Lecture 1.4

Phylogenetic Methods

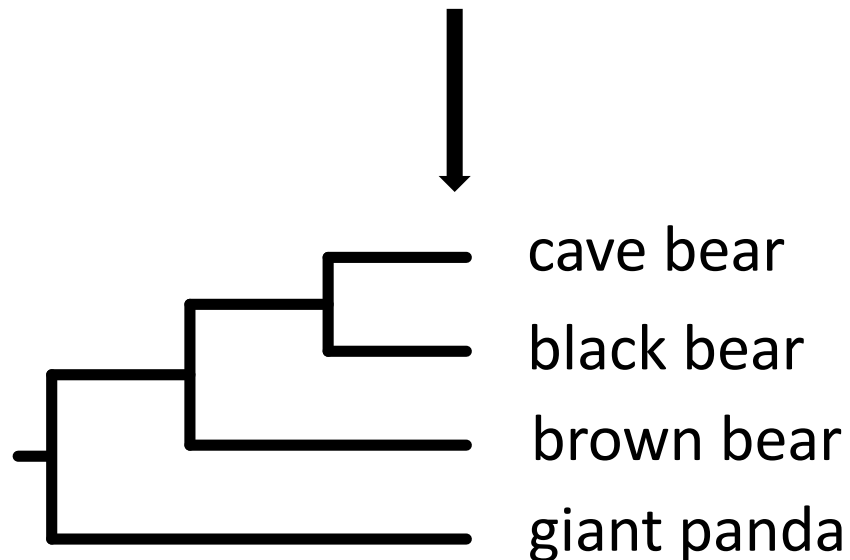# Maximum parsimony

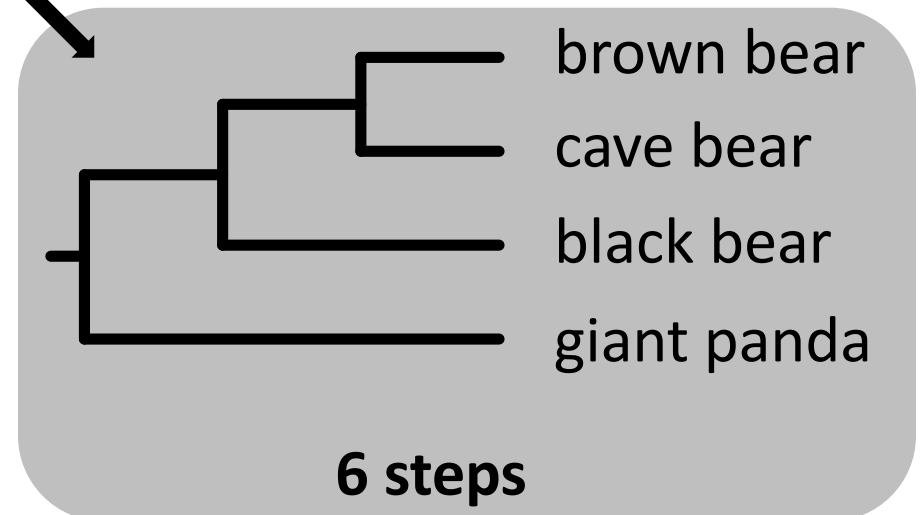brown bear **CGTTAGTACACT**
cave bear **CGATAGTTCACT**
black bear **CGTTAGTTTACC**
giant panda **CATTGGTTTACT**
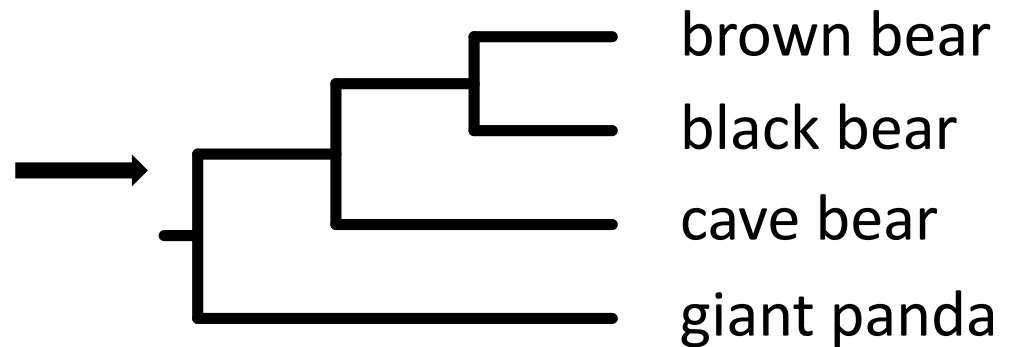
7 steps

7 steps

6 steps

2

Popular phylogenetic methods

1. Maximum parsimony

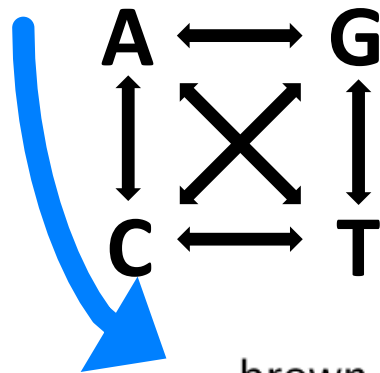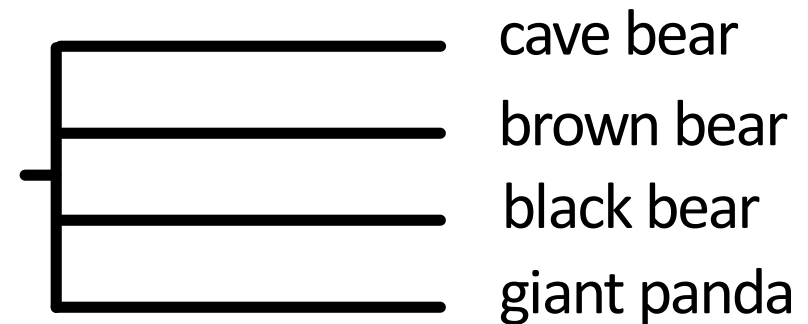2. Distance-based methods

3. Maximum likelihood   Model-based methods

4. Bayesian inference

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

cave bear
brown bear
black bear
giant panda

Neighbour joining

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

cave bear
brown bear
black bear
giant panda

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
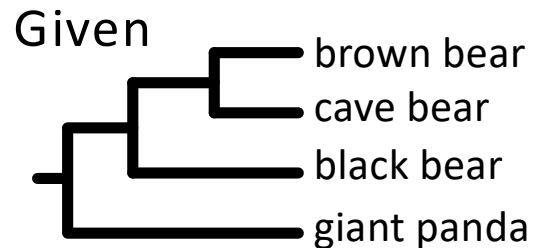 - Unable to implement sophisticated evolutionary models
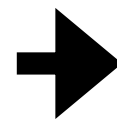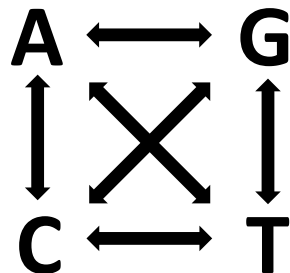
Maximum Likelihood

Maximum likelihood

Likelihood of hypothesis *H =*
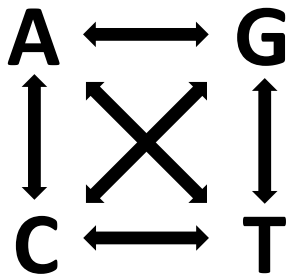
$$P(D|H)$$

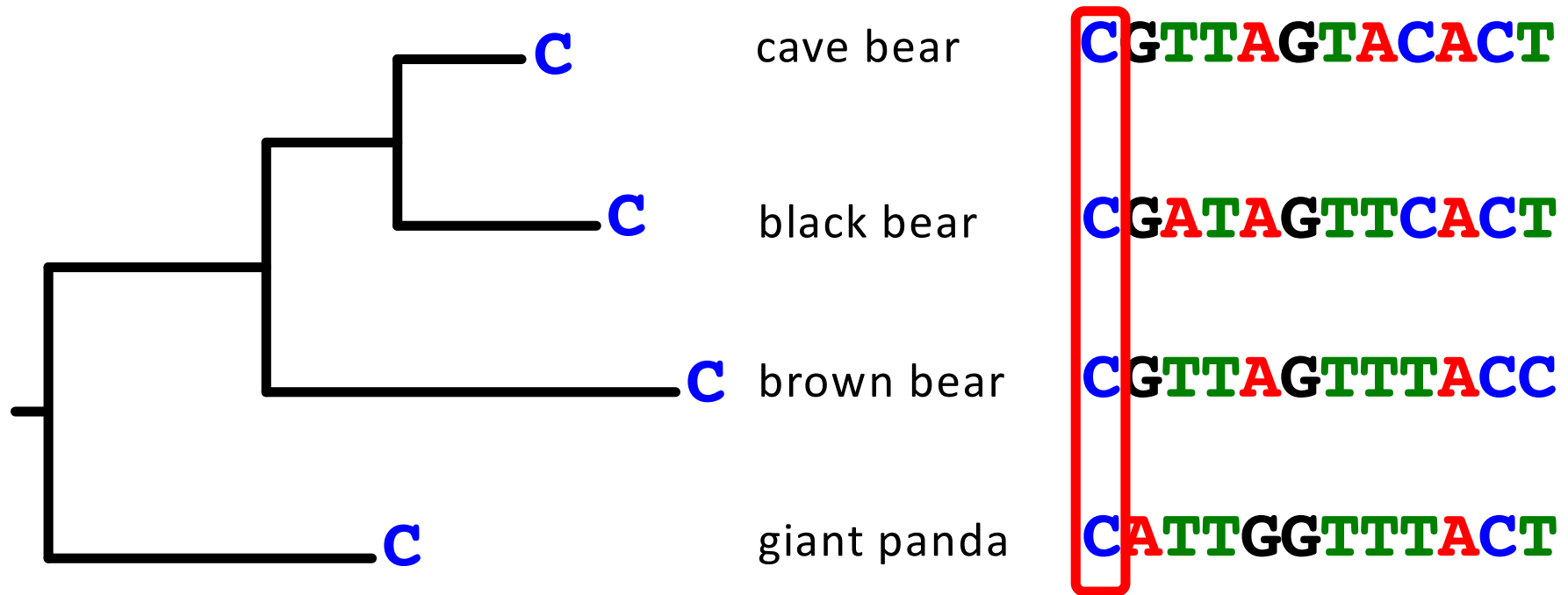Probability of the data, given the hypothesis

Given

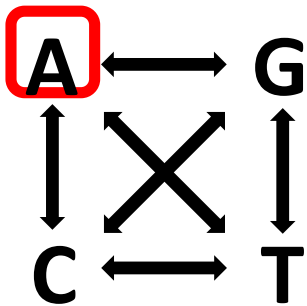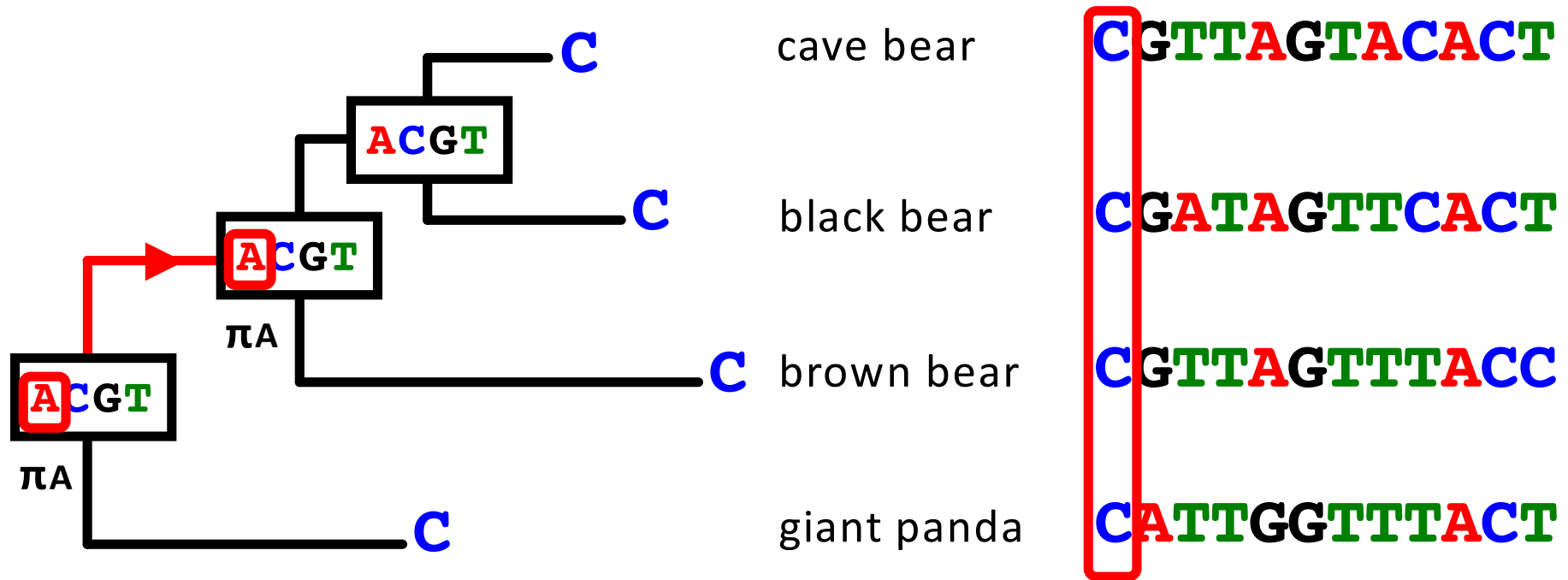|		Probability of?
brown bear		brown bear	CGTTAGTACACT
cave bear	A ⟷ G	cave bear	CGATAGTTCACT
black bear	C ⟷ T	black bear	CGTTAGTTTACC
giant panda	+	giant panda	CATTGGTTTACT

Maximum likelihood

Maximum likelihood

Maximum likelihood

Maximum likelihood

Maximum likelihood

cave bear **CGTTAGTACACT**

black bear **CGATAGTTCACT**

brown bear **CGTTAGTTTACC**

giant panda **CATTGGTTTACT**
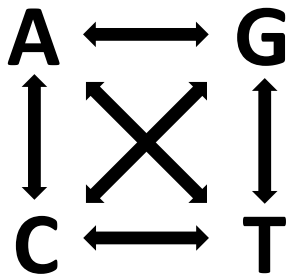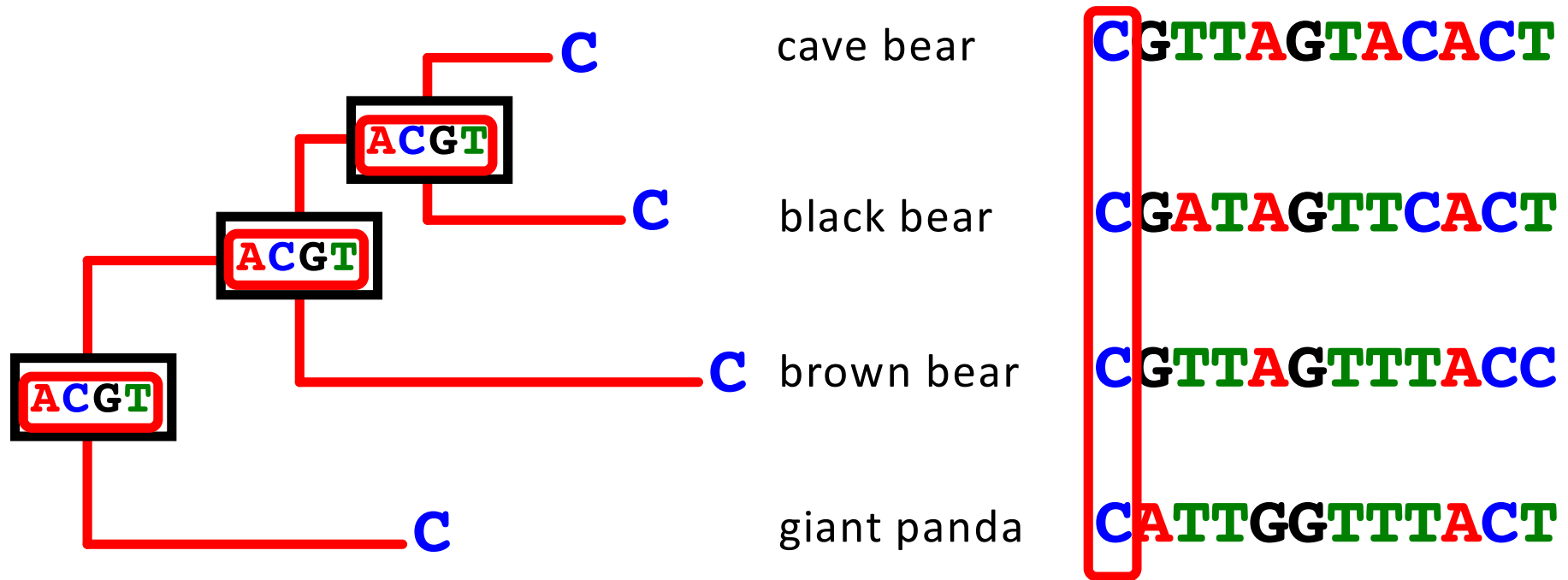
Likelihood is summed over all possibilities

Maximum likelihood

cave bear — CGTTAGTACACT

black bear — CGATAGTTCACT

brown bear — CGTTAGTTTACC

giant panda — CATTGGTTTACT
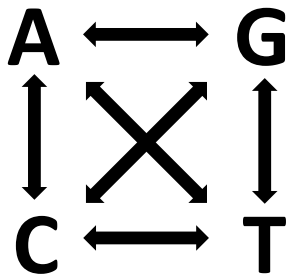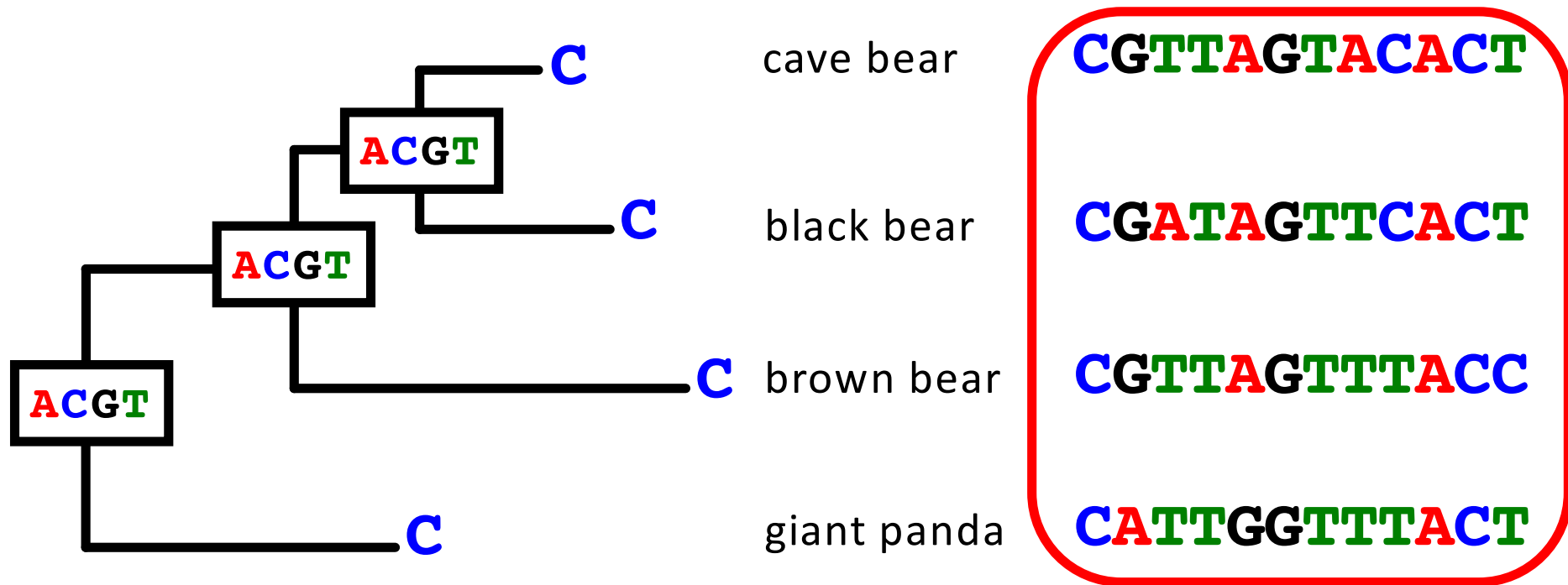
Likelihood is multiplied across all sites

Very low probability of observing any particular alignment

Maximum likelihood

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

brown bear
black bear
cave bear
giant panda

lnL = -1203.83

cave bear
black bear
brown bear
giant panda

lnL = -1241.47

brown bear
cave bear
black bear
giant panda
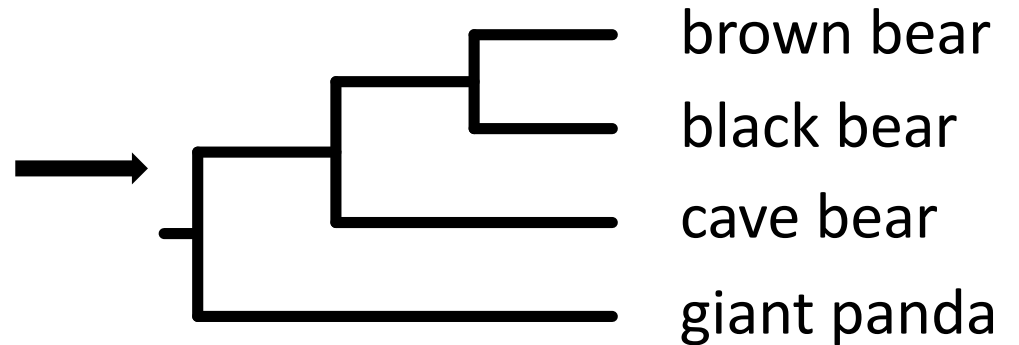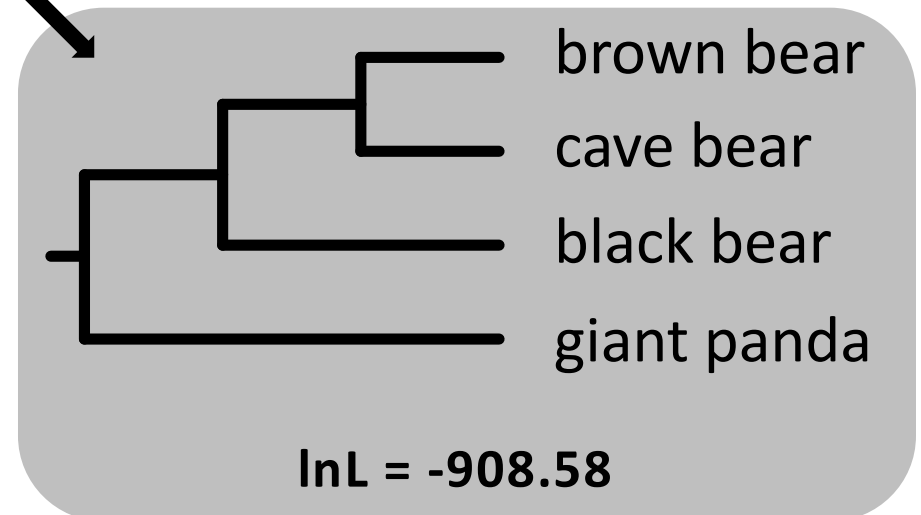
lnL = -908.58

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

Likelihood optimisation

Heuristic search

likelihood

random starting tree

brown bear
cave bear
black bear
giant panda

brown bear
cave bear
black bear
giant panda

brown bear
cave bear
black bear
giant panda

Heuristic search

likelihood

make a change to
a parameter

brown bear
cave bear
black bear
giant panda

brown bear
cave bear
black bear
giant panda

brown bear
cave bear
black bear
giant panda

Heuristic search

likelihood

accept
improvements

brown bear
cave bear
black bear
giant panda

brown bear
cave bear
black bear
giant panda

brown bear
cave bear
black bear
giant panda

Heuristic search
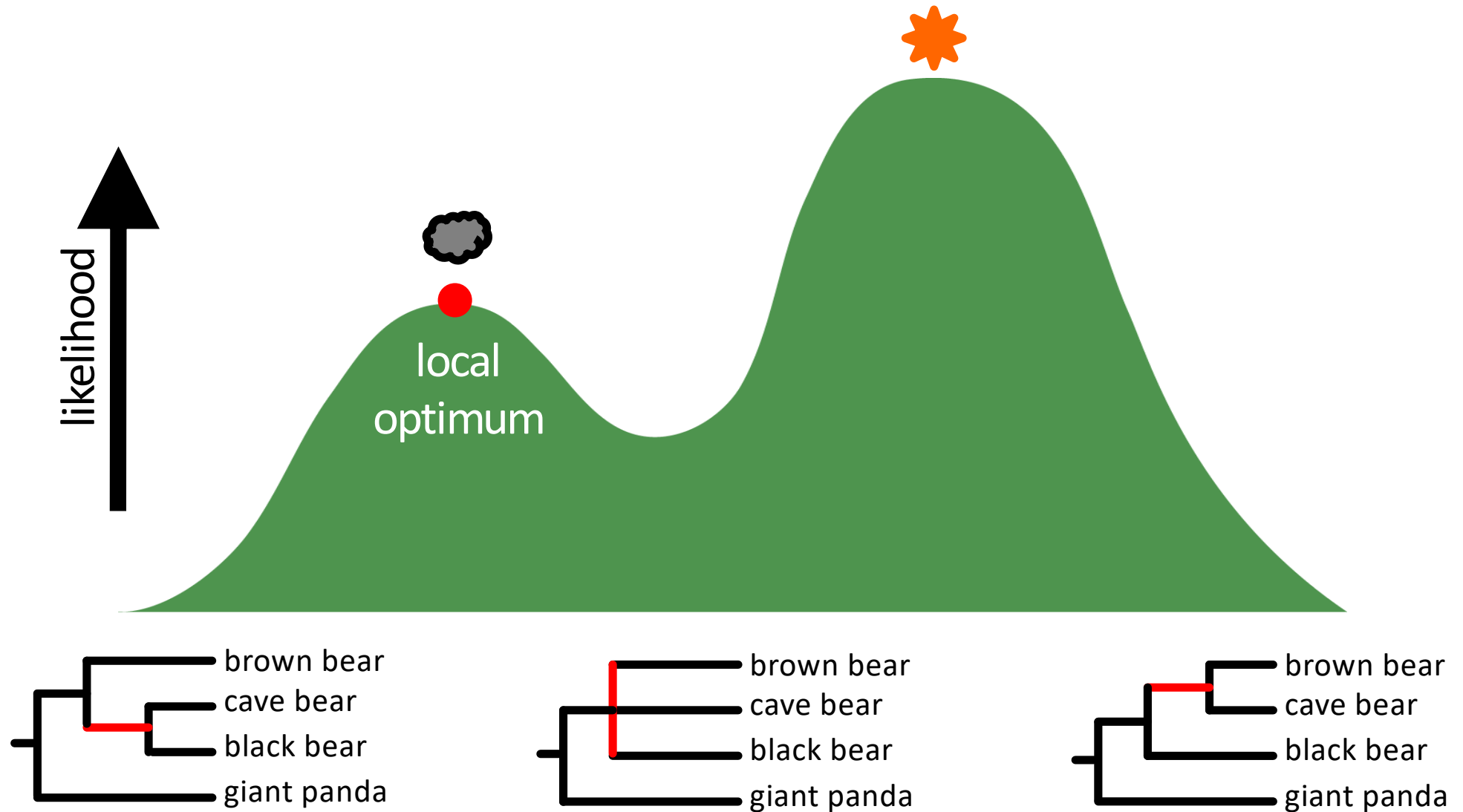
Heuristic search

Maximum-likelihood estimates

A single set of maximum-likelihood estimates of model parameters

A single maximum-likelihood tree

brown bear
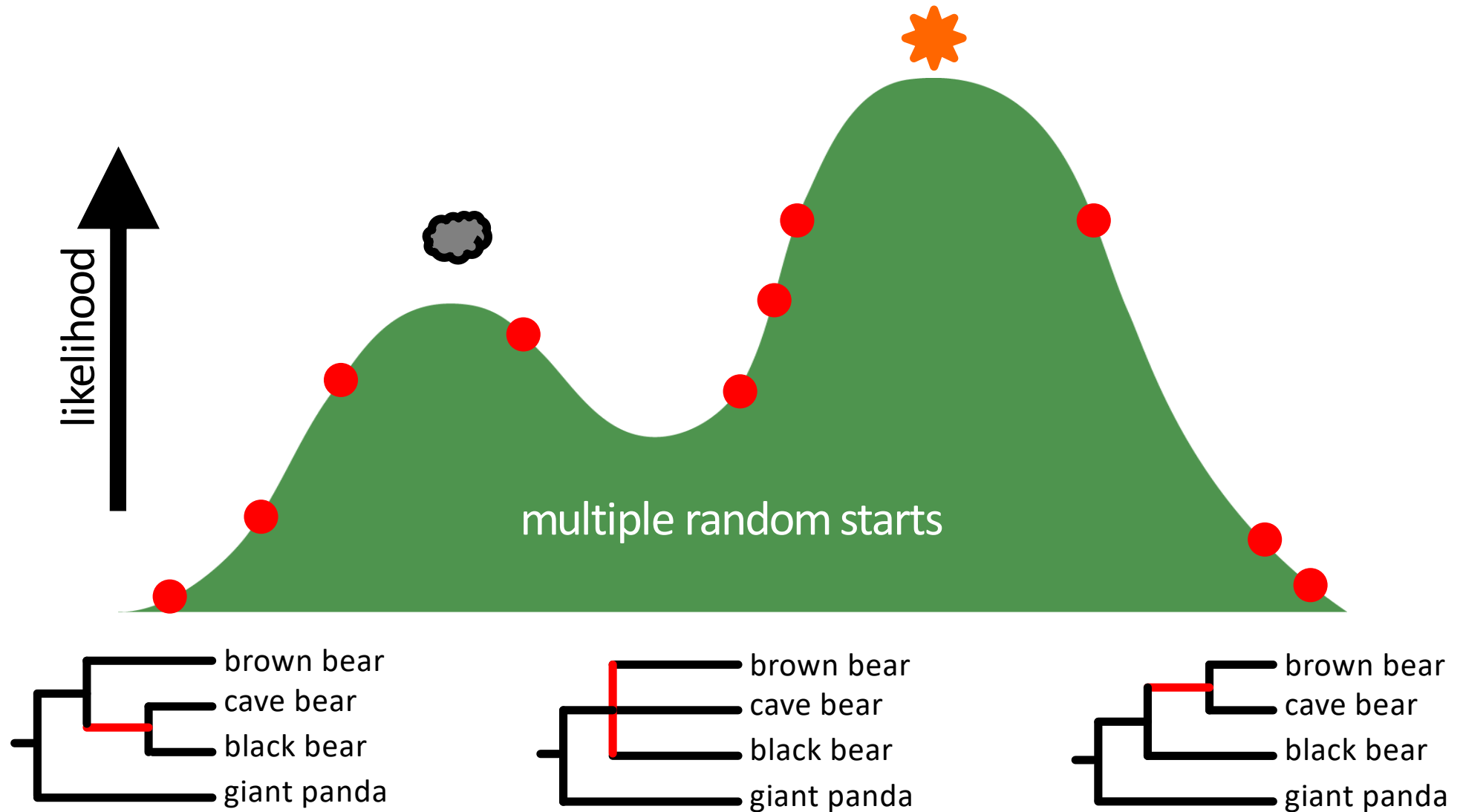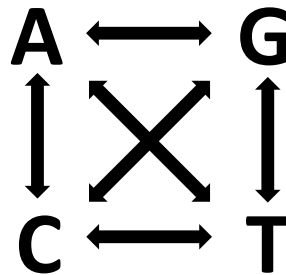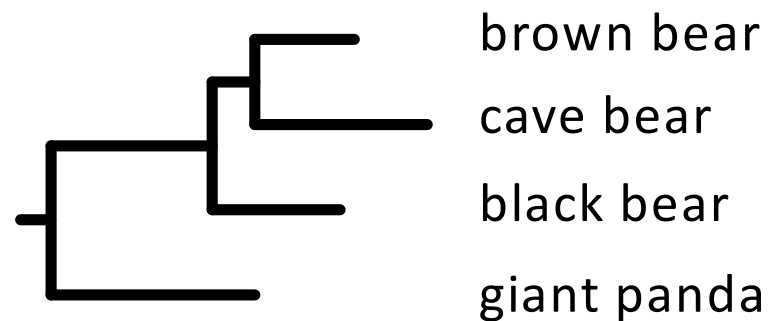
cave bear

black bear

giant panda

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
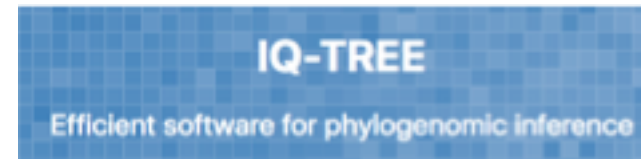

PhyML

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
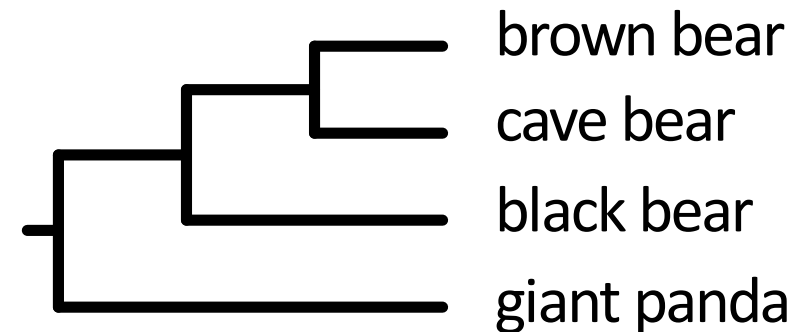
    - Distance-based methods

    - Maximum likelihood

Bootstrapping

Bootstrapping

ML tree

0.90

- brown bear
- cave bear
- black bear
- giant panda
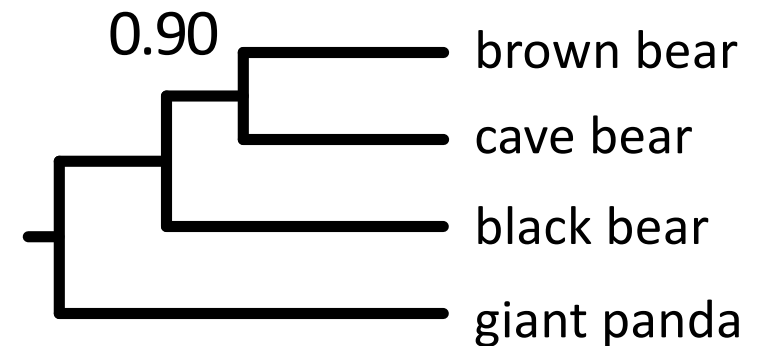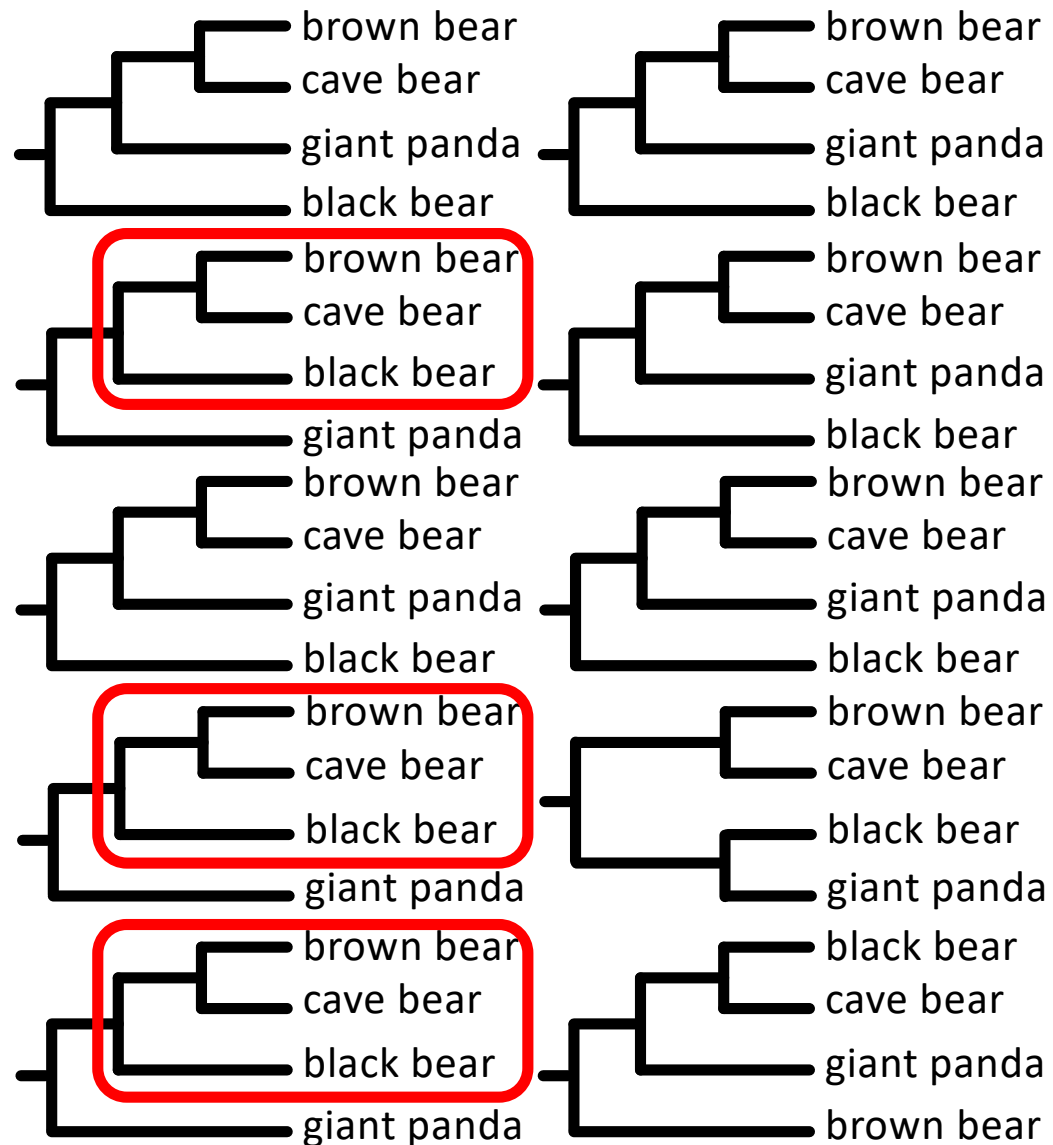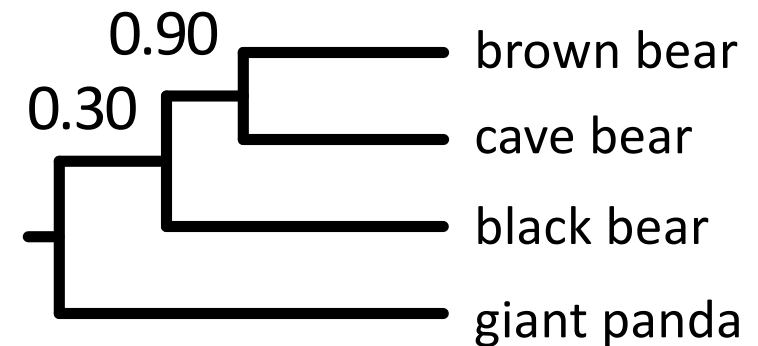
Bootstrapping

ML tree

34
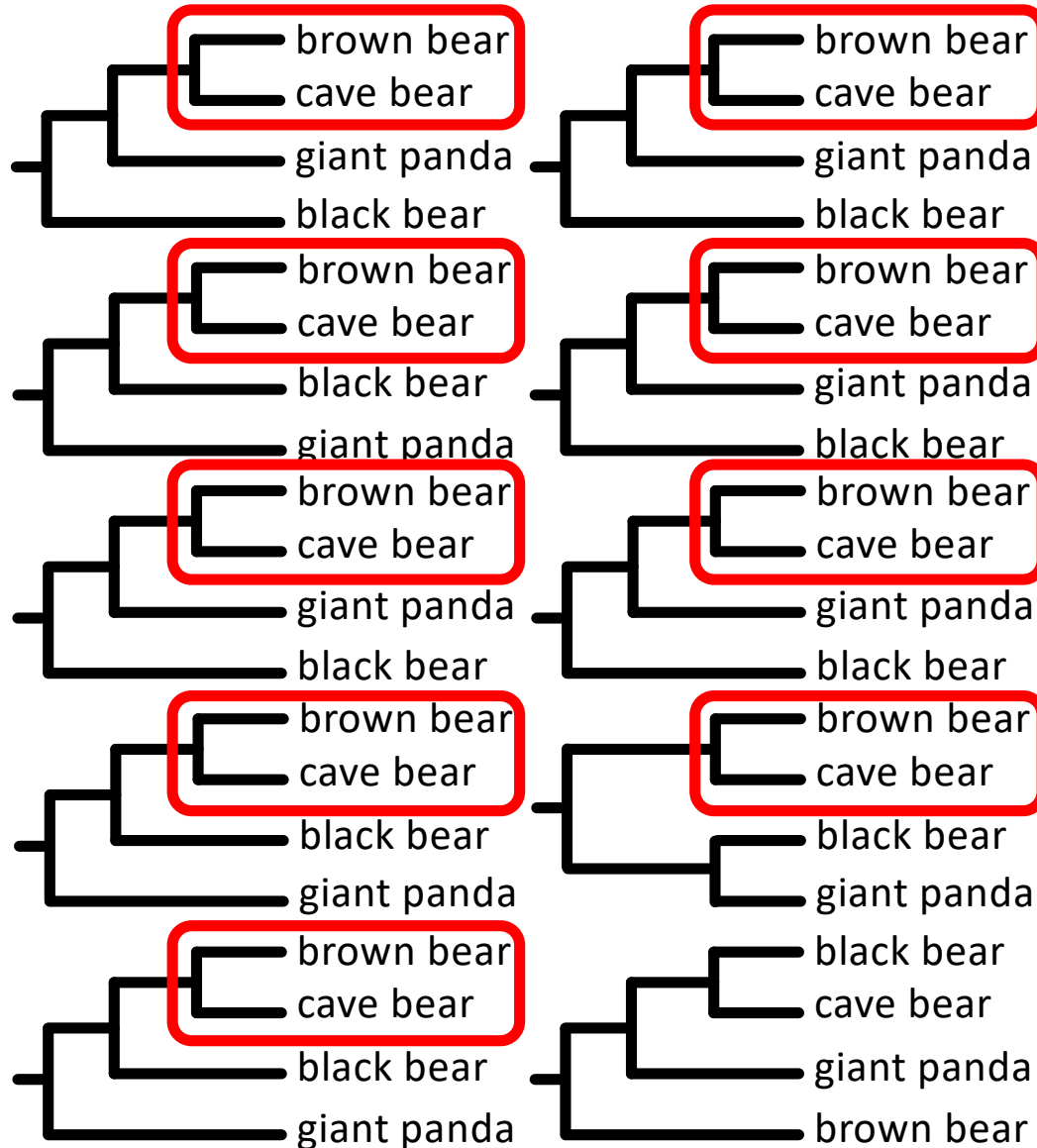
Interpreting bootstrap values

- **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*

Useful references

- **Phylogeny estimation and hypothesis testing using maximum likelihood**

  Huelsenbeck & Crandall (1997) *Annu Rev Ecol Syst,* 28: 437–466.