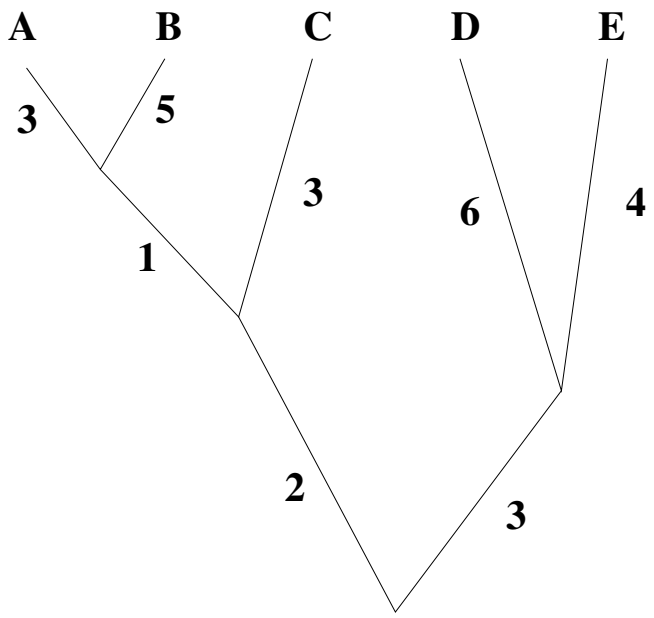
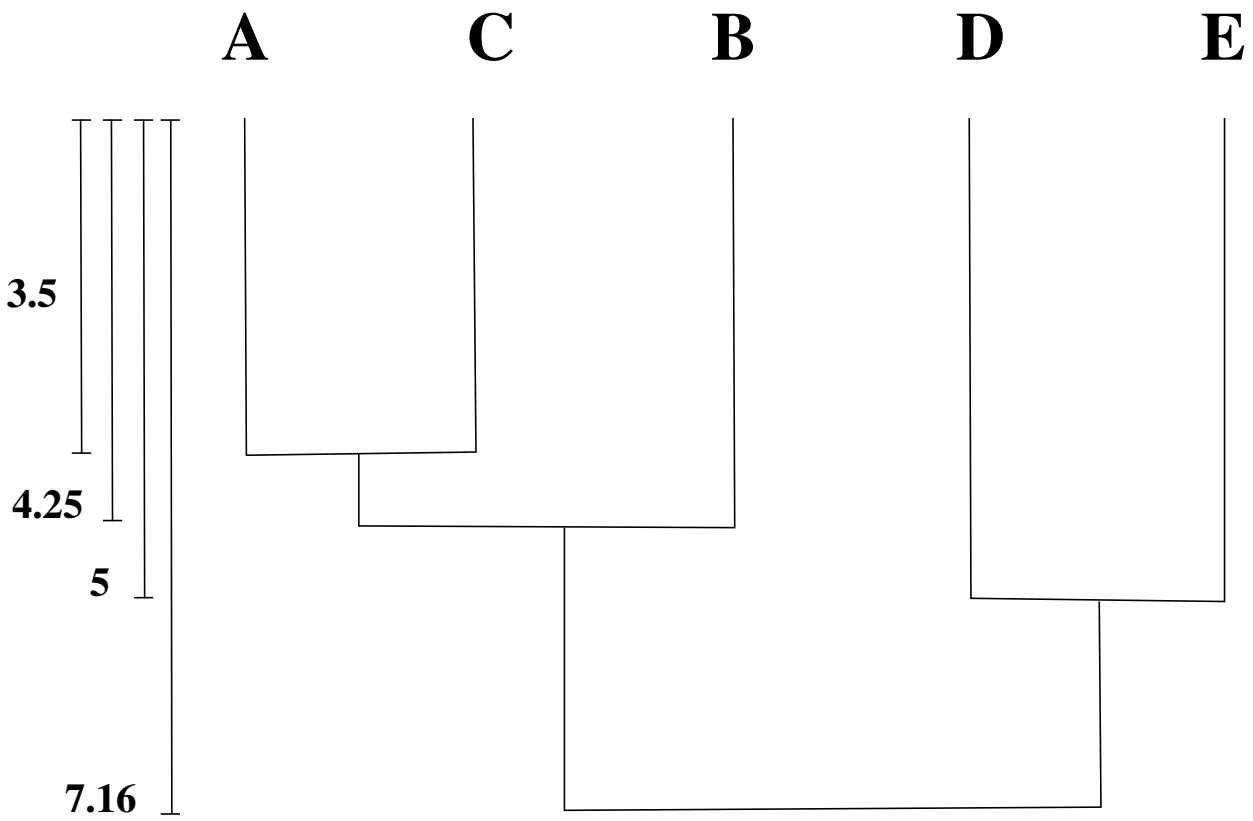
Example



	A	B	C	D
B	8	-	-	-
C	7	9	-	-
D	15	17	14	-
E	13	15	12	10

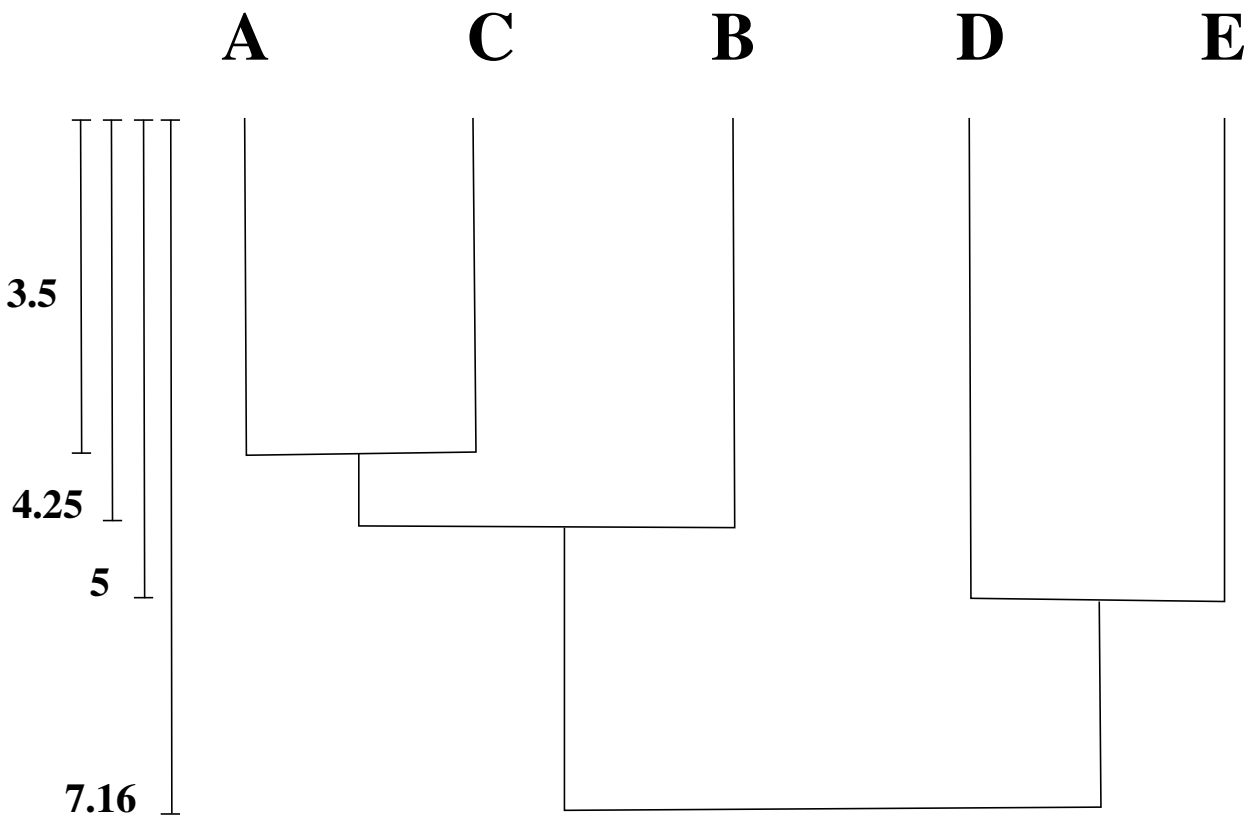
Example

Yields ...



Example

Yields ...



Hence have inferred the **wrong** tree !!

What went wrong?

Strict additivity of the distances does not hold. Why?

What went wrong?

Strict additivity of the distances does not hold. Why?

We have implicitly assumed a molecular clock holds.

This is not justified in this case.
(it seldom is)

Distance methods

- Very fast!!
- The reduction of the data to a single number (the genetic distance) leads to a loss of information
- UPGMA assumes that the branch lengths correlate with the phenotypic distance between taxa and should correspond to some proportional measure of time
    - Need to apply corrections (e.g. JC) for this last assumption
- UPGMA assumes a constant rate of evolution (molecular clock)
- UPGMA is prone to errors if the distances deviate (even slightly) from a clock-like tree

Clocklike vs. non-clocklike trees

A clocklike tree (or ultrametric tree) is a rooted tree in which the total branch length from the root to any tip is equal

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.

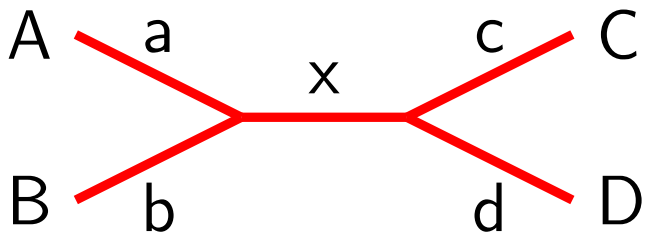
Need to add more information

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.

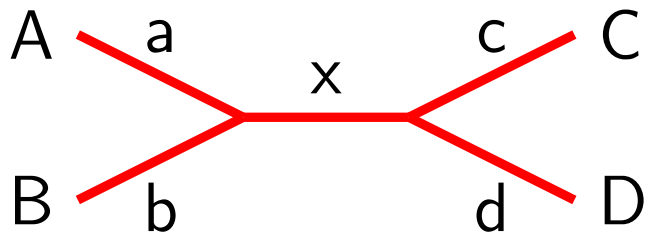
Neighbor relationship

Consider a four taxa, unrooted tree.



Neighbor relationship

Consider a four taxa, unrooted tree.



Then it must be true that,

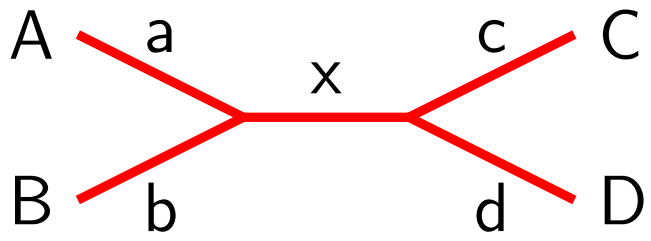
$$d_{AB} + d_{CD} < d_{AC} + d_{BD}$$

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

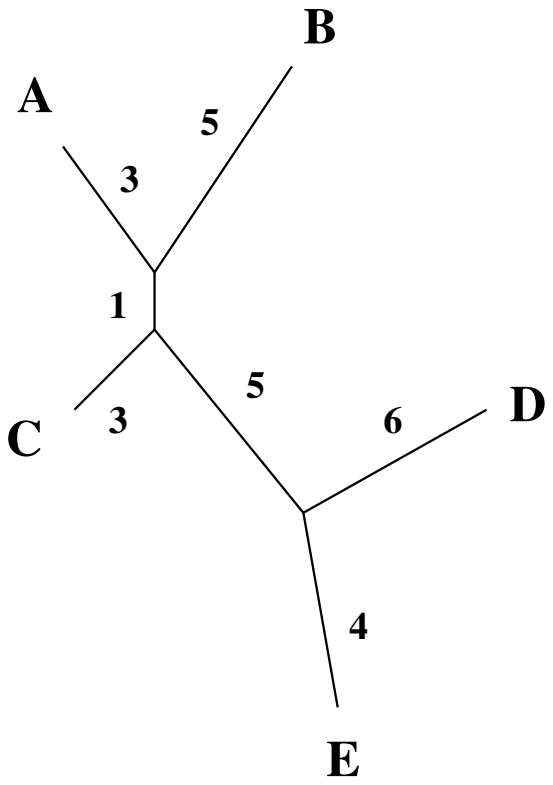
Either x_1 and x_2 are neighbors,
or x_1 and x_3 are neighbors,
or x_1 and x_4 are neighbors.

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 i 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

- 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