

Lecture 1.2:

Substitution Models

Popular phylogenetic methods

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

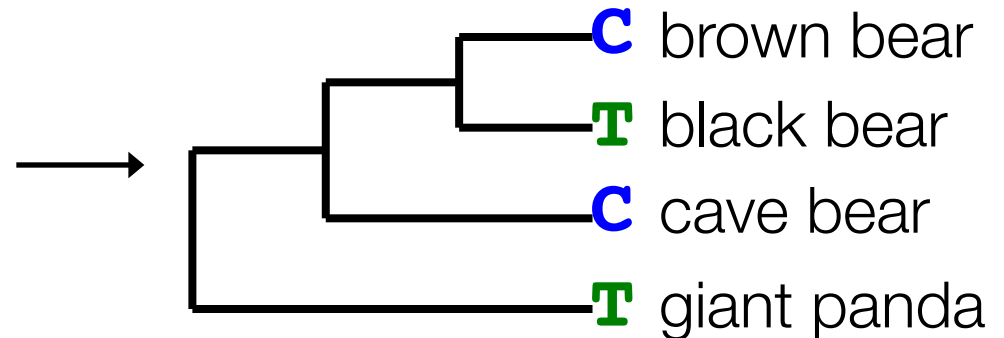
Model-based methods



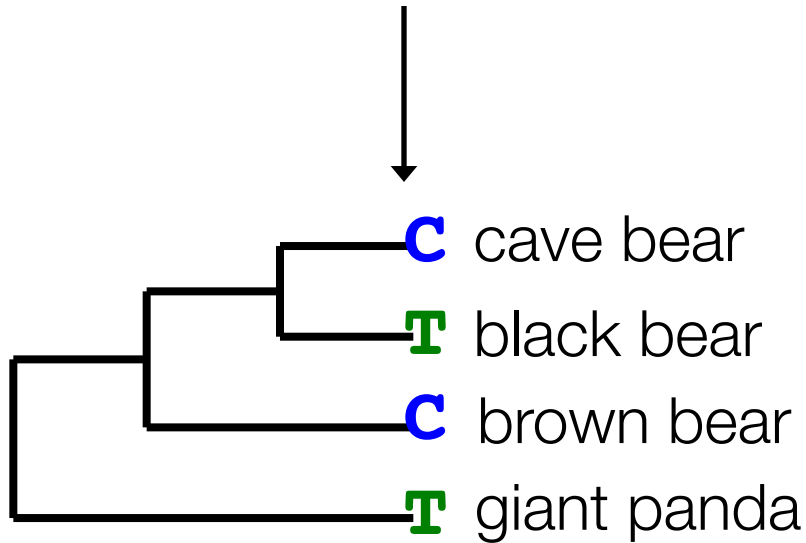
Maximum Parsimony

Maximum parsimony

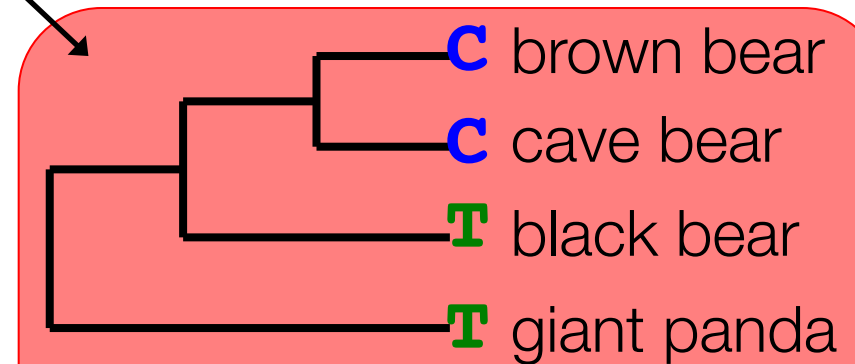
brown bear **C****G****T****T****A****G****T****A****C****A****C****T**
cave bear **C****G****A****T****A****G****T****T****C****A****C****T**
black bear **C****G****T****T****A****G****T****T****T****A****C****C**
giant panda **C****A****T****T****G****G****T****T****T****A****C****T**



2 steps (7 overall)



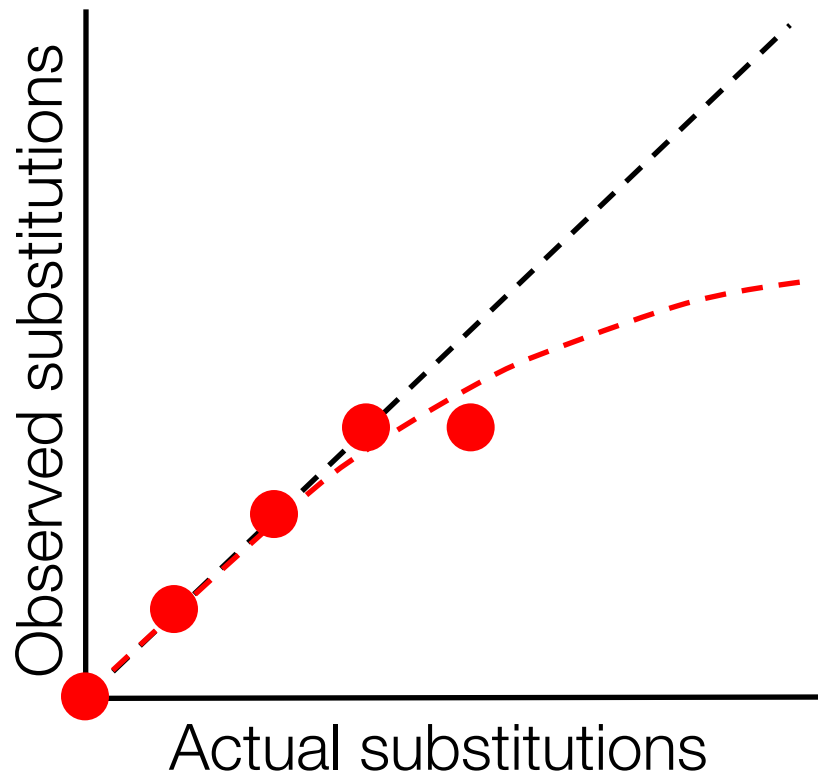
2 steps (7 overall)



1 step (6 overall)

Maximum parsimony

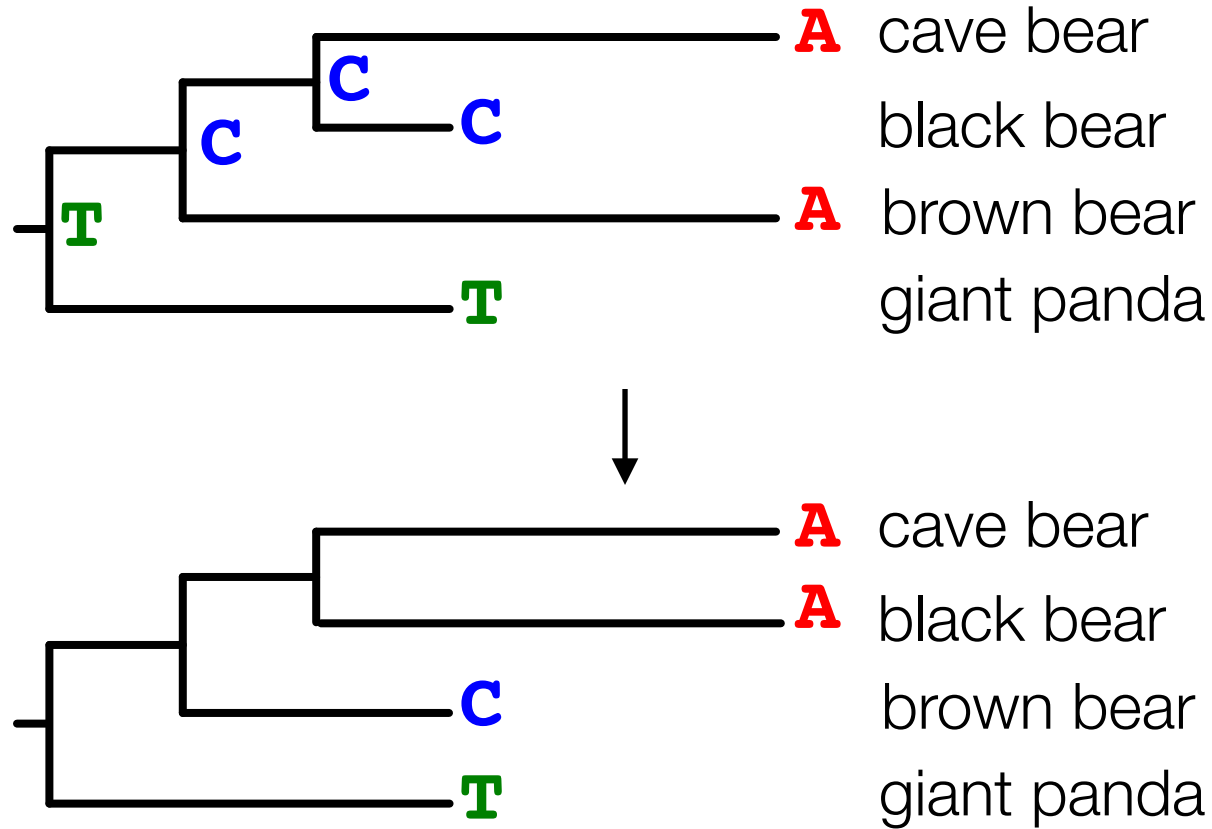
- Identifies the tree topology that can explain the sequence data, using the smallest number of inferred substitution events
- Commonly used for morphological data
- Now rarely used for analysing genetic data
 - Cannot estimate evolutionary rates or timescales
 - Effects of multiple substitutions



- 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

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

Long-branch attraction



Weaknesses

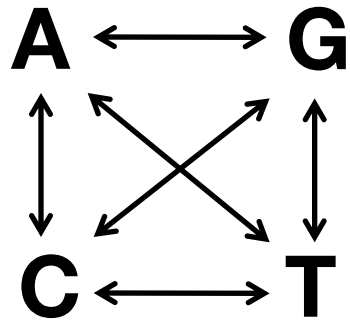
- Maximum parsimony does not correct for multiple substitutions at the same site
- This leads to a problem known as ‘long-branch attraction’
 - Long branches in the tree tend to group together

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

Evolutionary Models

Nucleotide substitution models

Rate Matrix



Base Frequencies

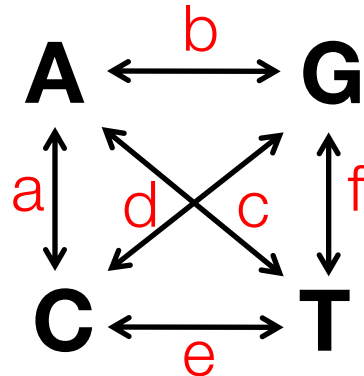
$$\pi_A + \pi_C + \pi_G + \pi_T = 1$$

Site Rates

$$+ I + G$$

Nucleotide substitution models

Rate Matrix



Base Frequencies

$$\pi_A + \pi_C + \pi_G + \pi_T = 1 \quad + \text{I} + \text{G}$$

Site Rates

JC

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

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

No I or G

0 free

parameters

HKY

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

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

No I or G

4 free

parameters

GTR

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

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

No I or G

8 free

parameters

GTR+I+G

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

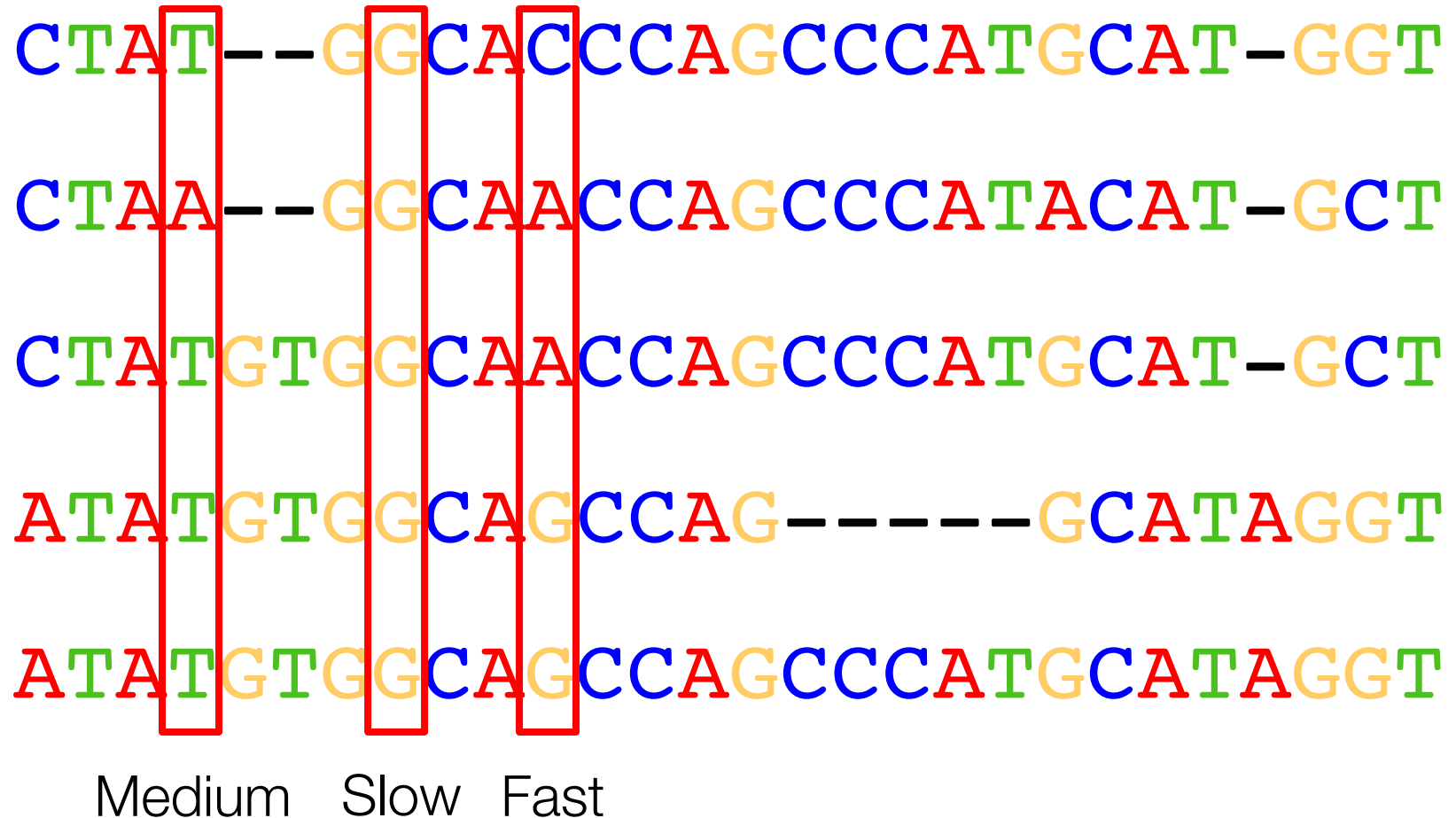
$$\pi_A, \pi_C, \pi_G, \pi_T$$

I, G

10 free

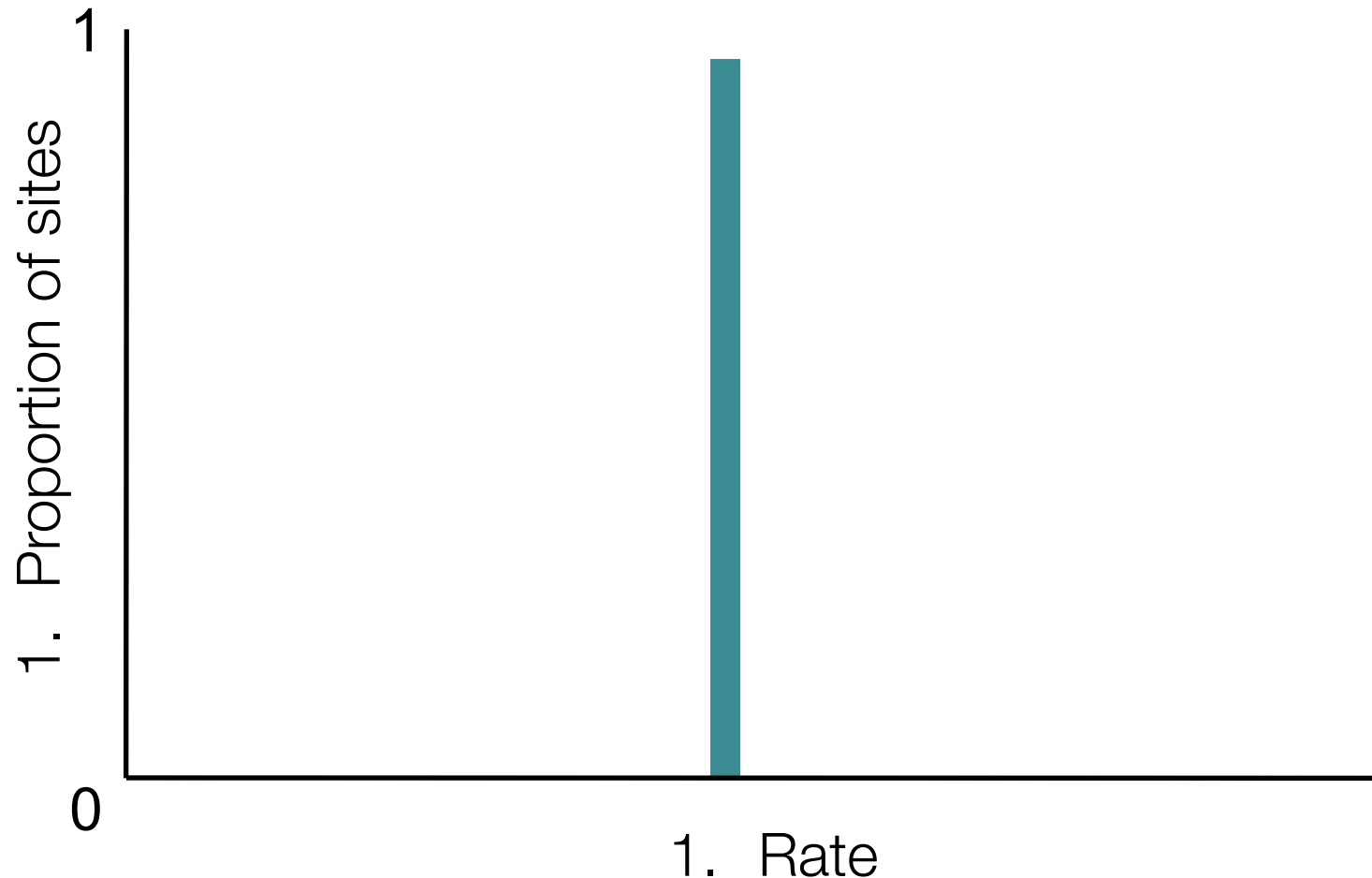
parameters

Rate variation across sites



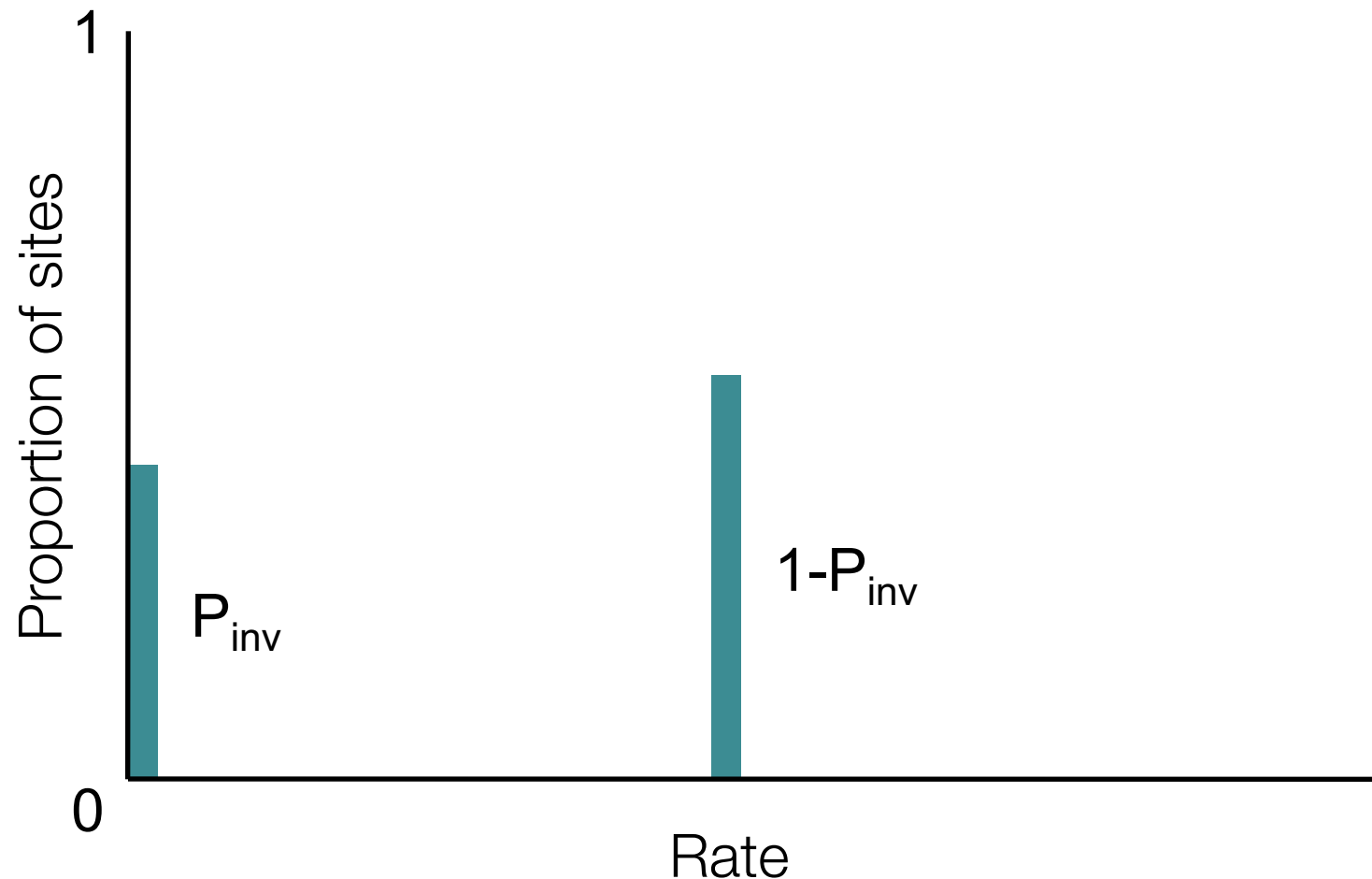
1. Rate variation among sites

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



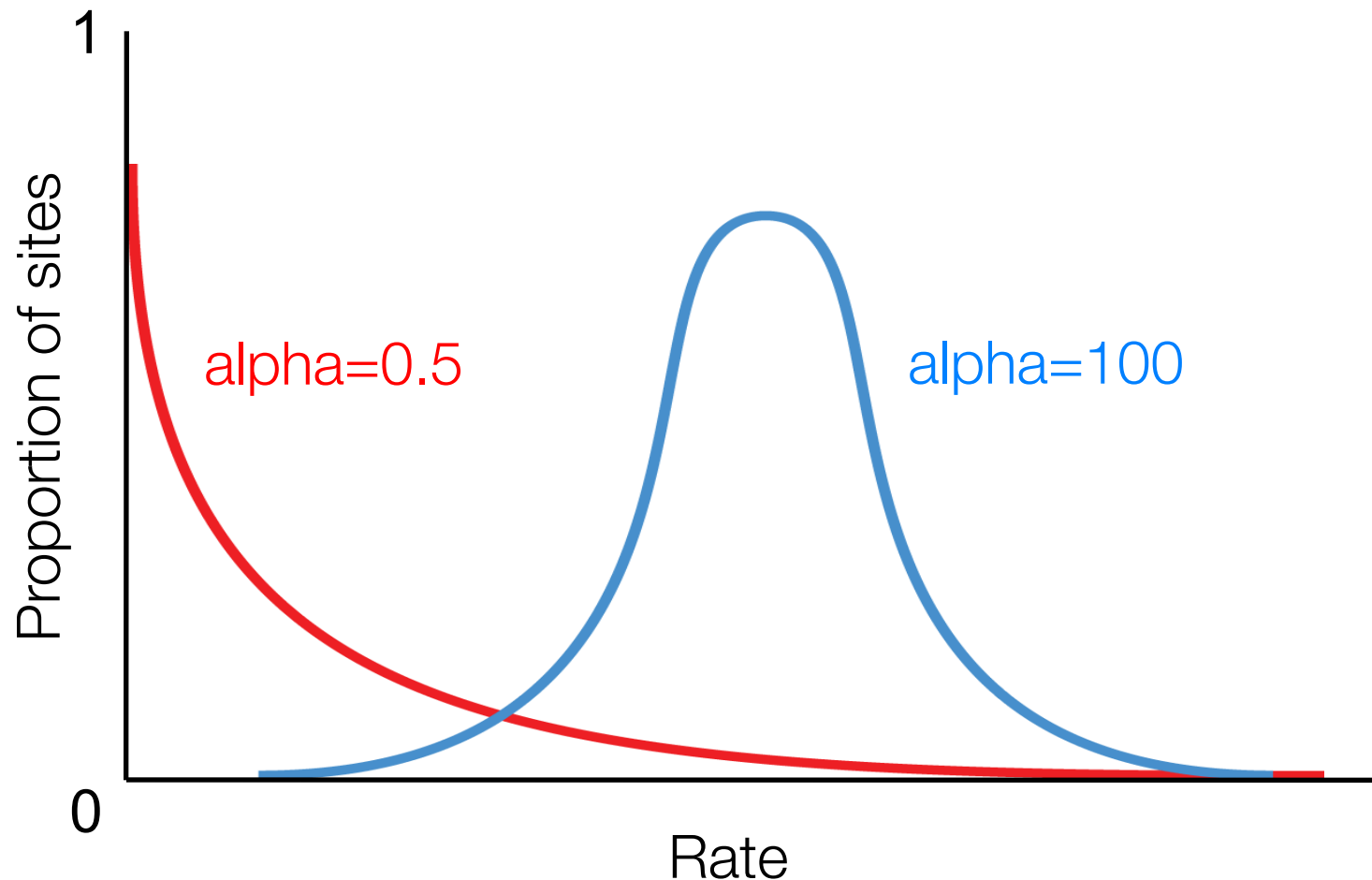
Rate variation among sites

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



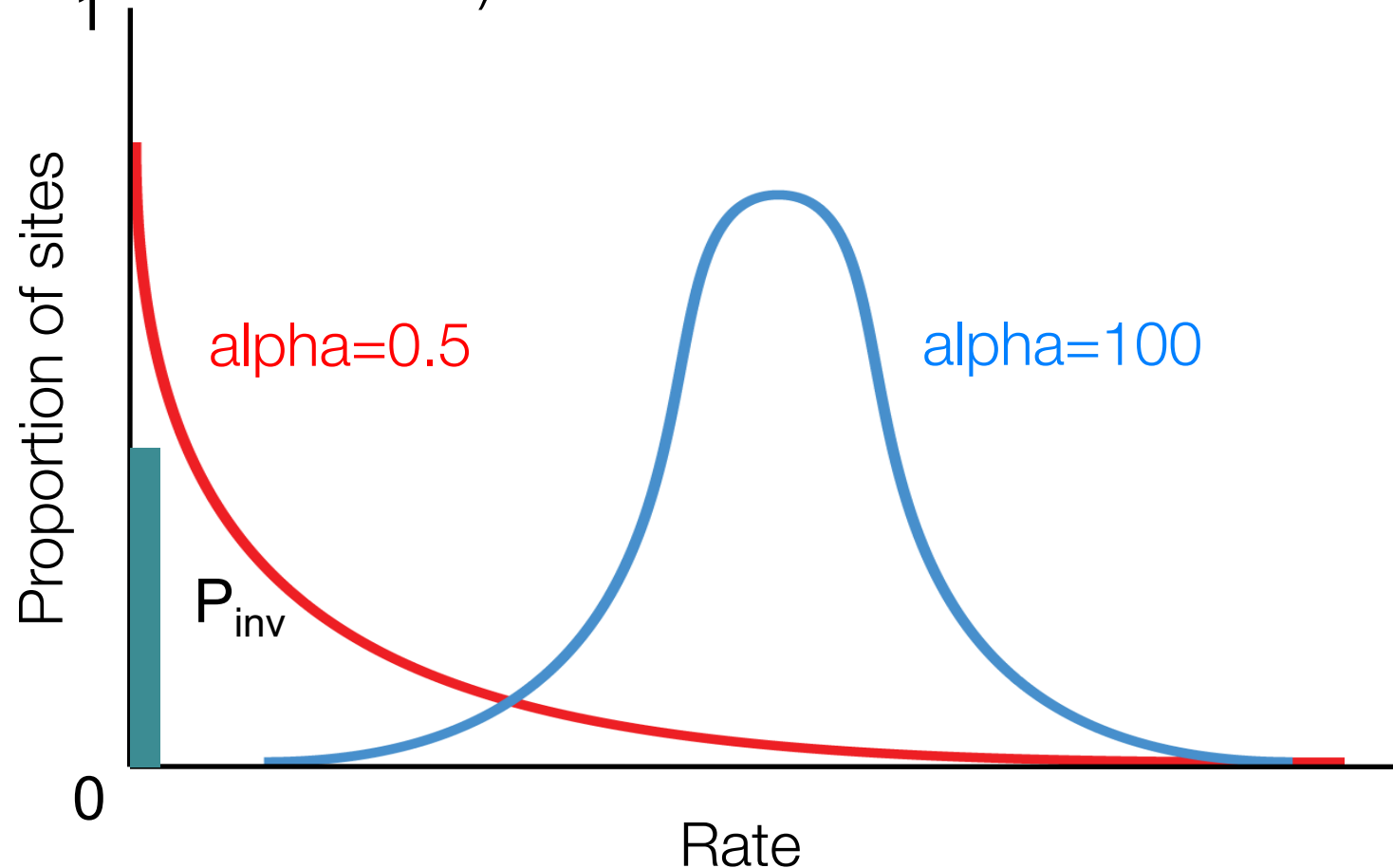
Rate variation among sites

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



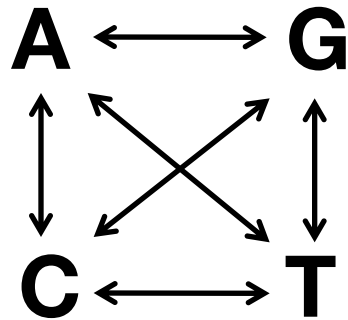
Rate variation among sites

- Gamma-distributed rate variation among sites and a proportion of invariable sites (e.g., **JC+G+I**, **GTR+G+I**, **HKY₁+G+I** models)



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

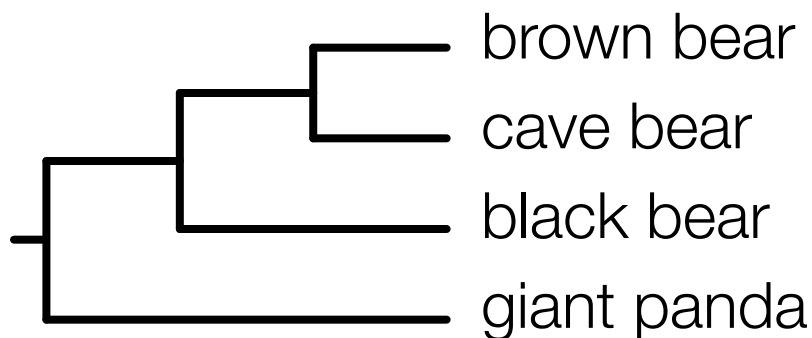
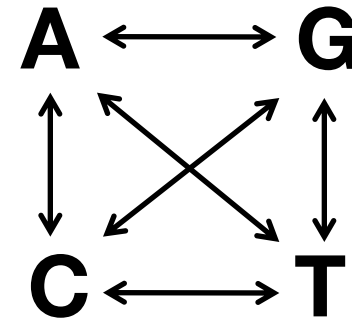
Amino acid substitution matrices

- 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

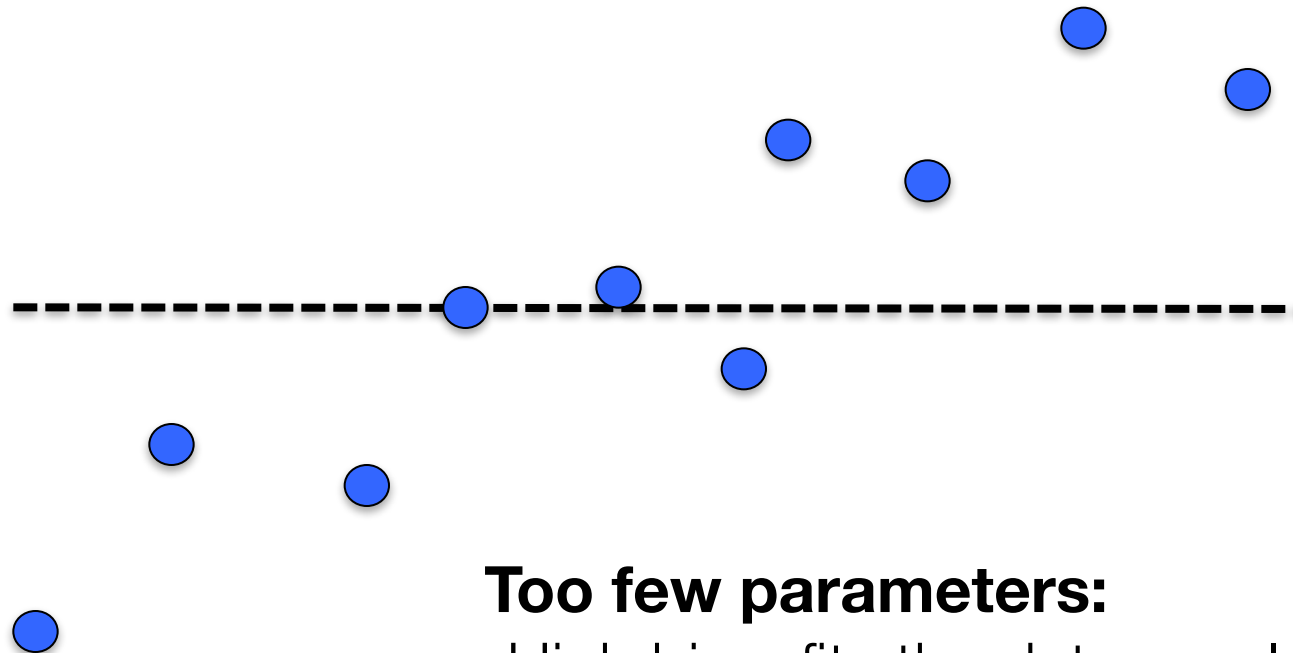
π_A π_C π_G π_T



CGTTAGTACACT
CGATAGTTCACACT
CGTTAGTTTACC
CATTGGTTTACT

Model Selection

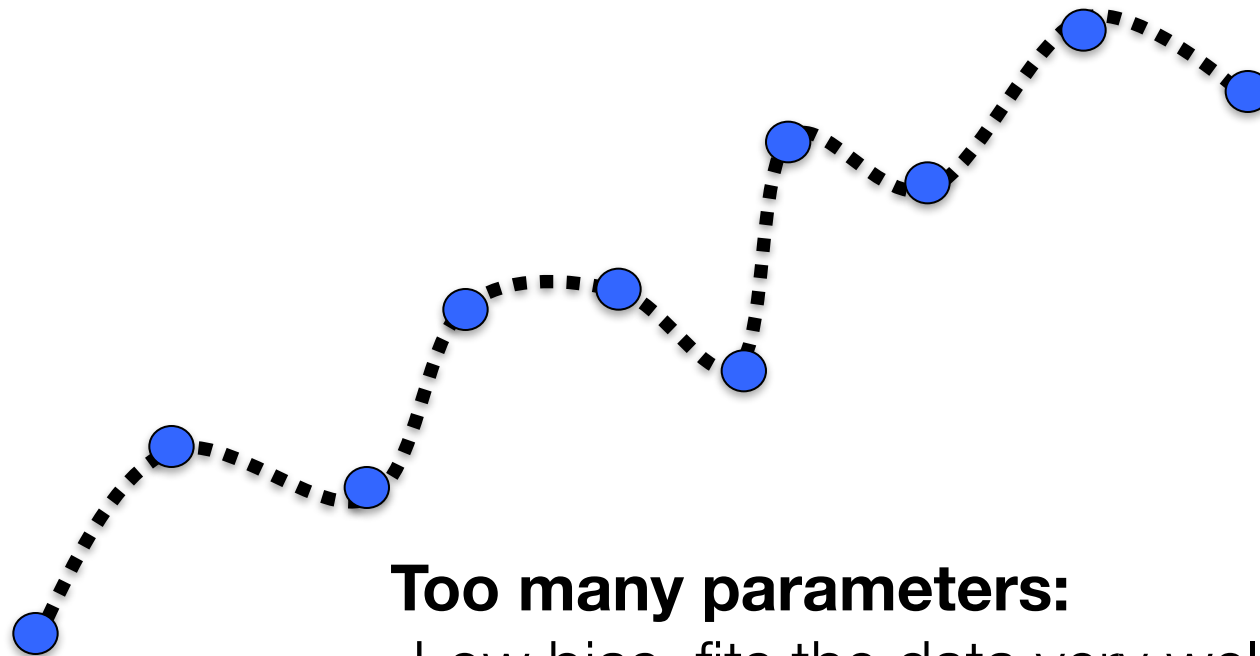
Model selection



Too few parameters:

- High bias, fits the data poorly
- Low variance in parameter estimate

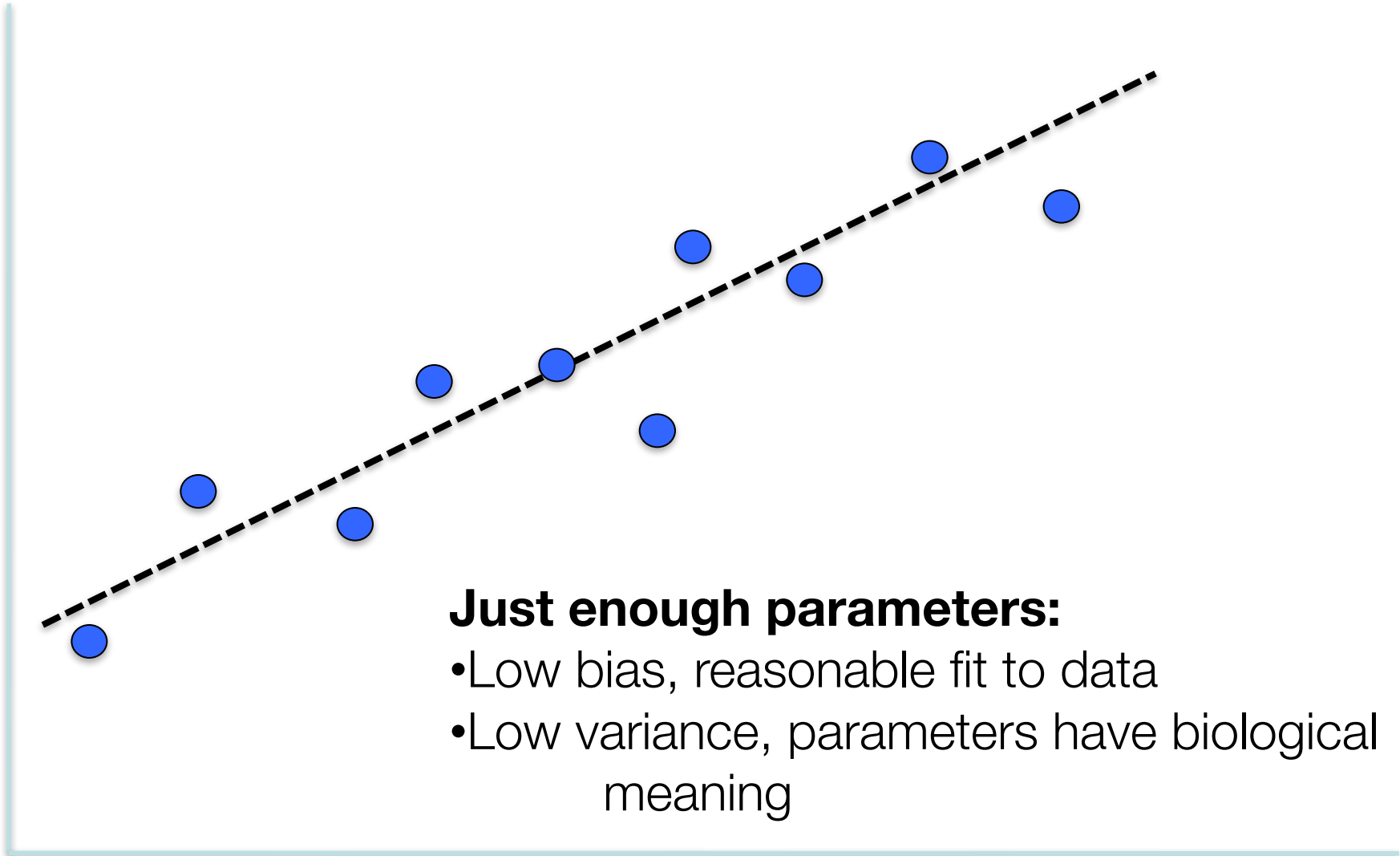
Model selection



Too many parameters:

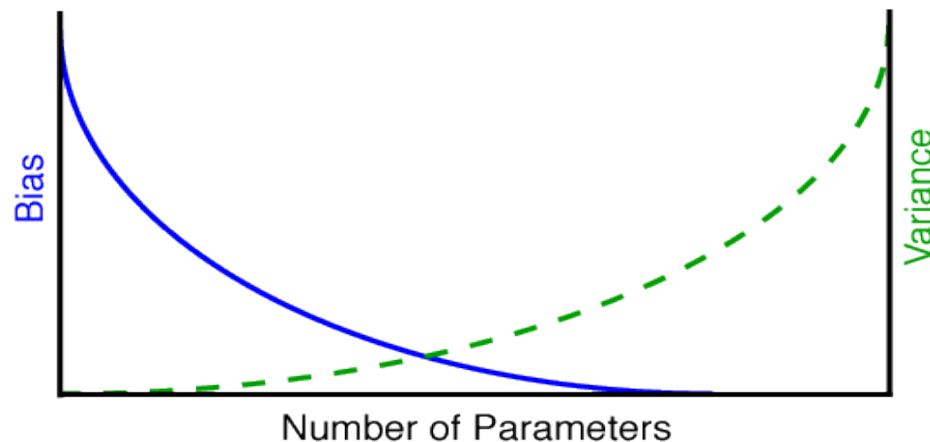
- Low bias, fits the data very well
- Too many parameters, tell us little about the biological process that gave rise to the data

Model selection



Model selection

- Adding more parameters *always* improves the fit of the model to the observed data
- More parameters \rightarrow higher R^2 and better likelihood
- But it doesn't necessarily improve the model!
- 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)**
 $AIC = -2\ln(\text{likelihood}) + k$
 - **Bayesian information criterion (BIC)**

Likelihood-ratio test

- **Likelihood ratio = $2(\ln L_1 - \ln L_0)$**

L_0 is the likelihood of the null model

L_1 is the likelihood of the alternative model

- Used to compare nested models, such as:
 - HKY vs GTR substitution model
 - Strict clock vs unconstrained model

GTR

a, b, c, d, e, f

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

HKY

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

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

Likelihood-ratio test

- Test statistic is χ^2 -distributed
(d.f. = diff. in number of parameters)
- When multiple models are compared hierarchically, outcome can depend on order of tests
- χ^2 approximation
- Might be inappropriate when null model involves fixing a parameter at boundary of possible values
- Performs poorly when competing models are not nested

Akaike information criterion

- **$AIC = -2\ln L + 2p$**

L is the likelihood under the model

p is the number of parameters in the model

- Balances likelihood against number of parameters
- Prefer models with smaller AIC values
- Can be used to compare non-nested models, such as:
 - HKY+I vs GTR+G substitution model

Bayesian information criterion

- **$\text{BIC} = -2\ln L + p\ln(n)$**

L is the likelihood under the model

p is the number of parameters in the model













n is the sample size (sequence length)

- Stronger penalty on number of parameters
- Prefer models with smaller BIC values

Data Partitioning

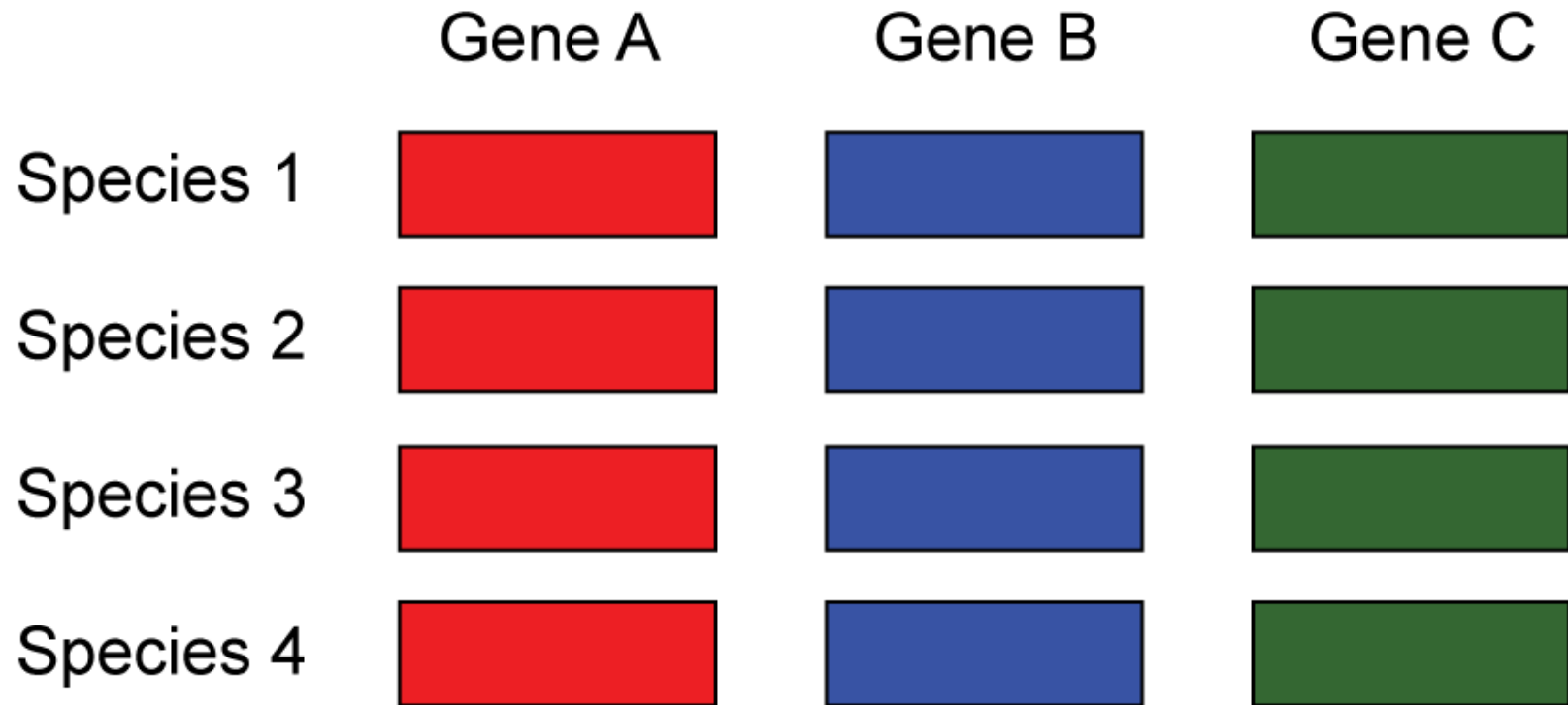
Data partitioning

- Single substitution model across 3 genes

	Gene A	Gene B	Gene C
Species 1			
Species 2			
Species 3			
Species 4			

Data partitioning

- Separate substitution model for each gene



Data partitioning

- Separate substitution model for each gene and codon position

