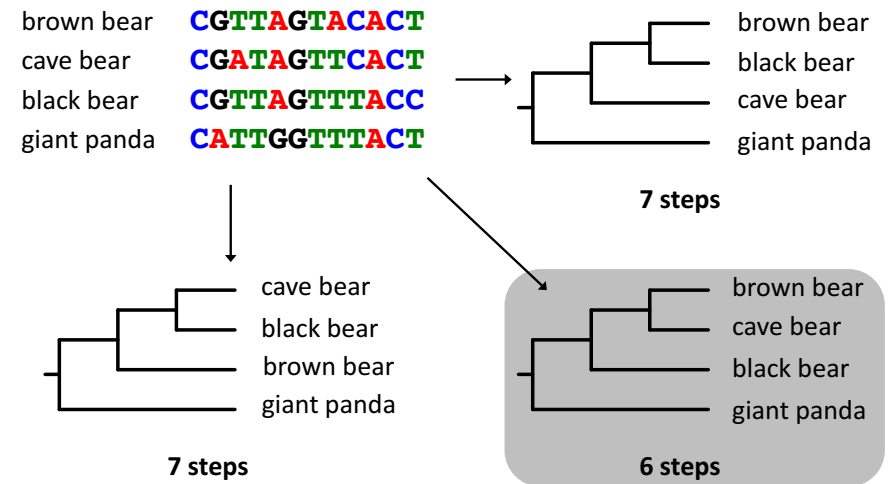


Lecture 1.4

Phylogenetic Methods

Simon Ho

Maximum parsimony



2

Popular phylogenetic methods

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

Model-based methods

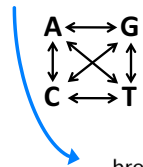
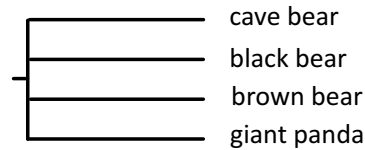


3

Distance-Based Methods

Neighbour joining

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



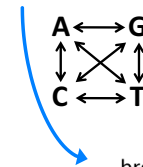
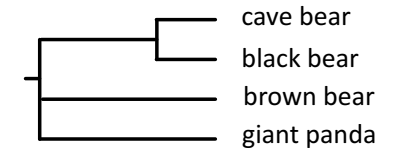
	brown bear	cave bear	black bear	giant panda
brown bear	-			
cave bear	.1	-		
black bear	.3	.3	-	
giant panda	.4	.5	.4	-

Clustering
algorithm

5

Neighbour joining

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



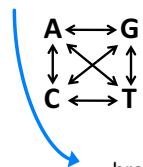
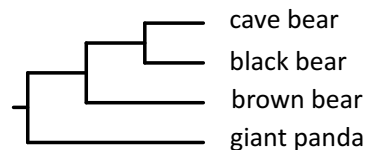
	brown bear	cave bear	black bear	giant panda
brown bear	-			
cave bear	.1	-		
black bear	.3	.3	-	
giant panda	.4	.5	.4	-

Clustering
algorithm

6

Neighbour joining

brown bear **CGTTAGTACACT**
 cave bear **CGATAGTTCACCT**
 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

7

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

8

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

9

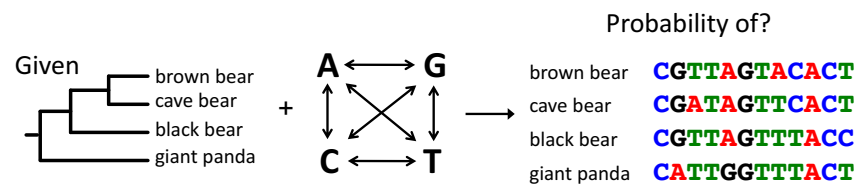
Maximum Likelihood

Maximum likelihood

Likelihood of hypothesis $H =$

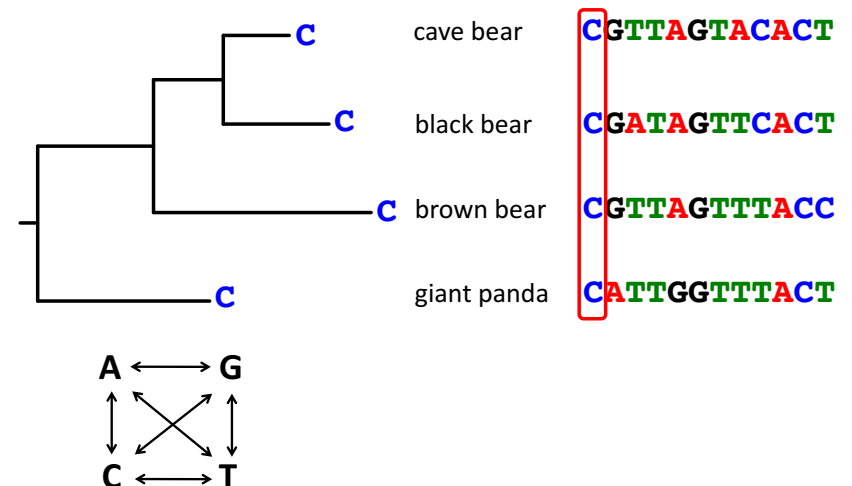
$$P(D/H)$$

Probability of the data, given the hypothesis



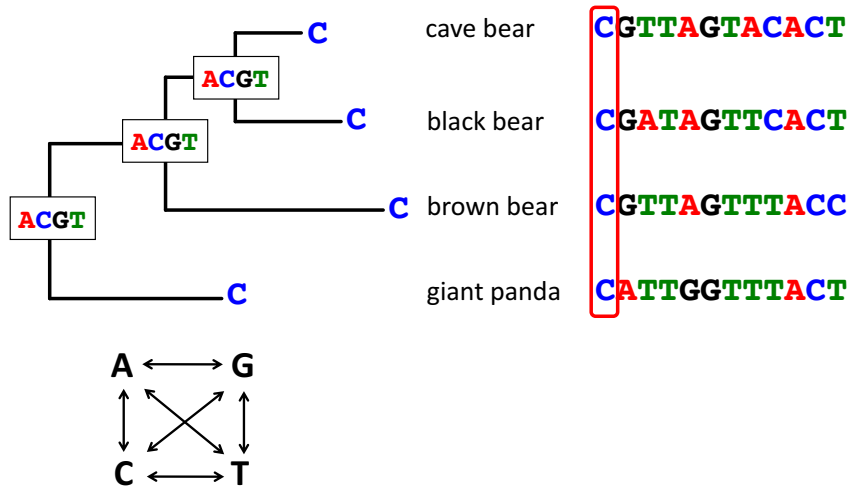
11

Maximum likelihood



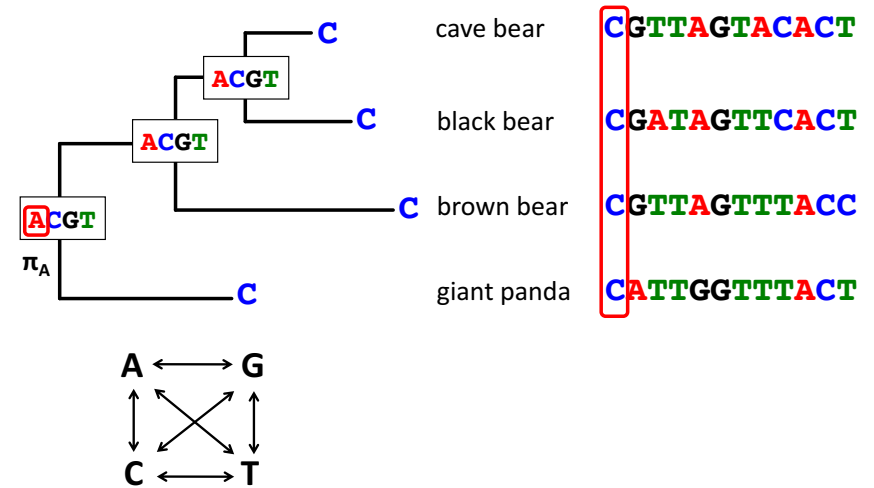
12

Maximum likelihood



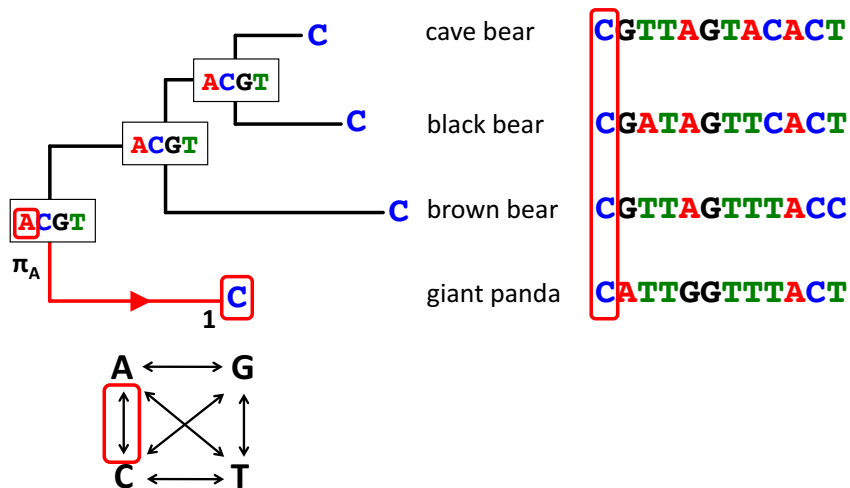
13

Maximum likelihood



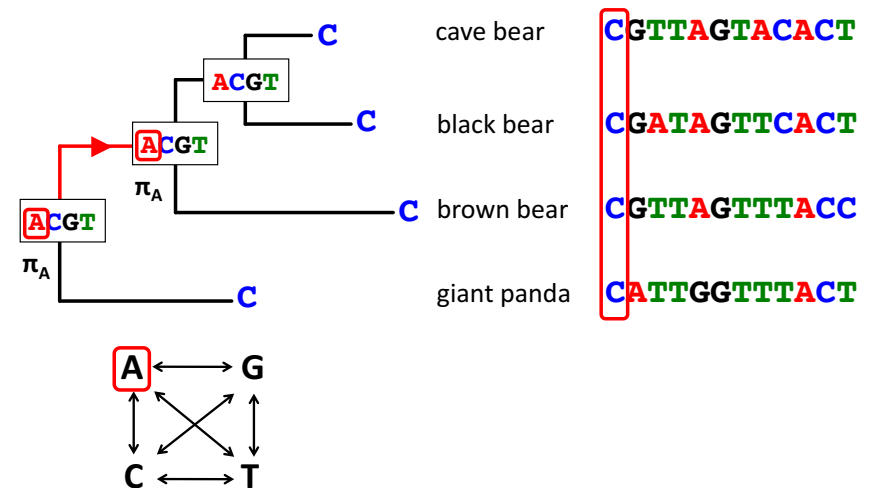
14

Maximum likelihood



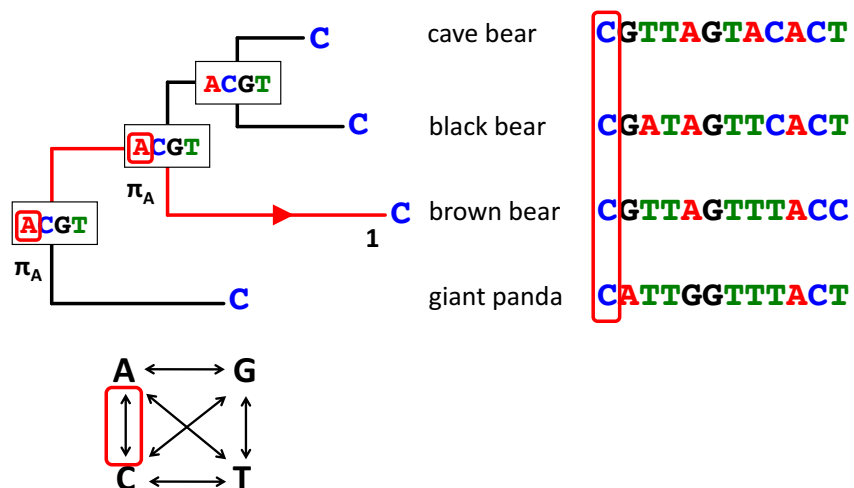
15

Maximum likelihood



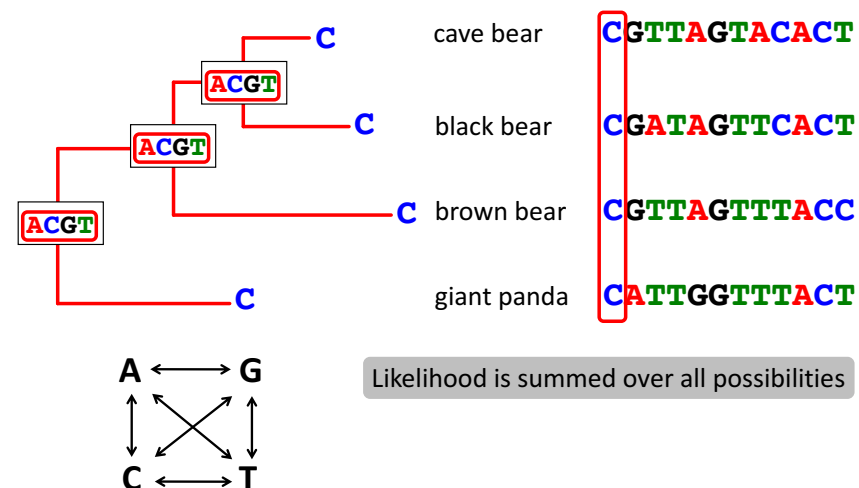
16

Maximum likelihood



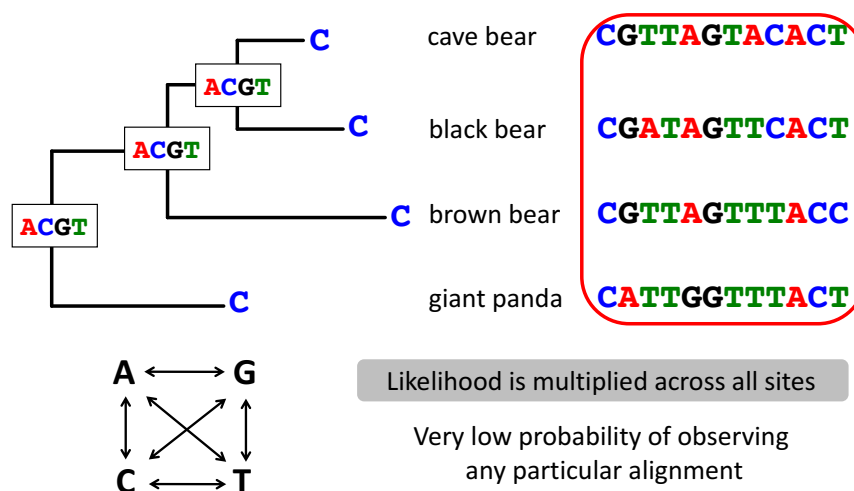
17

Maximum likelihood



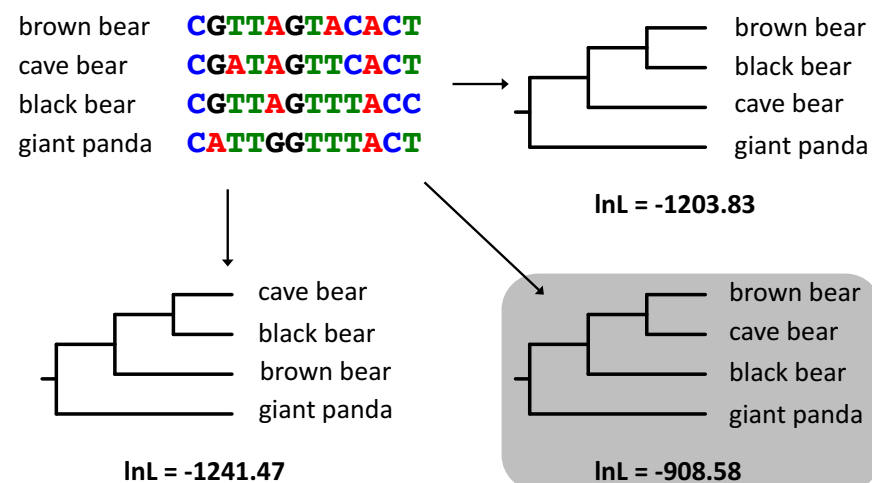
18

Maximum likelihood



19

Maximum likelihood



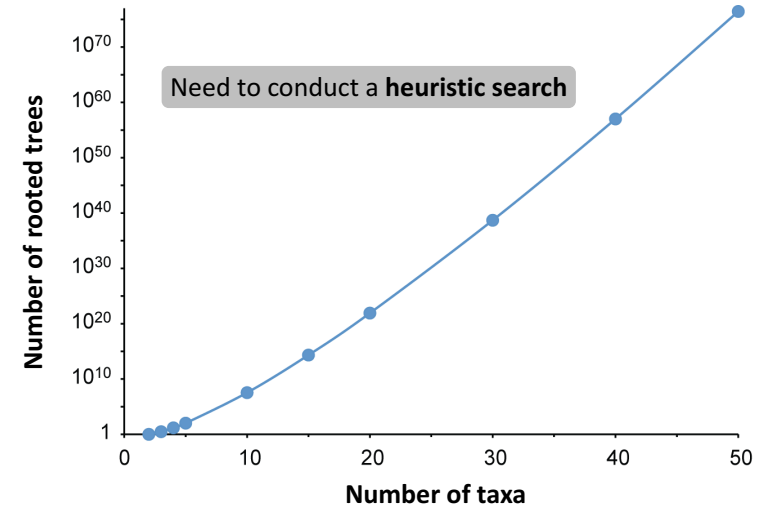
20

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

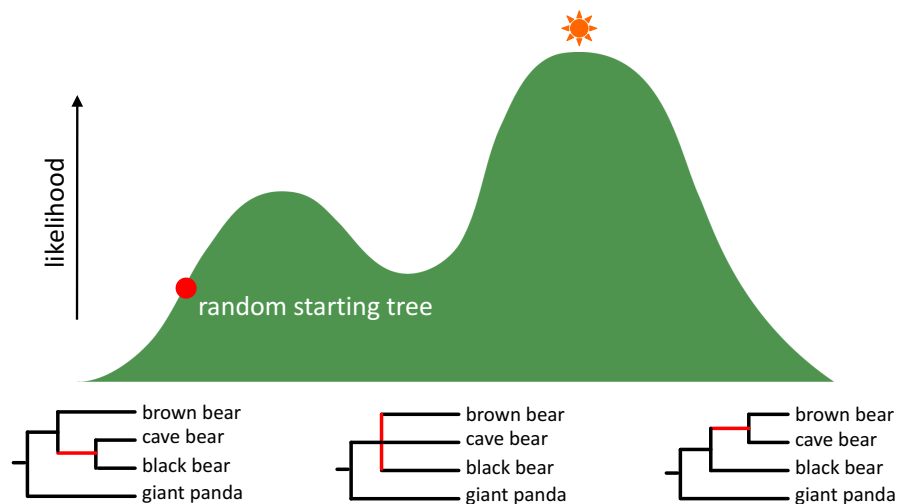
21

Likelihood optimisation



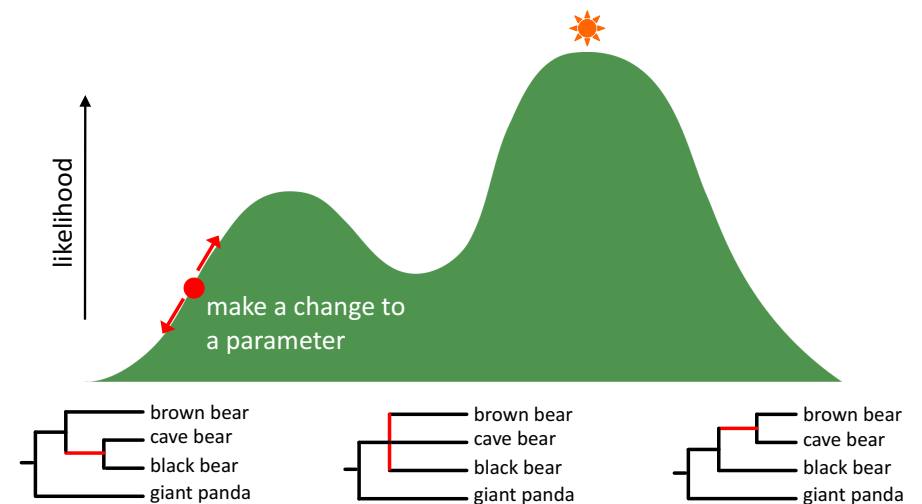
22

Heuristic search



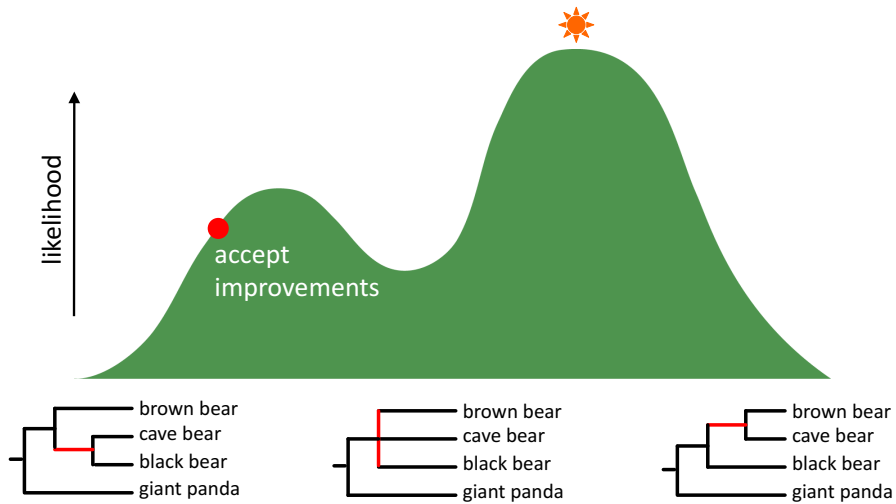
23

Heuristic search



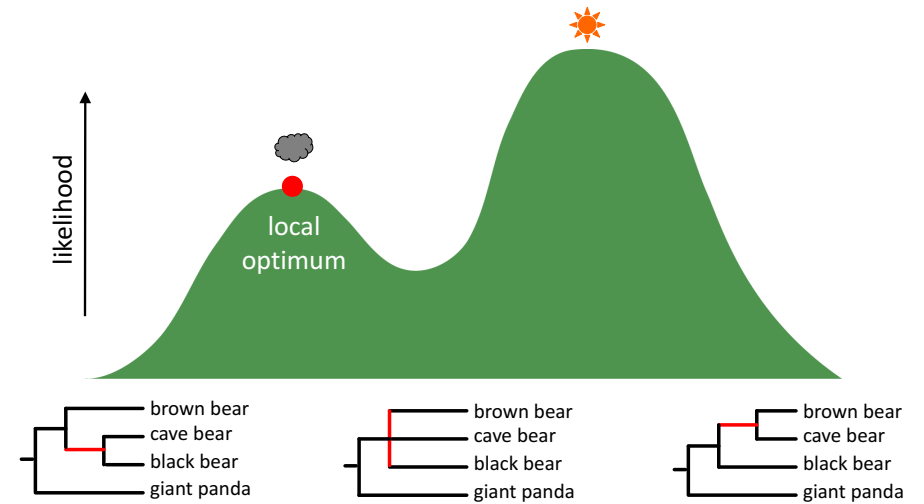
24

Heuristic search



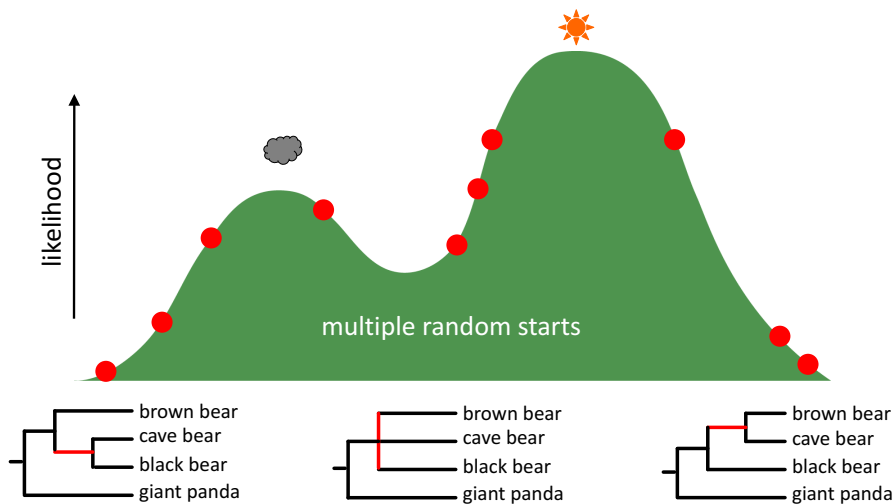
25

Heuristic search



26

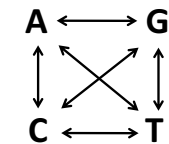
Heuristic search



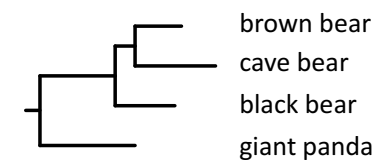
27

Maximum-likelihood estimates

A single set of maximum-likelihood estimates of model parameters



A single maximum-likelihood tree



28

Strengths and weaknesses

- **Strengths**

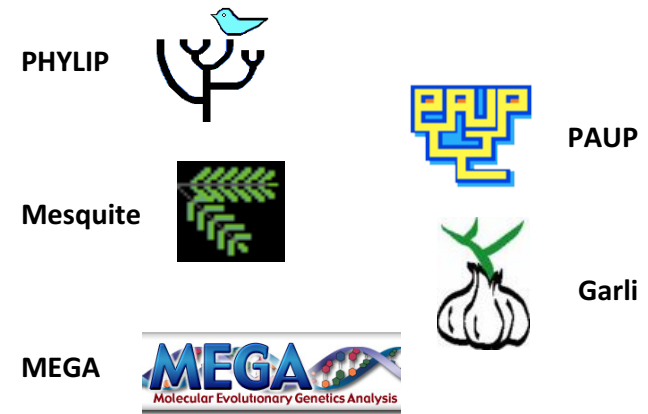
- Rigorous statistical method
- Deals with multiple substitutions and long-branch attraction
- Highly robust to violations of assumptions

- **Weaknesses**

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

29

Software



30

RAXML

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



31

ExaML

- **Exa**scale **M**aximum **L**ikelihood
- Phylogenetic inference on supercomputers
- New MPI parallelisation approach
- Koslov, Aberer, & Stamatakis (2015) *Bioinformatics*



32

Bootstrapping

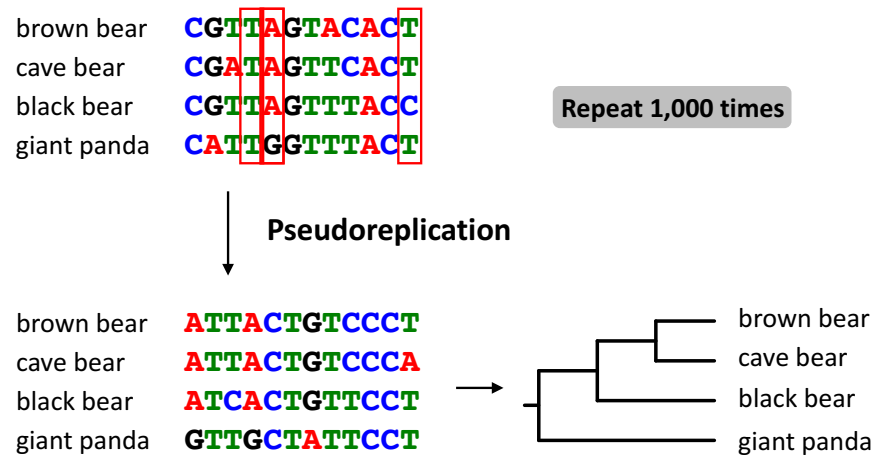
Nonparametric bootstrap

- 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



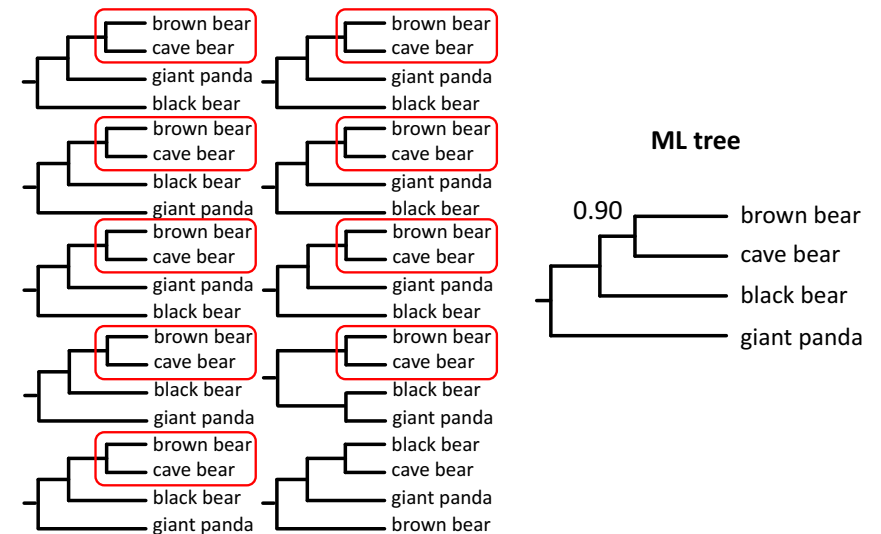
34

Bootstrapping



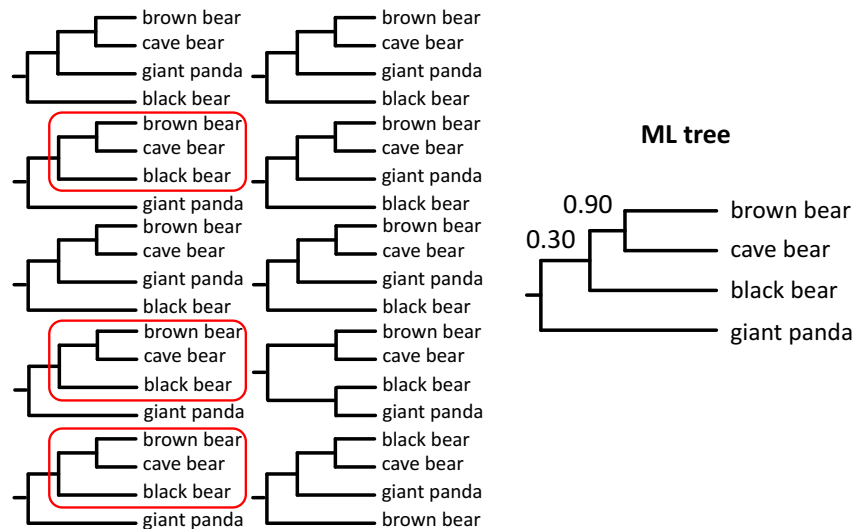
35

Bootstrapping



36

Bootstrapping



37

Interpreting bootstrap values

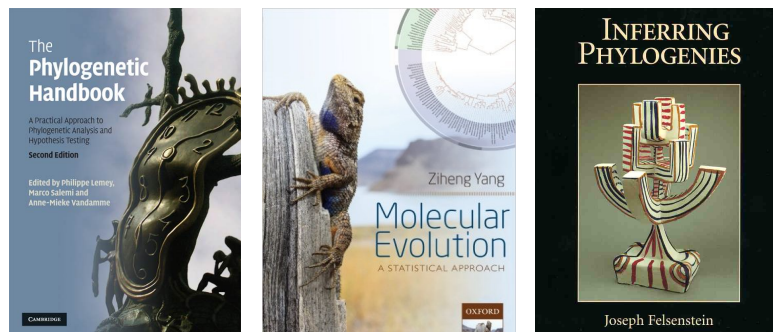
- **Felsenstein (1985)**
bootstrapping provides a confidence interval that contains the *phylogeny that would be estimated from repeated sampling of many characters from the underlying set of all characters*
- Bootstrap values are **measures of repeatability**
 - High when the data set is large
 - Not meaningful when analysing genome-scale data

Soltis & Soltis (2003) *Stat Sci*

38

Useful references

- **Phylogeny estimation and hypothesis testing using maximum likelihood**
Huelsenbeck & Crandall (1997) *Annu Rev Ecol Syst*, 28: 437–466.



39

Phylogenetic methods

	Algorithm-based	Optimality criterion	Other
No explicit substitution model	Distance-based methods	Maximum parsimony	
$ \begin{array}{ccc} A & \longleftrightarrow & G \\ \updownarrow & \times & \updownarrow \\ C & \longleftrightarrow & T \end{array} $	Distance-based methods	Maximum likelihood	Bayesian inference

40