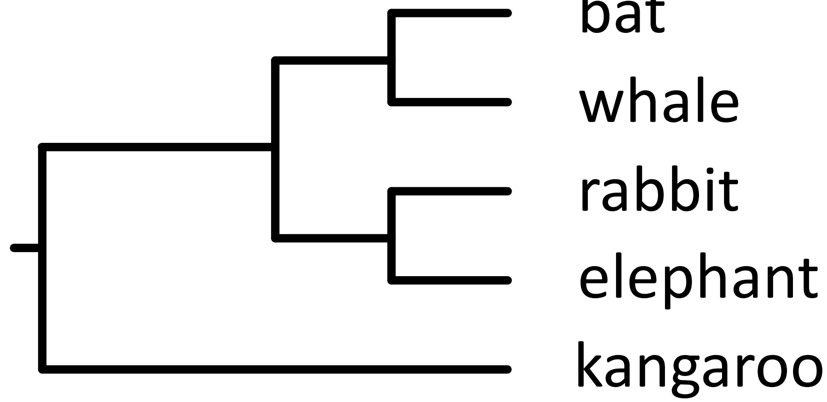
Lecture 1.4

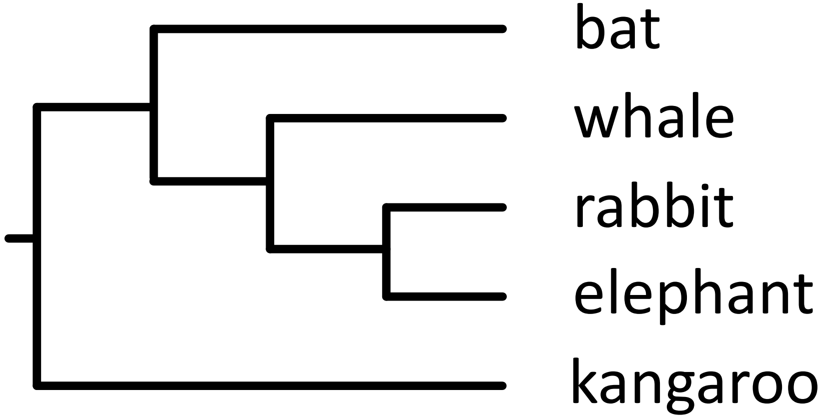
Phylogenetic Methods

Maximum parsimony

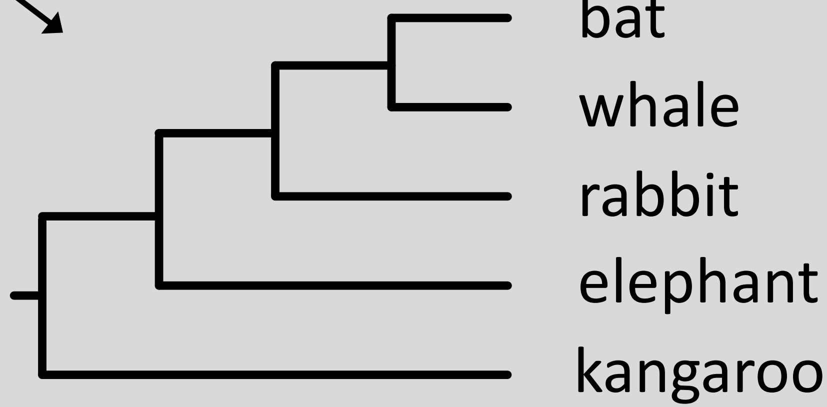
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



7 steps



8 steps



6 steps

Popular phylogenetic methods

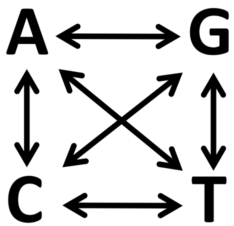
1. Maximum parsimony
2. Distance-based methods
3. Maximum likelihood
4. Bayesian inference

Model-based methods

Distance-Based Methods

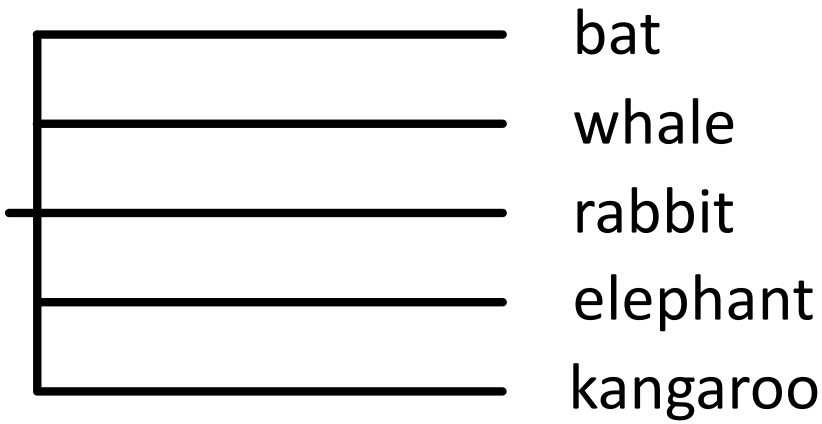
Neighbour joining

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



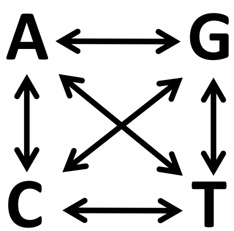
	bat	whale	rabbit	elephant	kangaroo
bat	–				
whale	.15	–			
rabbit	.20	.25	–		
elephant	.35	.40	.35	–	
kangaroo	.55	.60	.55	.55	–

Clustering algorithm

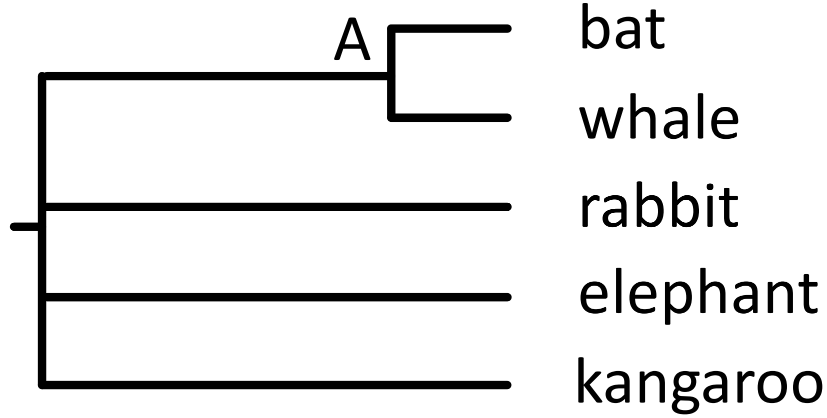


Neighbour joining

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



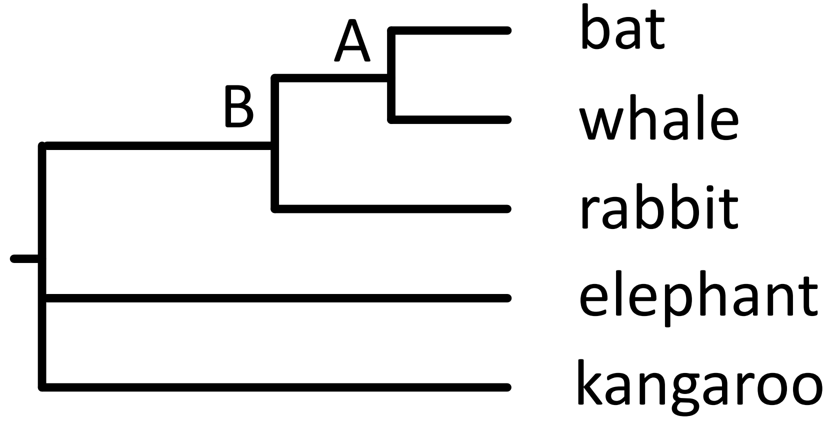
	bat	whale	rabbit	elephant	kangaroo
bat	–				
whale	.15	–			
rabbit	.20	.25	–		
elephant	.35	.40	.35	–	
kangaroo	.55	.60	.55	.55	–

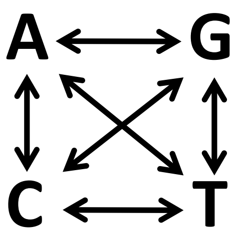


Clustering algorithm

Neighbour joining

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



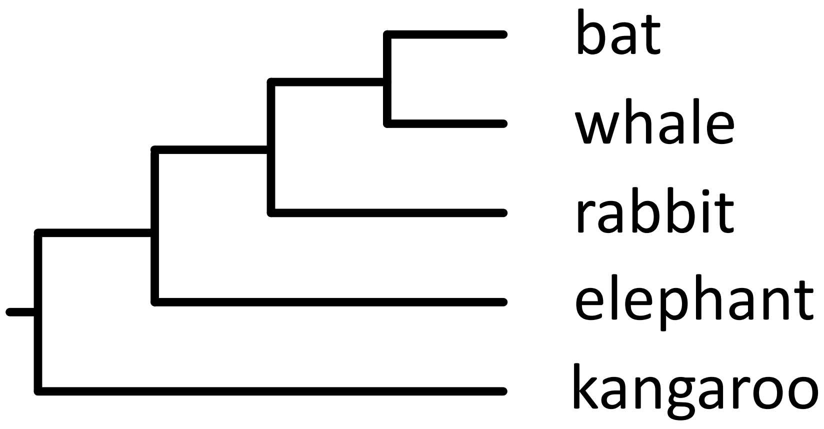


	A	rabbit	elephant	kangaroo
A	–			
rabbit	$\boxed{.15}$	–		
elephant	.30	.35	–	
kangaroo	.50	.55	.60	–

Clustering algorithm

Neighbour joining

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



Distance-based methods

- **Clustering algorithms**
  - Unweighted pair group method with arithmetic mean (UPGMA)
  - Neighbour joining
- **Tree searching using optimality criteria**
  - Minimum evolution
  - Least-squares inference

Strengths and weaknesses

- **Strengths**
  - Very quick method
  - Deals with multiple substitutions and long-branch attraction

- **Weaknesses**
  - Does not use all information in alignment
  - Loss of information in pairwise comparisons
  - Unable to implement sophisticated evolutionary models

Maximum Likelihood

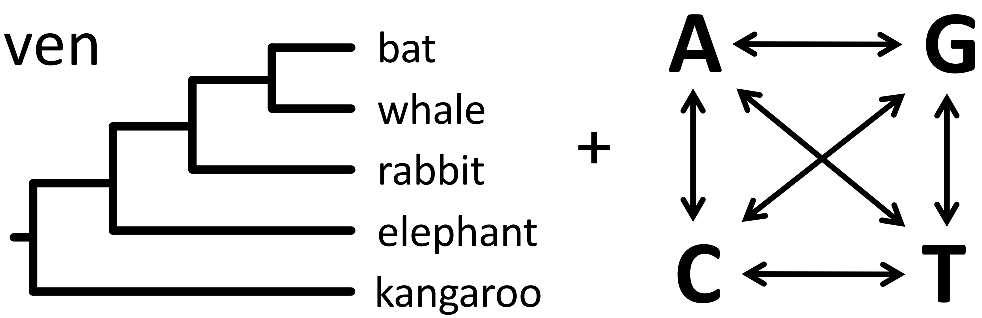
Maximum likelihood

Likelihood of hypothesis *H* =

$$\boxed{P(D|H)}$$

Probability of the data, given the hypothesis

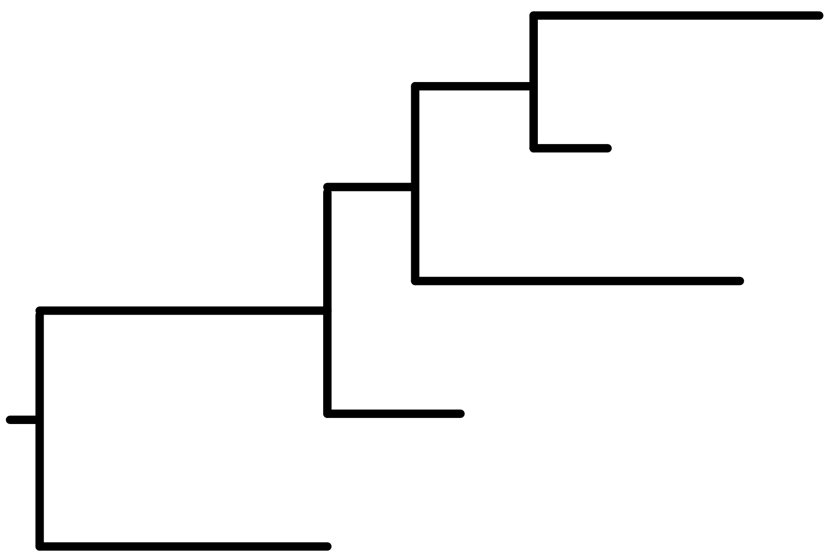
Given



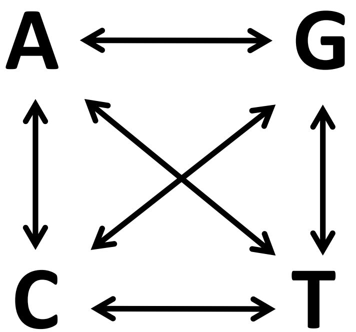
Probability of?

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

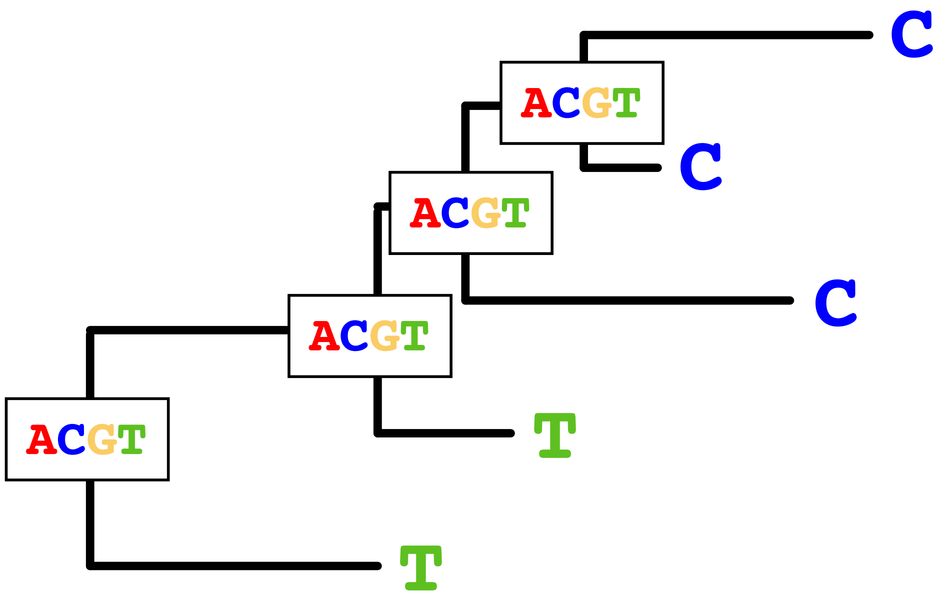
Maximum likelihood



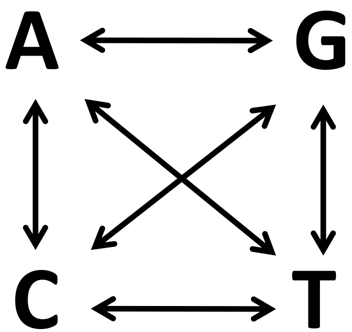
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



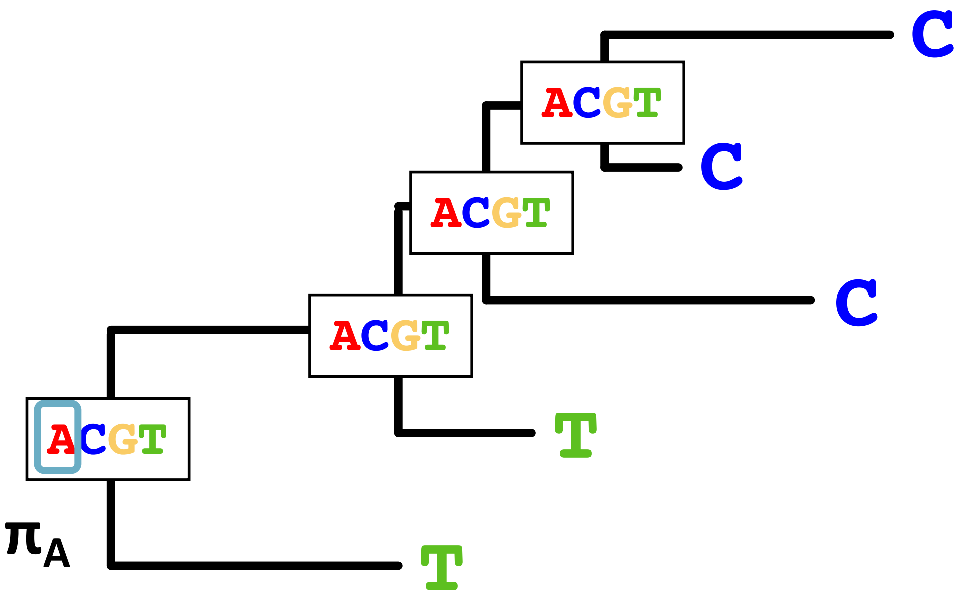
Maximum likelihood



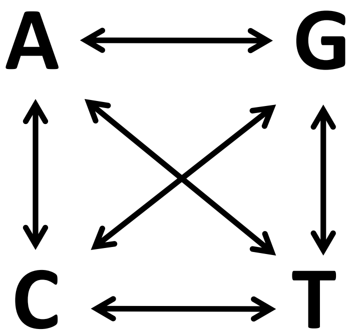
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



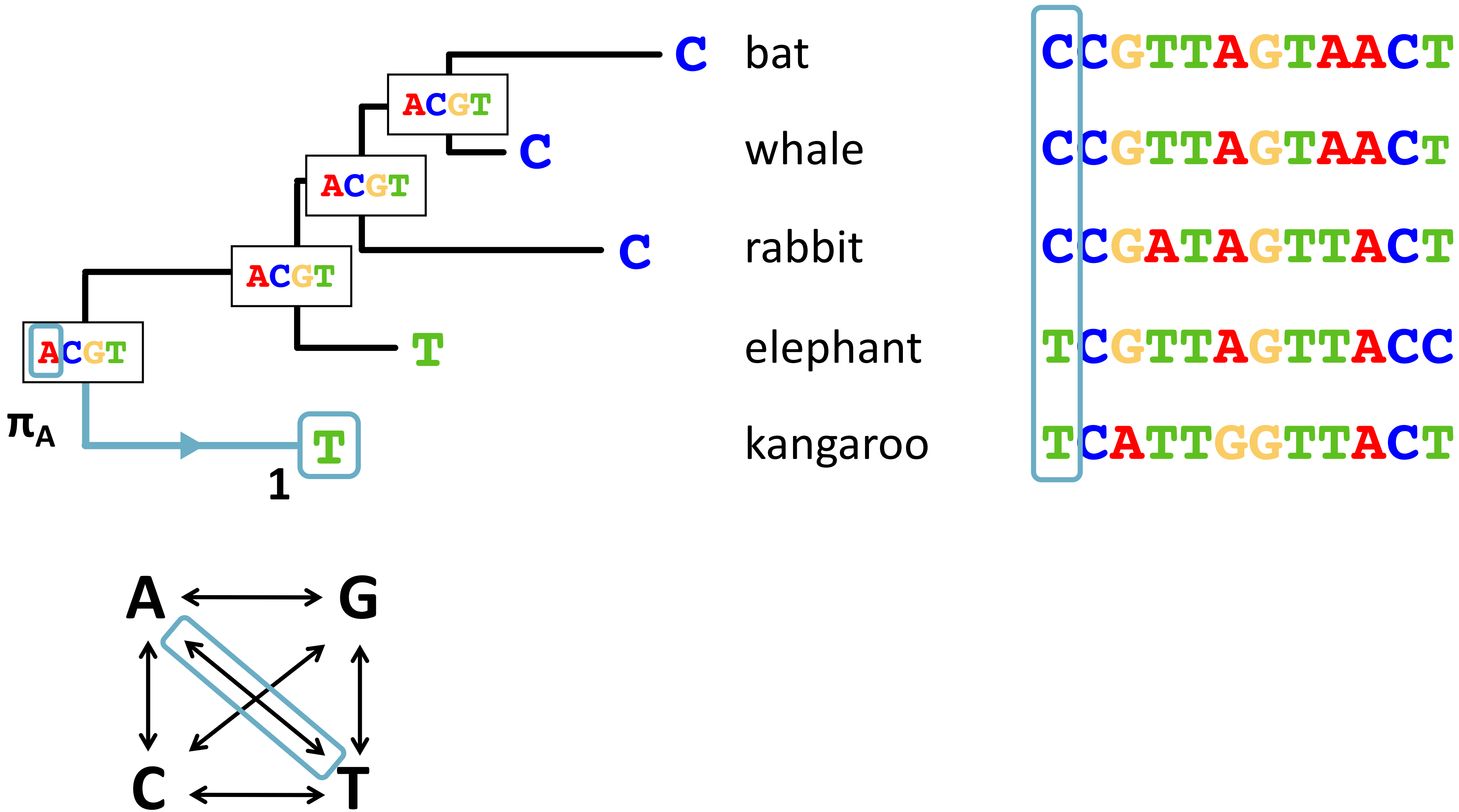
Maximum likelihood



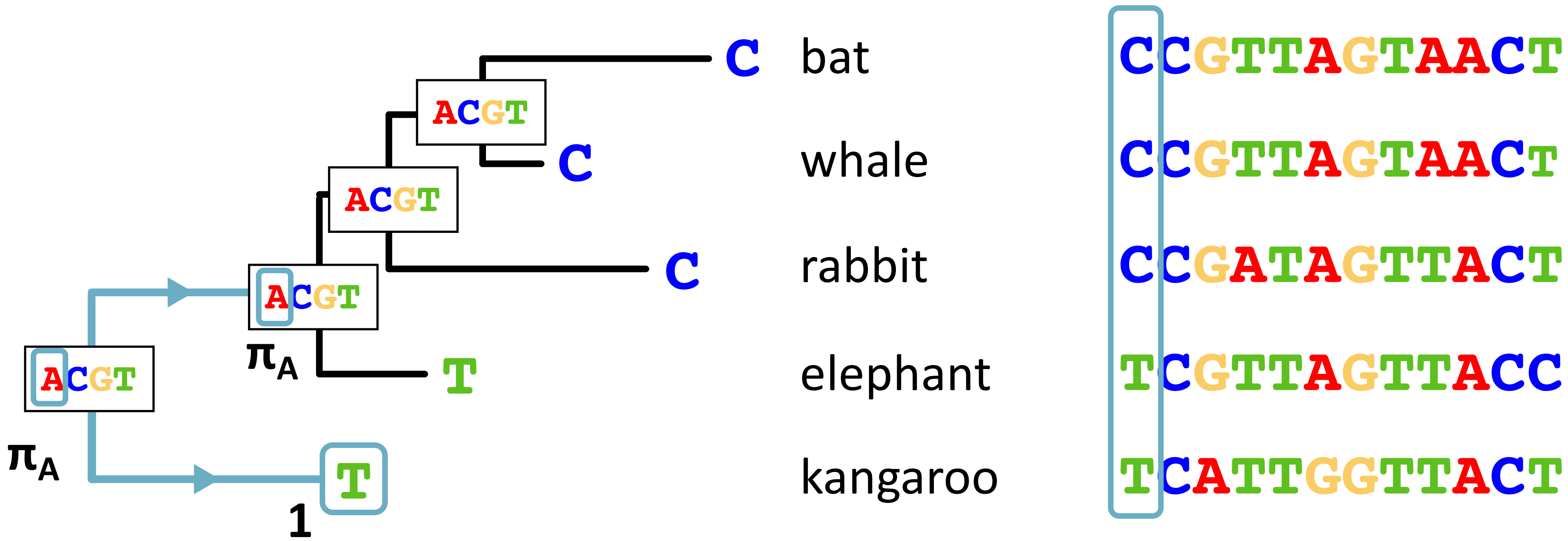
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

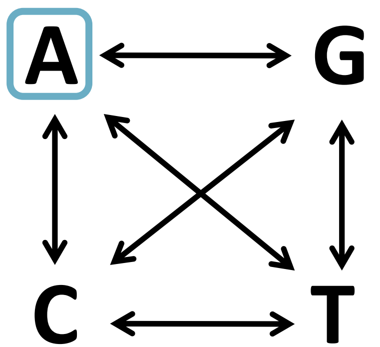


Maximum likelihood

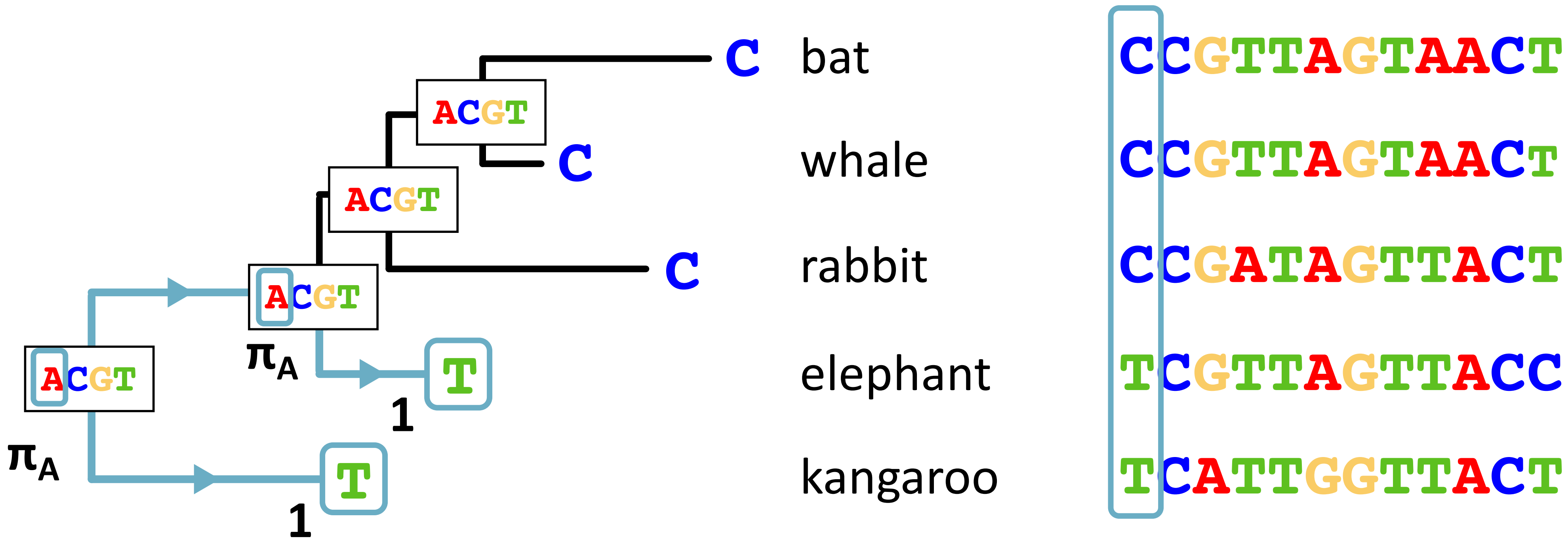


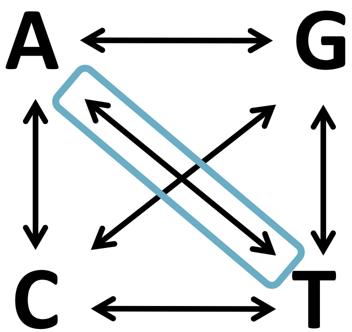
Maximum likelihood



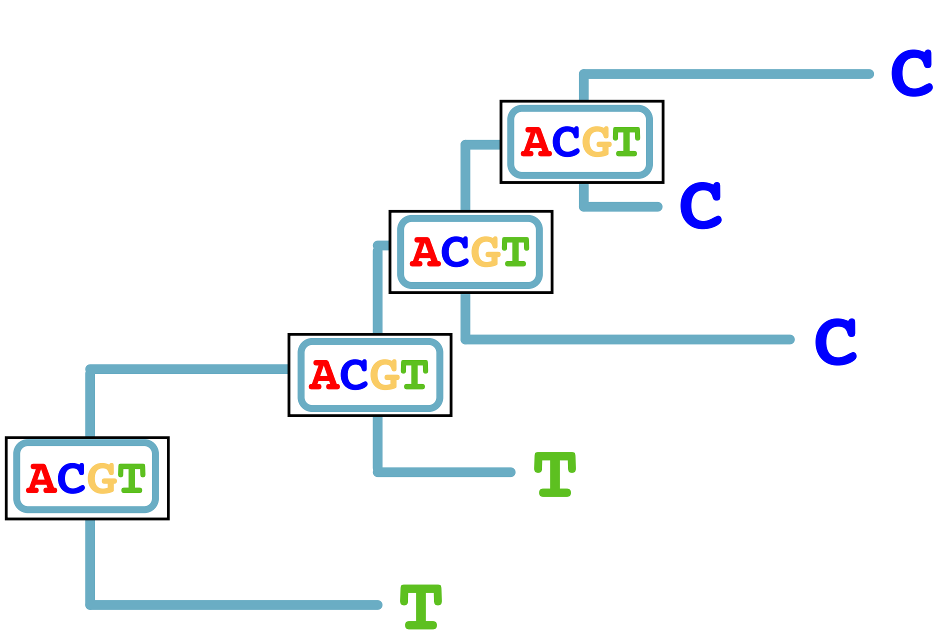


Maximum likelihood

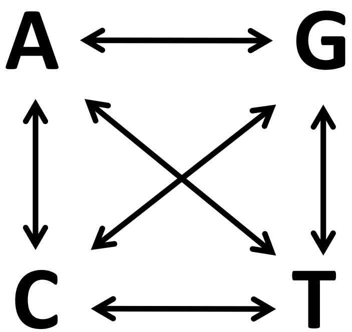




Maximum likelihood

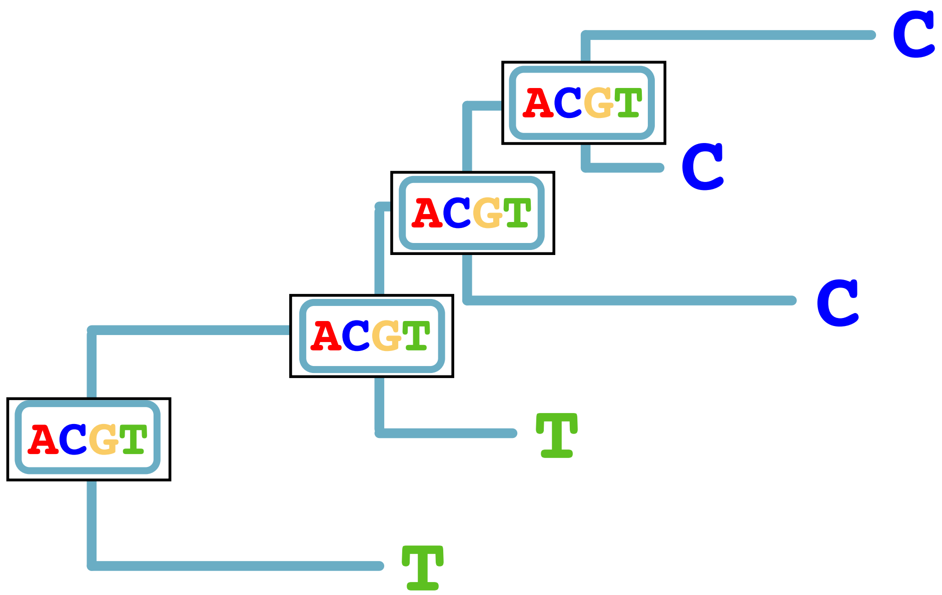


bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

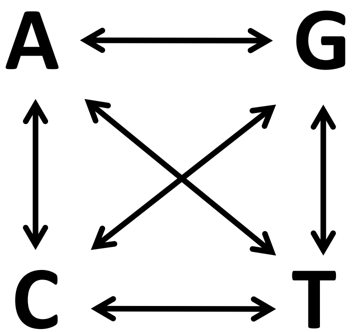


Likelihood is summed over all possibilities

Maximum likelihood



bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

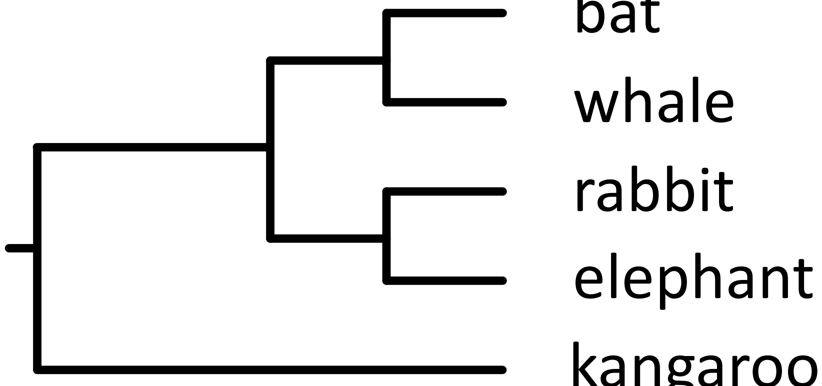


Likelihood is multiplied across all sites

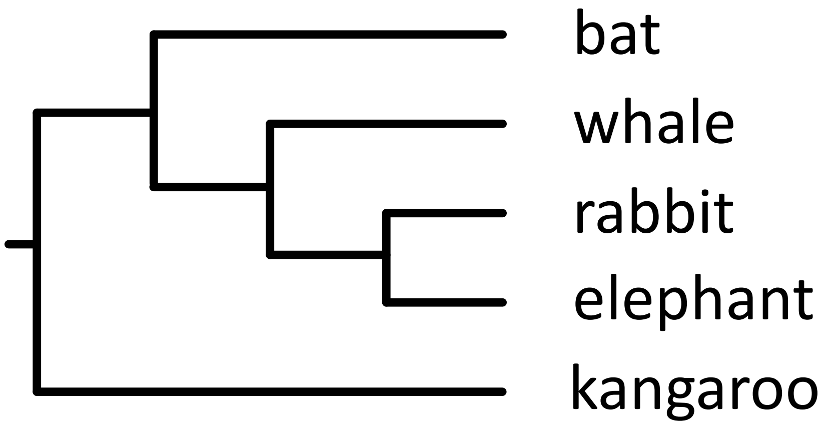
Very low probability of observing any particular alignment

Maximum likelihood

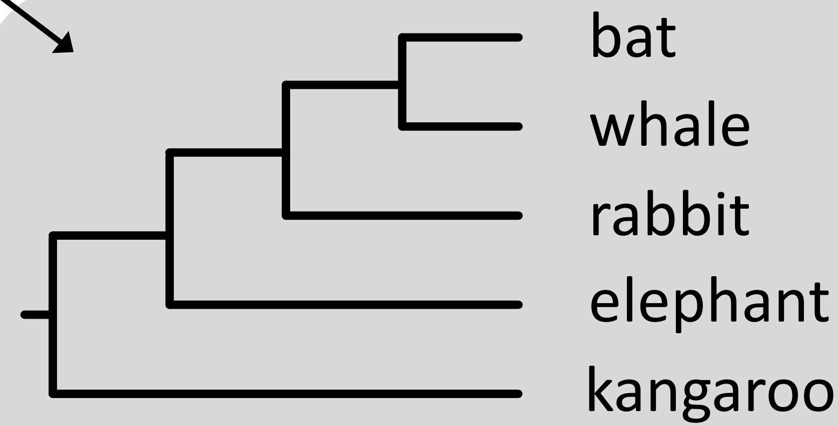
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT



lnL = -1203.83



lnL = -1241.47



lnL = -1008.58

Likelihood optimisation

- Search through the space of possible trees (including branch lengths) and model parameter values
- Calculate the likelihood for these
- Find best tree and model parameter values
- Multivariate optimisation

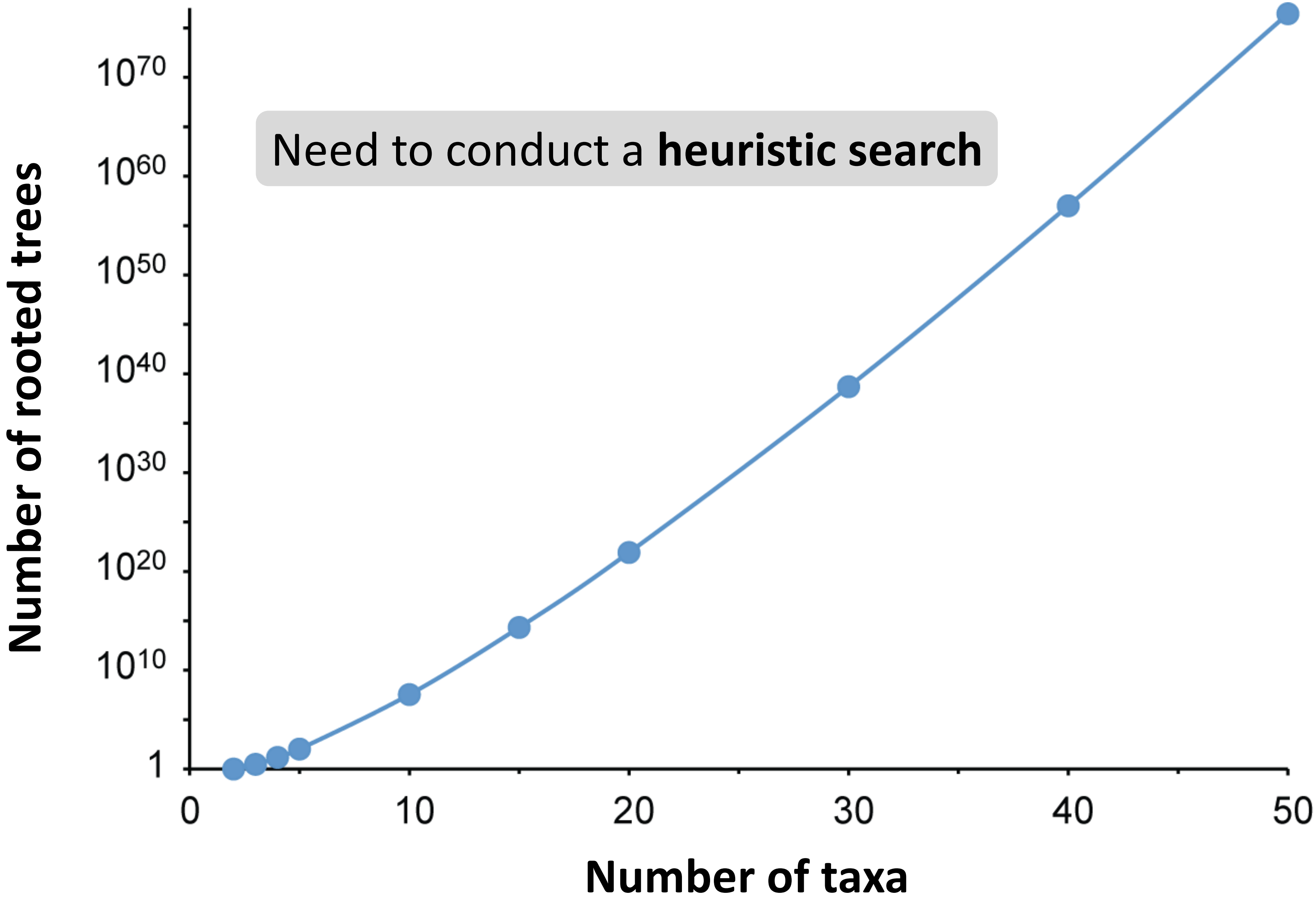
Finding the best tree

- For n taxa, the number of possible unrooted trees (B_n) is:

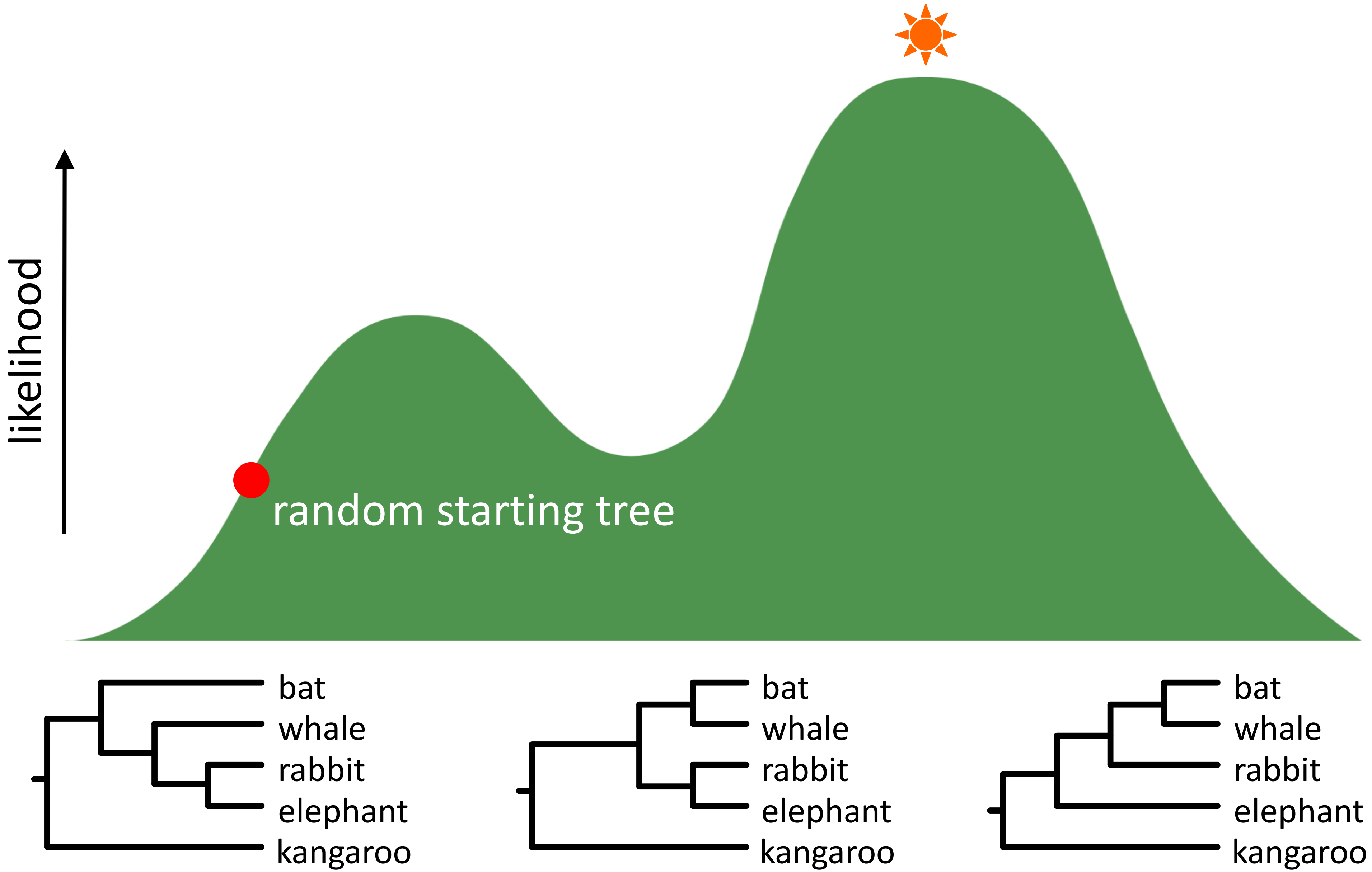
$$B_n = 1 \times 3 \times 5 \times \ldots \times (2n-5) = \prod_{i=3}^{n} (2i-5)$$

- For example:
    - 4 taxa → 3 trees
    - 5 taxa → 15 trees
    - 10 taxa → 2,027,025 trees

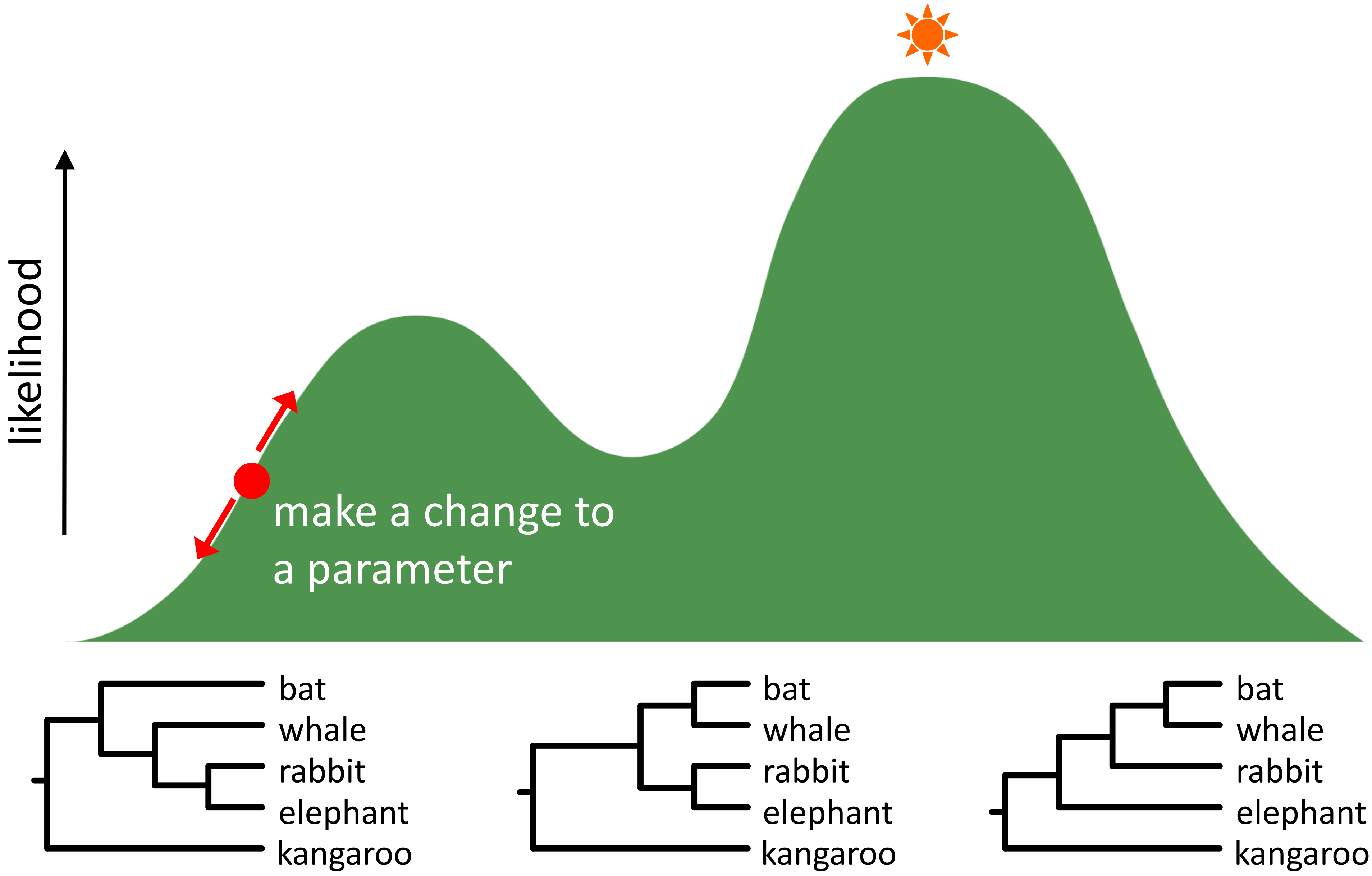
Finding the best tree



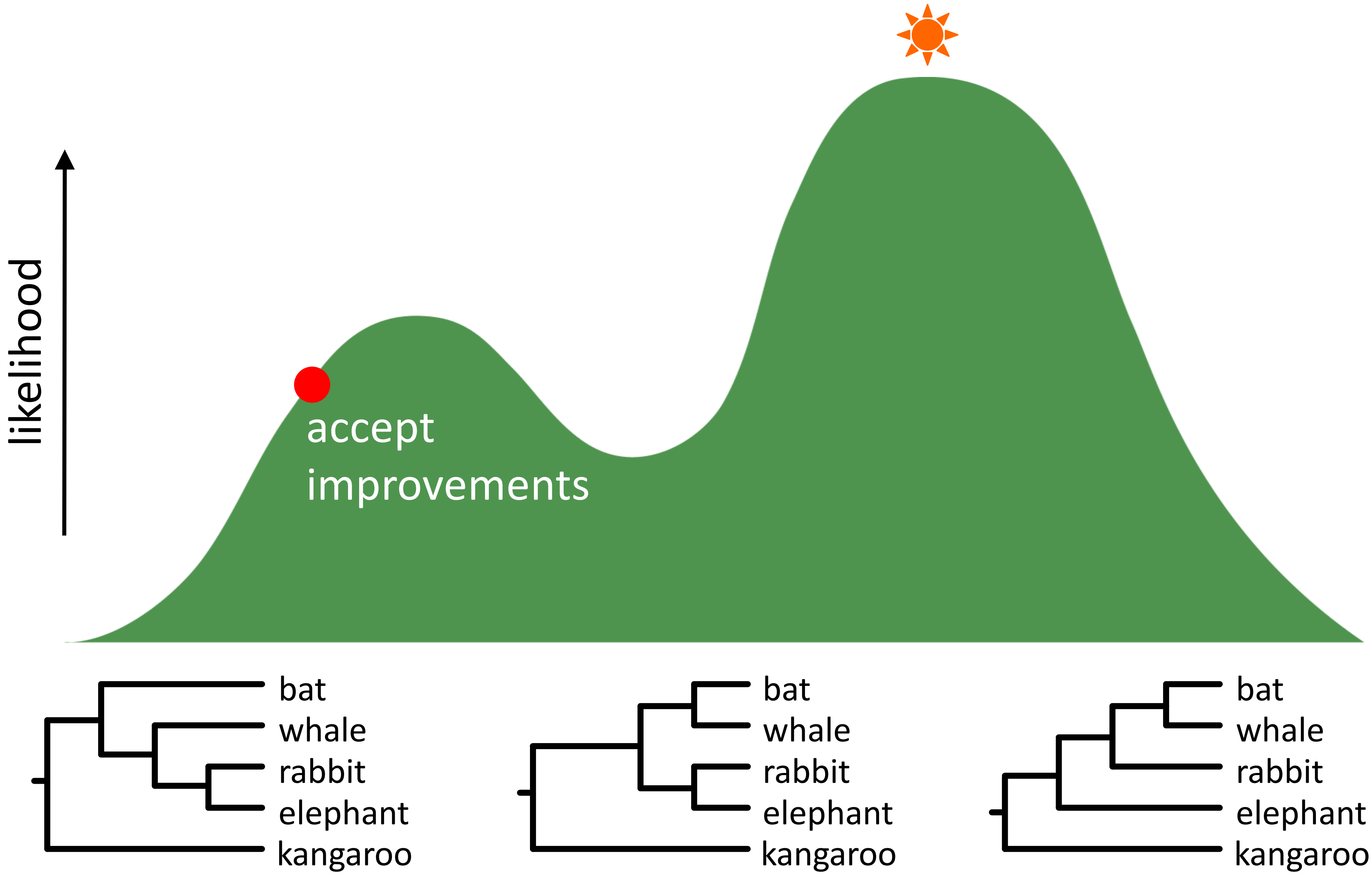
Heuristic search



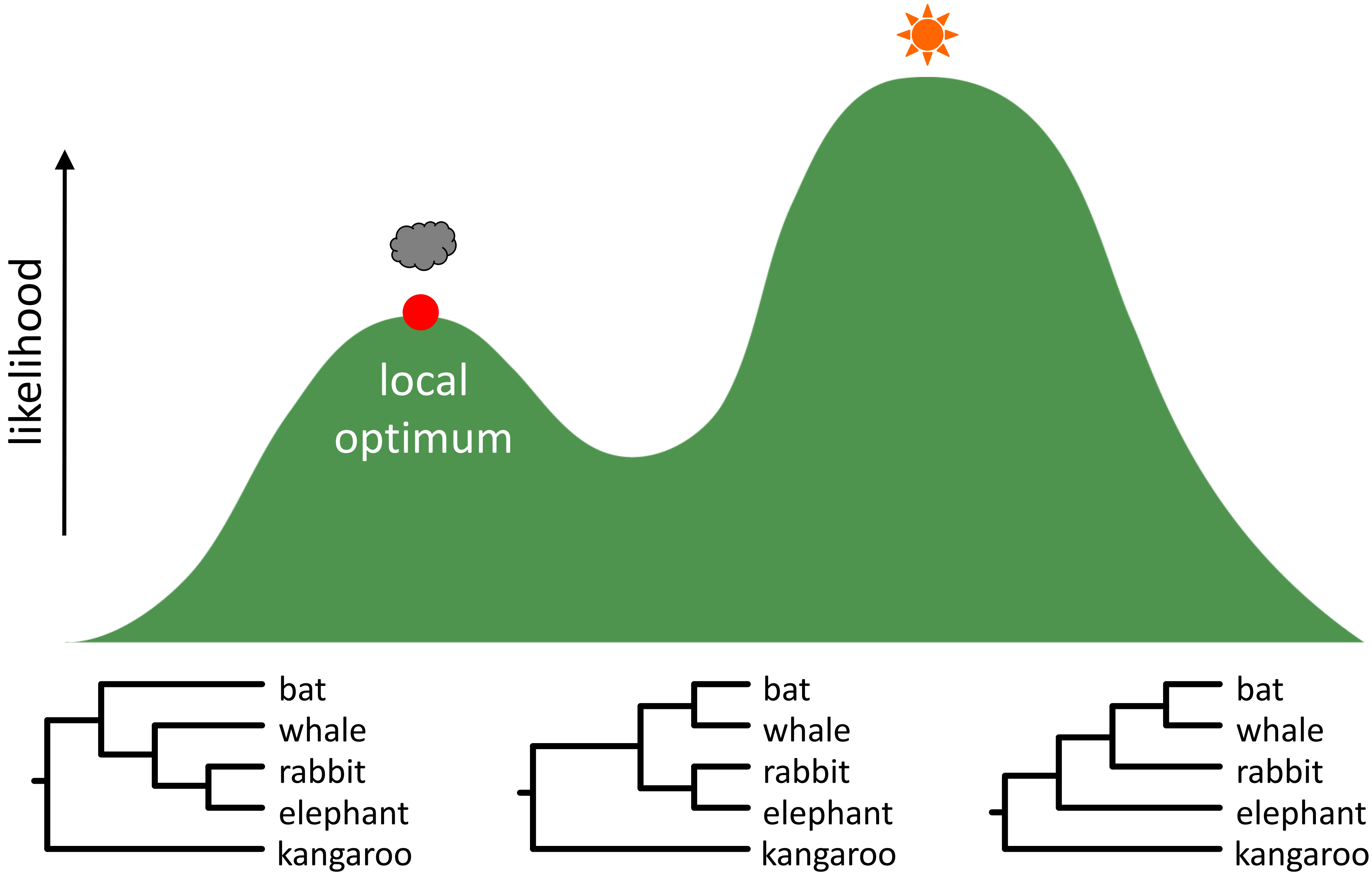
Heuristic search



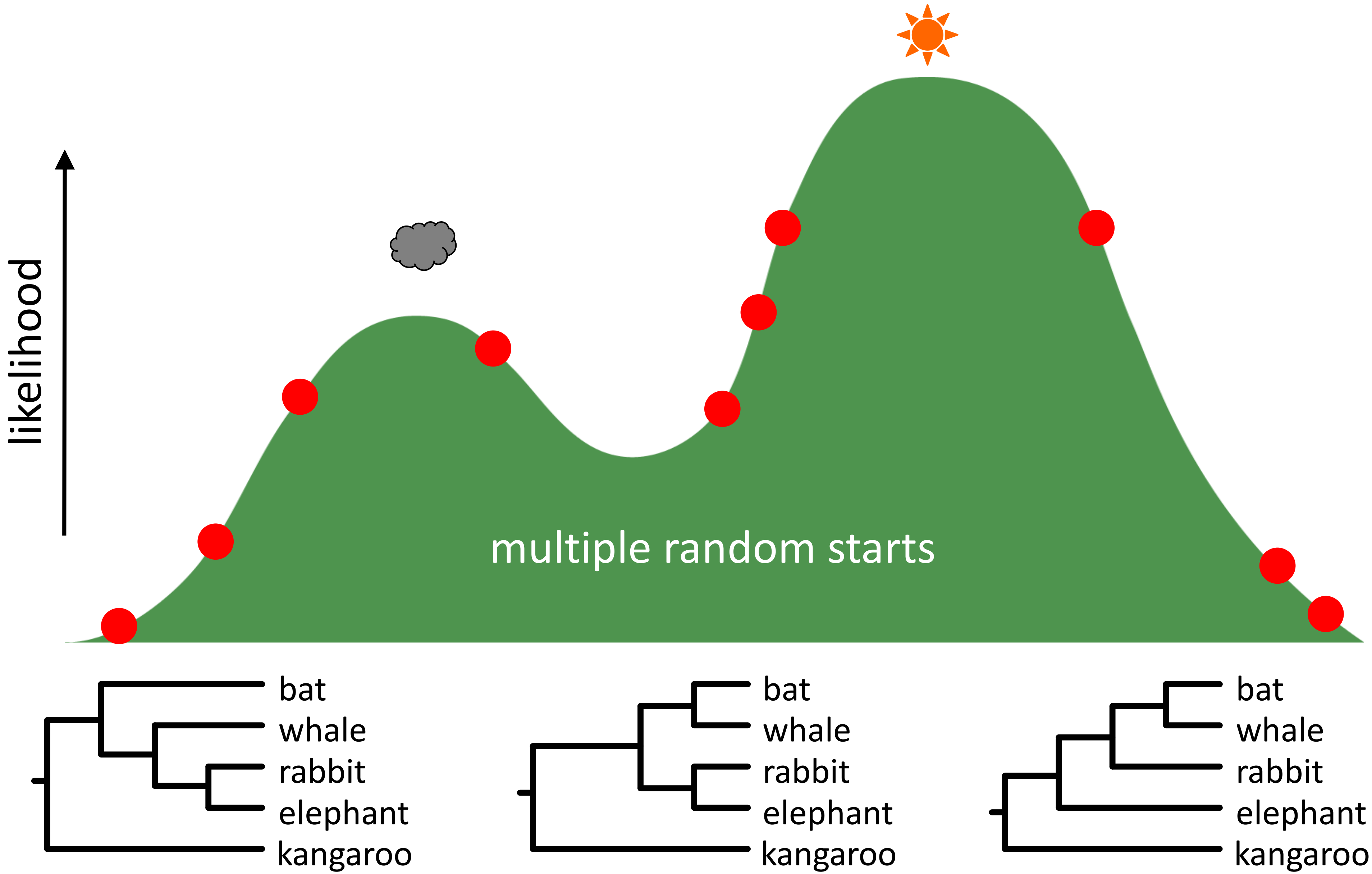
Heuristic search



Heuristic search

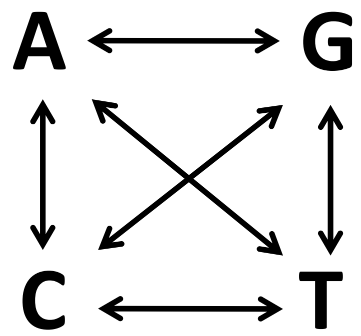


Heuristic search

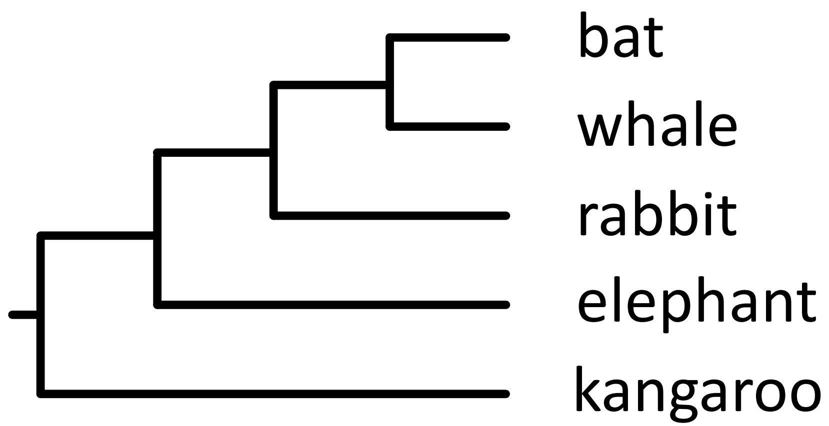


Maximum-likelihood estimates

A single set of maximum-likelihood estimates of model parameters



A single maximum-likelihood tree



Strengths and weaknesses

- **Strengths**
  - Rigorous statistical method
  - Deals with multiple substitutions and long-branch attraction
  - Robust to violations of assumptions

- **Weaknesses**
  - Generally not feasible to implement very parameter-rich models
  - Searching tree space can be difficult

Software

RAxML

The Exelixis Lab
Enabling Research in Evolutionary Biology

PhyML

ATGC
Montpellier Bioinformatics Platform

MEGA

MEGA

PAML

IQ-TREE

Bootstrapping

Nonparametric bootstrap

- Uncertainty in the estimate of the tree is inferred indirectly using **bootstrapping analysis**
- “pull oneself up by one's bootstraps”
- Bootstrapping analysis can be used in various phylogenetic methods:
  - Maximum parsimony
  - Distance-based methods
  - Maximum likelihood

Bootstrapping

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

Randomly sample sites (with replacement)

bat	T
whale	T
rabbit	A
elephant	T
kangaroo	T

Bootstrapping

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

bat	TG
whale	TG
rabbit	AG
elephant	TG
kangaroo	TG

Bootstrapping

bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

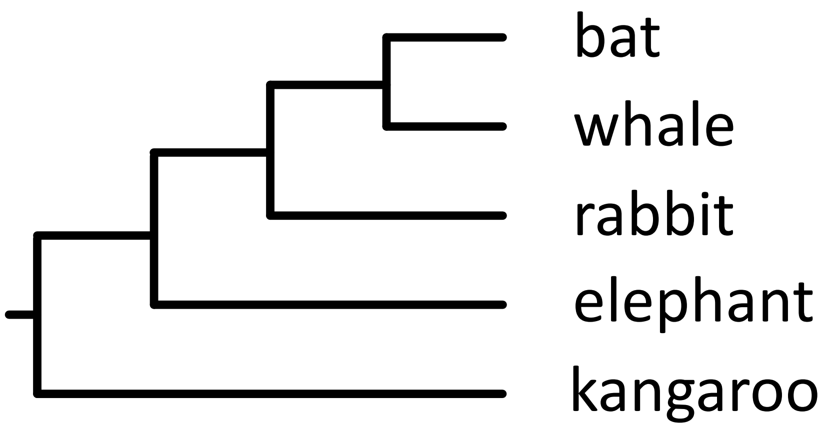
bat	TGCCCTTAGCAC
whale	TGCCCTTAGCAC
rabbit	AGCCCATAGCAC
elephant	TGCTCTCAGCAT
kangaroo	TGCTCTTAACGT

Bootstrapping

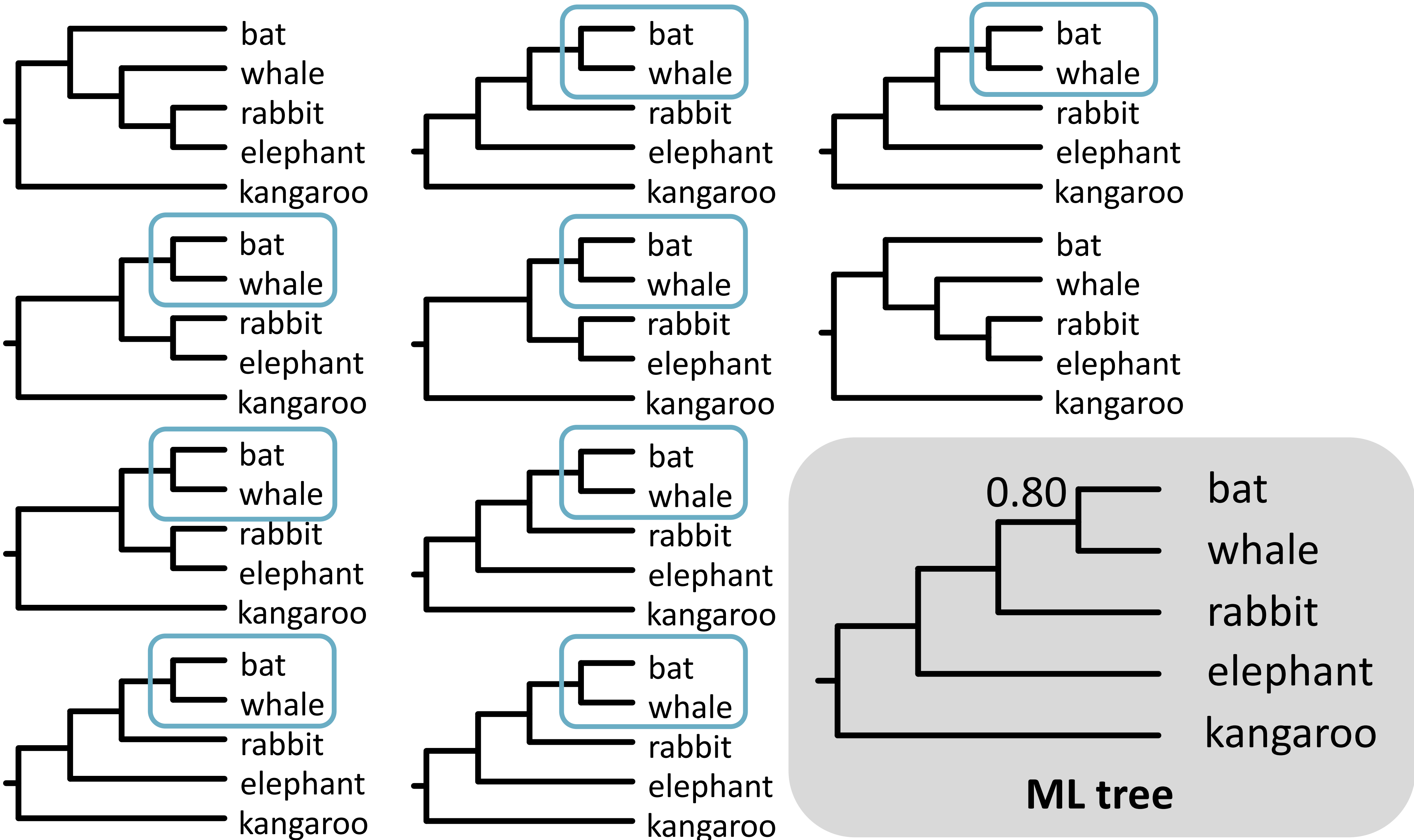
bat	CCGTTAGTAACT
whale	CCGTTAGTAACT
rabbit	CCGATAGTTACT
elephant	TCGTTAGTTACC
kangaroo	TCATTGGTTACT

Repeat 1,000 times

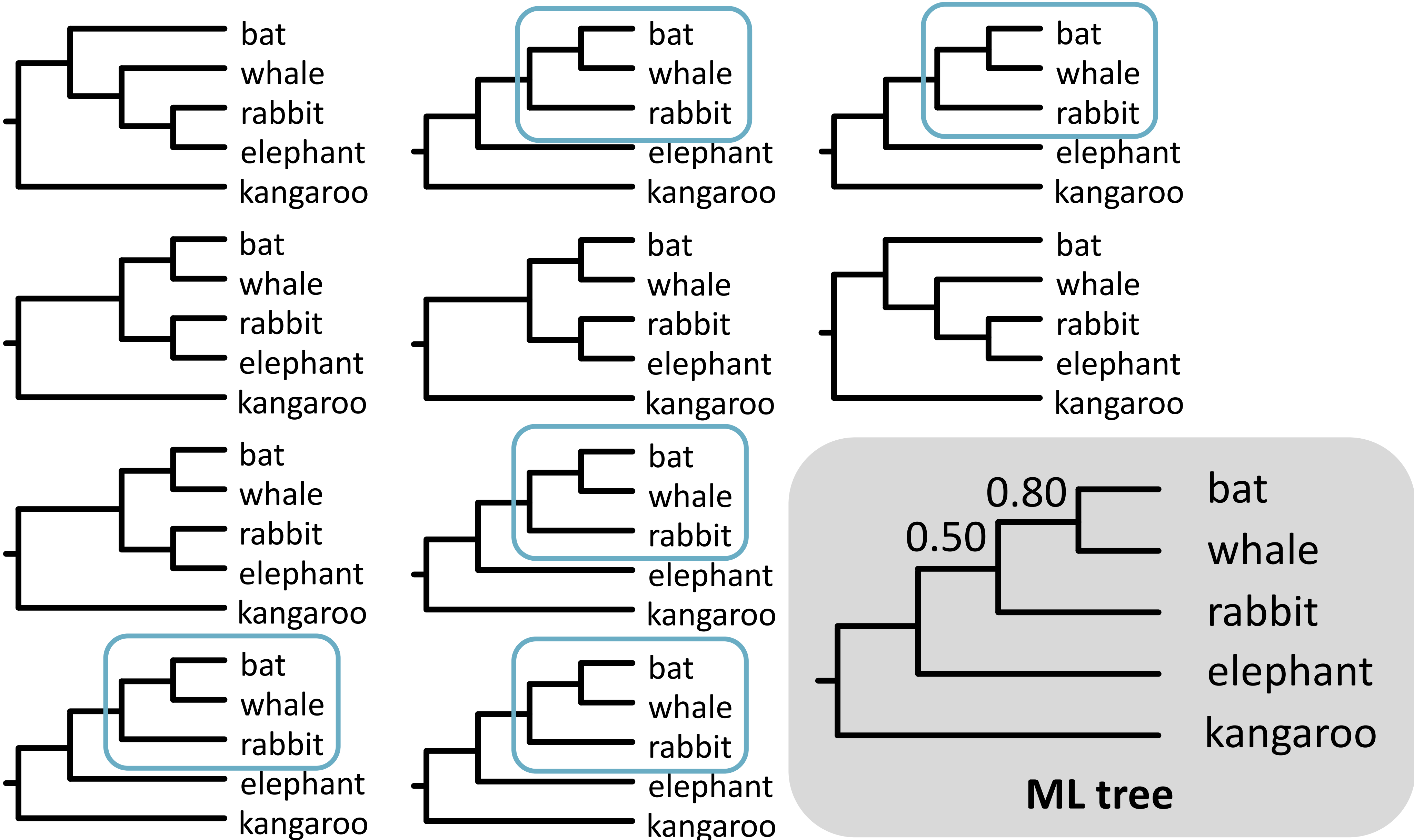
bat	TGCCCTTAGCAC
whale	TGCCCTTAGCAC
rabbit	AGCCCATAGCAC
elephant	TGCTCTCAGCAT
kangaroo	TGCTCTTAACGT



Bootstrapping



Bootstrapping



Interpreting bootstrap values

- **Felsenstein (1985)**
  bootstrapping provides a confidence interval that contains the *phylogeny that would be estimated from repeated sampling of many characters from the underlying set of all characters*

- Bootstrap values are **measures of repeatability**
  - High when the data set is large
  - Not meaningful when analysing genome-scale data

Soltis & Soltis (2003) *Stat Sci*

Methods in practice

- **Maximum parsimony**
  - Commonly used to analyse morphological data
  - Rarely used to analyse molecular data
- **Distance-based methods**
  - Popular in some fields of research
  - Used to analyse very large data sets with many taxa
- **Maximum likelihood**
  - Widely used, lost some ground to Bayesian methods but is experiencing a resurgence

Useful references

- **Molecular phylogenetics: principles and practice**
  Yang & Rannala (2012) *Nature Reviews Genetics* 13: 303–314.