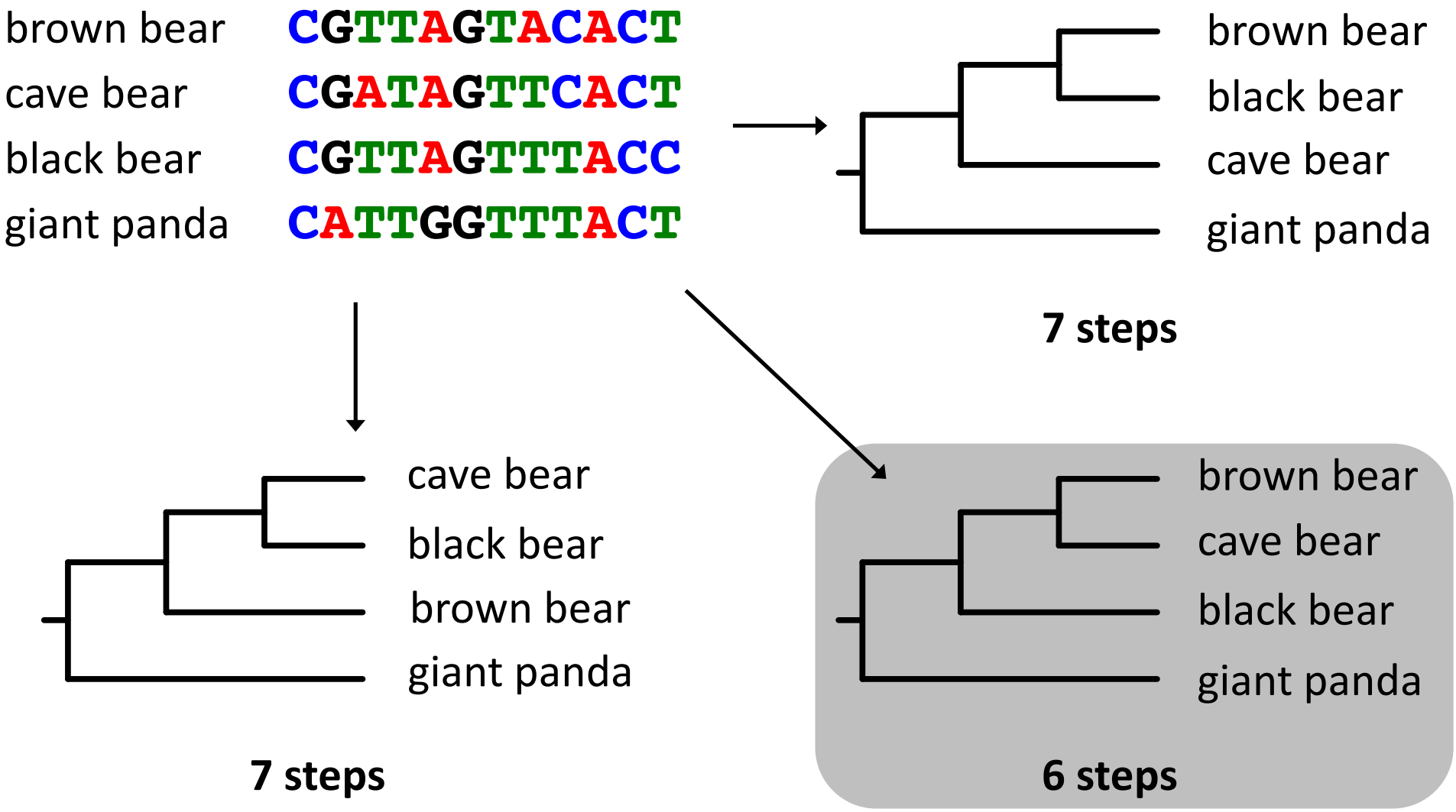
Lecture 1.4

Phylogenetic Methods

Maximum parsimony



Popular phylogenetic methods

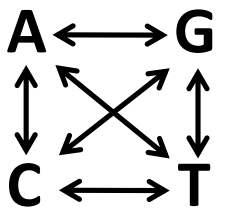
1. Maximum parsimony
2. Distance-based methods
3. Maximum likelihood
4. Bayesian inference

Model-based methods

Distance-Based Methods

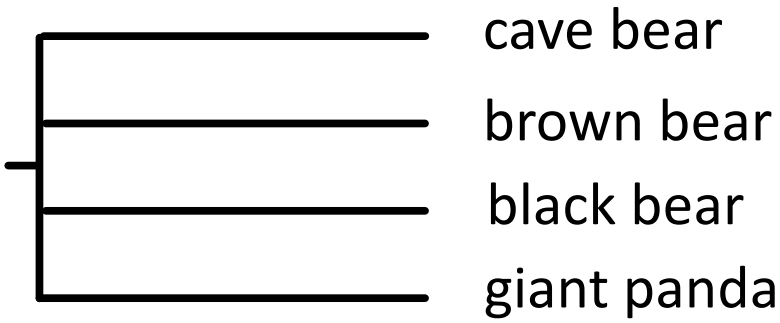
Neighbour joining

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



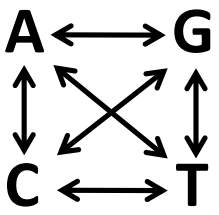
	brown bear	cave bear	black bear	giant panda
brown bear	–			
cave bear	.1	–		
black bear	.3	.3	–	
giant panda	.4	.5	.4	–

Clustering algorithm



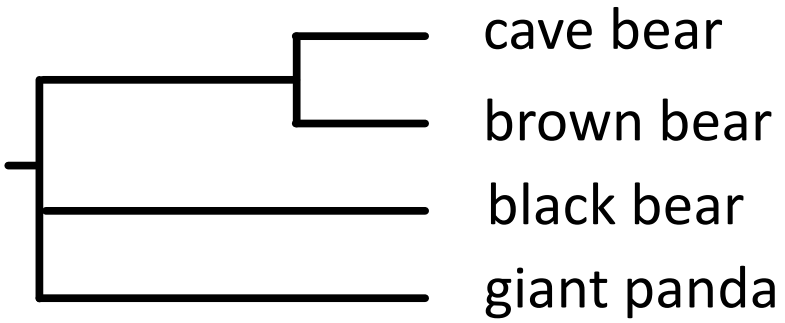
Neighbour joining

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



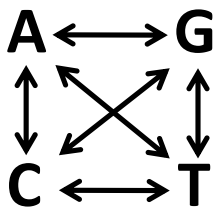
	brown bear	cave bear	black bear	giant panda
brown bear	–			
cave bear	.1	–		
black bear	.3	.3	–	
giant panda	.4	.5	.4	–

Clustering algorithm



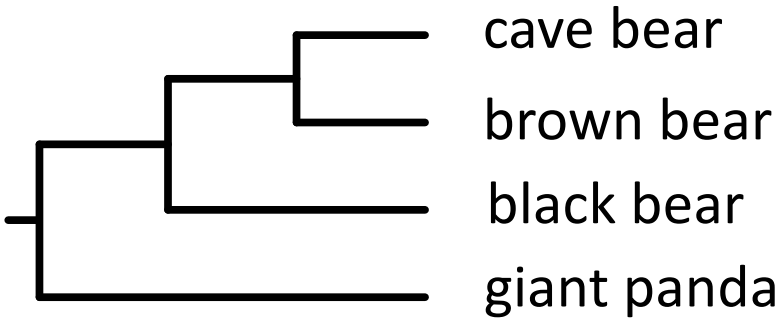
Neighbour joining

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



	brown bear	cave bear	black bear	giant panda
brown bear	–			
cave bear	.1	–		
black bear	.3	.3	–	
giant panda	.4	.5	.4	–

Clustering algorithm



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

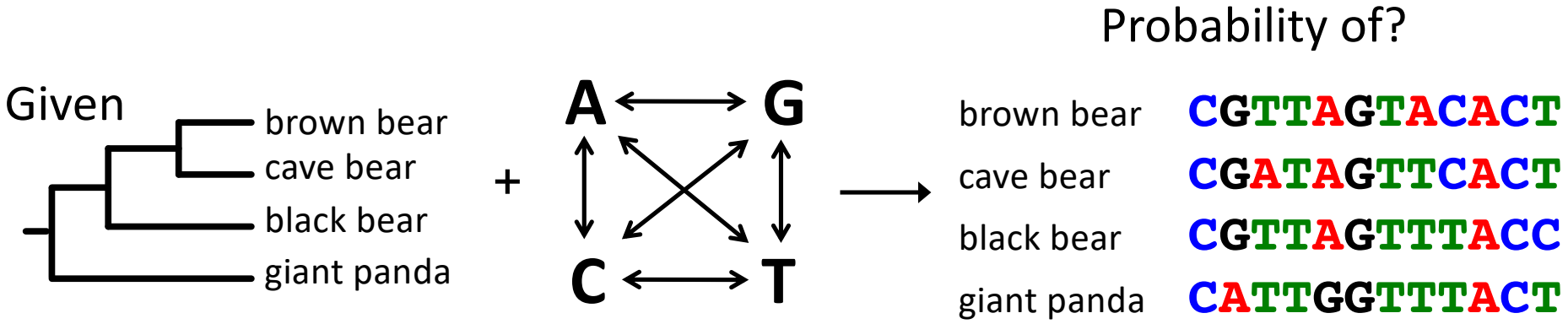
Maximum Likelihood

Maximum likelihood

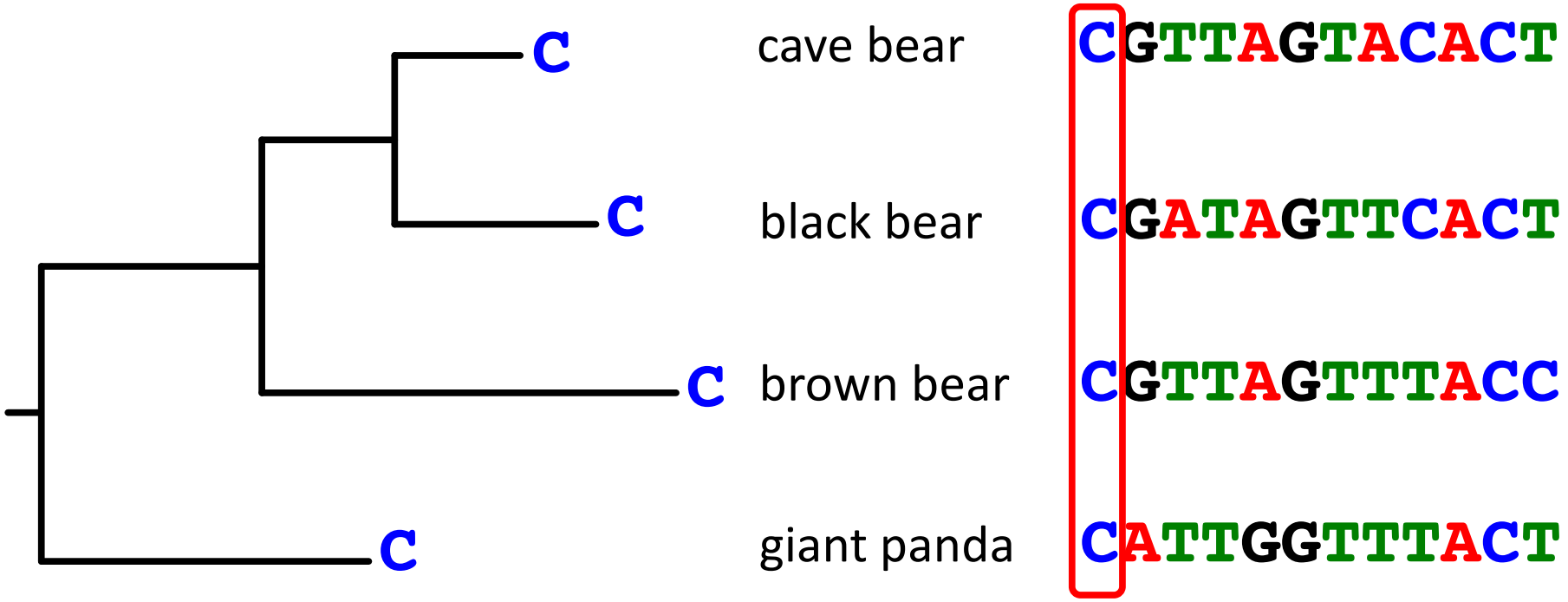
Likelihood of hypothesis *H* =

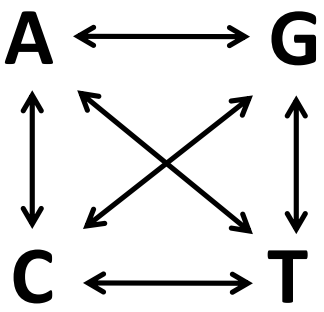
$$P(D|H)$$

Probability of the data, given the hypothesis

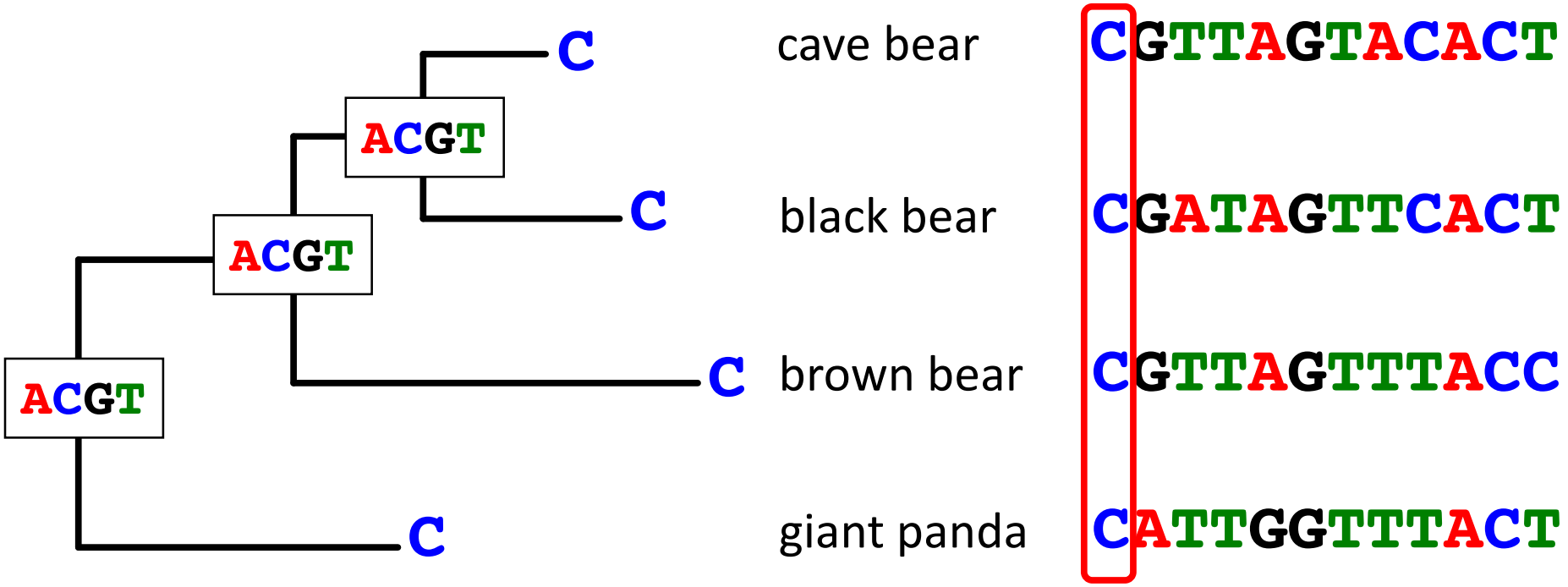


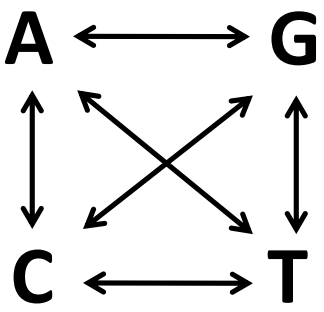
Maximum likelihood



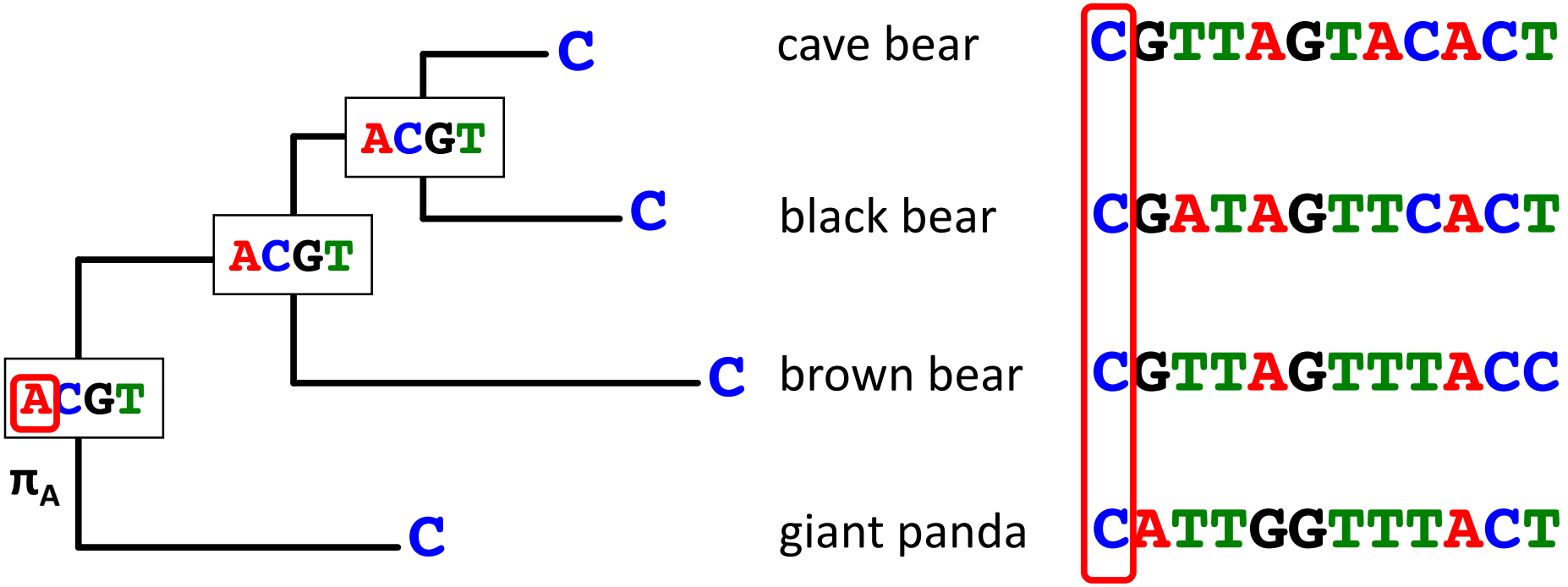


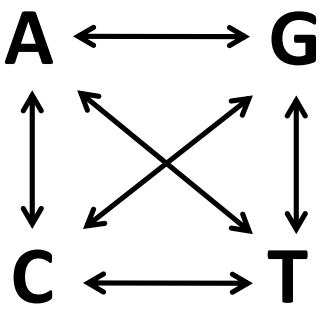
Maximum likelihood



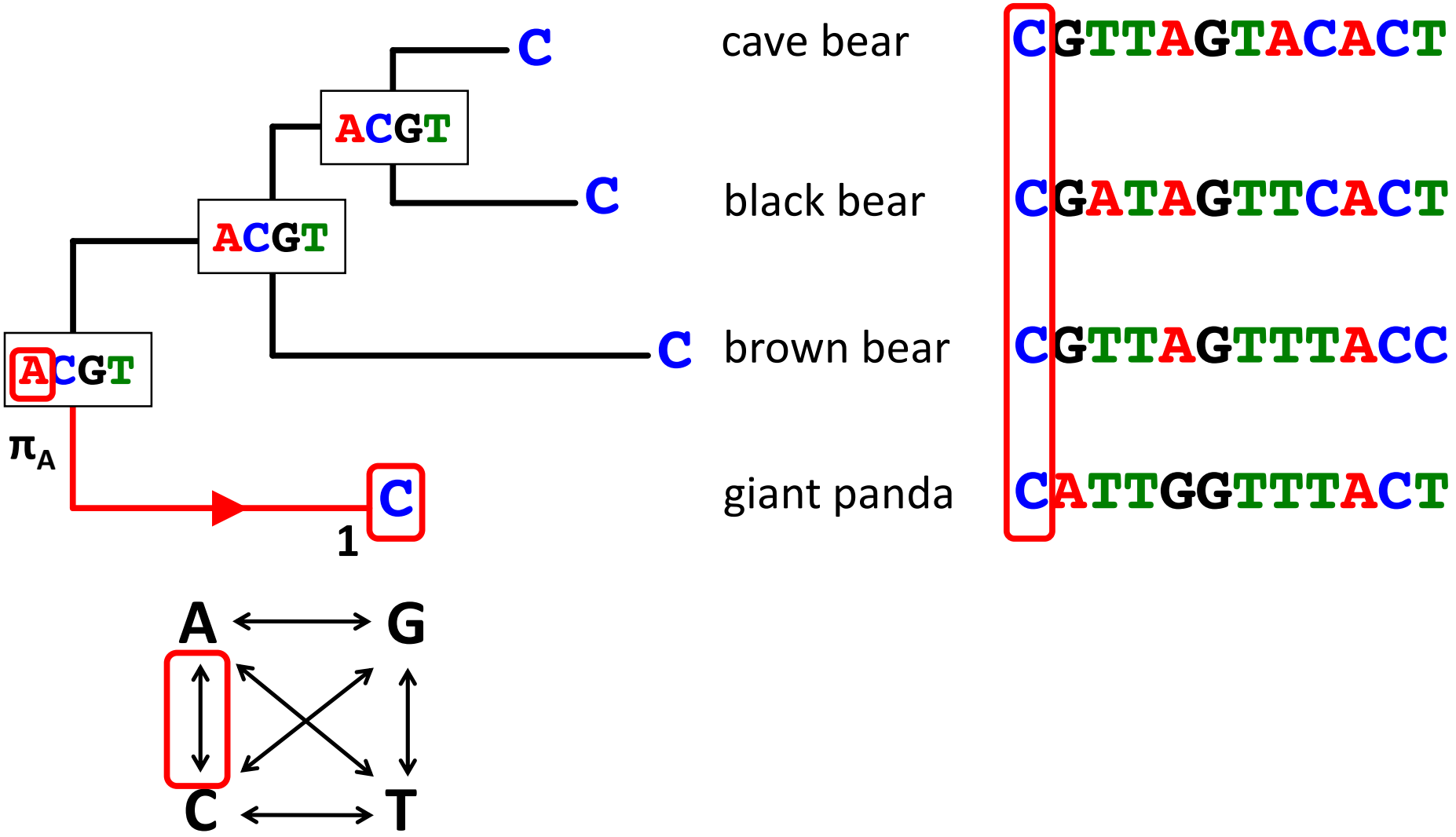


Maximum likelihood

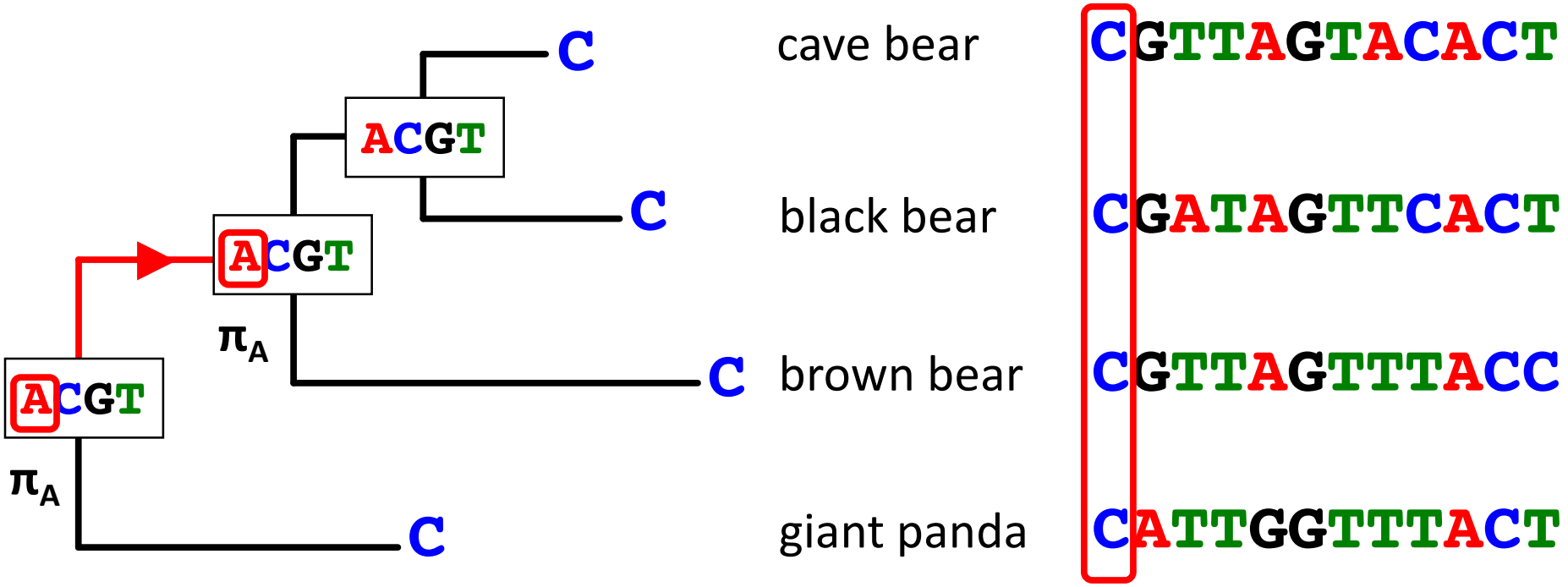


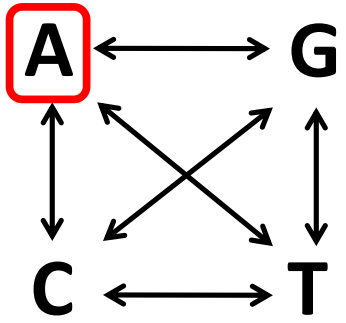


Maximum likelihood

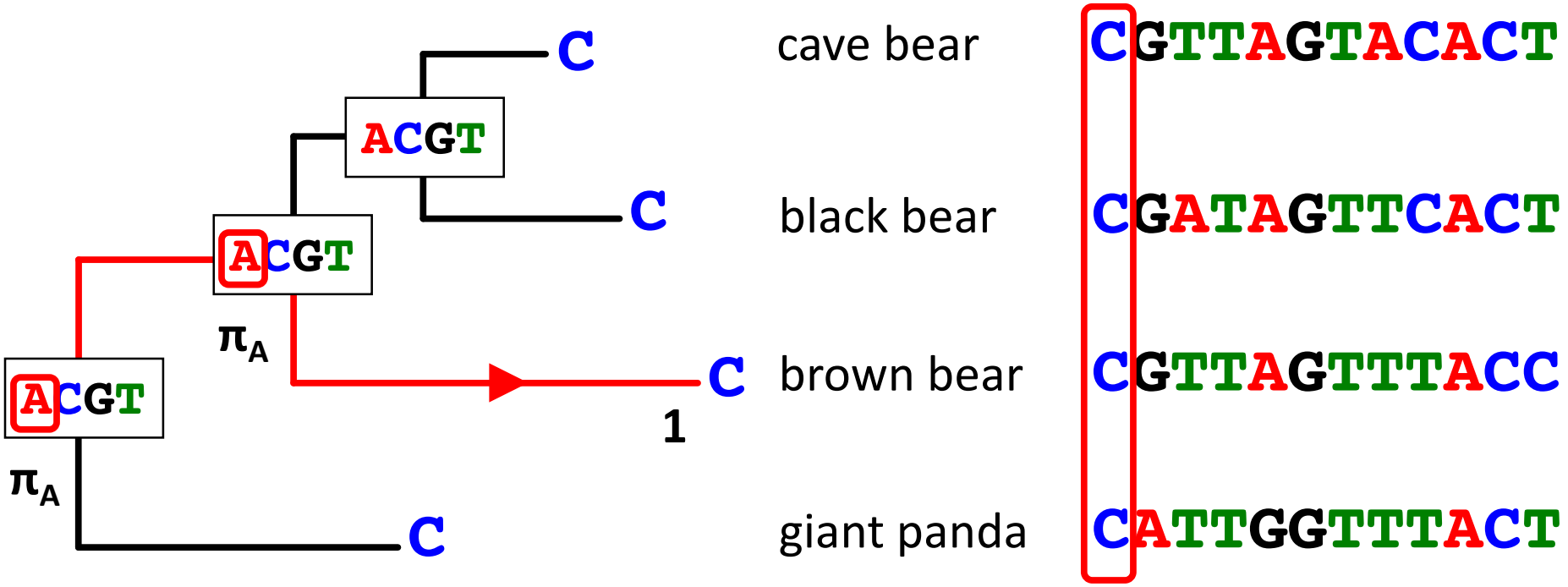


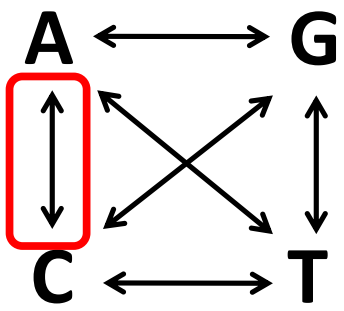
Maximum likelihood



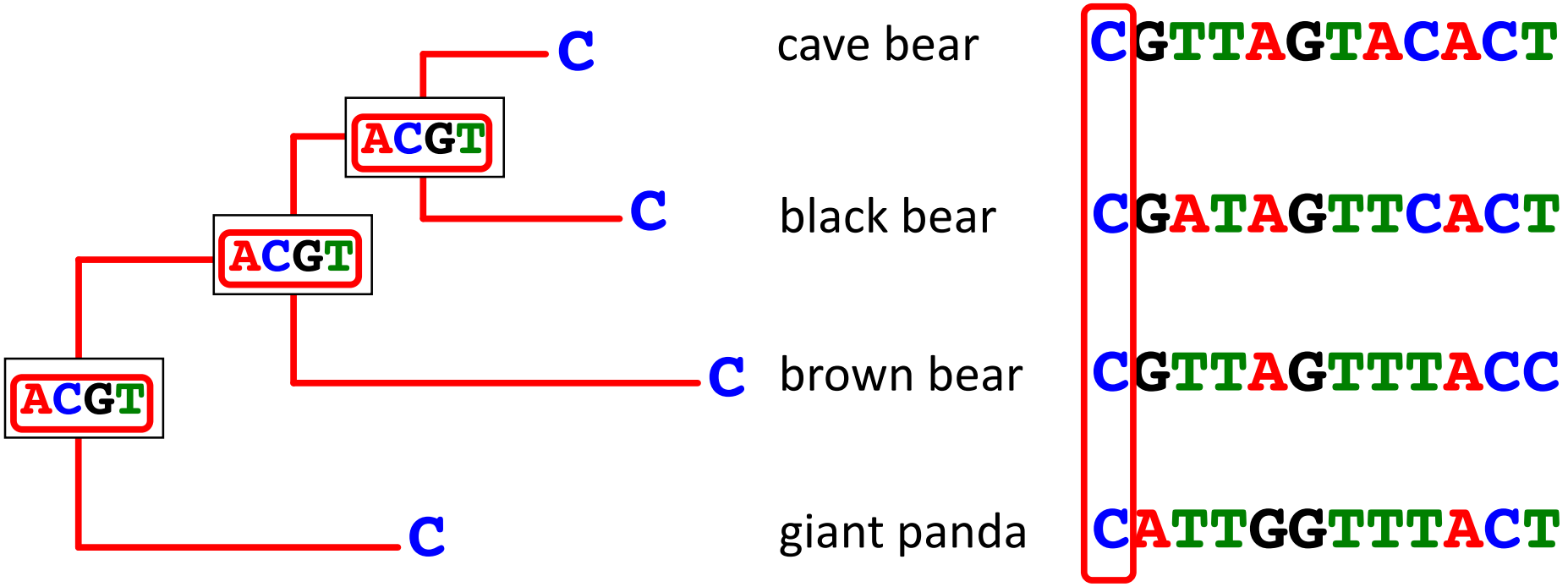


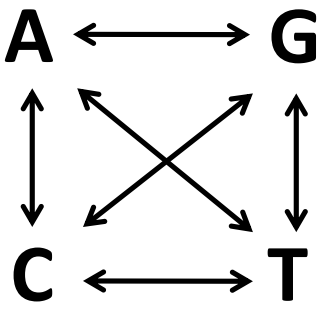
Maximum likelihood





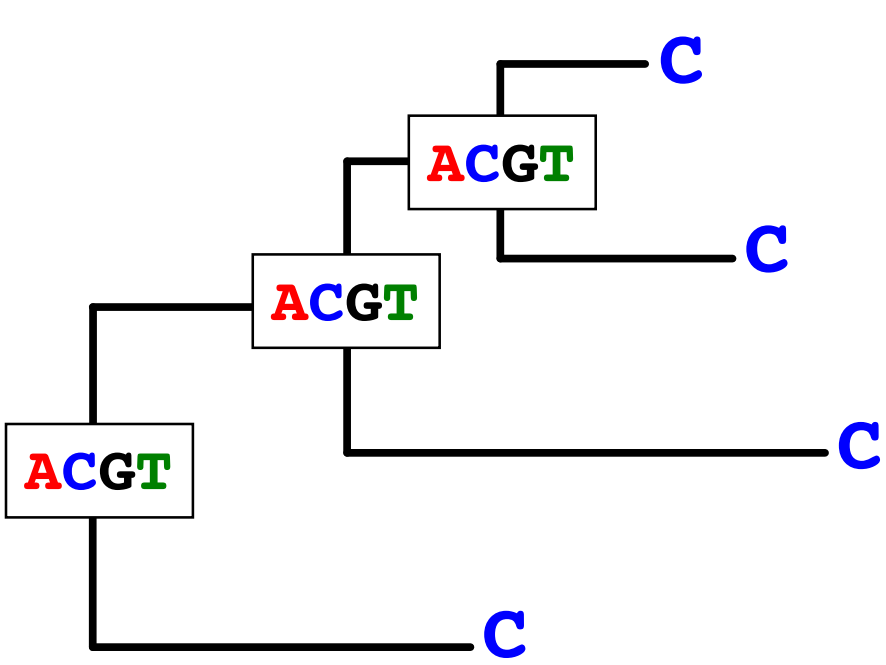
Maximum likelihood





Likelihood is summed over all possibilities

Maximum likelihood

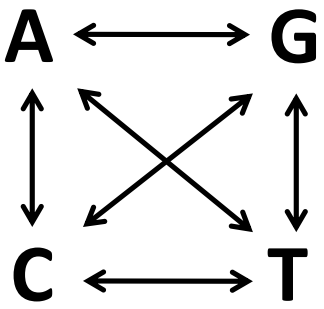


cave bear CGTTAGTACACT

black bear CGATAGTTCACT

brown bear CGTTAGTTTACC

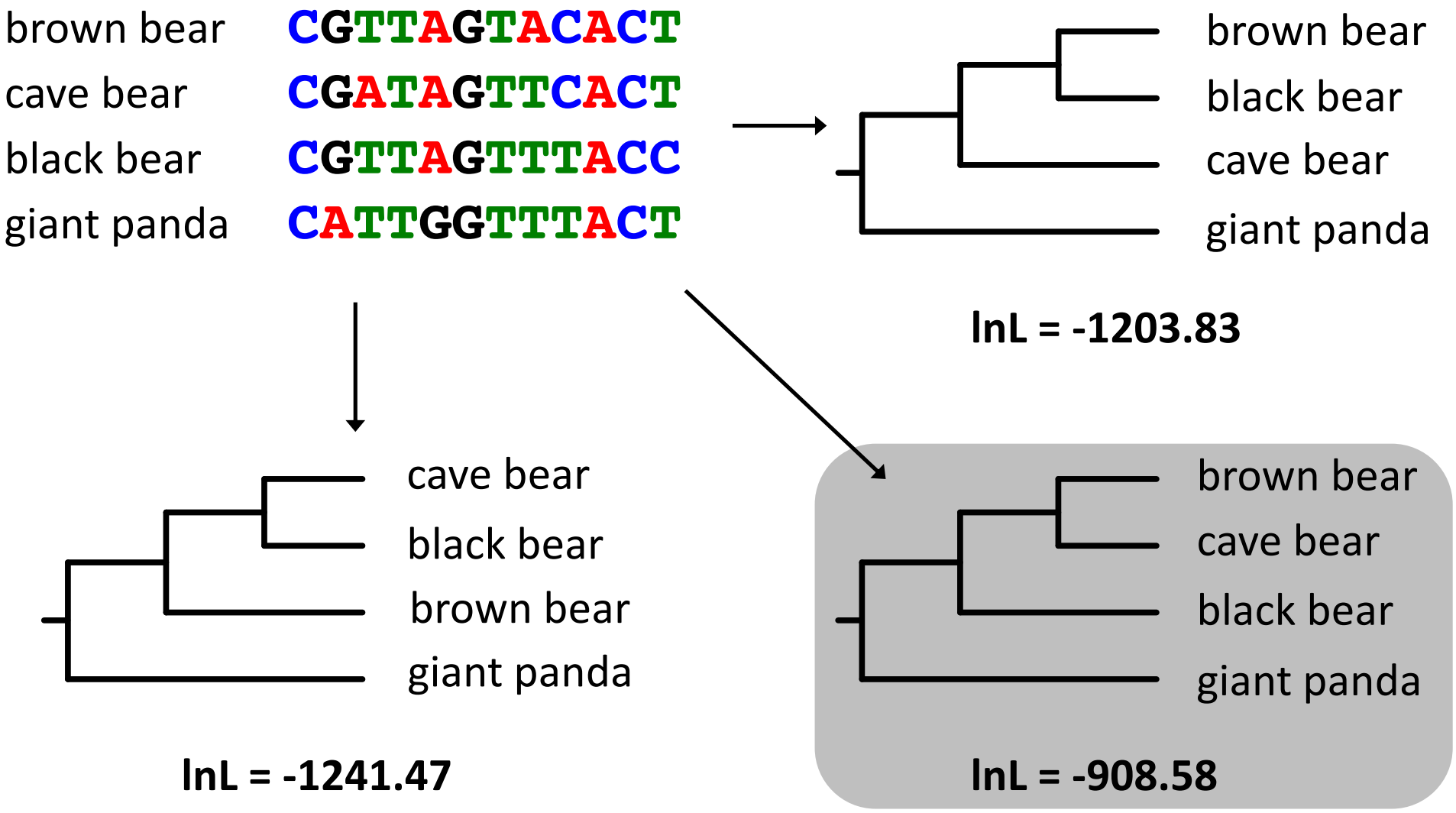
giant panda CATTGGTTTACT



Likelihood is multiplied across all sites

Very low probability of observing any particular alignment

Maximum likelihood



Likelihood optimisation

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

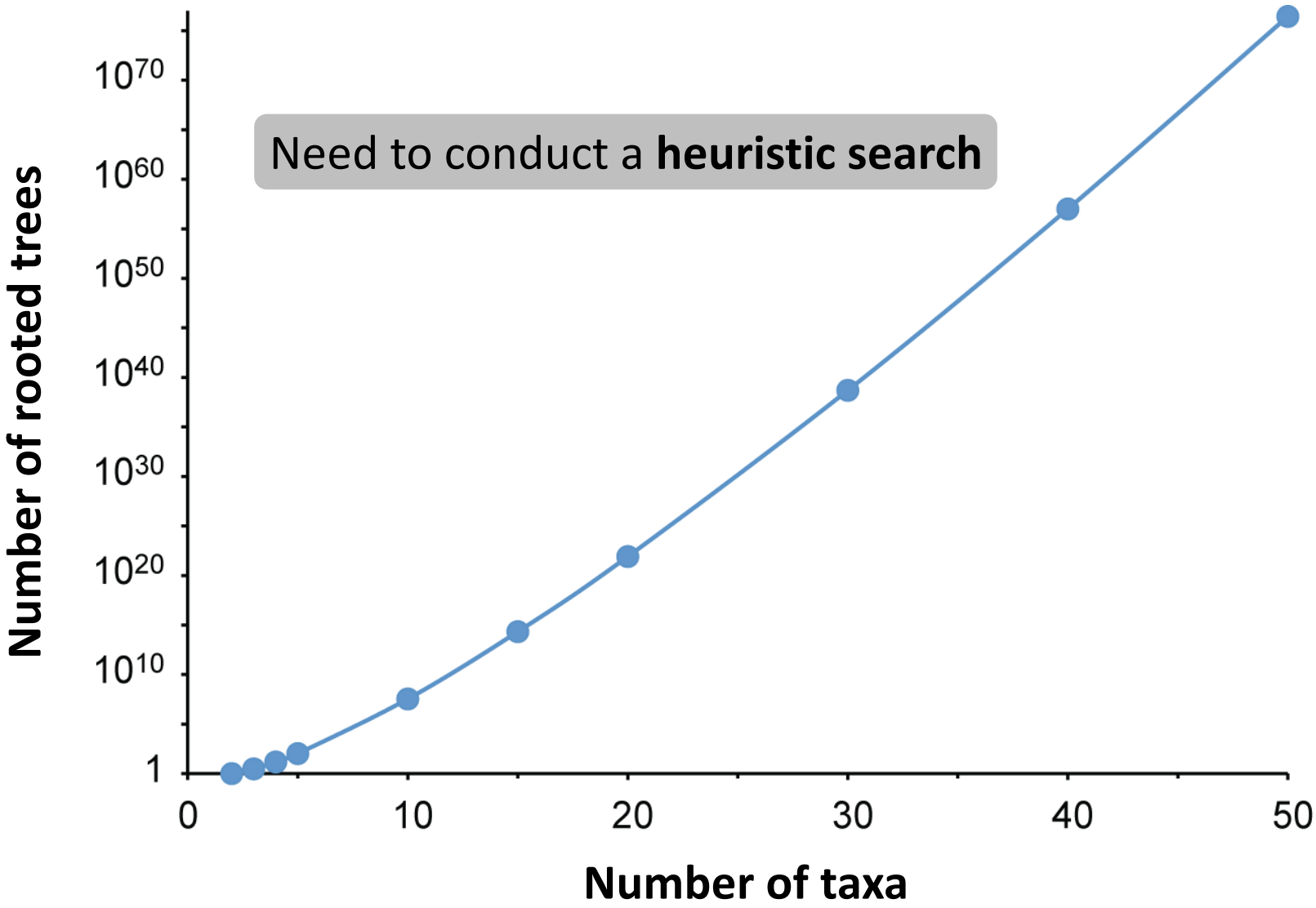
Finding the best tree

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

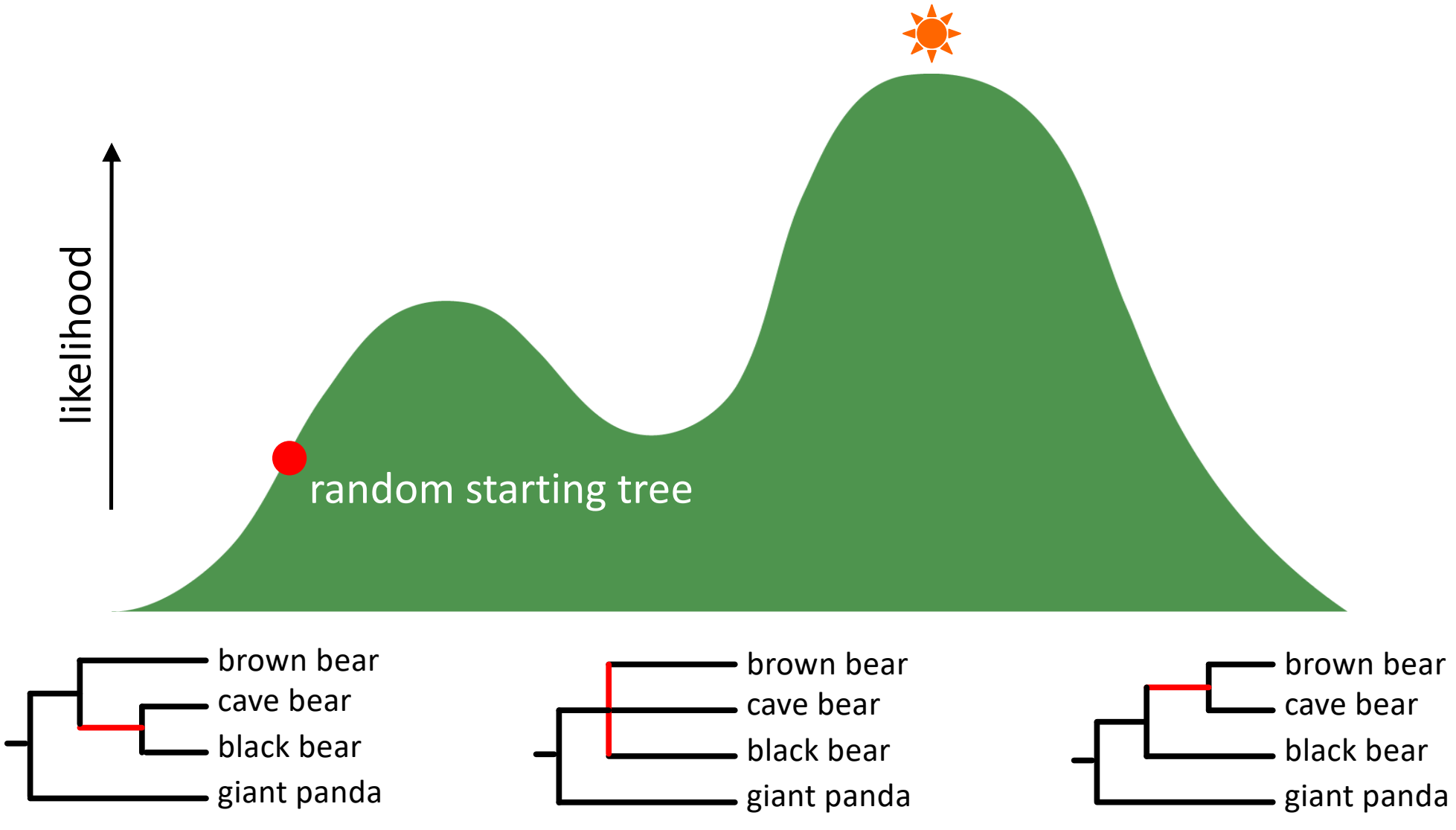
$$B_n = 1 \times 3 \times 5 \times \ldots \times (2n-5) = \prod_{i=3}^{n} (2i-5)$$

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

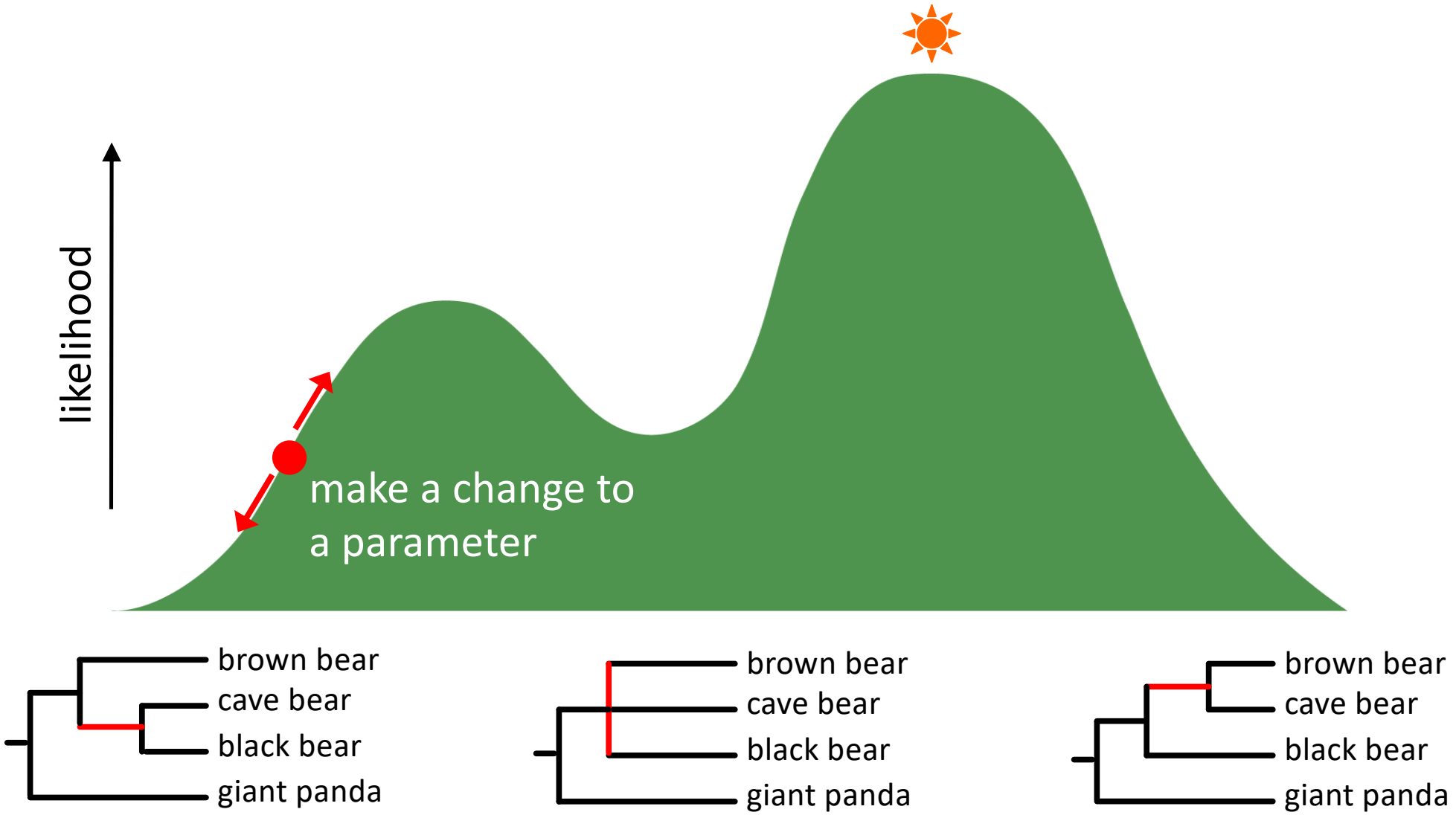
Finding the best tree



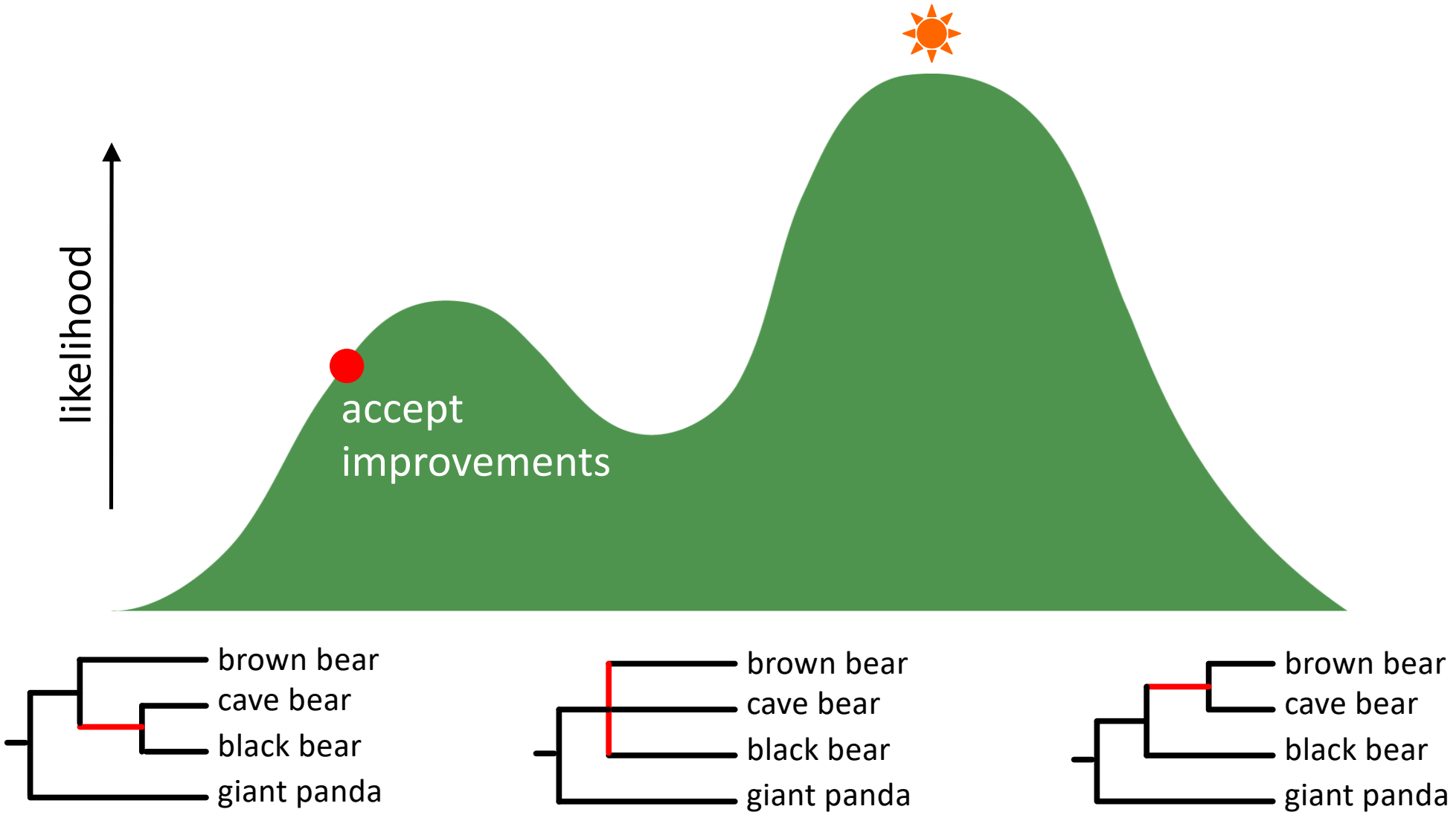
Heuristic search



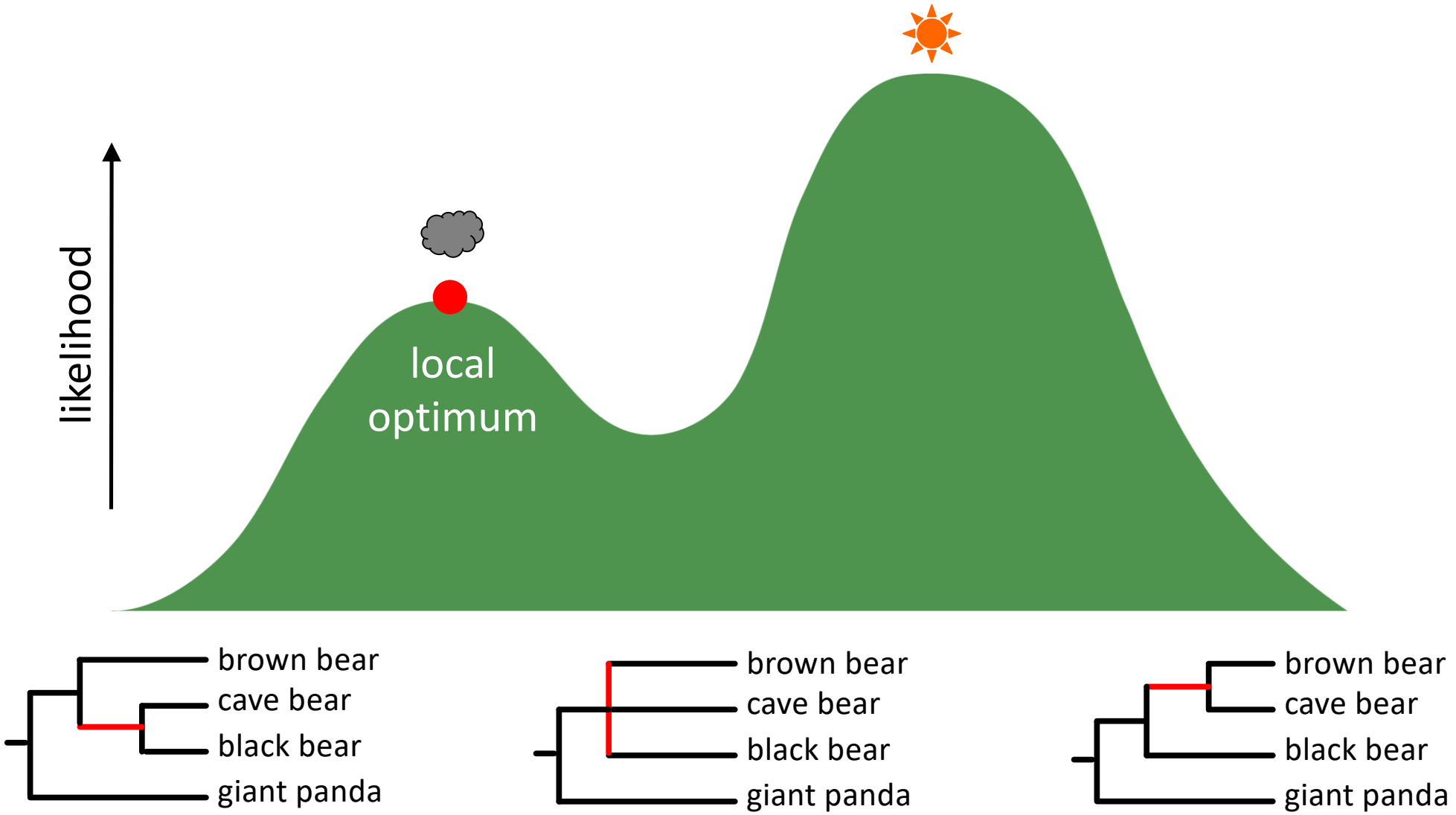
Heuristic search



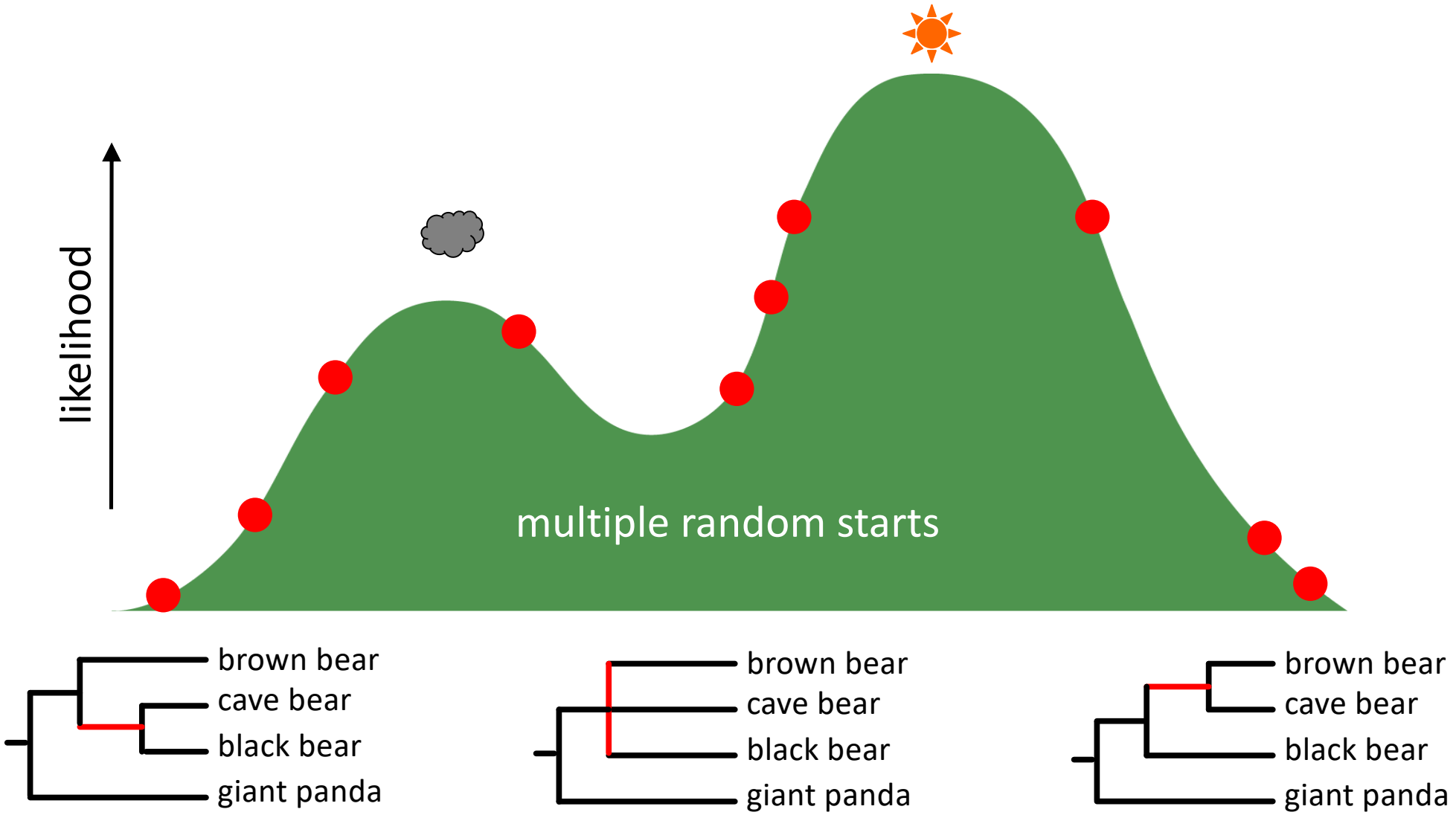
Heuristic search



Heuristic search

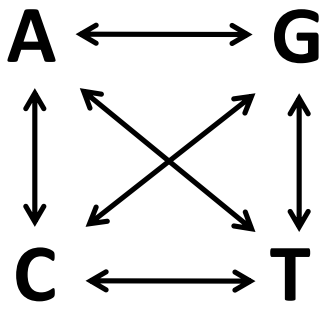


Heuristic search

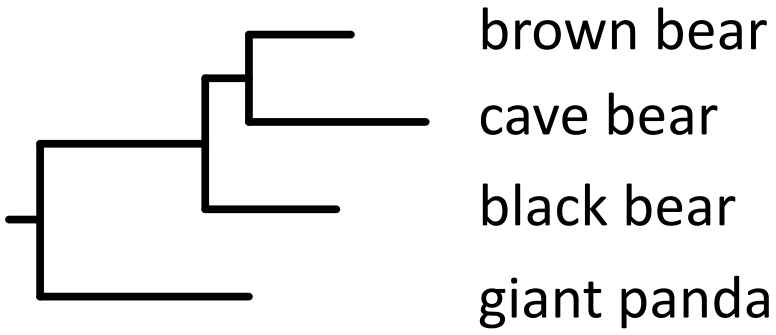


Maximum-likelihood estimates

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
  - Highly robust to violations of assumptions

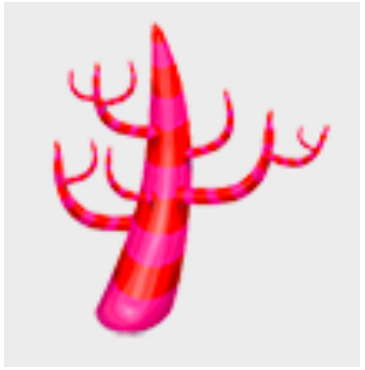
- **Weaknesses**
  - Not feasible to implement very parameter-rich models
  - Searching tree space can be difficult

Software

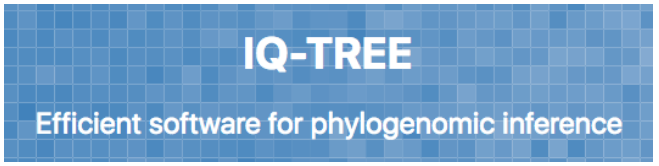
RAxML

PhyML

MEGA



PAML



IQ-TREE

RAxML

- **R**andomized **Ax**elerated **M**aximum **L**ikelihood
- Compile to suit your processor architecture
- Can run sequentially or in parallel
- Rapid bootstrapping (Stamatakis *et al*. 2008)

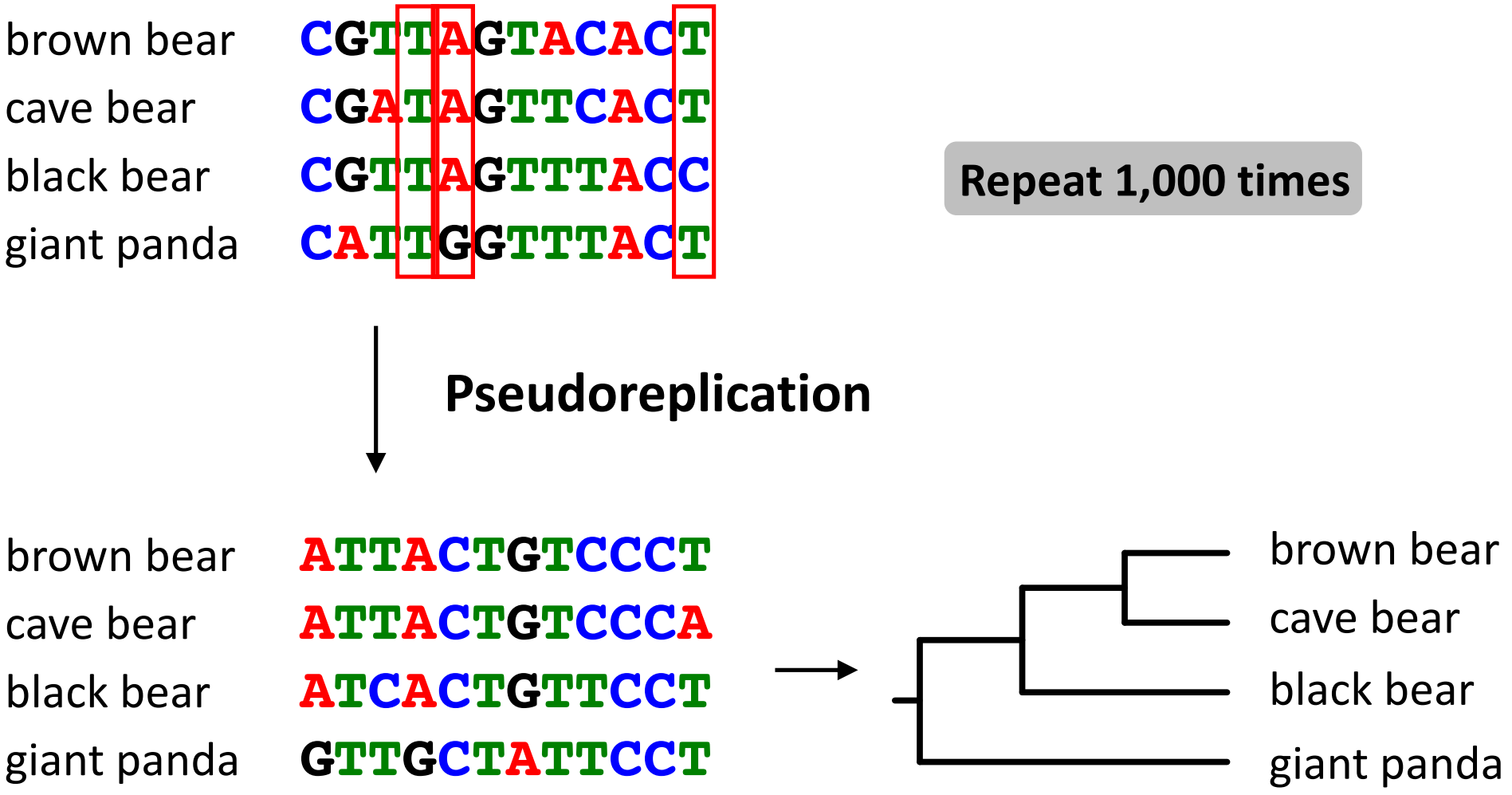
Bootstrapping

Nonparametric bootstrap

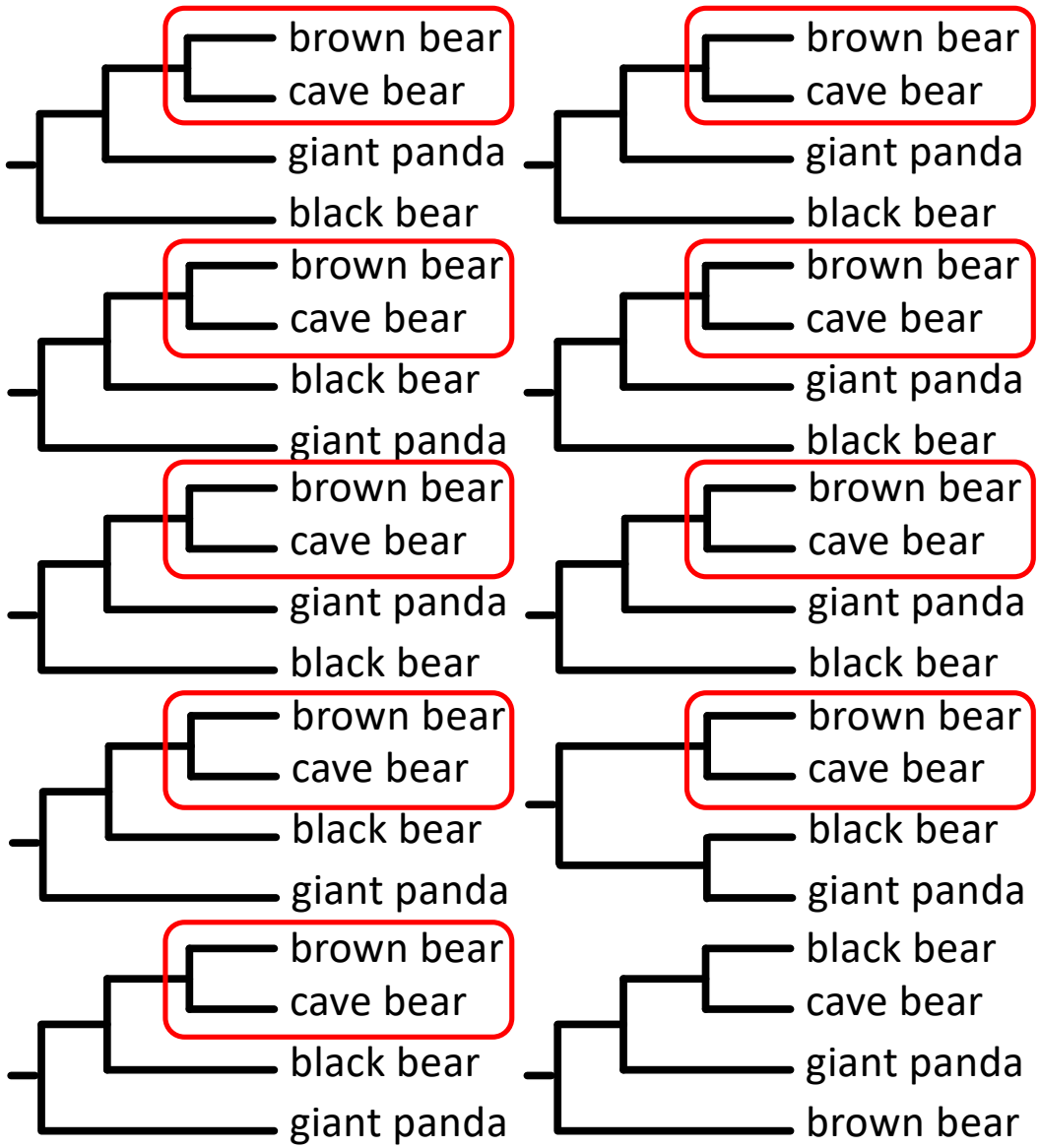
- Uncertainty in the estimate of the tree is inferred indirectly using **bootstrapping analysis**
- “pull oneself up by one's bootstraps”

  - Bootstrapping analysis can be used in various phylogenetic methods:
    - Maximum parsimony
    - Distance-based methods
    - Maximum likelihood

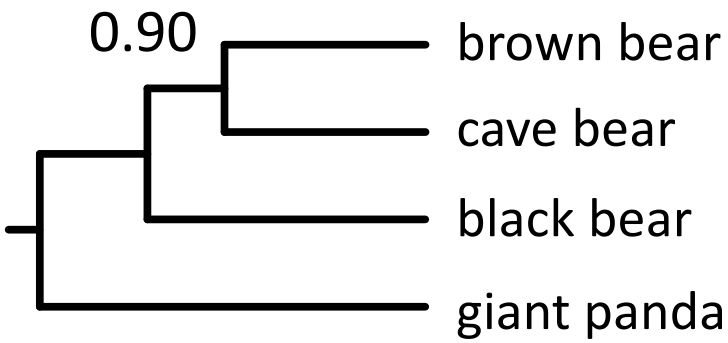
Bootstrapping



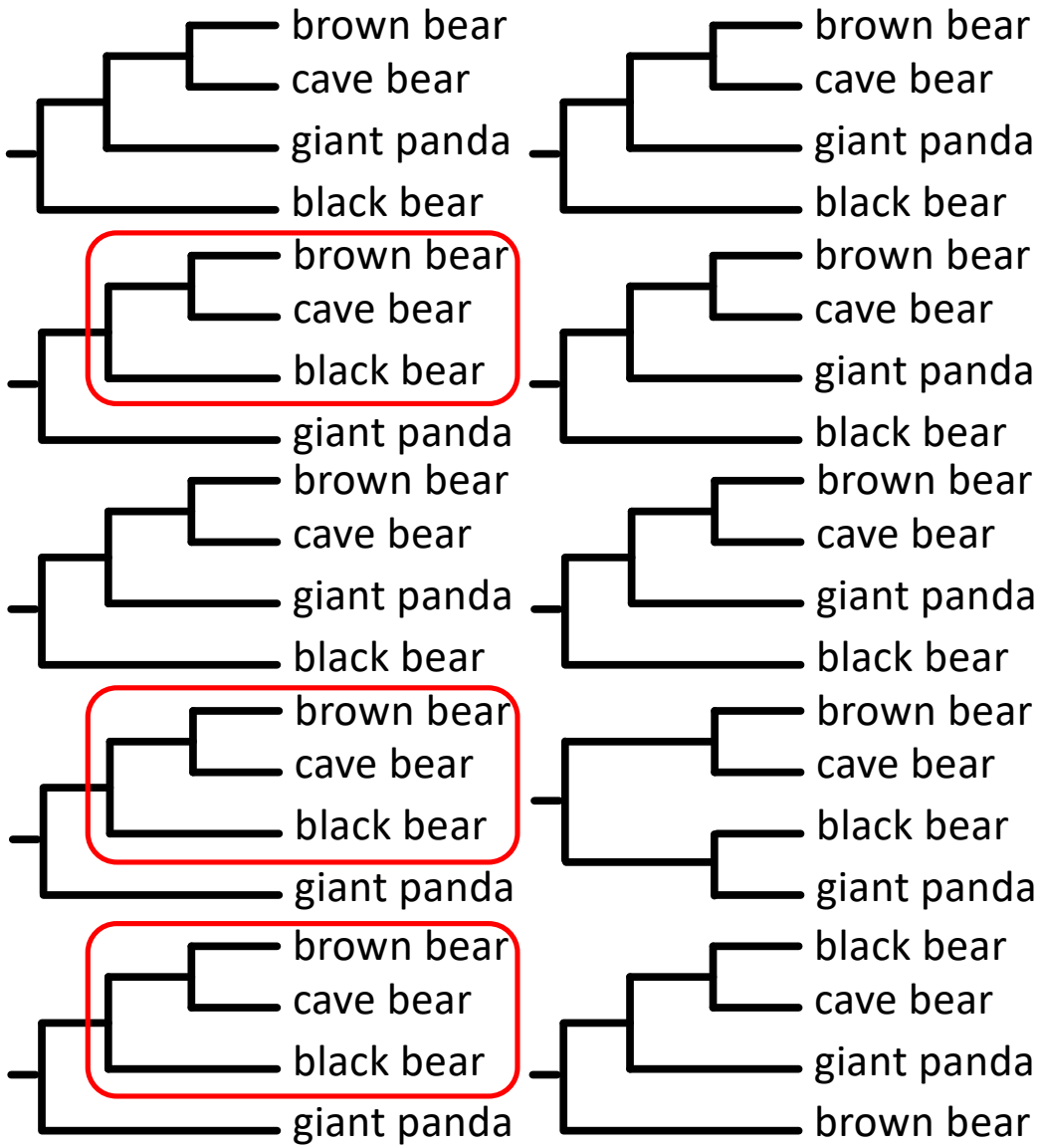
Bootstrapping



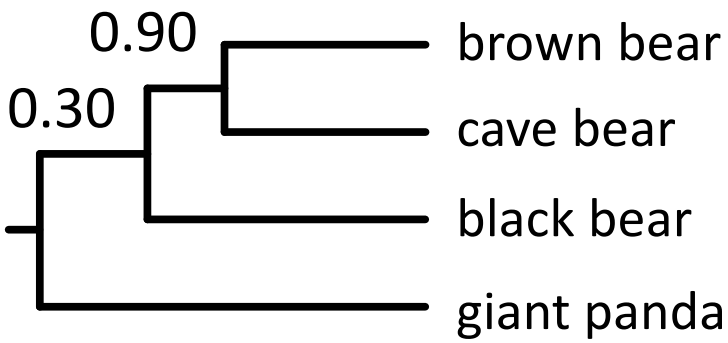
ML tree



Bootstrapping



ML tree



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, but has been losing ground to Bayesian methods

Useful references

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