

# Bioinformatics 1--lecture 14

Trees

# Phylogenetic trees

What is a phylogenetic tree?

A model of evolutionary relationships -- common ancestors and speciation events.

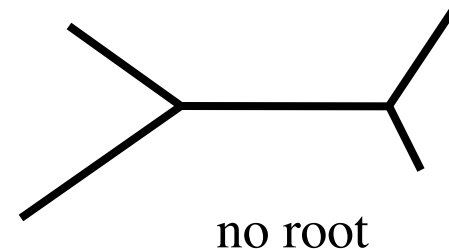
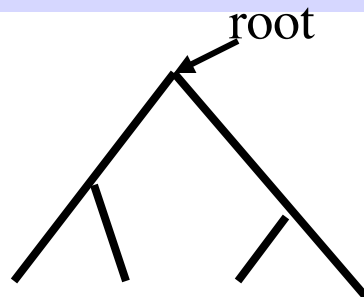
Why build phylogenetic trees?

To trace the branch order of "taxa" (taxon = a gene, a species, a population, etc.)

To understand the evolution of traits

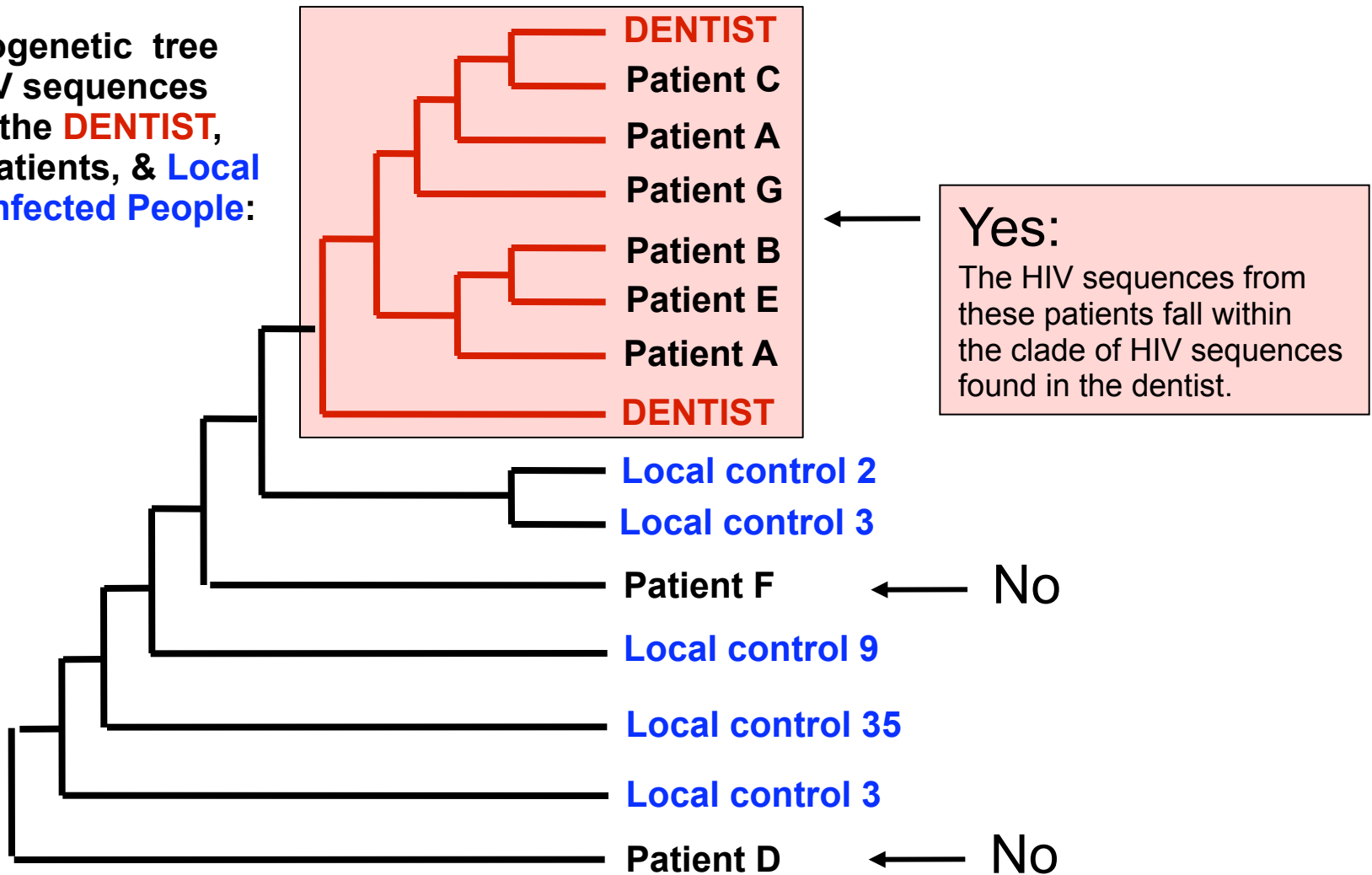
As part of a multiple sequence alignment algorithm

Trees can be  
"rooted" or  
"unrooted"



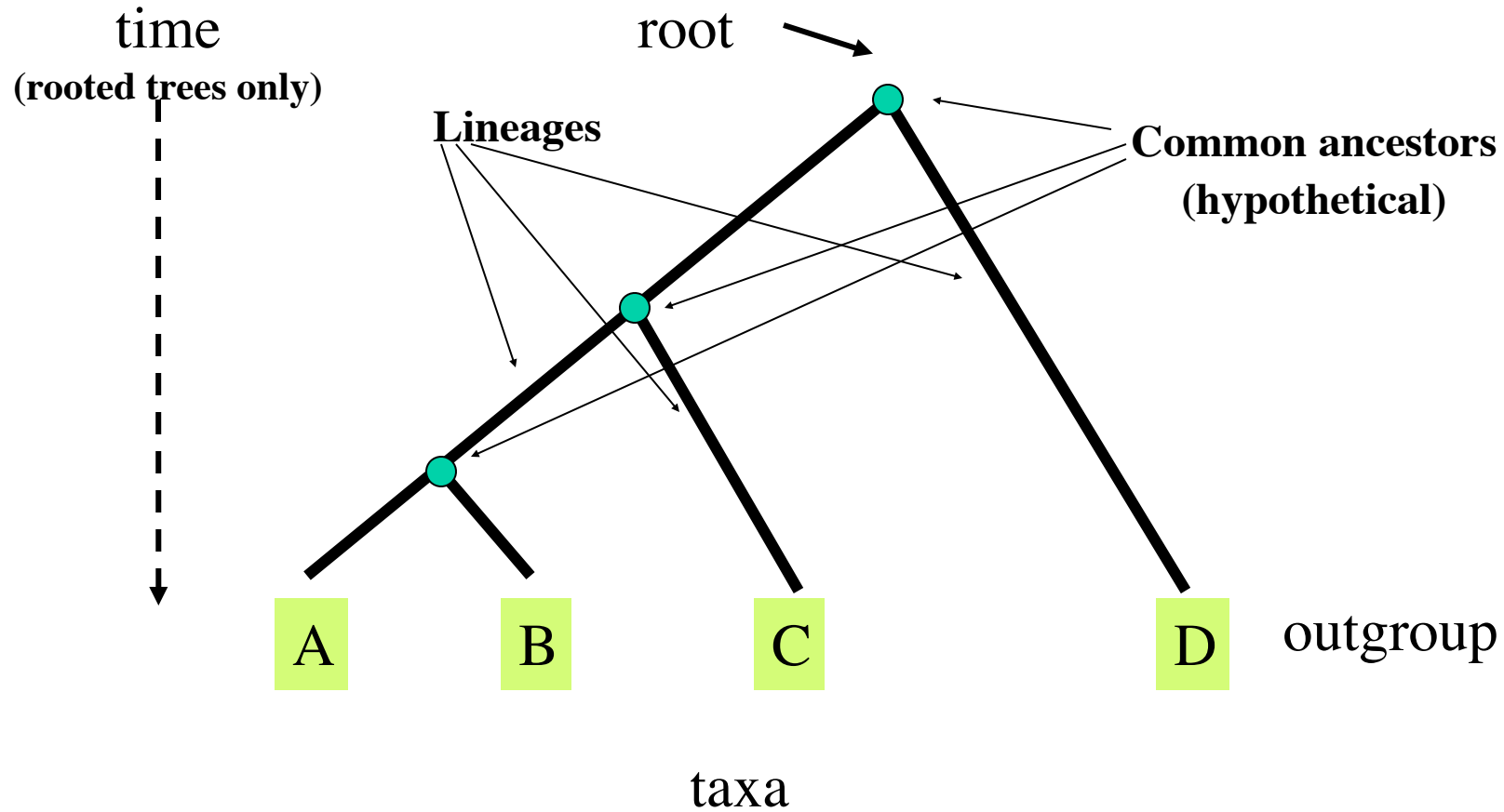
# Did the *Florida Dentist* infect his patients with HIV?

Phylogenetic tree of HIV sequences from the **DENTIST**, his Patients, & **Local HIV-infected People**:



From Ou et al. (1992) and Page & Holmes (1998)

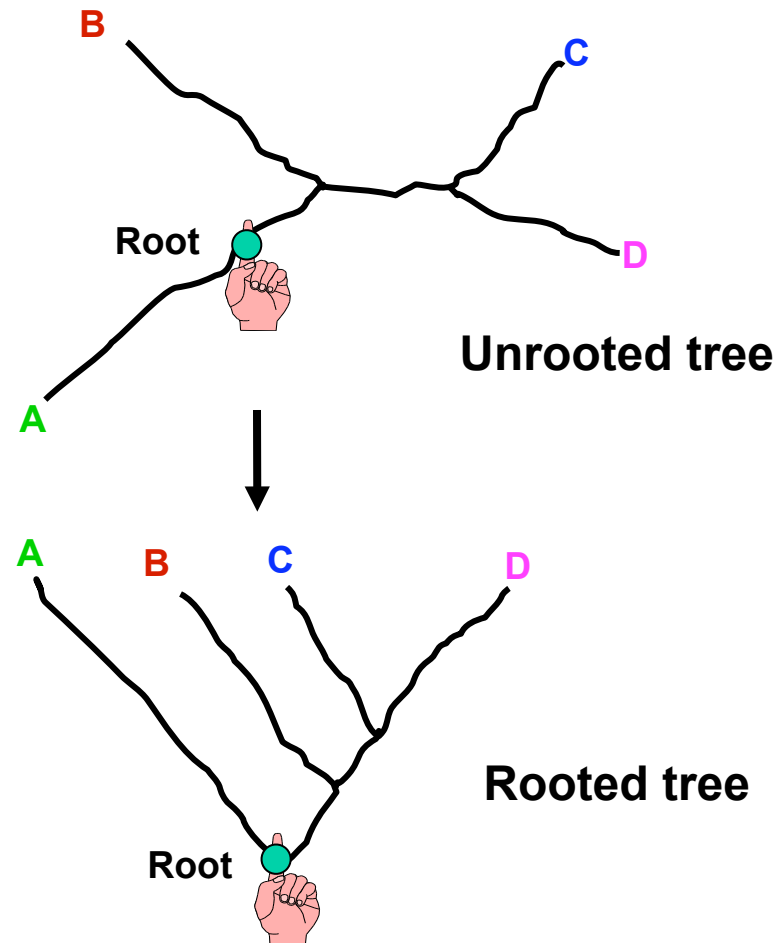
# Tree Terminology



**Taxa are observed species or genes.**

## Inferring evolutionary *relationships* between the taxa requires rooting the tree:

To root a tree mentally, imagine that the tree is made of string. Grab the string at the root ● and tug on it until the ends of the string (the taxa) fall opposite the root:

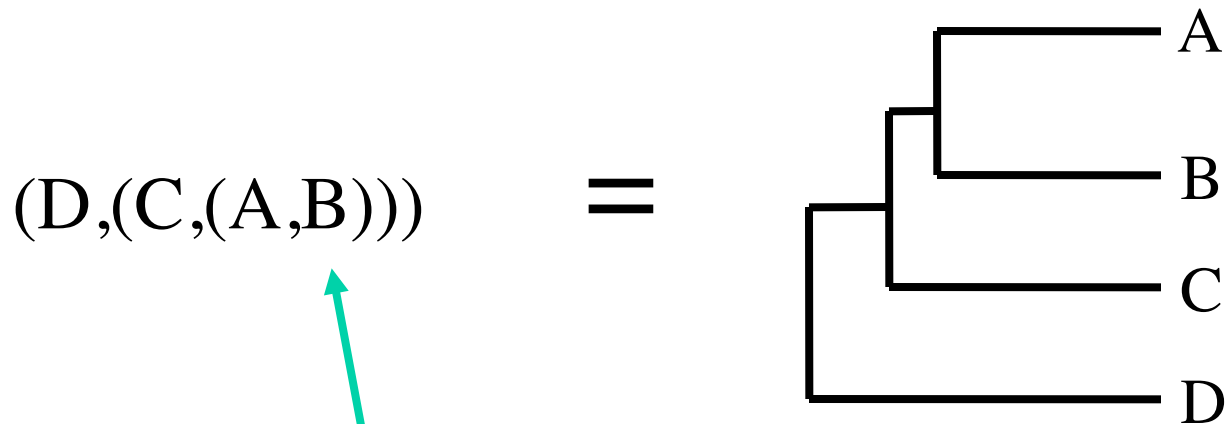


## Newick notation

( / ( / ( / ) ) )

Trees can be represented in plain text Newick or "parenthesis" notation.

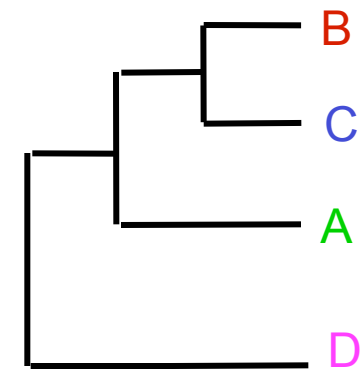
Each set of parentheses represents a branch-point (bifurcation), the comma separates left and right lineages.



Parenthesis notation can contain sequence labels too.

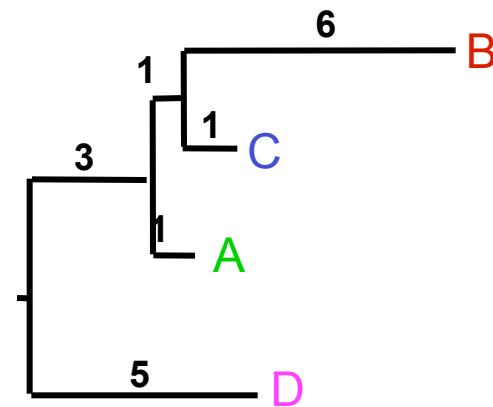
# Evolutionary time

Cladogram



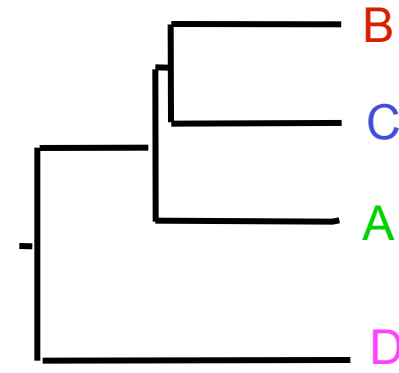
no meaning

Phylogram



genetic change

Ultrametric tree



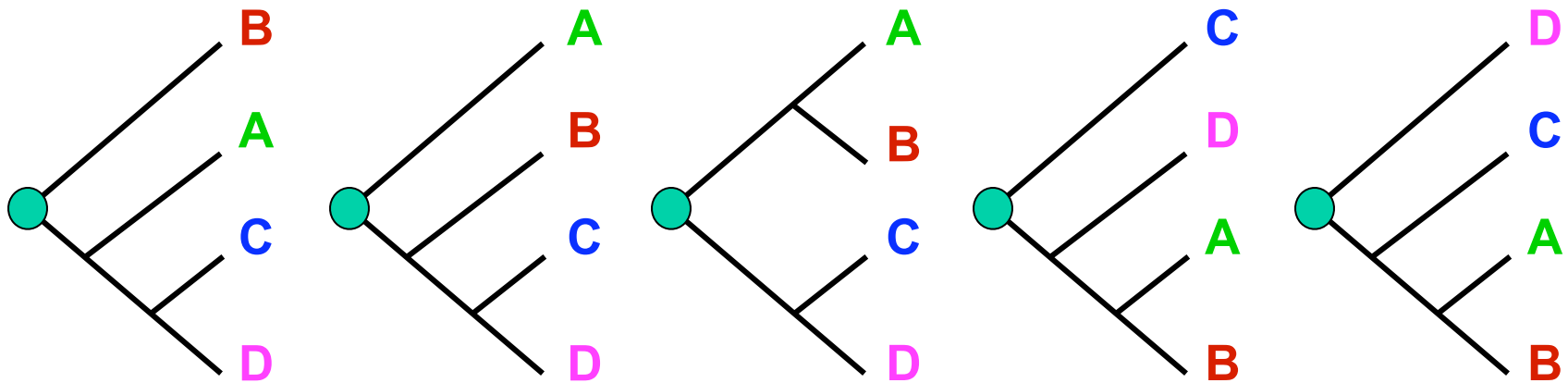
time

$(D:5,(A:1,(C:1,B:6):1):3)$

parenthesis (notation can have both labels and distances.

Where the tree is rooted changes its meaning.

*Each of these trees is possible by choosing a different root.*

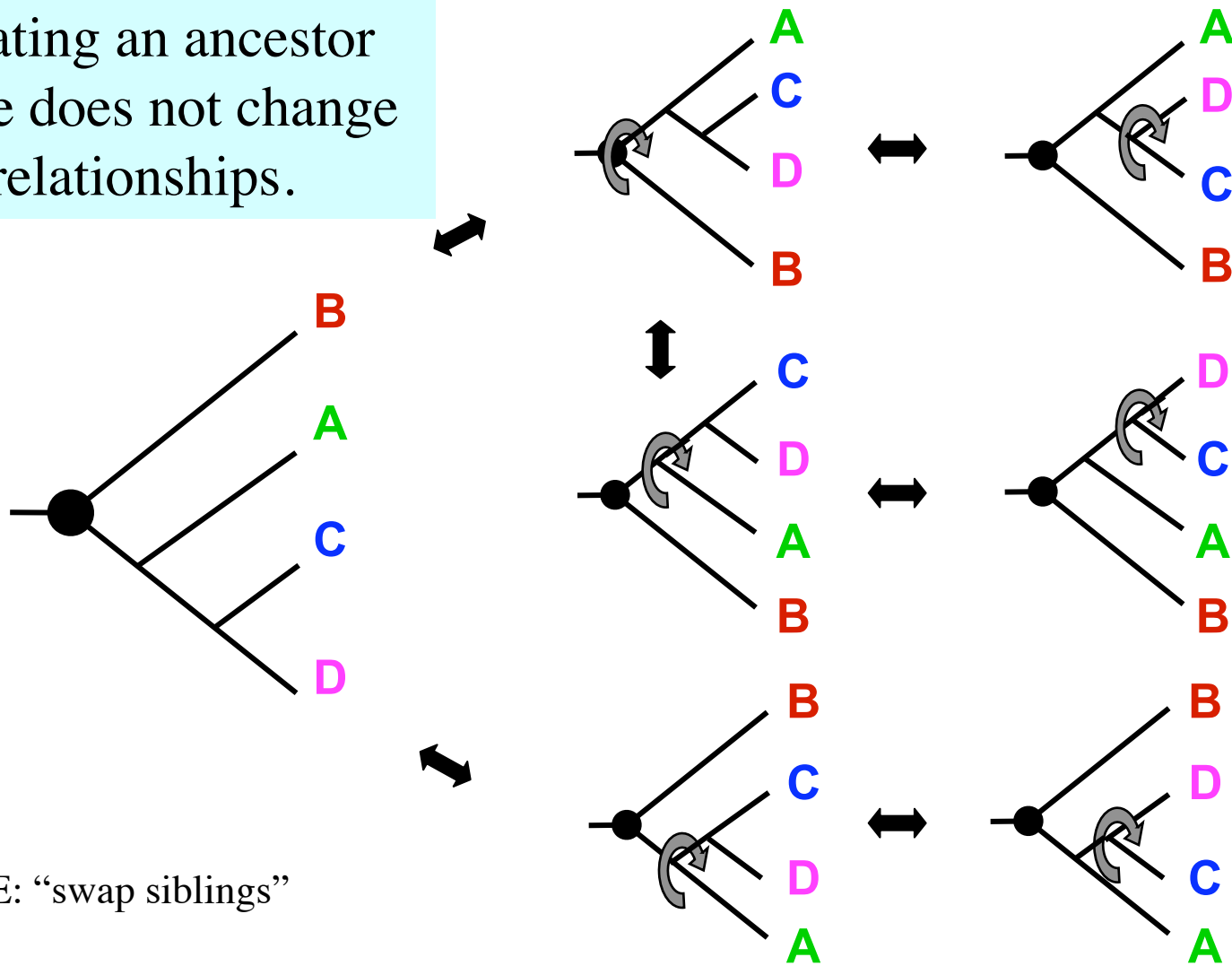


This one says  
C and D  
branched *late*.

This one says C  
and D branched  
*early*.

# Taxon order doesn't matter.

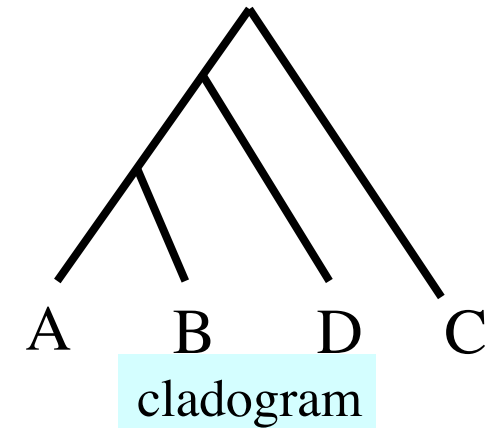
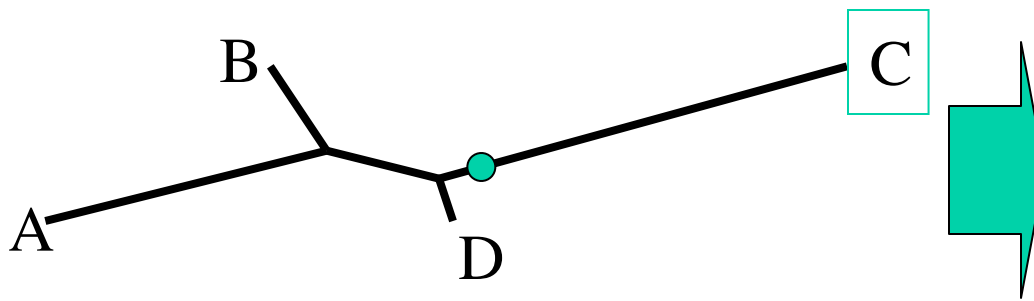
Rotating an ancestor node does not change the relationships.



UGENE: "swap siblings"

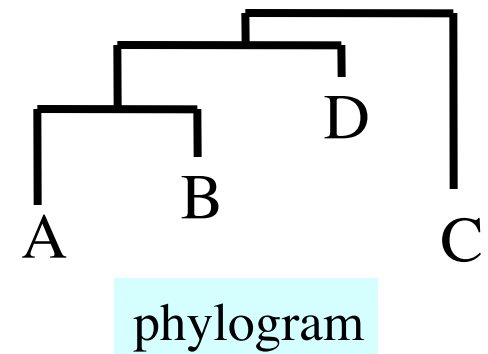
# Two strategies for rooting a tree:

1. Choose the **midpoint** between the two most distant branches.



2. Choose one taxon as the "out group." (it branches first.)

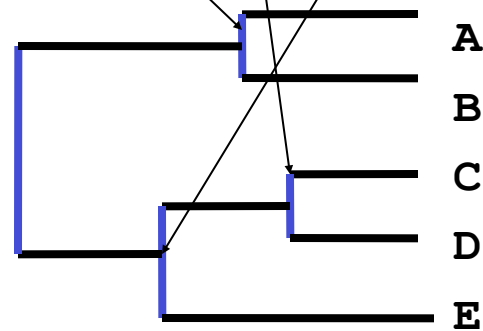
A good outgroup is not too distant from the rest of the tree.



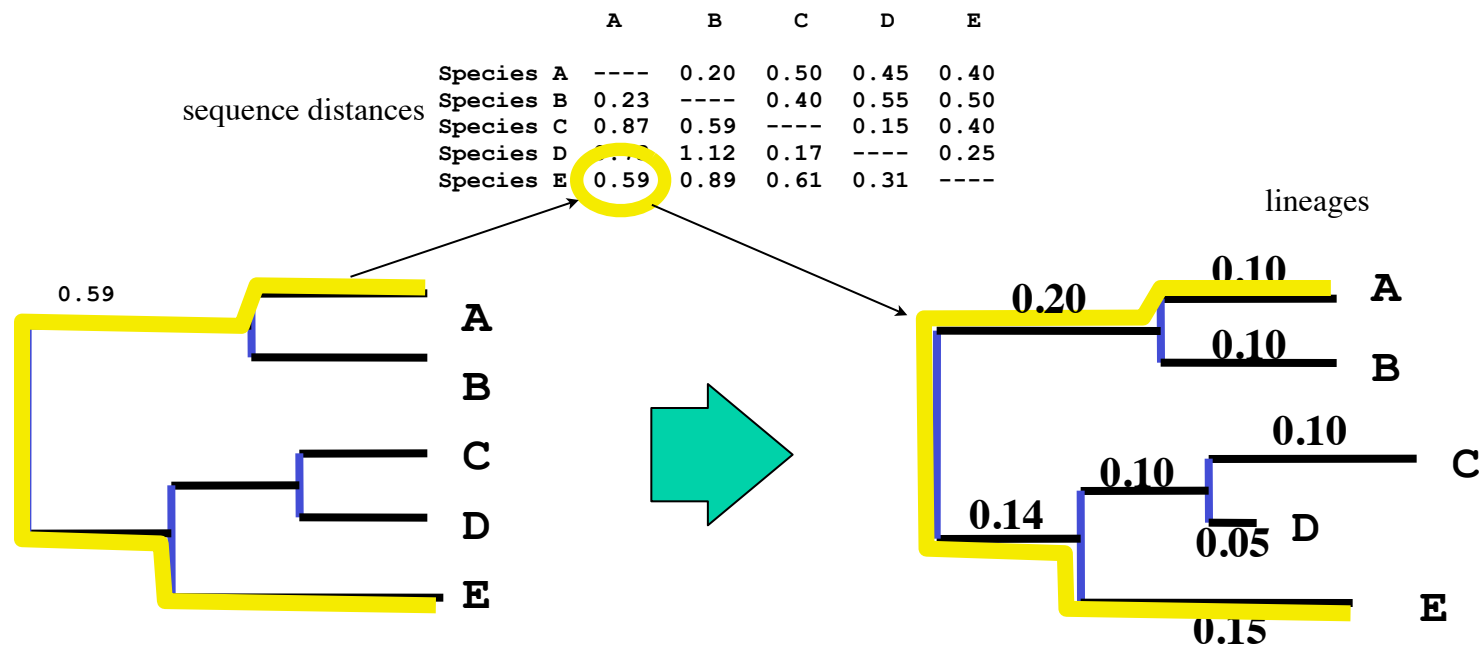
# Neighbor joining: cladogram

Choose the closest neighbors. Add a node between them.  
Choose the next closest, and so on.

	A	B	C	D	E
Species A	----	<b>0.20</b>	0.50	0.45	<b>0.40</b>
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	<b>0.15</b>	0.40
Species D	0.73	1.12	0.17	----	<b>0.25</b>
Species E	0.59	0.89	0.61	0.31	----

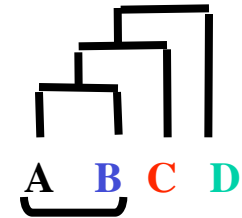


# cladogram to phylogram: Fitch-Margoliash





Problem statement: Given a tree and a set of sequence distances, derive lineages such that the **tree distances** maximally match the sequence distances.

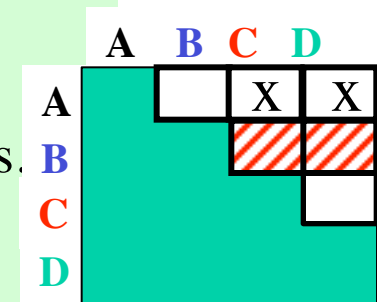
# Fitch-Margoliash algorithm for calculating the branch lengths



1. Find the most closely-related pair of sequences, **A** and **B**

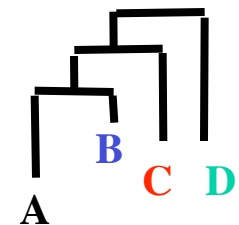
2. Calculate the average distance from **A** to all other sequences.

 then from **B** to all other sequences. 



3. Adjust the position of the common ancestor node for **A** and **B** so that the difference between the **averages** is equal to the difference between the **A** and **B branch lengths**, while the sum of the branch lengths is still equal to  $d(A,B)$ .

$$d(A)-d(B) = (d(A,C)+d(A,D))/2 - (d(B,C)+d(B,D))/2$$



NOTE: the difference between the averages may be greater than  $D(A,B)$ , making step 3 *impossible*.

# In class: create a rooted phylogram with 4 taxa

<b>A</b>	<b>T</b> T <b>G</b> <b>A</b> <b>C</b> <b>C</b> <b>A</b> <b>G</b> <b>A</b> <b>C</b> <b>C</b> <b>T</b> <b>G</b> <b>T</b> <b>G</b> <b>G</b> <b>T</b> <b>C</b> <b>C</b> <b>G</b>
<b>B</b>	<b>T</b> T <b>G</b> <b>A</b> <b>A</b> <b>C</b> <b>A</b> <b>G</b> <b>A</b> <b>C</b> <b>C</b> <b>T</b> <b>G</b> <b>C</b> <b>G</b> <b>G</b> <b>T</b> <b>C</b> <b>G</b> <b>G</b>
<b>C</b>	<b>T</b> <b>A</b> <b>G</b> <b>A</b> <b>A</b> <b>A</b> <b>G</b> <b>A</b> <b>C</b> <b>C</b> <b>T</b> <b>G</b> <b>T</b> <b>C</b> <b>G</b> <b>T</b> <b>A</b> <b>G</b> <b>G</b>
<b>D</b>	<b>G</b> <b>T</b> <b>G</b> <b>C</b> <b>A</b> <b>A</b> <b>A</b> <b>G</b> <b>T</b> <b>C</b> <b>C</b> <b>T</b> <b>G</b> <b>T</b> <b>G</b> <b>T</b> <b>A</b> <b>T</b> <b>C</b> <b>G</b>

	<b>A</b>	<b>B</b>	<b>C</b>	<b>D</b>
<b>A</b>		.15	.3	.3
<b>B</b>			.25	.45
<b>C</b>				.45
<b>D</b>				

pdist

$$K(A,B) = -3/4 \ln [1 - 4/3 \text{pdist}(A,B)]$$

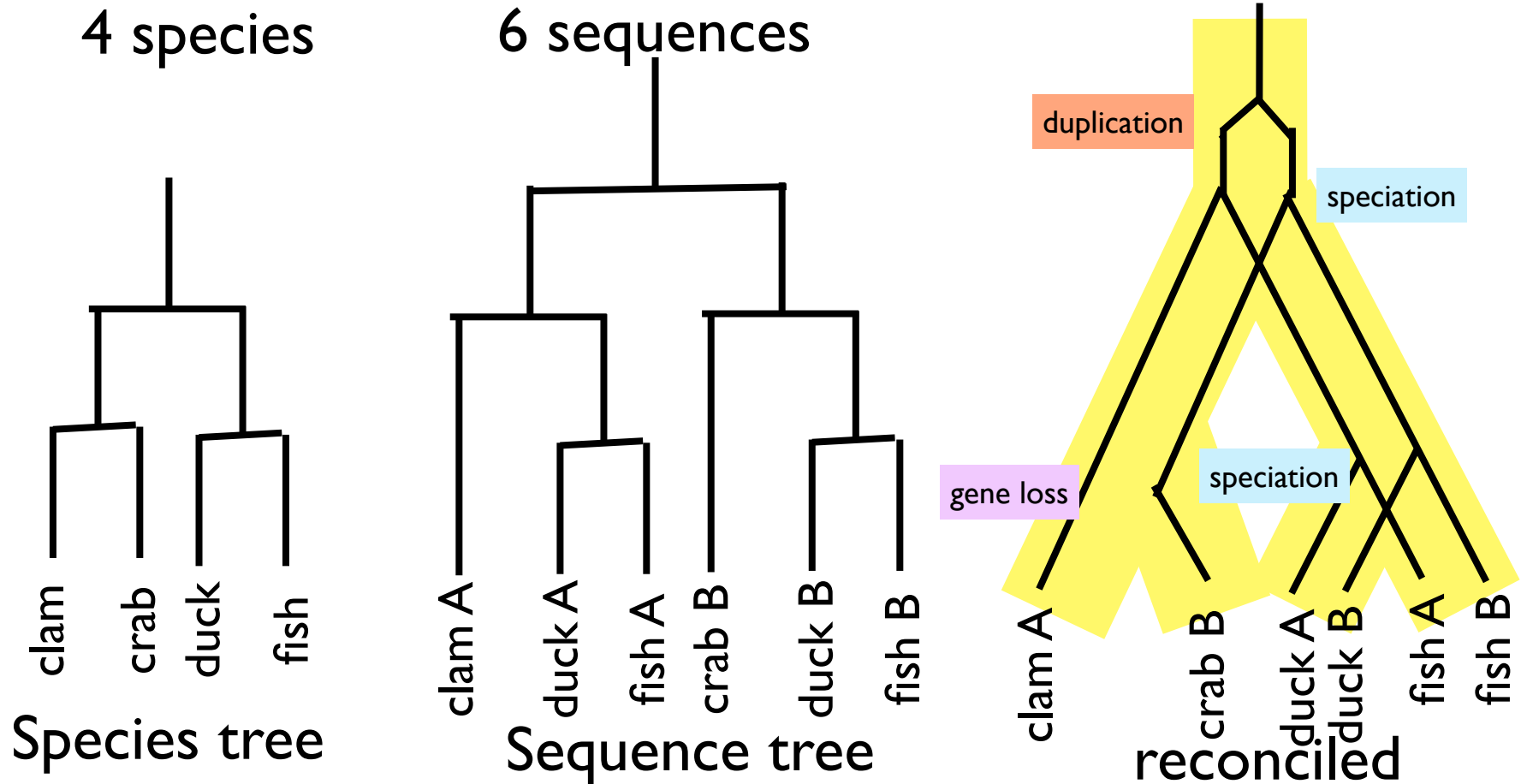
## Directions:

1. Make a distance matrix. (p-distance, then convert to **J-C distance**)
2. Use **Neighbor-joining** to make a tree.
3. Adjust branch lengths using **Fitch-Margoliash**.
4. Choose the root using the **Midpoint method**.

# Orthologs/paralogs

Orthologs: homologs originating from a speciation event

Paralogs: homologs originating from a gene duplication event.



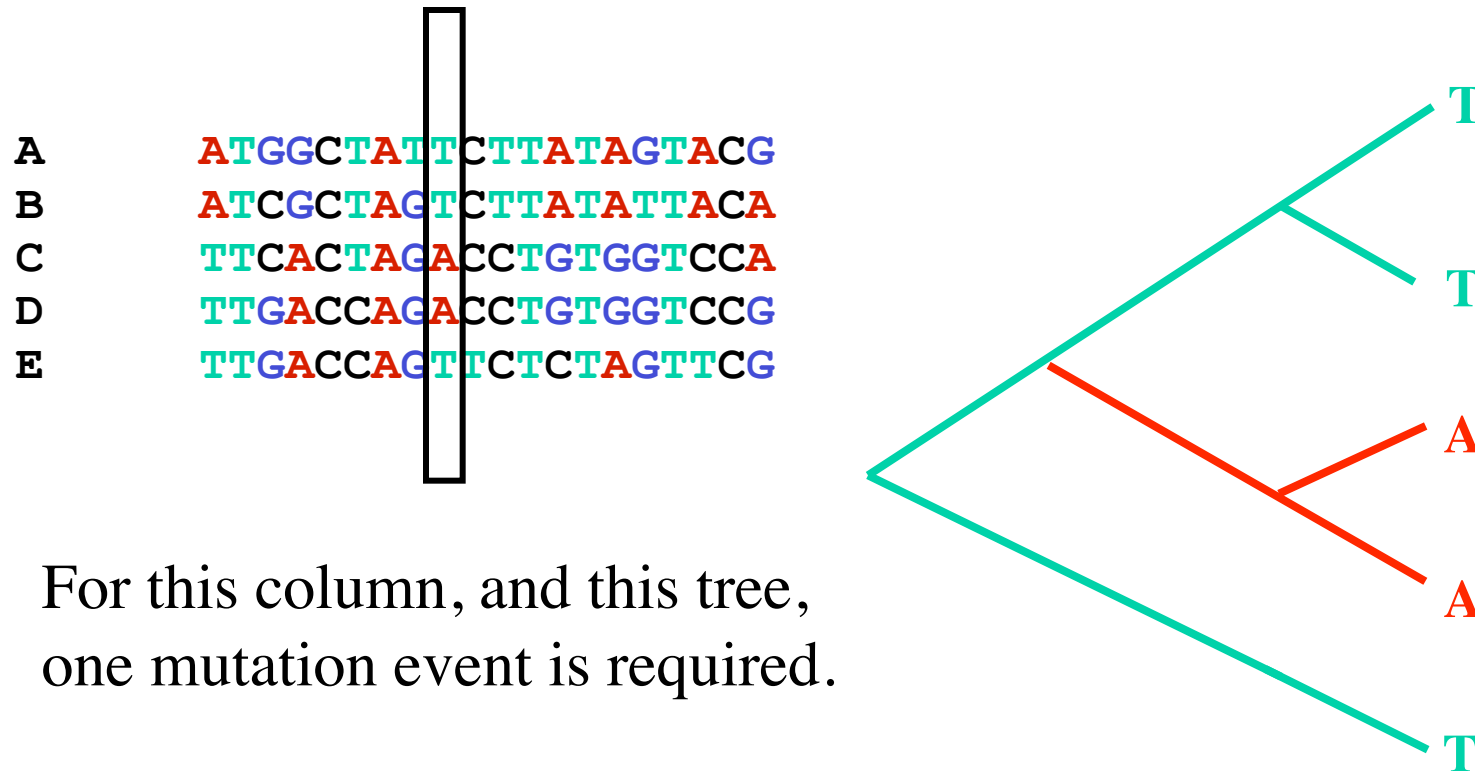
clam A and fish A are *orthologs*. clam A and crab B are *paralogs*.

# How do I know it's a paralog?

- If it's a **paralog**, then at some point in evolutionary history, a species existed with two identical genes in it.
  - One may have been lost since then. (Descendants are still paralogs!)
  - Paralogs can be from **different species**.
- Paralogous genes have **more than the expected sequence divergence**.
  - Because they are more likely to have different functions
  - Because they diverged earlier than the speciation event.
- Without *species information* or *functional information*, it's impossible to tell orthologs from paralogs.

# Maximum parsimony -- it's “character-building”

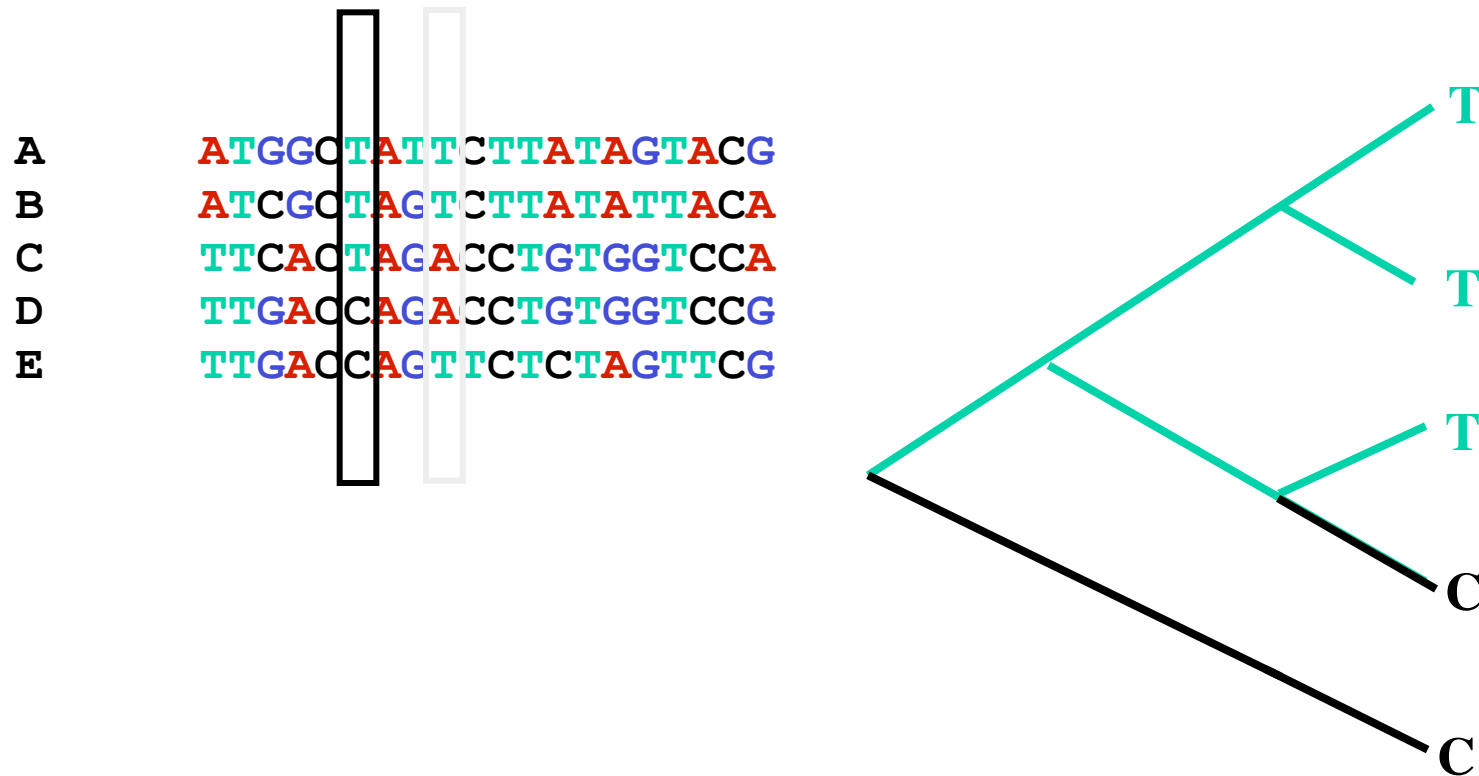
**Optimality criterion:** The ‘most-parsimonious’ tree is the one that requires the fewest number of evolutionary events (e.g., nucleotide substitutions, amino acid replacements) to explain the sequences.



For this column, and this tree,  
one mutation event is required.

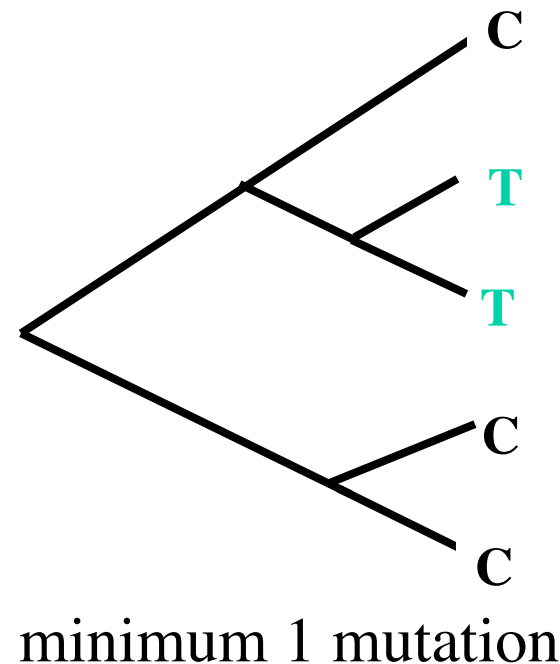
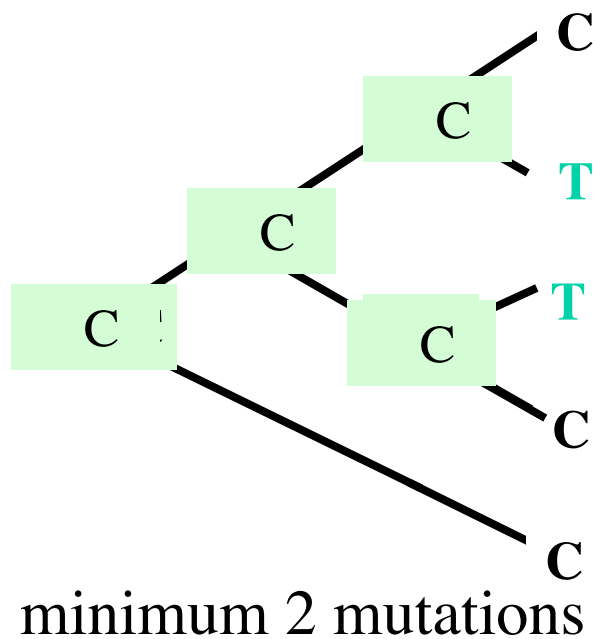
# character-based tree-building

For this other column, the same tree requires **two** mutation events. A different tree would require only one.



# Finding the minimum number of mutations

Given a tree and a set of taxa, one-letter each (1) choose optional characters for each ancestor. (2) Select the root character that minimizes the number of mutations by selecting each and propagating it through the tree.



Parsimony tips:  
Ignore non-informative sites

- No mismatches ---> 0 mutations, all trees
- 1 mismatch --> 1 mutation, all trees.
- all different --> all trees equivalent.



# Which method do I use?

Sequence similarity

strong

weak

very weak

Method to use

distance

parsimony

maximum likelihood