Molecular Evolution & Phylogenetics

Complexity of the search space, distance matrix methods, maximum parsimony

Jean-Baka Domelevo Entfellner
2nd EANBiT residential training
KEMRI-Wellcome Trust, Kilifi, 11 July 2019

biosciences
eastern and central **africa**

EANBiT
Eastern Africa Network for Bioinformatics Training

ILRI
INTERNATIONAL
LIVESTOCK RESEARCH
INSTITUTE

Learning Objectives

- understand the complexity of the space of all phylogenetic trees on *n* taxa

- understand what are distance matrix methods for phylogenetic inference

- understand what is the maximum parsimony paradigm and how it is used to infer trees

Learning Outcomes

- be able to give reasonable estimates of the number of trees on n taxa, perhaps with the mathematical formula

- be able to build "manually" the UPGMA tree (distance method) from a small distance matrix

- be able to calculate the most parsimonious history on a given labeled tree

Strategies in the quest for "the best" phylogenetic tree
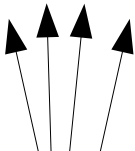
The tree inference problem

Problem: Assuming common descent, how to derive the "most probably correct" tree from the knowledge of the traits in the extant taxa (leaves)?
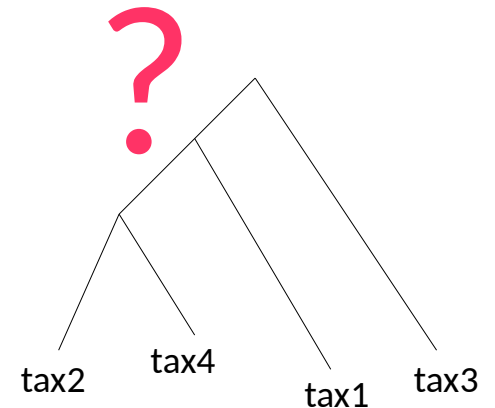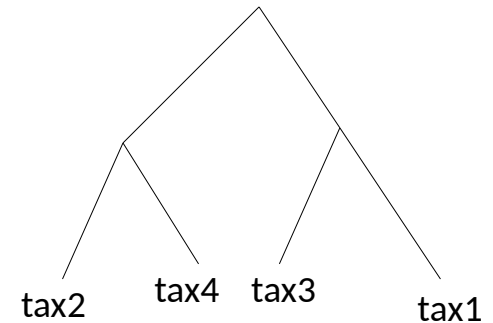
tax1: ACGG
tax2: AAGG
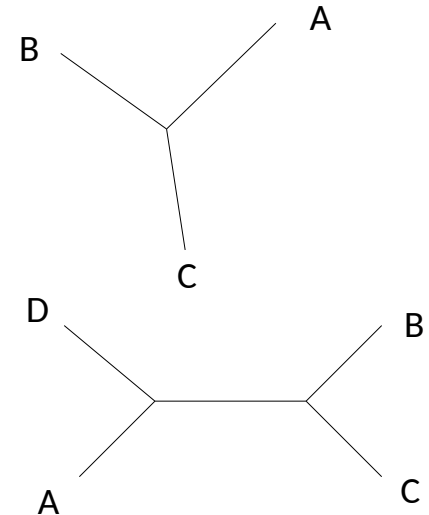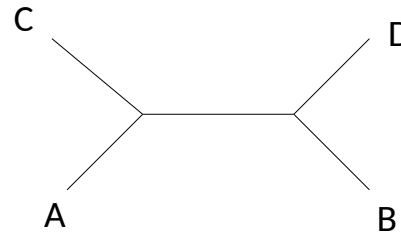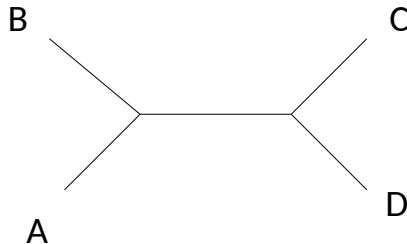tax3: AAGT
tax4: GAGG

input data =
aligned nucl.

phylogenetic inference

tax2   tax4 tax3   tax1

?

tax2   tax4   tax1 tax3

So many trees! Let's count them.

Fast growth of the number of binary tree topologies with *n* taxa:

- 1 unrooted binary tree topology with 3 taxa:

- 3 unrooted binary tree topologies with 4 taxa:

- 15 unrooted binary tree topologies with 5 taxa

- (2*n*-5)!! = 3 x 5 x 7 x 9 x … x (2*n*-5) unrooted binary tree topologies with *n* taxa

- (2*n*-3) x (2*n*-5)!! = (2*n*-3)!! rooted binary tree topologies with *n* taxa

So many trees!

- for n = 10 taxa, approx. 2 million unrooted binary trees

- for n = 20 taxa, approx. 2.2×10^{20} unrooted binary trees

- and then extra information is in the branch lengths!

- looking for "the right tree" is searching an infinite space.

- one has to devise strategies, either "exact" (algorithms) or including some amount of "sub-optimal guessing" (heuristics) to perform **tree inference**

Several families of methods for phylogenetic tree inference

- **distance matrix methods** try and devise a tree from an input being a matrix of pairwise distances between taxa

- **minimum evolution (ME) a.k.a. maximum parsimony (MP)** methods try and find the tree with minimal number of evolutionary events (character changes) along its branches

- **maximum likelihood (ML)** methods try and find the tree maximising the probability of the data observed on the leaves, under a certain evolutionary model

- **Bayesian** methods try and associate trees with posterior probabilities (of the model, i.e. the tree, having considered the data), and ultimately pick the tree maximizing that posterior probability

Molecular Evolution & Phylogenetics

Distance matrix methods

Distance matrix methods: pros and cons

- advantage: can be used on virtually any data (traits), as long as one knows how to calculate pairwise distances from them

- advantage: distance matrix methods are **fast**, accommodate even large numbers of taxa

- shortcoming: most often, calculated distances differ from patristic distances (sum of brach lengths on the path)

- shortcoming: loss of signal from the data, (complexity of a taxa summarised into its pairwise distances with others)

- remark: some methods produce constrained trees (e.g. respecting the molecular clock)

Simplest distance matrix method: UPGMA

UPGMA (Sokal&Michener 1958): "Unweighted Pair Group Method using arithmetic Averages"

- progressively grouping taxa into clusters, building the tree bottom-up (from leaves to root)

- the distance between two clusters is the average distance between pairs of sequences from each cluster

- initially as many clusters as individual sequences. Place corresponding nodes at height 0.

- iteratively choose the two clusters C_i and C_j with minimum distance d_{ij}, define a new cluster C_k containing all sequences from clusters C_i and C_j, place the corresponding node at height $d_{ij}/2$ and link it with the two nodes representing C_i and C_j

- continue until the root is reached, when all clusters are connected.
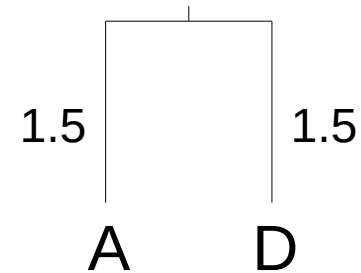
UPGMA at work: 1/3

	A	B	C	D
A	-	9	7	3
B		-	2	6
C			-	5
D				-

```
        ┌───────┐
   1.0  │       │  1.0
        │       │
        B       C
```
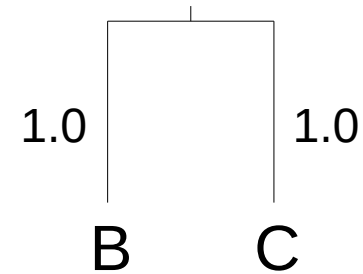
- new cluster BC
- d(BC, A) = (9+7)/2 = 8
- d(BC,D) = (6+5)/2 = 5.5

UPGMA at work: 2/3

	A	BC	D
A	-	8	3
BC		-	5.5
D			-

1.5   1.5

A   D

- new cluster AD
- d(BC, AD) = (8+5.5)/2 = 6.75

UPGMA at work: 3/3

From previous slide: d(BC, AD) = 6.75

2.375

1.875

1.0

1.0

1.5

1.5

B

C

A

D

d(BC, AD)/2 =
6.75/2 = 3.375
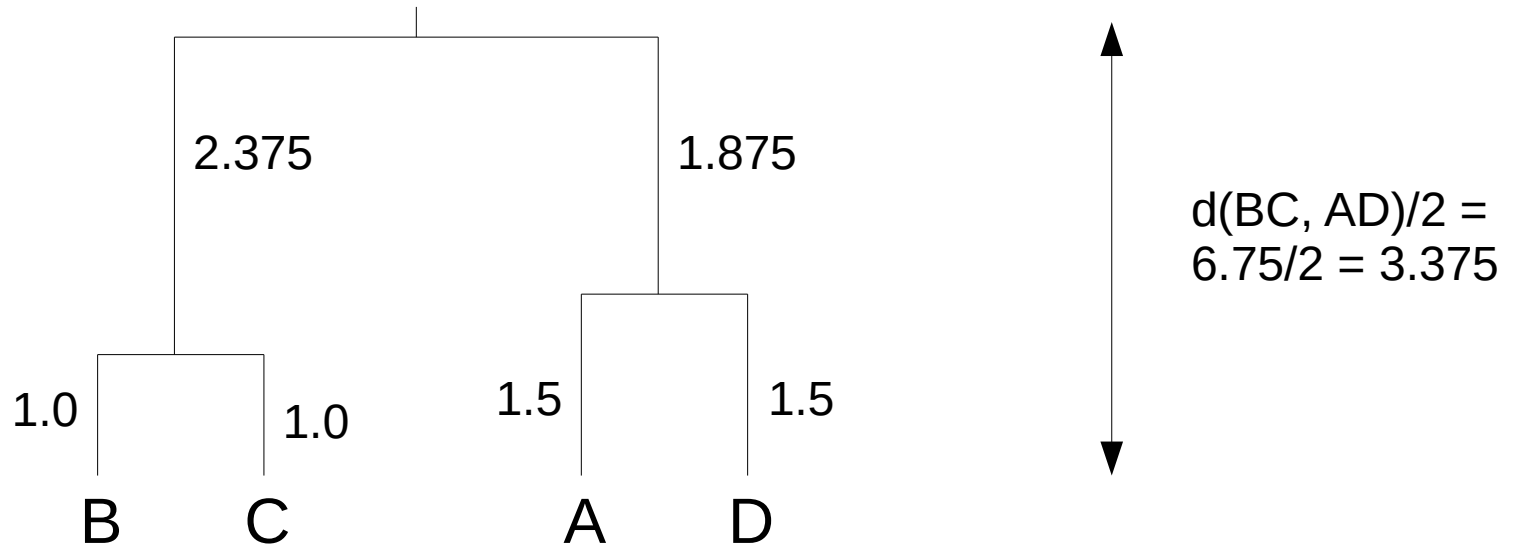
UPGMA warning: arithmetic averages

The distance between two clusters is the average distance between pairs *(i,j)* of taxa with *i* in one cluster and *j* in the other.

For instance, in the prep. of cluster ABC from clusters AB and C, we use a **weighted average**:

d(ABC, E) = 2/3 * d(AB,E) + 1/3 * d(C,E)

Properties of the UPGMA tree inference

- trees reconstructed by the UPGMA algorithm are **ultrametric** (same length from the root to any leaf)

- ⇒ if the **molecular clock hypothesis** holds (constant mutation rate), then UPGMA reconstructs the true tree.
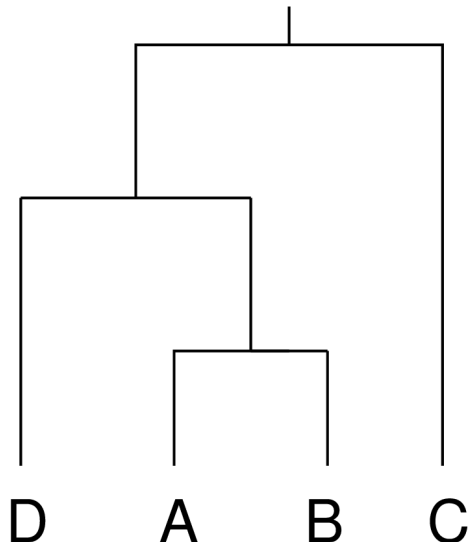
An ultrametric tree
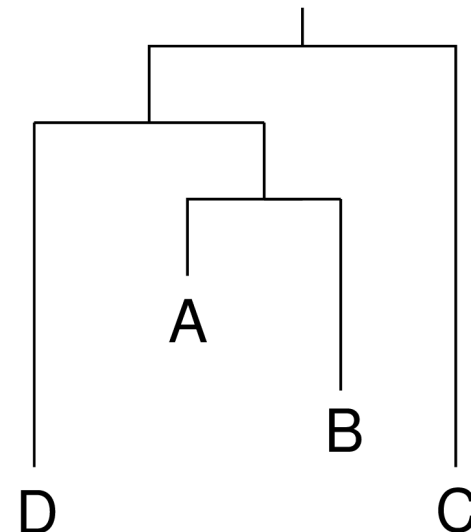
A non-ultrametric tree

Additivity and the NJ algorithm

In case the distance between any pair of leaves is equal to the length of the path between them, the tree is said to be **additive**.

- UPGMA always yields an additive tree, but the final distances do not always equal the original distances from the distance matrix: if additivity holds but not ultrametricity, then UPGMA will not reconstruct the correct tree

- Saitou & Nei (1987) and Studier & Keppler (1988) developed a method that always reconstructs the correct tree if the set of distances is additive: the **Neighbour-Joining (NJ) algorithm**

- NJ more complex than UPGMA, using corrected (modified) distances
- most popular distance method, used e.g. to build quick guide trees for multiple sequence alignment algorithms
- refined implementation: BioNJ (Gascuel 1997)

Distance methods: a summary

- two main algorithms: UPGMA and NJ

- methods suitable for a wide range of data: only need pairwise distances between taxa

- these algorithms are fast: (n − 1) internal nodes created one by one, i.e. linear complexity

- greedy algorithms: no costly exploration of the set of all tree topologies

Rooting a tree with an outgroup

- UPGMA yields rooted trees

- NJ and other methods to be seen later produce unrooted trees. How to root them?

- in case the molecular clock property holds, one can choose for the root the midpoint on the longest path between any two leaves

- otherwise, one can root using an **outgroup**

- an outgroup is an extra taxon for which we know its MRCA (Most Recent Common Ancestor) with any taxon in the tree clearly predates the MRCA of all other taxa in the tree

- ⇒ the node where it connects with the rest of the tree is the inferred root

Maximum Parsimony (MP)

a.k.a

Minimum Evolution (ME)

Minimum Evolution Principle

- Minimum Evolution principle goes for the evolutionary history accounting for (i.e. "explaining") the extant sequences with **as few evolutionary events** (character substitutions, deletions and insertions) **as possible**.

- Minimum count of events: "Parsimony cost" or "Parsimony steps"

- similar to trying and get a tree with smallest total branch length

- a principle (dogma), not a scientific truth

- beware! true number of substitutions ≥ inferred parsimony cost (think of the possible evolutionary history A → T → A)

- be cautious with MP phylogenies on very divergent data

Different trees, different parsimony costs


cost=2 cost=2 cost=1
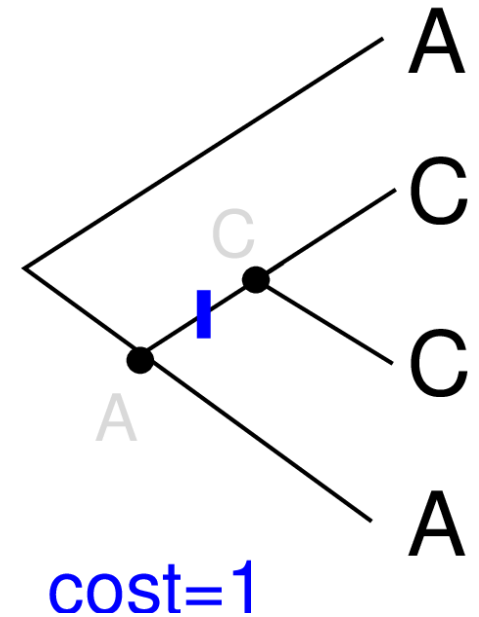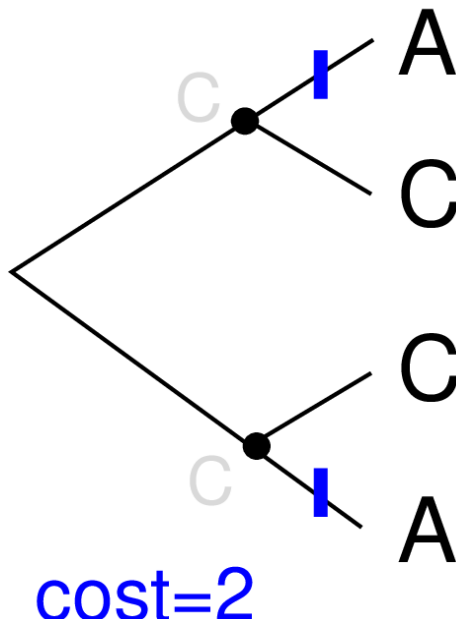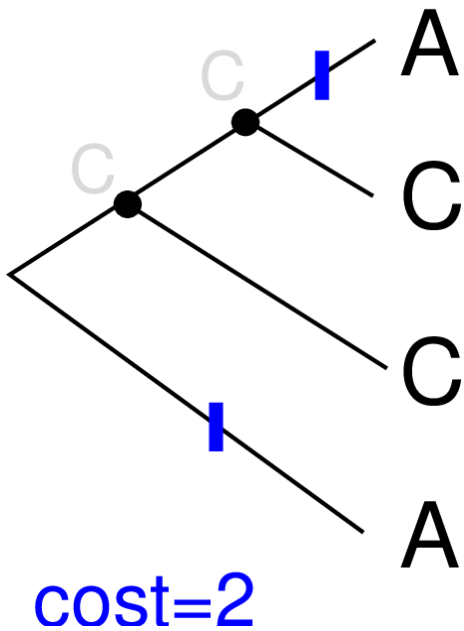
Different sites "support" different trees

- a site is a column in a multiple sequence alignment (one aligned character per taxa)

- phylogeny built from a Multiple Sequence Alignment (MSA)
  $\Rightarrow$ must take into account all sites

- total parsimony cost = $\sum\limits_{site\ i} cost_i$ ("independent sites" paradigm)
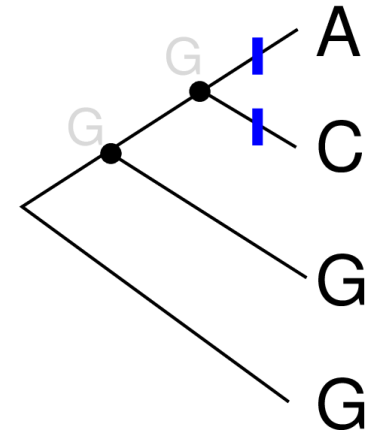
Some sites "don't choose" (1/2)

Some sites are **uninformative** from the viewpoint of MP: they are useless in deciding about the most parcimonious tree.
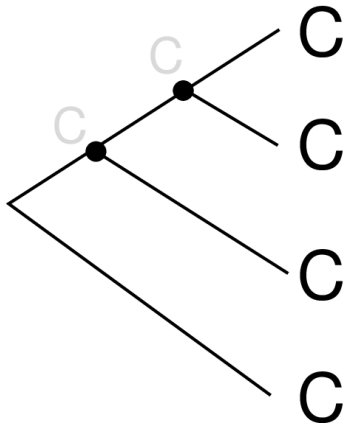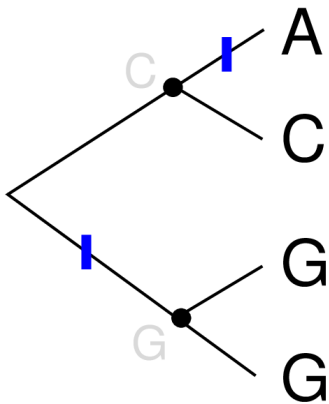

conserved site:
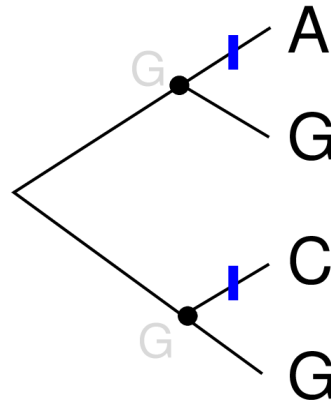cost = 0 for all trees
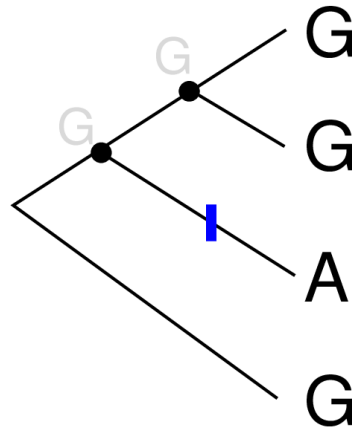
(n–1) consensus:
cost = 1 for all trees

(n–2) consensus
with unique chars:
cost = 2 for all trees
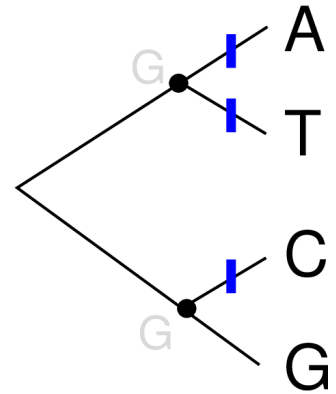
Some sites "don't choose" (2/2)

More **uninformative** sites:

(n−2) consensus
with unique chars:
cost = 2 for all trees

(n−2) consensus
with unique chars:
cost = 2 for all trees

no two pairs of
identical characters.
cost = 3 for all trees

Informative site for Maximum Parsimony

A site (column from an alignment) is **informative for MP inference methods** if and only if it contains at least **two different characters in at least two copies each.**

⇒ MP methods "ignore" potentially many sites

Core to Maximum Parsimony analysis: Fitch's algorithm

- published by Fitch in 1971

- fast algorithm to calculate the parsimony score for any given site on any **given tree**.

- gives both the maximum parsimony score and a most parsimonious history (ancestral characters) in linear time.

- **doesn't search the space** of the different phylogenies: mere calculation on a given tree

Fitch's algorithm, step by step

- let s = 0 be the score counter initialization
- let c_k be the set of "acceptable" characters on node k
- for leaves k, c_k is the singleton containing the character seen in the alignment

- moving up from leaves to root (post-order traversal), when k is a node having sons i and j, let $c_k = c_i \cap c_j$ if that intersection is not empty, otherwise $c_k = c_i \cup c_j$ and s = s + 1

- once the root is reached, s is the total parsimony score and the "acceptable sets" on the nodes define most parsimonious histories
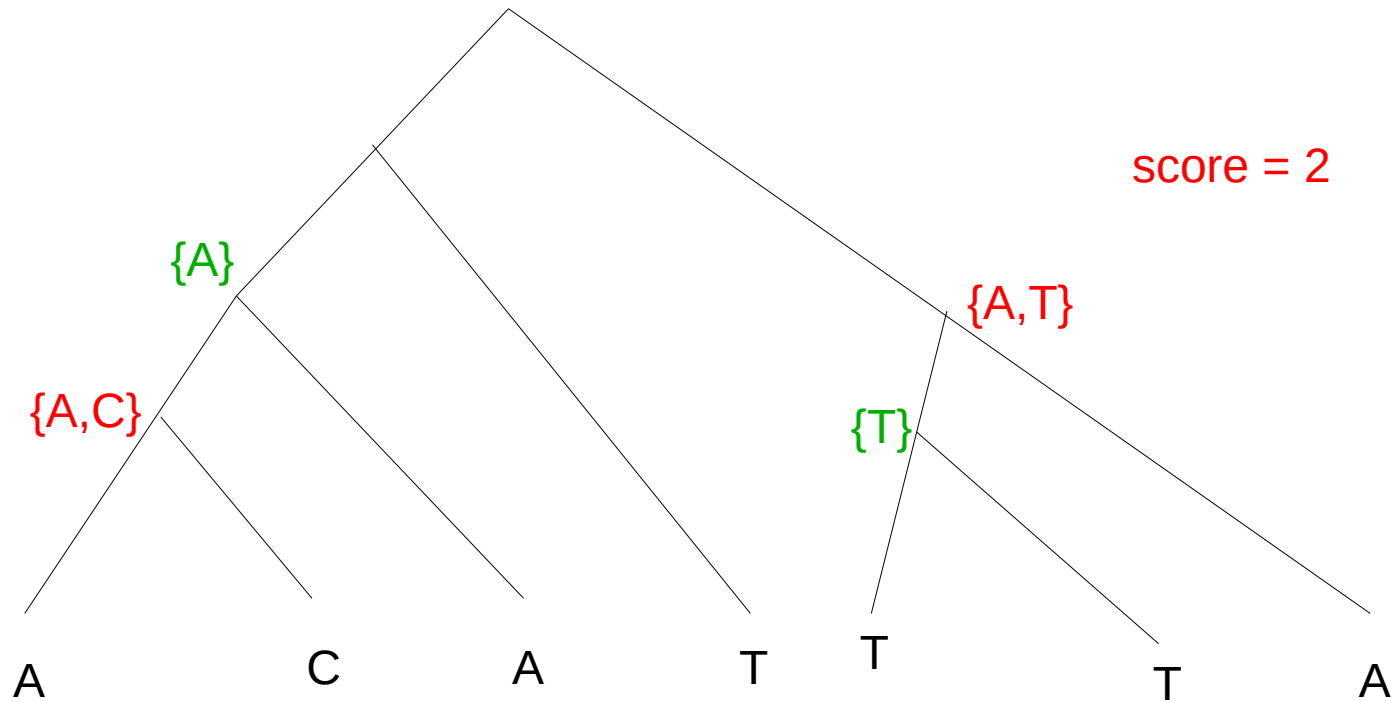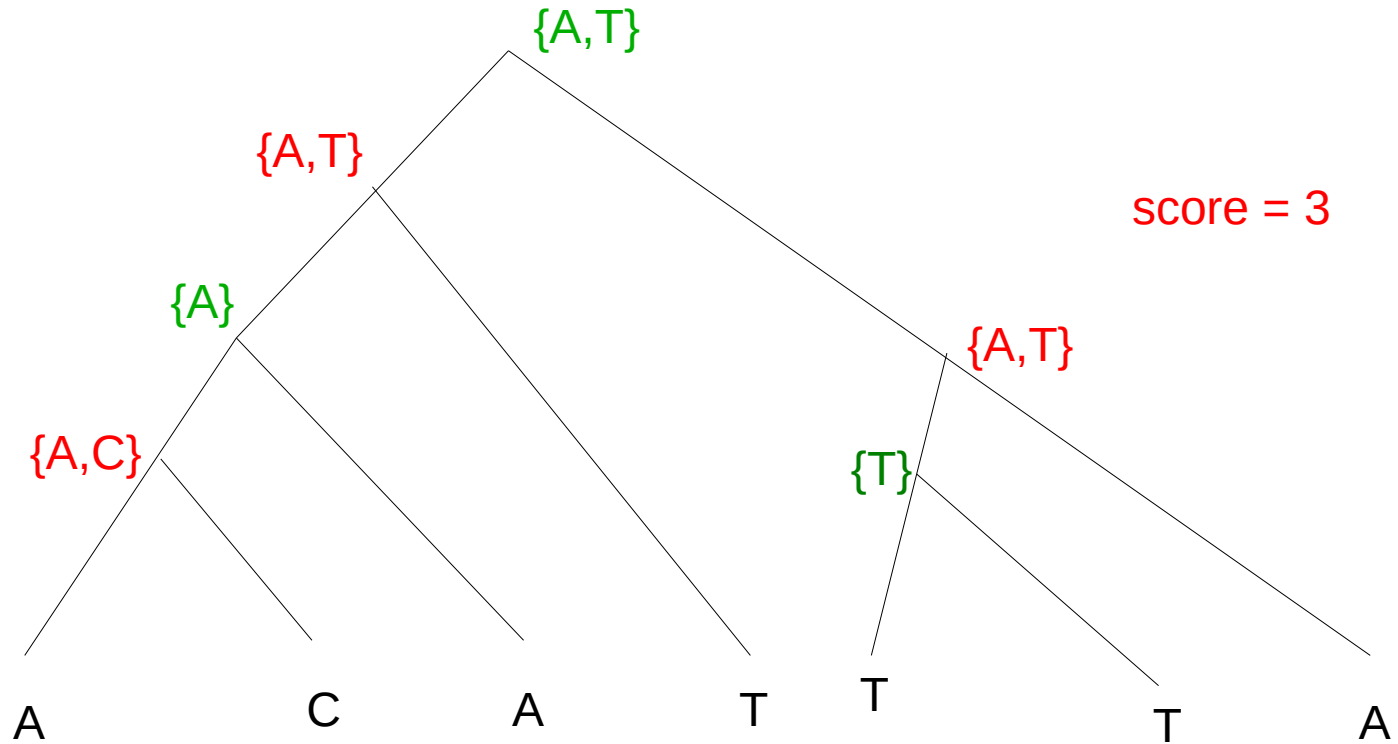
Fitch's algorithm: example application
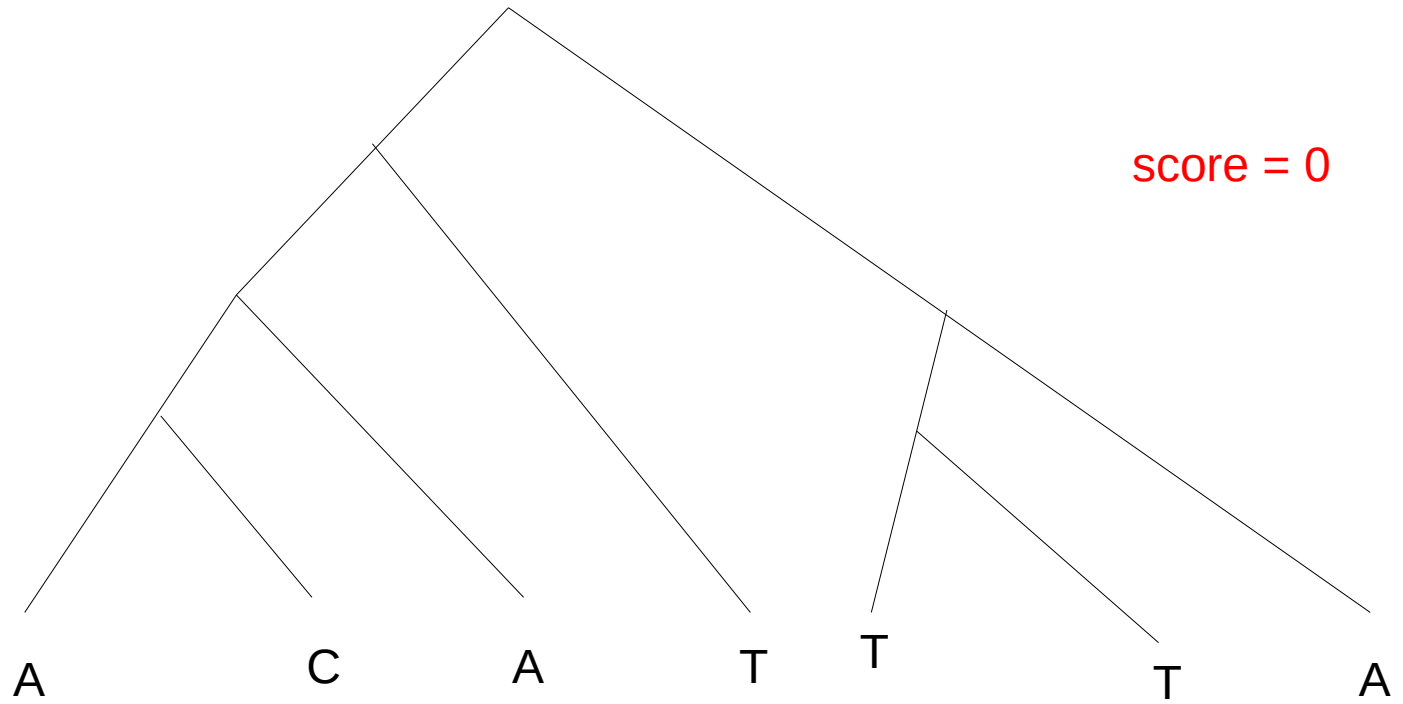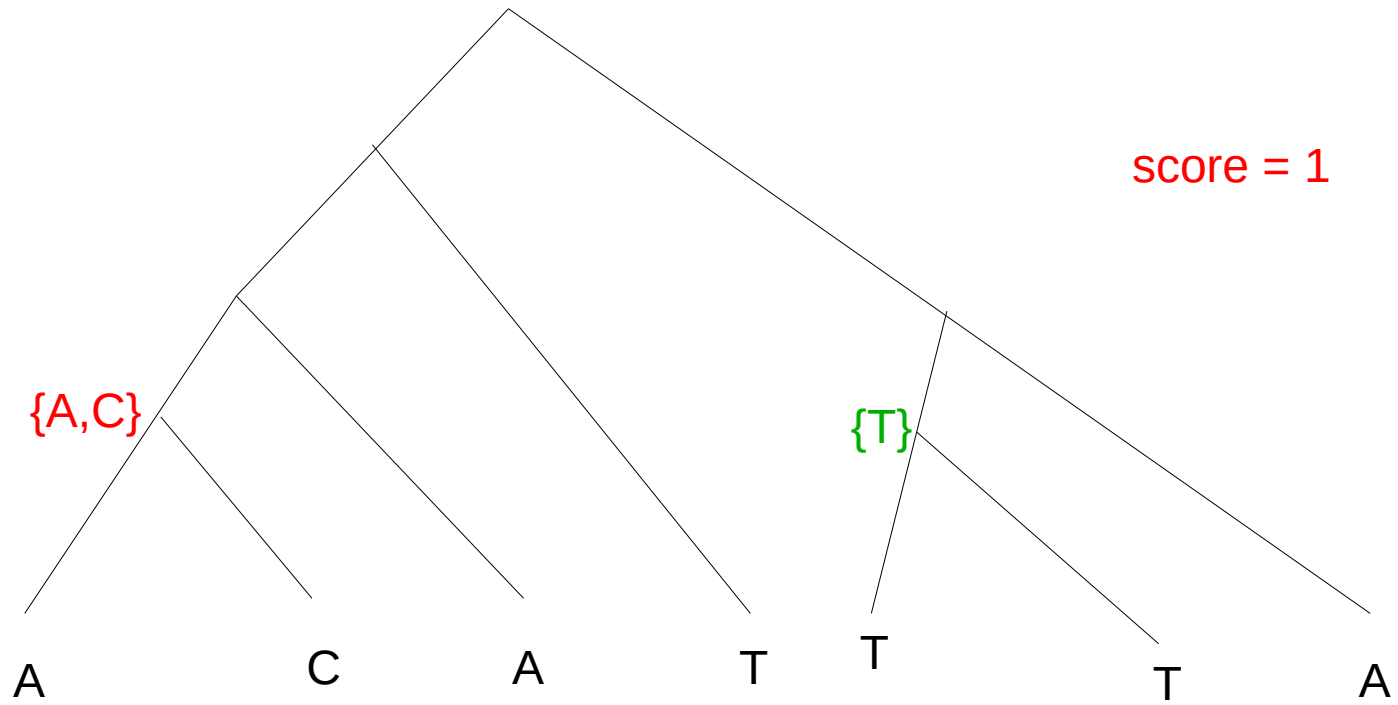
score = 0

A   C   A   T   T   T   A

Fitch's algorithm: example application

score = 1

{A,C}

{T}

A   C   A   T   T   T   A

Fitch's algorithm: example application

score = 2

{A}

{A,T}

{A,C}

{T}

A   C   A   T   T   T   A

Fitch's algorithm: example application

score = 3

Fitch's algorithm: a most parsimonious history

score = 3
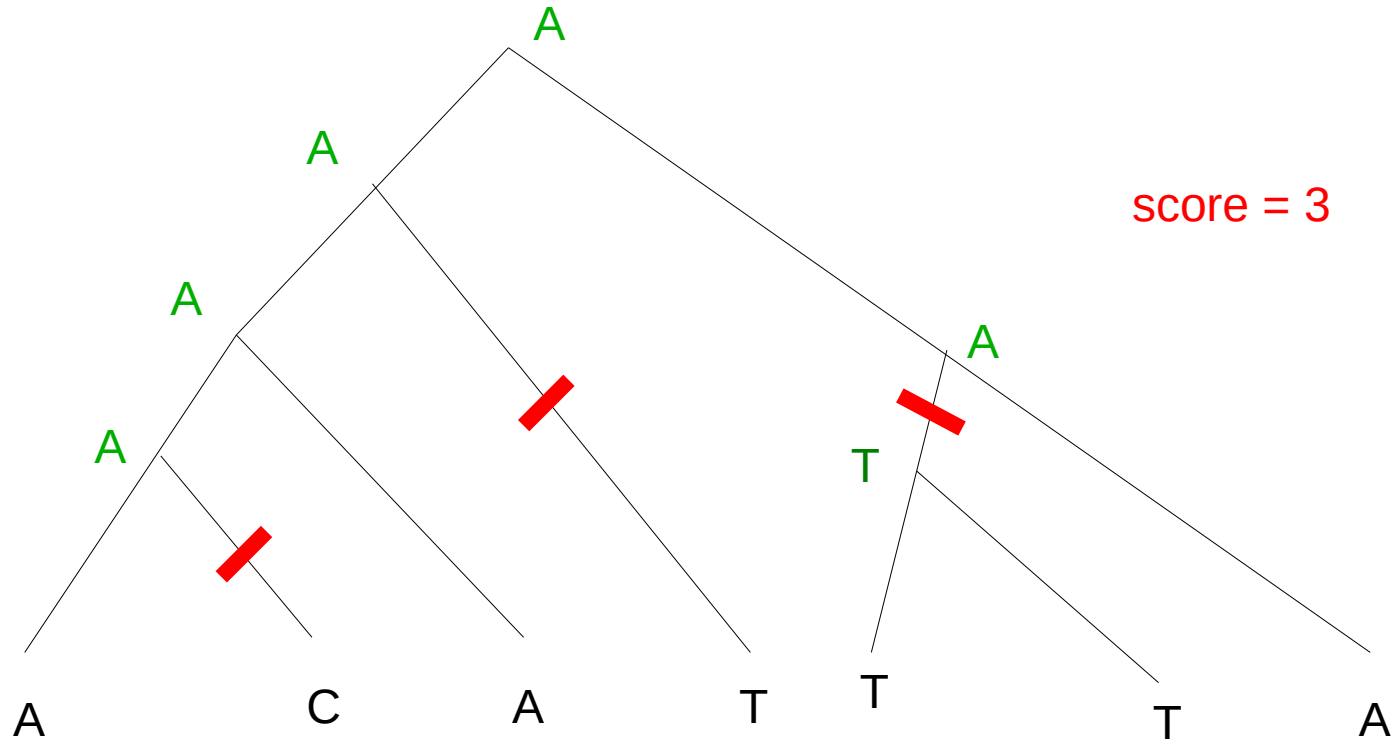
Programme for the next session...

- heuristics for an efficient travel in the space of trees

- the likelihood of a tree

- the Maximum Likelihood framework

- Bayesian methods

- assessing the quality of a tree: bootstrap method and local confidence statistics