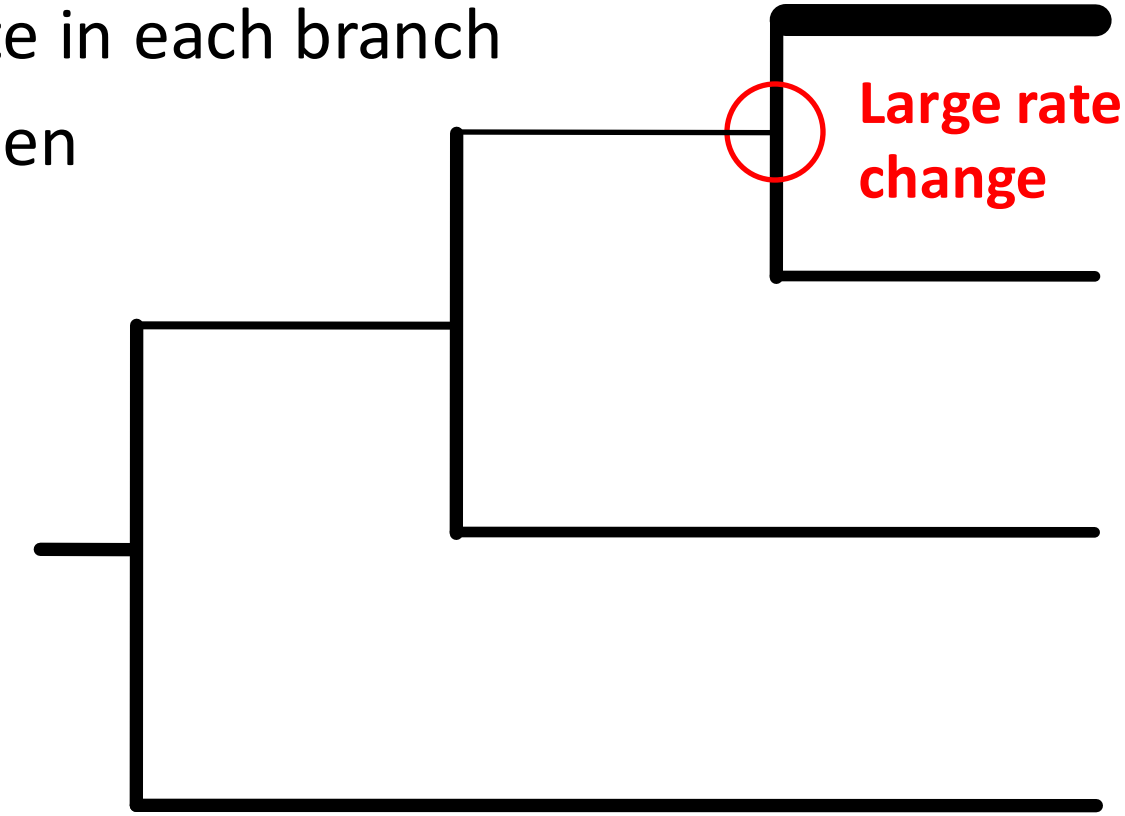
- User-defined discrete clocks
  - Fixed tree topology
- Dirichlet process prior (DPP) model
  - Branch rates drawn from gamma distribution
  - Concentration parameter (alpha) governs number of rate categories and number of branches assigned to each rate category

Relaxed clocks

- We know that life-history characteristics:
  - Have effects on rates of molecular evolution
  - Are usually heritable to some degree
    - Treat molecular rate as a heritable trait
    - Relaxed clocks generally assume that closely related species share similar rates

Relaxed clocks

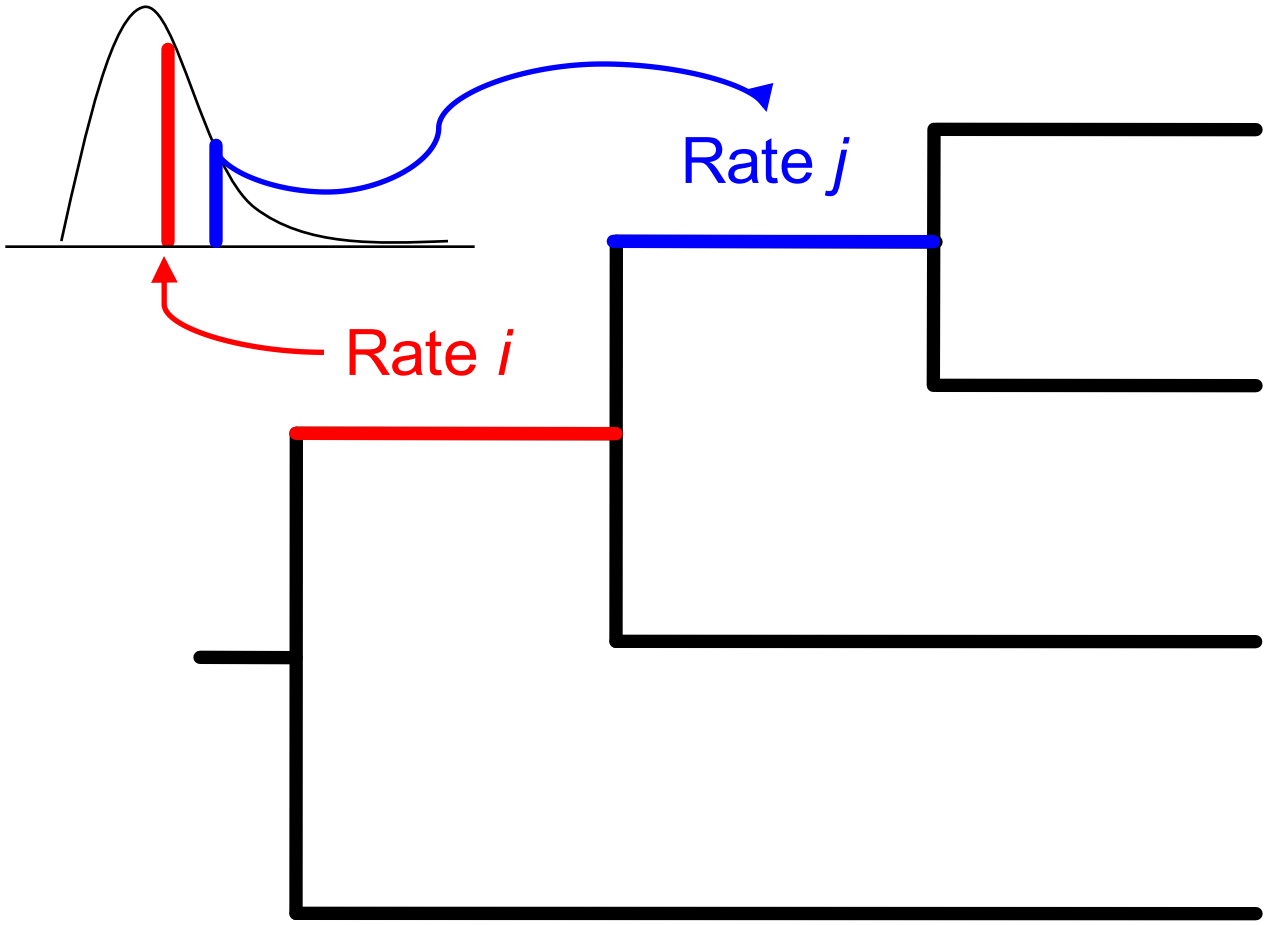
- Allow a different rate in each branch
- Penalise large, sudden changes in rates



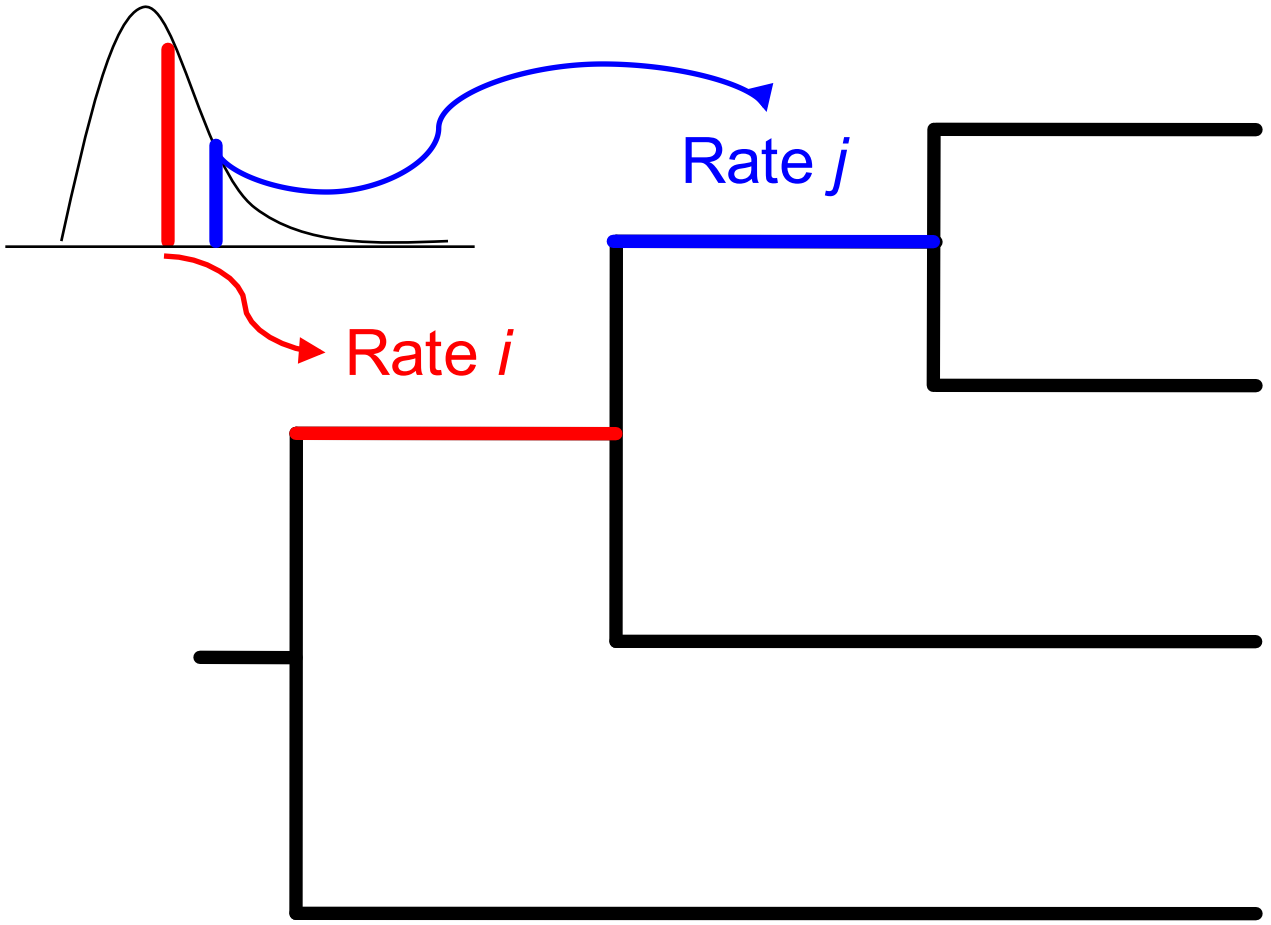
Bayesian relaxed clocks

- Allow a different rate in each branch
- Statistical models of rates among branches
- Rates can be autocorrelated or uncorrelated
  - **Autocorrelated**: rates in neighbouring branches are related
  - **Uncorrelated**: rates identically and independently distributed among branches

Autocorrelated relaxed clocks

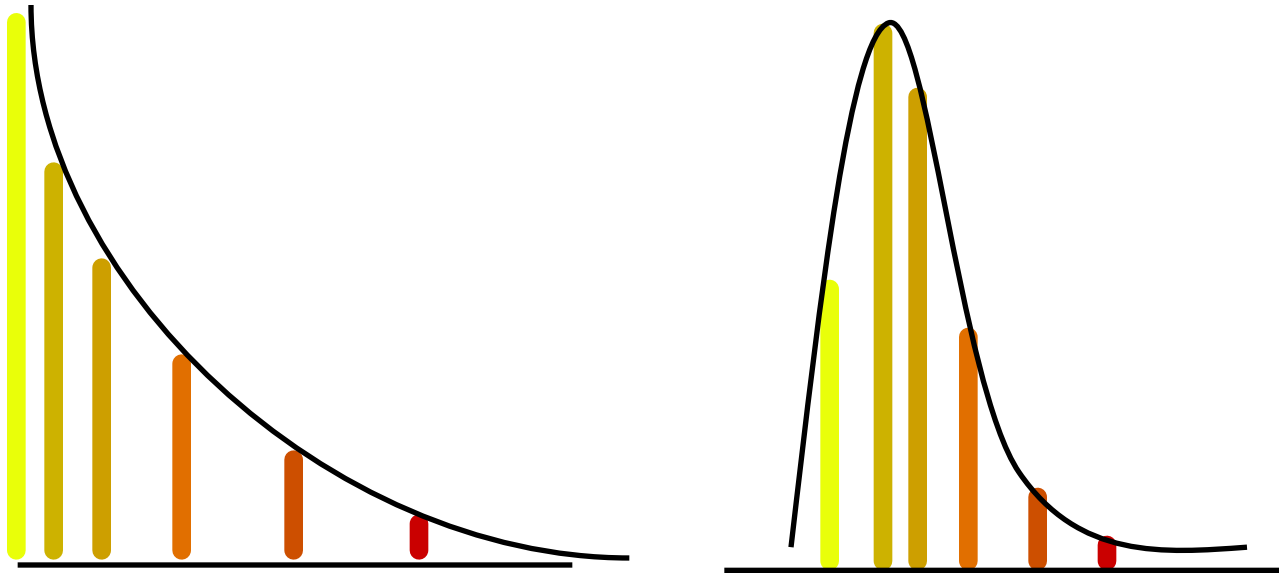


Uncorrelated relaxed clocks



Uncorrelated relaxed clocks

- Models available in BEAST
  - **Exponential distribution**
    Most rates are quite low
  - **Lognormal distribution**
    Most rates cluster around the mean



Lognormal uncorrelated relaxed clocks

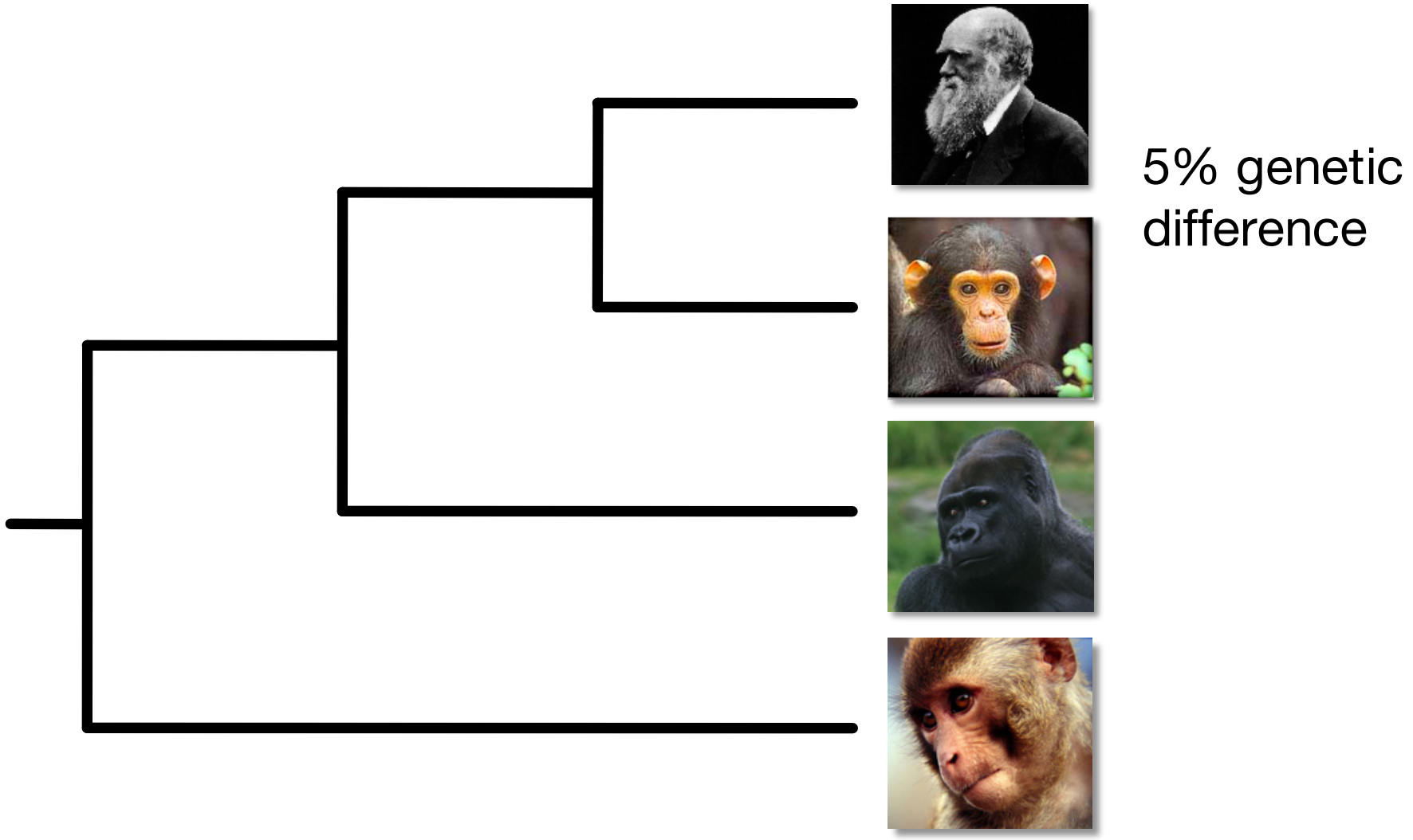
- In the uncorrelated lognormal relaxed clock, two statistics can be obtained:

1. **Coefficient of variation of rates**
   Measures the rate variation among branches
   A value of 0 indicates clocklike evolution

2. **Covariance of rates**
   Measures autocorrelation of rates between adjacent branches

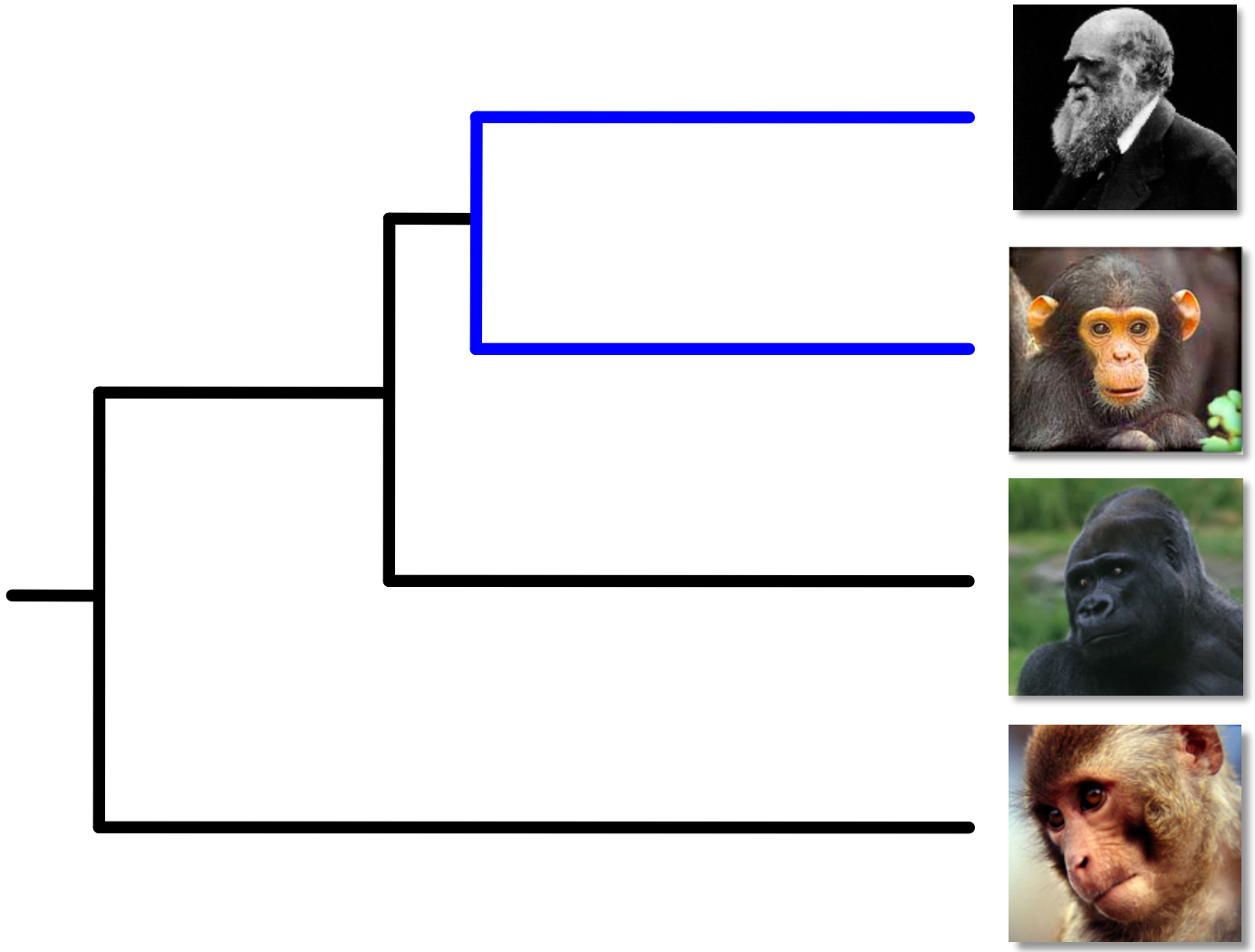
Calibrating the Molecular Clock

Calibrating the molecular clock

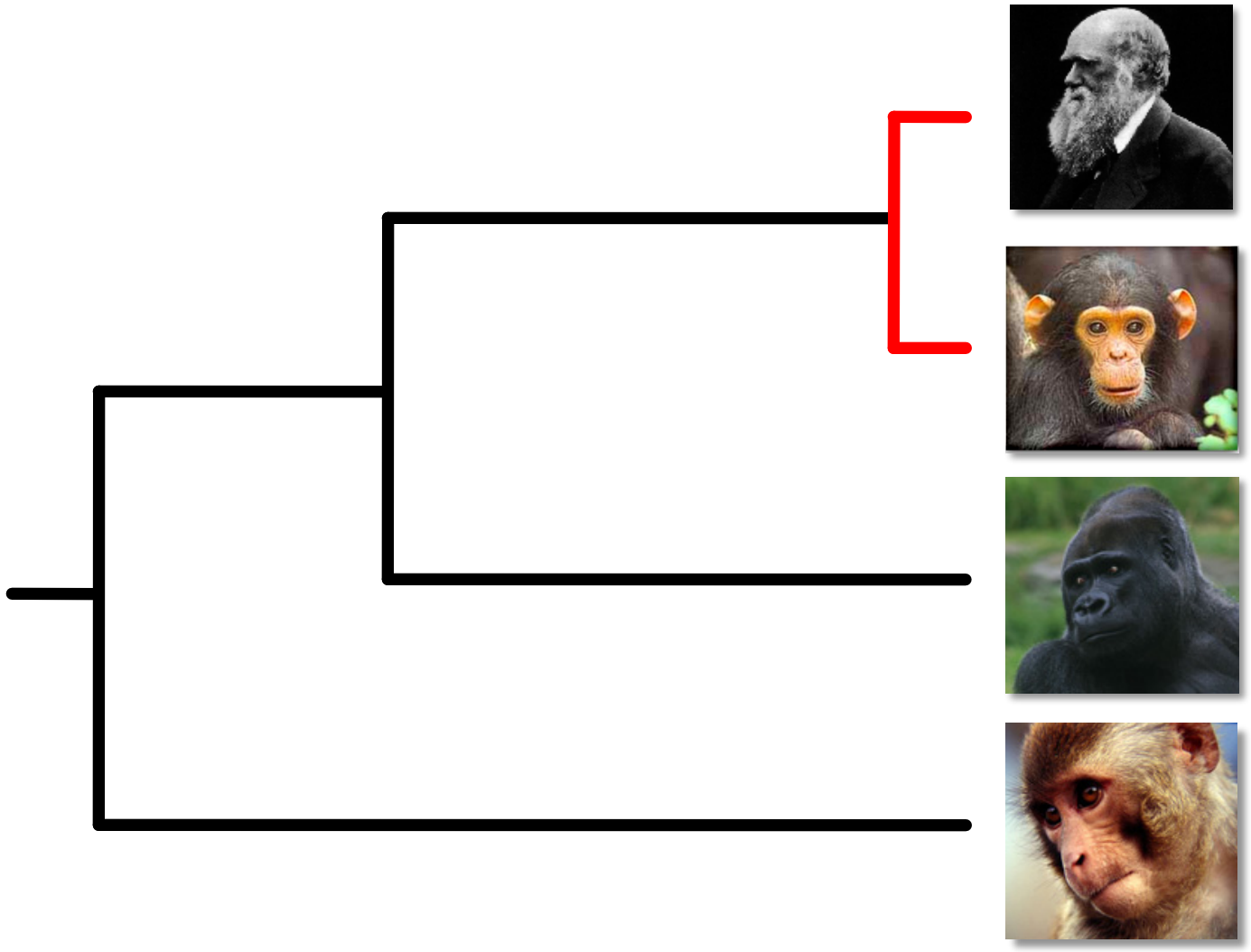


5% genetic difference

Calibrating the molecular clock



Calibrating the molecular clock

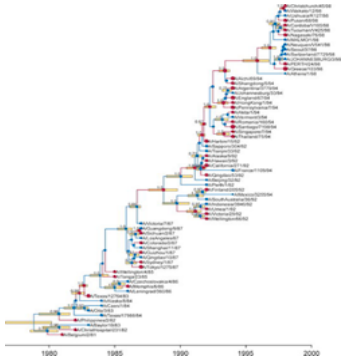


Calibrating information

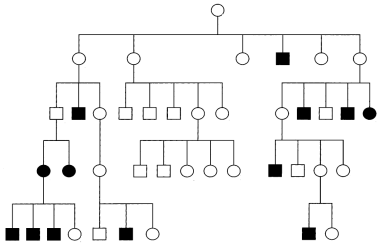
- Information about the rate
  - Substitution rate obtained from an independent study
- Information about time

Fossil record

Biogeography

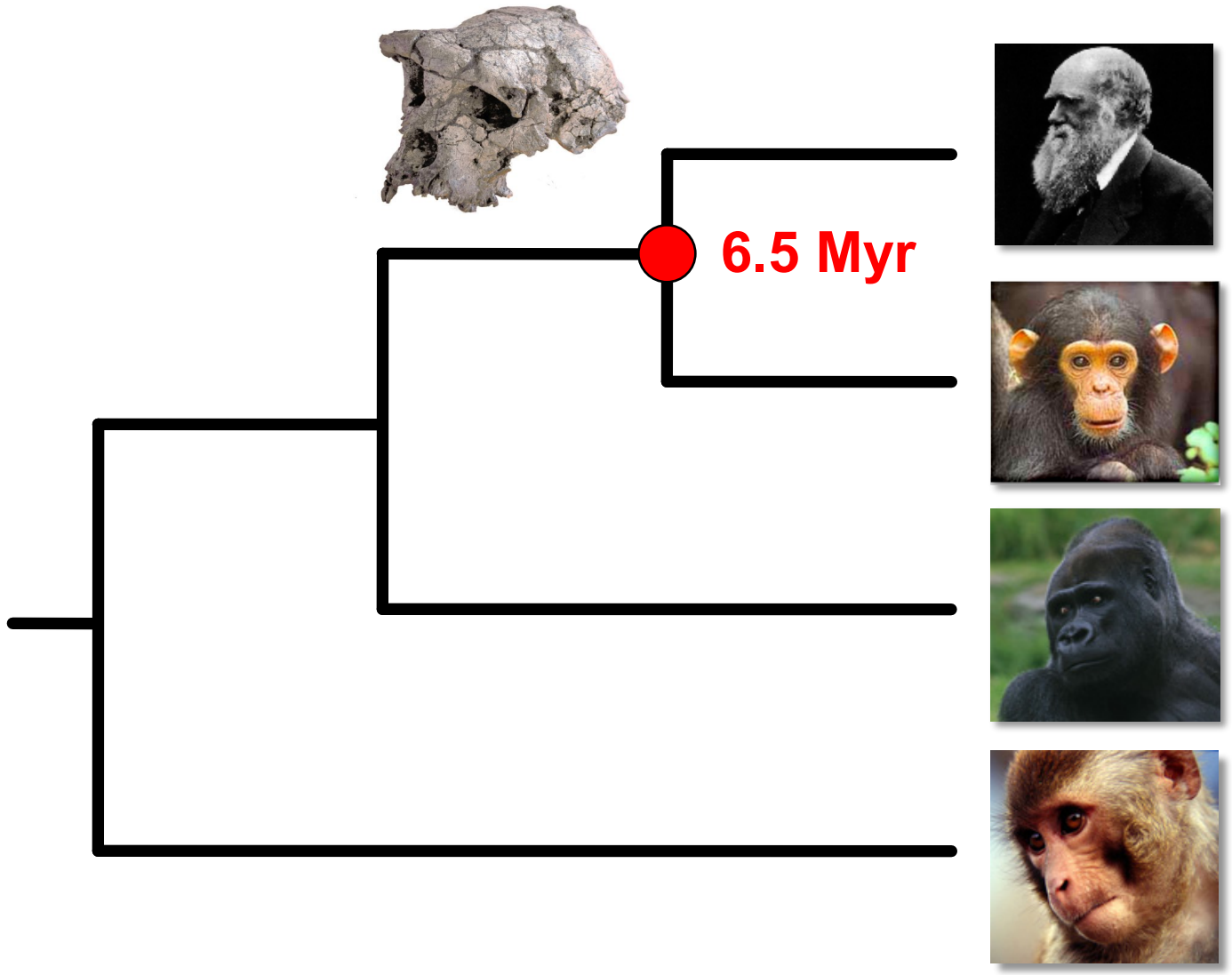


Sampling times



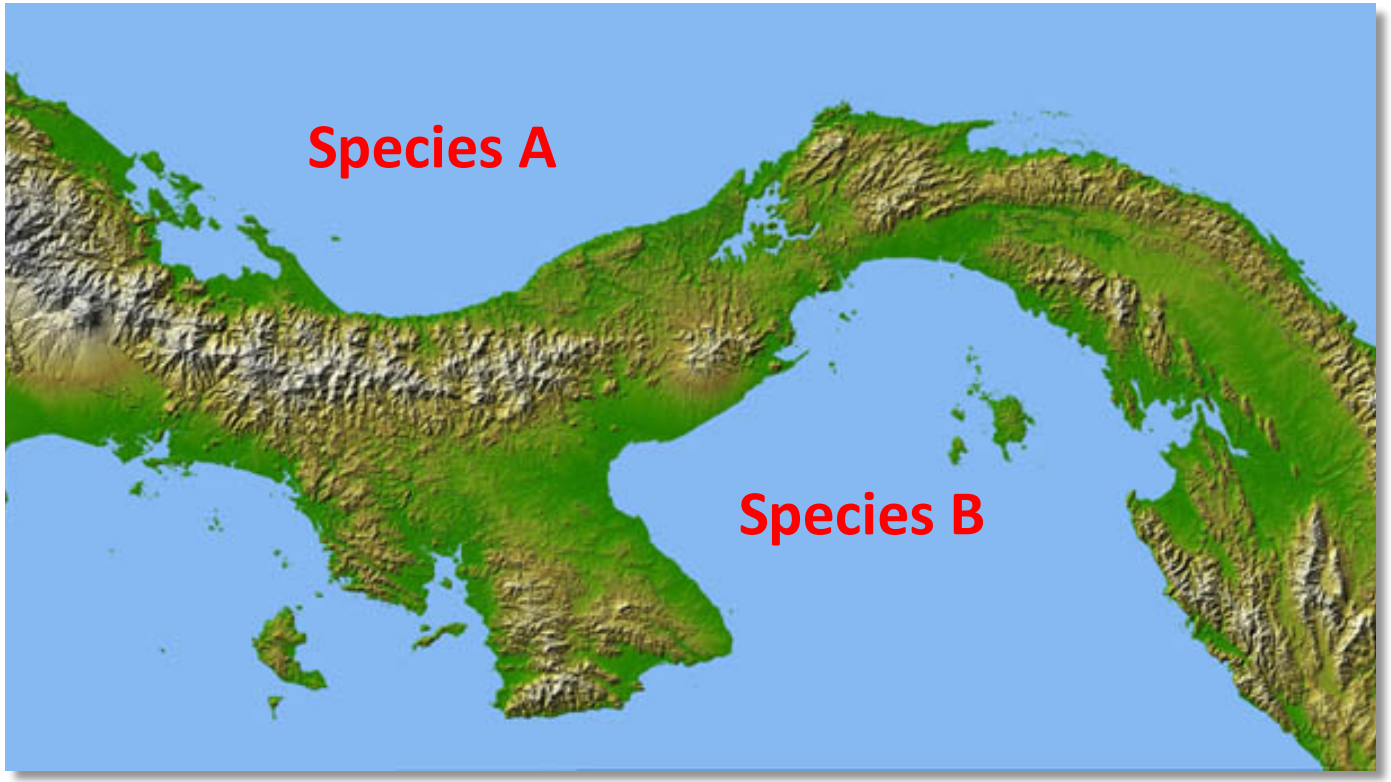
Pedigrees

Calibration: Fossil record

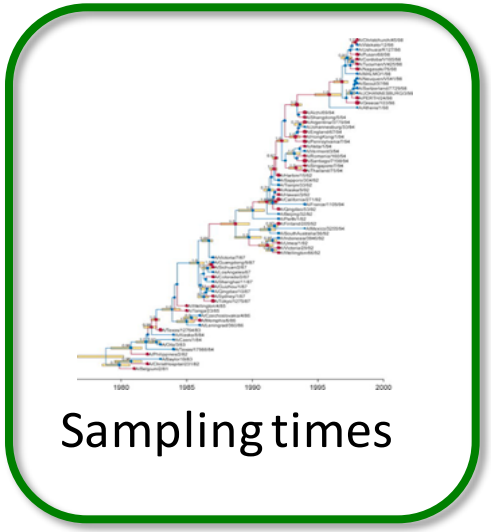


Calibration: Biogeography

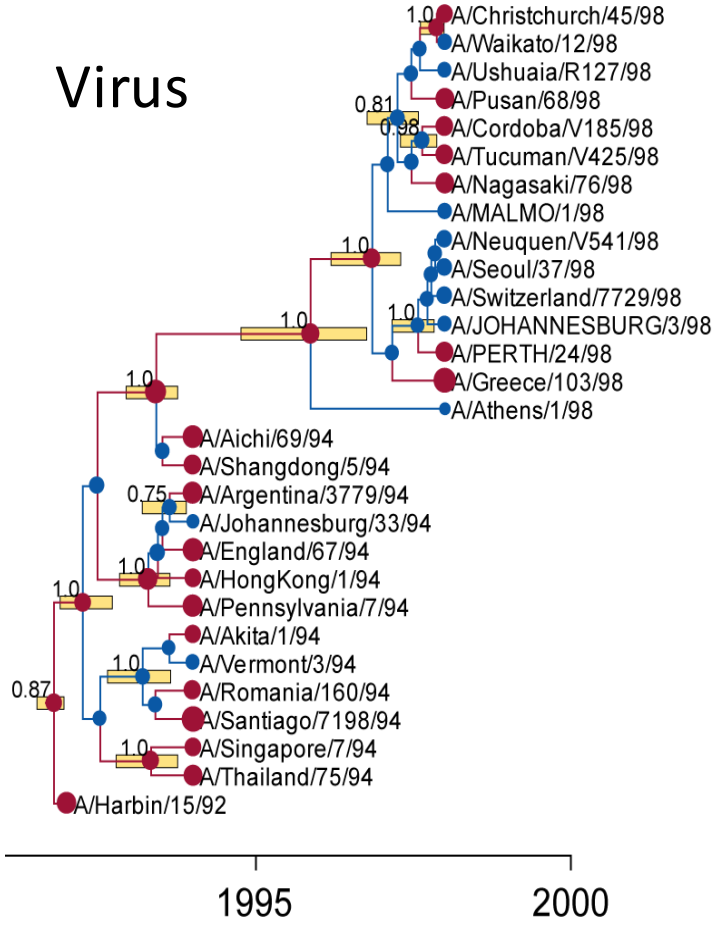
Biogeography



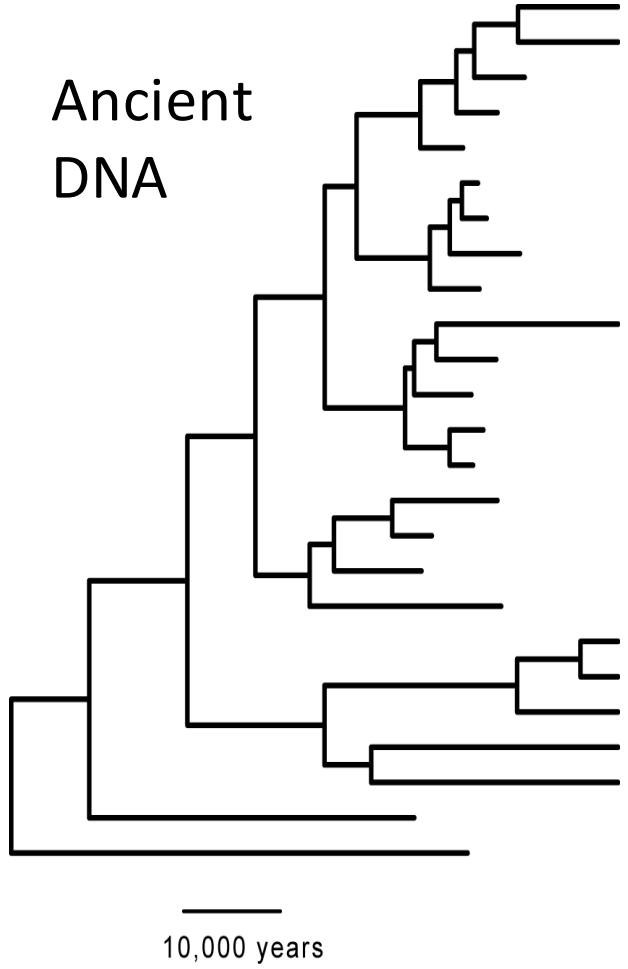
Calibration: Sampling times



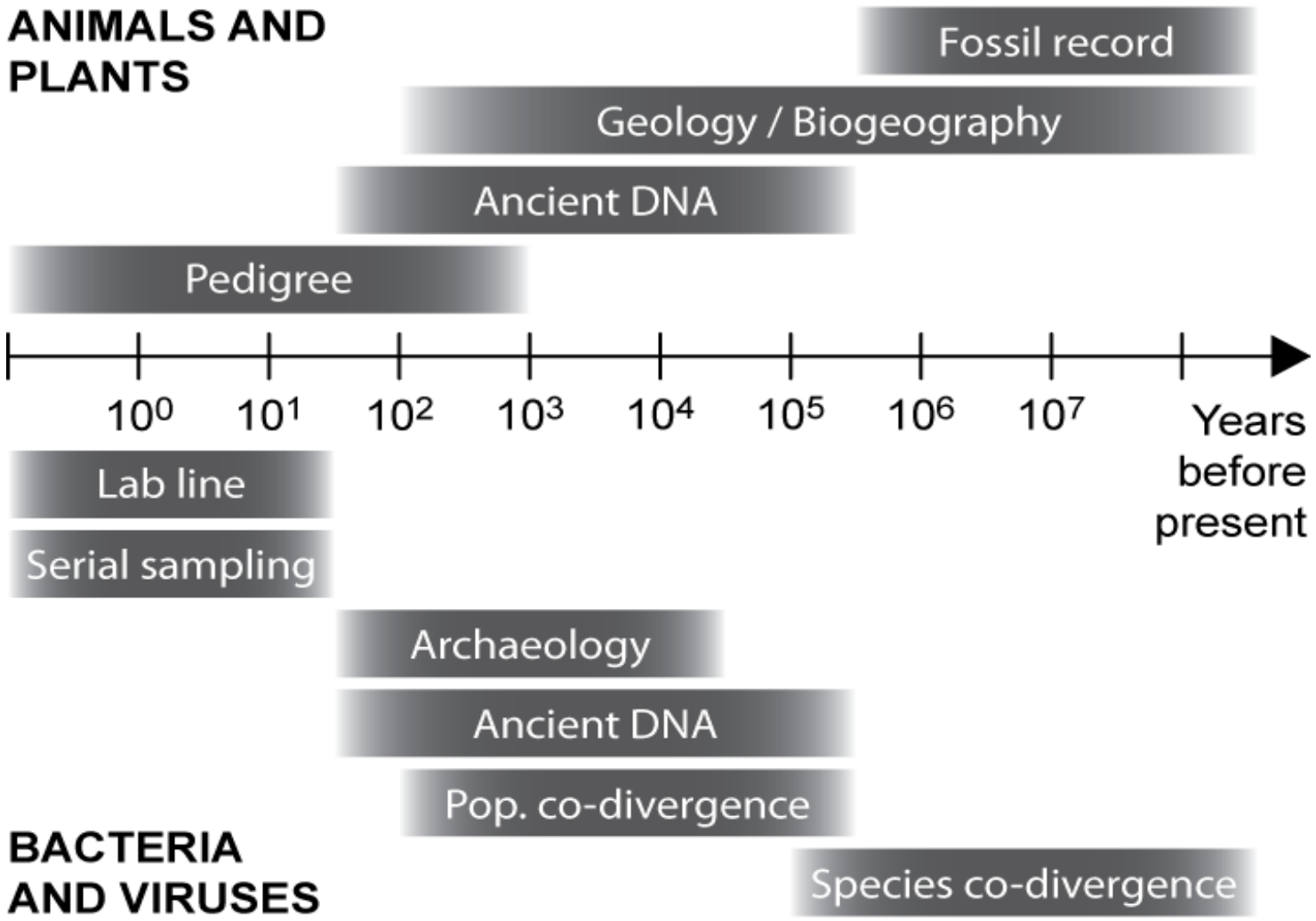
Virus



Ancient DNA

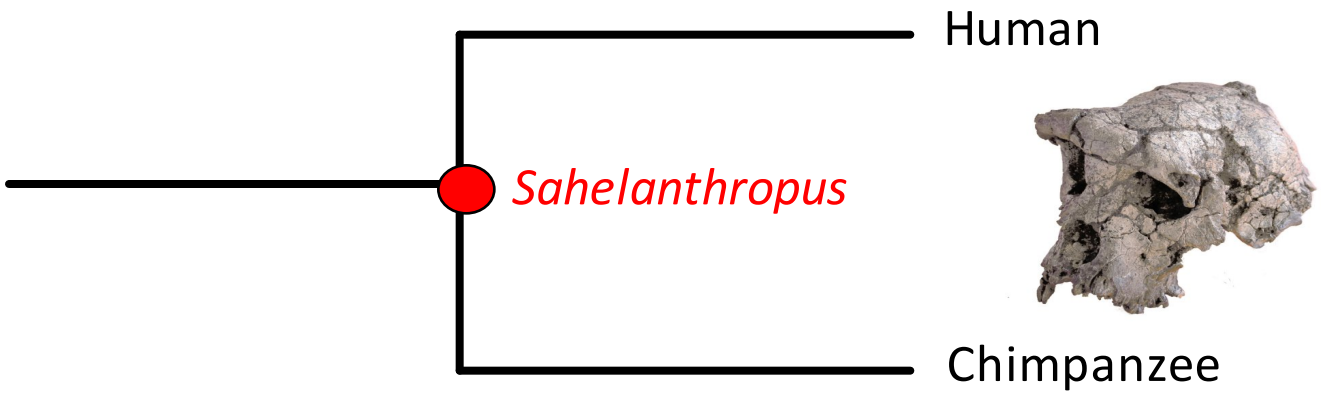


Calibrations



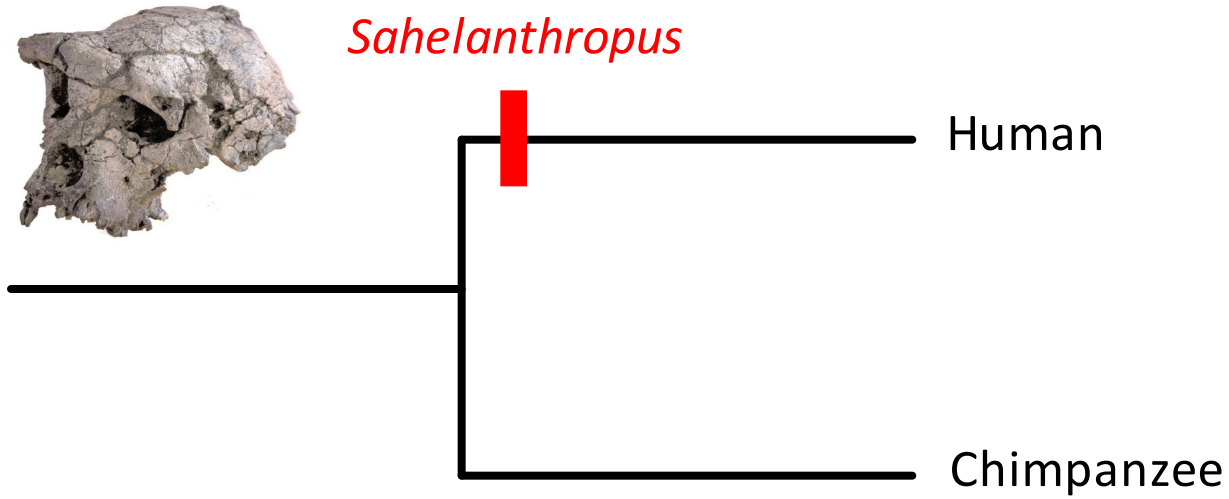
Point calibrations

- Traditional approach
- Artificial precision: ignores the uncertainty arising from preservational biases, isotopic dating errors, *etc.*



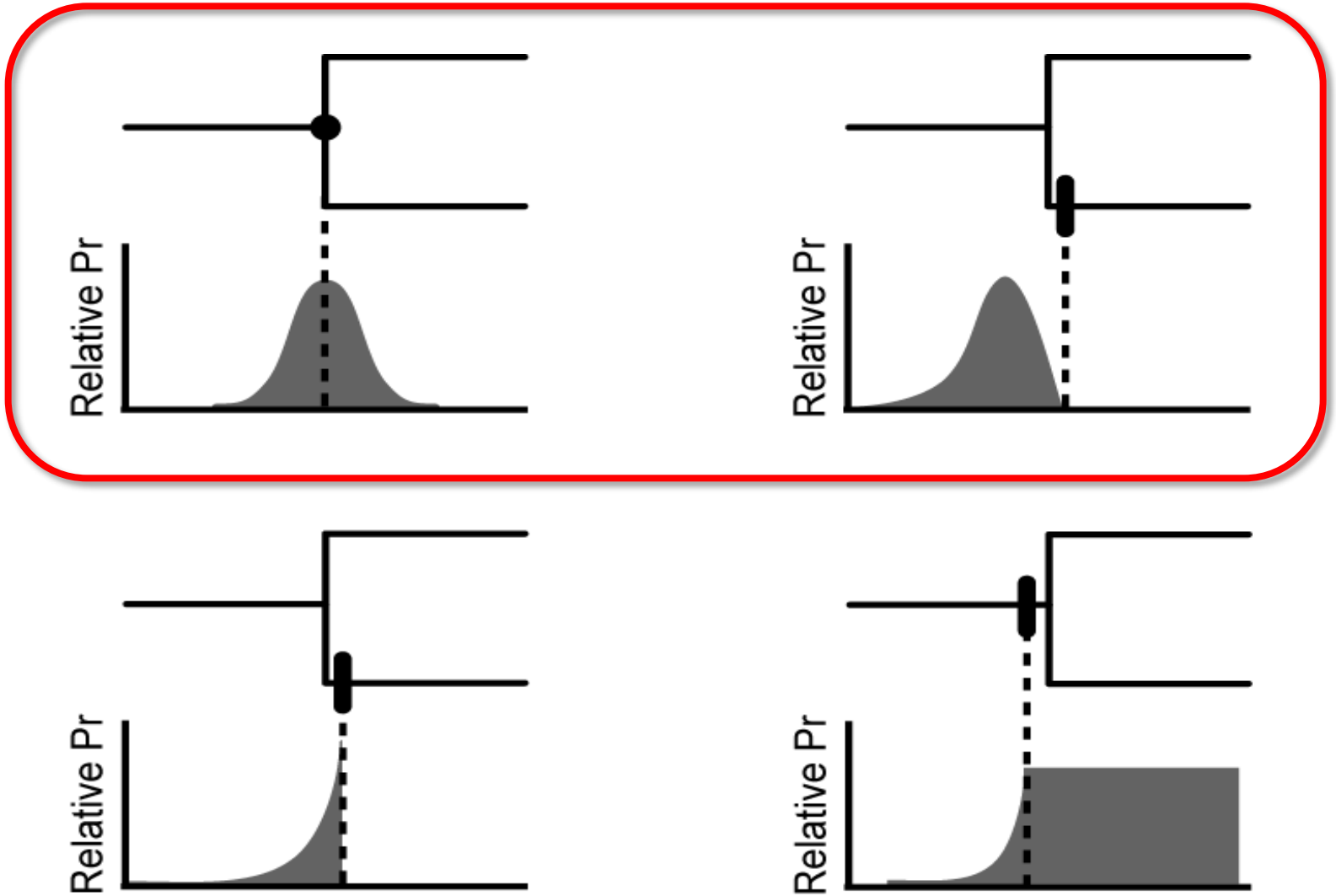
Hard calibration bounds

- Minimum or maximum age constraints



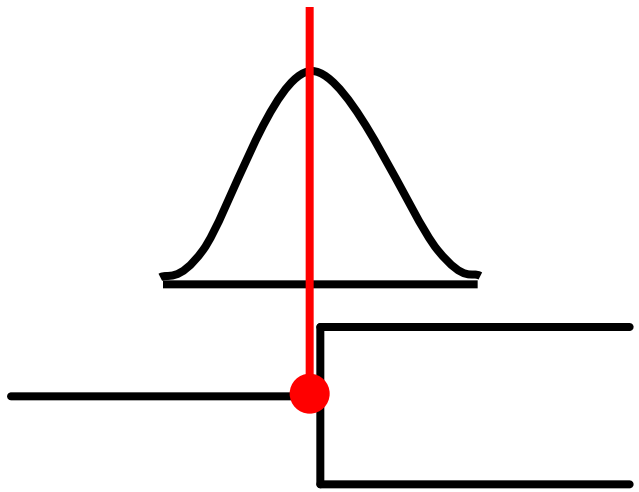
  - Discards potentially useful information
  - Inadequate information for estimating divergence times

Parametric prior distribution



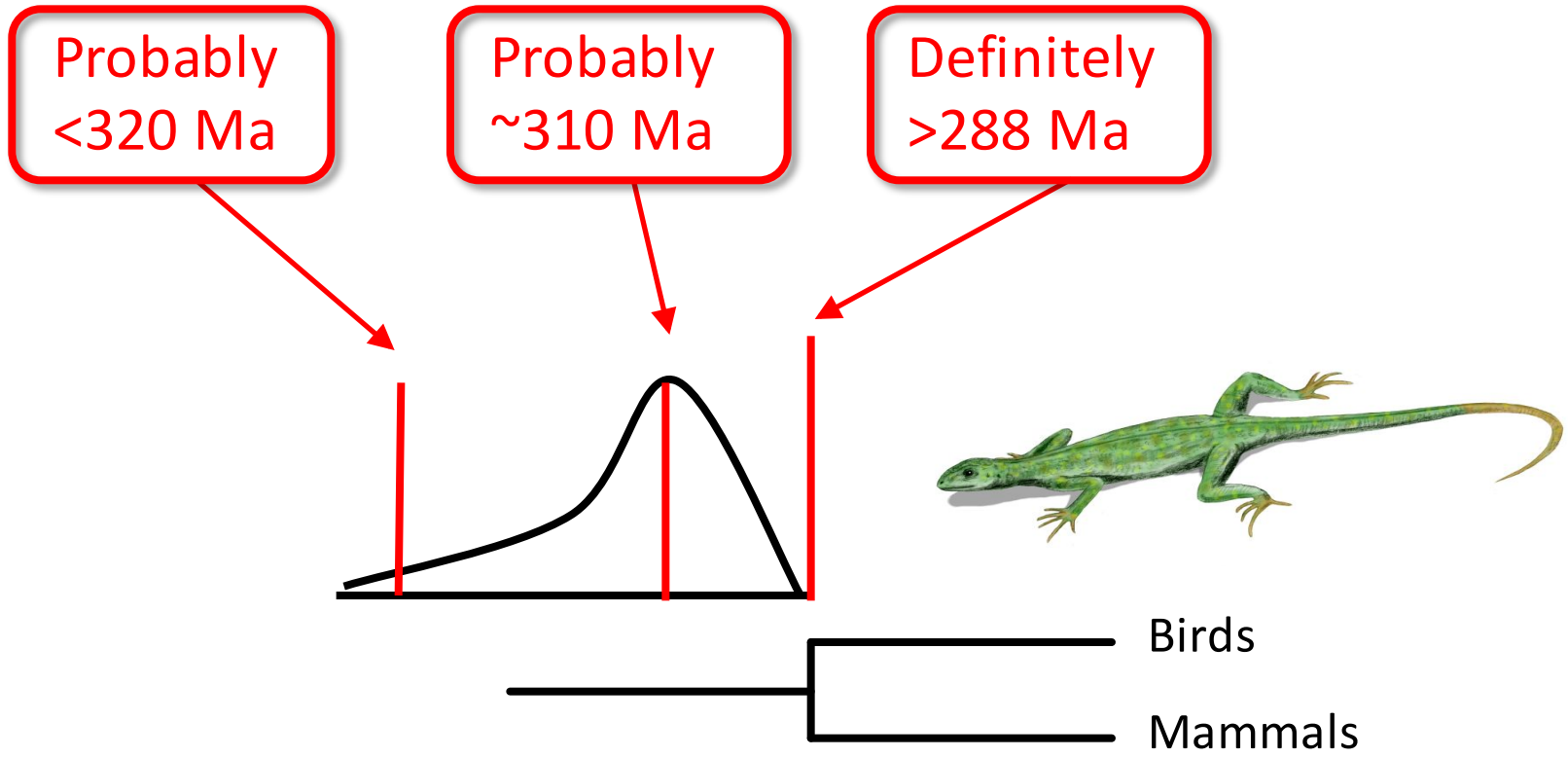
Normal prior

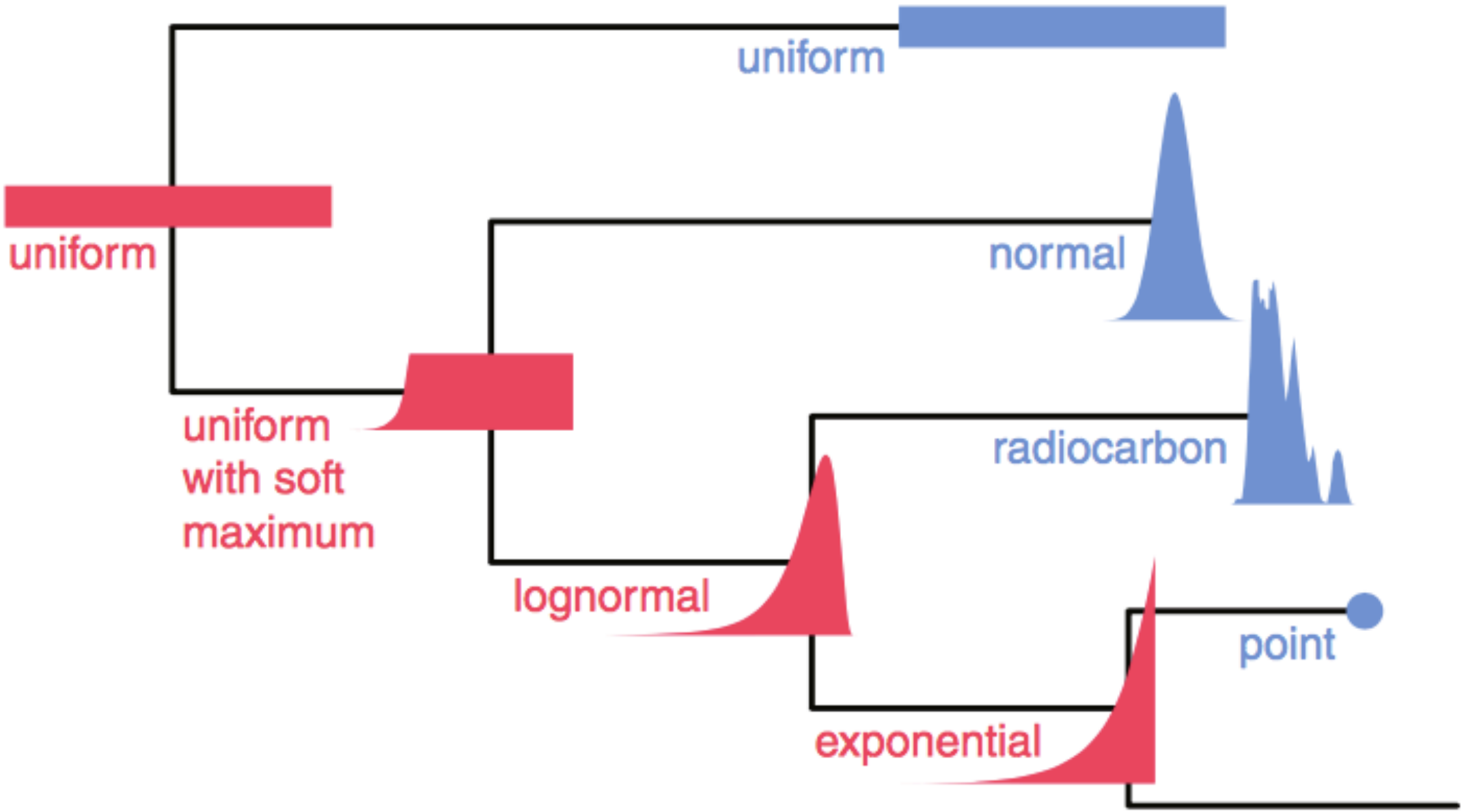
- Two parameters (mean, s.d.)
- Useful for describing secondary calibrations
- Can be used for some biogeographic and fossil calibrations



Lognormal prior

- Two parameters (mean, s.d.) and an offset (minimum value)
- Example: bird-mammal split



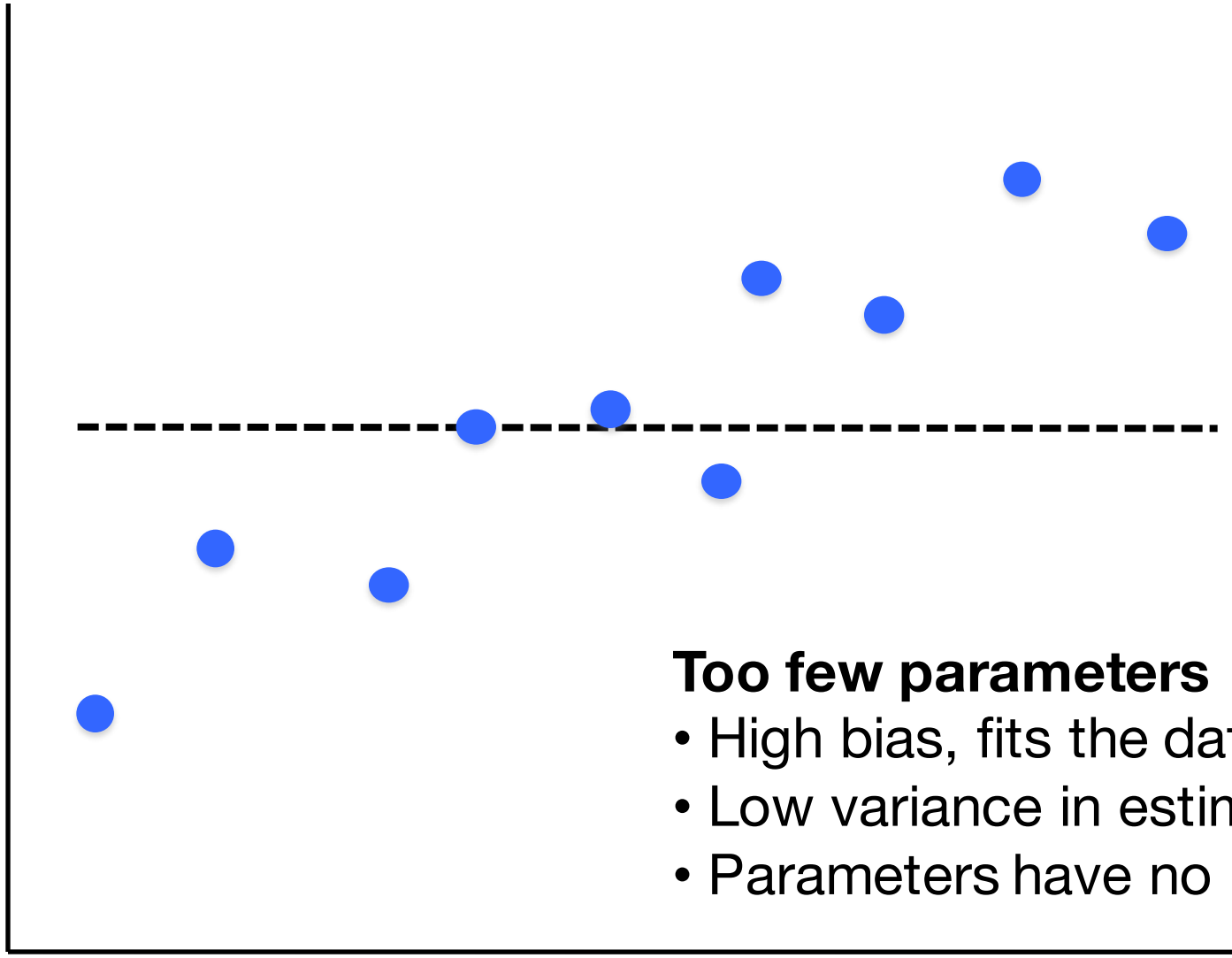

uniform
uniform
normal
uniform
with soft
maximum
radiocarbon
lognormal
point
exponential

Choosing calibrations

- The age estimates for poorly supported clades should be interpreted carefully
- Interactions among calibrations can be inspected by running the analysis without data (sampling directly from the joint prior)
- Careful selection of clock models can improve the estimates

Model Selection

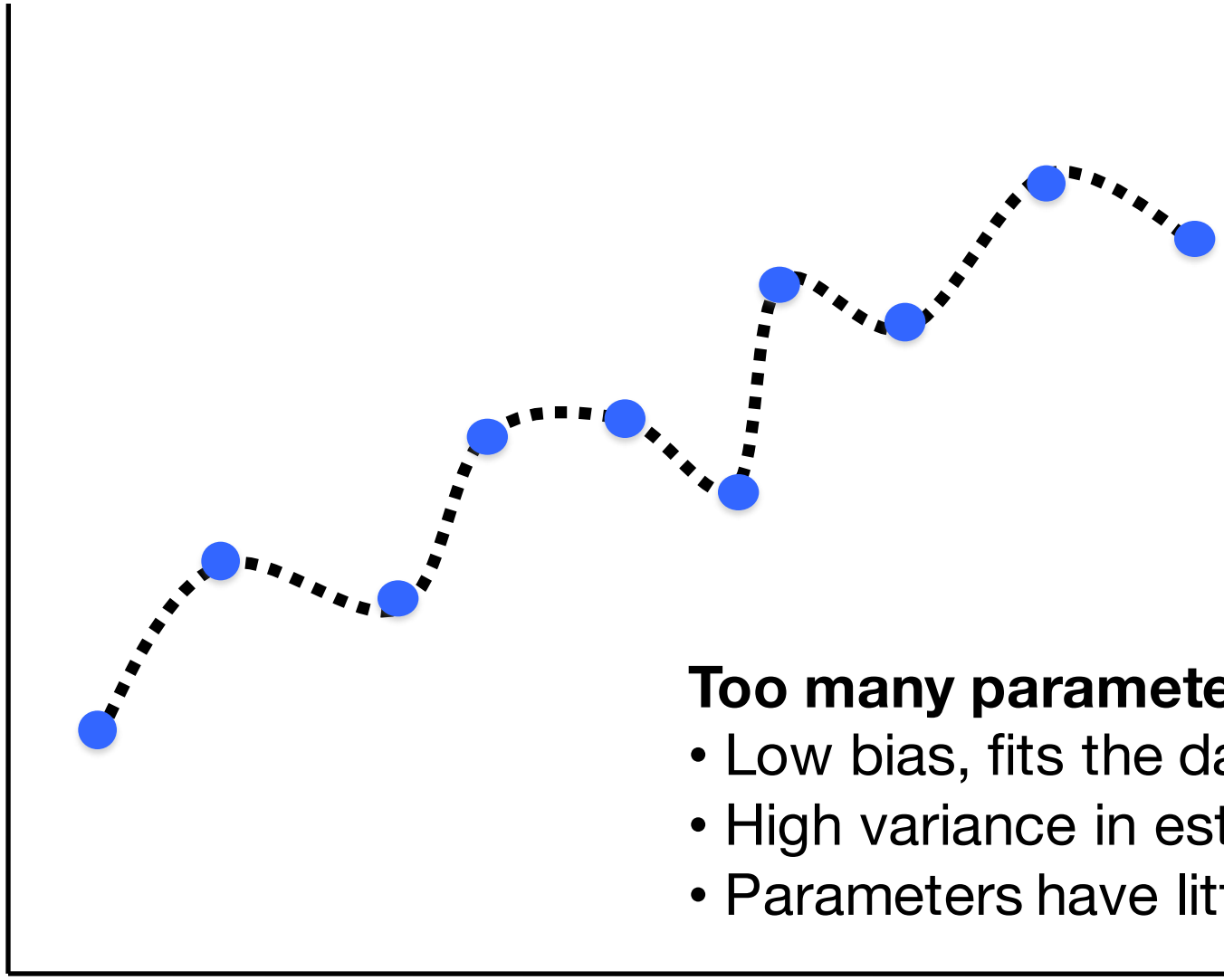
Model selection



Too few parameters

- High bias, fits the data poorly
- Low variance in estimate
- Parameters have no biological meaning

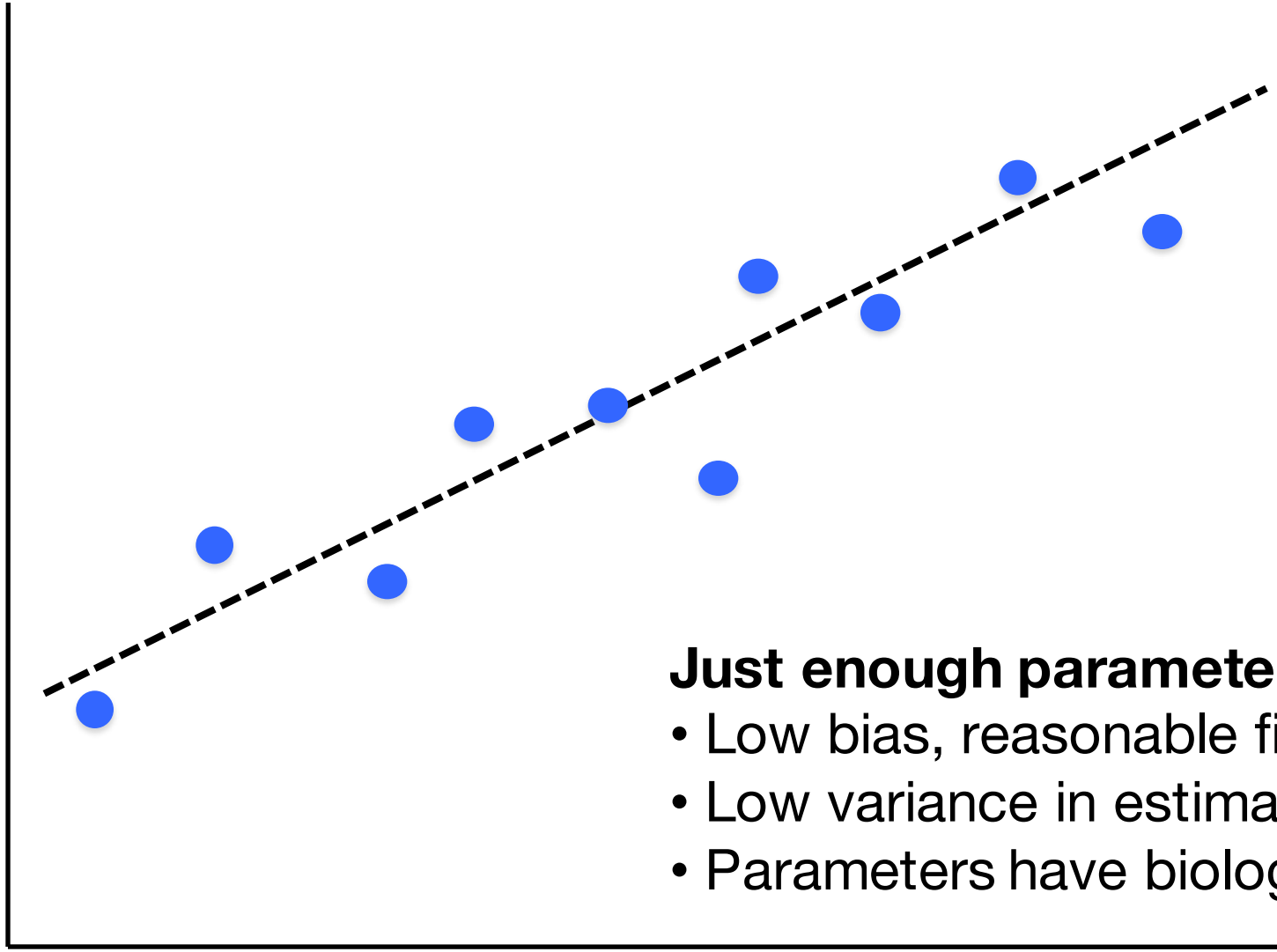
Model selection



Too many parameters

- Low bias, fits the data very well
- High variance in estimates
- Parameters have little biological meaning

Model selection

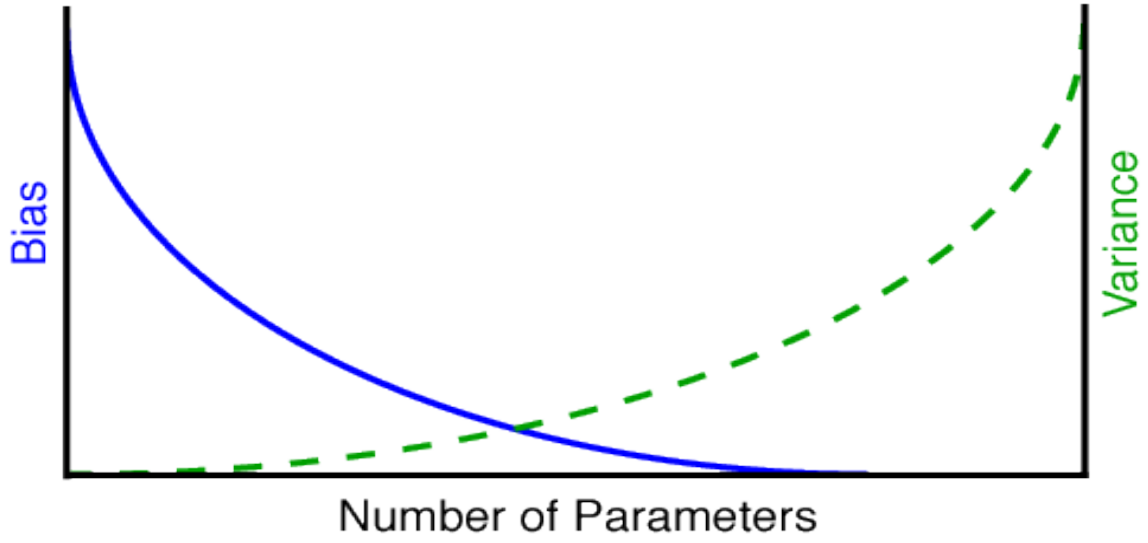


Just enough parameters

- Low bias, reasonable fit to data
- Low variance in estimates
- Parameters have biological meaning

Model selection

- Adding more parameters *always* improves the fit of the model to the observed data
- But more parameters leads to greater variance in the estimates of those parameters
- Goal is to find the best balance between bias and variance



Model selection

- Adding a parameter to the model:
  - Is the improvement in likelihood worth the cost of adding a parameter?

- Model selection methods
  - **Likelihood-ratio test (LRT)**
    Used to compare nested models
  - **Akaike information criterion (AIC)**
    AIC = -2ln(likelihood) + k
  - **Bayesian information criterion (BIC)**

Bayesian model selection

- Bayesian model selection is usually based on the marginal likelihood of the data, conditioned on the model:

$$Pr(D|M)$$

- This is the weighted average of the likelihood.
- Weights are given by the prior distribution (proper priors).

Marginal likelihood of the model

Bayesian model selection

- Compare marginal likelihoods of competing models.

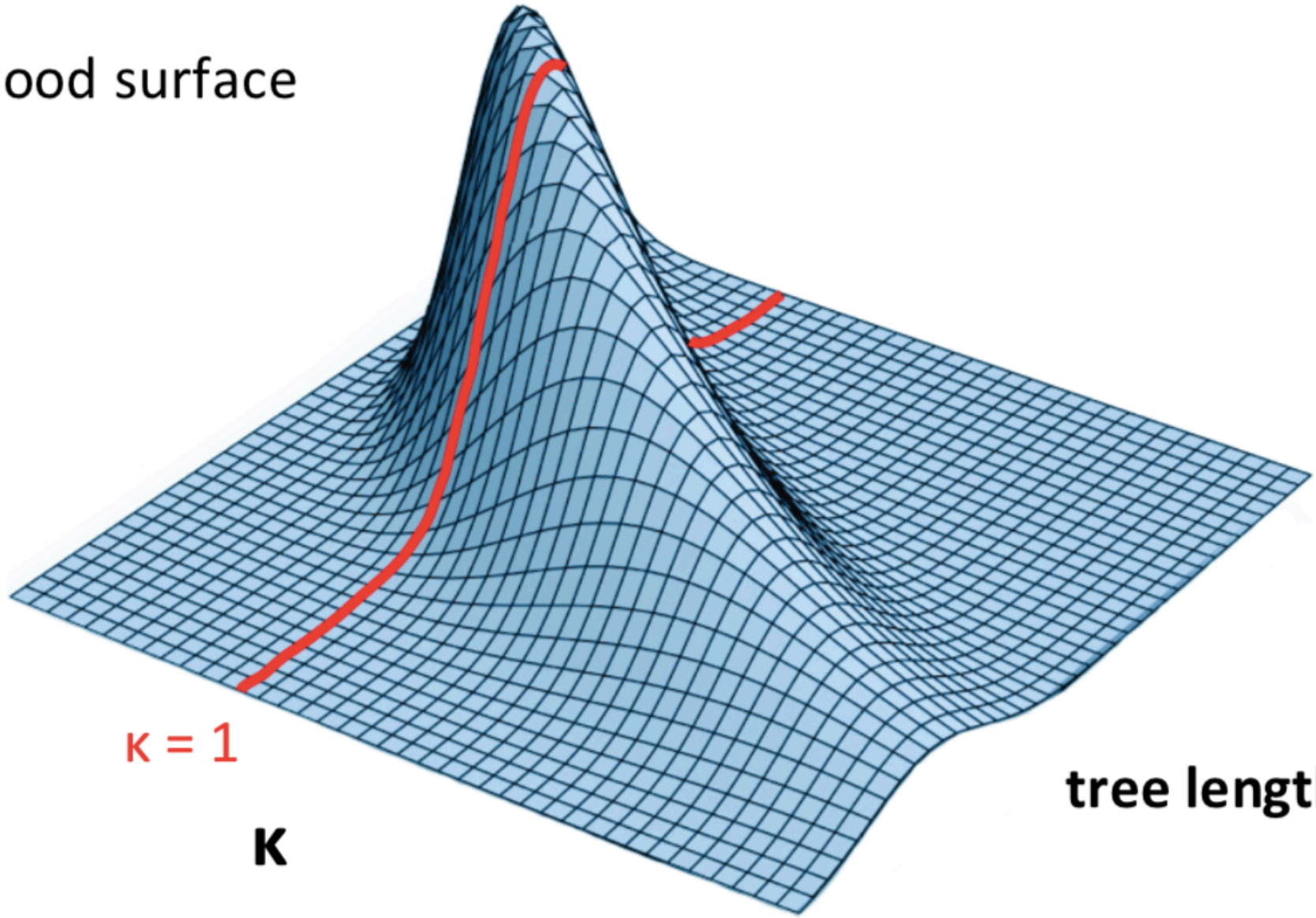
- Ratio of the marginal likelihoods is the **Bayes factor**.

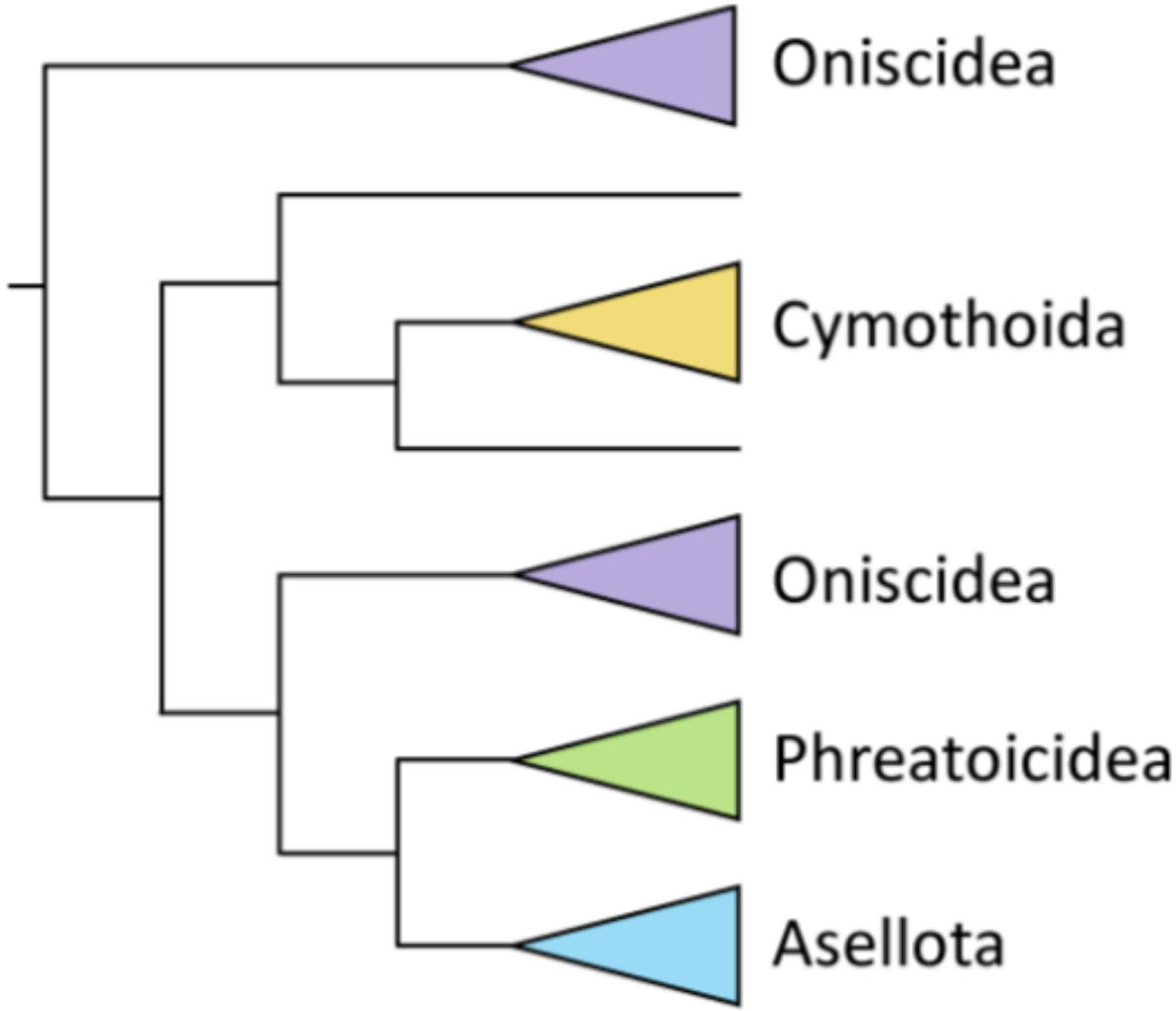
$$BF = \frac{\Pr(D|M1)}{\Pr(D|M2)}$$

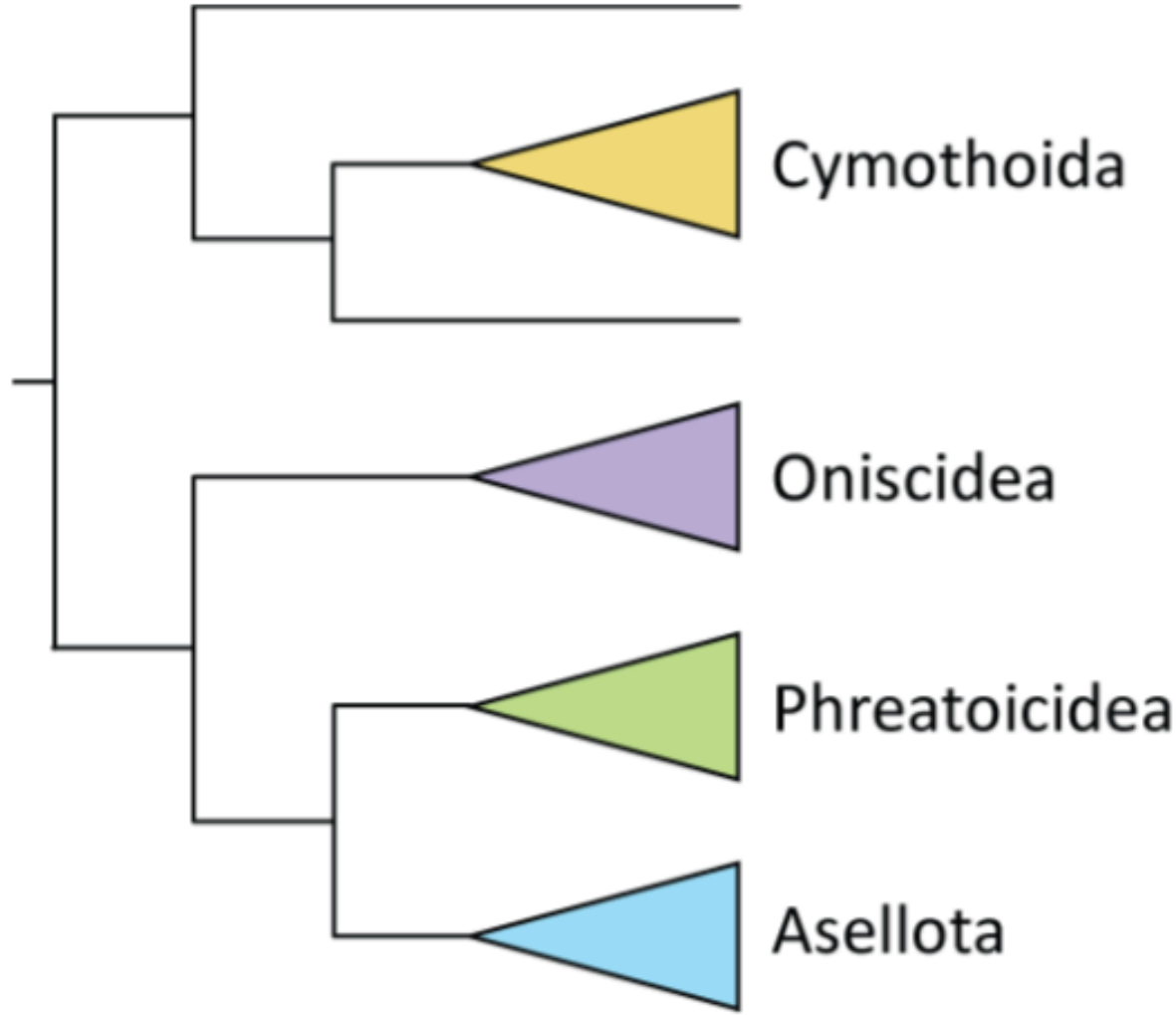
$$logBF = \log \Pr(D|M1) - \log \Pr(D|M)$$

- Models do not need to be nested.
- Do not need to correct for the number of parameters.
- **Priors need to be proper**

Likelihood surface







marginal log*L* = -13085

marginal log*L* = -13089

log*BF* = 4

Bayesian model selection

- Interpreting Bayes factor

BF	log*BF*	Evidence against M_2
1 – 3	0 – 1	Not worth mentioning
3 – 20	1 – 3	Positive
20 – 150	3 – 5	Strong
> 150	> 5	Very strong

Kass & Raftery (1995) *J Am Stat Assoc*

Estimating the marginal likelihood

Infinite variance

- Harmonic mean estimator
  - Calculated from likelihood values sampled from the MCMC.
  - Easy to calculate from output of Bayesian analysis.

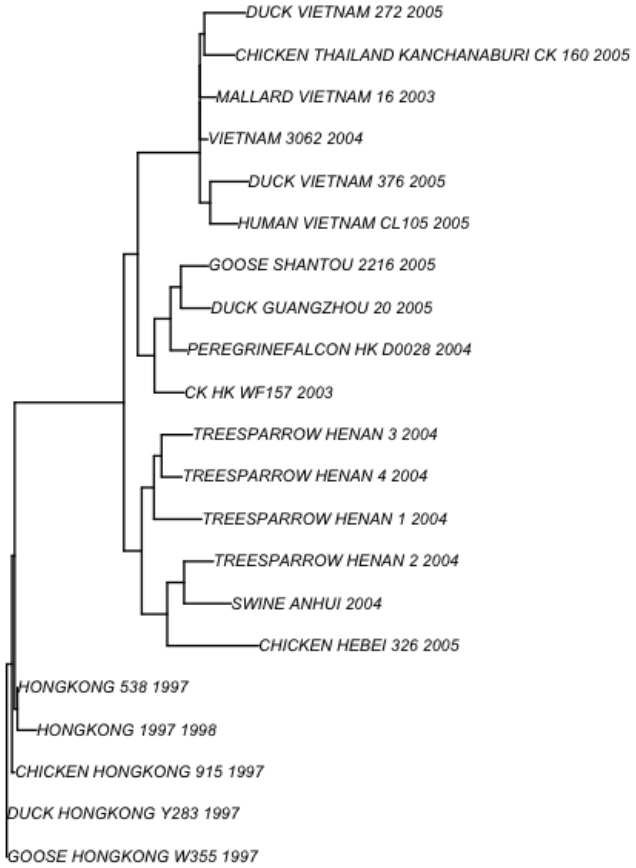
Computationally intensive

- Better methods
  - Path sampling (thermodynamic integration).
  - Stepping stone sampling.
  - Generalised stepping stone sampling.

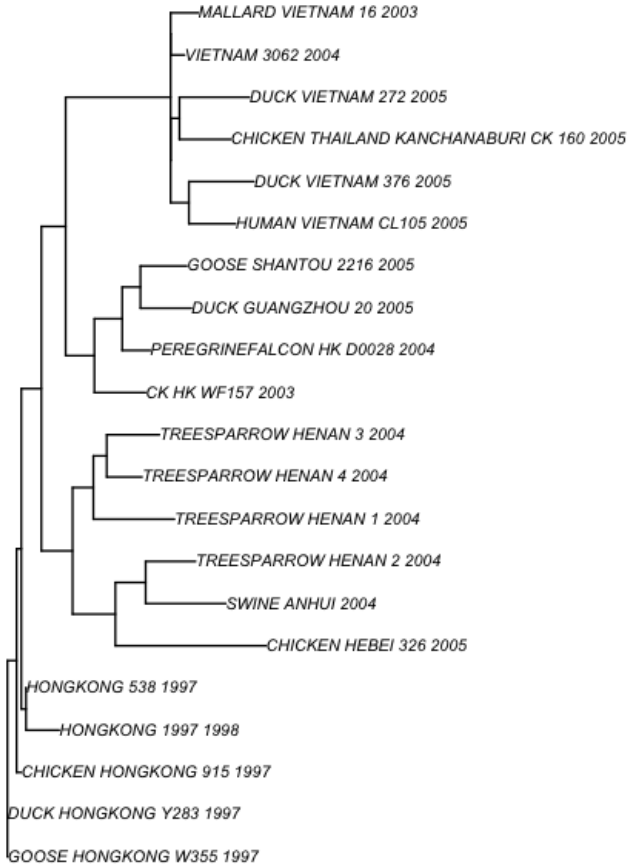
Model averaging

- Maximum likelihood methods
  - Score a set of models (e.g. AIC).
  - Average estimates using model weights.

HKY+G
AIC=8725.4

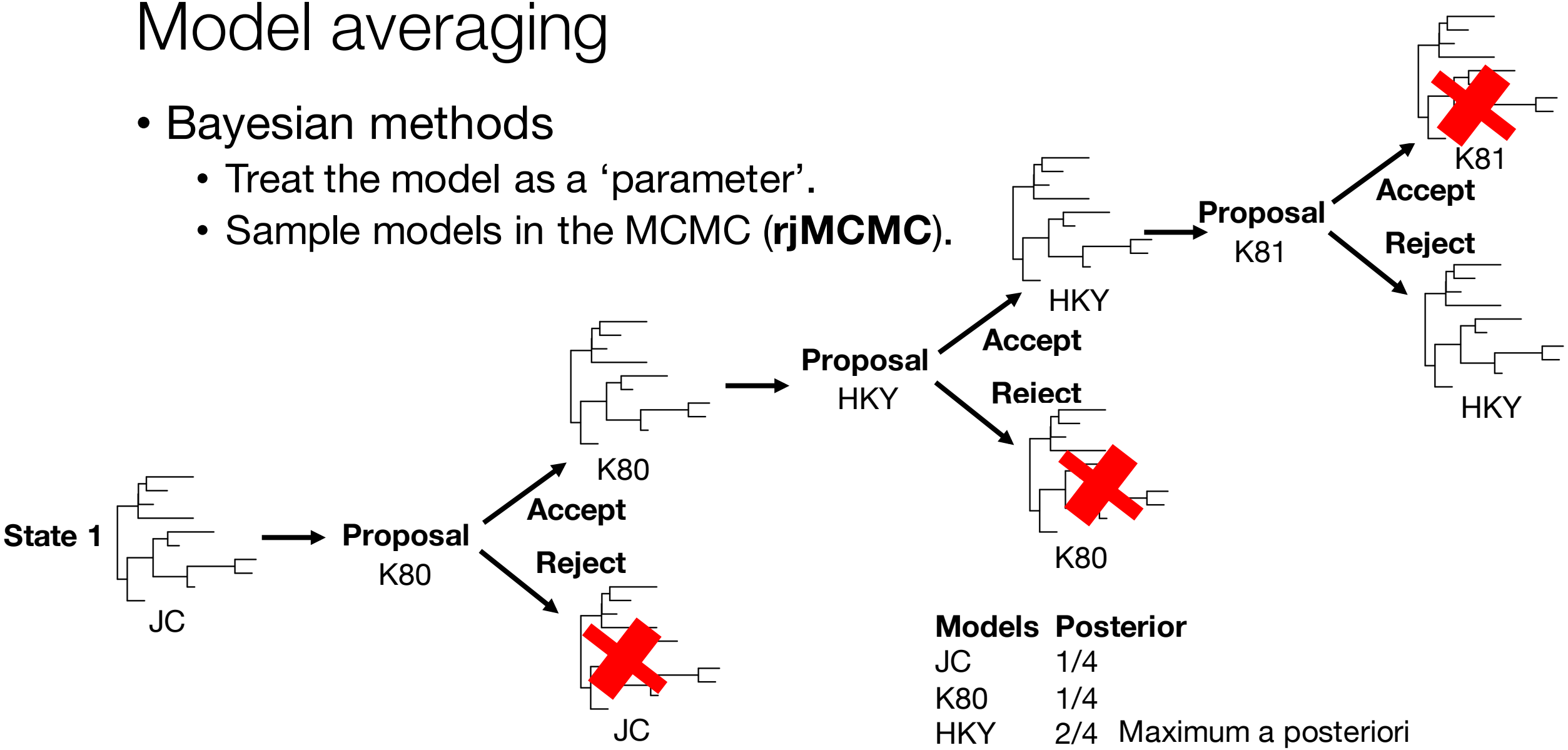


HKY
AIC=8825.4



Model averaging

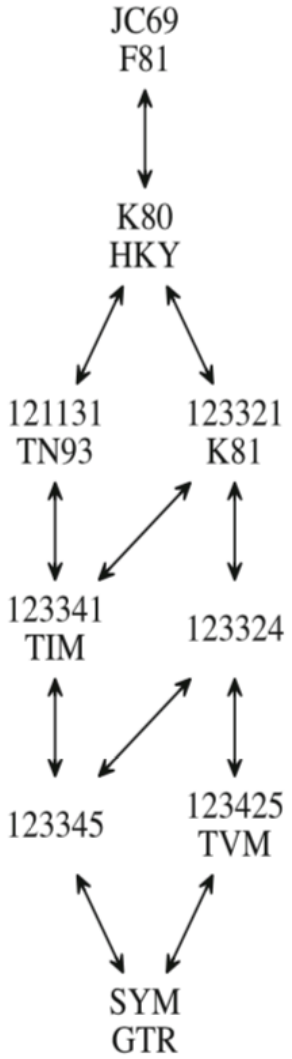
- Bayesian methods
  - Treat the model as a ‘parameter’.
  - Sample models in the MCMC (**rjMCMC**).



Models	Posterior	
JC	1/4	
K80	1/4	
HKY	2/4	Maximum a posteriori (MAP) model

Model averaging

- The MCMC needs to visit models with different parameters.
- Use reversible jump MCMC (rjMCMC).
- Moving in model space requires special operators and proposal probabilities.



Standard Metropolis-Hastings ratio:

$$\frac{\Pr(M,\theta',\tau',v'|data)}{\Pr(M,\theta,\tau,v|data)}$$

, where ' indicates a proposed state.

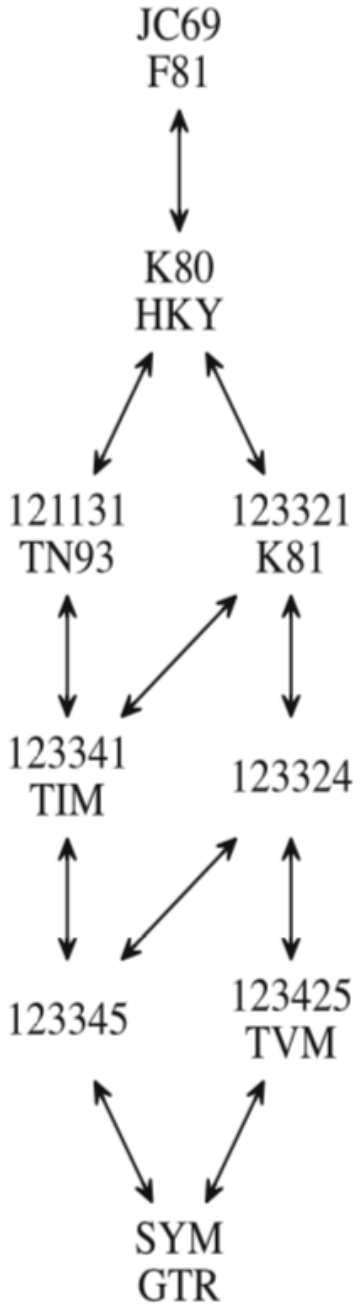
Model averaging

- The MCMC needs to visit models with different parameters.
- Use reversible jump MCMC (rjMCMC).
- Moving in model space requires special operators and proposal probabilities.

Reversible jump acceptance probability=

Posterior ratio * proposal ratio * Jacobian

$$\frac{\Pr(M', \theta', \tau', \nu' | data)}{\Pr(M, \theta, \tau, \upsilon | data)} * \frac{\Pr(M \to M')}{\Pr(M' \to M)} * Jacobian$$



Model averaging

- Assess support of models depending on their posterior probability (i.e. the proportion of times they were visited in the MCMC).
- No need to select a single model.
- Some estimates (e.g. trees and branch lengths) are averaged estimates over models.
- Parameters that are unique to a model cannot be averaged (e.g. shape of Gamma distribution).

References

- Posada, D., & Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic biology*, *53*(5), 793-808.
- Bromham, L., Duchêne, S., Hua, X., Ritchie, A. M., Duchêne, D. A., & Ho, S. Y. (2018). Bayesian molecular dating: opening up the black box. *Biological Reviews*, *93*(2), 1165-1191.
- Huelsenbeck, J. P. et al. (2004). Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular biology and evolution*, *21*(6), 1123-1133.
- Baele, G. et al. (2015). Genealogical working distributions for Bayesian model testing with phylogenetic uncertainty. *Systematic biology*, *65*(2), 250-264.
- Chen, M. H., Kuo, L., & Lewis, P. O. (Eds.). (2014). *Bayesian Phylogenetics: methods, algorithms, and applications*. CRC Press.
- Bouckaert, R. R., & Drummond, A. J. (2017). bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC evolutionary biology*, *17*(1), 42.