

Need to add more information

Can we use some property that does not depend on a molecular clock assumption. The answer is, of course, yes.

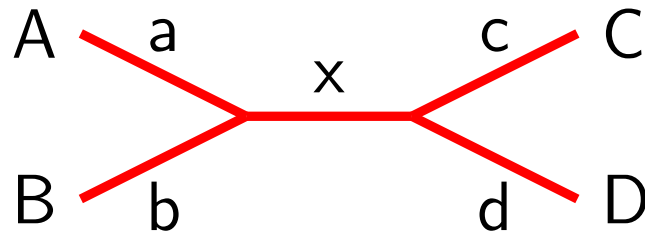
Need to add more information

Can we use some property that does not depend on a molecular clock assumption. The answer is, of course, yes.

One (of several) is the four point or neighbor relationship.

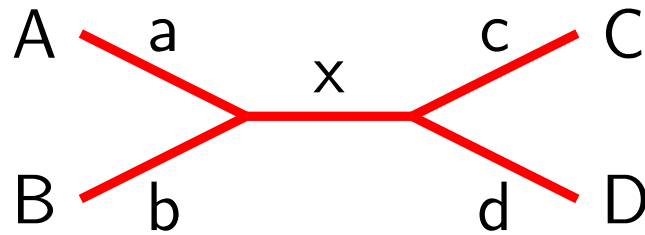
Neighbor relationship

Consider a four taxa, unrooted tree.



Neighbor relationship

Consider a four taxa, unrooted tree.



Then it must be true that,

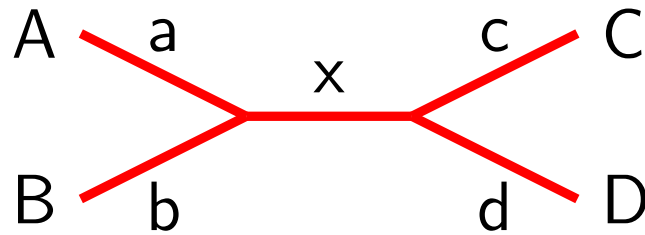
$$d_{AB} + d_{CD} < d_{AC} + d_{BD}$$

and

$$d_{AB} + d_{CD} < d_{AD} + d_{BC}$$

Neighbor relationship

Consider a four taxa, unrooted tree.



Then it must be true that,

$$d_{AB} + d_{CD} < d_{AC} + d_{BD}$$

and

$$d_{AB} + d_{CD} < d_{AD} + d_{BC}$$

because these are both equal to
 $a + b + c + d < a + b + c + d + 2x$

Neighbor relationship

So all that is required is to consider all three possible neighbor relationships among four taxa.

If you have taxa x_1 through x_4 then

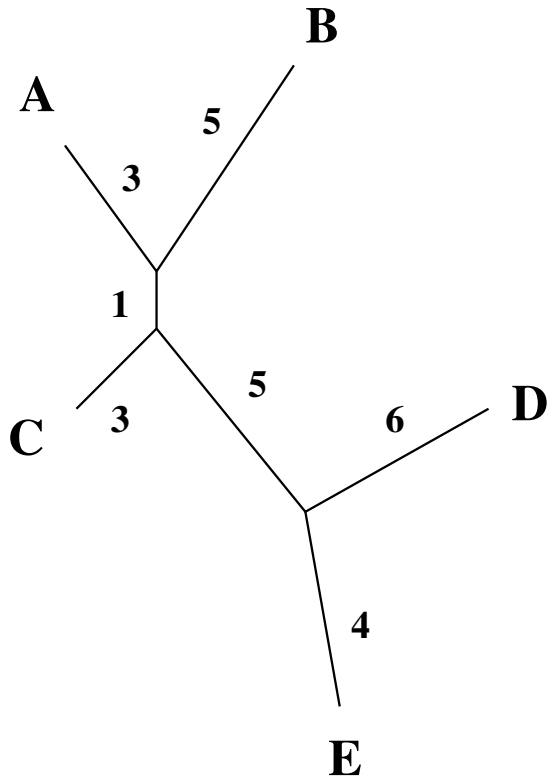
Either x_1 and x_2 are neighbors,
or x_1 and x_3 are neighbors,
or x_1 and x_4 are neighbors.

Because of the neighbor relationship, one of these pairs will have a shorter distance than the other two (ie: be true neighbors).

A popular method that implements this method is the “Neighbor-joining method” of Saitou and Nei (1987).

Neighbor relationship

For the previous example the NJ method gives the tree as ...



Why the difference in how the tree is drawn?

Neighbor relationship

This method does not assume a constant rate of change.

Therefore there is no “molecular clock” .

Neighbor relationship

This method does not assume a constant rate of change.

Therefore there is no “molecular clock” .

Hence there can be no information about the location of the root without the addition of an outgroup.

Neighbor joining

- The topology that gives the least total branch length at each step is preferred
- Does not assume a constant rate of evolution
- Leads to an unrooted tree
- A modified distance matrix is constructed in order to adjust for differences in the rate of evolution of each taxon

Neighbor joining

Formally this method is implemented by modifying the matrix according to the overall net difference in species i from all other taxa.

$$r_i = \sum_k d_{ik}$$

Rate corrected matrix:

$$M_{ij} = d_{ij} - (r_i + r_j)/(n - 2)$$

where n is the number of taxa.

Neighbor joining

The taxa will be joined first with the smallest M_{ij} and the branch lengths to the new node, u , are defined as

$$l_{iu} = d_{ij}/2 + (r_i - r_j)/(2n - 4)$$

$$l_{ju} = d_{ij} - l_{iu}$$

at the next step

$$d_{ku} = (d_{ik} + d_{jk} - d_{ij})/2$$

Distance methods

Fitch and Margoliash, 1970 Science 155:279

- Pairwise clustering algorithm
- Yields an unrooted tree
- Unlike the UPGMA and the Neighbor Joining methods it does not proceed by adding taxa one at a time
- The Fitch Margoliash method uses least squares method to choose the best tree.

Distance methods

Fitch and Margoliash, 1970 Science 155:279

- This method minimizes the sum

$$\sum (d - d')^2 / d^2$$

where d is the observed distance and d' is the distance implied by the tree (the expected).

Least squares – example

	Observed		
	A	B	C
A	-	2.1	4.0
B	-	-	3.5
C	-	-	-

Consider two trees with branches of unit length (hypotheses)

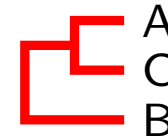


Expected

	A	B	C
A	-	2	4
B	-	-	4
C	-	-	-

$$\frac{(2.1 - 2)^2}{2.1^2} + \frac{(4 - 4)^2}{4^2} + \frac{(3.5 - 4)^2}{3.5^2}$$

$$= 0.02$$



Expected

	A	B	C
A	-	4	2
B	-	-	4
C	-	-	-

$$\frac{(2.1 - 4)^2}{2.1^2} + \frac{(4 - 2)^2}{4^2} + \frac{(3.5 - 4)^2}{3.5^2}$$

$$= 1.09$$

Distance methods

- Distance methods have the advantage of being very fast
- Can be tested easily (bootstrap)
- When distances are not computed according to the correct model they can lead to incorrect values
- When using sequence data, the accuracy of these methods decreases as the number of substitutions increases requiring correction
- If evolutionary rates vary from site to site corrections are needed
- In the case of very large trees distance methods will perform poorly in comparison with other methods

Parsimony methods

- **A given tree is more likely if it requires less rather than more evolutionary changes in character states**
- **Find the tree that minimizes the number of evolutionary changes required**
- **Parsimony works best when changes are rare**

a → a

inference is clear

a → b

inference is clear

a → b → a

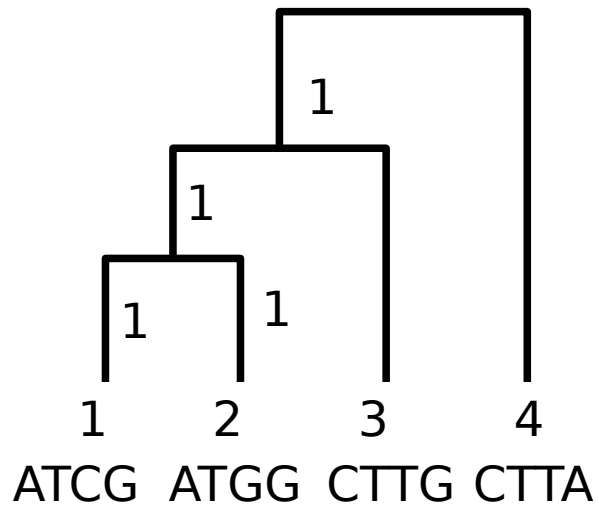
inference is lost unless "b" is known

Homoplasy

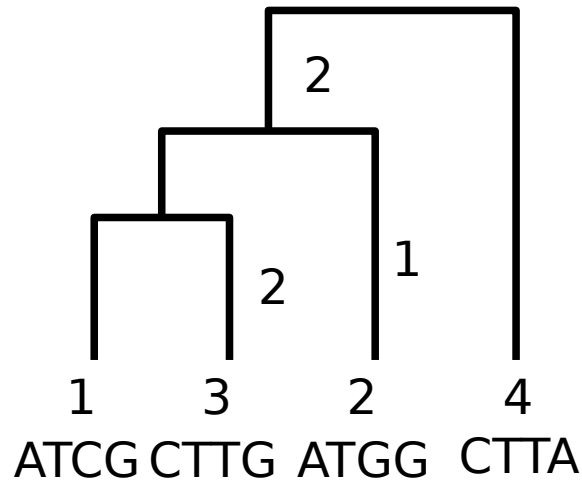
Parsimony methods

Building the tree corresponding to this alignment:

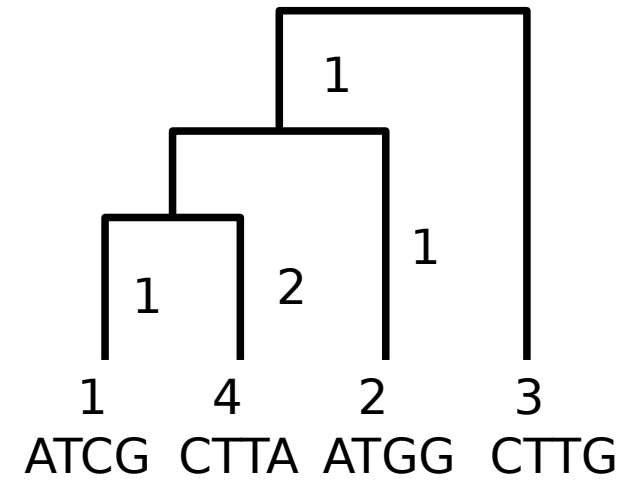
species 1 ATCG
species 2 ATGG
species 3 CTTG
species 4 CTTA



Score= 4



Score= 5

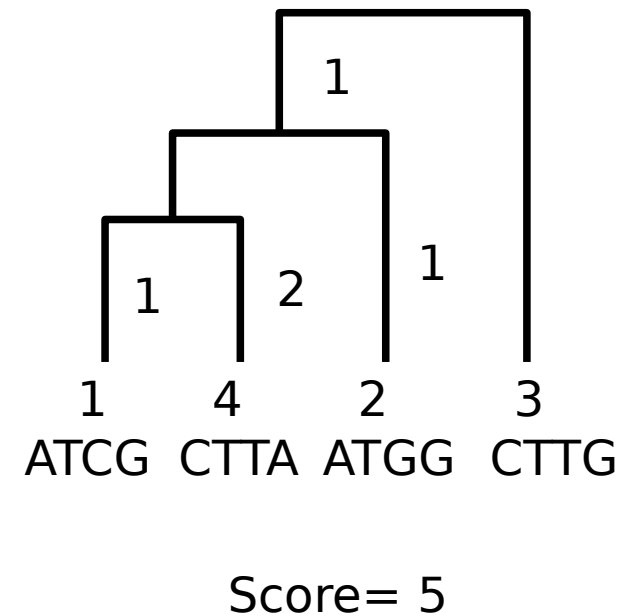
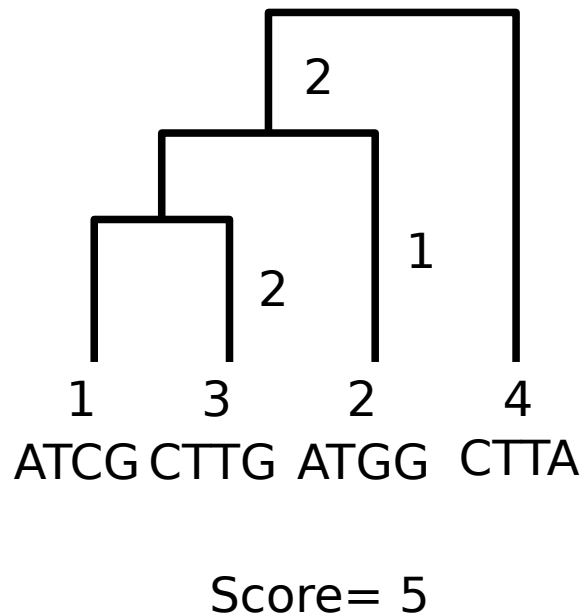
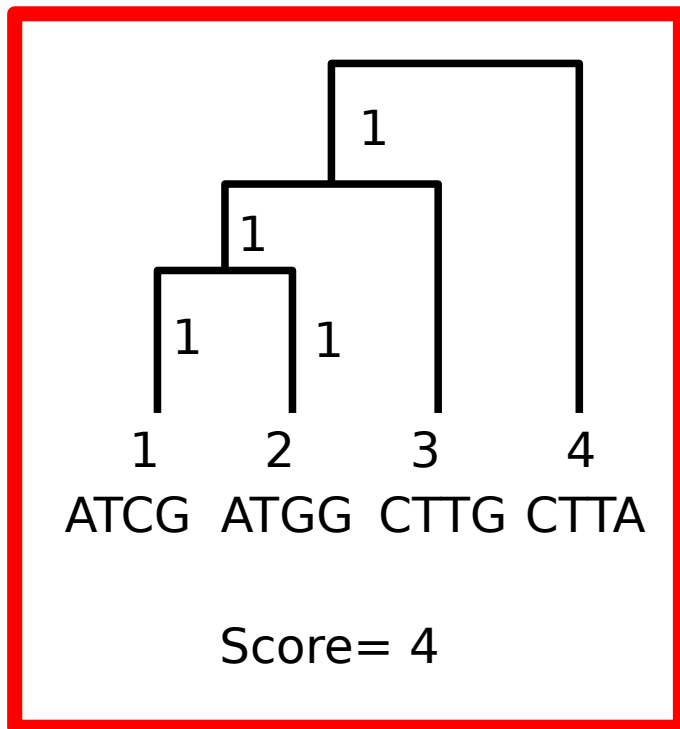


Score= 5

Parsimony methods

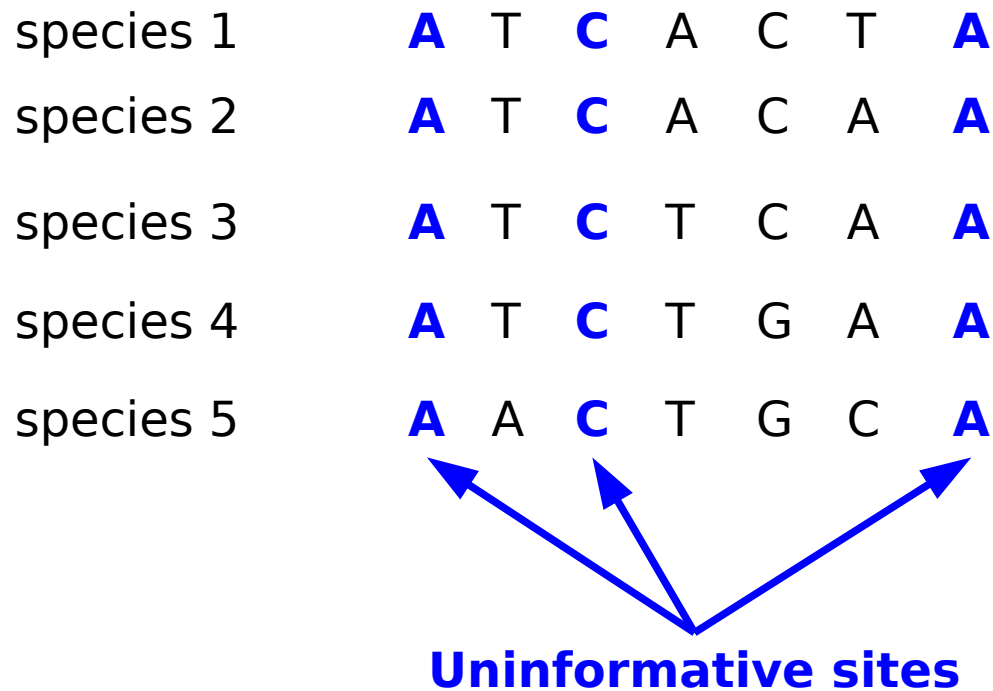
Building the tree corresponding to this alignment:

species 1 ATCG
species 2 ATGG
species 3 CTTG
species 4 CTTA



Parsimony methods

In order to infer the **topology** of the tree parsimony methods use only informative sites, discarding uninformative sites



Parsimony methods

In order to infer the **topology** of the tree parsimony methods use only informative sites, discarding uninformative sites

species 1	A	T	C	A	C	T	A
species 2	A	T	C	A	C	A	A
species 3	A	T	C	T	C	A	A
species 4	A	T	C	T	G	A	A
species 5	A	A	C	T	G	C	A

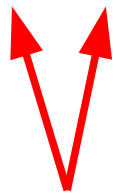
Uninformative sites



Parsimony methods

In order to infer the **topology** of the tree parsimony methods use only informative sites, discarding uninformative sites

species 1	A	T	C	A	C	T	A
species 2	A	T	C	A	C	A	A
species 3	A	T	C	T	C	A	A
species 4	A	T	C	T	G	A	A
species 5	A	A	C	T	G	C	A

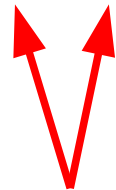


Informative sites

Parsimony methods

In order to infer the **topology** of the tree parsimony methods use only informative sites, discarding uninformative sites

species 1	A	T	C	A	C	T	A
species 2	A	T	C	A	C	A	A
species 3	A	T	C	T	C	A	A
species 4	A	T	C	T	G	A	A
species 5	A	A	C	T	G	C	A



Informative sites

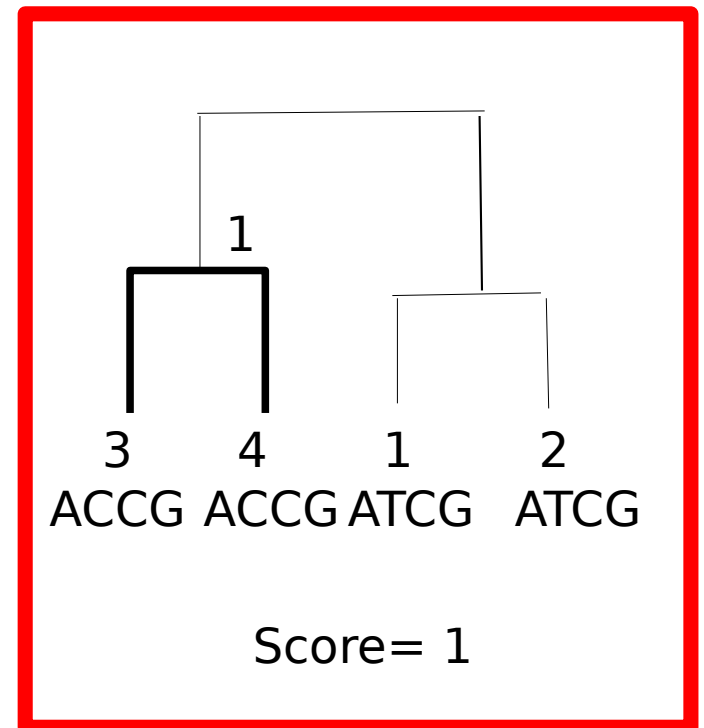
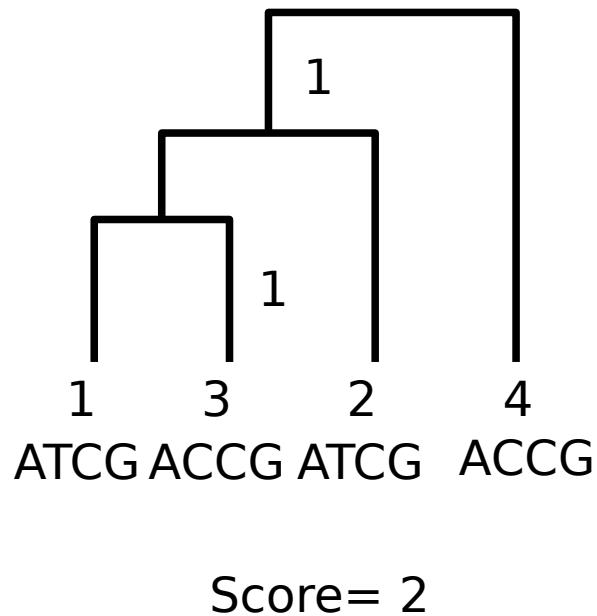
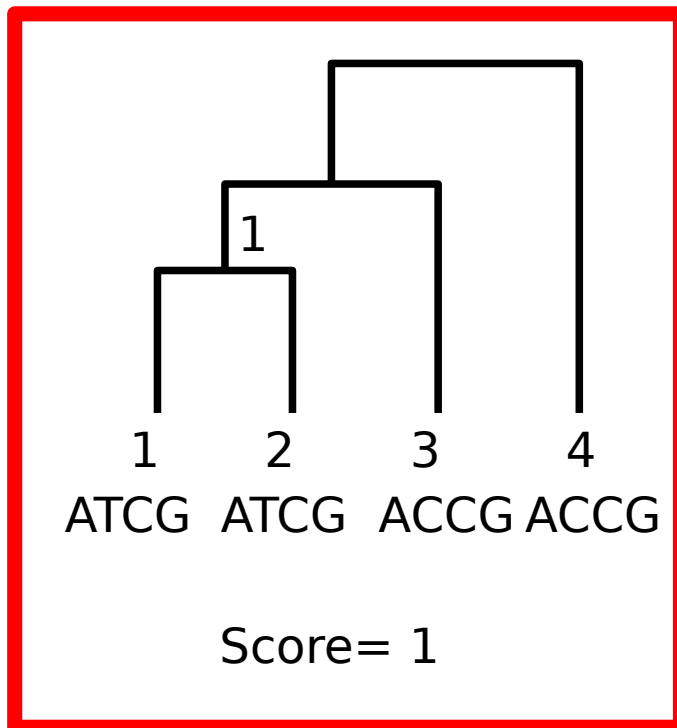
Do not forget that ALL the sites are used to compute the branch lengths

Parsimony methods

Building the tree corresponding to this alignment:

species 1 ATCG
species 2 ATCG
species 3 ACCG
species 4 ACCG

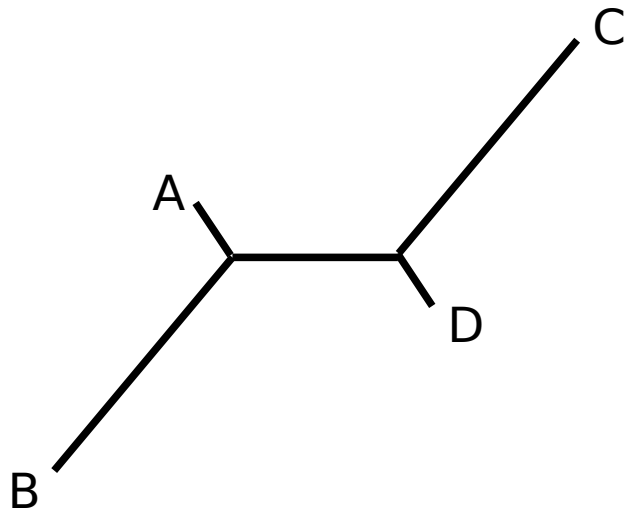
In some cases more than one parsimonious tree can be found



Parsimony methods

Long branch attraction

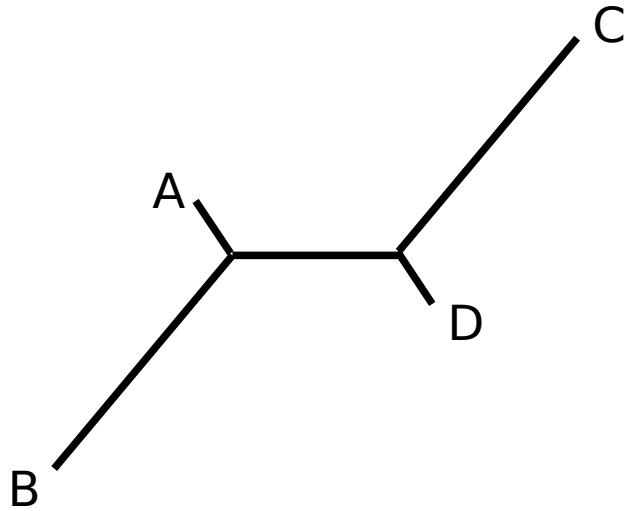
True tree



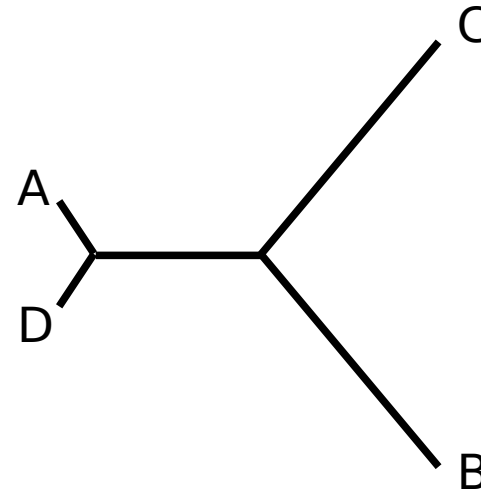
Parsimony methods

Long branch attraction

True tree



Maximum parsimony tree



Maximum parsimony

- The maximum parsimony result will converge to a maximum likelihood result when the rates of evolution along each branch are small
- Parsimony does not require constancy in the rates of change between branches if the number of substitutions per site is small. If the number is large serious errors will be made.
- If the number of substitution per site is small, a large proportion of the substitutions are autapomorphic and therefore uninformative (distance method may perform better in such a case)