Lecture 1.2

Evolutionary Models

Popular phylogenetic methods

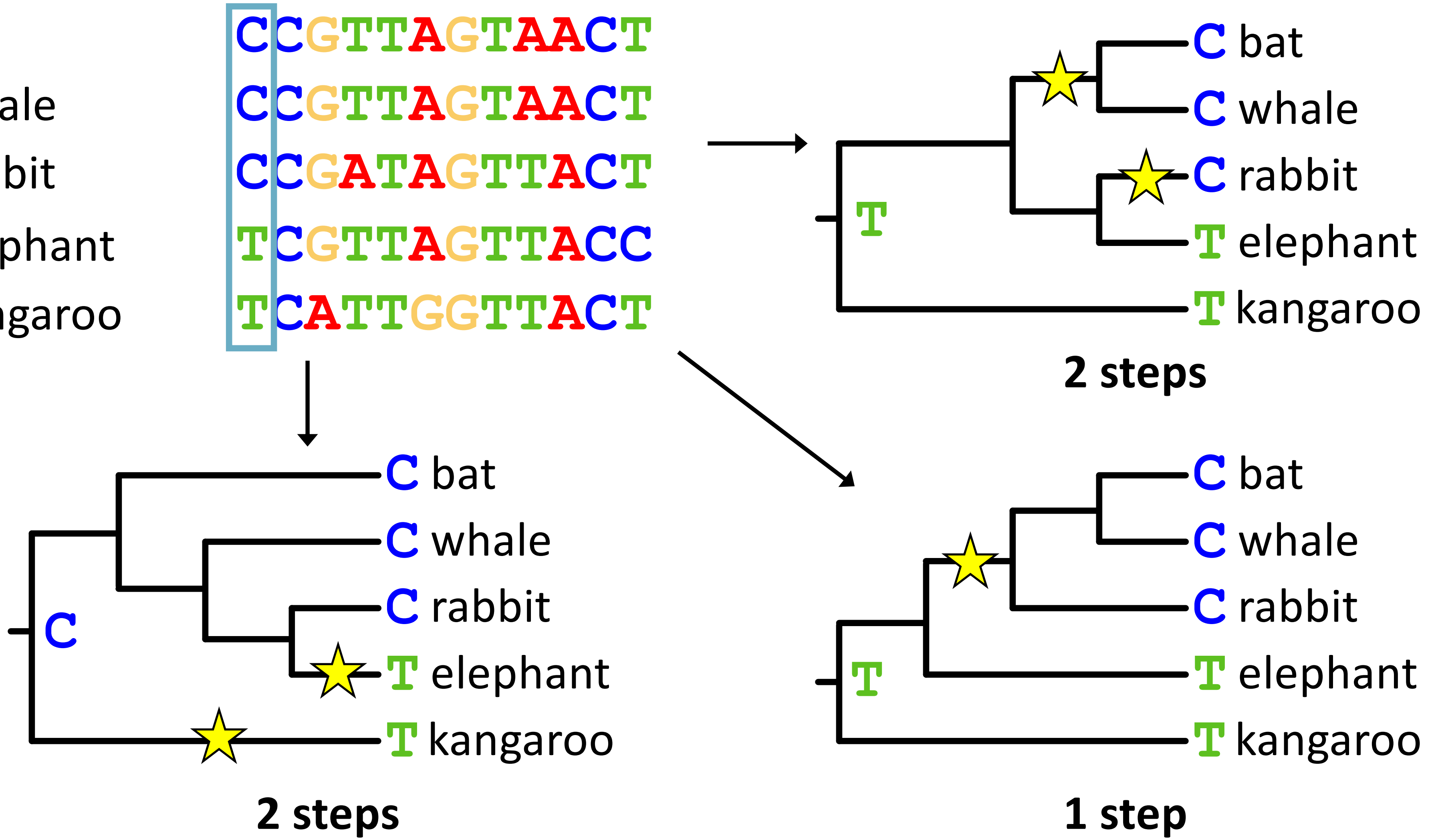
1. Maximum parsimony
2. Distance-based methods
3. Maximum likelihood
4. Bayesian inference

Model-based methods

Maximum Parsimony

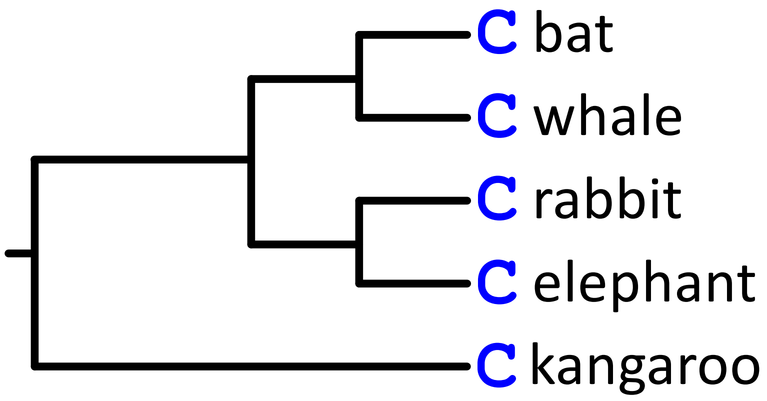
Maximum parsimony

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

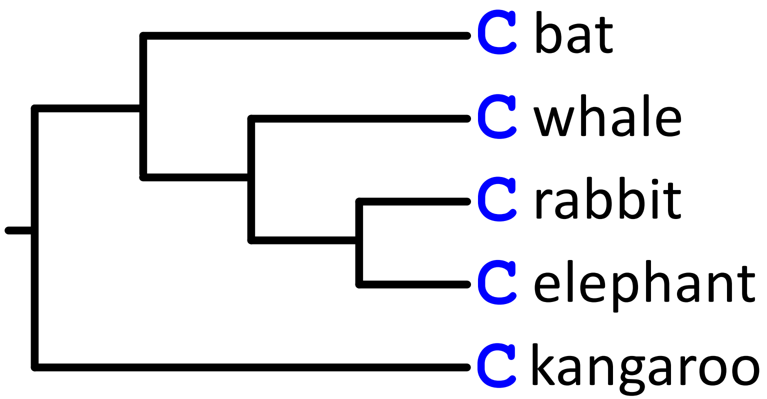


Maximum parsimony

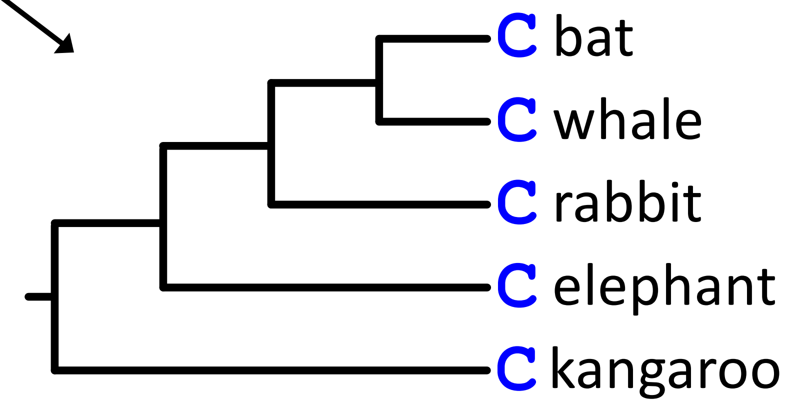
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



0 steps



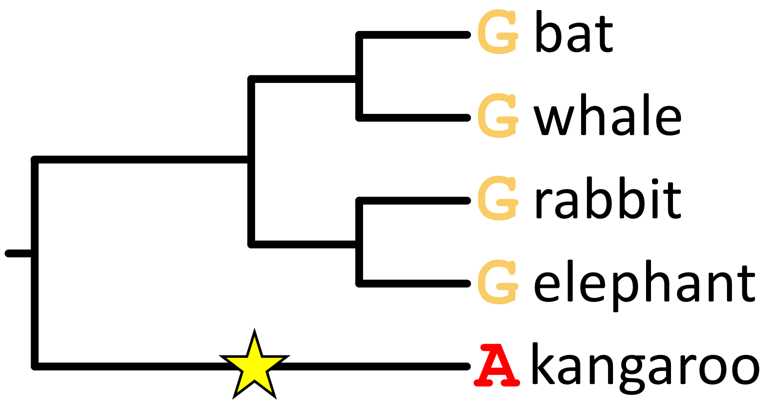
0 steps



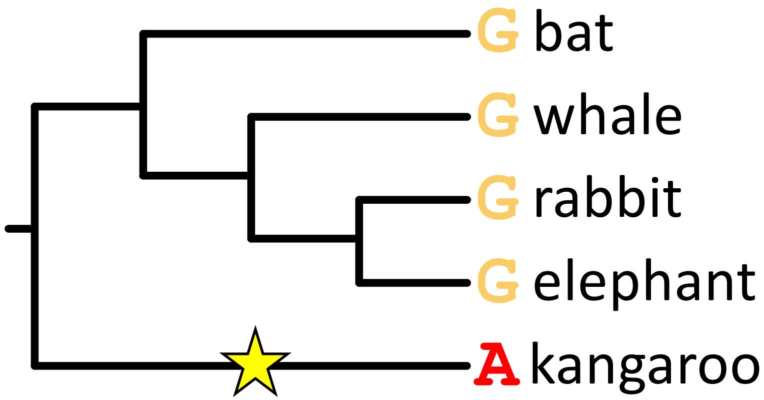
0 steps

Maximum parsimony

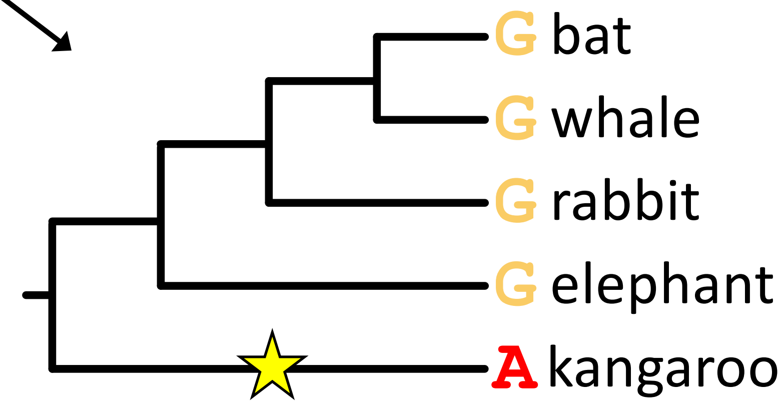
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



1 step



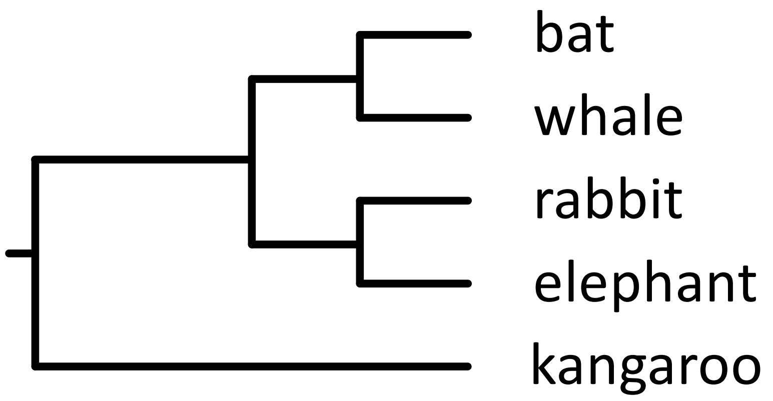
1 step



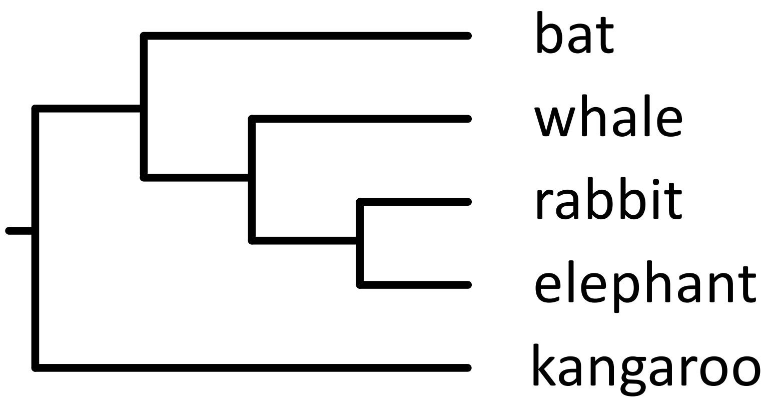
1 step

Maximum parsimony

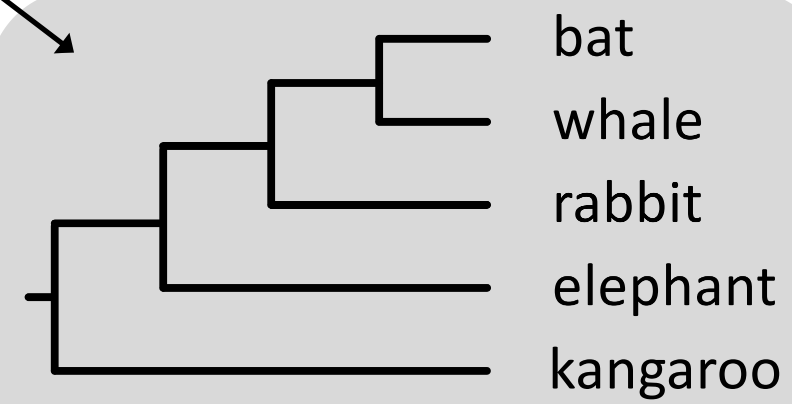
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



7 steps



8 steps



6 steps

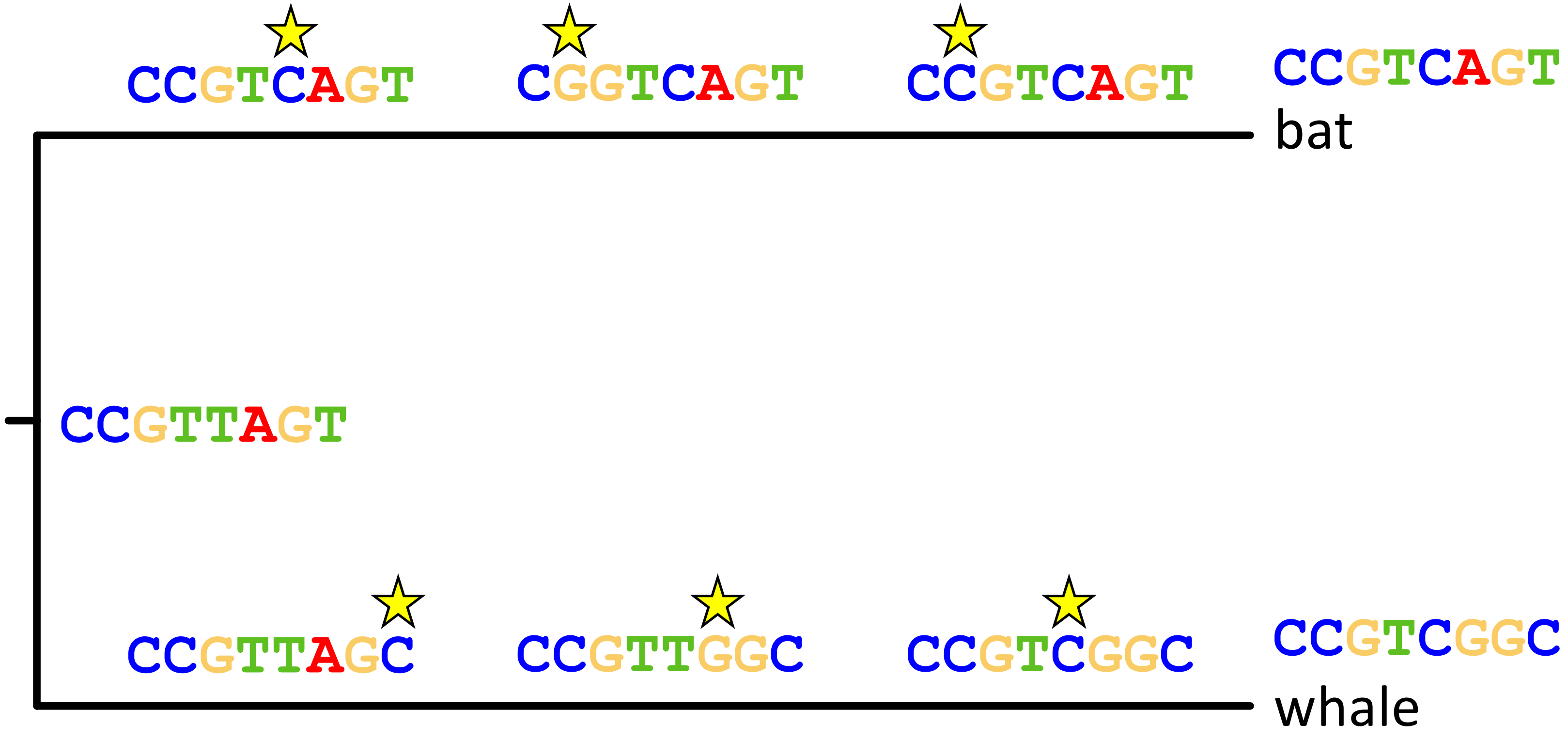
Maximum parsimony

- 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
    - Effects of multiple substitutions
    - Computationally intensive
    - Cannot estimate evolutionary rates or timescales

Multiple hits



Multiple hits

- Maximum parsimony does not account for multiple hits
- Leads to a problem known as **long-branch attraction**
  - Long branch = many substitutions
  - Similarities arise by chance
  - Long branches cluster together
- Multiple hits cause loss of evolutionary signal (**substitution saturation**)

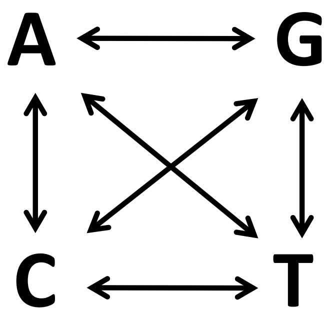
We can correct for multiple hits using substitution models

Substitution Models

Nucleotide substitution models

- Model describing the process of DNA sequence evolution
  - Parameters describing the relative rates of the different pairwise mutations (A → G, G → T, etc.)
  - Parameters describing the frequencies of the four nucleotides

rate matrix

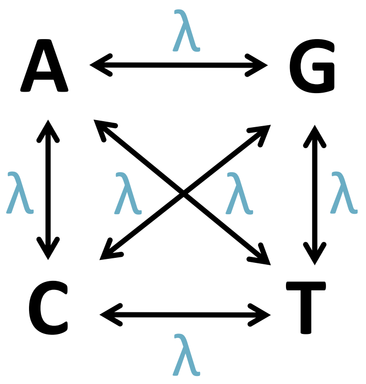


base frequencies

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

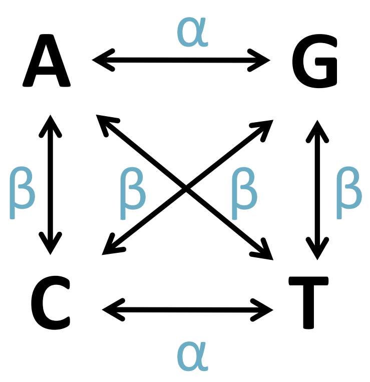
Nucleotide substitution models

JC



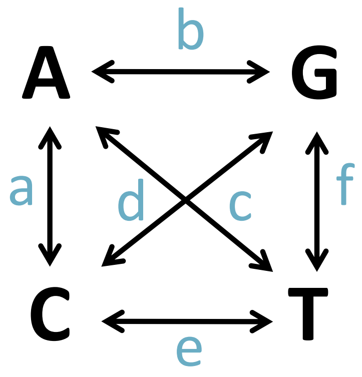
$\pi_A=\pi_C=\pi_G=\pi_T$

HKY



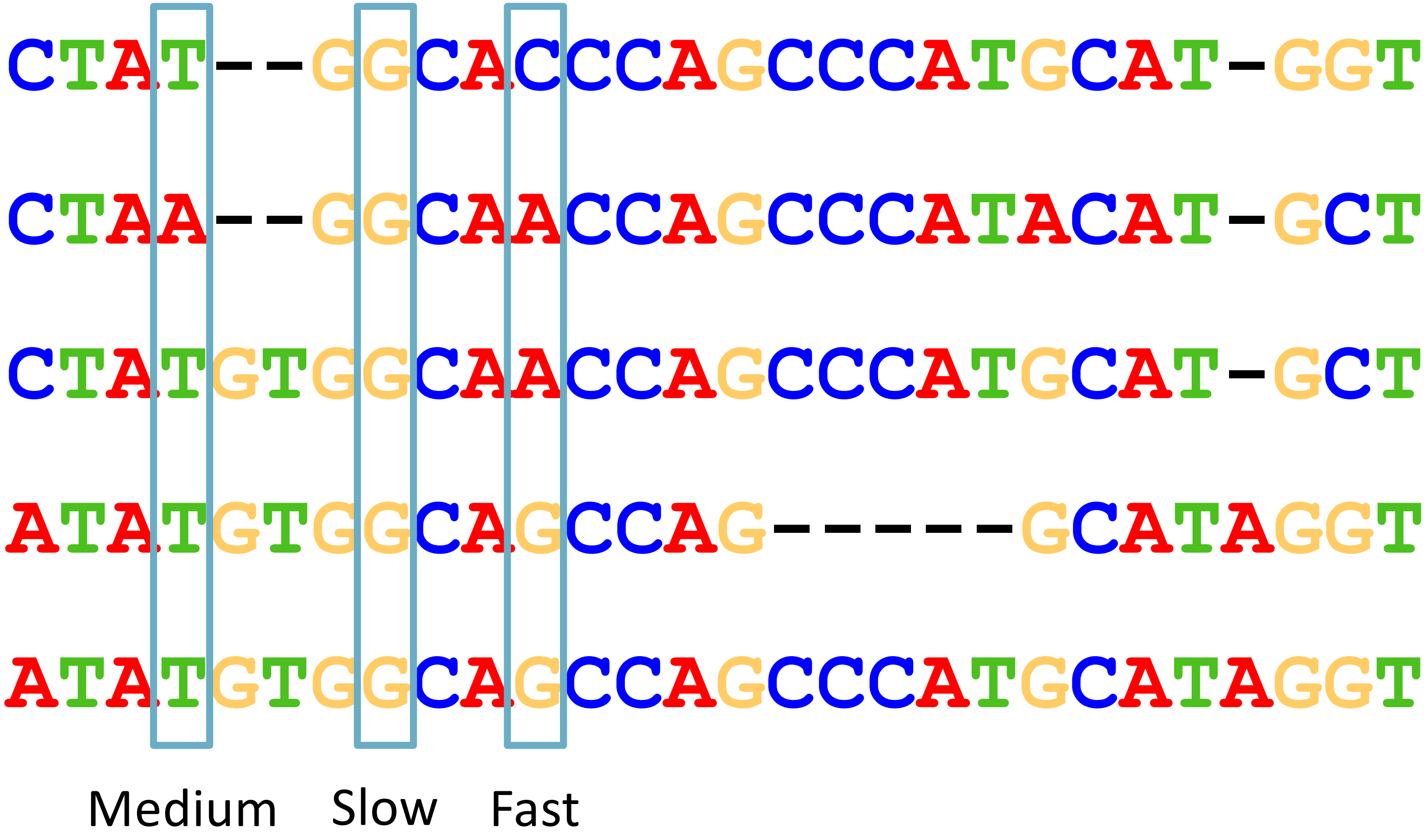
$\pi_A, \pi_C, \pi_G, \pi_T$

GTR



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

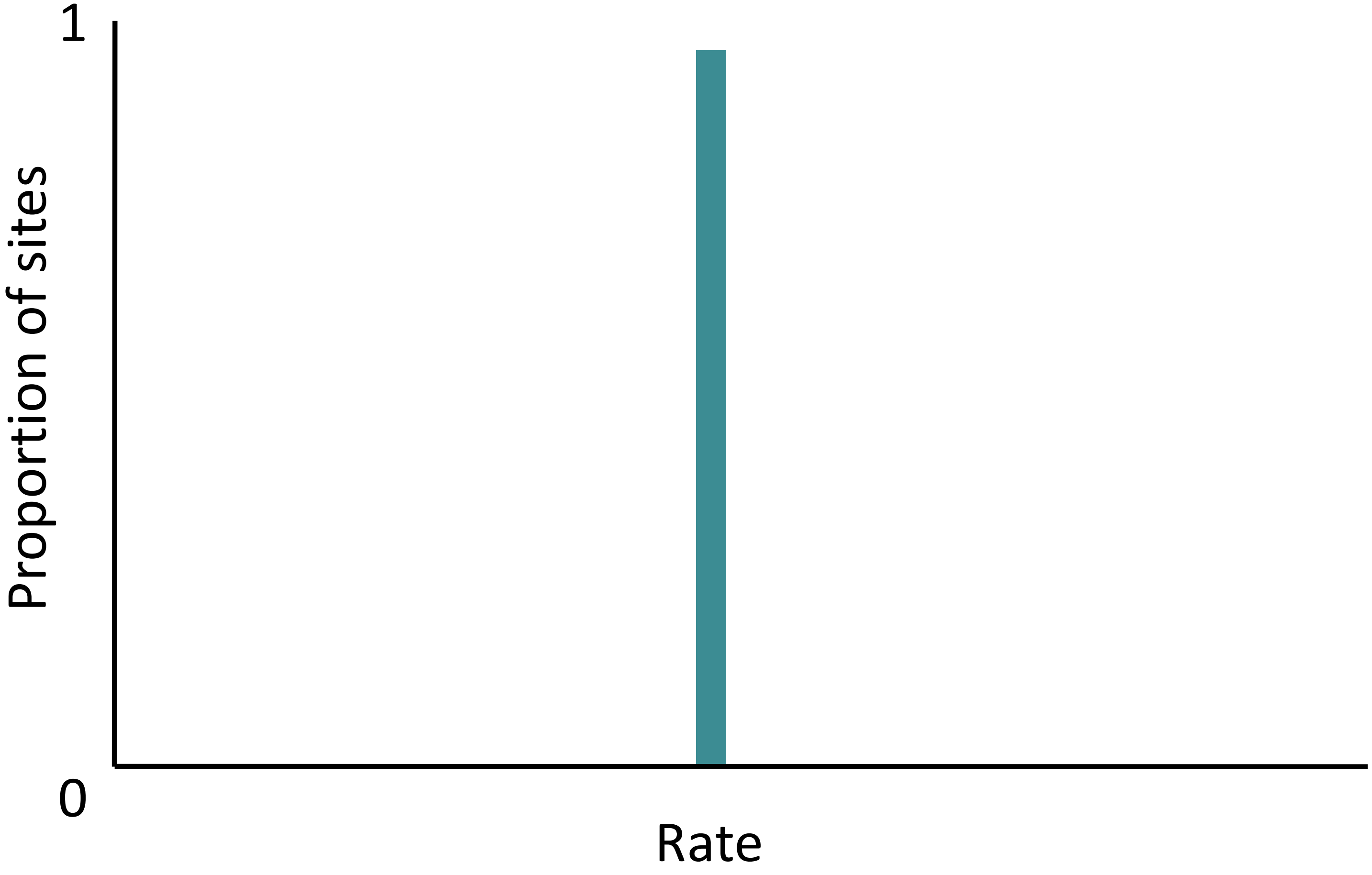
Rate variation across sites



```
CTAT--GGCACCCAGCCCATGCAT-GGT
CTAA--GGCAACCAGCCCATACAT-GCT
CTATGTGGCAACCAGCCCATGCAT-GCT
ATATGTGGCAGCCAG-----GCATAGGT
ATATGTGGCAGCCAGCCCATGCATAGGT
```

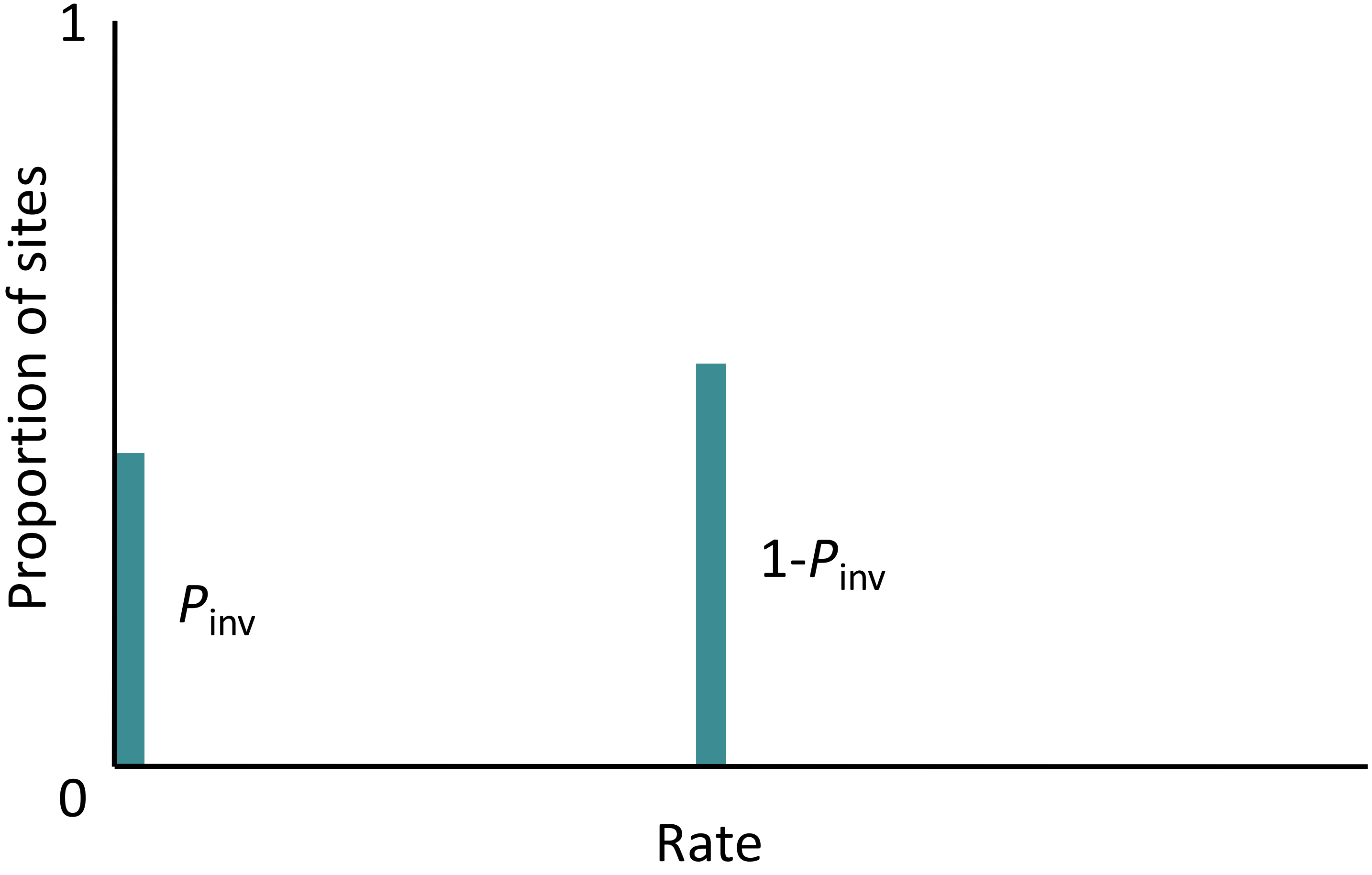
Rate variation across sites

- Equal rates among sites



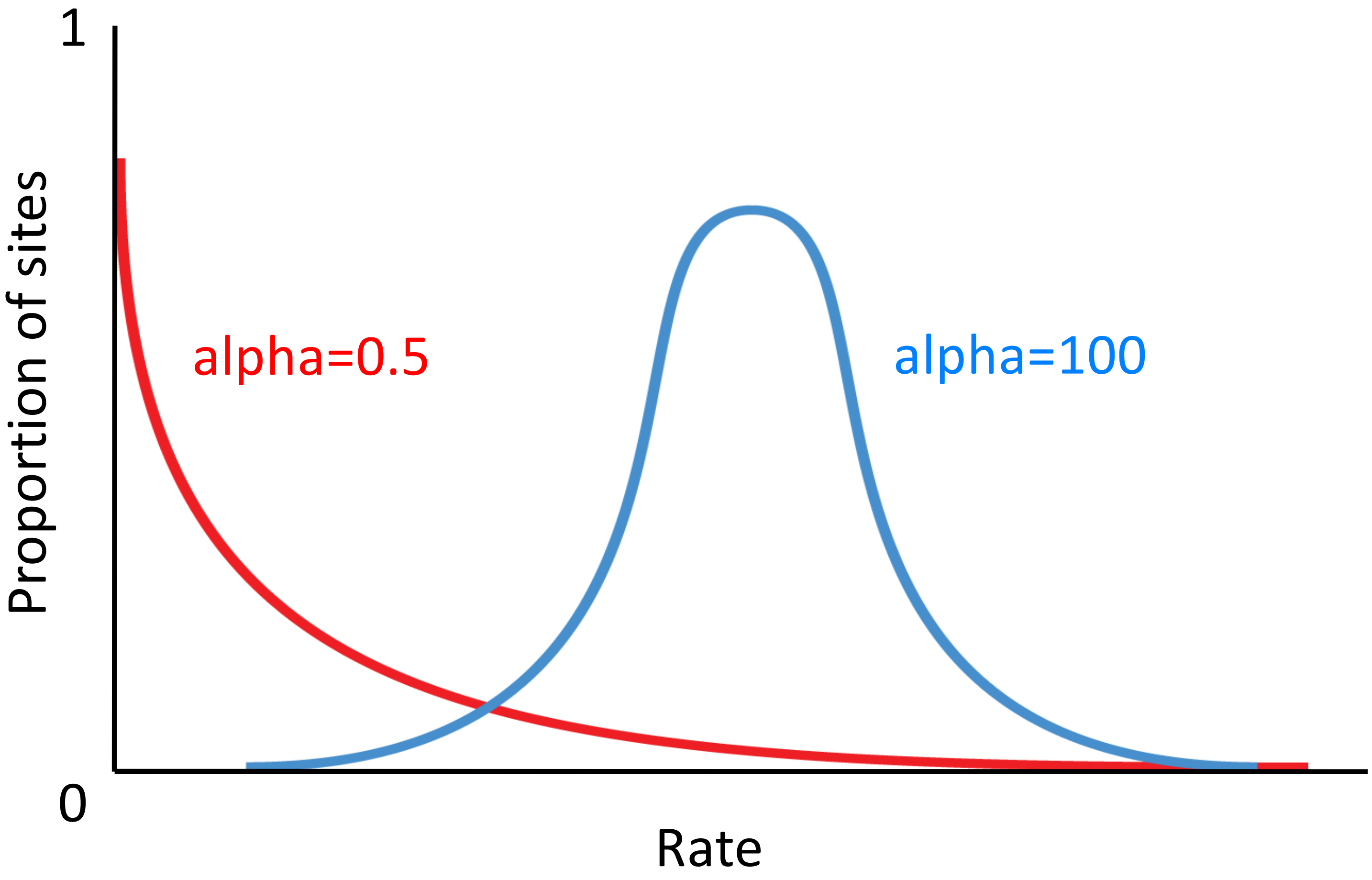
Rate variation across sites

- Proportion of invariable sites (**+I** models)



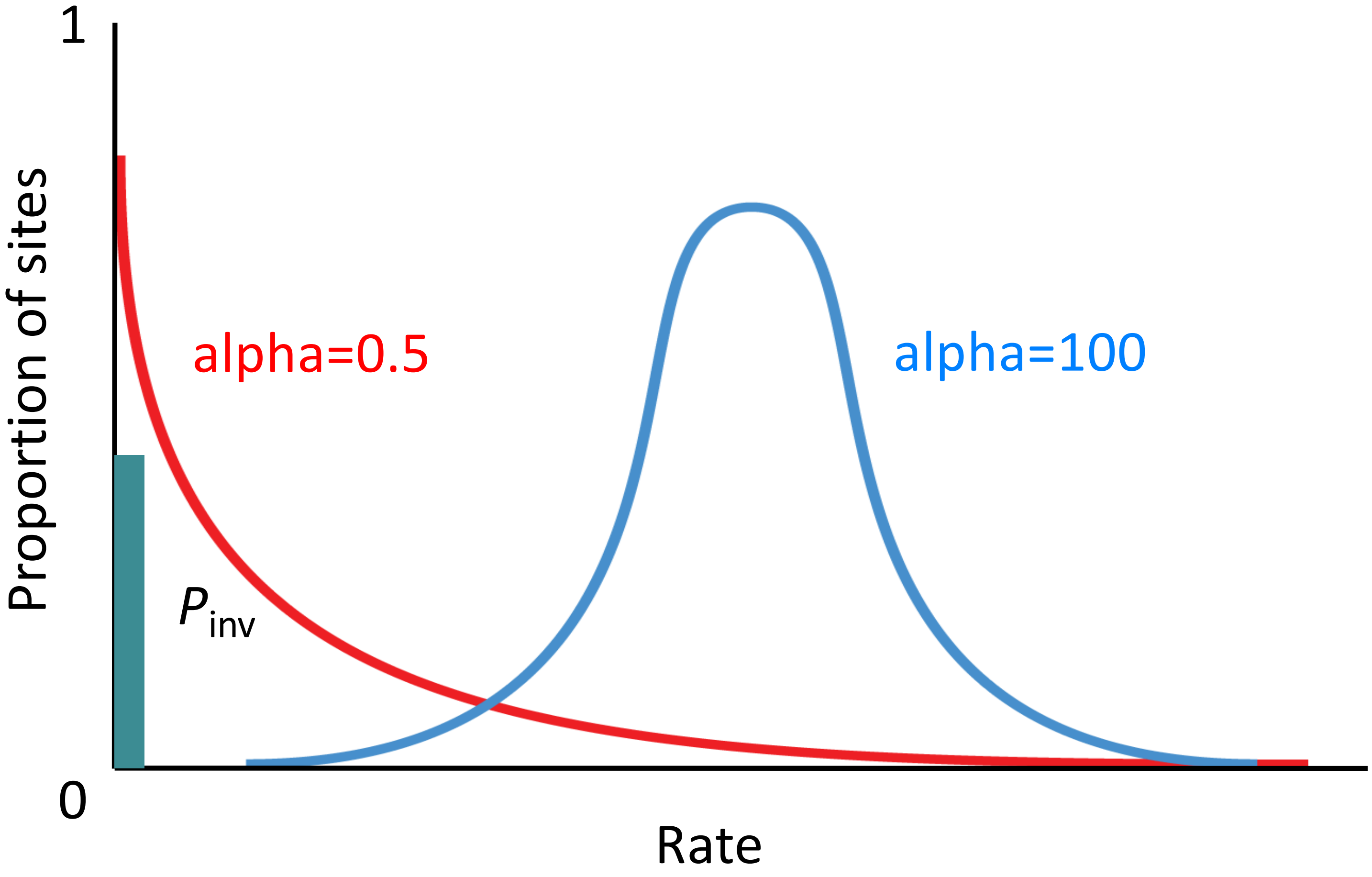
Rate variation across sites

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



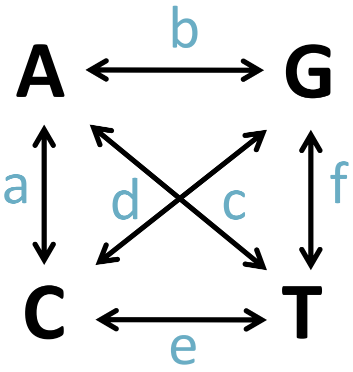
Rate variation across sites

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



Nucleotide substitution models

rate matrix



base frequencies

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

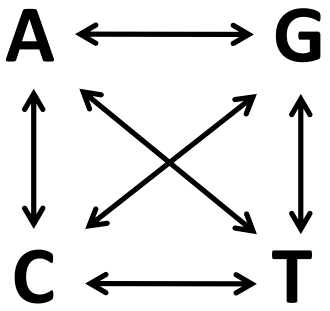
site rates

+ I + G

most complex time-reversible model:	**GTR+I+G** a, b, c, d, e, f π_A, π_C, π_G, π_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

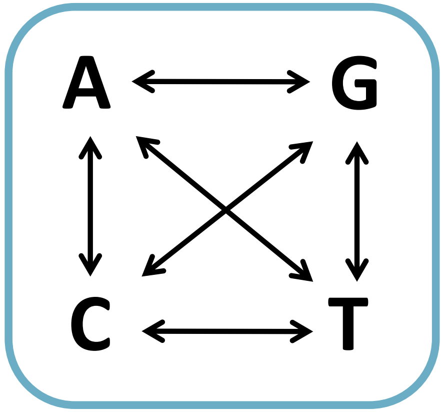
Proportion of invariable sites

- 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

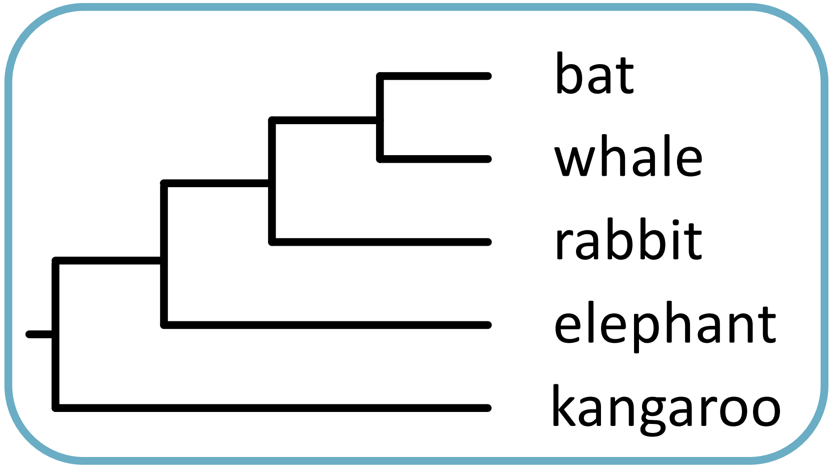
Fundamental assumptions

reversible



stationary

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



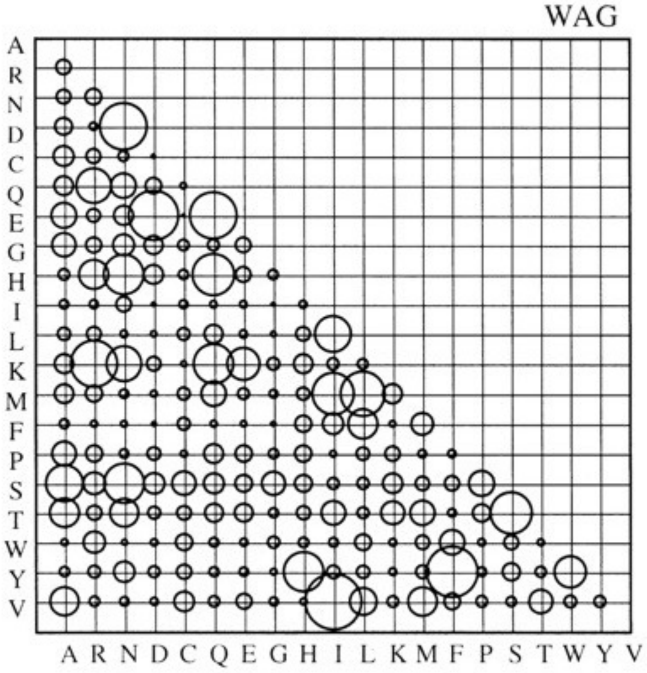
homogeneous

CCGTTAGTAACT
CCGTTAGTAACT
CCGATAGTTACT
TCGTTAGTTACC
TCATTGGTTACT

independent across sites

Amino acid substitution matrices

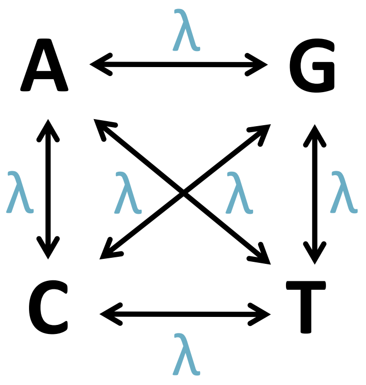
- 20x20 matrix of substitution probabilities
- Too many parameters to estimate
  - GTR model for DNA: 6 parameters
  - GTR model for proteins: 190 parameters
- Estimate substitution probabilities using large data set
  - PAM
  - BLOSUM
  - JTT
  - WAG



Model Selection

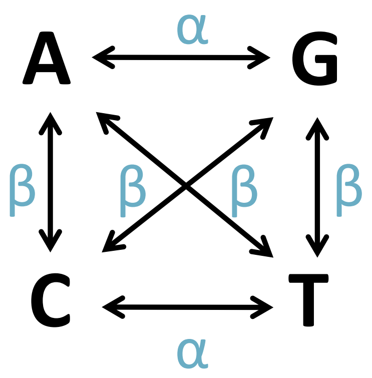
Nucleotide substitution models

JC



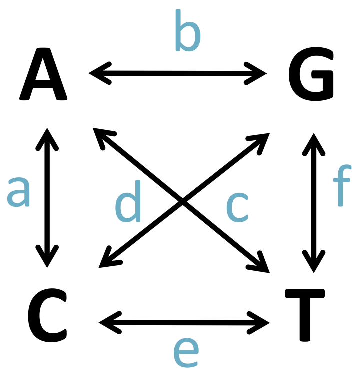
$\pi_A=\pi_C=\pi_G=\pi_T$

HKY



π_A, π_C, π_G, π_T

GTR



π_A, π_C, π_G, π_T

How do we choose a model for our data set?

Model selection

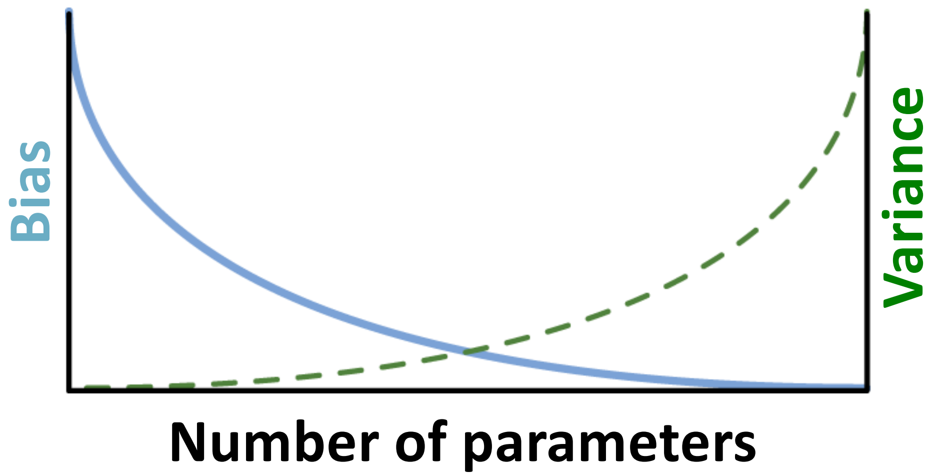
1. **Subjective model selection**
   - Pick a model that seems sensible
   - Balance the number of parameters against the amount of data
   - Biological motivation

2. **Objective model selection**
   - Use information theory and let a computer do it for you
   - Statistical motivation

Model selection

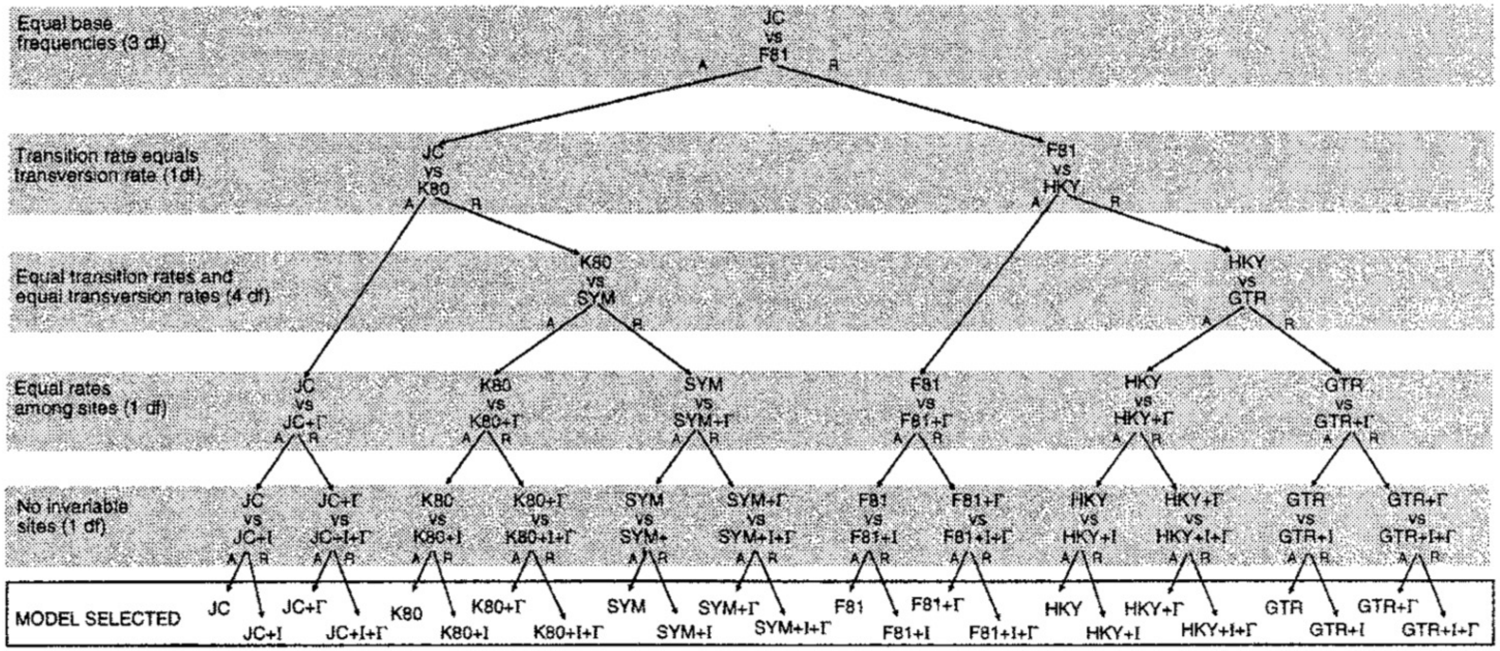
- Adding more parameters *always* improves the fit of the model to the observed data (reduces bias in estimates)
- But more parameters leads to greater variance in the estimates of those parameters

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



Model selection

- General approach is to balance model fit (likelihood) against model complexity (number of parameters)
  - **Likelihood-ratio test (LRT)**
    Used to compare nested models



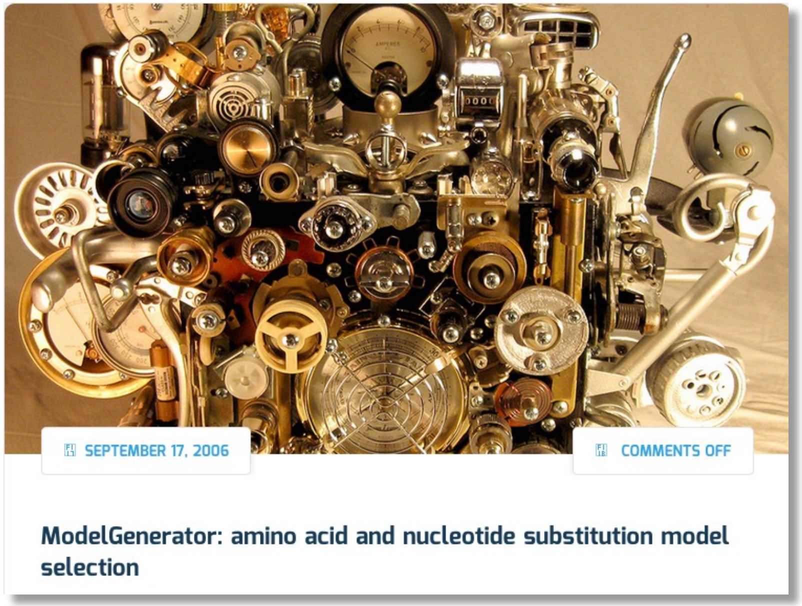
Posada & Crandall (2001) *Syst Biol*

Model selection

- General approach is to balance model fit (likelihood) against model complexity (number of parameters)
  - **Likelihood-ratio test (LRT)**
    Used to compare nested models
  - **Akaike information criterion (AIC)**
    AIC = -2ln(likelihood) + 2k
  - **Bayesian information criterion (BIC)**
    BIC = -2ln(likelihood) + $k$ln(n)

Model selection

- Software for selecting substitution models
  - *MEGA*
  - *MODELTEST*
  - *MODELGENERATOR*
  - ModelFinder (in *IQ-TREE*)



SEPTEMBER 17, 2006

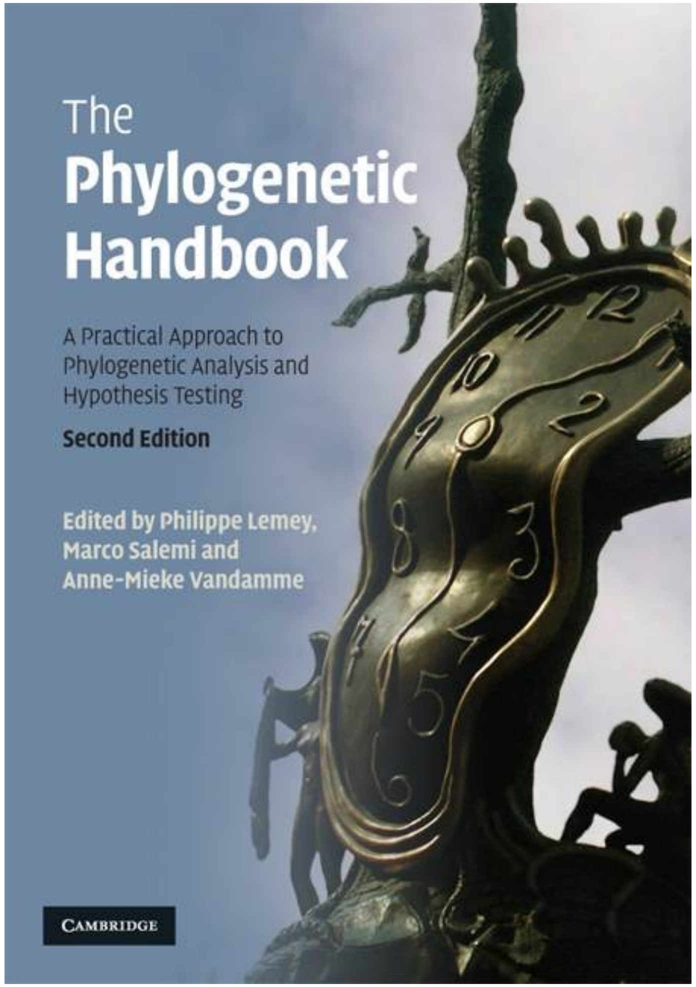
COMMENTS OFF

ModelGenerator: amino acid and nucleotide substitution model selection

Phylogenetic estimates are often robust to choice of substitution 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.


The Phylogenetic Handbook
A Practical Approach to Phylogenetic Analysis and Hypothesis Testing
Second Edition
Edited by Philippe Lemey, Marco Salemi and Anne-Mieke Vandamme
CAMBRIDGE