# A (short) introduction to phylogenetics

**Thibaut Jombart**, Caitlin Collins

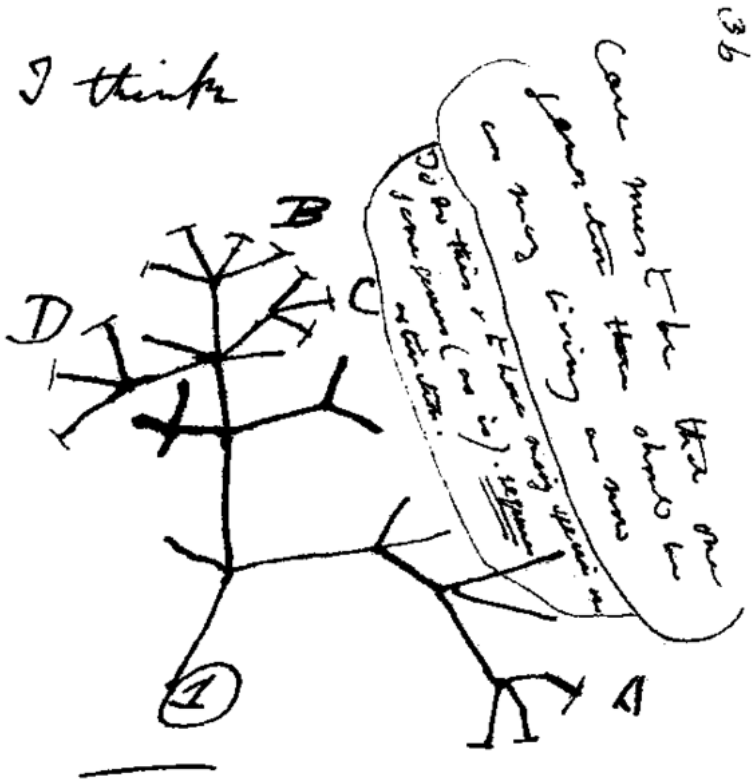
MRC Centre for Outbreak Analysis and Modelling
Imperial College London

*Genetic data analysis using* R, University of Leuven
27-10-2014

# Outline

# Outline

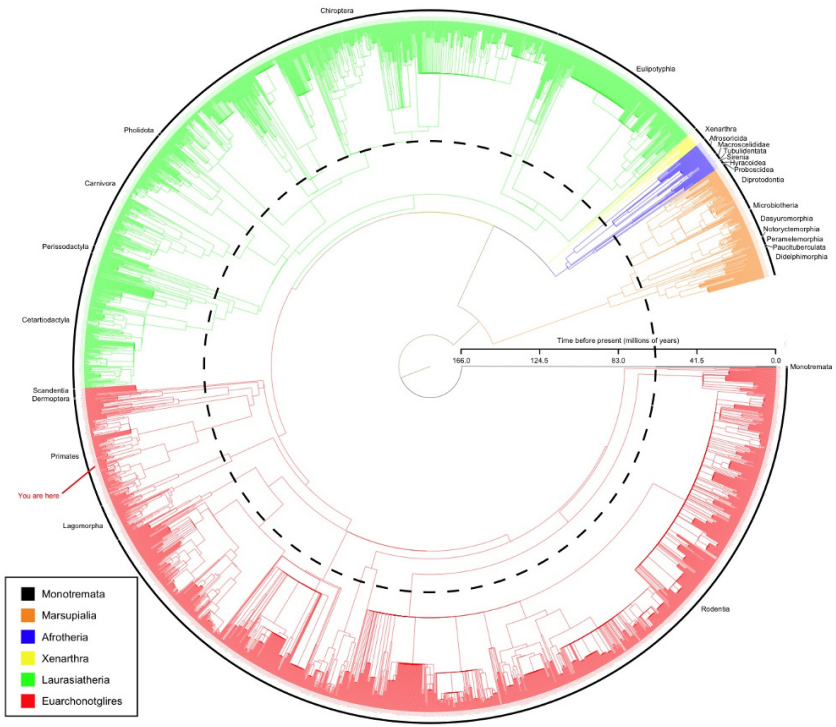
# Phylogenetics: from the origins...




C. Darwin, Notebook, 1837.

*'From the first growth of the tree, many a limb and branch has decayed and dropped off; and these fallen branches of various sizes may represent those whole orders, families, and genera which have now no living representatives, and which are known to us only in a fossil state.'*

# Phylogenetics: ...to the present




Bininda-Emonds *et al.*, 2007, Nature.

- phylogenetic trees are part of the standard toolbox of genetic data analysis
- represent the evolutionary history of a set of (sampled) taxa

# And the main difference is...

**Current trees look better!**

(and some other minor differences)

# And the main difference is...

**Current trees look better!**

(and some other minor differences)

# And the main difference is...

**Current trees look better!**

(and some other minor differences)

# About the minor differences...

- DNA sequencing revolution
- huge data banks freely available (e.g. GenBank)
- easier, cheaper, faster to obtain DNA sequences
- increasing number of full genomes available

**Different ways to exploit this information.**

# About the minor differences...

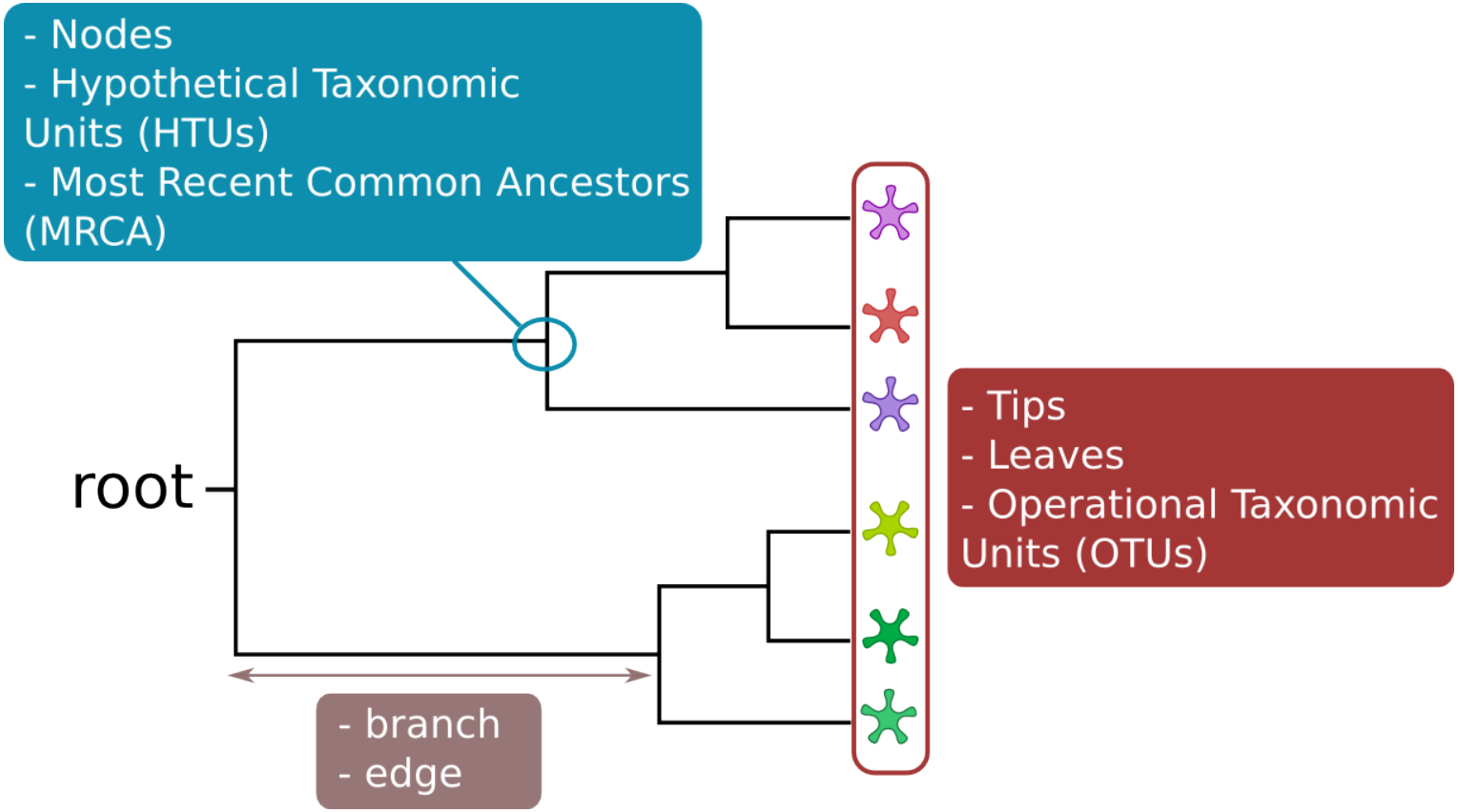
- DNA sequencing revolution
- huge data banks freely available (e.g. GenBank)
- easier, cheaper, faster to obtain DNA sequences
- increasing number of full genomes available

**Different ways to exploit this information.**

# Outline

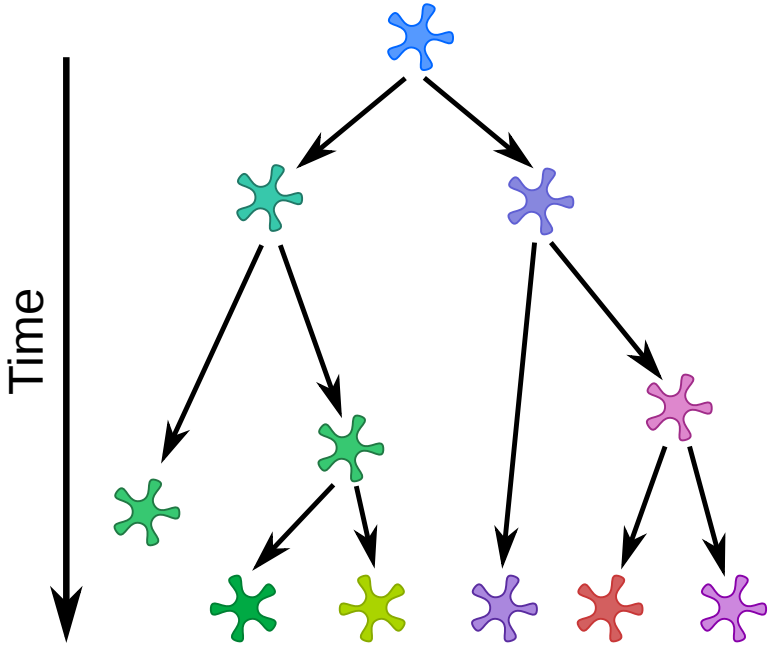
# Phylogenetic trees: some useful terms

**Phylogenetic tree**: representation of evolutionary relationships between a set of taxa.



# Accumulated substitutions tell us about the genealogy

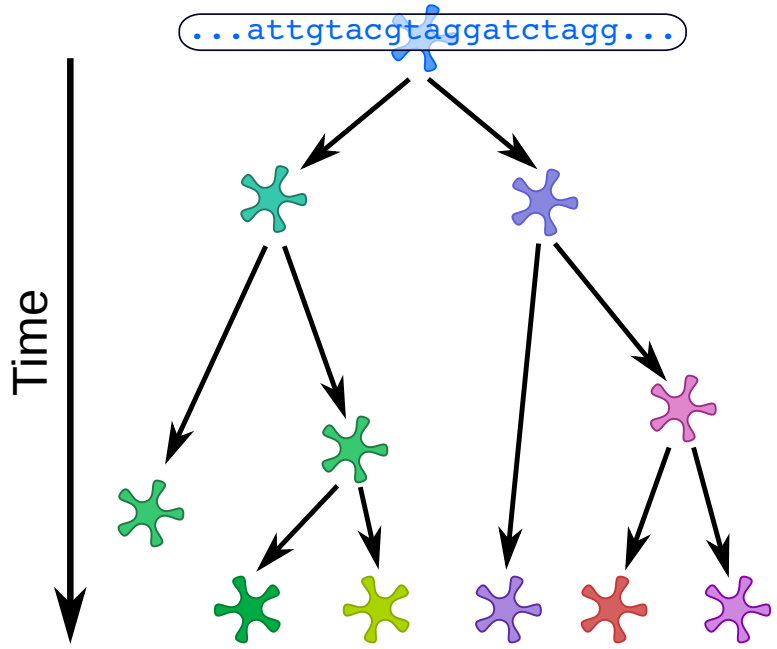
**Substitution**: replacement of a nucleotide (e.g. a $\rightarrow$ t)




True phylogeny

# Accumulated substitutions tell us about the genealogy

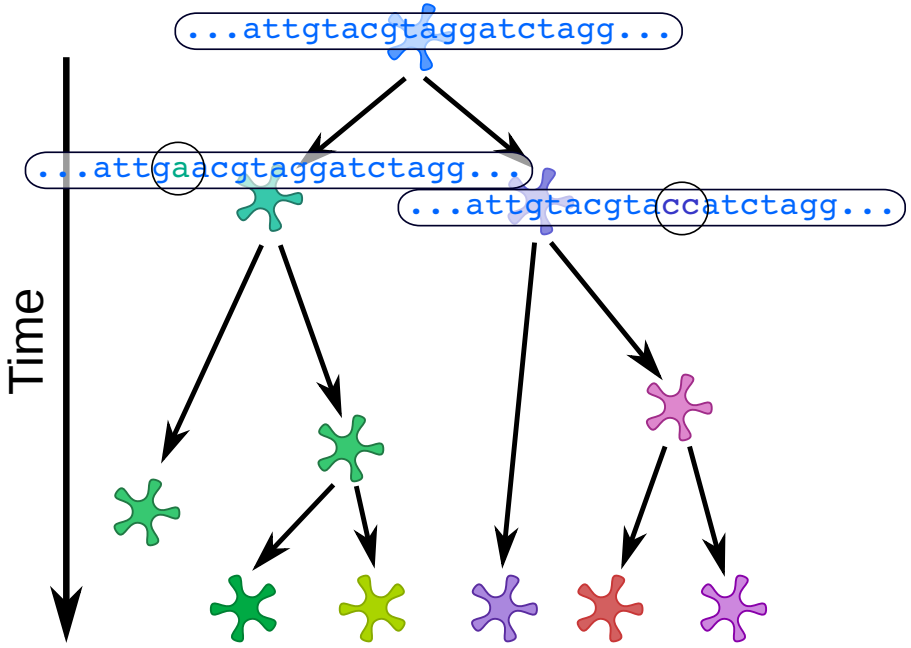
**Substitution**: replacement of a nucleotide (e.g. a $\rightarrow$ t)




True phylogeny

# Accumulated substitutions tell us about the genealogy

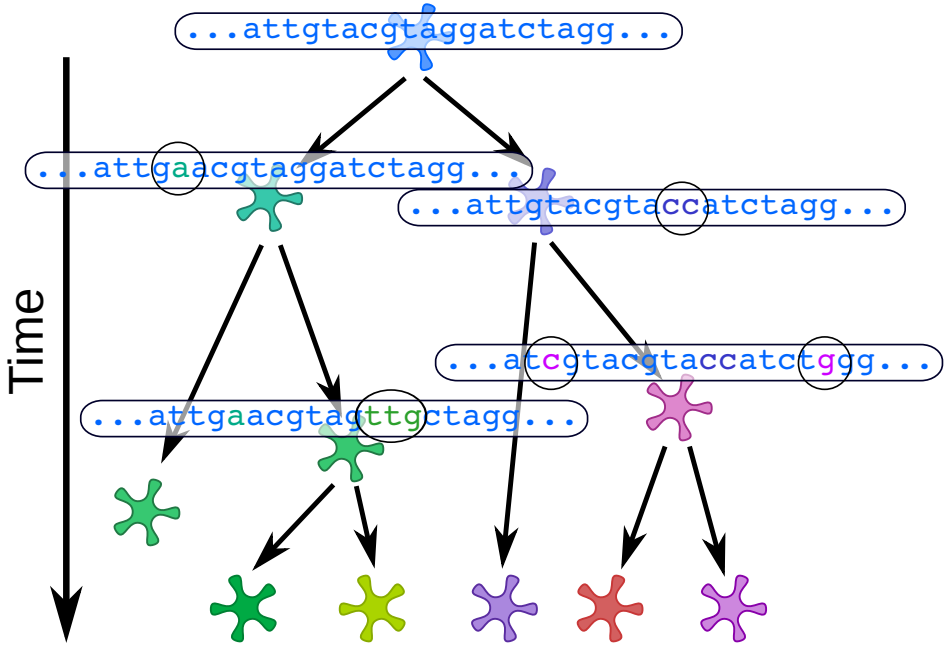
**Substitution**: replacement of a nucleotide (e.g. a $\rightarrow$ t)




True phylogeny

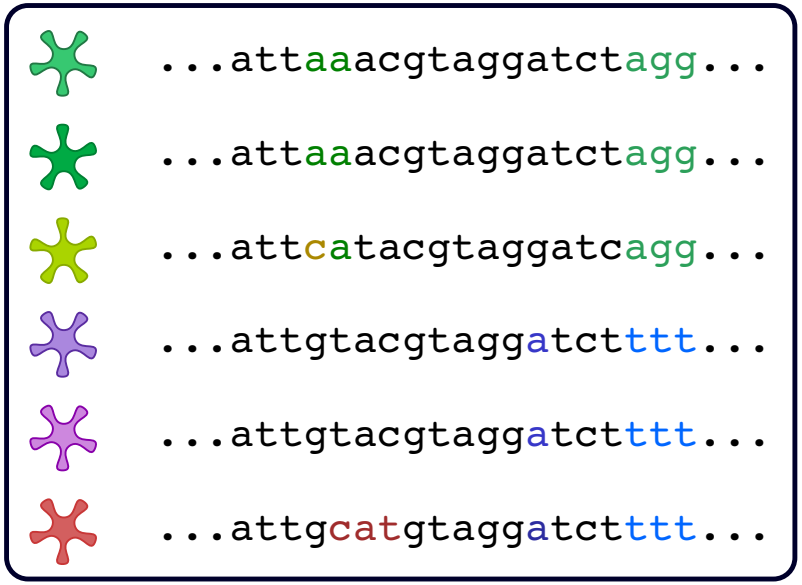
# Accumulated substitutions tell us about the genealogy

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

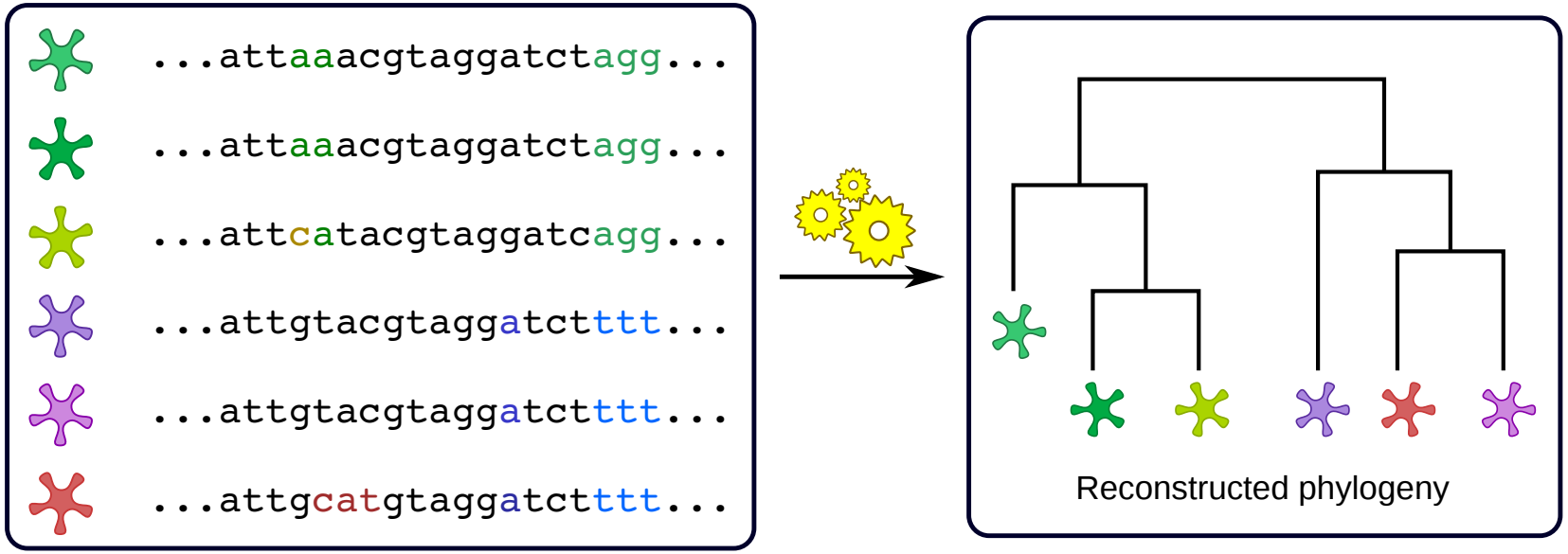



True phylogeny

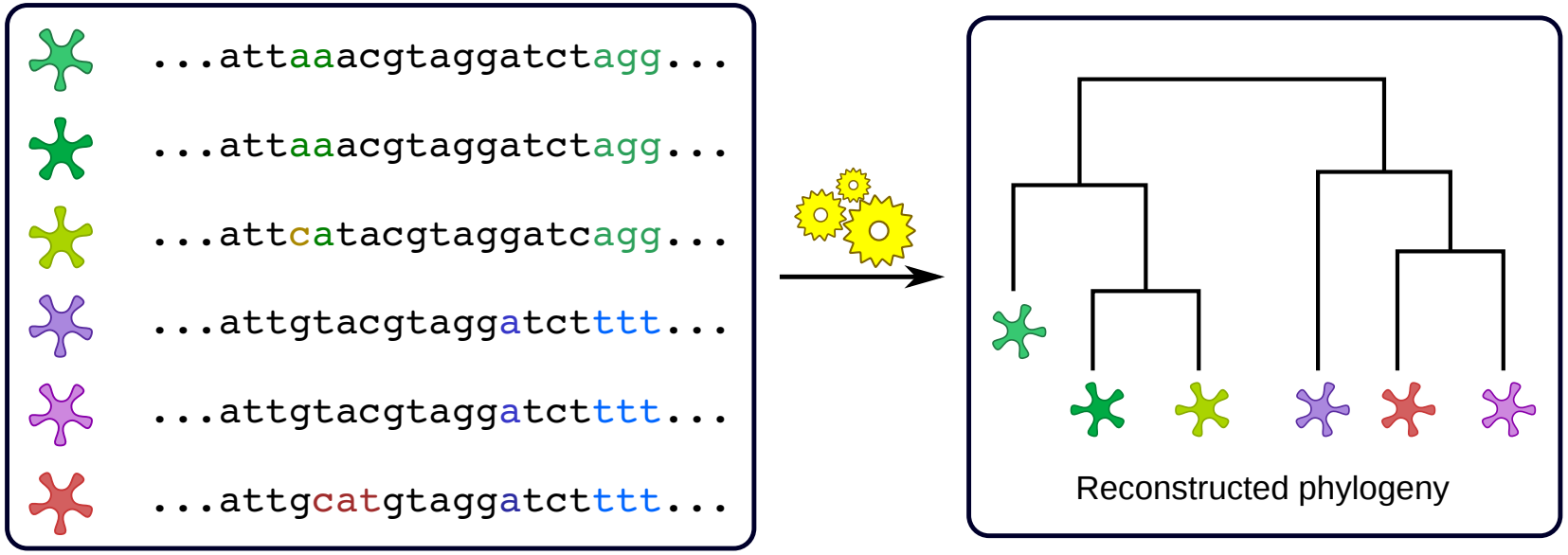
# From alignments to phylogenies



# From alignments to phylogenies



# From alignments to phylogenies




Different methods for achieving phylogenetic reconstruction.

# Workflow

## Prepare data

- align sequences: alignment software + manual refinement

## Build the tree

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

## Analyse the tree

- assess uncertainty
- test phylogenetic signal
- model trait evolution
- ...

# Workflow

## Prepare data

- align sequences: alignment software + manual refinement

## Build the tree

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

## Analyse the tree

- assess uncertainty
- test phylogenetic signal
- model trait evolution
- ...

# Workflow

## Prepare data

- align sequences: alignment software + manual refinement

## Build the tree

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

## Analyse the tree

- assess uncertainty
- test phylogenetic signal
- model trait evolution
- ...

# Workflow

## Prepare data

- align sequences: alignment software + manual refinement

## **Build the tree**

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

## Analyse the tree

- assess uncertainty
- test phylogenetic signal
- model trait evolution
- ...

# Outline

# Distance-based phylogenetic reconstruction

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

## Rationale

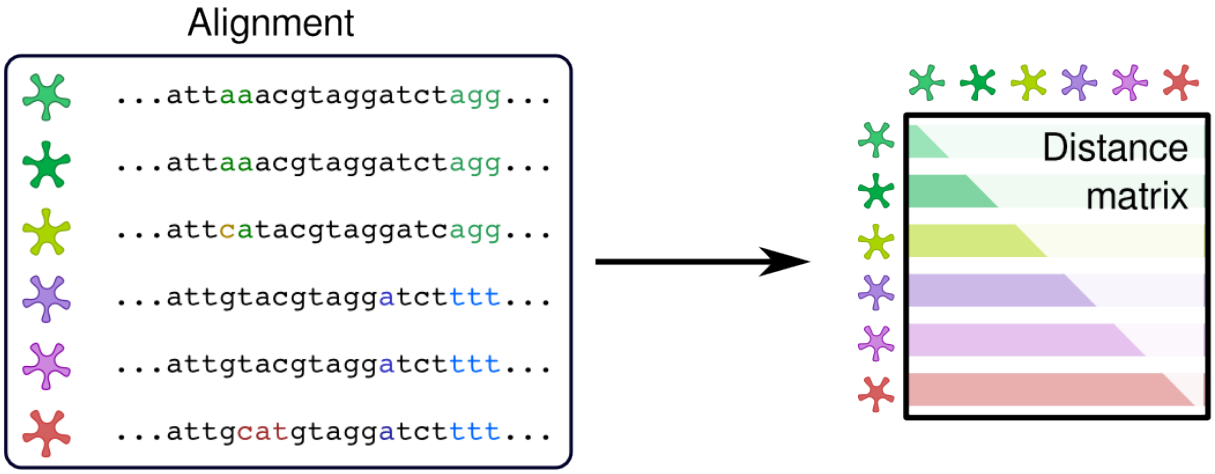
1. compute pairwise genetic distances $\mathbf{D}$
2. group closest sequences
3. update $\mathbf{D}$
4. go back to 2) until all sequences are grouped

# Distance-based phylogenetic reconstruction

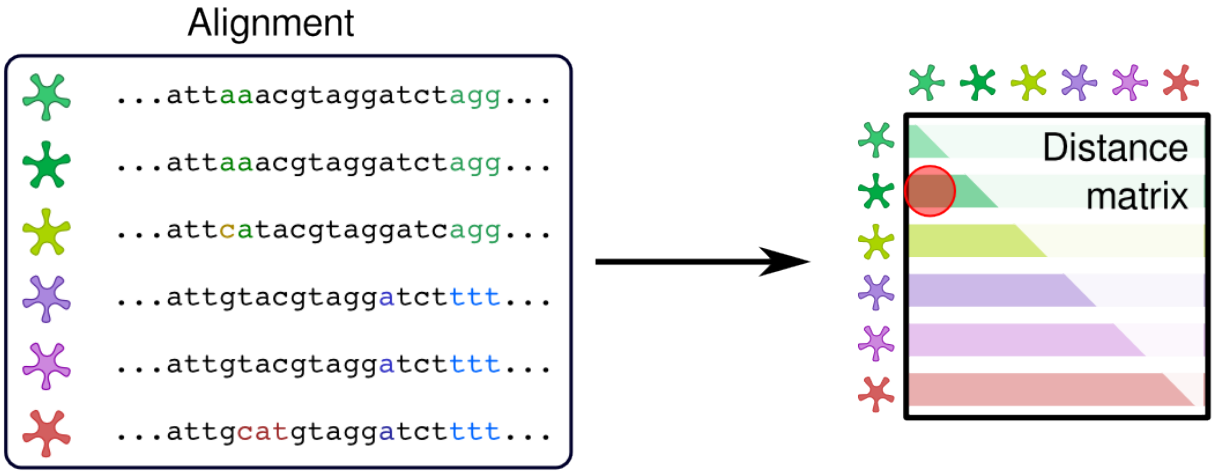
Alignment

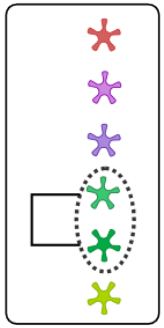
```
...attaaacgtaggatctagg...
...attaaacgtaggatctagg...
...attcatacgtaggatcagg...
...attgtacgtaggatctttt...
...attgtacgtaggatctttt...
...attgcatgtaggatctttt...
```

# Distance-based phylogenetic reconstruction

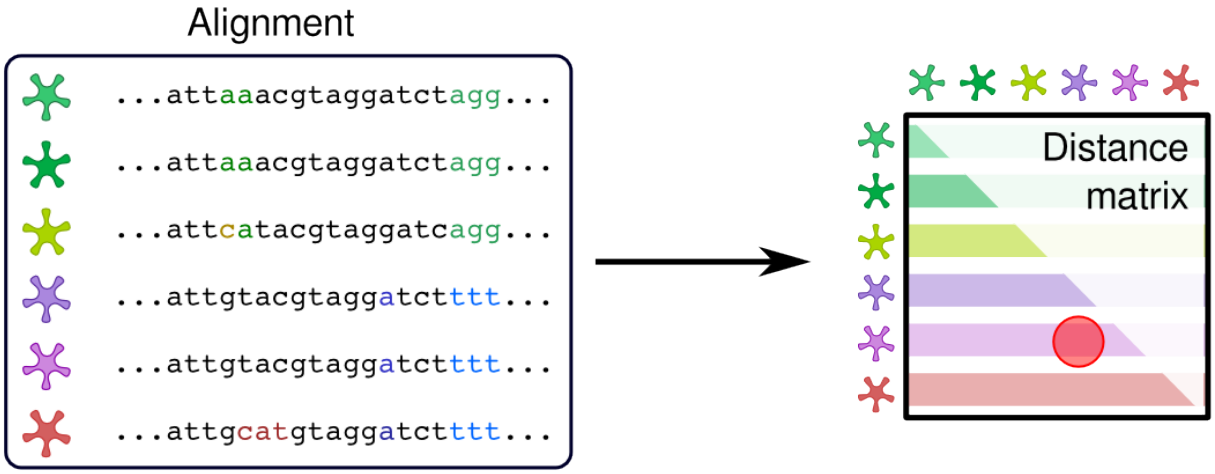


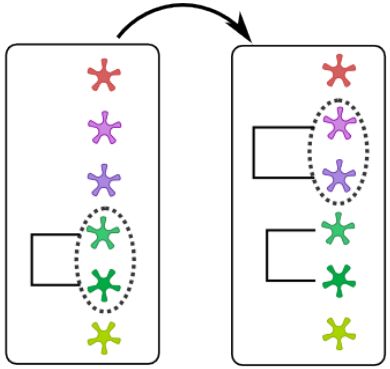
# Distance-based phylogenetic reconstruction



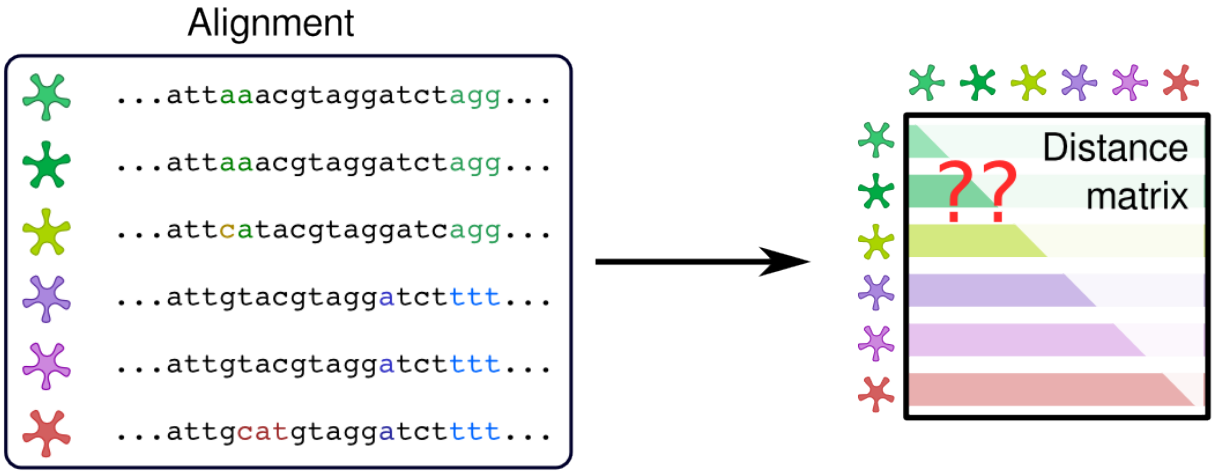



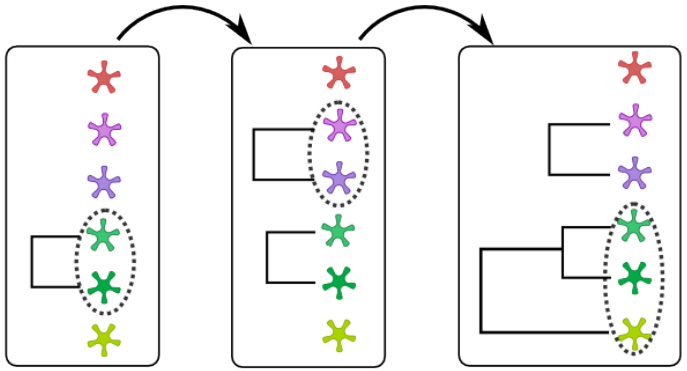
# Distance-based phylogenetic reconstruction



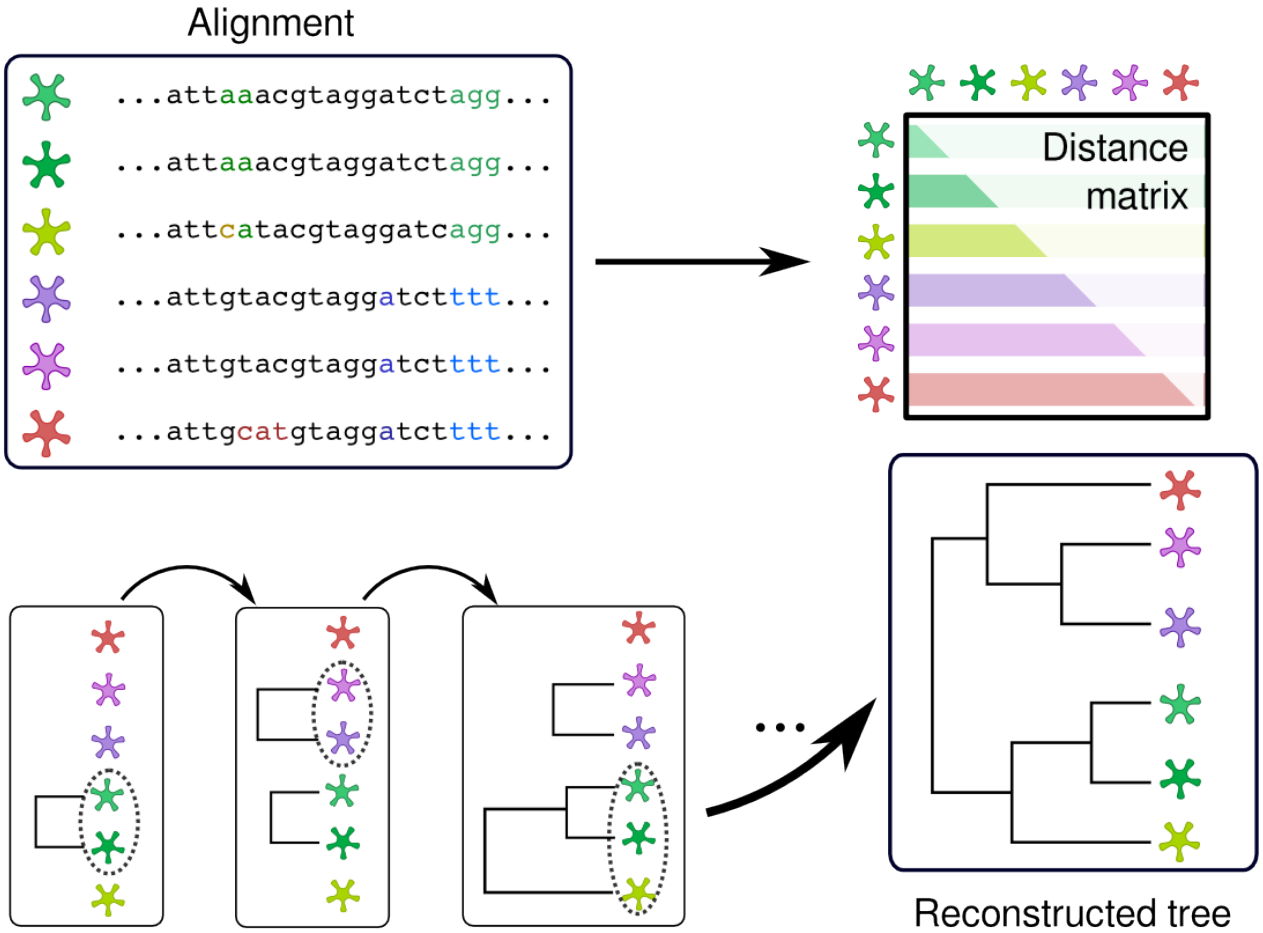



# Distance-based phylogenetic reconstruction



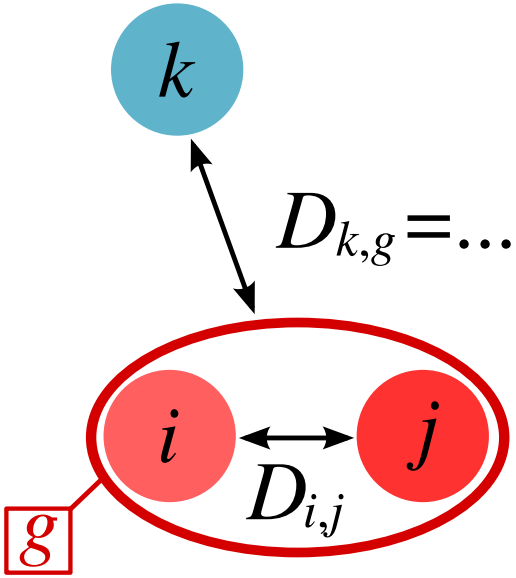



# Distance-based phylogenetic reconstruction



# What is the distance between a node and tips?

**Hierarchical clustering:**

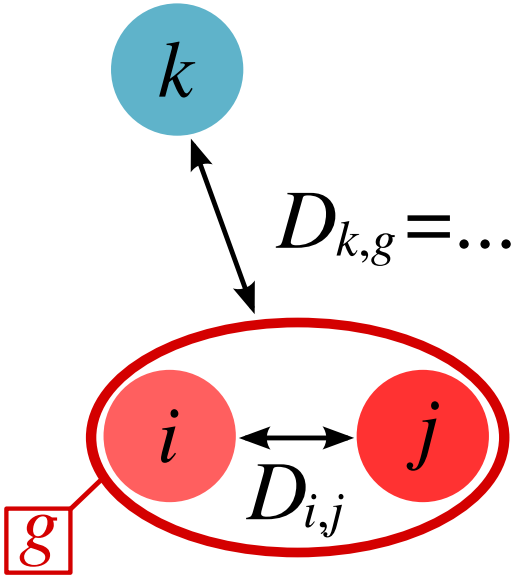



- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

**Neighbor joining:**
Transforms original distances to account for heterogeneous rates of evolution.

# What is the distance between a node and tips?

**Hierarchical clustering:**

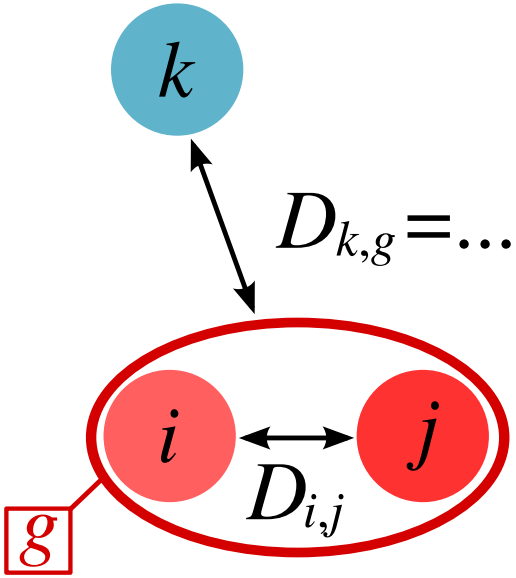



- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

**Neighbor joining:**
Transforms original distances to account for heterogeneous rates of evolution.

# What is the distance between a node and tips?

**Hierarchical clustering:**

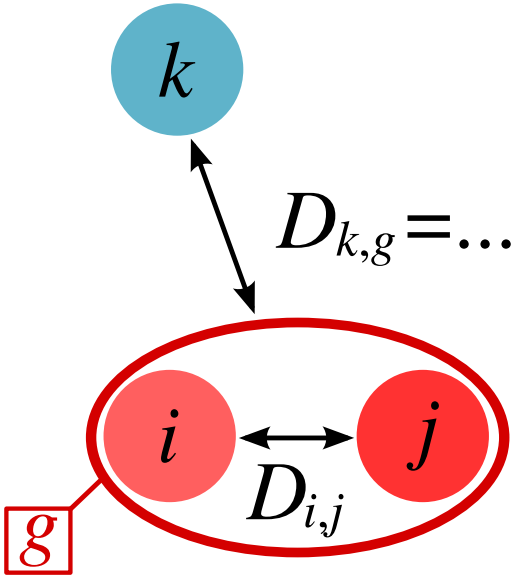



- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

**Neighbor joining:**
Transforms original distances to account for heterogeneous rates of evolution.

# What is the distance between a node and tips?

**Hierarchical clustering:**

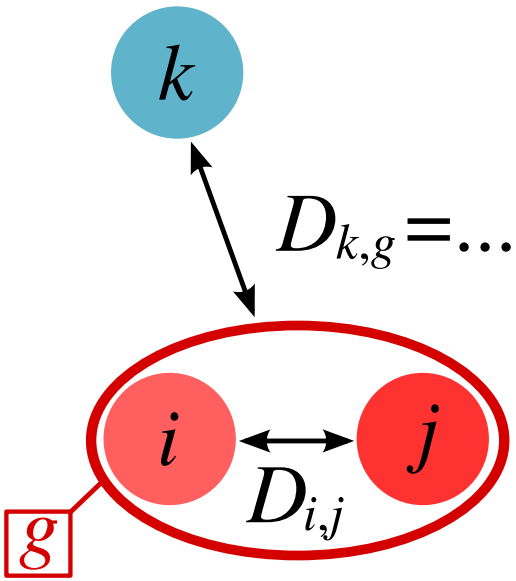



- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

**Neighbor joining:**
Transforms original distances to account for heterogeneous rates of evolution.

# What is the distance between a node and tips?

**Hierarchical clustering:**




- single linkage: $D_{k,g} = \min(D_{k,i}, D_{k,j})$
- complete linkage: $D_{k,g} = \max(D_{k,i}, D_{k,j})$
- UPGMA: $D_{k,g} = \frac{D_{k,i}+D_{k,j}}{2}$

**Neighbor joining:**
Transforms original distances to account for heterogeneous rates of evolution.

# Distance-based phylogenetic reconstruction

## Advantages

- simple
- flexible (many distances and clustering algorithms)
- fast and scalable (applicable to large datasets)

## Limitations

- sensitive to distance/clustering chosen
- evolutionary rates are not estimated
- no measure of uncertainty for the tree obtained

# Distance-based phylogenetic reconstruction

## Advantages

- simple
- flexible (many distances and clustering algorithms)
- fast and scalable (applicable to large datasets)

## Limitations

- sensitive to distance/clustering chosen
- evolutionary rates are not estimated
- no measure of uncertainty for the tree obtained

# Outline

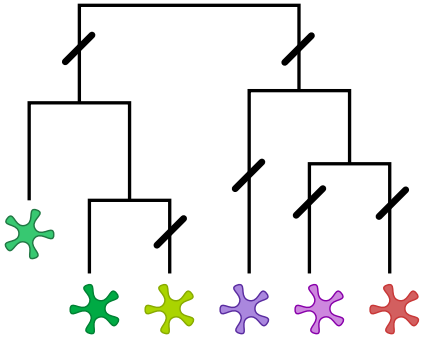
# Maximum parsimony phylogenies

Approaches relying on finding the tree with the **smallest number of character changes (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

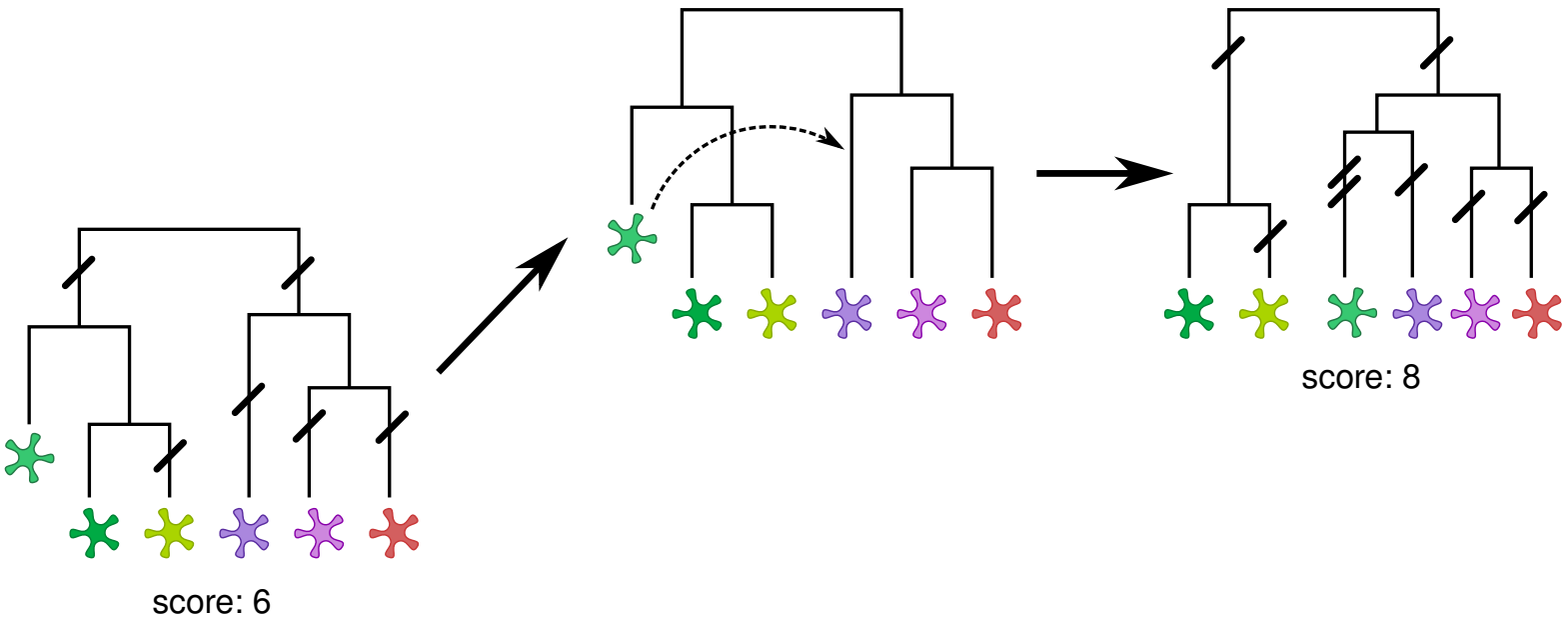
# Maximum parsimony phylogenies



score: 6

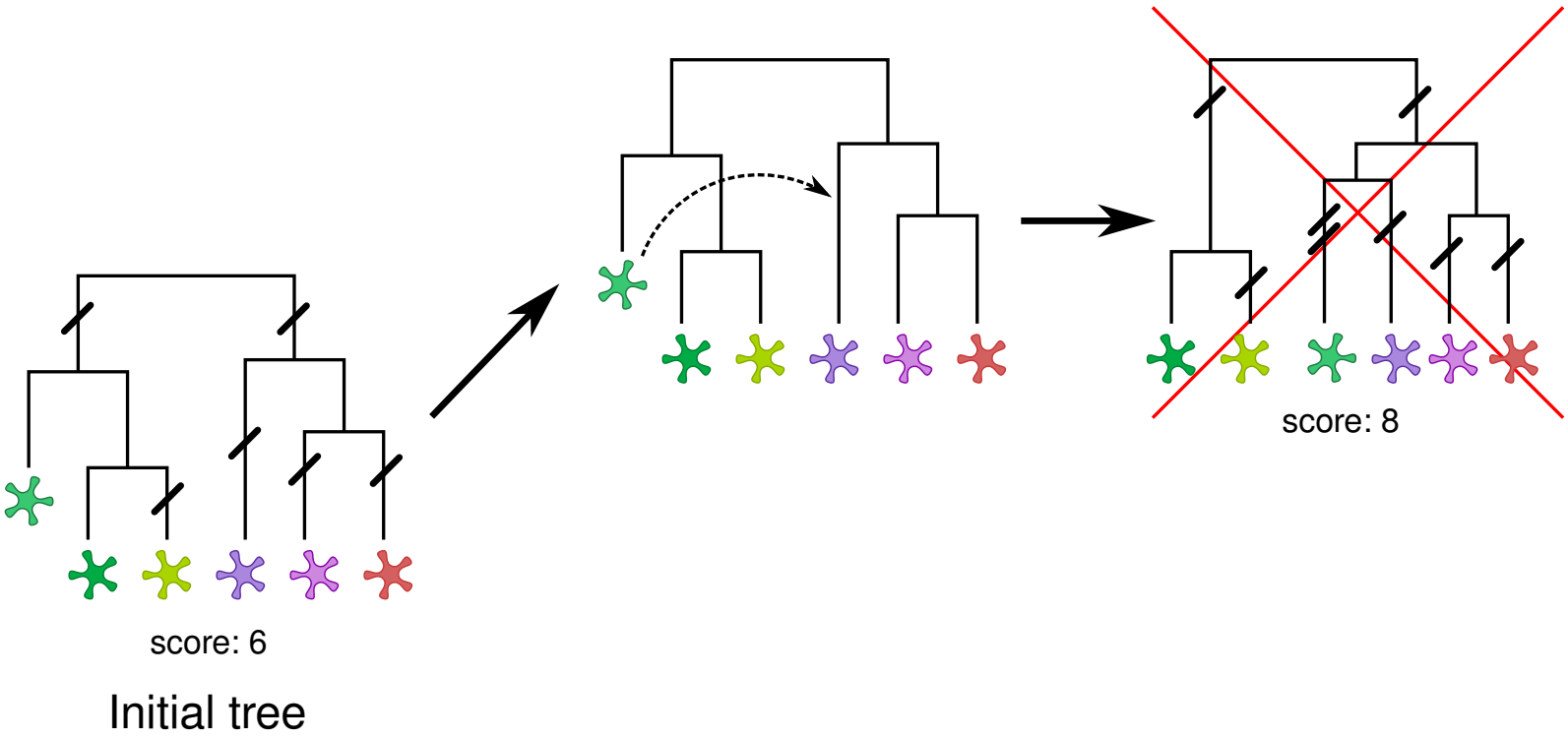
Initial tree

# Maximum parsimony phylogenies



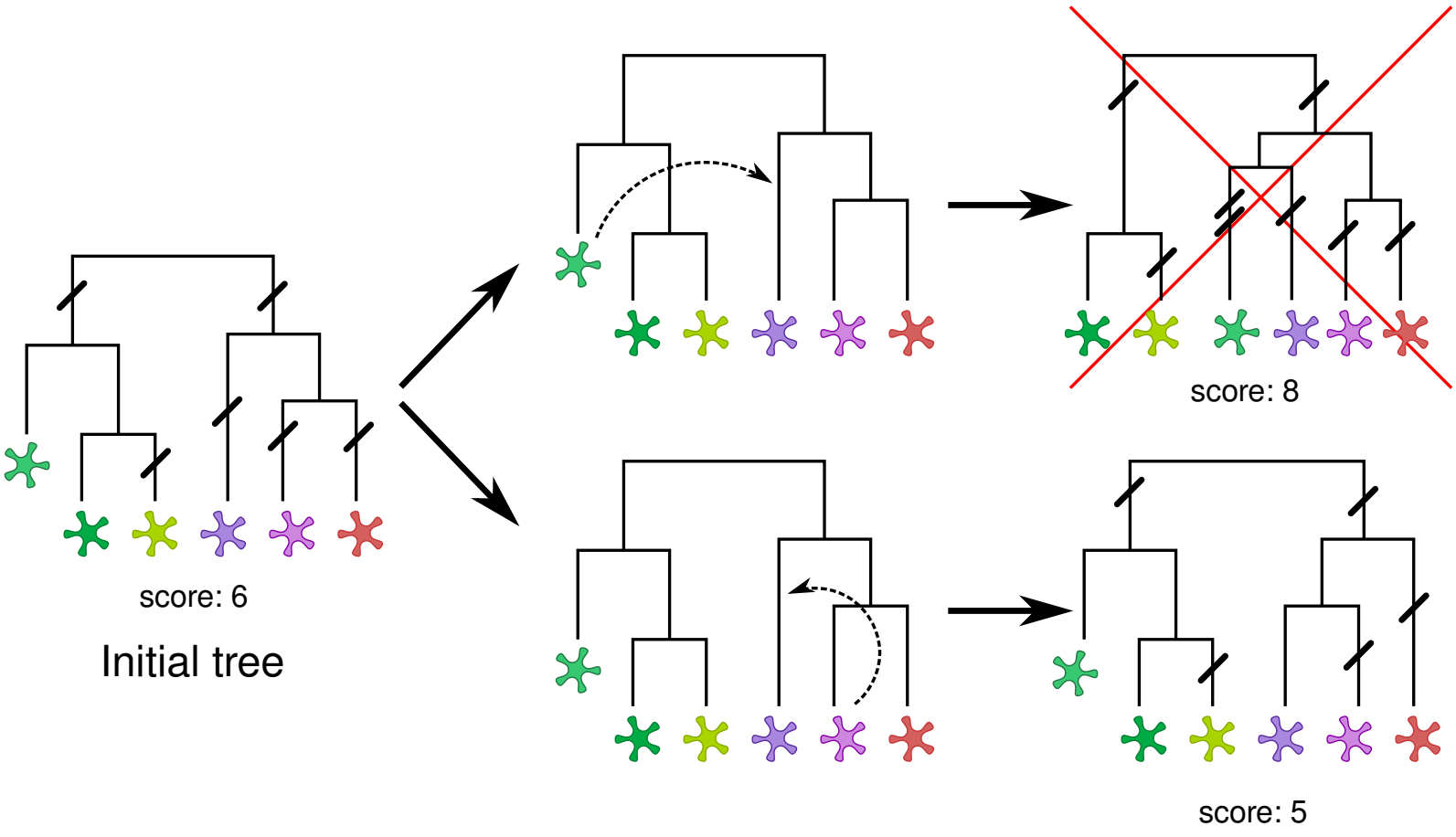

Initial tree

# Maximum parsimony phylogenies

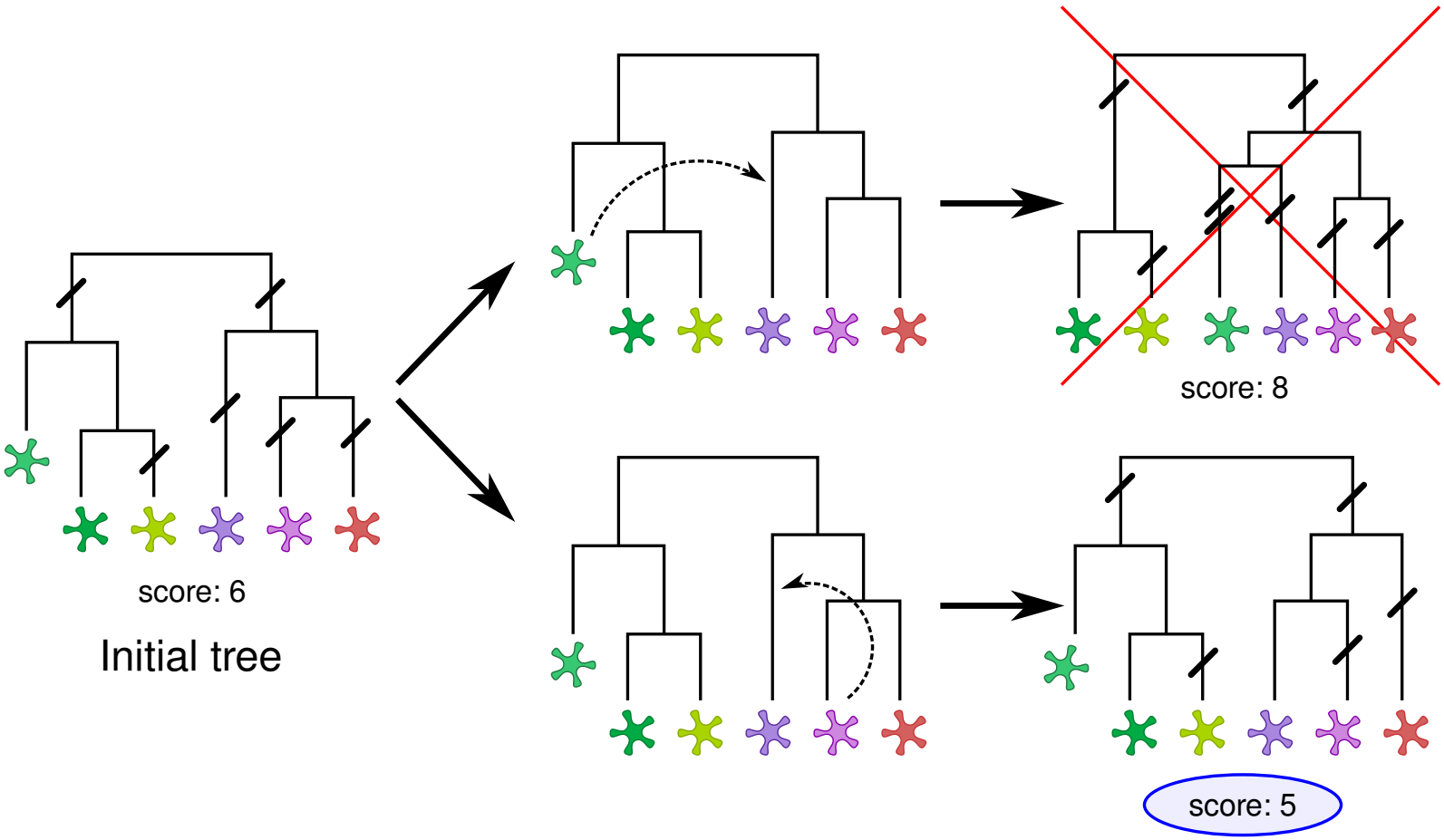



Initial tree

# Maximum parsimony phylogenies



# Maximum parsimony phylogenies



# Maximum parsimony phylogenies

## Advantages

- applicable to any discontinuous characters (not just DNA)
- intuitive explanation: 'simplest' evolutionary scenario

## Limitations

- evolutionary rates are not estimated
- no measure of uncertainty for the tree obtained
- computer-intensive
- different types of substitutions ignored
- evolution not necessarily parsimonious
- sensitive to heterogeneous rates of evolution (*long branch attraction*)

# Maximum parsimony phylogenies

## Advantages

- applicable to any discontinuous characters (not just DNA)
- intuitive explanation: 'simplest' evolutionary scenario

## Limitations

- evolutionary rates are not estimated
- no measure of uncertainty for the tree obtained
- computer-intensive
- different types of substitutions ignored
- evolution not necessarily parsimonious
- sensitive to heterogeneous rates of evolution (*long branch attraction*)

# Outline

# Likelihood-based phylogenies (ML / Bayesian)

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

- **ML**: find tree and evolutionary rates with highest likelihood
- **Bayesian**: find tree and evolutionary rates 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)

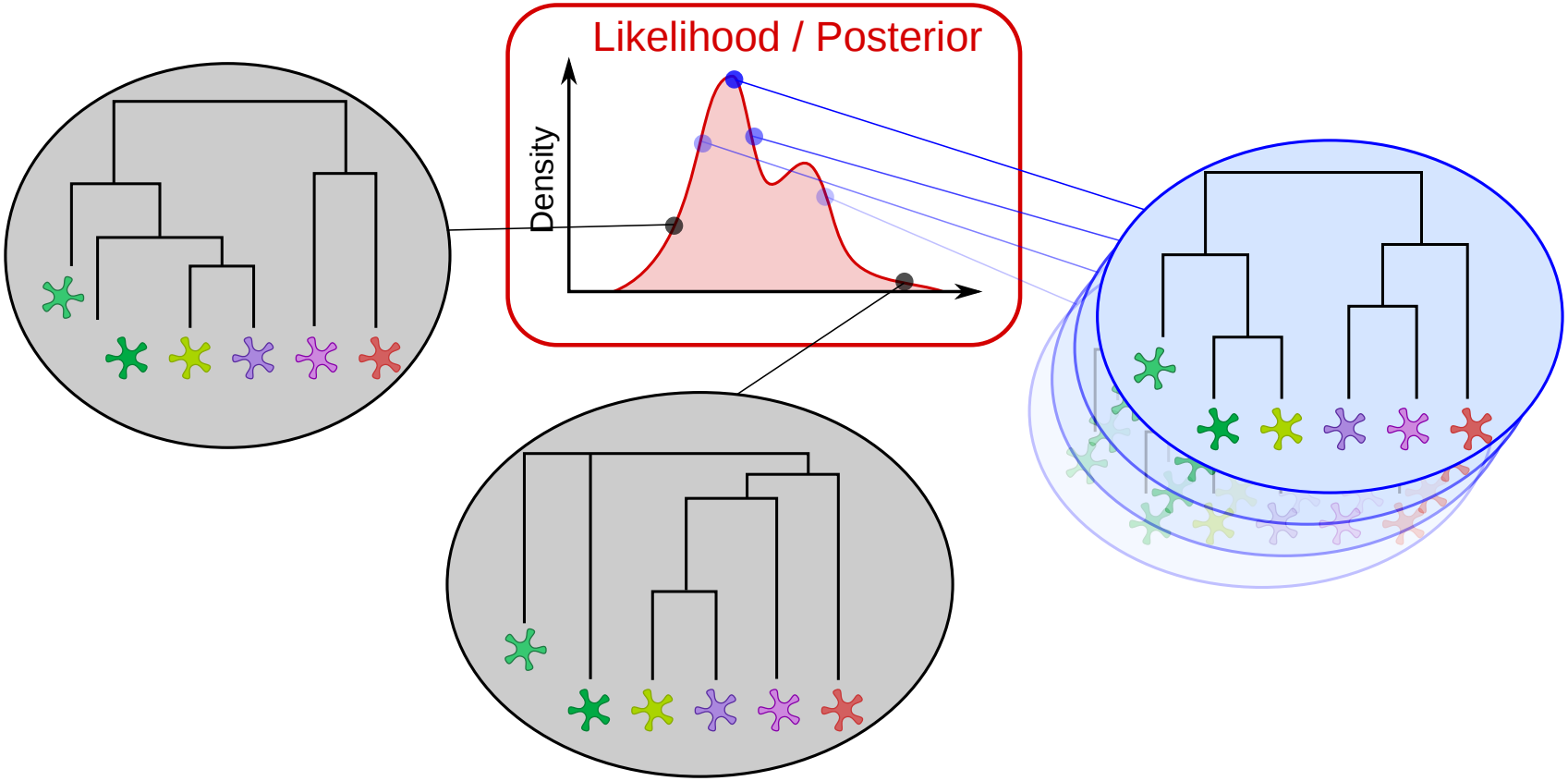
Approaches relying on a **model of sequence evolution**:

- **ML**: find tree and evolutionary rates with highest likelihood
- **Bayesian**: find tree and evolutionary rates 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
- (Bayesian) several trees $\rightarrow$ measure of uncertainty

## Limitations

- computer-intensive
- choice of the model of evolution
- (ML) no measure of uncertainty for the tree obtained
- (Bayesian) need to find a consensus tree

# Likelihood-based phylogenies (ML / Bayesian)

## Advantages

- very flexible
- consistent with a model of evolution
- statistically consistent (model comparison)
- parameter estimation
- (Bayesian) several trees $\rightarrow$ measure of uncertainty

## Limitations

- computer-intensive
- choice of the model of evolution
- (ML) no measure of uncertainty for the tree obtained
- (Bayesian) need to find a consensus tree

# Outline

# How do we know the tree is robust?

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

## Approaches

- ML: model selection to compare trees (whole tree)
- Bayesian methods: between-samples variability (individual nodes)
- any method: bootstrap (individual nodes)

# How do we know the tree is robust?

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

## Approaches

- ML: model selection to compare trees (whole tree)
- Bayesian methods: between-samples variability (individual nodes)
- any method: bootstrap (individual nodes)

# How do we know the tree is robust?

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

## Approaches

- ML: model selection to compare trees (whole tree)
- Bayesian methods: between-samples variability (individual nodes)
- any method: bootstrap (individual nodes)

# How do we know the tree is robust?

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

## Approaches

- ML: model selection to compare trees (whole tree)
- Bayesian methods: between-samples variability (individual nodes)
- any method: bootstrap (individual nodes)

# How do we know the tree is robust?

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

## Approaches

- ML: model selection to compare trees (whole tree)
- Bayesian methods: between-samples variability (individual nodes)
- **any method: bootstrap (individual nodes)**

# Bootstrapping phylogenies

- 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

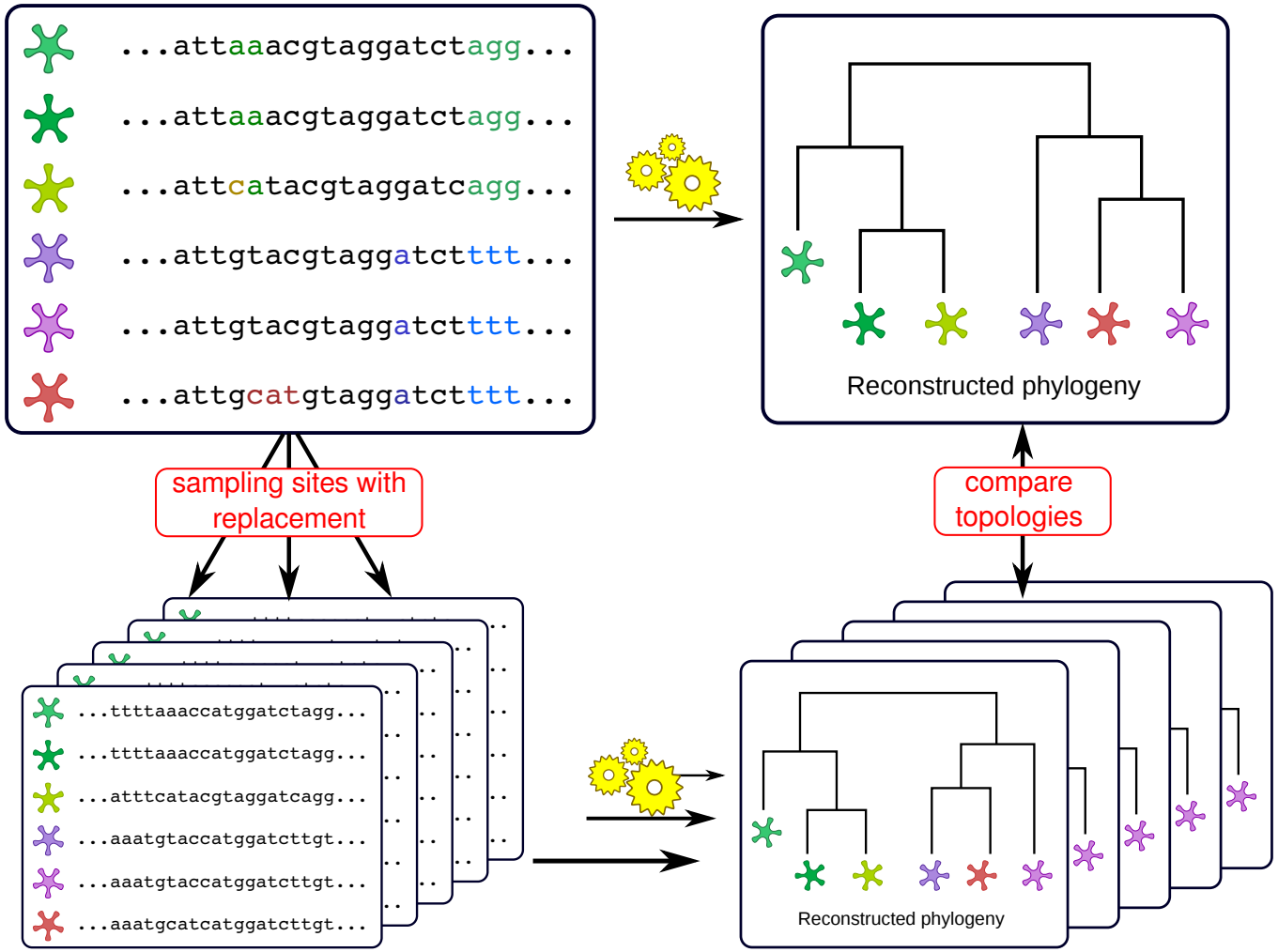
# Bootstrapping phylogenies

- 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

## Limitations

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

# Bootstrapping phylogenies

## Advantages

- standard
- simple to implement

## Limitations

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

# Outline

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)
- ...

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)
- ...

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)
- ...

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)
- ...

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)
- ...

# This is only the beginning

Many things can be done with trees

- estimate divergence time
- model trait evolution (phylogenetic comparative method)
- reconstruct ancestral states
- measure diversity
- infer past demographics/effective population size (coalescence)
- ...

# Time to get your hands dirty!

The pdf of the practical is online:

`http://adegenet.r-forge.r-project.org/`

or

Google → adegenet → documents → "Workshop Leuven, October 2014"