Lecture 2

Data and methods

Data in phylogenetics

1. **Data preparation**
   - Sample taxonomic groups and genomic regions
   - Alignment
   - Data filtering

2. **Phylogenetic inference**
   - Method selection
   - Parameter estimation (including the tree)
   - Additional analysis and interpretation

Data in phylogenetics

- Select data to optimize signal:noise
  - Slowly evolving regions for ancient evolutionary events
  - Regions that evolve quickly for recent evolutionary events
- Homoplasy
  - Organisms have similarities that do not reflect evolutionary history
- Take advantage of available resources

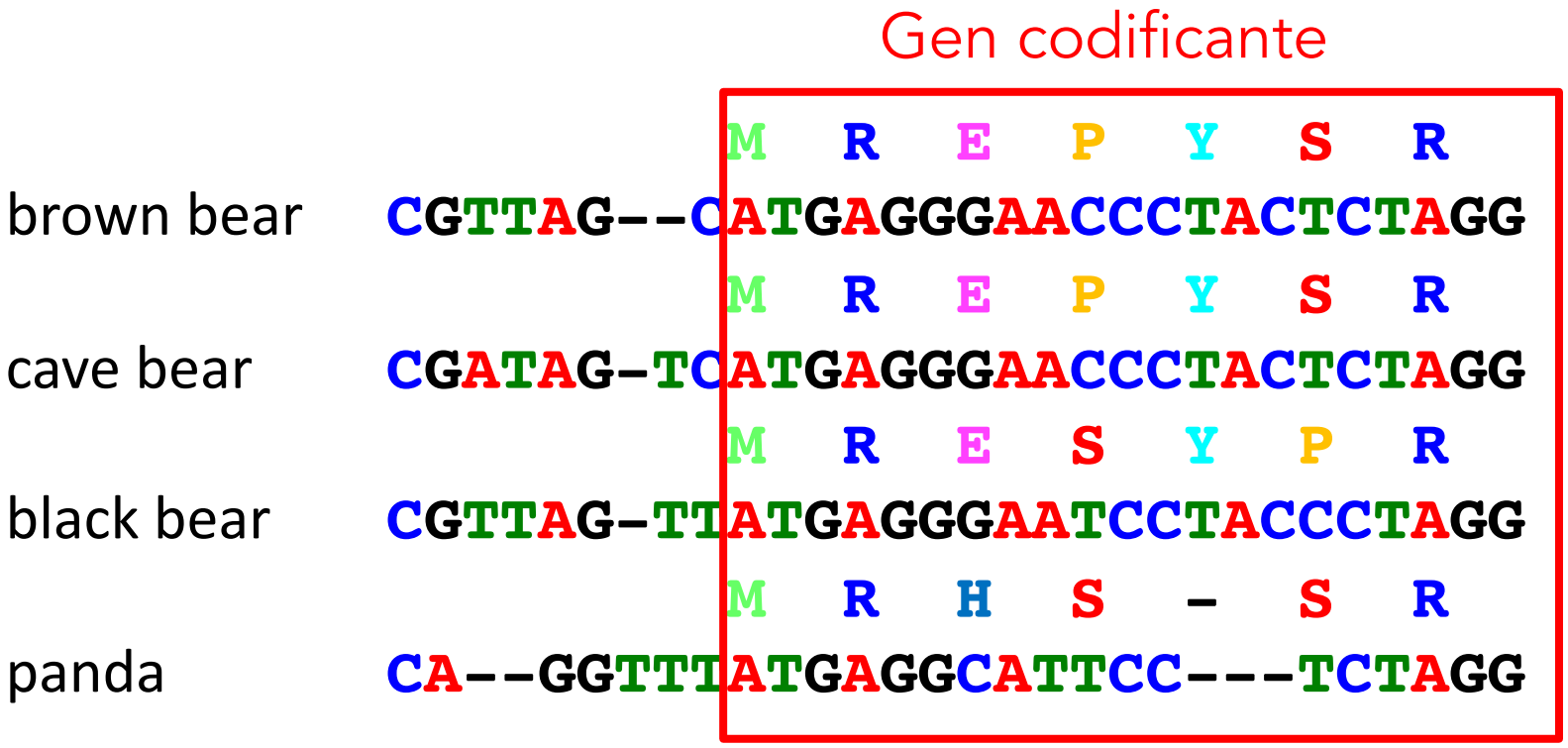
Types of data

- **Sequences**
  - Nucleotides
  - Amino acids
- **Binary data** (presence absence of genomic features)
- **Microsatelites** (number of repeats)
- **Single Nucleotide Polimorphisms** (SNP)
- **Reduced representation sequences**

Sequence types

- Coding regions
  - Ribosomal RNA
  - Protein coding
- Non-coding regiosn
  - Intergenic regions
  - Introns
- Amino acids

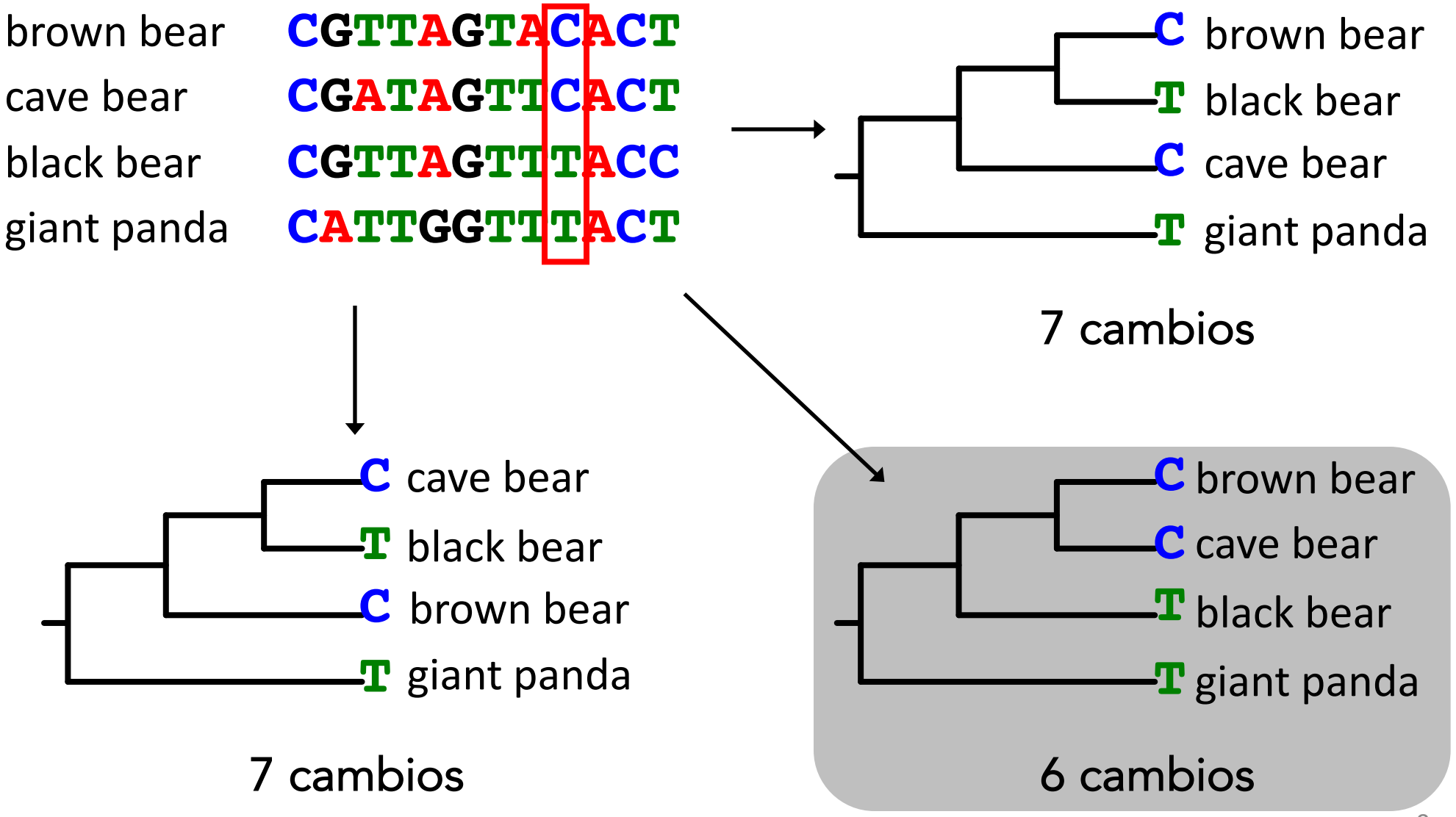
Sequence types



Phylogenetic methods

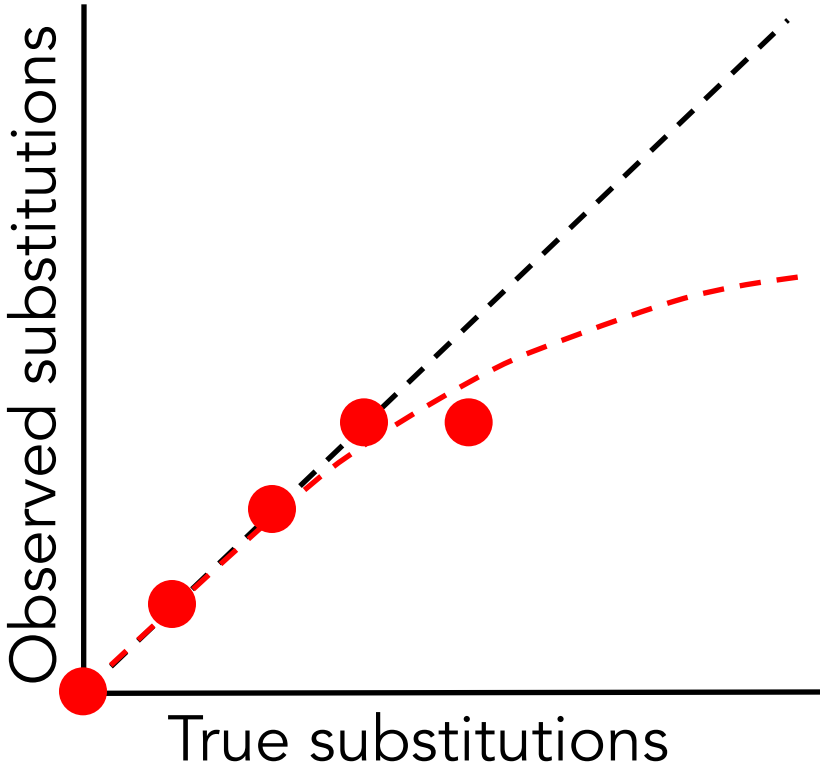
Maximum parsimony

Maximum parsimony



Maximum parsimony

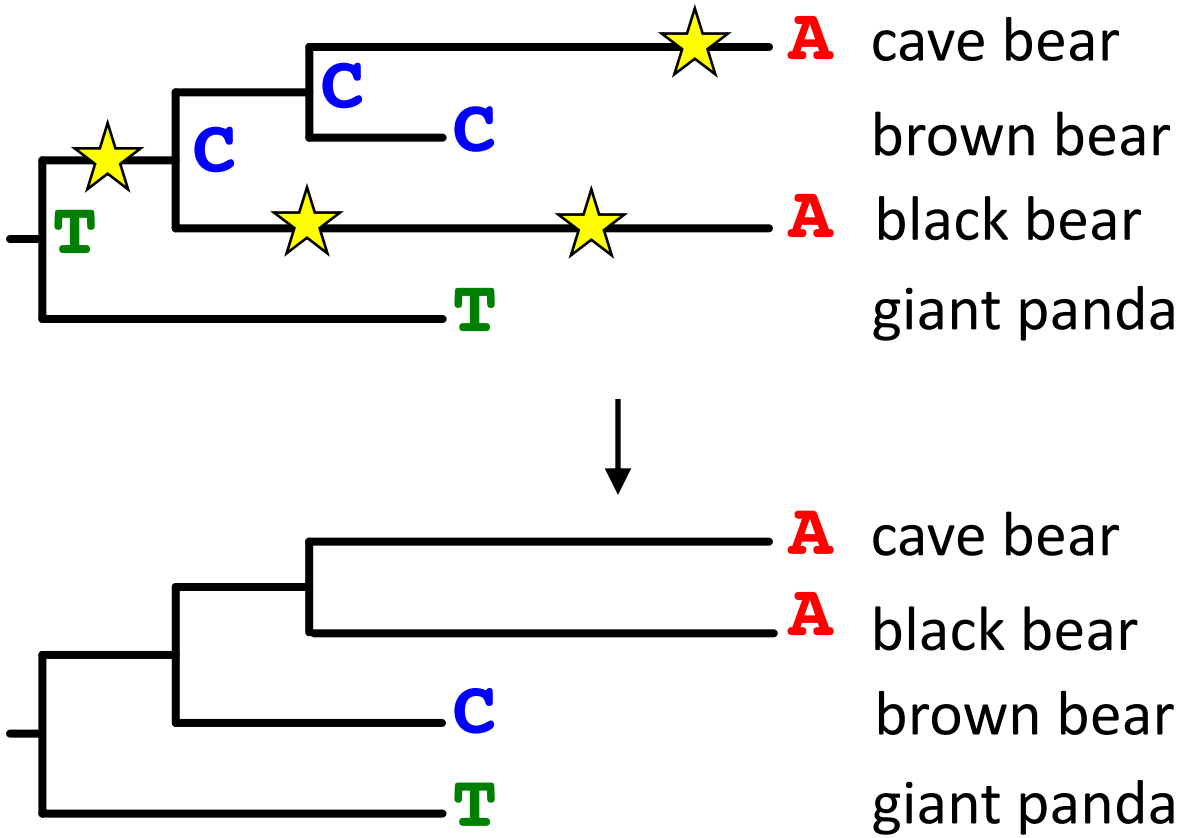
- Identifies the topology that explains the data with the minimum possible number of evolutionary changes
- Often use for analysis of morphological data
- Nowadays rarely used for analyses of molecular data
  - Does not allow estimation of molecular rates or times of divergence
  - Has undesired effects when there have been multiple molecular evolutionary events




```
A  A  A  A  A
A  T  T  T  T
C  C  G  G  G
A  A  A  A  A
T  T  T  T  T
T  T  T  T  T
A  A  A  A  A
G  G  G  G  G
T  T  T  A  C
```

- Maximum parsimony does not take into account multiple evolutionary events at one site
- This leads to a problem called **long branch attraction**
  - Long branches = multiple molecular substitutions
  - Similarities (homoplasy) emerge stochastically
  - Long branches are grouped

Long branch attraction



We can use statistical models to correct for multiple events

Popular methods in phylogenetics

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

Statistical methods

Maximum likelihood

Phylogenetic likelihood

Probability	Model
Tree 1	0.1
Tree 2	0.7
Tree 3	0.15
Tree 4	0.05
Sum	1

Phylogenetic likelihood

Probability	Model
Tree 1	0.1
Tree 2	0.7
Tree 3	0.15
Tree 4	0.05
Sum	1

A mathematical function gives us the probability of each tree:

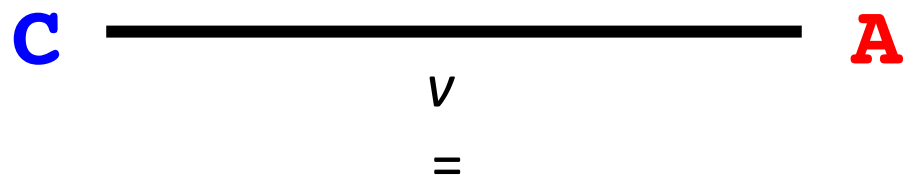
The phylogenetic likelihood function

Phylogenetic likelihood

- A molecular substitution is a stochastic event

Phylogenetic likelihood

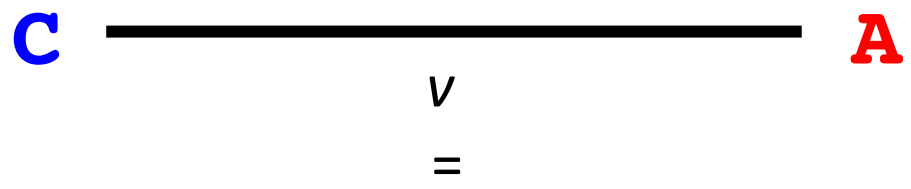
- A molecular substitution is a stochastic event
  - **We are interested in the probability of transition**



Hypothesis on the number of changes

Phylogenetic likelihood

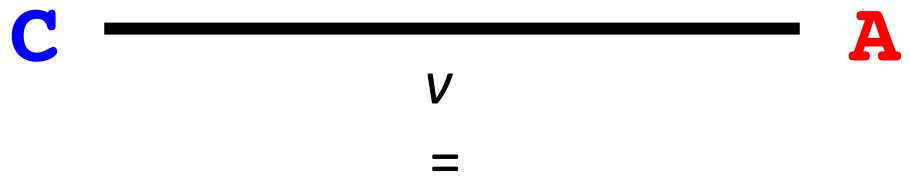
- A molecular substitution is a stochastic event
  - **We are interested in the probability of transition**



- The **Poisson Distribution** describes discrete stochastic events

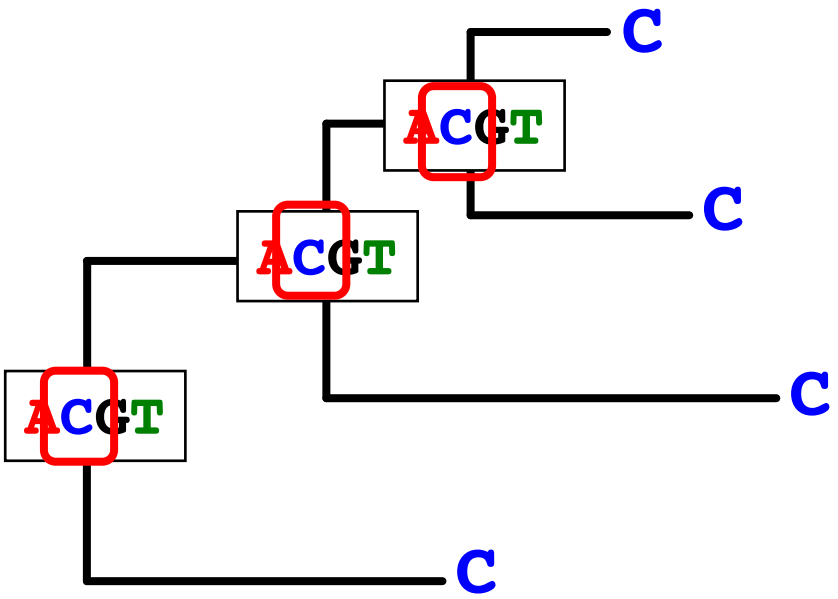
Phylogenetic likelihood

- A molecular substitution is a stochastic event
  - **We are interested in the probability of transition**



- The **Poisson Distribution** describes discrete stochastic events
  - The transition probability is given by the equation: e^{Qv}

The likelihood of a hypothesis

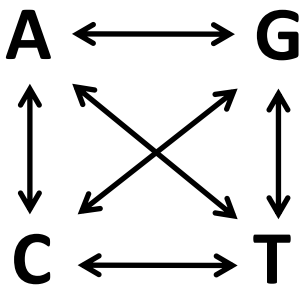


cave bear CGTTAGTACACT

black bear CGATAGTTCACT

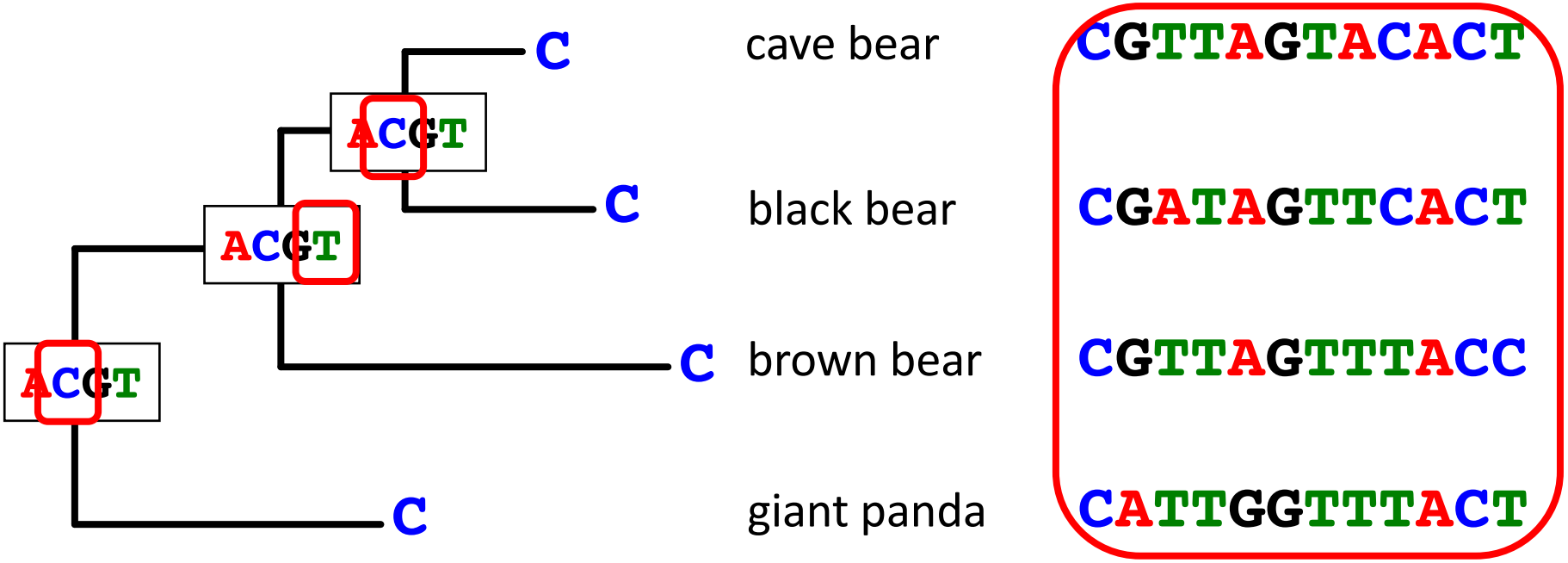
brown bear CGTTAGTTTACC

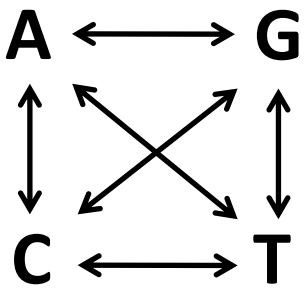
giant panda CATTGGTTTACT



The likelihood is multiplied across all sites

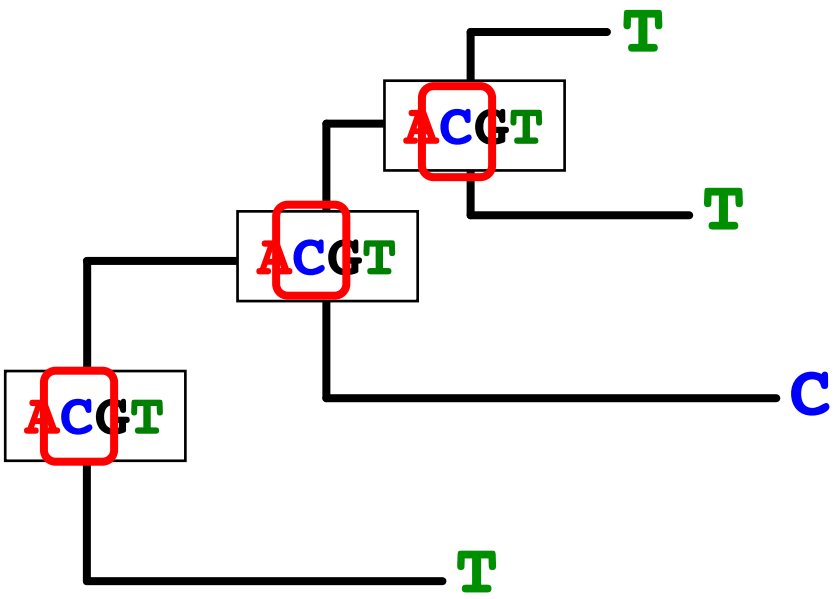
The likelihood of a hypothesis





The likelihood is multiplied across all sites

The likelihood of a hypothesis



cave bear

black bear

brown bear

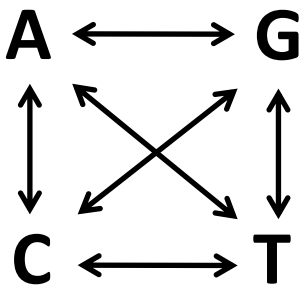
giant panda

CGTTAGTACACT

CGATAGTTCACT

CGTTAGTTTACC

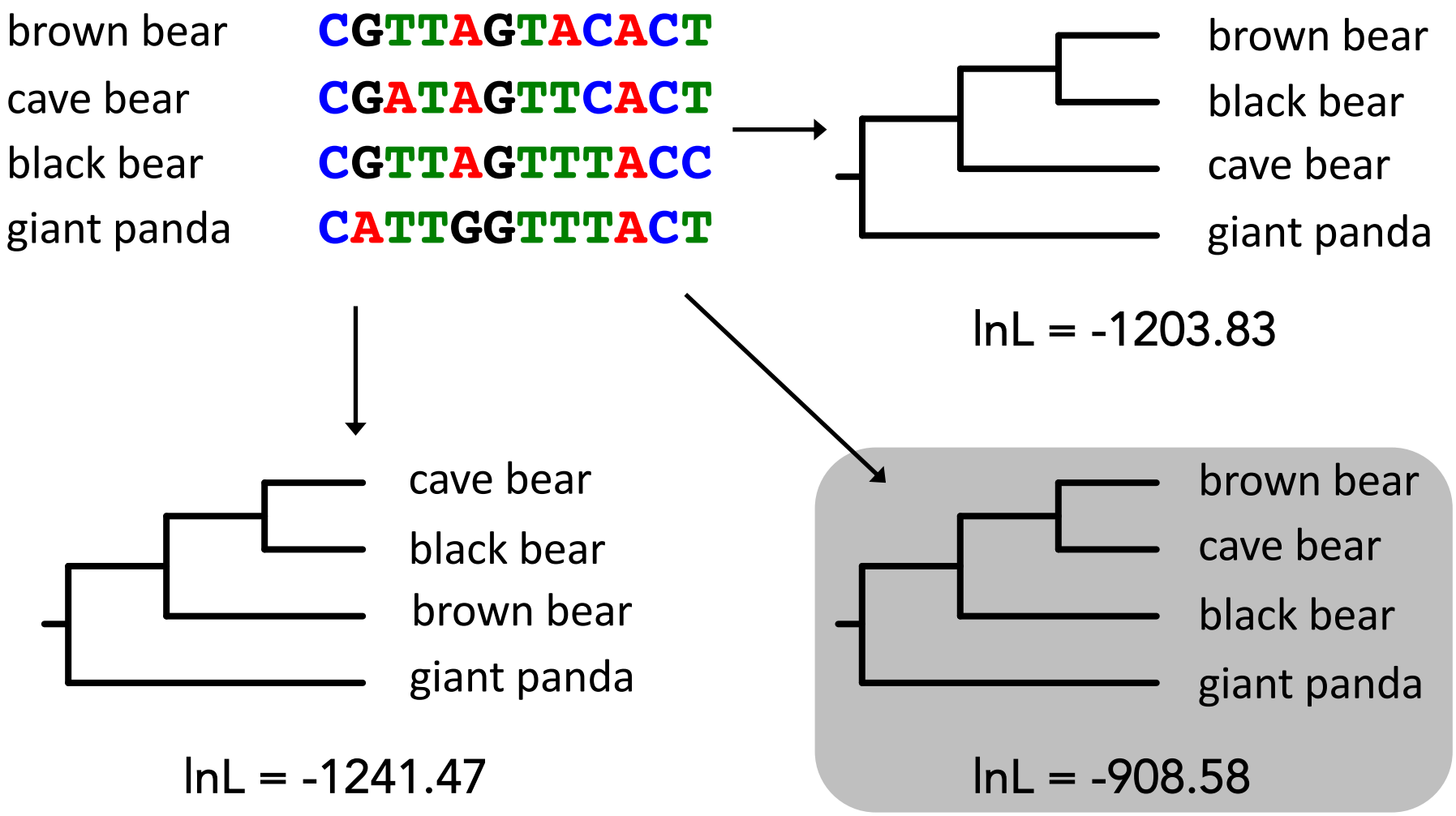
CATTGGTTTACT



The likelihood is multiplied across all sites

There is a low probability of observing any particular alignment

Maximum likelihood



Likelihood optimization

- Search the space of possible trees and parameters
- Calculate the likelihood of each
- Find the case with the maximum likelihood
- Optimize multiple variables

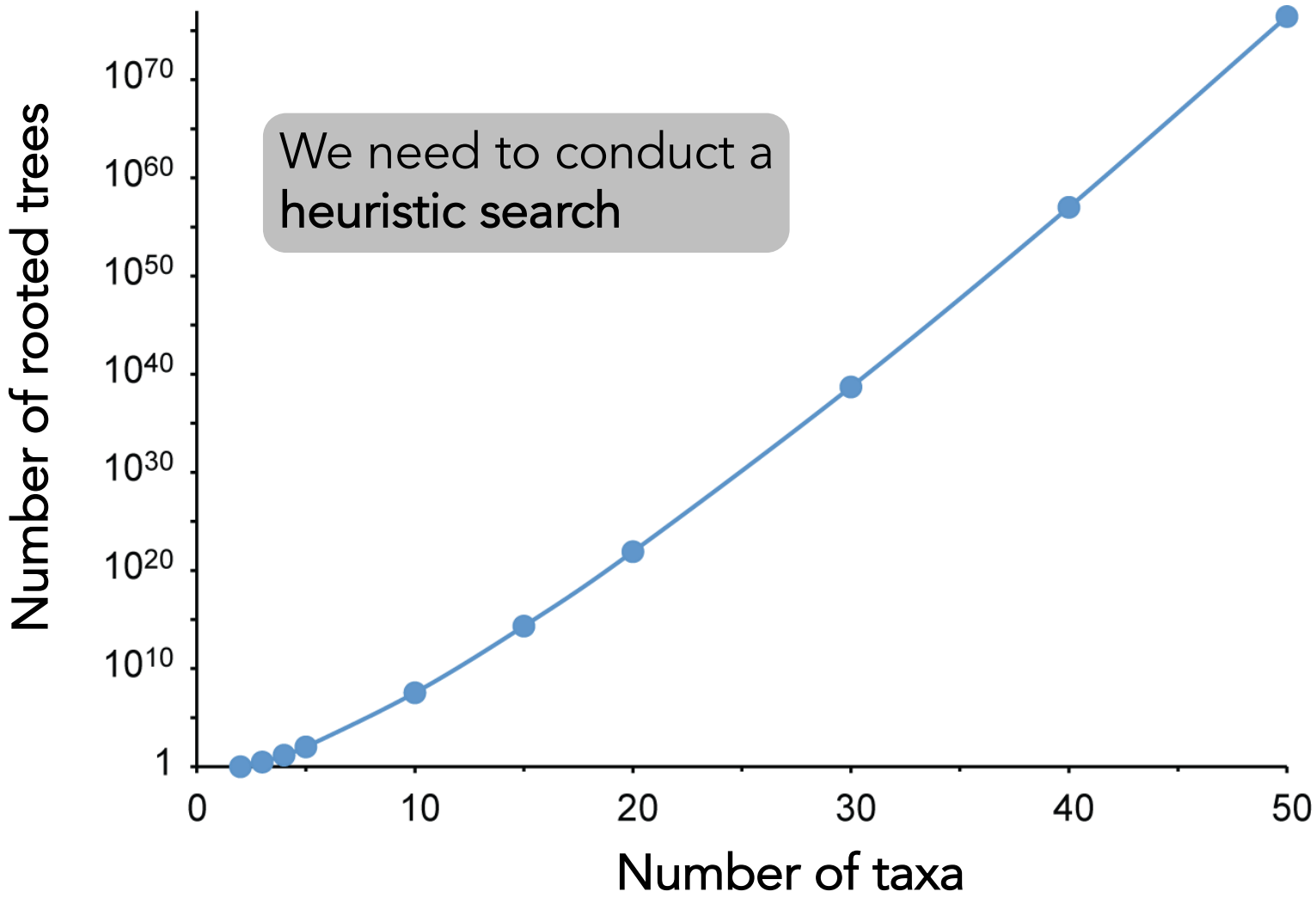
How to find the best tree

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

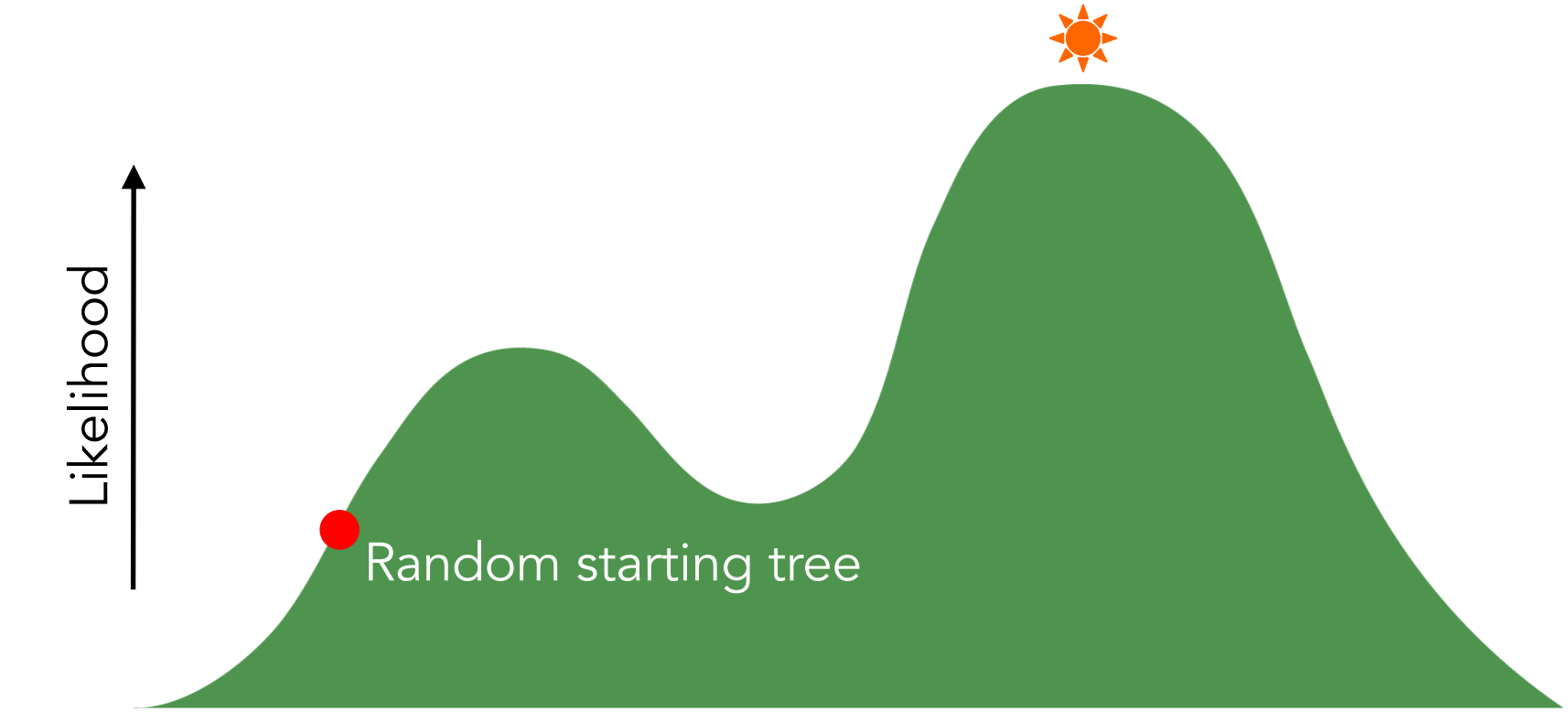
$$B_n = 1 \times 3 \times 5 \times \ldots \times (2n-5) = \prod_{i=3}^{n} (2i-5)$$

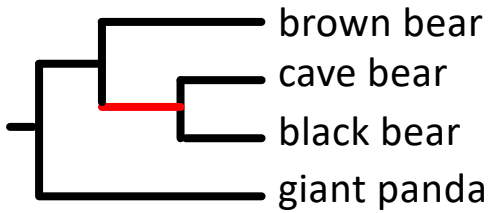
- For example:
  - 4 taxa → 3 trees
  - 5 taxa → 15 trees
  - 10 taxa → 2,027,025 trees

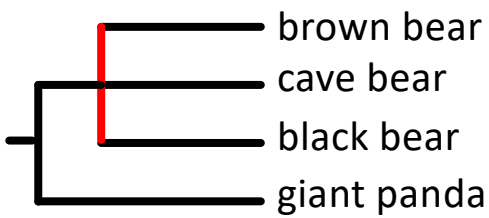
How to find the best tree

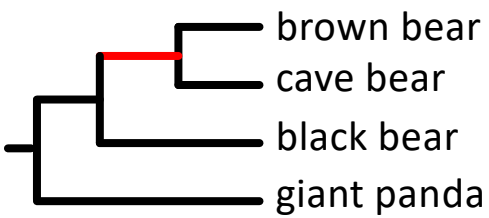


Heuristic search

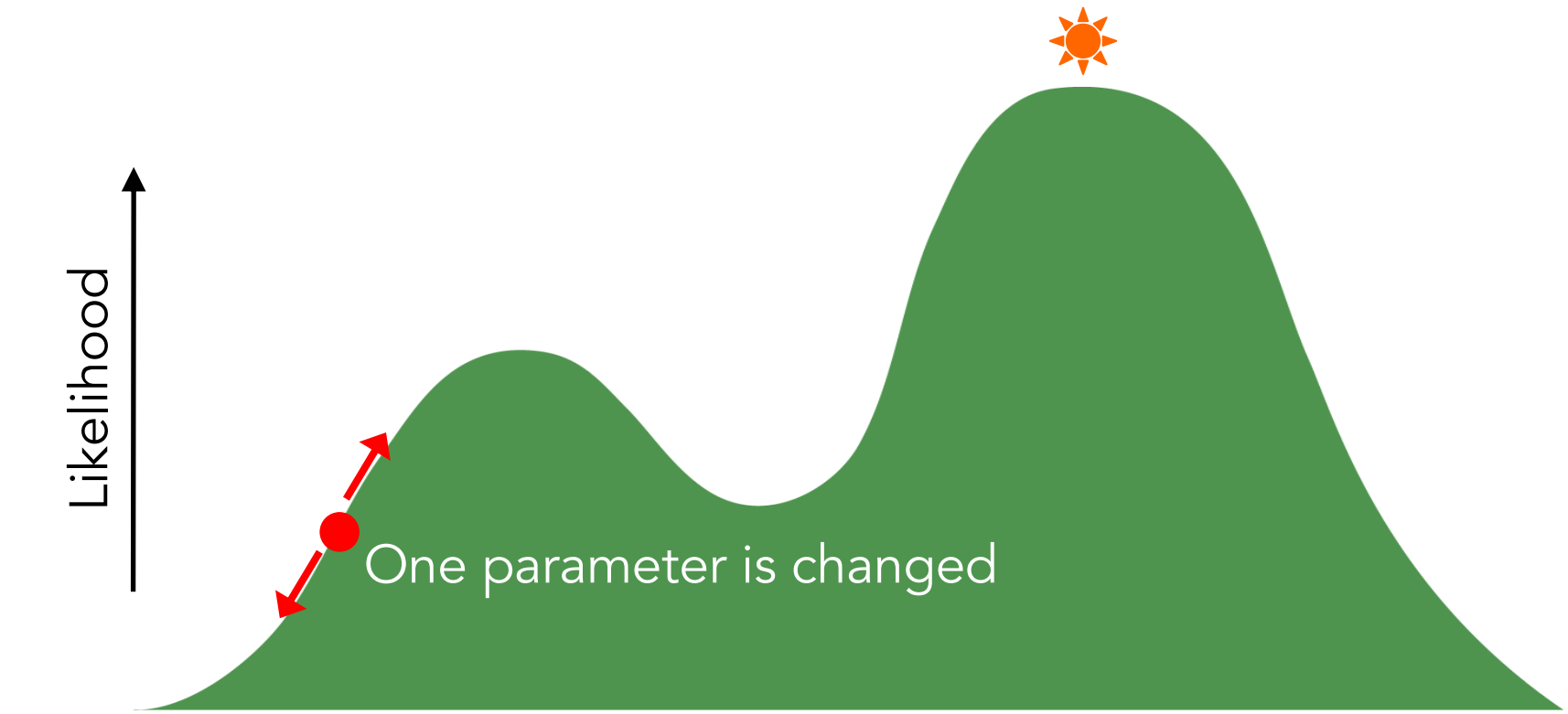


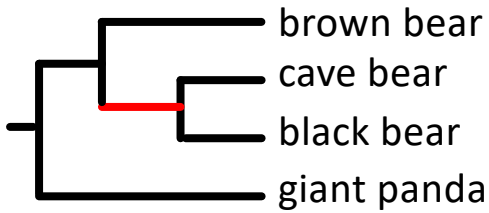


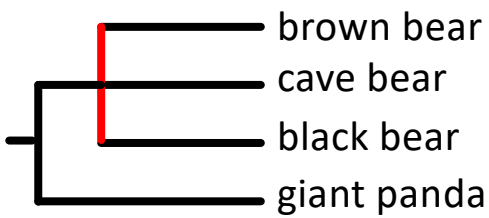


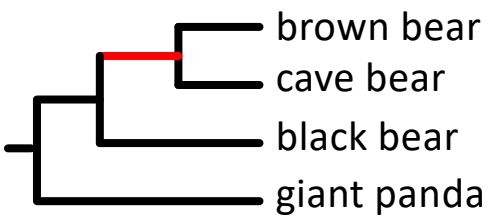


Heuristic search

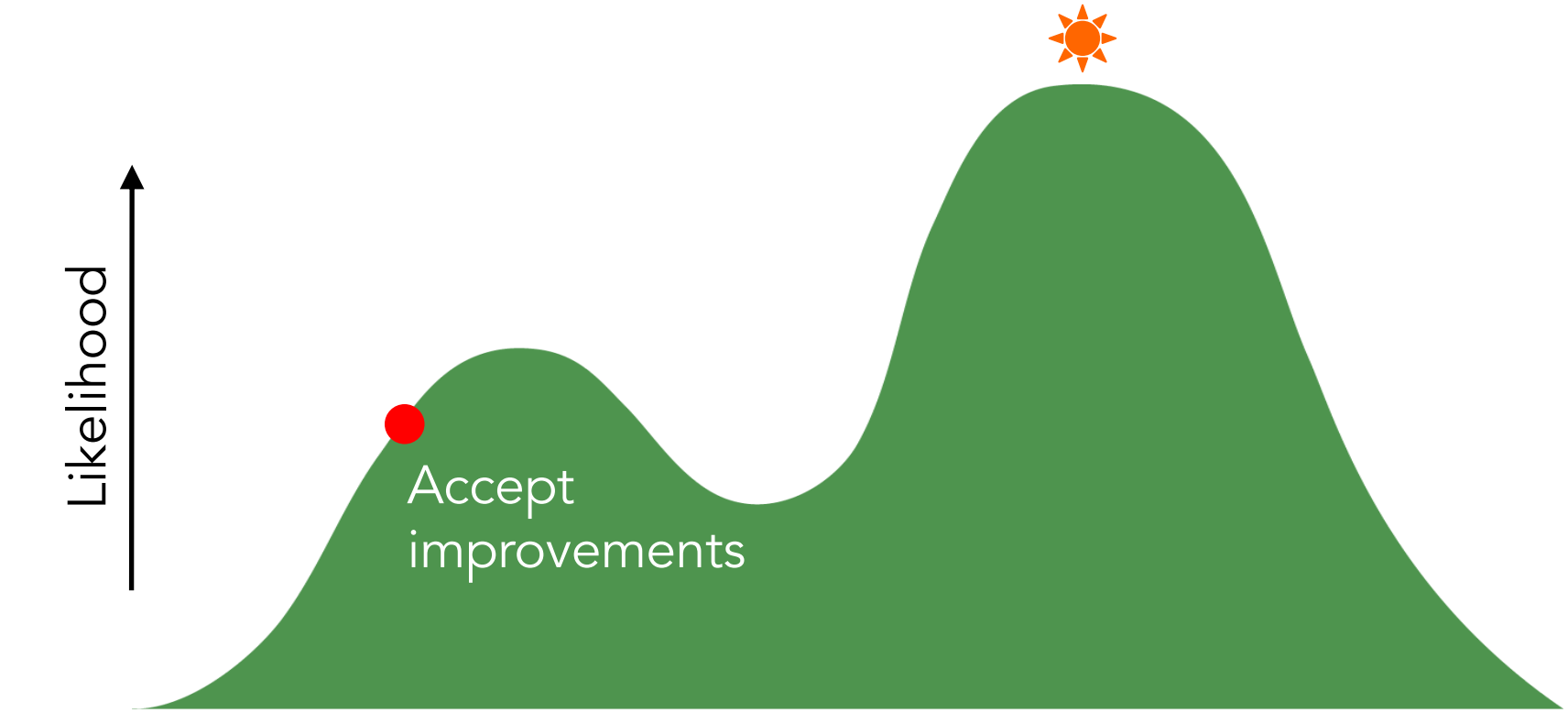


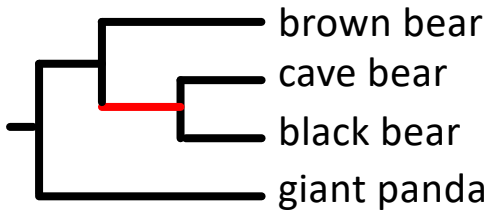


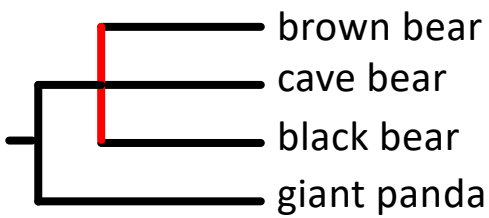


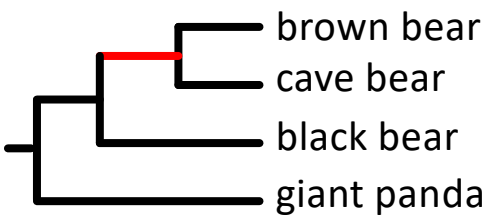


Heuristic search

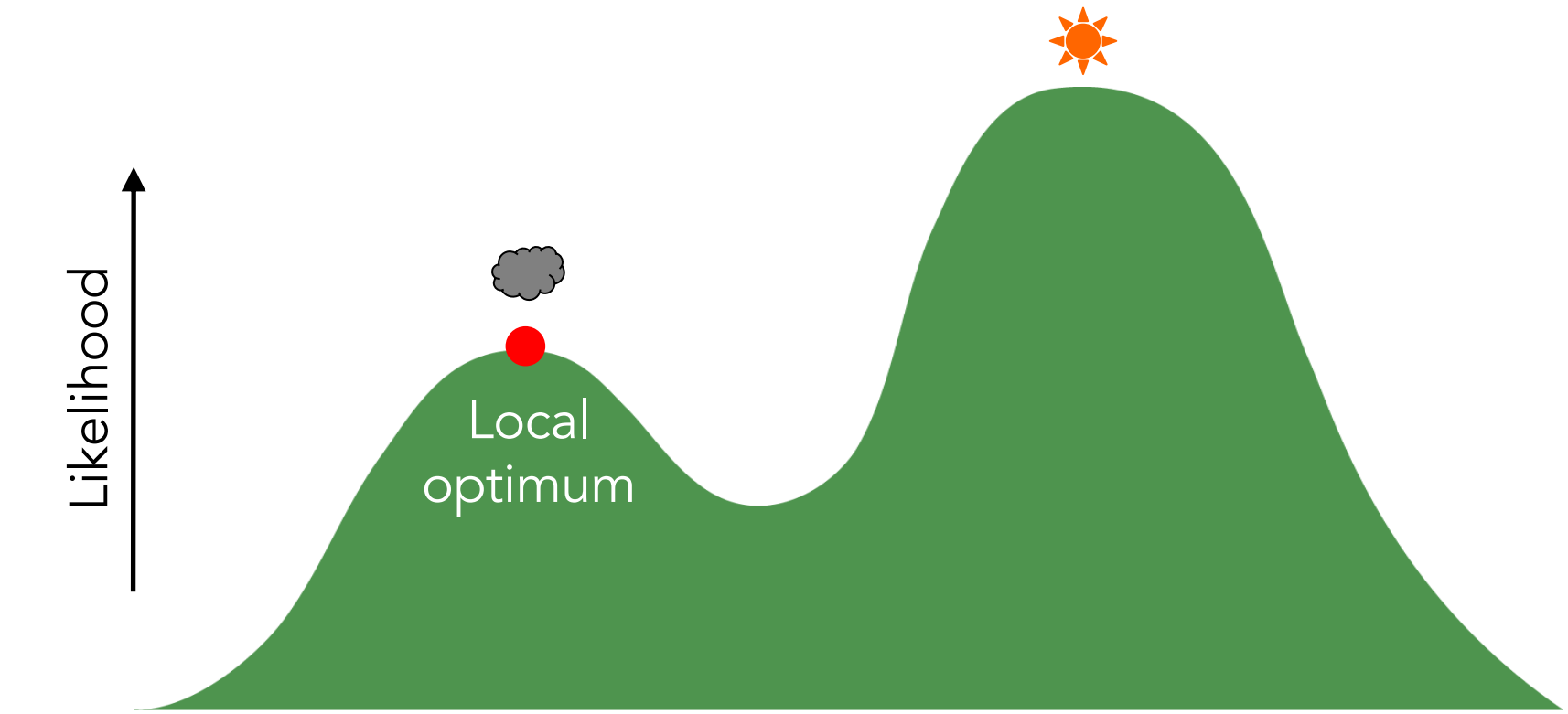


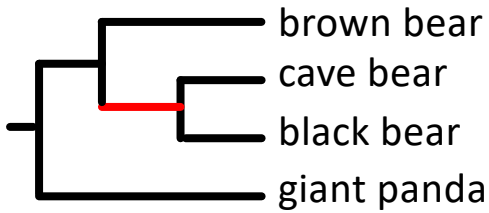


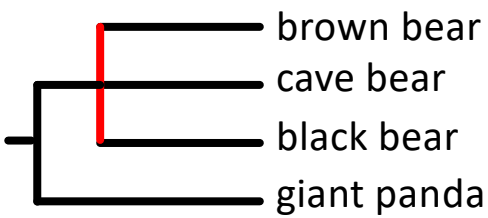


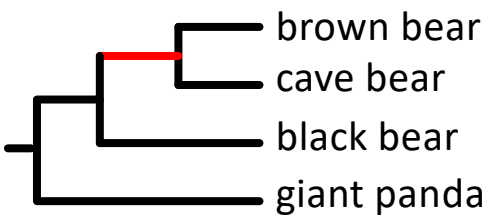


Heuristic search

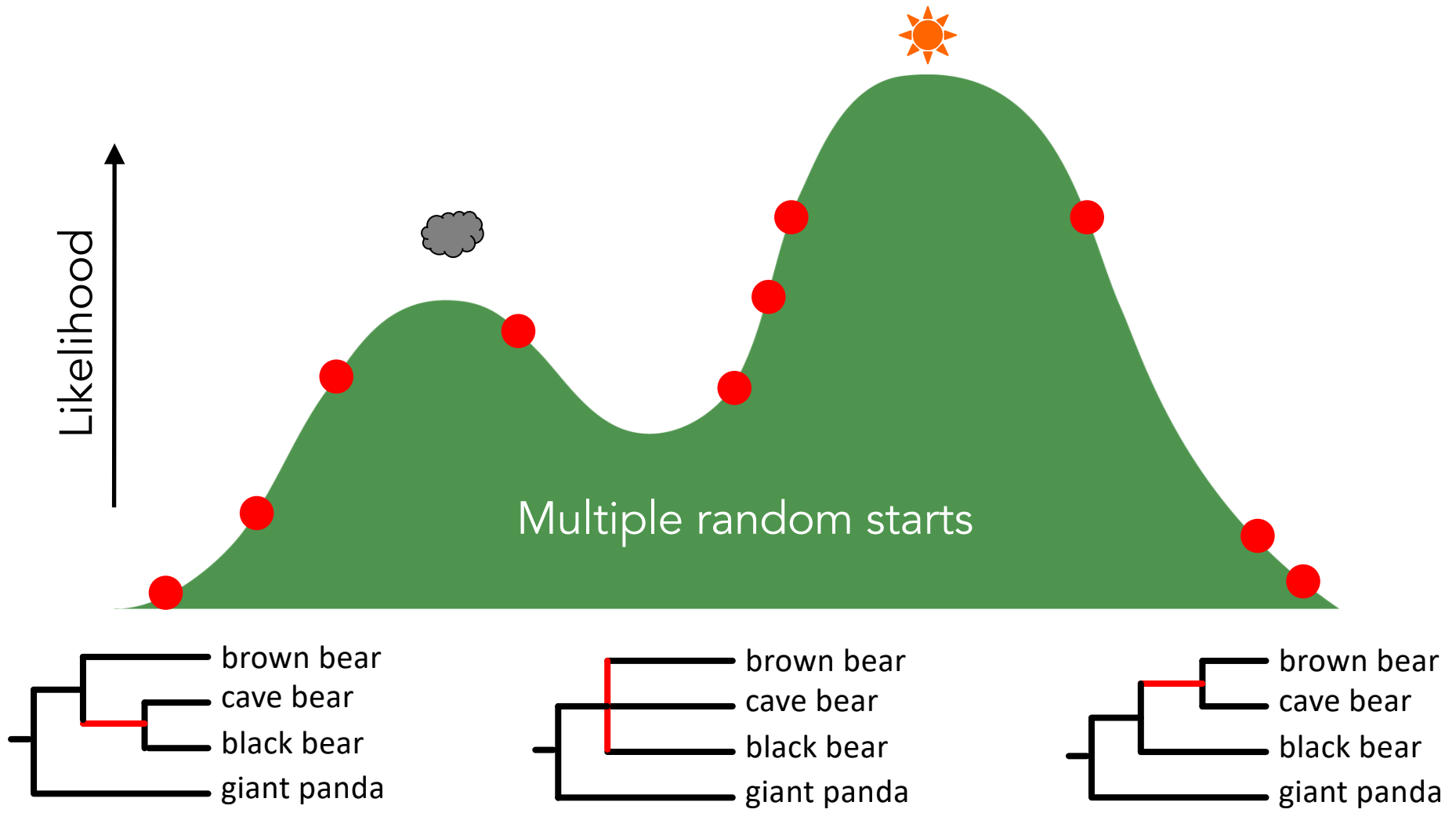






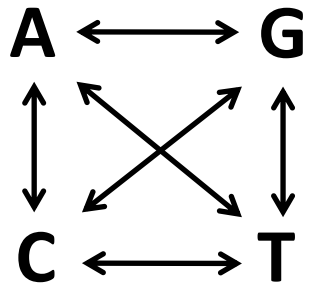


Heuristic search

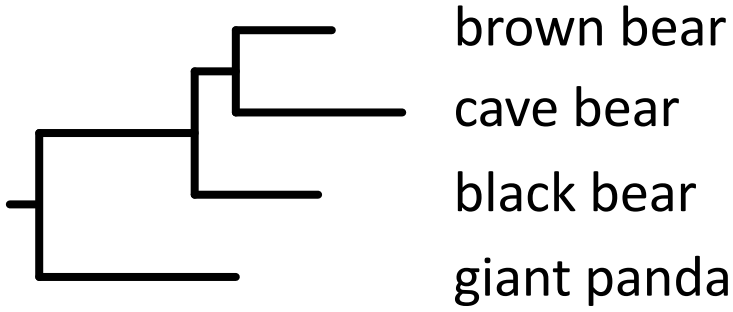


Maximum likelihood estimates

Parameters of a substitution model



A phylogenetic tree with branch lengths



Why the term *likelihood*?

Likelihoods are attributes of models

Probability	Model 1
Tree 1	0.1
Tree 2	0.7
Tree 3	0.15
Tree 4	0.05
Sum	1

Why the term *likelihood*?

Likelihoods are attributes of models

Probability	Model 1	Modelo 2	Modelo 3
Tree 1	0.1	0.2	0.05
Tree 2	0.7	0.29	0.35
Tree 3	0.15	0.5	0.4
Tree 4	0.05	0.01	0.2
Sum	1	1	1

Why the term *likelihood*?

Likelihoods are attributes of models
They don't sum to 1 across models

Probability	Model 1	Modelo 2	Modelo 3
Tree 1	0.1	0.2	0.05
Tree 2	0.7	0.29	0.35
Tree 3	0.15	0.5	0.4
Tree 4	0.05	0.01	0.2
Sum	1	1	1

$P(D|H)$

Why the term *likelihood*?

Likelihoods are attributes of models
They don't sum to 1 across models

Probability	Model 1	Modelo 2	Modelo 3
Tree 1	0.1	0.2	0.05
Tree 2	0.7	0.29	0.35
Tree 3	0.15	0.5	0.4
Tree 4	0.05	0.01	0.2
Sum	1	1	1

Probability is an attribute
of the data
Sums to 1 within a model

Strengths and weaknesses

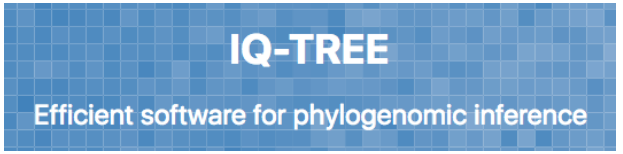
- **Strengths**
  - It is a rigorous statistical method
  - Can largely correct for multiple substitutions and long branches
  - Robust to violation of assumptions
- **Weaknesses**
  - Difficult to use when the model has many parameters
  - Can be difficult to explore the space of possible trees

Software

RAxML

PAML

PhyML



IQ-TREE

MEGA

Phylogenetic methods in practice

- **Maximum parsimony**
  - Often used for analyses of morphological data
  - Rarely used for analyses of molecular data
- **Maximum likelihood**
  - Widely used but partially replaced by Bayesian inference methods