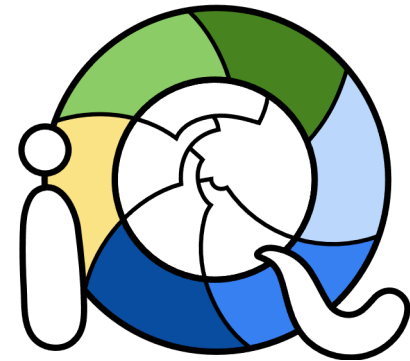

Lecture 1.2

Evolutionary Models

Popular phylogenetic methods

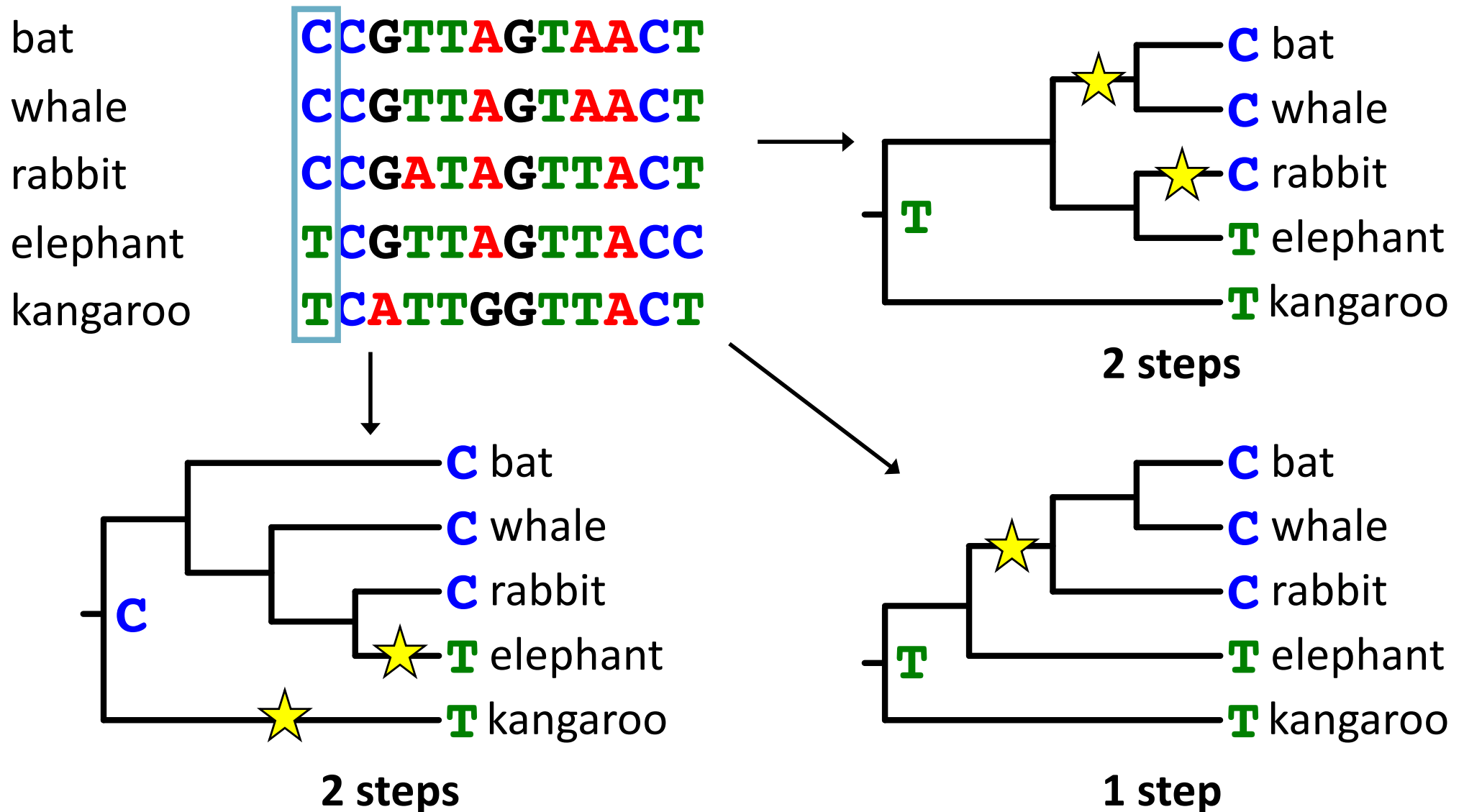
1. Maximum parsimony
2. Distance-based methods
3. Maximum likelihood
4. Bayesian inference

Model-based methods

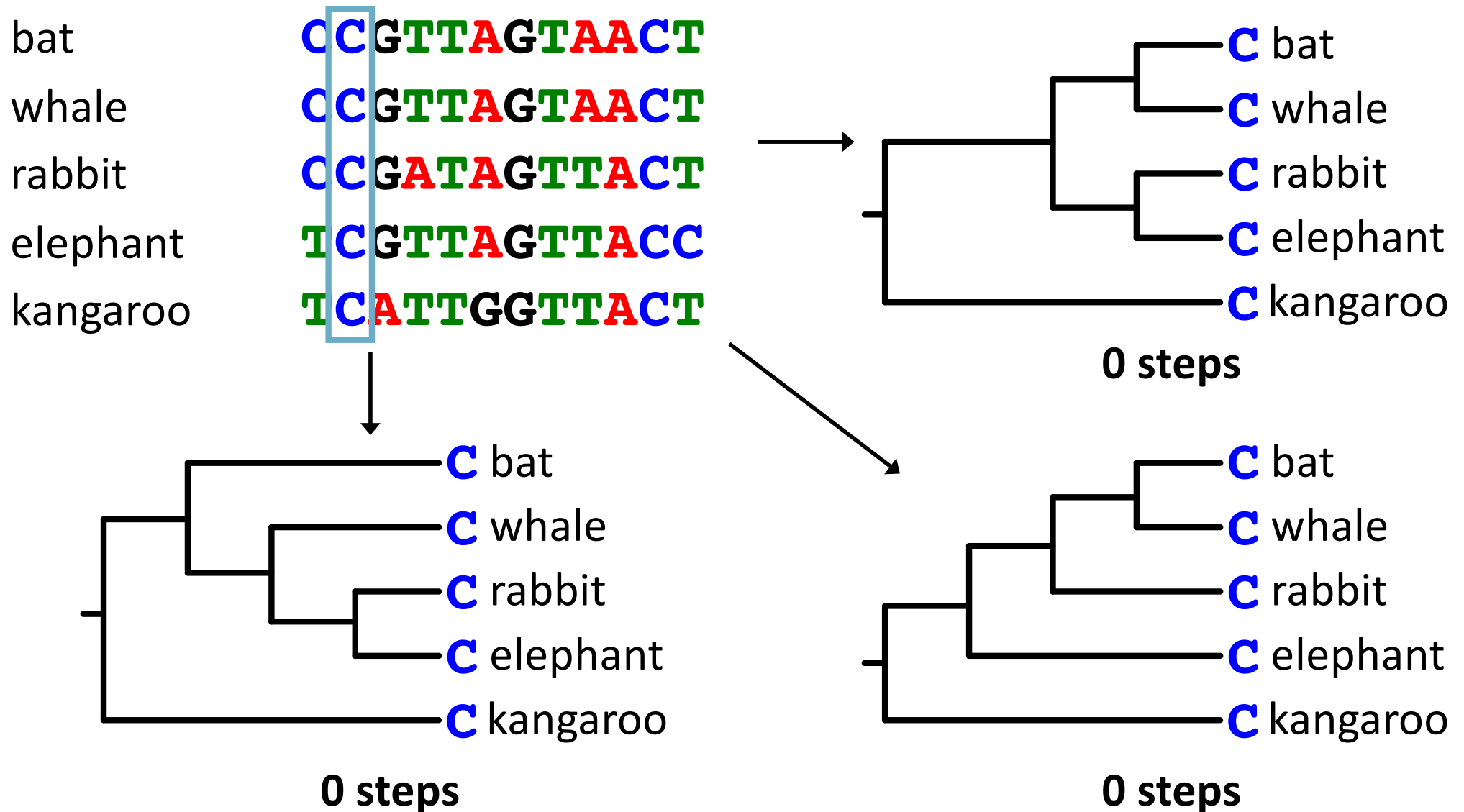


Maximum Parsimony

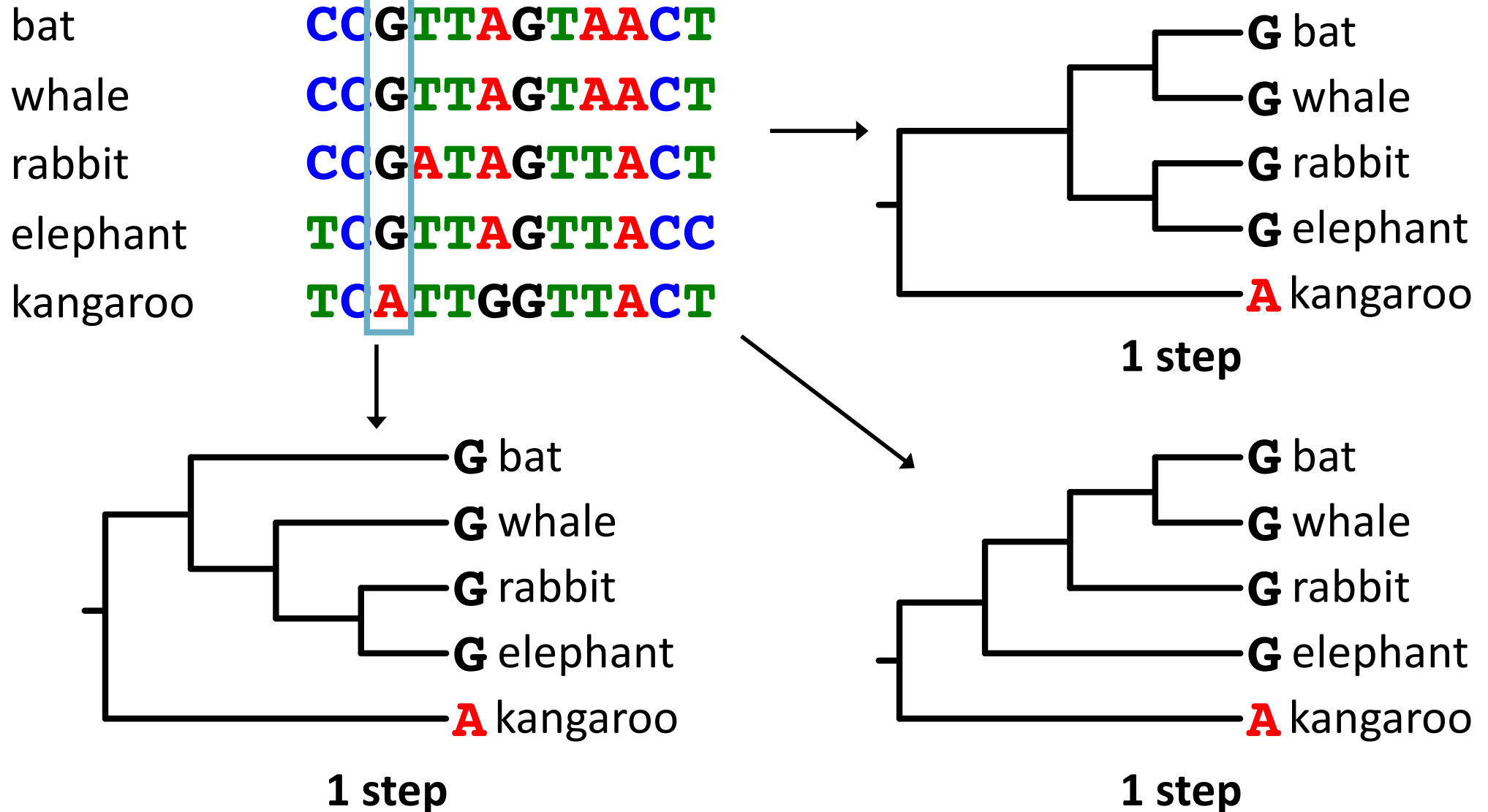
Maximum parsimony



Maximum parsimony



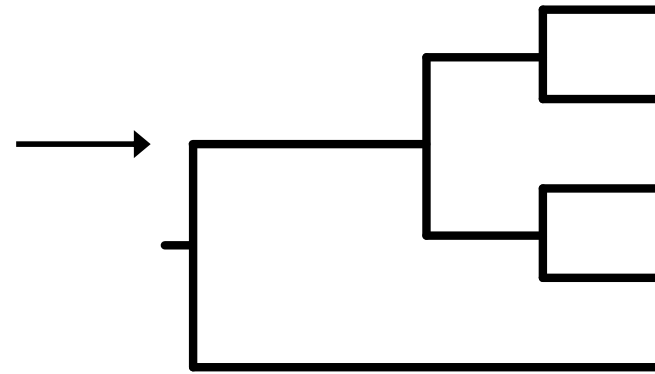
Maximum parsimony



Maximum parsimony

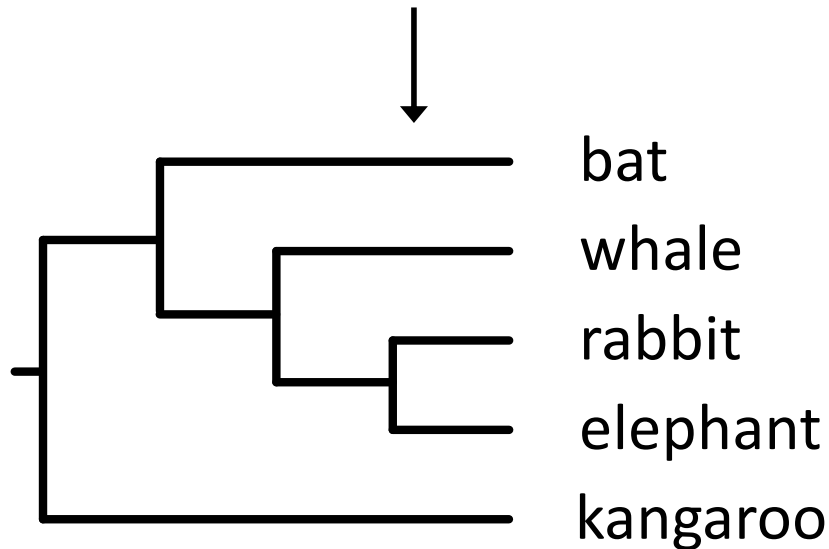
bat
whale
rabbit
elephant
kangaroo

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

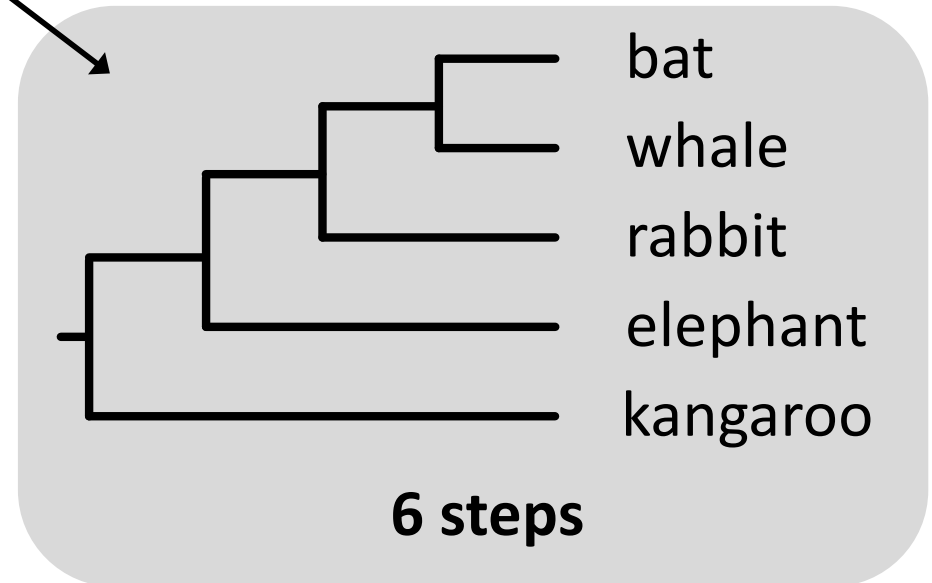


bat
whale
rabbit
elephant
kangaroo

7 steps



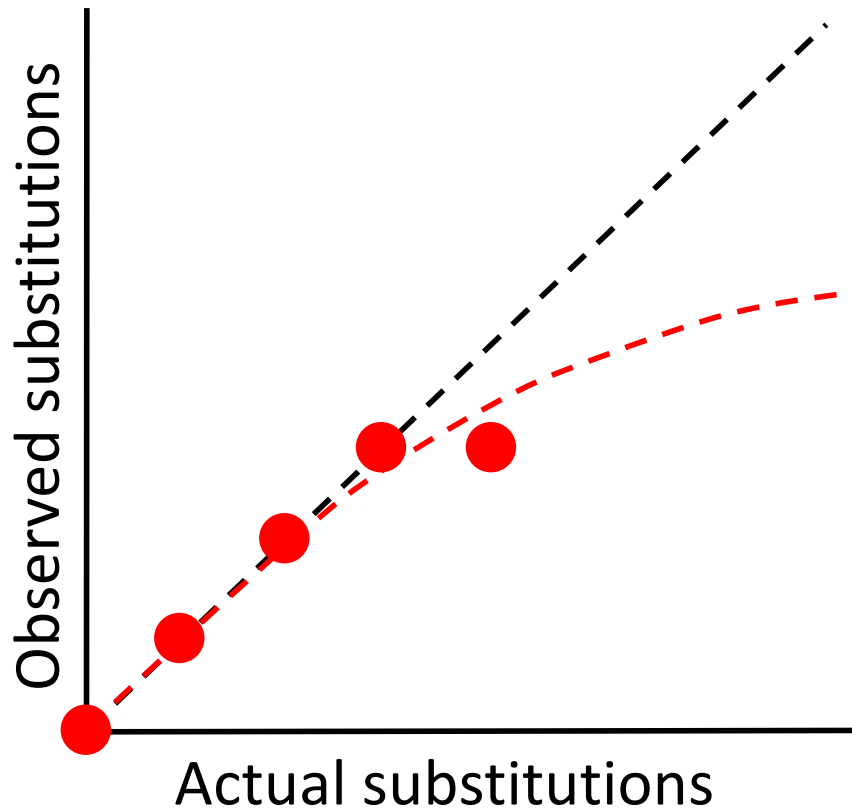
8 steps



6 steps

Maximum parsimony

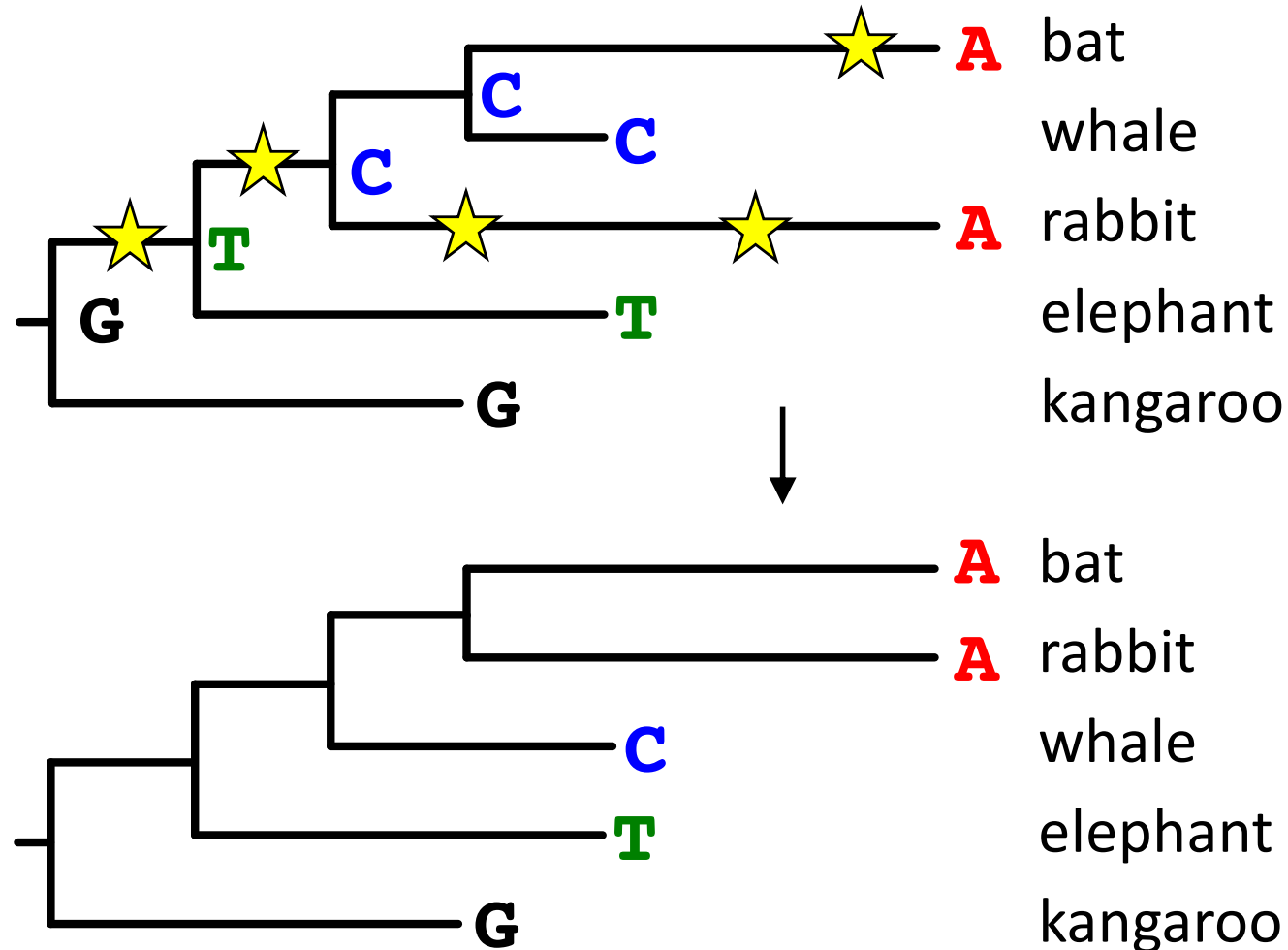
- Identifies the tree topology that can explain the sequence data, using the smallest number of inferred substitution events
- Commonly used for morphological data
- Now *rarely used* for analysing genetic data
 - Cannot estimate evolutionary rates or timescales
 - Effects of multiple substitutions



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

- Maximum parsimony does not correct for multiple substitutions at the same site
- This leads to a problem known as **long-branch attraction**
 - Long branch = many substitutions
 - Similarities arise by chance
 - Long branches cluster together

Long-branch attraction

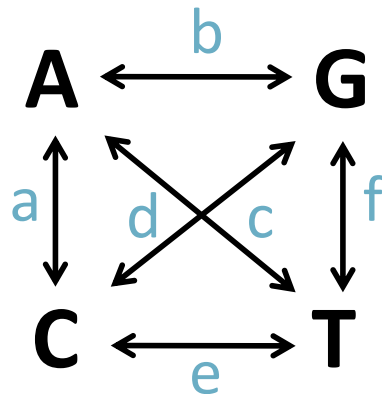


We can correct for multiple hits using substitution models

Substitution Models

Nucleotide substitution models

Rate Matrix



Base Frequencies

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

JC

$$a=b=c=d=e=f$$

$$\pi_A = \pi_C = \pi_G = \pi_T$$

HKY

$$a=c=d=f, b=e$$

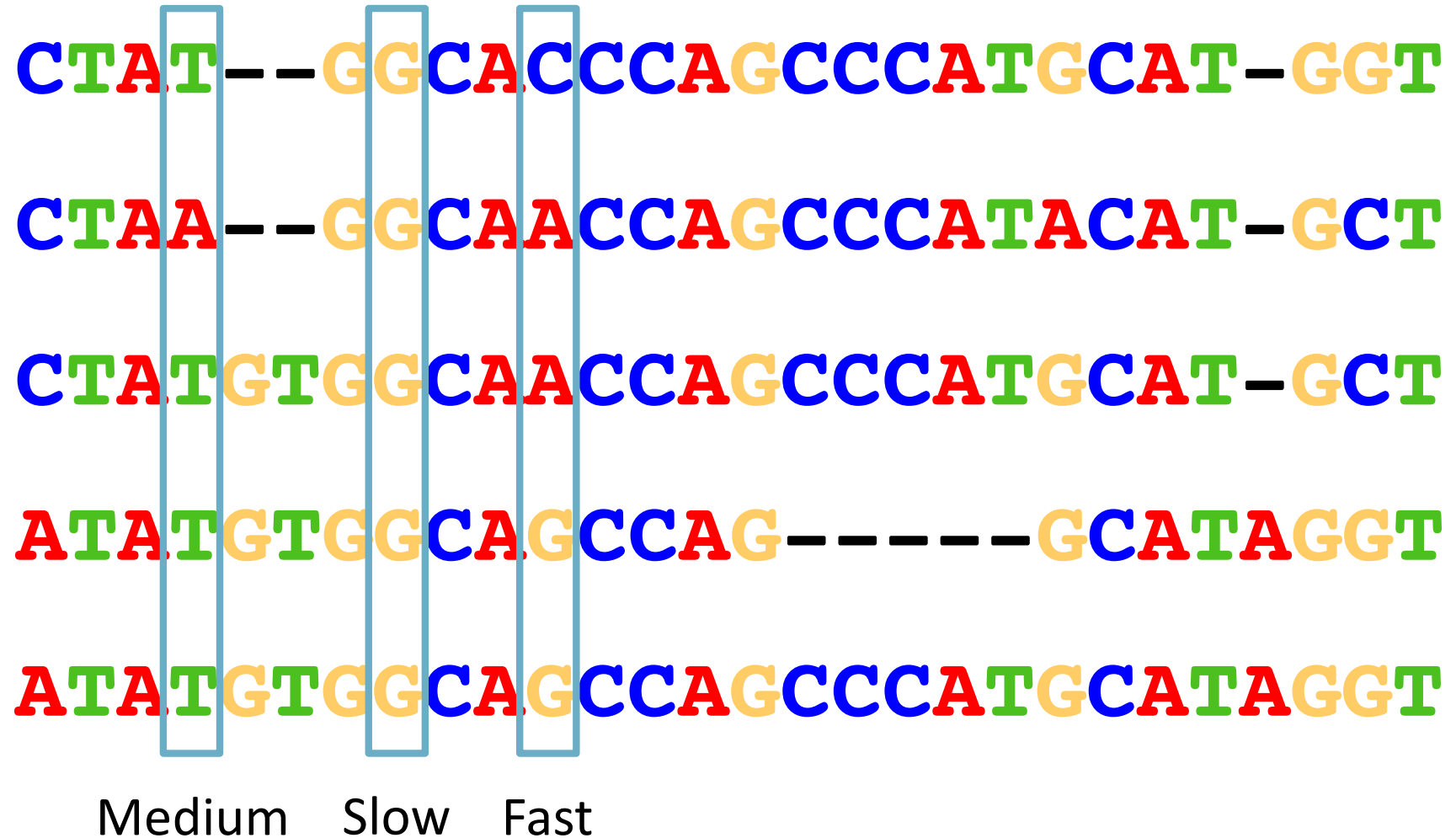
$$\pi_A, \pi_C, \pi_G, \pi_T$$

GTR

$$a, b, c, d, e, f$$

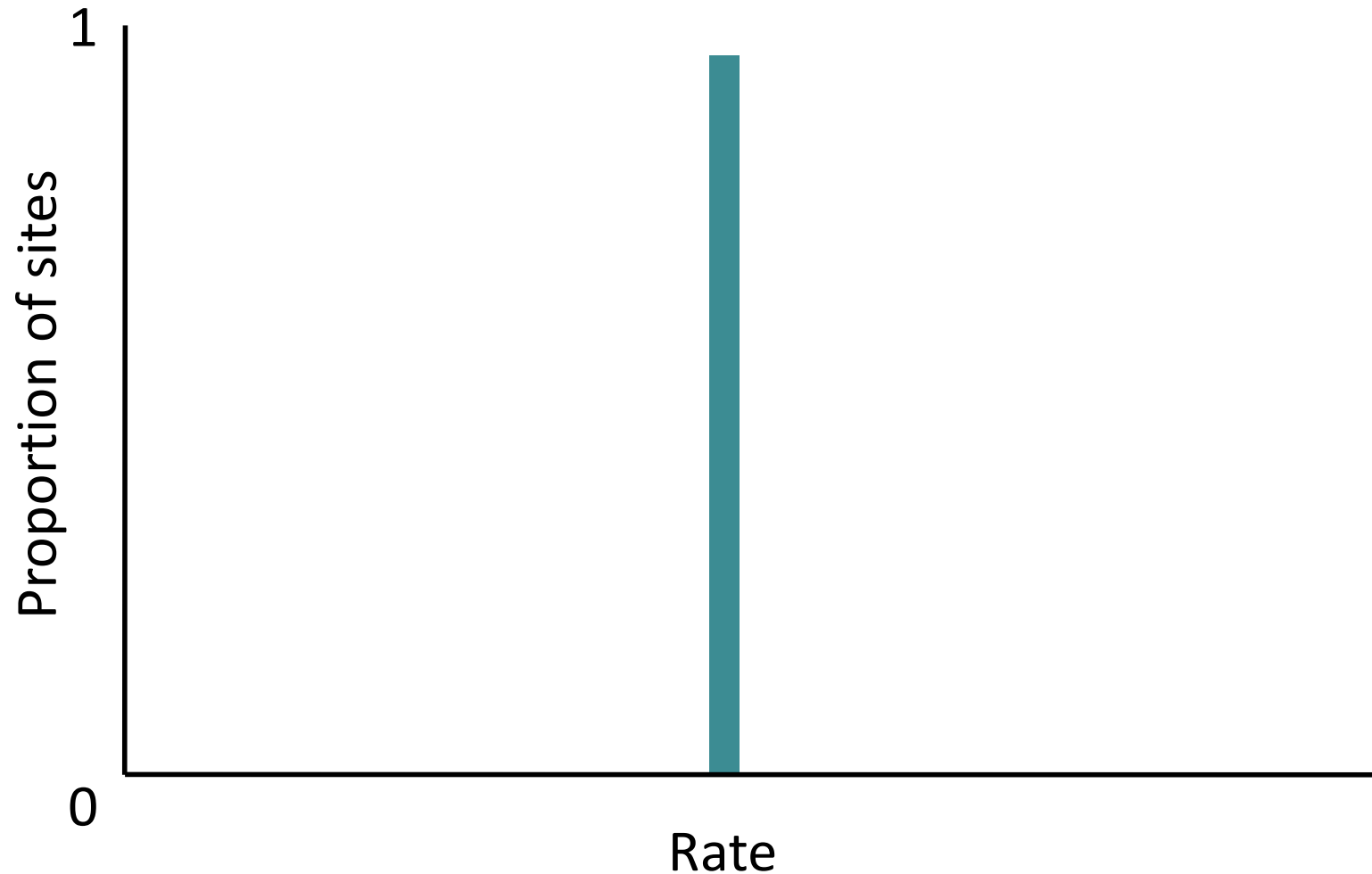
$$\pi_A, \pi_C, \pi_G, \pi_T$$

Rate variation across sites



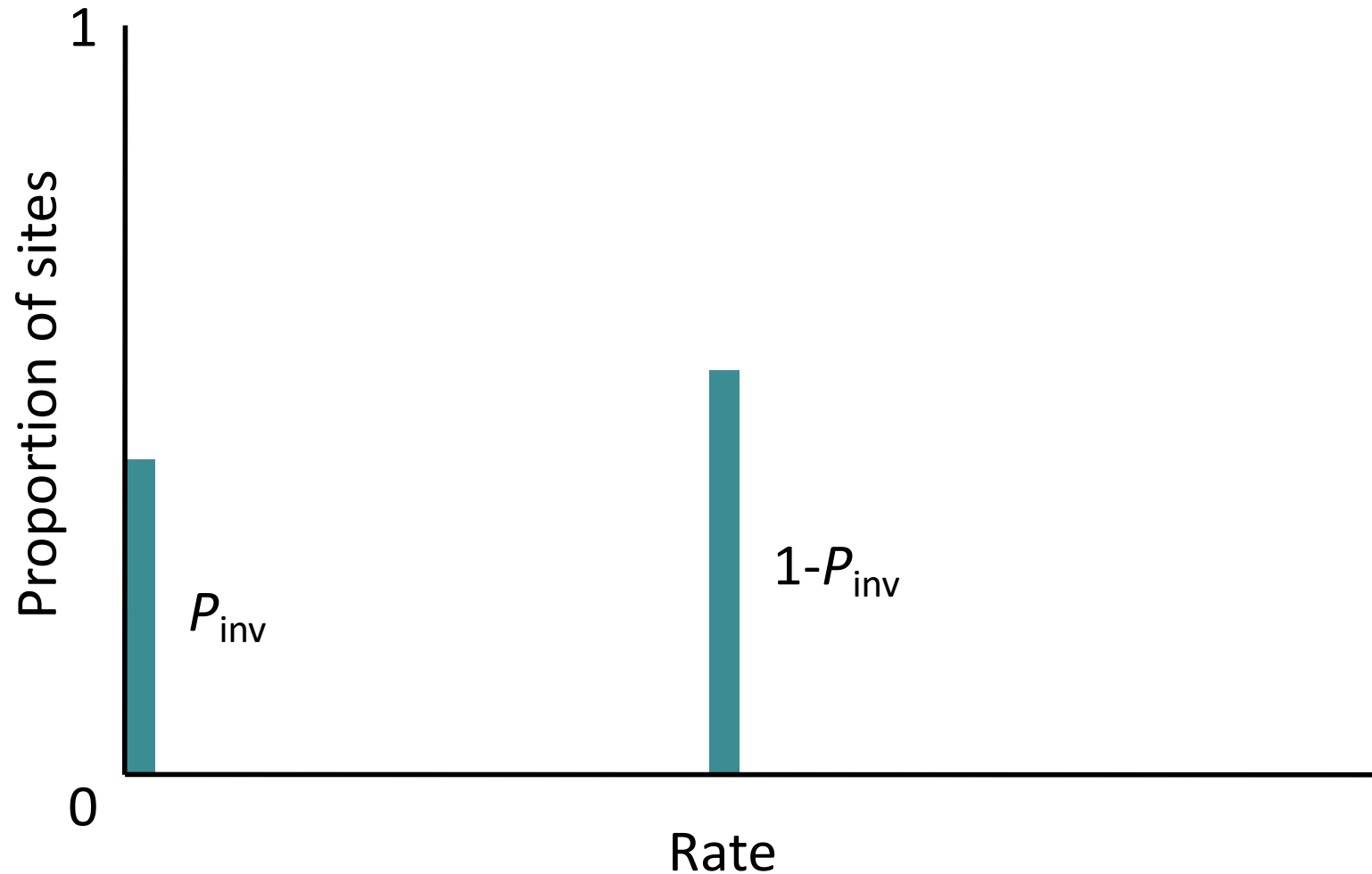
Rate variation across sites

- Equal rates among sites



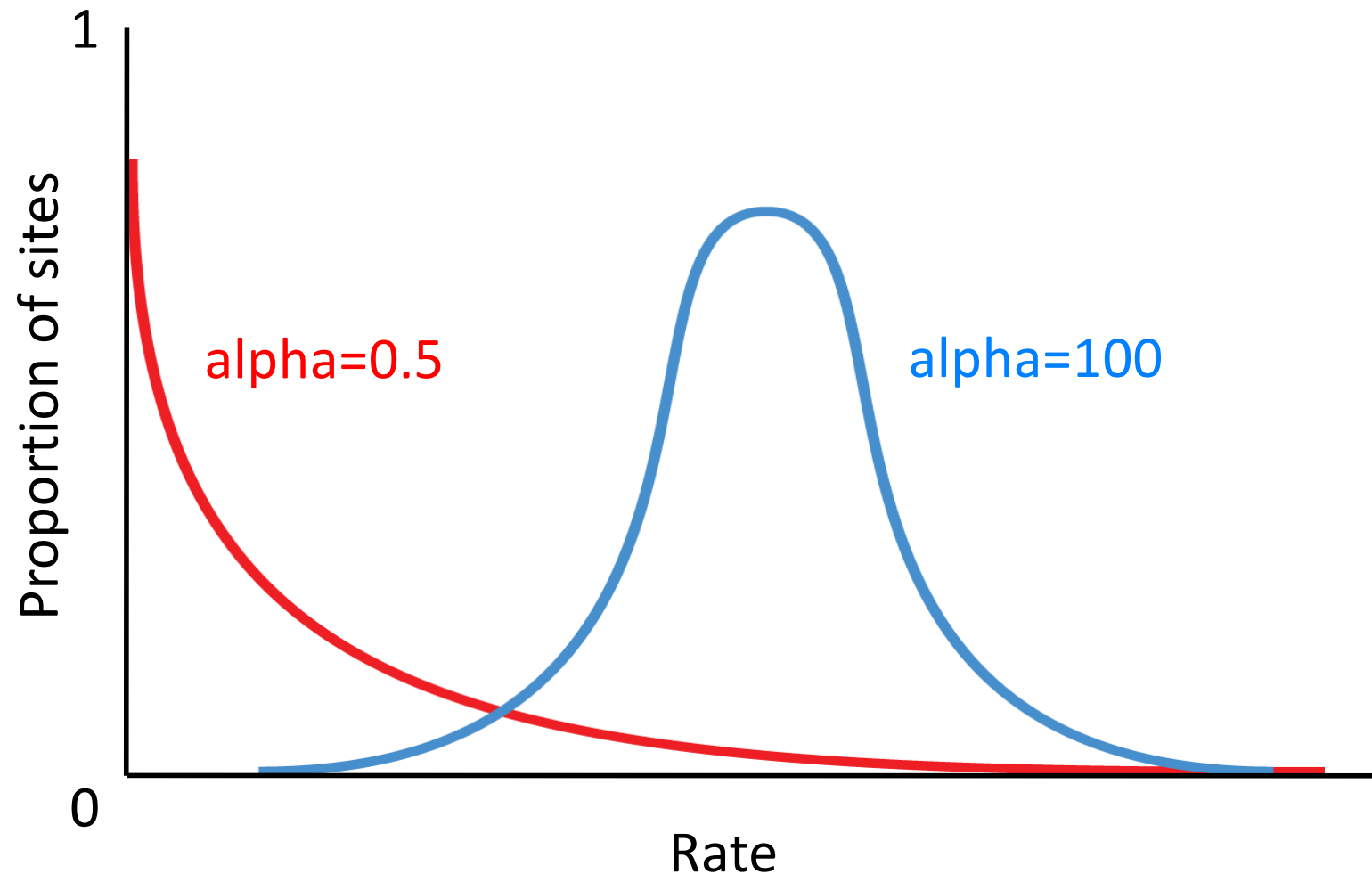
Rate variation across sites

- Proportion of invariable sites (+I models)



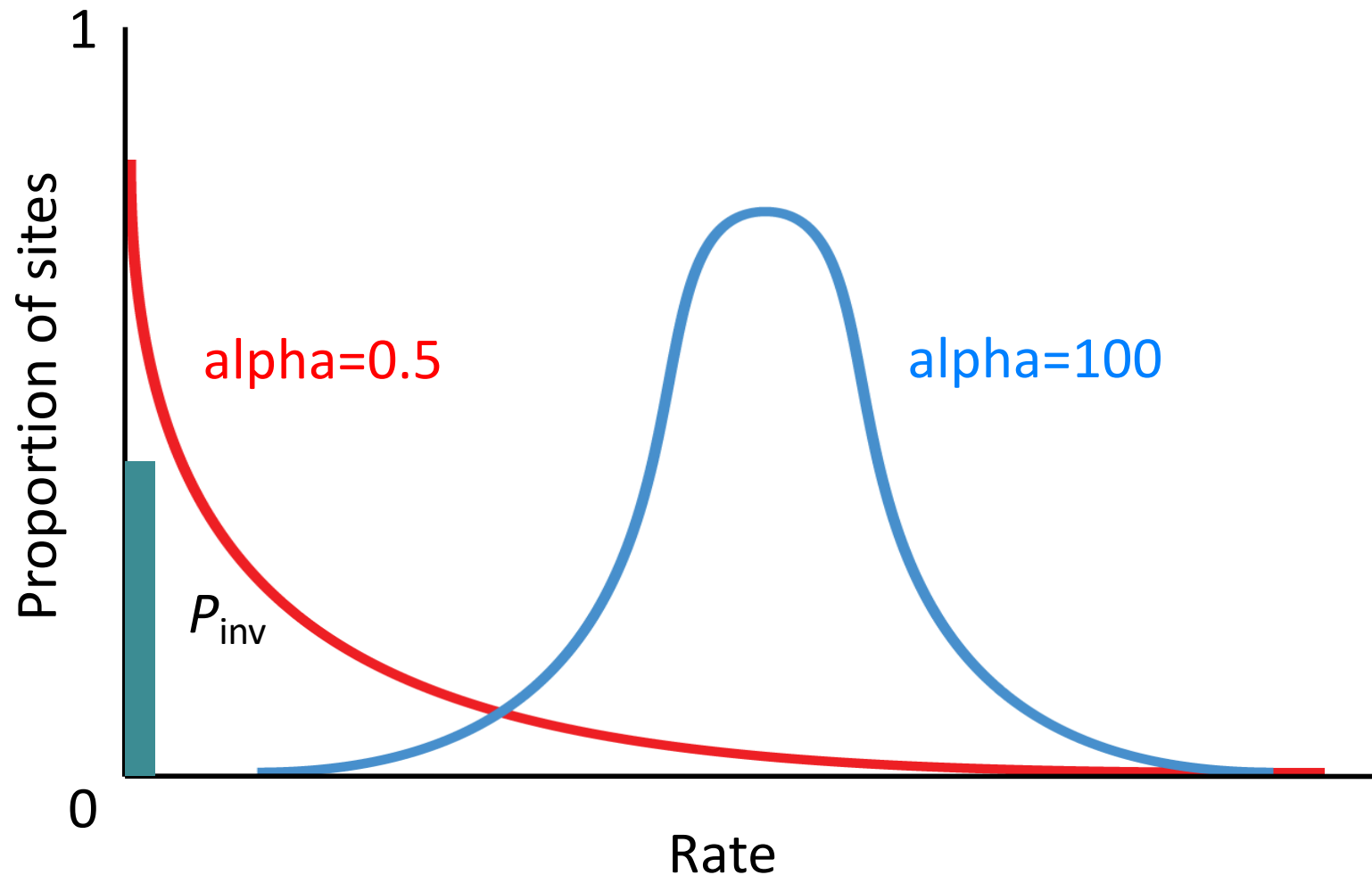
Rate variation across sites

- Gamma-distributed rate variation across sites (**+G** models)



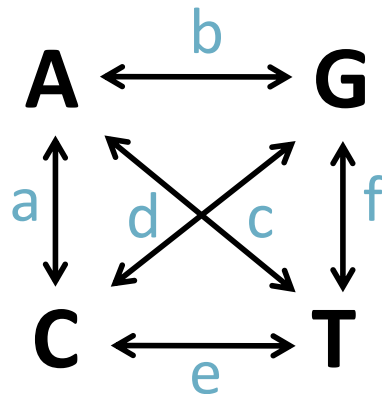
Rate variation across sites

- Gamma-distributed rate variation across sites and a proportion of invariable sites (**+G+I** models)



Nucleotide substitution models

Rate Matrix



Base Frequencies

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

Site Rates

$$+ I + G$$

JC

$$a=b=c=d=e=f$$

$$\pi_A = \pi_C = \pi_G = \pi_T$$

HKY

$$a=c=d=f, b=e$$

$$\pi_A, \pi_C, \pi_G, \pi_T$$

GTR

$$a, b, c, d, e, f$$

$$\pi_A, \pi_C, \pi_G, \pi_T$$

GTR+I+G

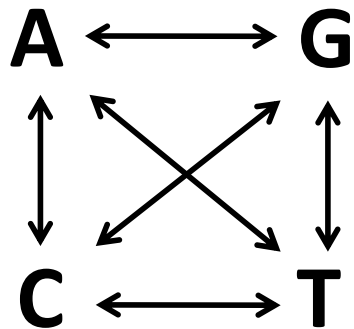
$$a, b, c, d, e, f$$

$$\pi_A, \pi_C, \pi_G, \pi_T$$

$$I, G$$

Nucleotide substitution models

Rate Matrix



Base Frequencies

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

Site Rates

$$+ I + G$$

#Models

203

x

15

x

4

= 12,180

In phylogenetics, we typically consider a small subset of these

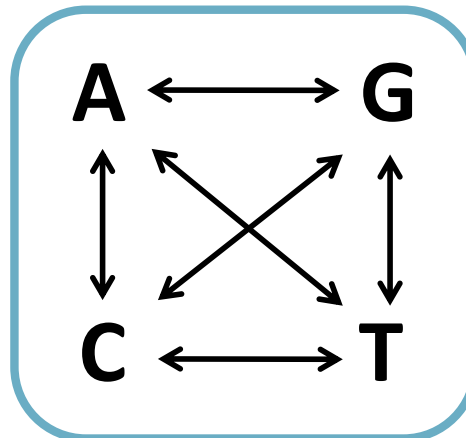
Proportion of invariable sites

- Often overestimated in analyses of intraspecific data
- Unable to distinguish between:
 - Sites that are **invariable** and unable to change
 - Sites that are **constant** and by chance have not mutated
- Not always biologically meaningful
- Slowly evolving sites taken into account by **+G**

Use +G models to account for rate variation across sites

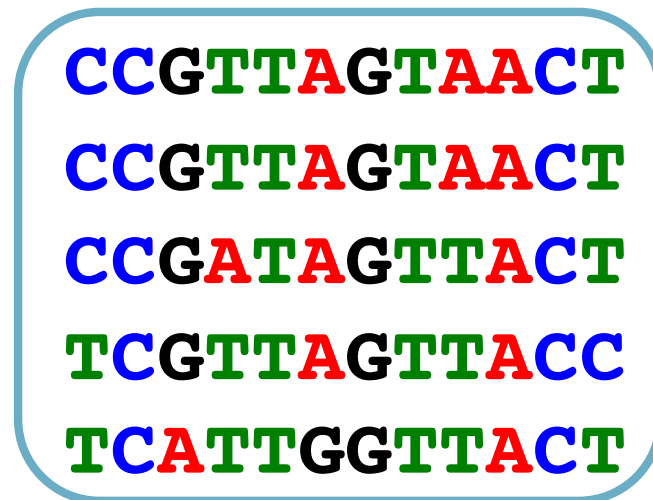
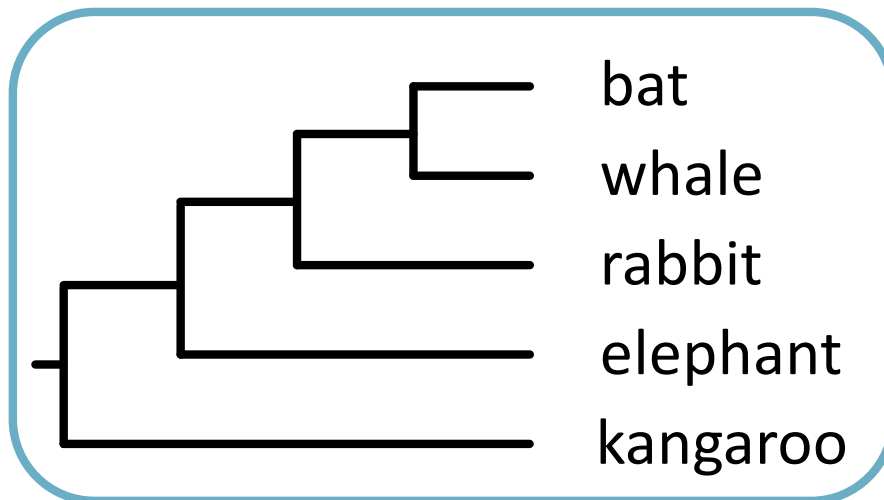
Fundamental assumptions

Reversible



Stationary

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

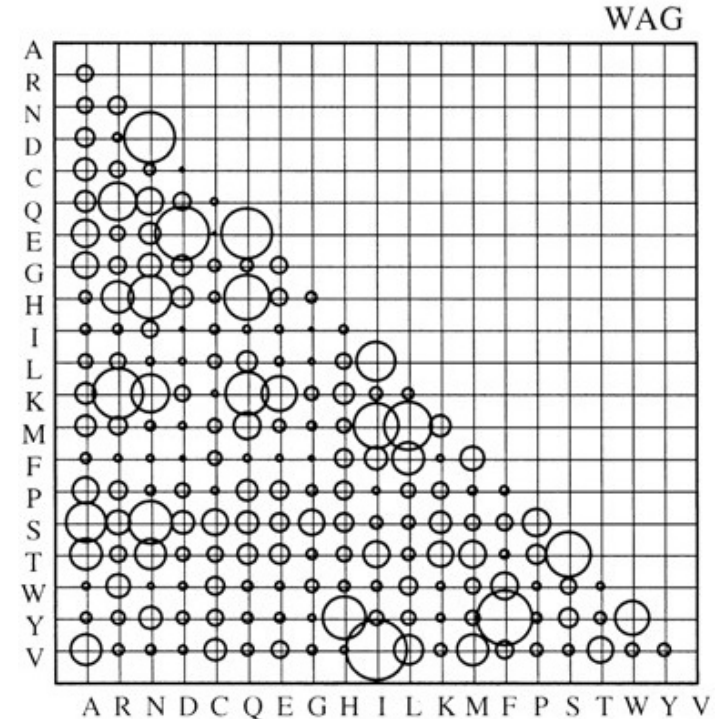


Homogeneous

Independent across sites

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 large data set
 - PAM
 - BLOSUM
 - JTT
 - WAG



Model Selection

Model selection

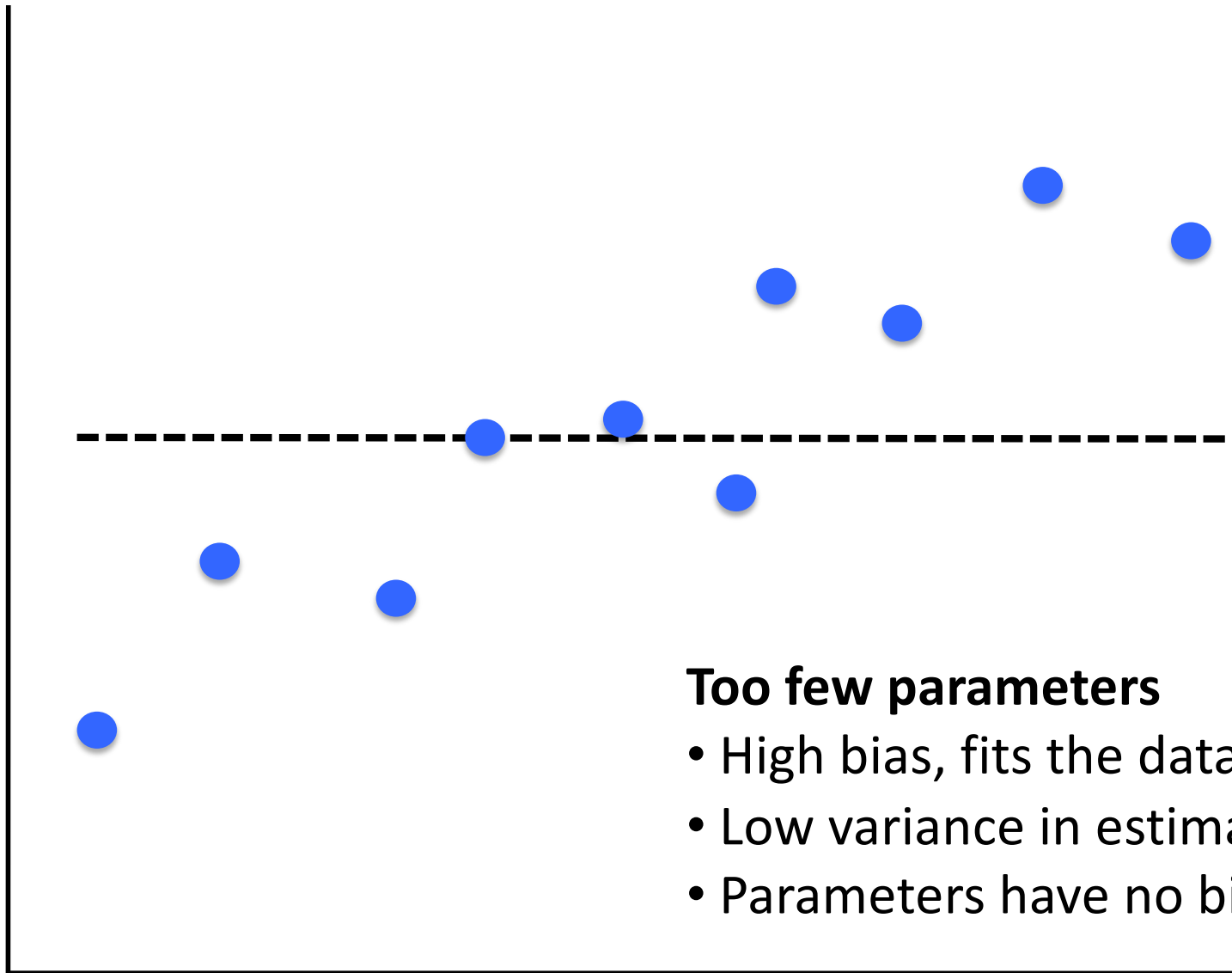
1. Subjective model selection

- Pick a model that seems sensible
- Balance the number of parameters against the amount of data
- Biological motivation

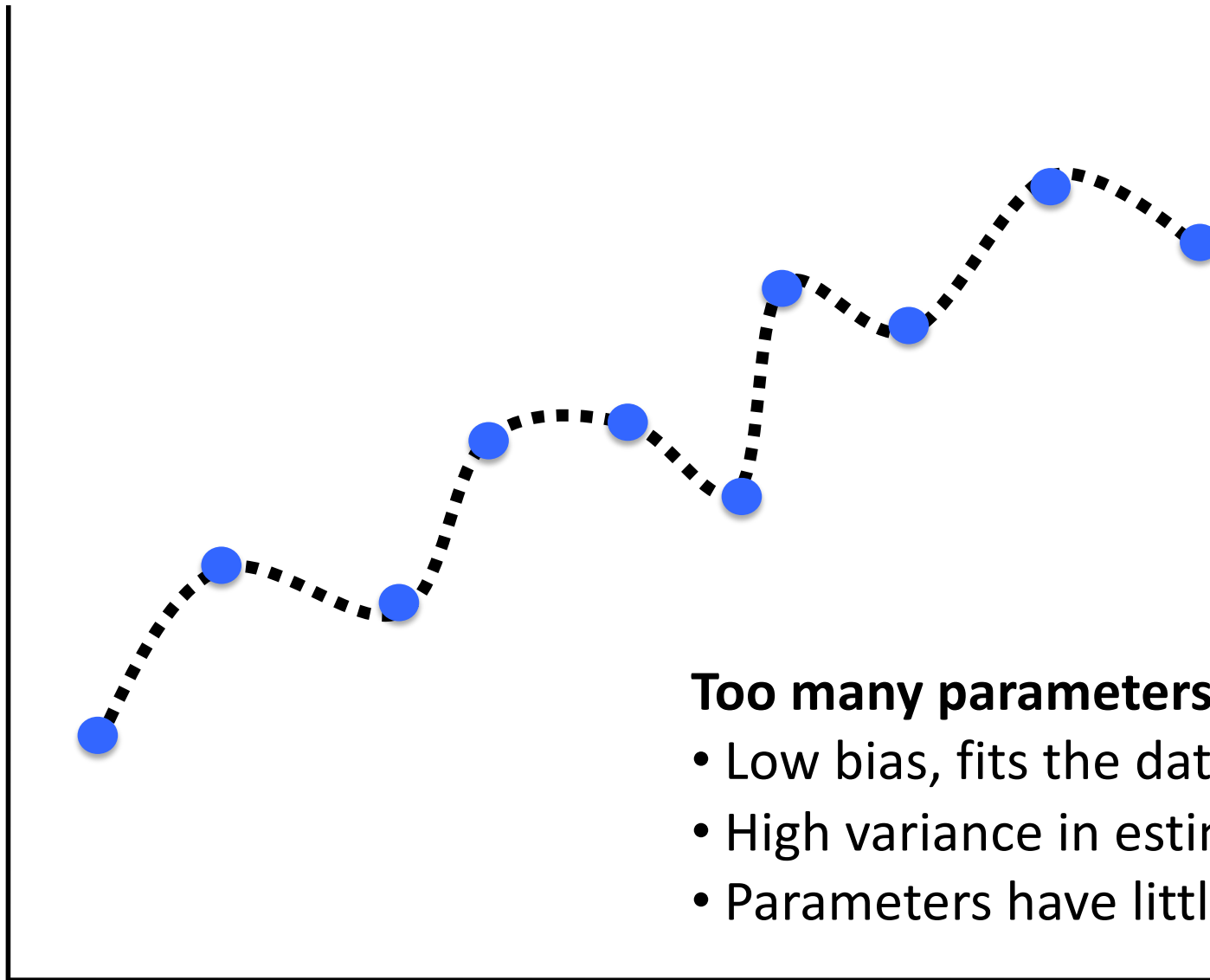
2. Objective model selection

- Use information theory and let a computer do it for you
- Statistical motivation

Model selection



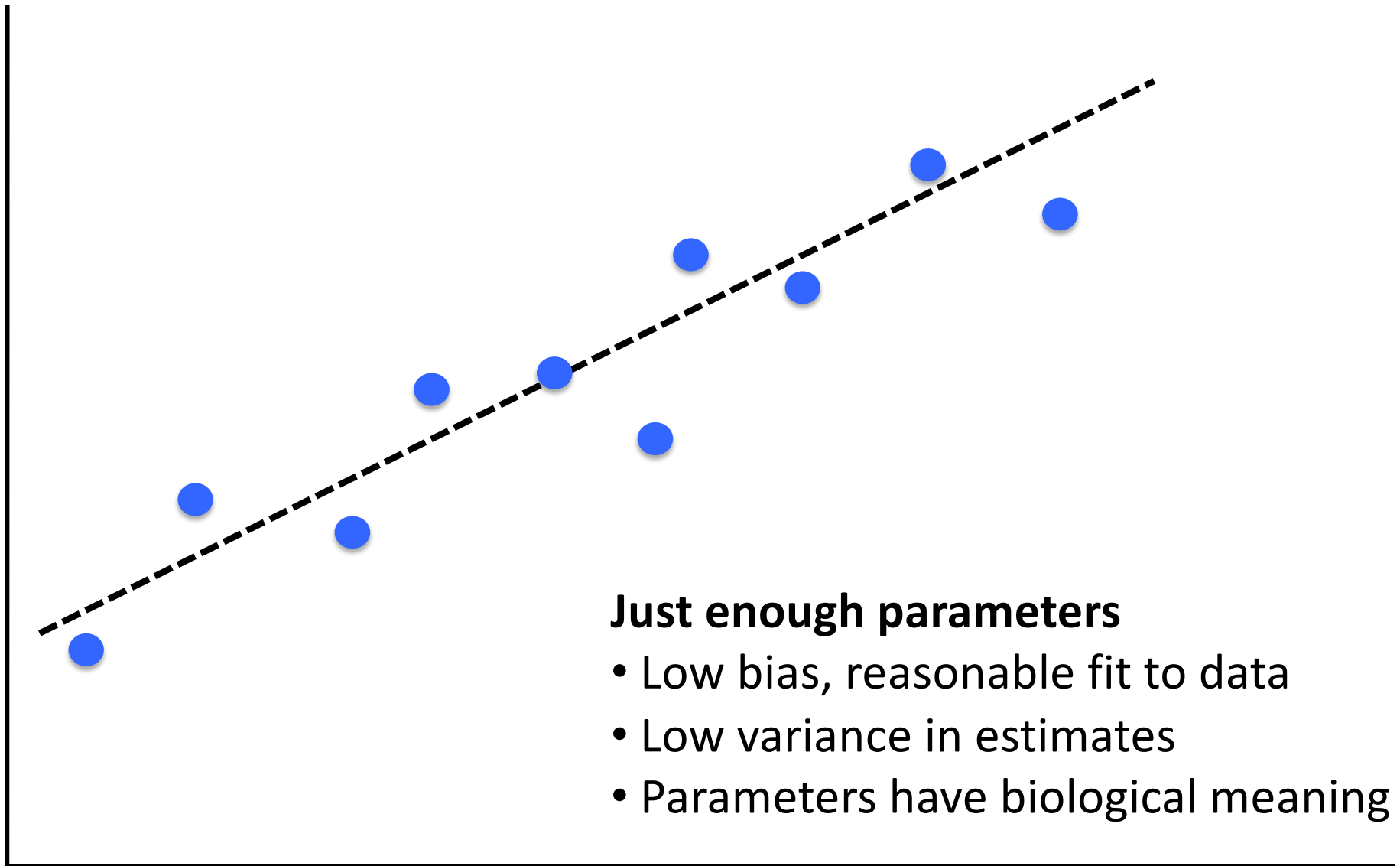
Model selection



Too many parameters

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

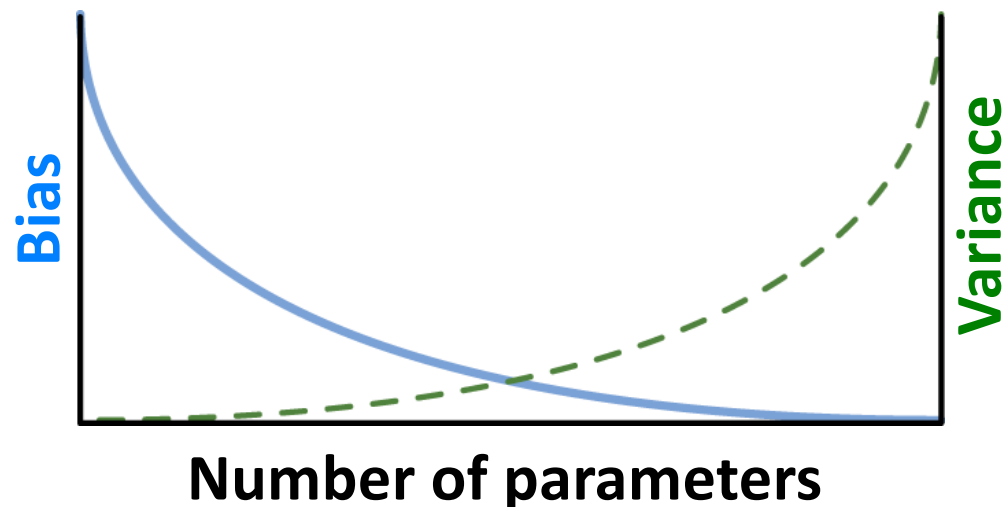
Model selection



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

Is the improvement in model fit worth the cost of adding a parameter?



Model selection

- **Likelihood-ratio test (LRT)**
Used to compare nested models
- **Akaike information criterion (AIC)**
 $AIC = -2\ln(\text{likelihood}) + 2k$
- **Bayesian information criterion (BIC)**
 $BIC = -2\ln(\text{likelihood}) + k\ln(n)$

Phylogenetic estimates are often robust to choice of model

Useful references

- **Model selection in phylogenetics**
Sullivan & Joyce (2005) *Annual Review of Ecology, Evolution, and Systematics*, 36: 445–466.
- **Model selection may not be a mandatory step for phylogeny reconstruction**
Abadi et al. (2019)
Nature Communications, 10: 934.

