Introduction to phylogenetics

Thibaut Jombart

MRC Centre for Outbreak Analysis and Modelling

4th May 2011

Outline

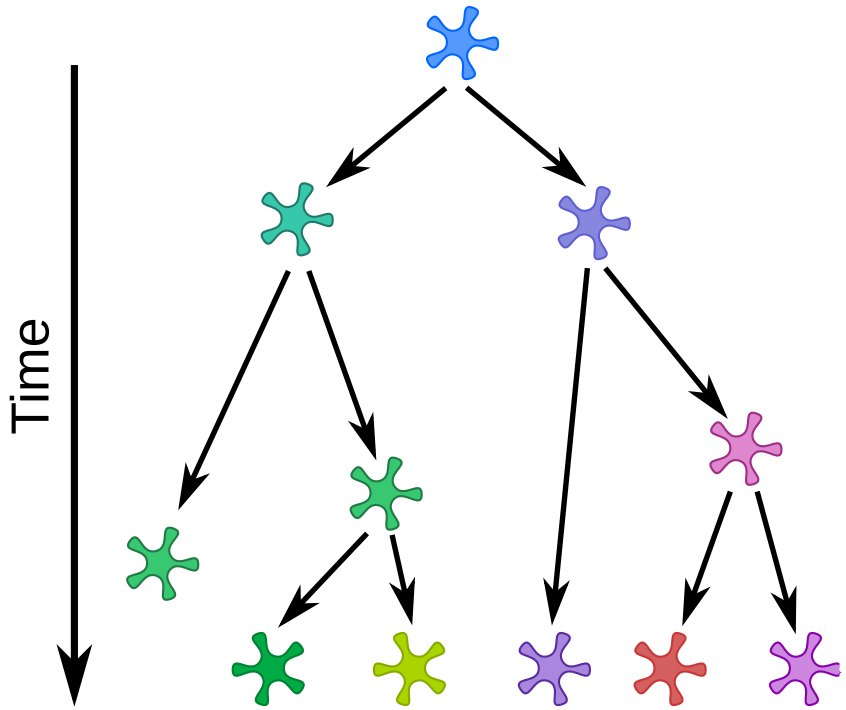
Outline

Reconstructing evolutionary history

Phylogenetic tree: representation of evolutionary relationships between a set of taxa (species, individuals, etc.)

Reconstructing evolutionary history

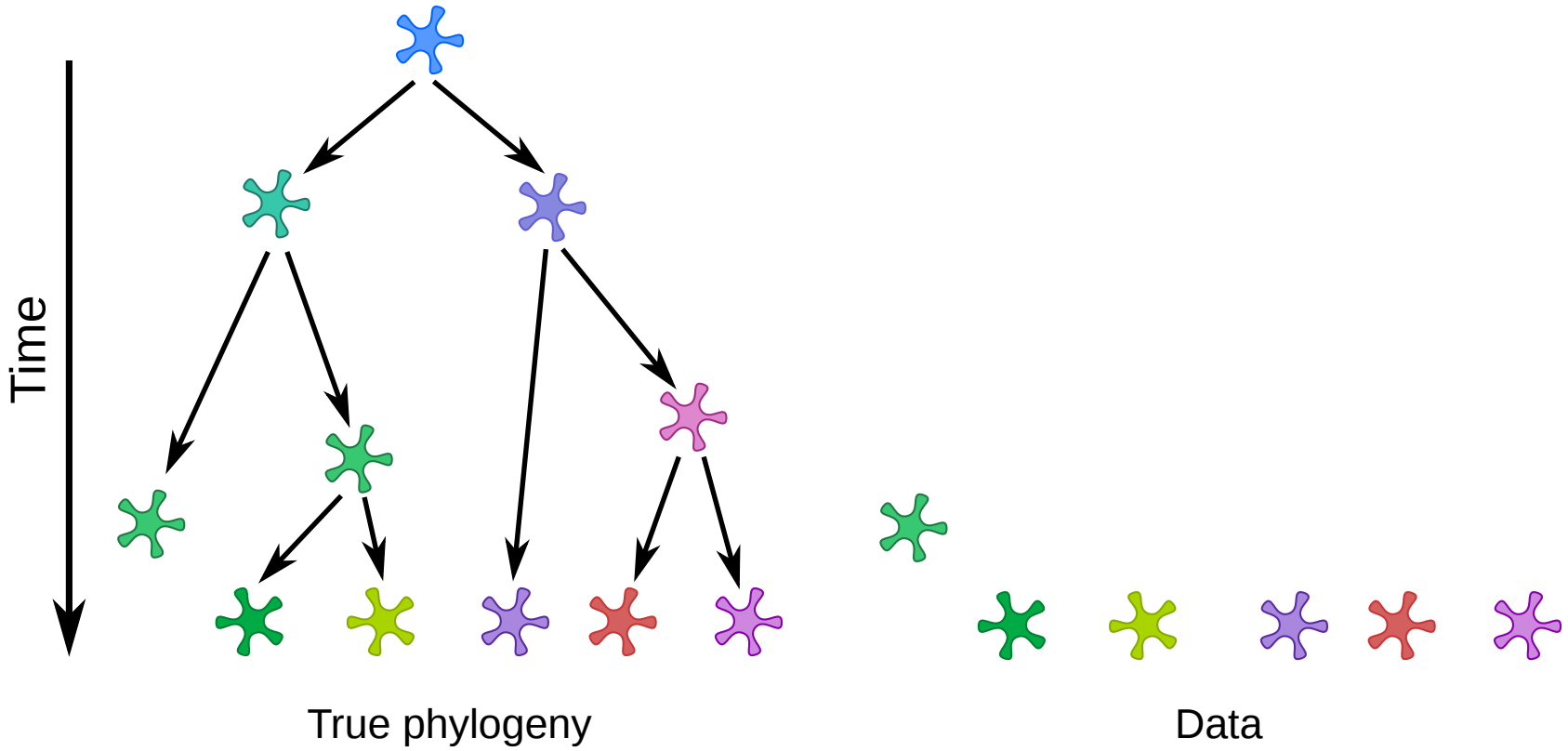
Phylogenetic tree: representation of evolutionary relationships between a set of taxa (species, individuals, etc.)



True phylogeny

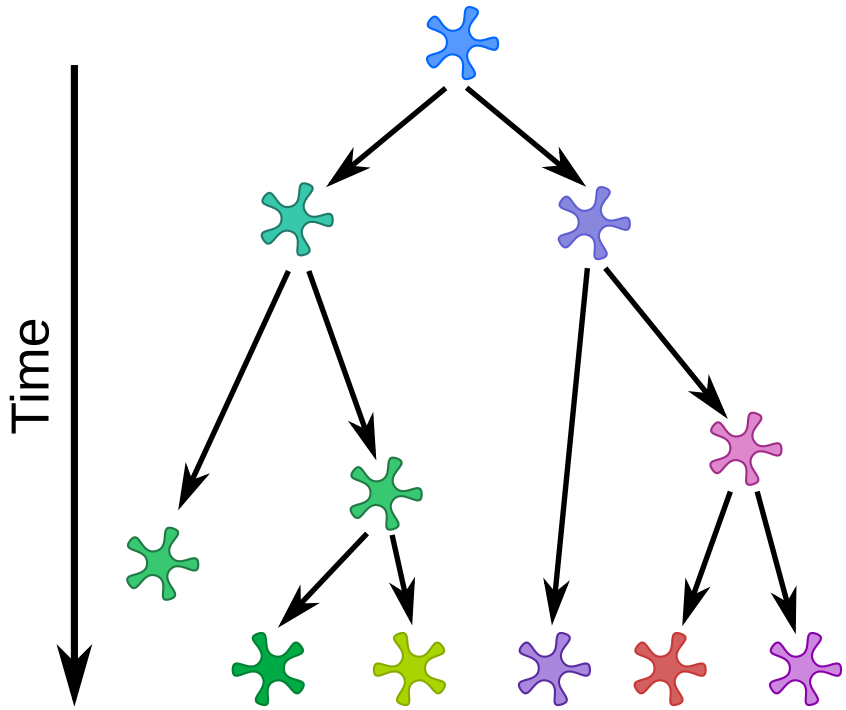
Reconstructing evolutionary history

Phylogenetic tree: representation of evolutionary relationships between a set of taxa (species, individuals, etc.)

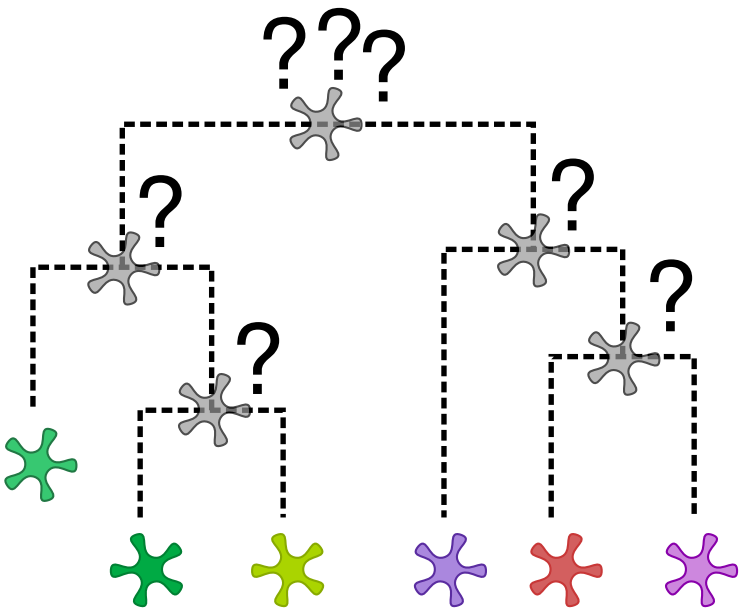


Reconstructing evolutionary history

Phylogenetic tree: representation of evolutionary relationships between a set of taxa (species, individuals, etc.)



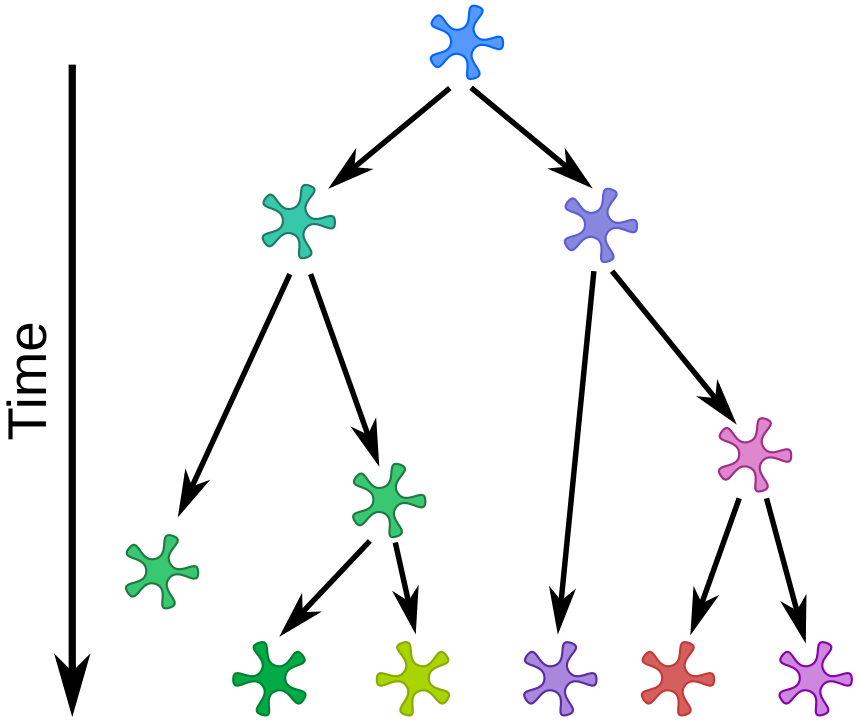
True phylogeny



Reconstructed phylogeny

Accumulation of substitutions

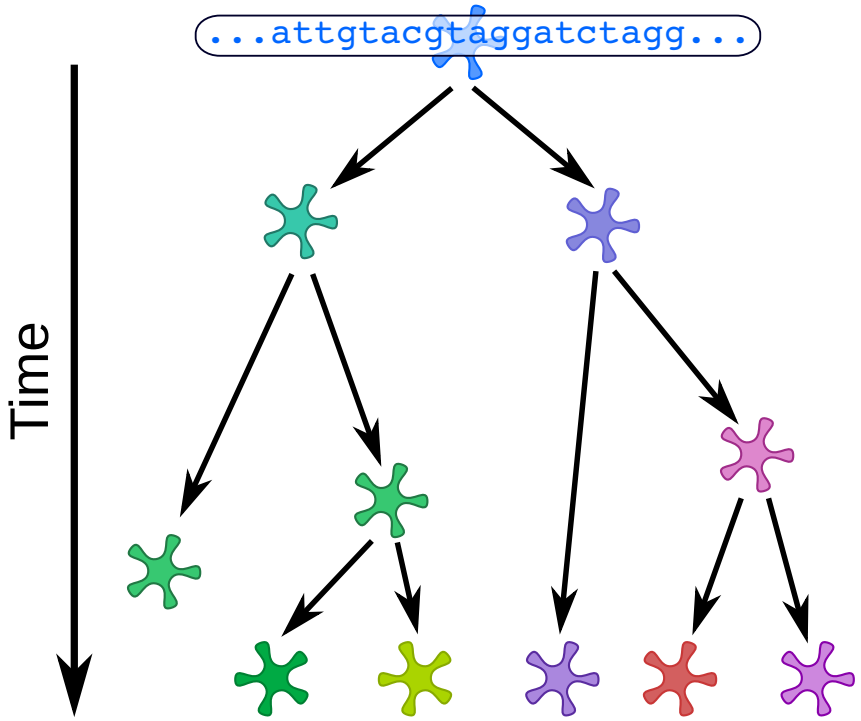
Substitution: replacement of a nucleotide (e.g. a $\rightarrow$ t)



True phylogeny

Accumulation of substitutions

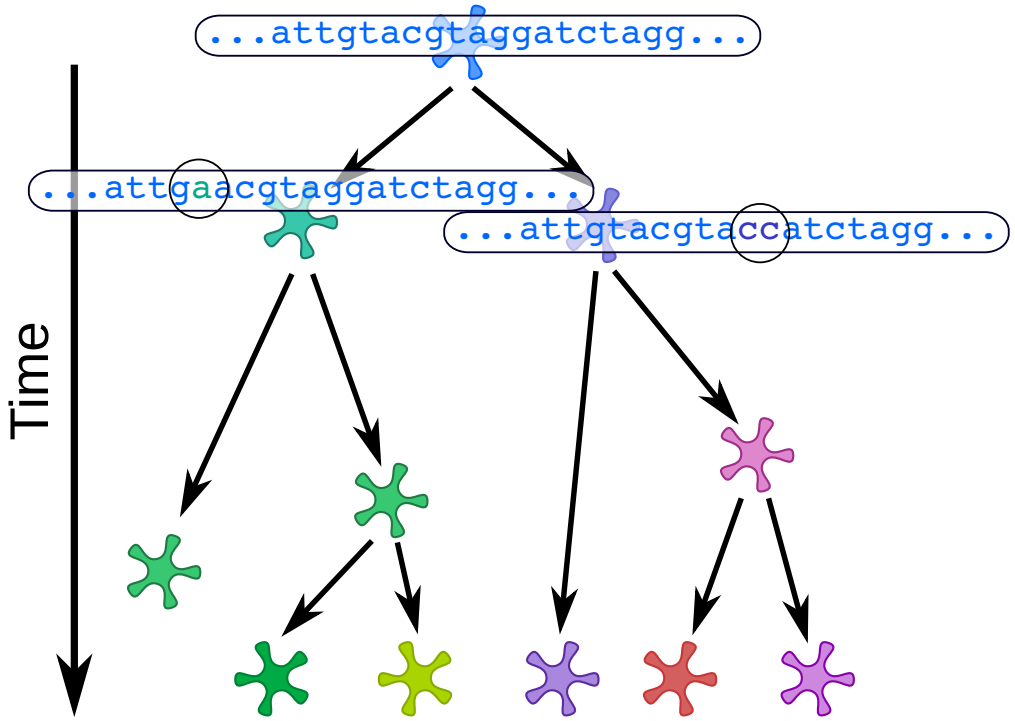
Substitution: replacement of a nucleotide (e.g. a $\rightarrow$ t)



True phylogeny

Accumulation of substitutions

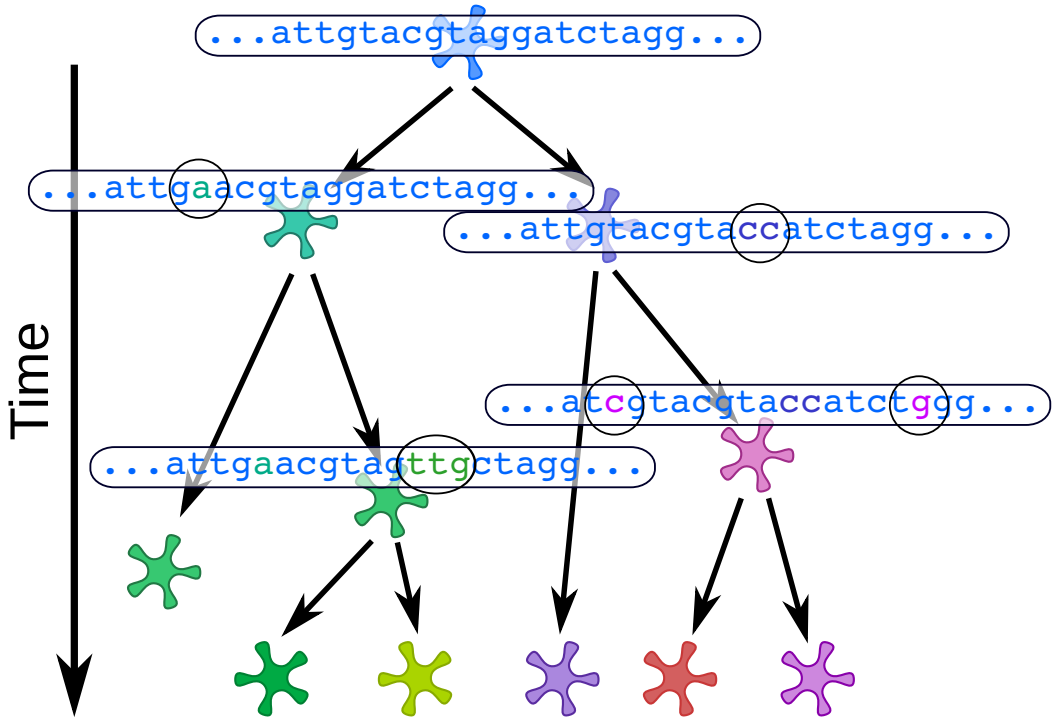
Substitution: replacement of a nucleotide (e.g. a $\rightarrow$ t)



True phylogeny

Accumulation of substitutions

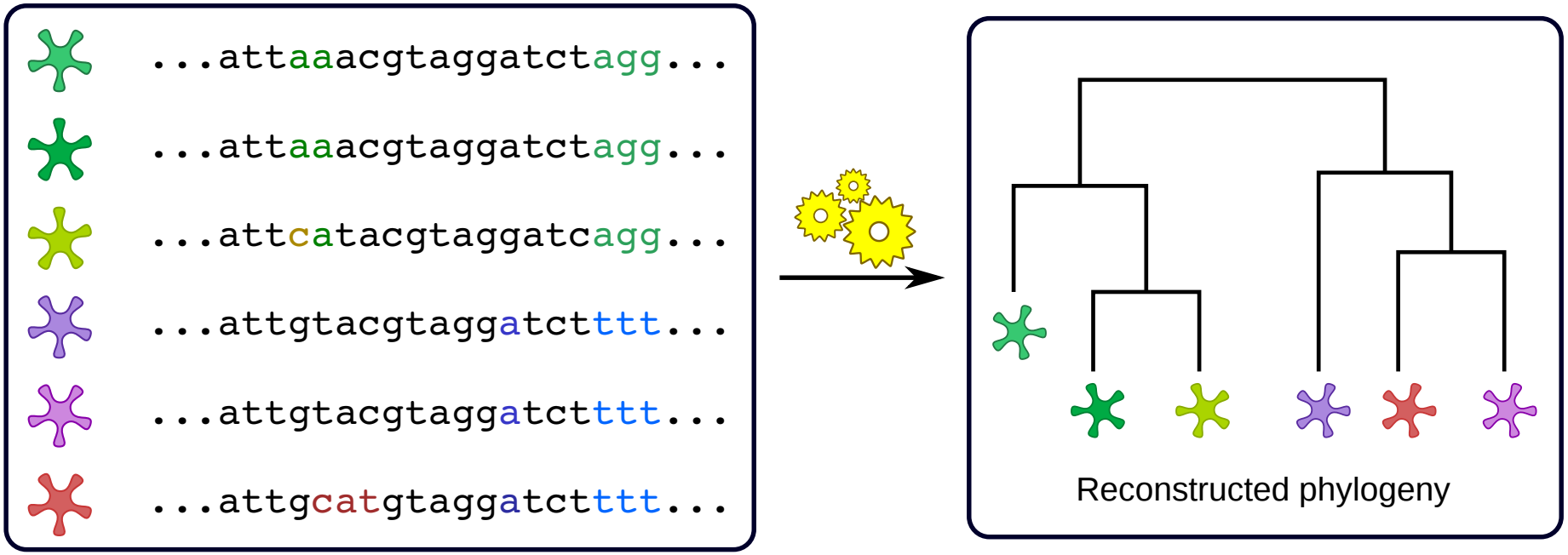
Substitution: replacement of a nucleotide (e.g. a $\rightarrow$ t)



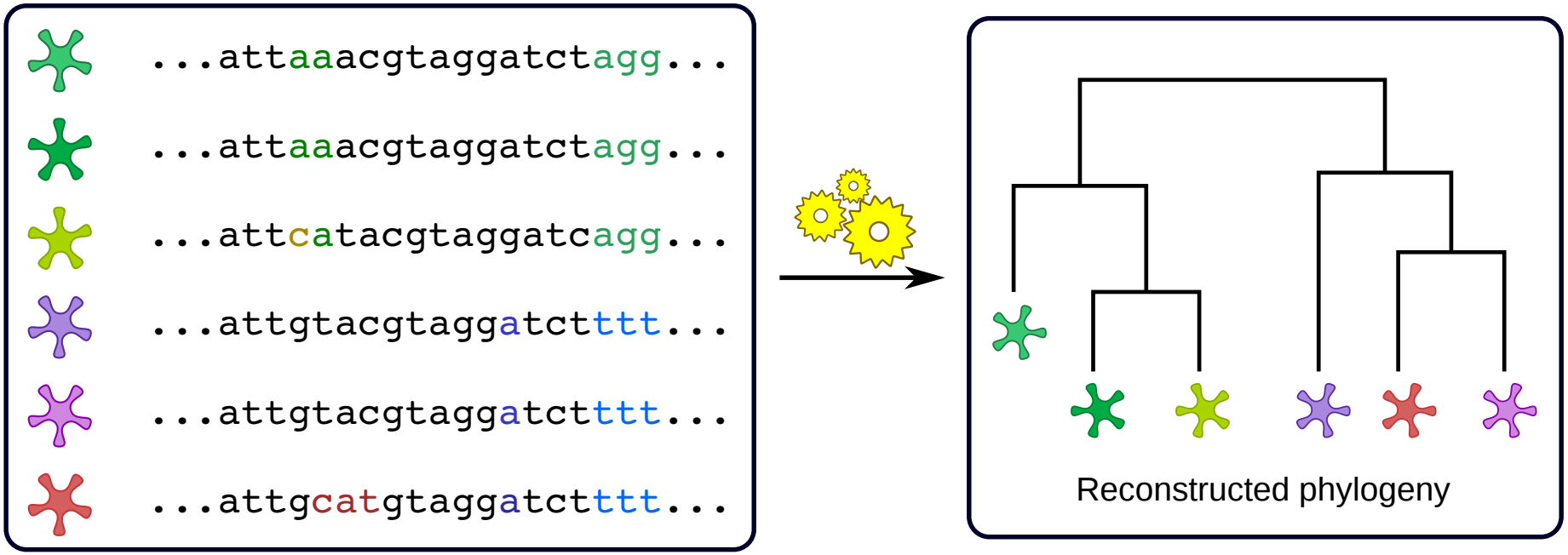
True phylogeny

From alignments to phylogenies

From alignments to phylogenies



From alignments to phylogenies



Different methods for achieving phylogenetic reconstruction.

Workflow

Prepare data

- select and retrieve sequences
- align sequences

Workflow

Prepare data

- select and retrieve sequences
- align sequences

Phylogenetic reconstruction

- distance-based methods
- maximum parsimony
- likelihood-based methods (ML, Bayesian)

Workflow

Prepare data

- select and retrieve sequences
- align sequences

Phylogenetic reconstruction

- distance-based methods
- maximum parsimony
- likelihood-based methods (ML, Bayesian)

Assess relevance of the tree

- bootstrap phylogeny
- for ML approaches: model selection

Outline

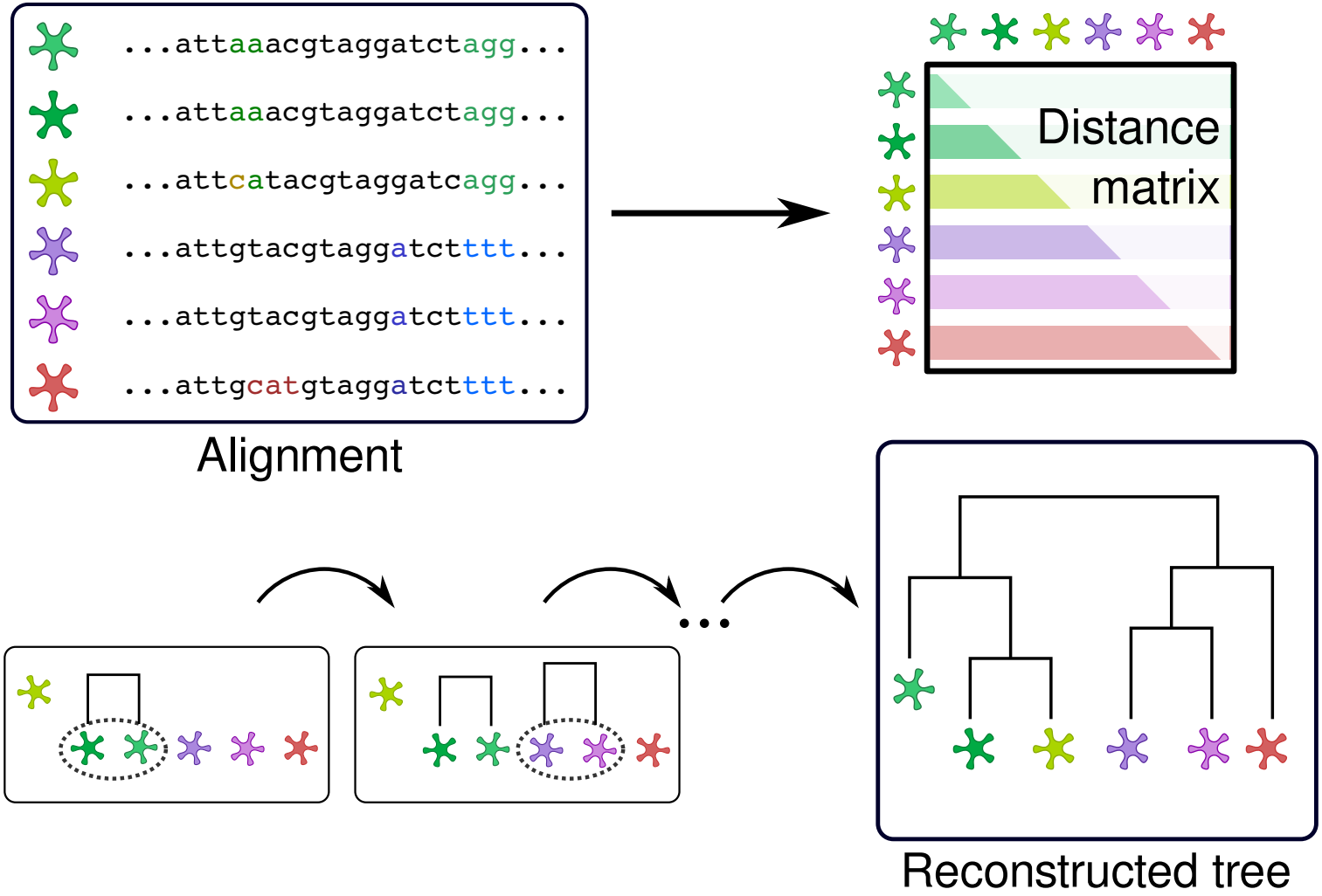
Distance-based phylogenetic reconstruction

Approaches relying on **hierarchical clustering** algorithms (e.g. Neighbor-Joining, UPGMA)

Rationale

1. compute pairwise genetic distances **D**
2. group closest sequences
3. update **D**
4. go back to 2) until all sequences are grouped

Distance-based phylogenetic reconstruction



Distance-based phylogenetic reconstruction

Advantages

- simple
- flexible (many distances and clustering algorithms)
- fast

Distance-based phylogenetic reconstruction

Advantages

- simple
- flexible (many distances and clustering algorithms)
- fast

Limitations

- sensitive to distance/clustering chosen
- no parameter estimation
- single tree produced

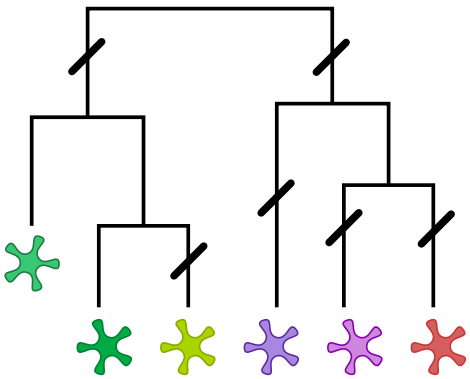
Maximum parsimony phylogenies

Approaches relying on finding the tree with the **smallest number of substitutions**

Rationale

1. start from a pre-defined tree
2. compute initial parsimony score
3. permute branches and compute parsimony score
4. accept new tree if the parsimony score is improved
5. go back to 3) until convergence

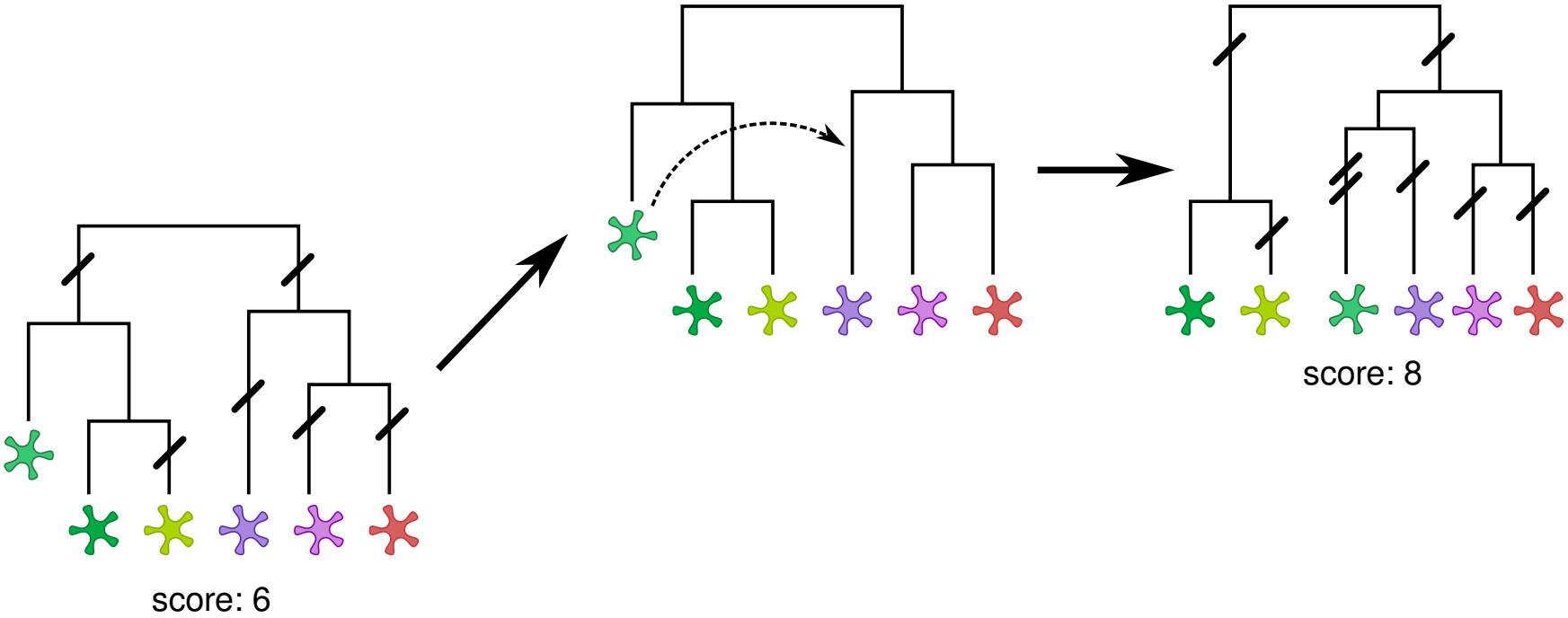
Maximum parsimony phylogenies



score: 6

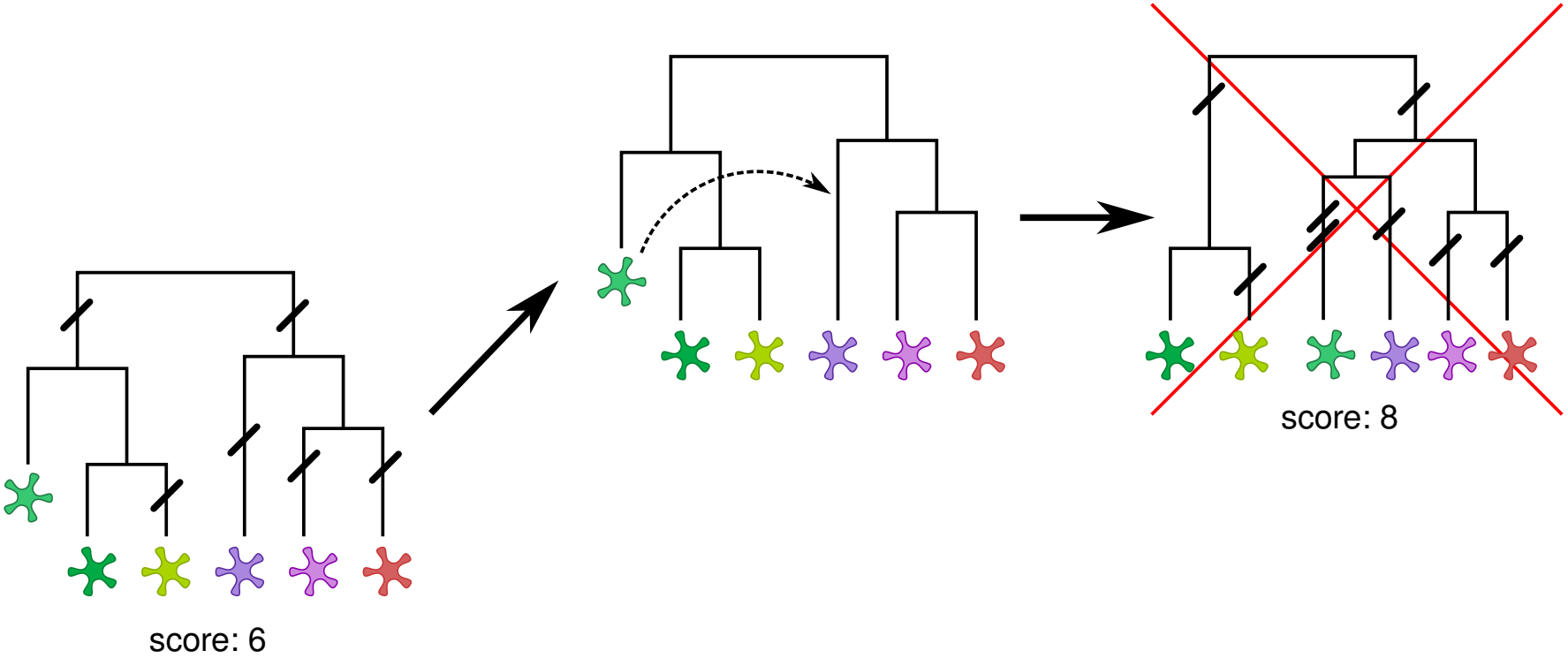
Initial tree

Maximum parsimony phylogenies



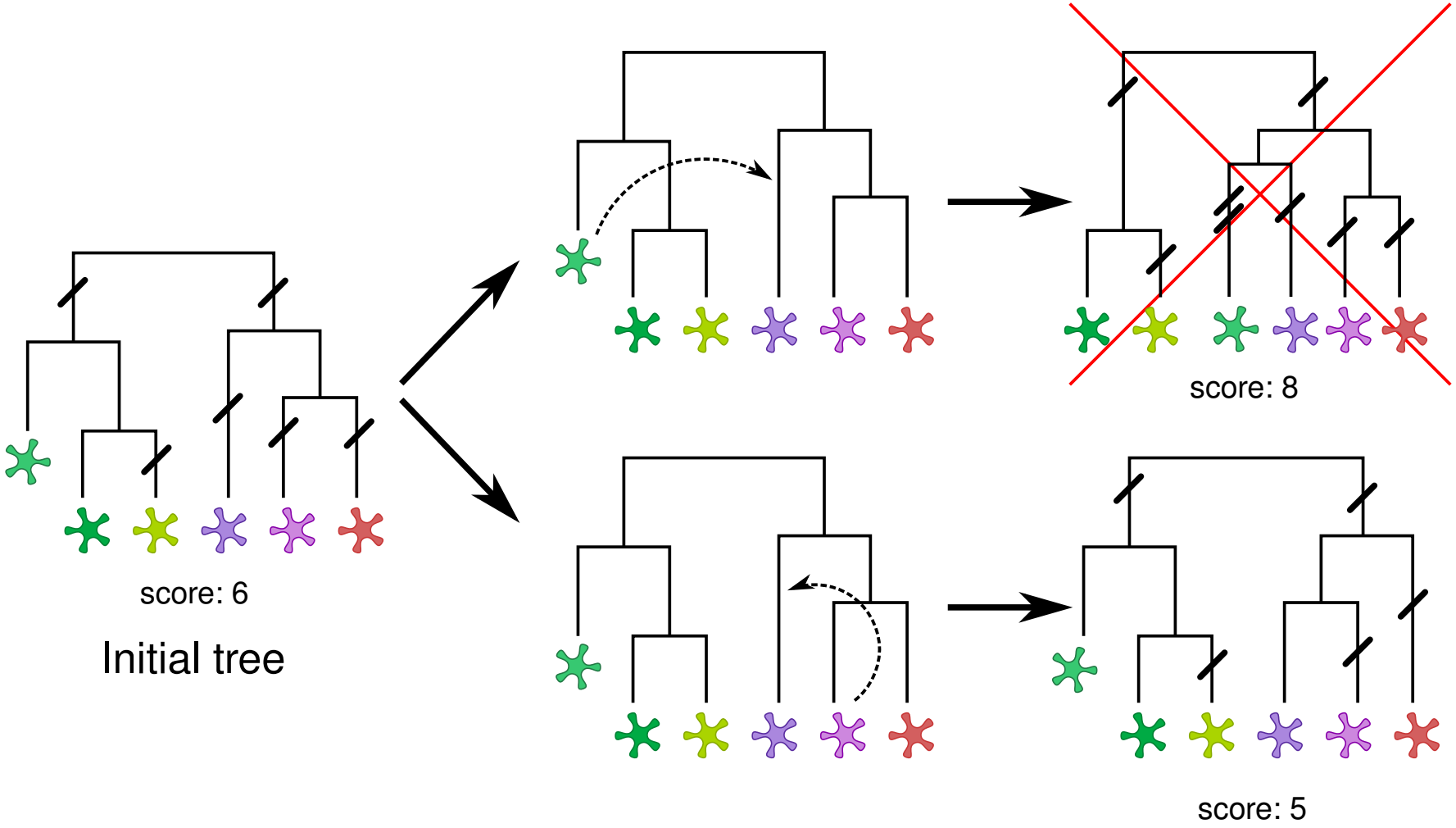
Initial tree

Maximum parsimony phylogenies

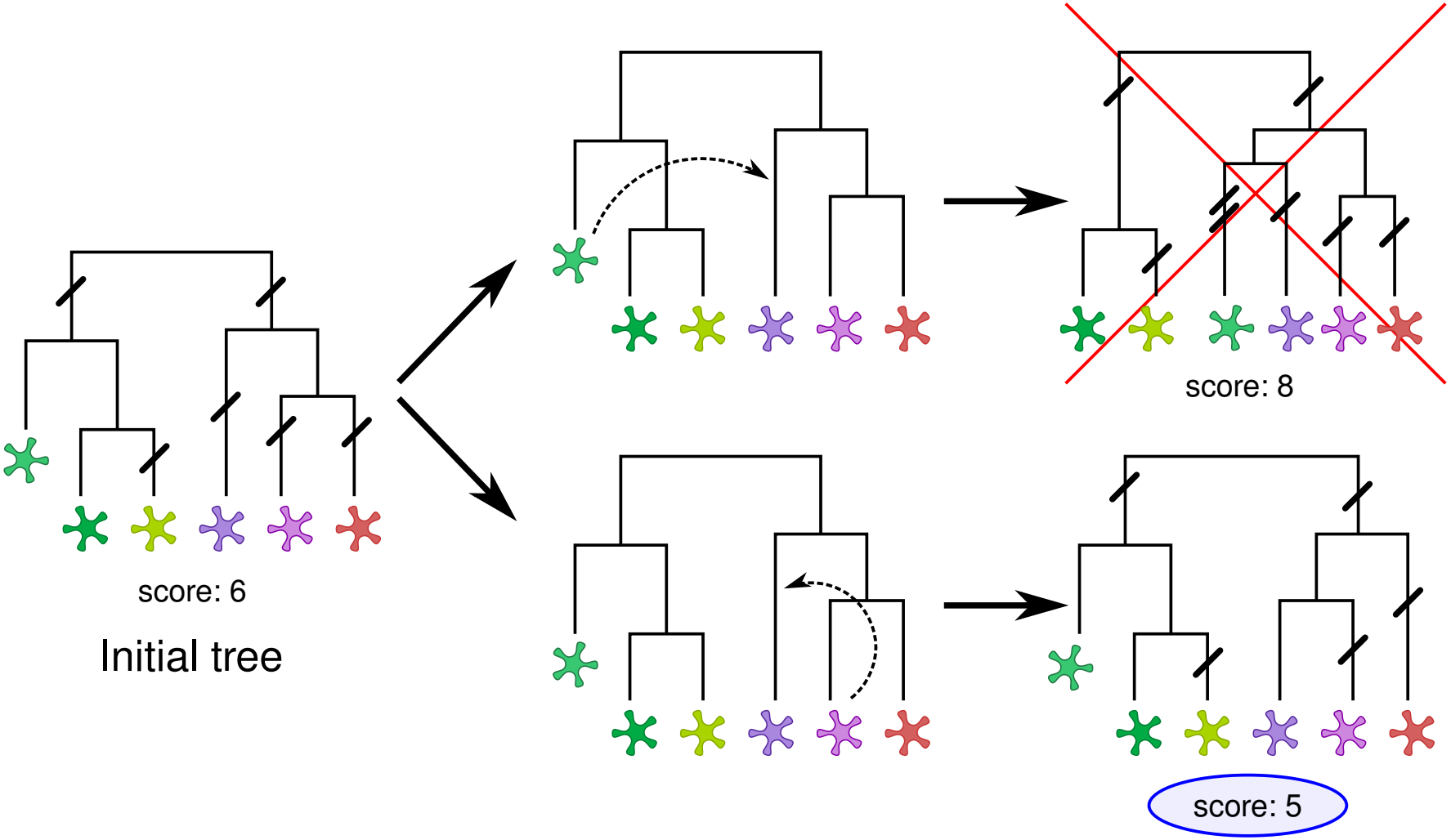


Initial tree

Maximum parsimony phylogenies



Maximum parsimony phylogenies



Maximum parsimony phylogenies

Advantages

- simple
- intuitive explanation / evolutionary meaning

Maximum parsimony phylogenies

Advantages

- simple
- intuitive explanation / evolutionary meaning

Limitations

- very limited flexibility
- no parameter estimation
- computer-intensive
- single tree produced
- simplistic (multipled substitutions not accounted for)

Likelihood-based phylogenies (ML / Bayesian)

Approaches relying on a probabilistic **model of sequence evolution**:

- **ML**: tree and parameter values giving maximum likelihood
- **Bayesian**: samples of trees and parameter values according to posterior probability

Likelihood-based phylogenies (ML / Bayesian)

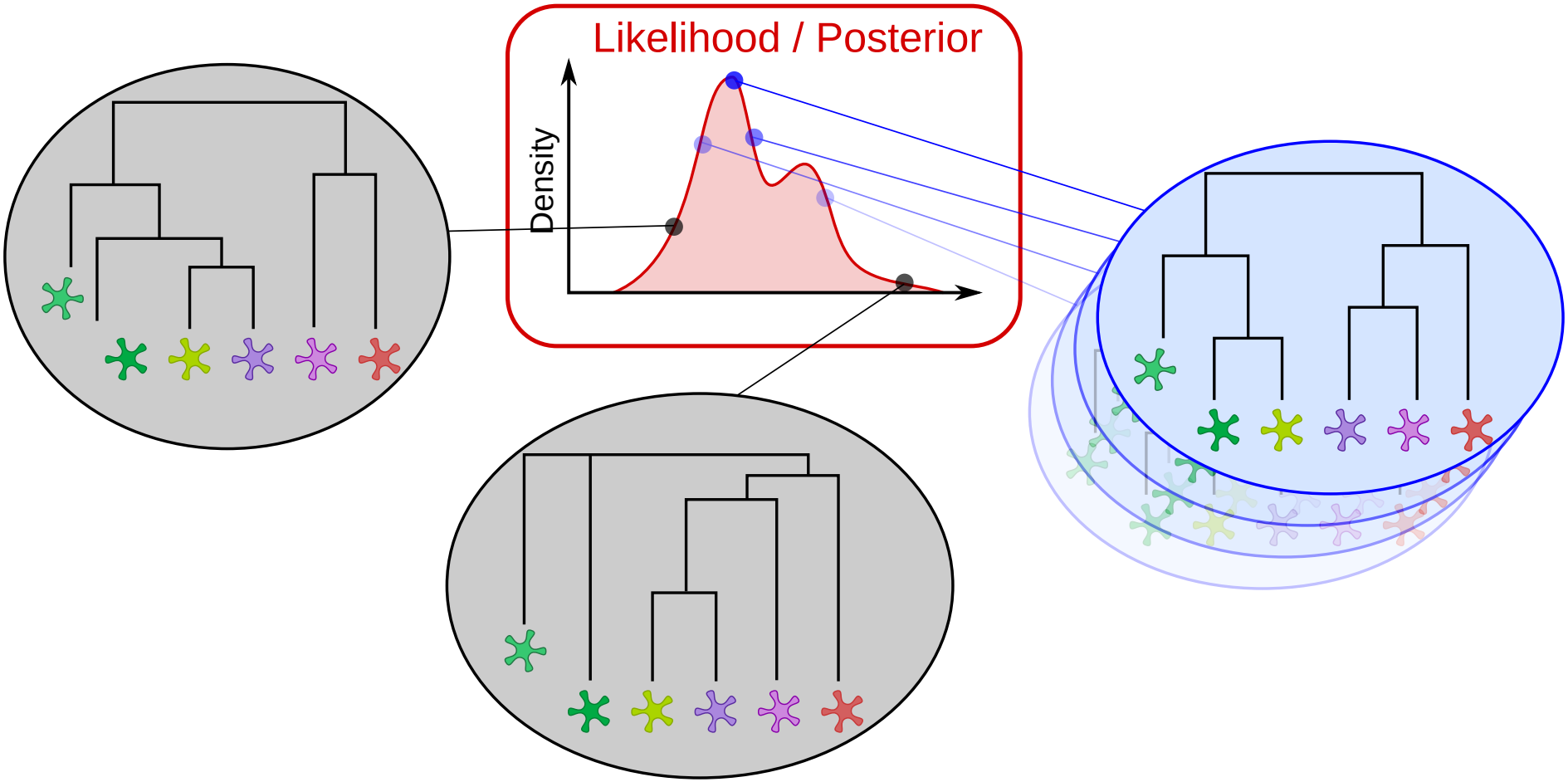
Approaches relying on a probabilistic **model of sequence evolution**:

- **ML**: tree and parameter values giving maximum likelihood
- **Bayesian**: samples of trees and parameter values according to posterior probability

Rationale

1. start from a pre-defined tree
2. compute initial likelihood/posterior
3. permute branches, sample new parameters and compute likelihood/posterior
4. accept new tree and parameters based on likelihood/posterior improvement
5. go back to 3) until convergence

Likelihood-based phylogenies (ML / Bayesian)



Likelihood-based phylogenies (ML / Bayesian)

Advantages

- very flexible
- consistent with a model of evolution
- statistically consistent (model comparison)
- parameter estimation

Likelihood-based phylogenies (ML / Bayesian)

Advantages

- very flexible
- consistent with a model of evolution
- statistically consistent (model comparison)
- parameter estimation

Limitations

- computer-intensive
- choice of the model of evolution

Outline

Various approaches

Main issue: assess the uncertainty of the tree topology / individual nodes

Various approaches

Main issue: assess the uncertainty of the tree topology / individual nodes

Approaches

- any phylogenetic reconstruction: bootstrap (individual nodes)
- ML approaches: model selection (whole topology)
- Bayesian methods: between-samples variability (individual nodes)

Bootstrapping phylogenies

- aims to assess **variability due to sampling the genome** and **conflicting signals**
- relies on analysing **resampled datasets**

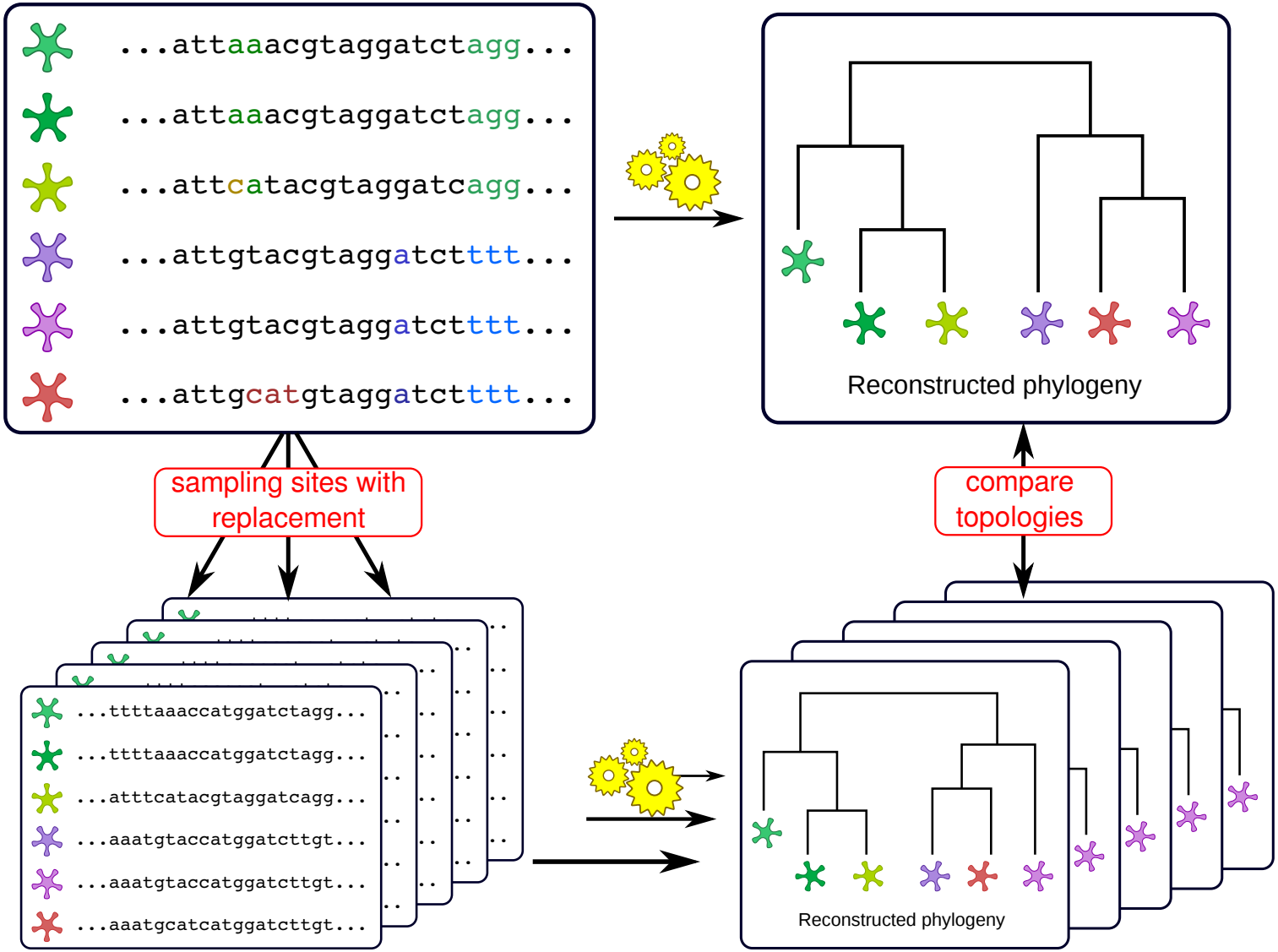
Bootstrapping phylogenies

- aims to assess **variability due to sampling the genome** and **conflicting signals**
- relies on analysing **resampled datasets**

Rationale

1. obtain a reference tree
2. resample the sites with replacement
3. obtain a tree for the resampled dataset
4. go back to 2) until the desired number of bootstrapped trees is attained
5. compute the frequency of each bifurcation of the reference tree occuring in bootstrapped trees

Bootstrapping phylogenies



Bootstrapping phylogenies

Advantages

- standard
- simple to implement

Bootstrapping phylogenies

Advantages

- standard
- simple to implement

Limitations

- possibly computer-intensive
- assumes that the genome has been sampled randomly (often wrong)

In practice...

Many software available for each step of the analysis:

- alignment: clustalw/clustalx, MUSCLE, ...
- refining alignments: Jalview, clustalx, seaview, ...
- phylogenetic reconstruction: **R**, **MEGA**, PAUP, MrBayes, BEAST, ...

R: apart from alignment, largest choice of methods for phylogenetics (and beyond).

MEGA: standalone software for getting sequences, alignment, phylogenetic reconstruction.