

Bioinformatics 1 -- lecture 9

Phylogenetic trees

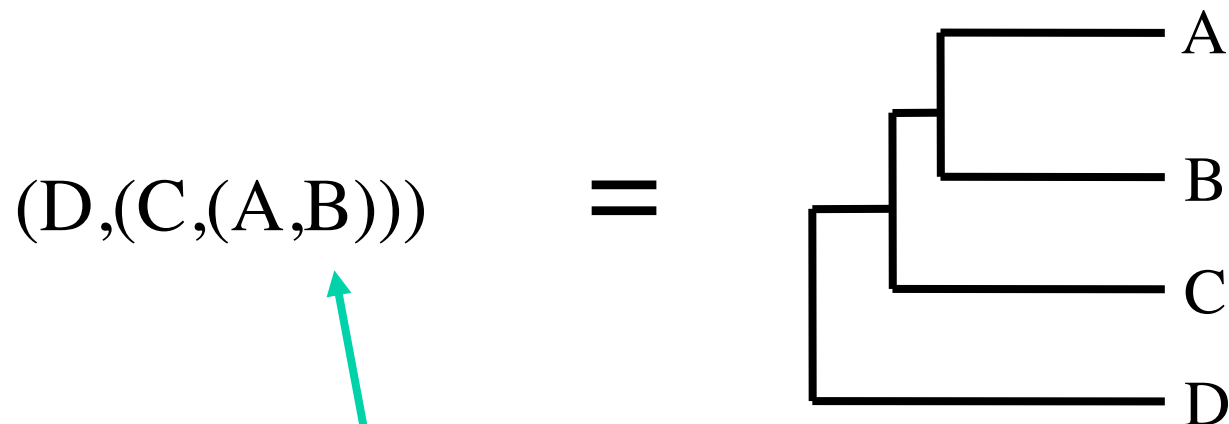
Distance-based tree building

Parsimony

(, (, (,)))

Trees can be represented in "parenthesis notation".

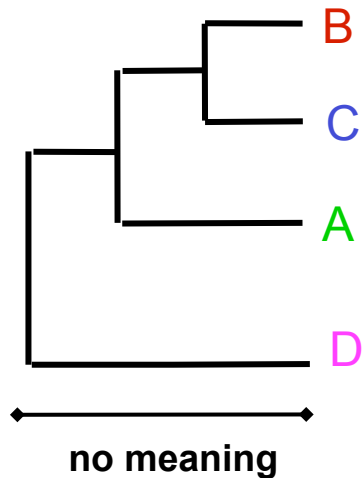
Each set of parentheses represents a branch-point (bifurcation), the comma separates left and right lineages.



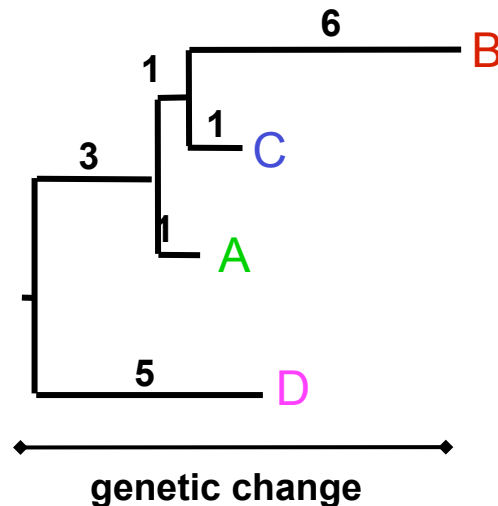
Parenthesis notation can contain sequence labels too.

Evolutionary time

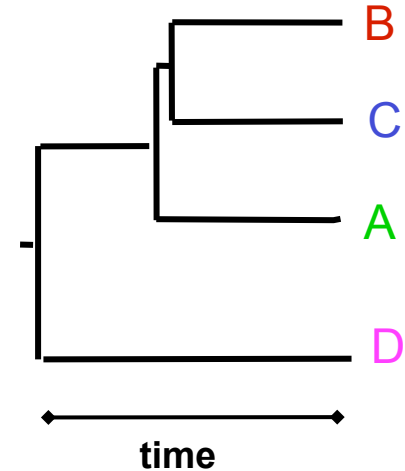
Cladogram



Phylogram



Ultrametric tree



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

parenthesis notation can have both labels and distances.

Distance metrics

METRIC DISTANCES between any two or three taxa (a, b, and c) have the following properties:

Property 1: $d(a, b) \geq 0$

Non-negativity

Property 2: $d(a, b) = d(b, a)$

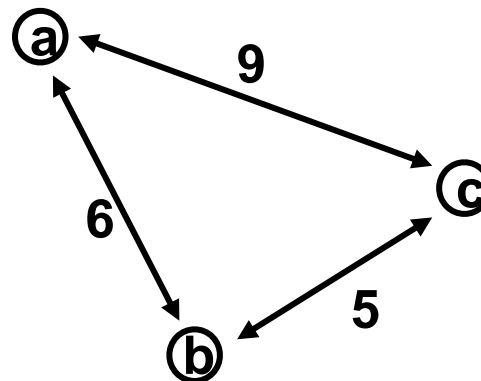
Symmetry

Property 3: $d(a, b) = 0$ if and only if $a = b$

Distinctness

Property 4: $d(a, c) \leq d(a, b) + d(b, c)$

Triangle inequality



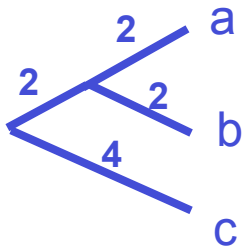
triangle inequality

Distance metrics

ULTRAMETRIC DISTANCES

....must satisfy the previous four conditions, plus:

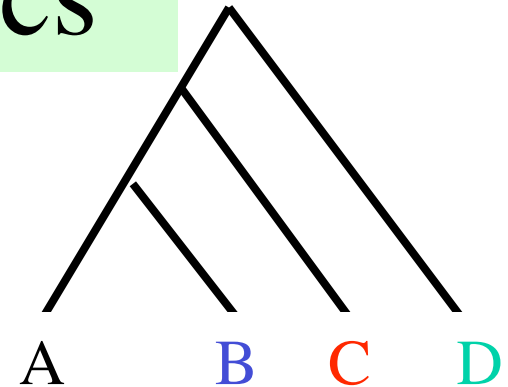
Property 5 *The distances from any branch point to the taxa in the clade defined by that branch point are equal.*



If distances are *ultrametric*, then the sequences are evolving in a perfectly **clock-like manner**. So any two sequences always have the same distance to their common ancestor.

Distance metrics

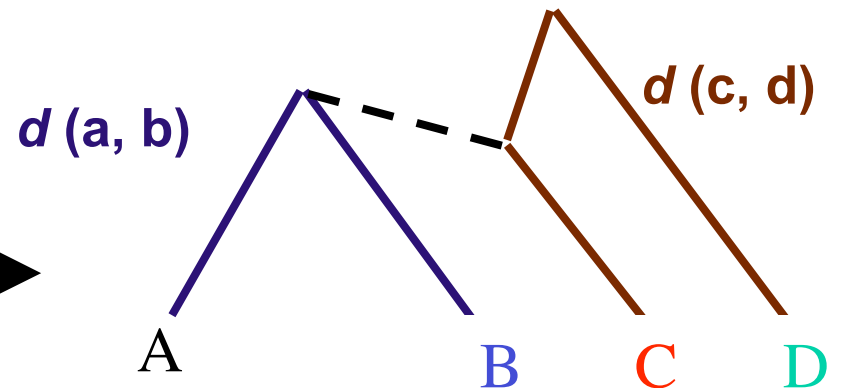
Additivity



Property 6: Example: if (a,b) are nearest neighbors,
 $d(a, b) + d(c, d) \leq \text{maximum} [d(a, c) + d(b, d), d(a, d) + d(b, c)]$

For distances to fit into an evolutionary tree, they must be additive. Estimated distances often fall short of these criteria, and thus can fail to produce correct evolutionary trees.

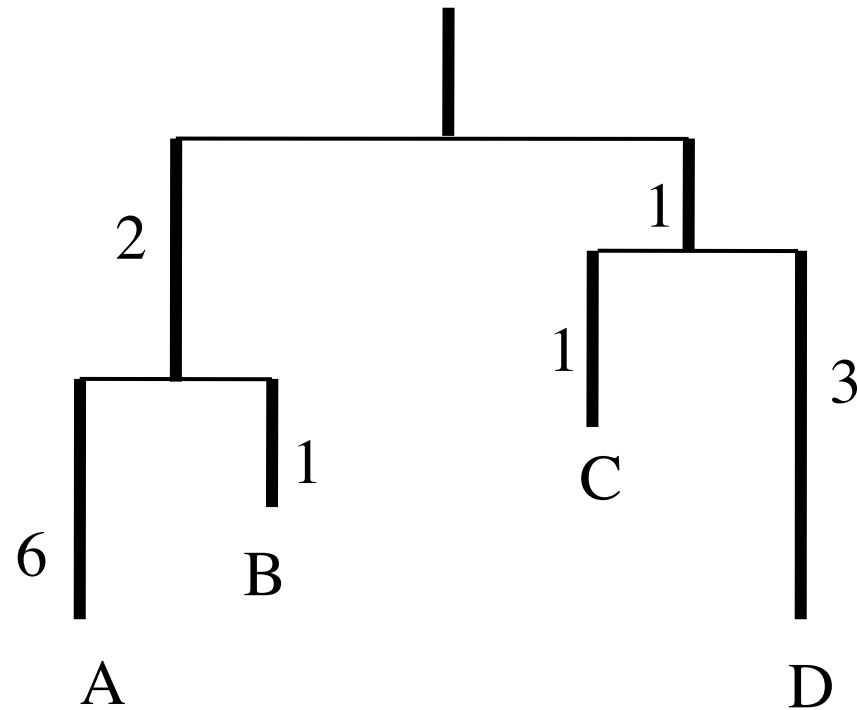
A lineage that goes *backwards*
in time violates additivity.



What's wrong with these distances?

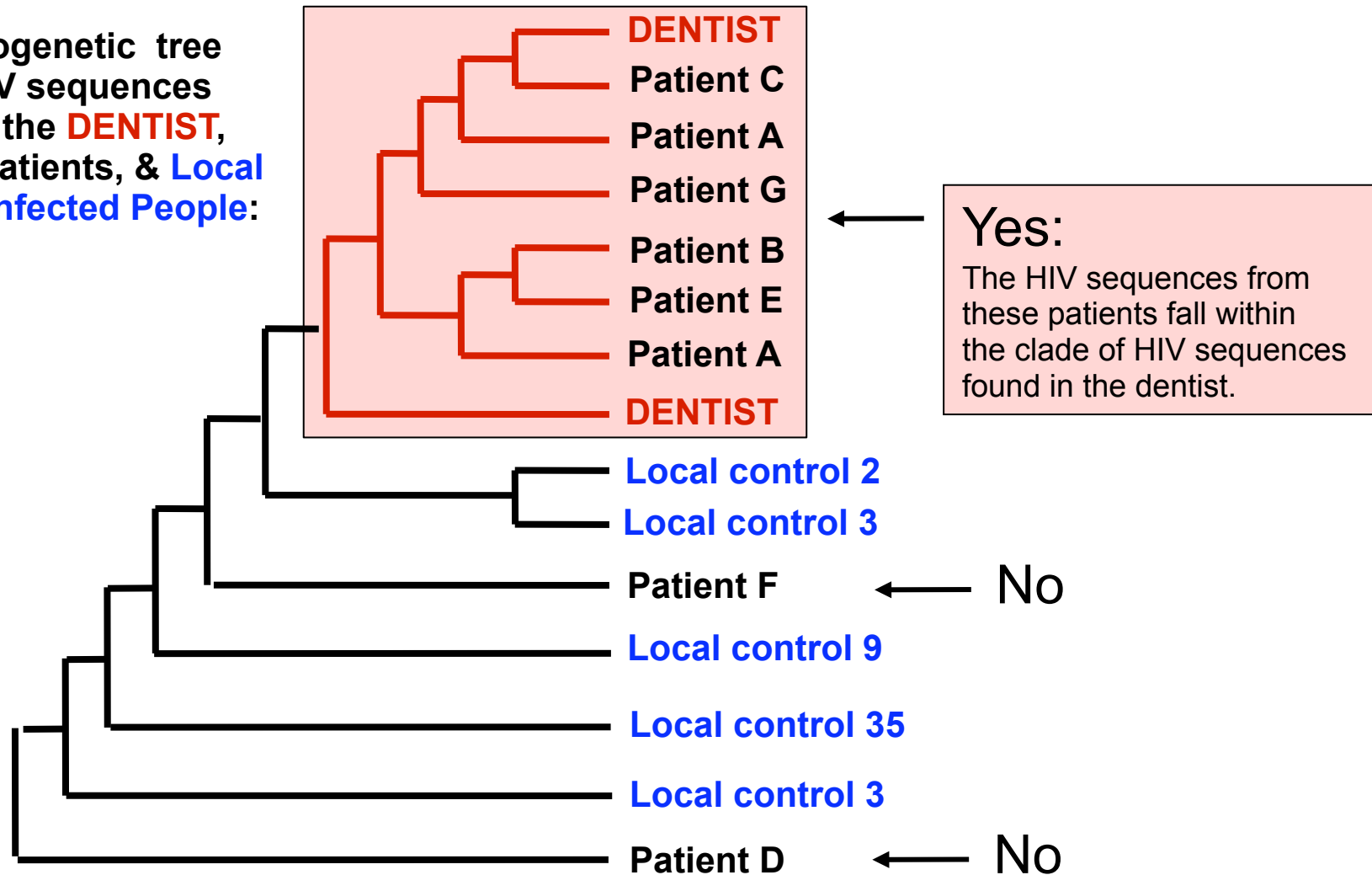
| | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 3 | 5 | 7 |
| B | 3 | 0 | 1 | 4 |
| C | 5 | 1 | 0 | 9 |
| D | 7 | 4 | 9 | 0 |

What's wrong with this tree?



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)

Character-based versus distance-based methods for tree building

Character-based methods: Use the aligned sequences directly during tree inference.

| Taxa | | Characters |
|-----------|--|-----------------------|
| Species A | | ATGGCTATTCTTATAGTACG |
| Species B | | ATCGCTAGTCTTATATTACA |
| Species C | | TTCACTAGACCTGTGGTCCA |
| Species D | | TTGACCAGACCTGTGGTCCG |
| Species E | | TTGACCAGTTCTCTAGTTTCG |

Distance-based methods: Transform the sequence data into pairwise distances, and then use the matrix during tree building, ignoring characters.

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

Calculating distances

Uncorrected p-distance: count the changes, divide by the length.

| | | | | | | | | | | | | | | | | | | | | |
|-----------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Species A | A | T | G | C | T | A | T | C | T | T | A | T | A | G | T | A | C | G | | |
| Species B | A | T | C | G | C | T | A | G | T | C | T | T | A | T | A | T | T | A | C | A |
| Species C | T | T | C | A | C | T | A | G | A | C | C | T | G | T | G | G | T | C | C | A |
| Species D | T | T | G | A | C | C | A | G | A | C | C | T | G | T | G | G | T | C | C | G |
| Species E | T | T | G | A | C | C | A | G | T | T | C | T | C | T | A | G | T | T | C | G |

$$D(A,B) = 4/20$$

Top: uncorrected p-distance, Bottom: Jukes-Cantor distance

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

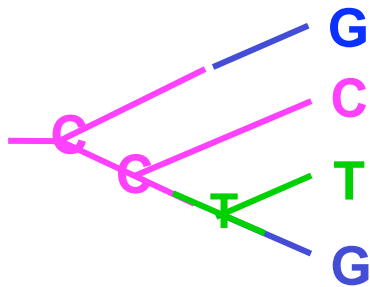
Jukes-Cantor correction:

$$K(A,B) = -3/4 \ln [1 - 4/3 D(A,B)]$$

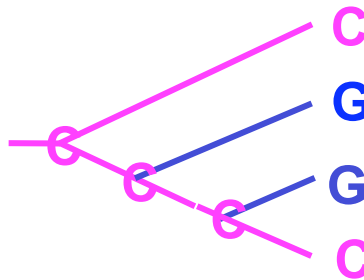
Homoplasy

Independent evolution of the same character.

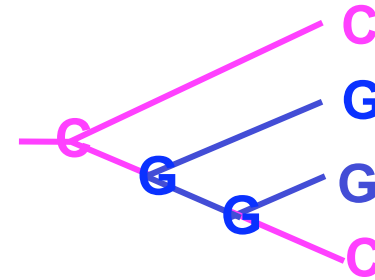
- (1) Convergent events (in either related or unrelated entities),
- (2) Parallel events (in related entities)
- (3) Reversals (in related entities)



(1)



(2)



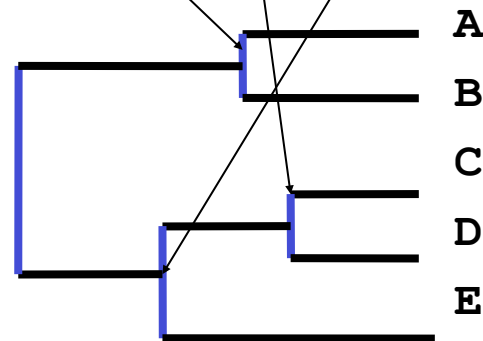
(3)

The **Jukes-Cantor correction** assumes homoplasy occurs at the rate predicted by random mutations.

Neighbor joining: a distance-based method

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

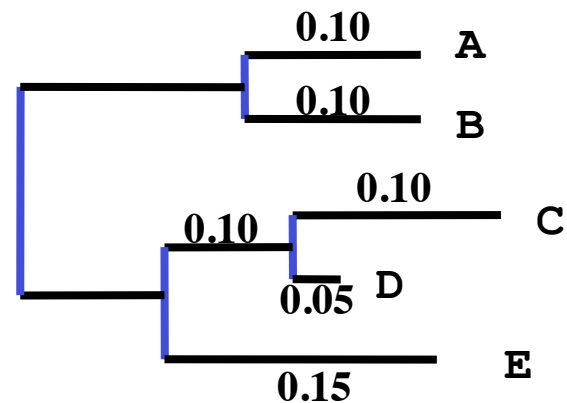
| | A | B | C | D | E |
|-----------|------|------|------|------|------|
| Species A | ---- | 0.20 | 0.50 | 0.45 | 0.40 |
| Species B | 0.23 | ---- | 0.40 | 0.55 | 0.50 |
| Species C | 0.87 | 0.59 | ---- | 0.15 | 0.40 |
| Species D | 0.73 | 1.12 | 0.17 | ---- | 0.25 |
| Species E | 0.59 | 0.89 | 0.61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





Neighbor joining: phylogram

Finally, **adjust the branch lengths** to fit the distances, if possible!

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

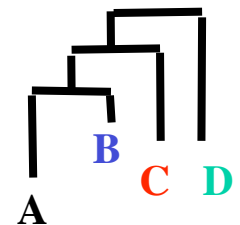
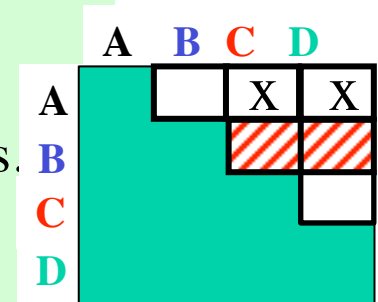
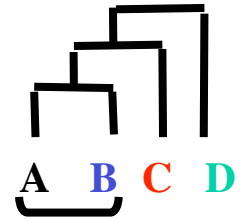


Fitch-Margoliash algorithm for calculating the branch lengths

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

| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | T | T | G | A | C | C | A | G | A | C | C | T | G | T | G | G | T | C | C | G |
| B | T | T | G | A | A | C | A | G | A | C | C | T | G | C | G | G | T | C | G | G |
| C | T | A | G | A | A | A | G | A | C | C | T | G | T | C | G | T | A | G | G | |
| D | G | T | G | C | A | A | A | G | T | C | C | T | G | T | G | T | A | T | C | G |

| | A | B | C | D | |
|---|---|-----|-----|-----|-------|
| A | | .15 | .3 | .3 | pdist |
| B | | | .25 | .45 | |
| C | | | | .45 | |
| D | | | | | |

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

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**.

Which method do I use?

Sequence similarity

strong

weak

very weak

Method to use

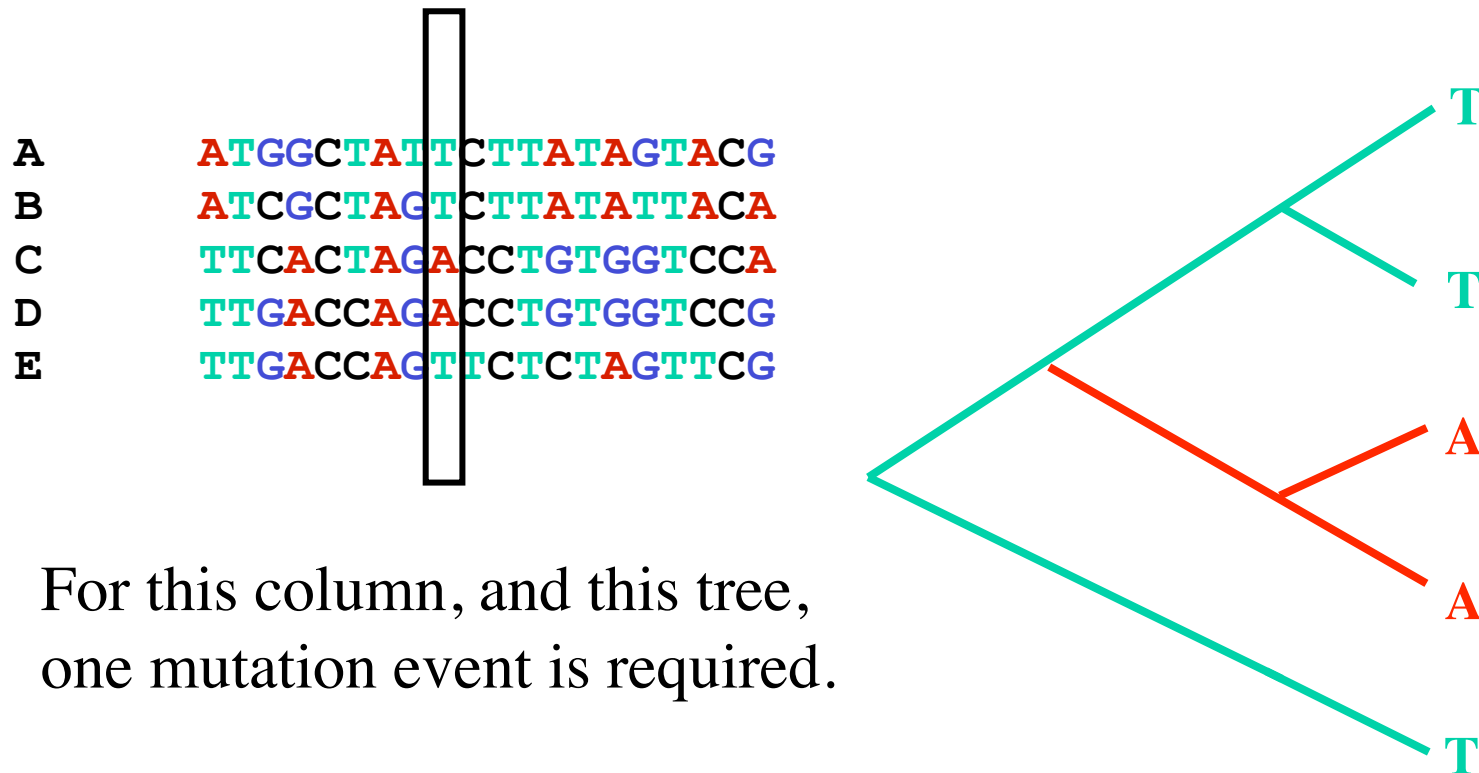
distance

parsimony

maximum likelihood

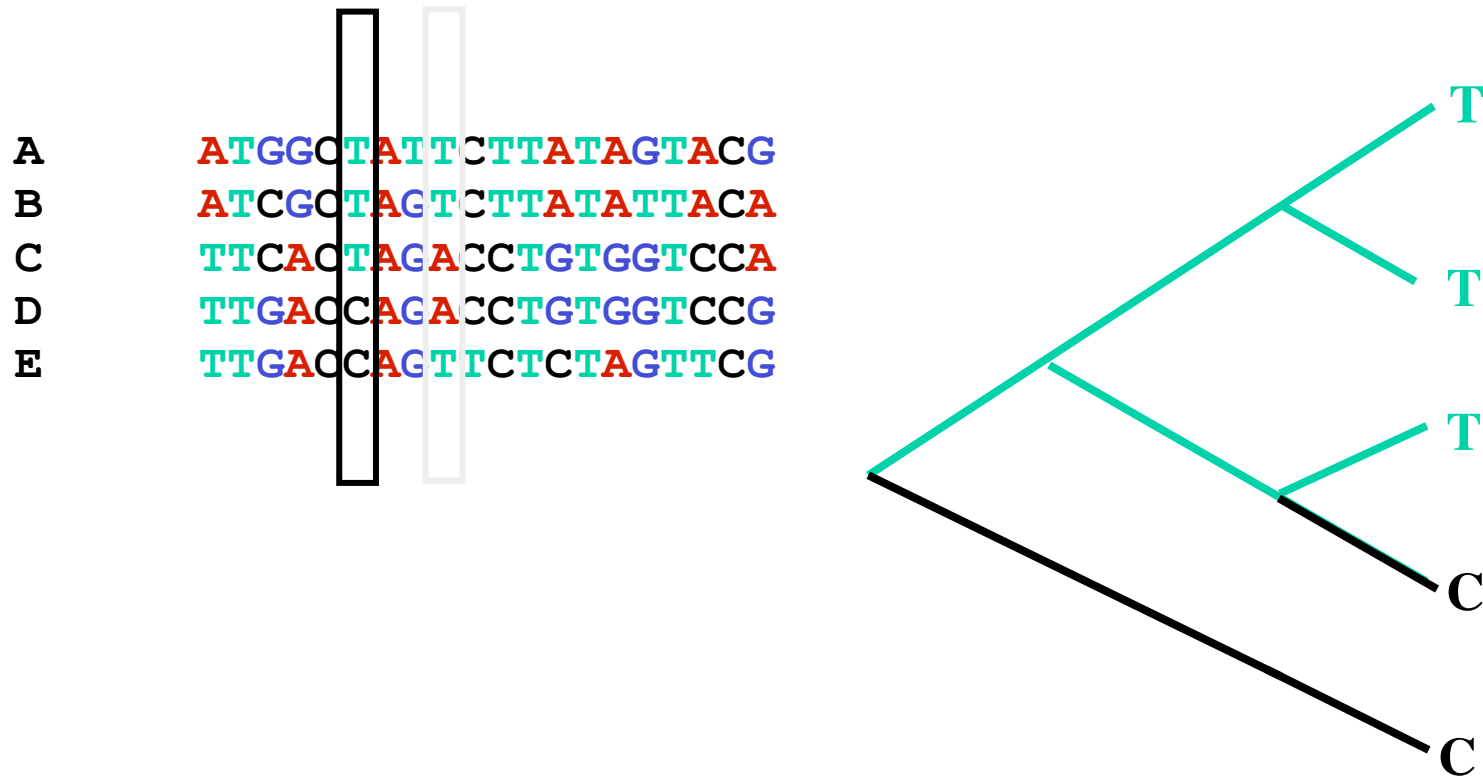
Maximum parsimony -- it's “character-building”

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.



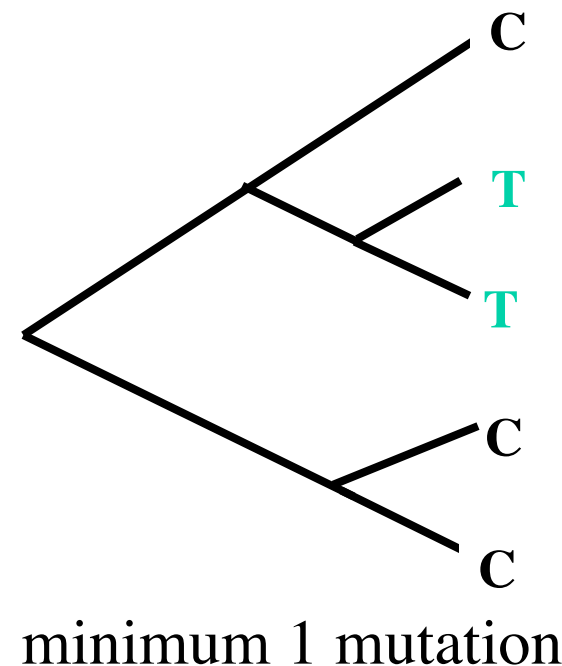
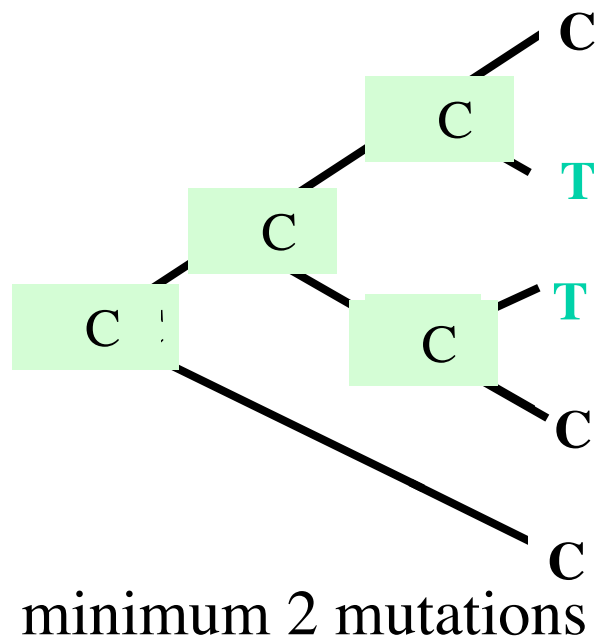
character-based tree-building

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.



Ignore non-informative sites

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

Max Unweighted Parsimony: Trying all trees

. . . . | 0 | 0
A A T G G C T A T T C T T A T A G T A C G
B A T G G C T A G T C T T A T A T T A C A
C T T C A C T A G A C C T G T G G T C C A
D T T G A C C A G A C C T G T G G T C C G
E T T G A C C A G T T C T C T A G T T C G

TOTALS

[illegible]