Lecture 3

Models and support

Bootstrapping

The bootstrap is non parametric

- Uncertainty in the tree estimate can be inferred indirectly using **bootstrap analysis**
- "Pulling oneself up by one's bootstraps"

- Bootstrap analysis can be performed when using a range of phylogenetic methods:
  - Maximum parsimony
  - Distance-matrix based methods
  - Maximum likelihood

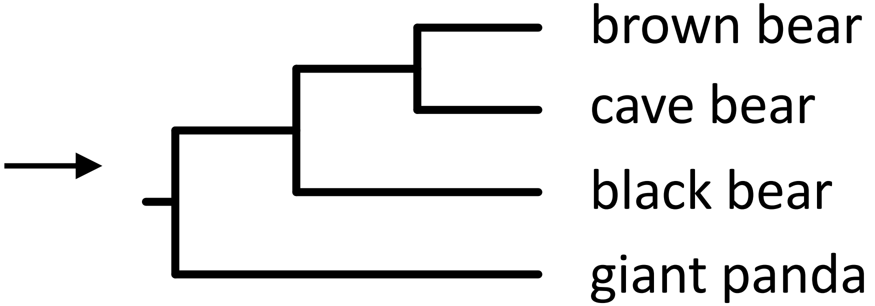
Bootstrap

brown bear	CGTTAGTACACT
cave bear	CGATAGTTCACT
black bear	CGTTAGTTTACC
giant panda	CATTGGTTTACT

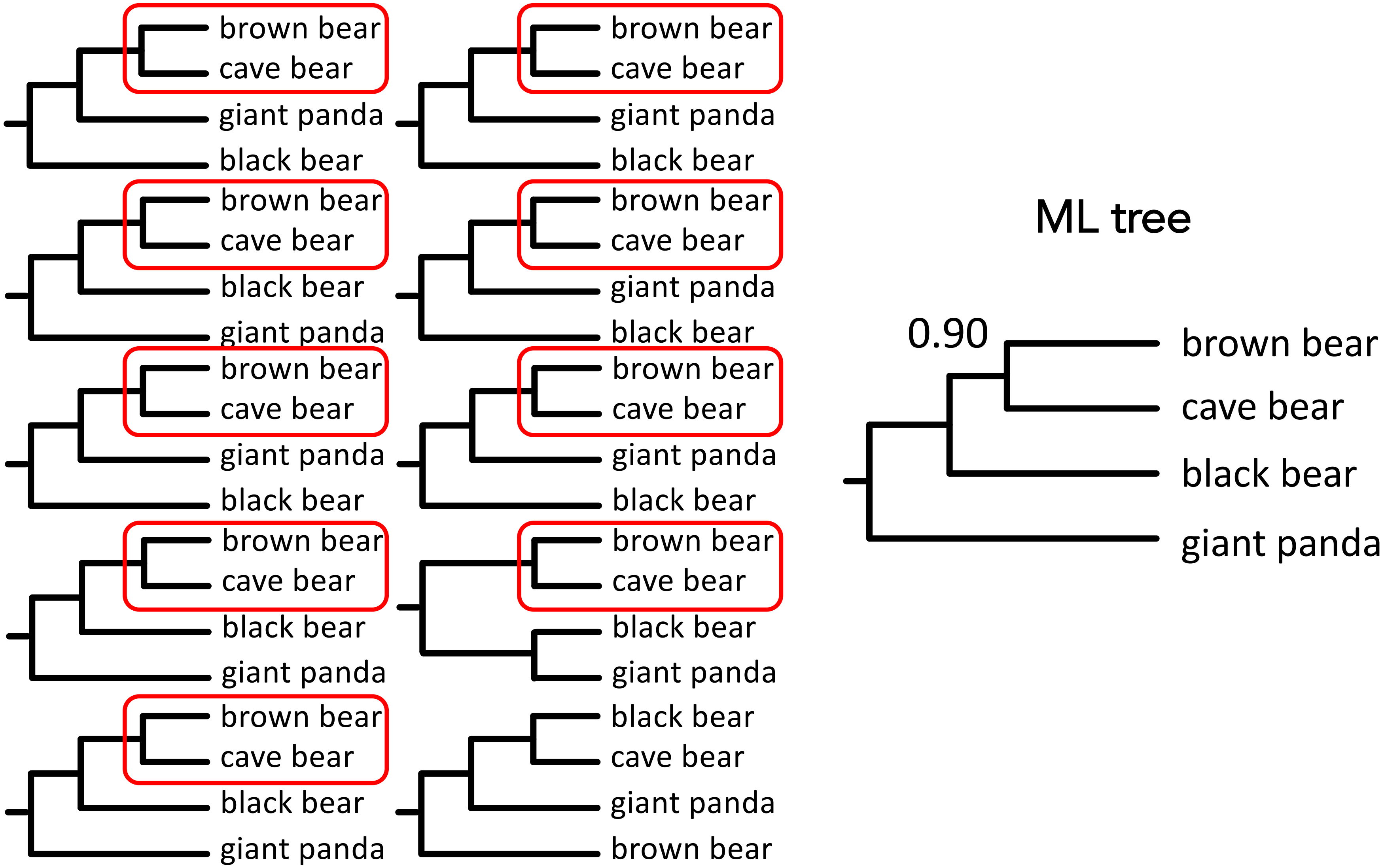
Repeat 1000 times

Pseudoreplicate

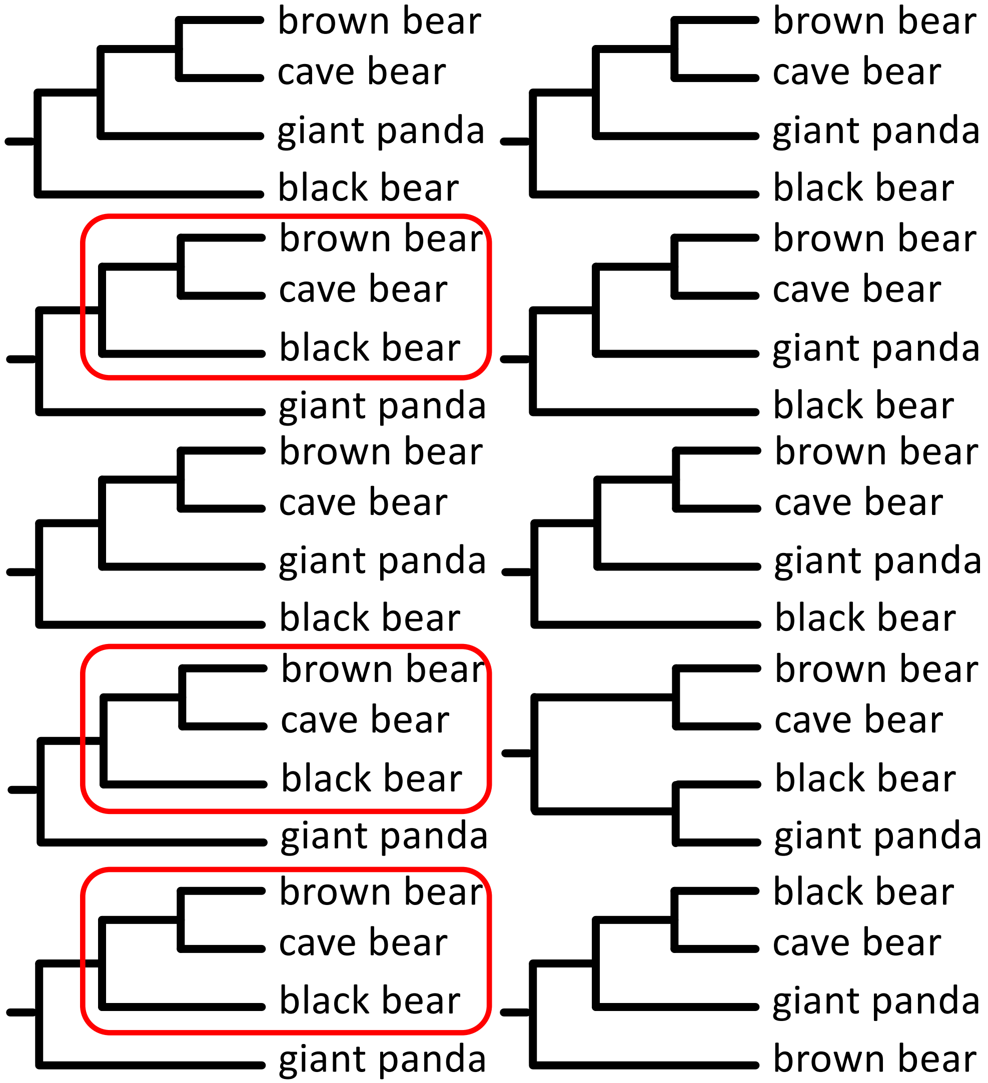
brown bear	ATTACTGTCCCT
cave bear	ATTACTGTCCCA
black bear	ATCACTGTTCCT
giant panda	GTTGCTATTCCT



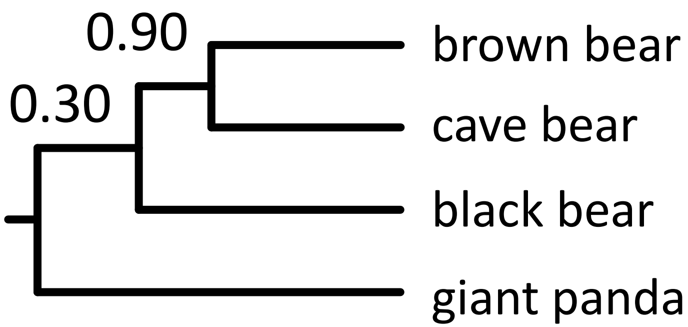
Bootstrap


brown bear
cave bear
giant panda
black bear
brown bear
cave bear
black bear
giant panda
brown bear
cave bear
giant panda
black bear
brown bear
cave bear
black bear
giant panda
brown bear
cave bear
black bear
giant panda
brown bear
cave bear
giant panda
black bear
brown bear
cave bear
giant panda
black bear
brown bear
cave bear
giant panda
black bear
brown bear
cave bear
black bear
giant panda
black bear
cave bear
giant panda
brown bear
ML tree
0.90
brown bear
cave bear
black bear
giant panda

Bootstrap



ML tree



Interpreting bootstrap values

- **Felsenstein (1985)**
  The bootstrap gives us a confidence interval that contains *the tree that would be estimated when repeatedly sampling sites from the existing distribution*

- Bootstrap values are **measures of repeatability**
  - High when lots of data are available
  - Has little meaning when genome-scale data are available

Soltis & Soltis (2003) *Stat Sci*

Popular methods in phylogenetics

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

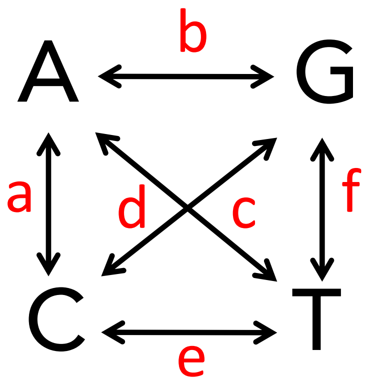
Statistical methods

MEGA

Substitution models

DNA substitution models

Rates 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

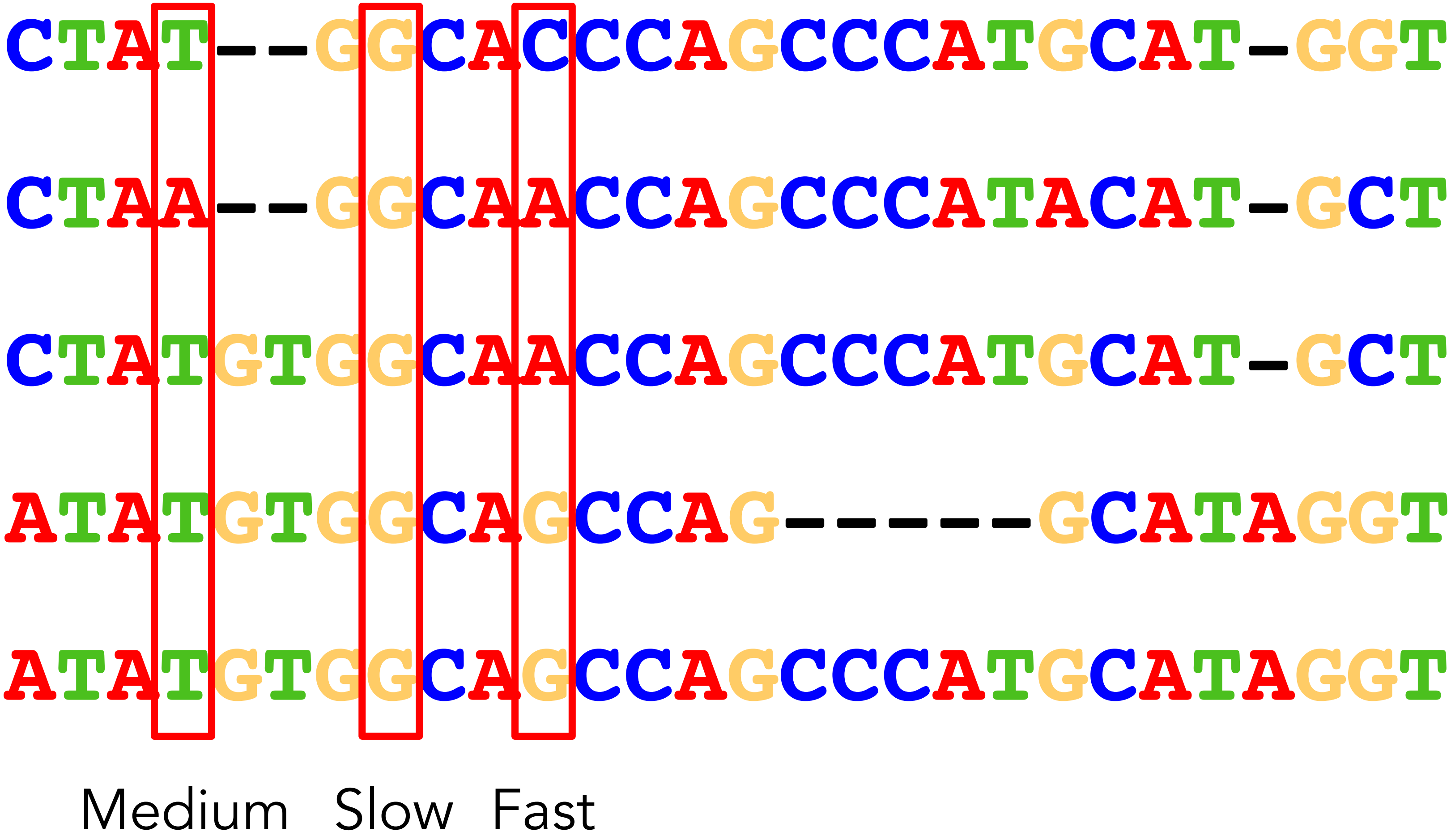
$\pi_A, \pi_C, \pi_G, \pi_T$

GTR

a, b, c, d, e, f

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

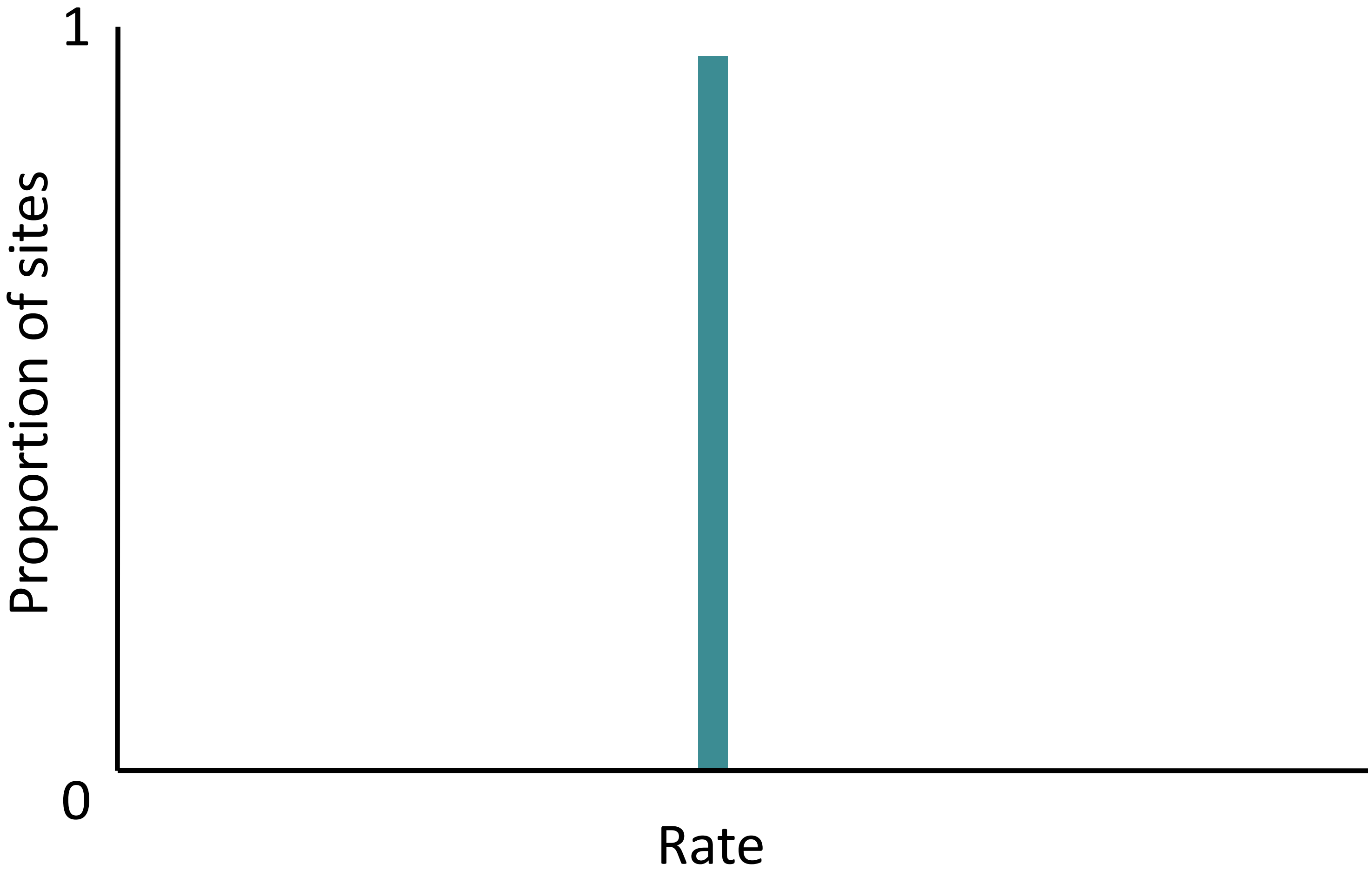
Rate variation across sites



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

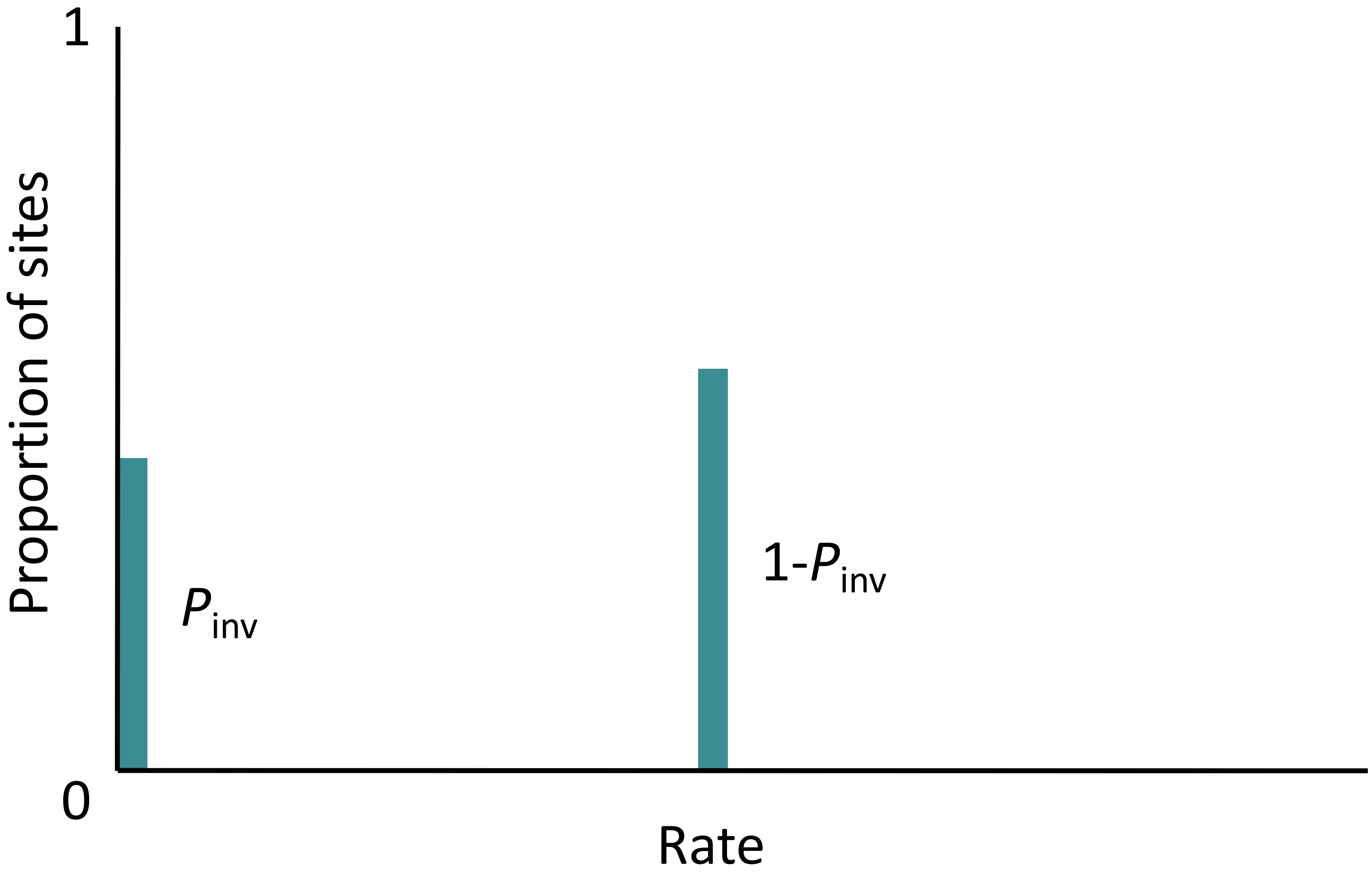
Rate variation across sites

- Identical among all sites



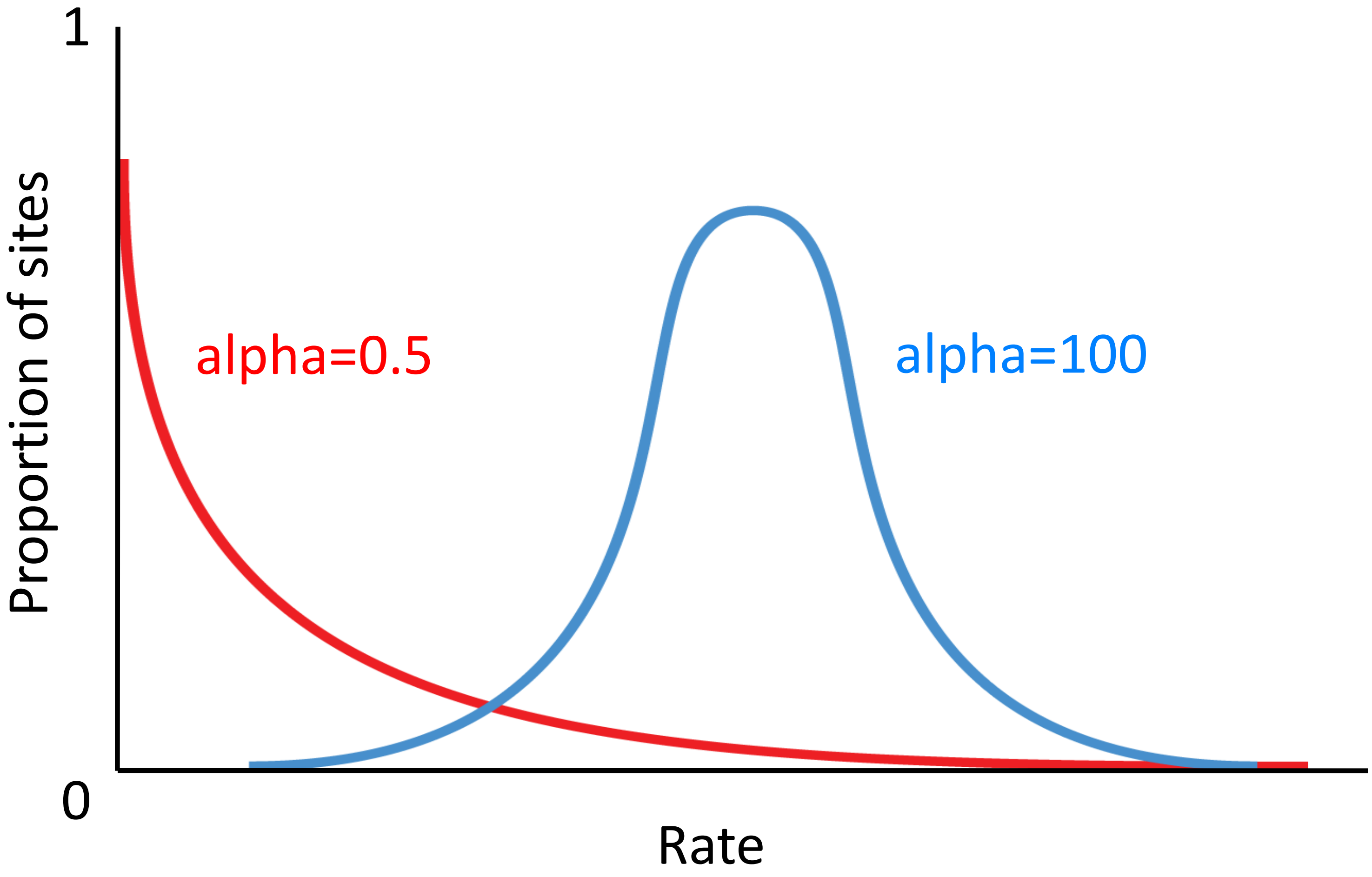
Rate variation across sites

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



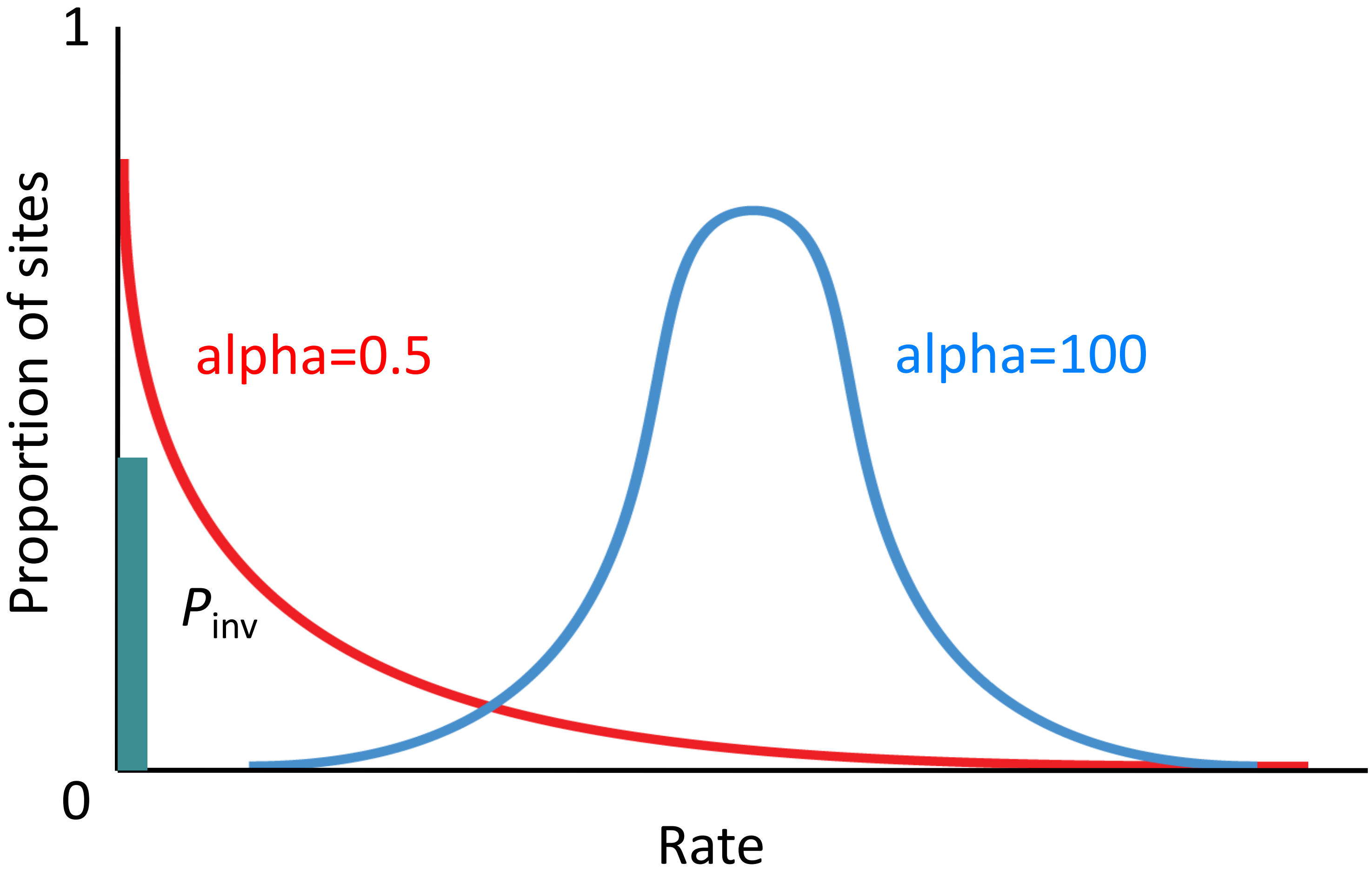
Rate variation across sites

- Gamma distributed rates across sites (**+G** models)



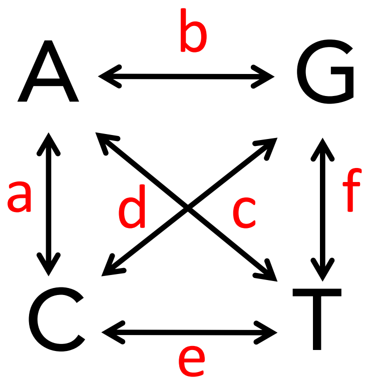
Rate variation across sites

- Rates across sites are assumed to follow a gamma distribution and a portion of invariable sites (**+G+I** models)



DNA substitution models

Rates 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

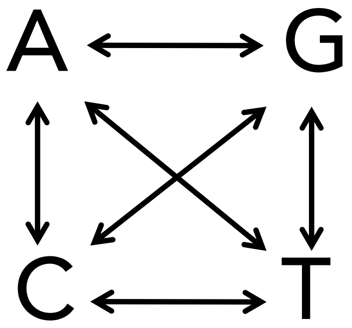
a, b, c, d, e, f

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

I, G

DNA substitution models

Rates matrix



Base frequencies

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

Site rates

+ I + G

Number of models

203 x 15 x 4 = 12,180

In phylogenetics we explore a small portion of these

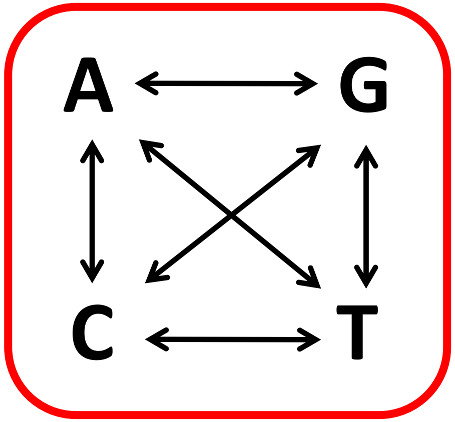
Proportion of invariable sites

- Often over-estimated in species-level analyses
- Do not distinguish:
  - Sites that are **invariable** and cannot change
  - Sites that are **constant** and for stochastic reasons do not have any subtitutions
- Little biological meaning
- Site rates can be adequately described using **+G**

We use +G models to account for variable rates across sites

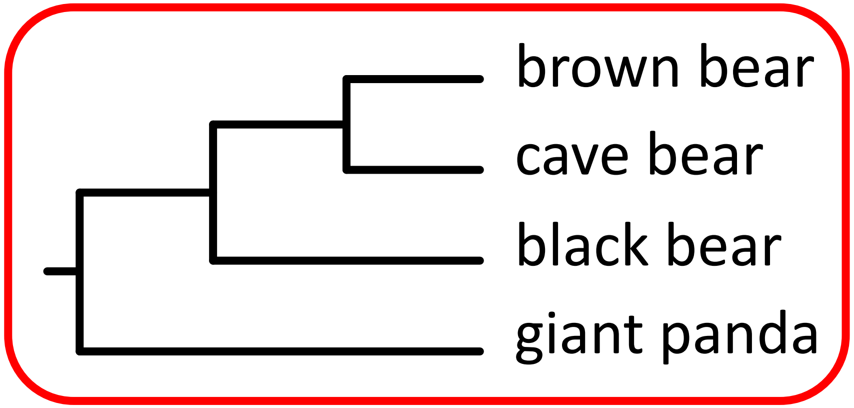
Fundamental assumptions

Reversible



Stationary

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



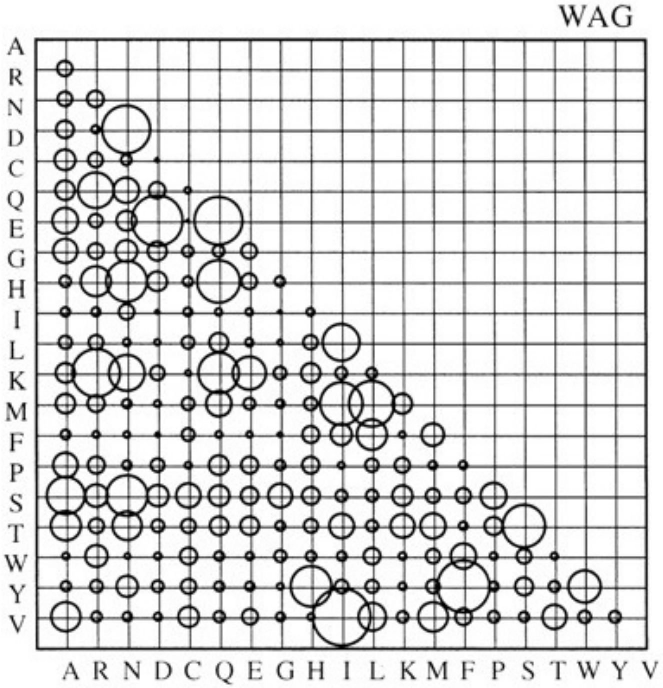
Homogeneous

```
CGTTAGTACACT
CGATAGTTCACT
CGTTAGTTTACC
CATTGGTTTACT
```

Independent sites

Amino acid substitution matrices

- Matrix has size 20x20
- Too many parameters to estimate
  - GTR model for DNA: 6 parameters
  - GTR model for proteins: 190 parameters
- Transition probabilities come from vast data sets
  - PAM
  - BLOSUM
  - JTT
  - WAG



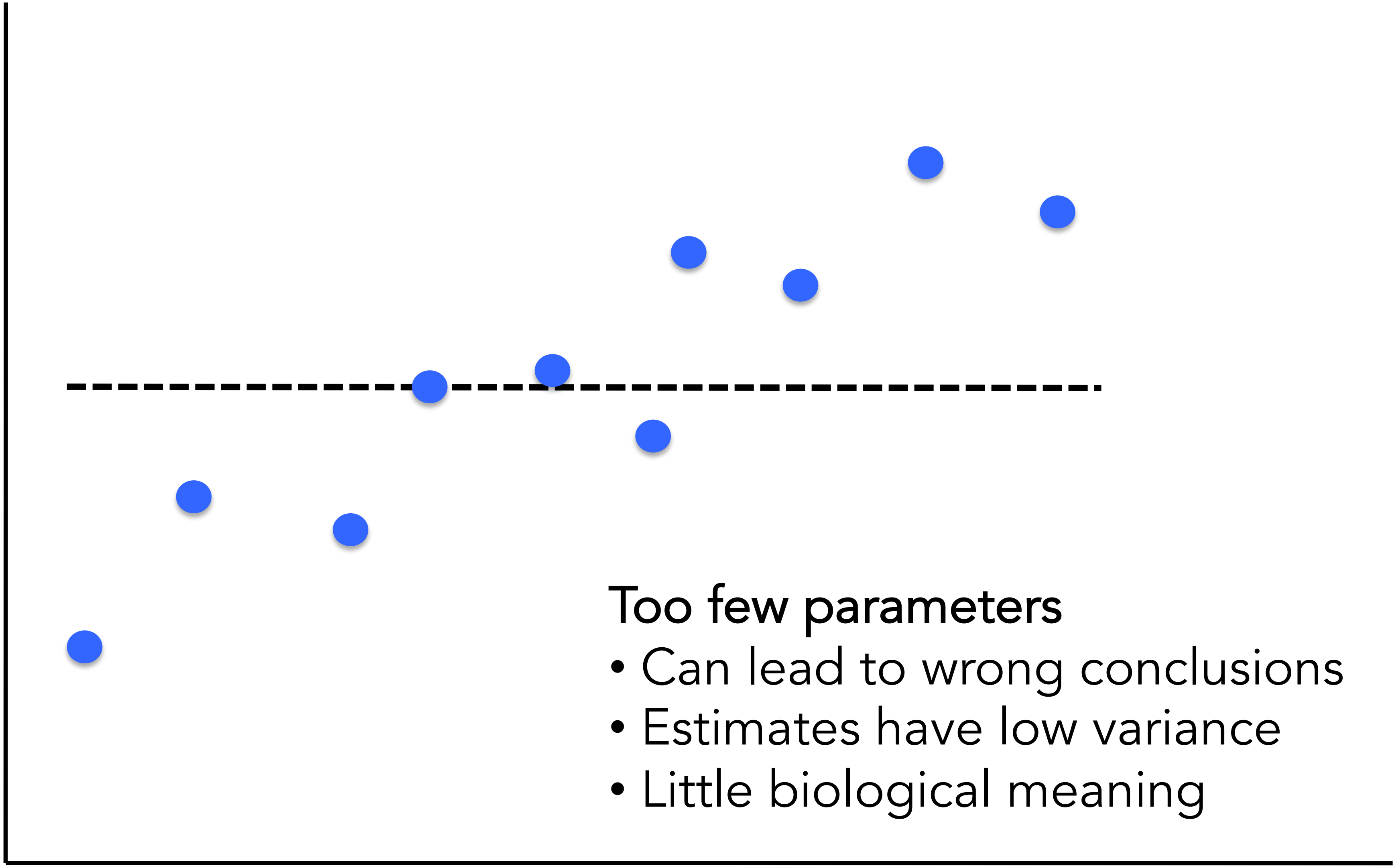
Model selection

Model selection

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

2. **Objective model selection**
   - Automated using information theory
   - Statistical motivation

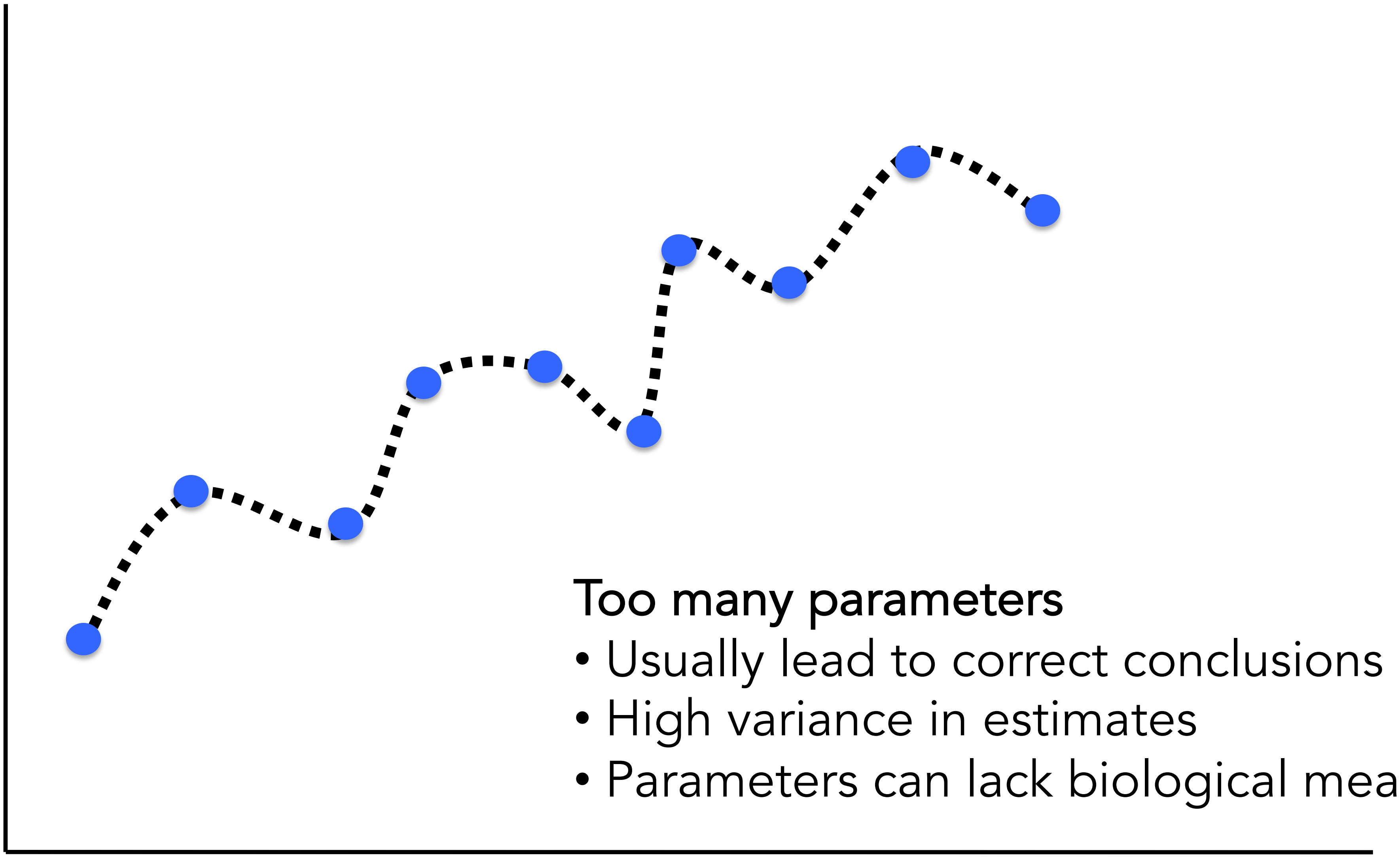
Model selection



Too few parameters

- Can lead to wrong conclusions
- Estimates have low variance
- Little biological meaning

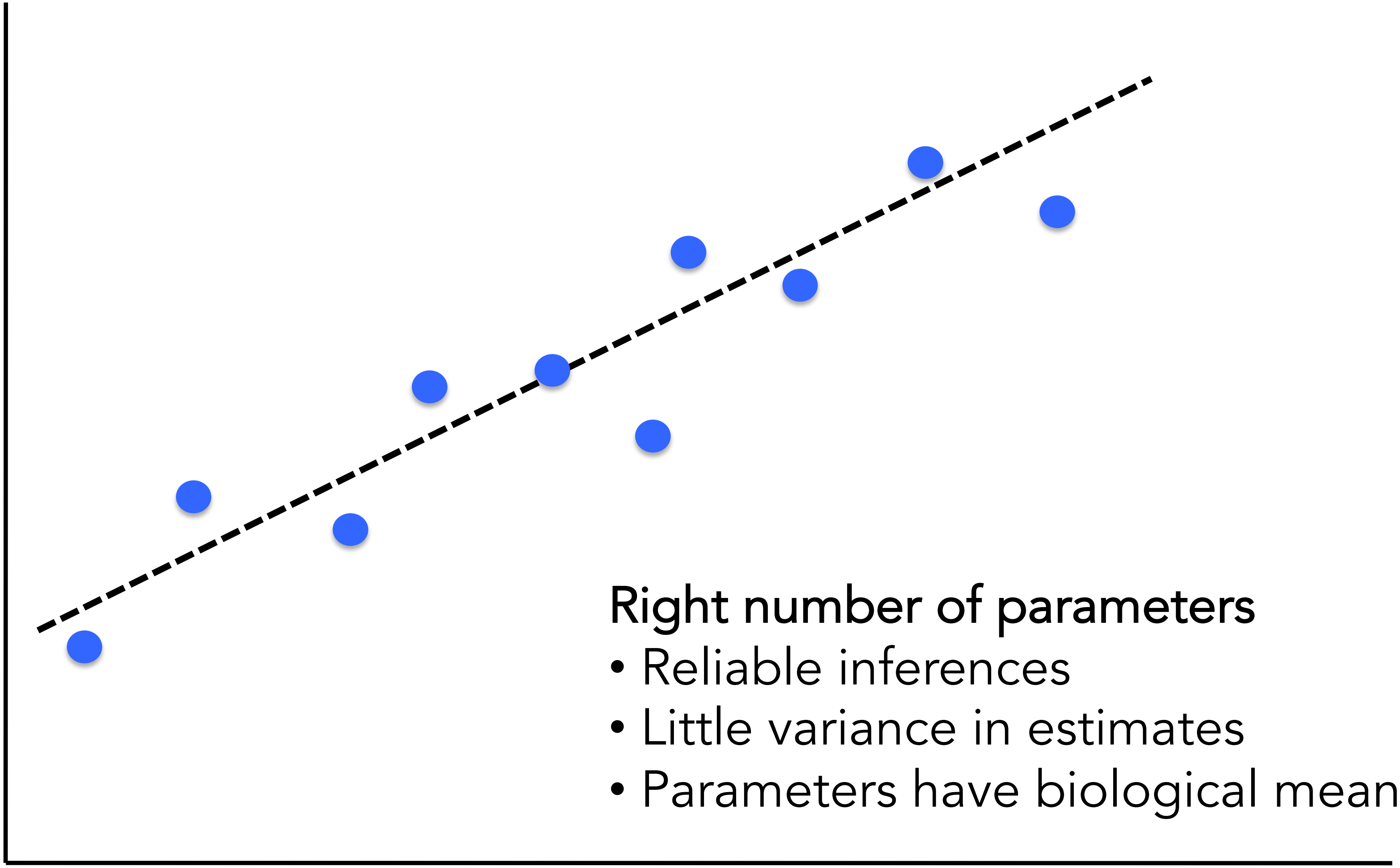
Model selection



Too many parameters

- Usually lead to correct conclusions
- High variance in estimates
- Parameters can lack biological meaning

Model selection



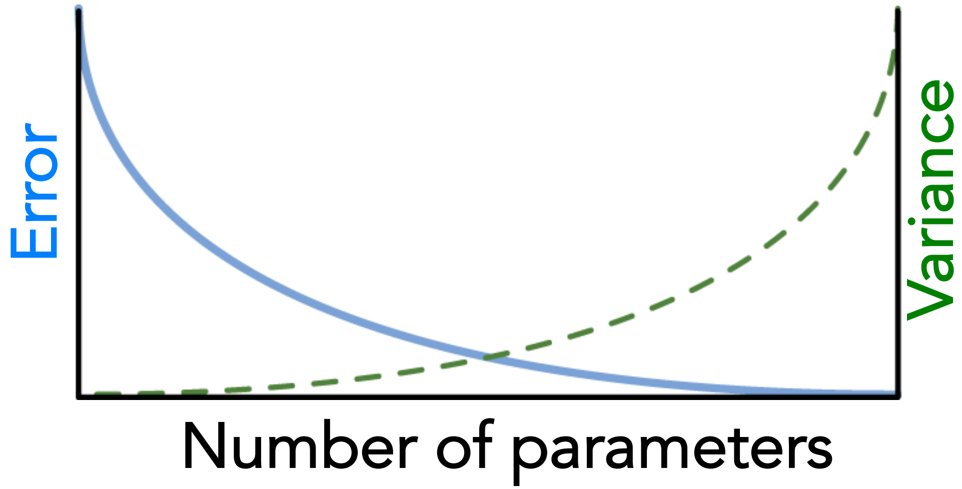
Right number of parameters

- Reliable inferences
- Little variance in estimates
- Parameters have biological meaning

Model selection

- Adding parameters *always* improves model fit
- But adding parameters leads to greater variance in estimates

Is the cost of additional parameters worthwhile?



Model selection

- **Likelihood-ratio test (LRT)**
  Used for comparing nested models

- **Akaike information criterion (AIC)**
  AIC = -2ln(likelihood) + 2k

- **Bayesian information criterion (BIC)**
  BIC = -2ln(likelihood) + $k$ln(n)

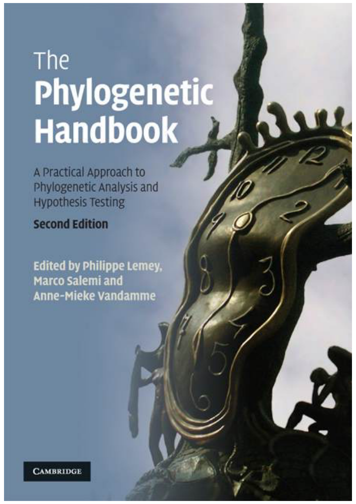
Substitution models in practice

- The tree topology is highly robust to the model used for inference
- GTR+G is acceptable for the majority of data sets

Useful references

- **Model selection in phylogenetics**
  Sullivan & Joyce (2005) *Annual Review of Ecology, Evolution, and Systematics*,
  36: 445–466.

- **The effects of partitioning on phylogenetic inference**
  Kainer & Lanfear (2015) *Molecular Biology and Evolution*, 32: 1611–1627.



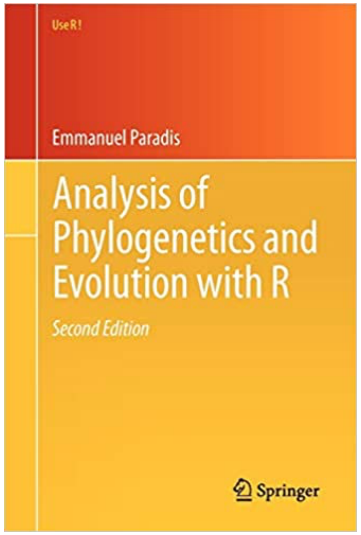
The Phylogenetic Handbook

A Practical Approach to Phylogenetic Analysis and Hypothesis Testing

Second Edition

Edited by Philippe Lemey, Marco Salemi and Anne-Mieke Vandamme

CAMBRIDGE



UseR!

Emmanuel Paradis

Analysis of Phylogenetics and Evolution with R

Second Edition

Springer