Phylogenetics trees

Tree types

Tree theory

Distance-based tree building

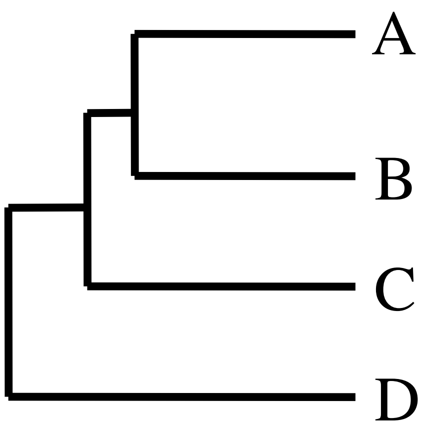
Parsimony

(, (, (,)))

Trees can be represented in "parenthesis notation".

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

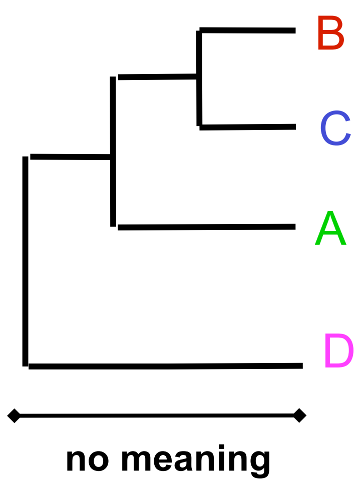
(D,(C,(A,B))) =



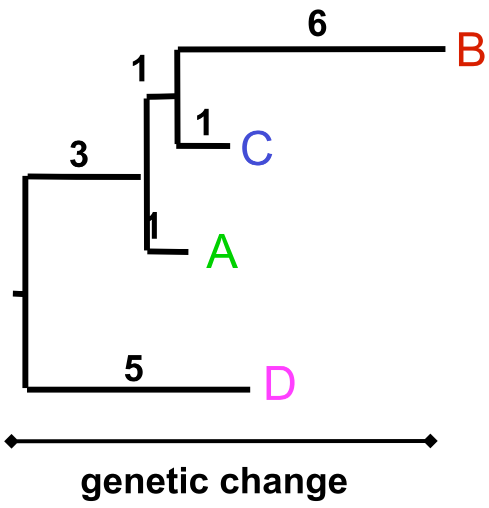
Parenthesis notation can contain sequence labels too.

Evolutionary time

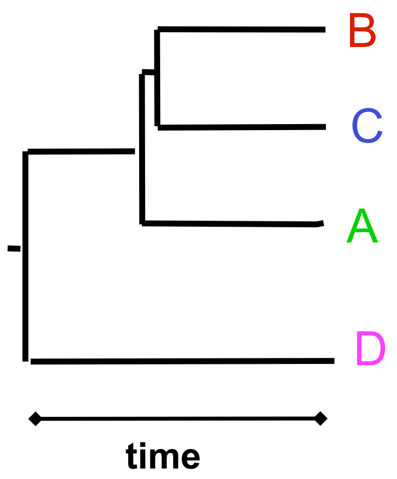
Cladogram



Phylogram



Ultrametric tree

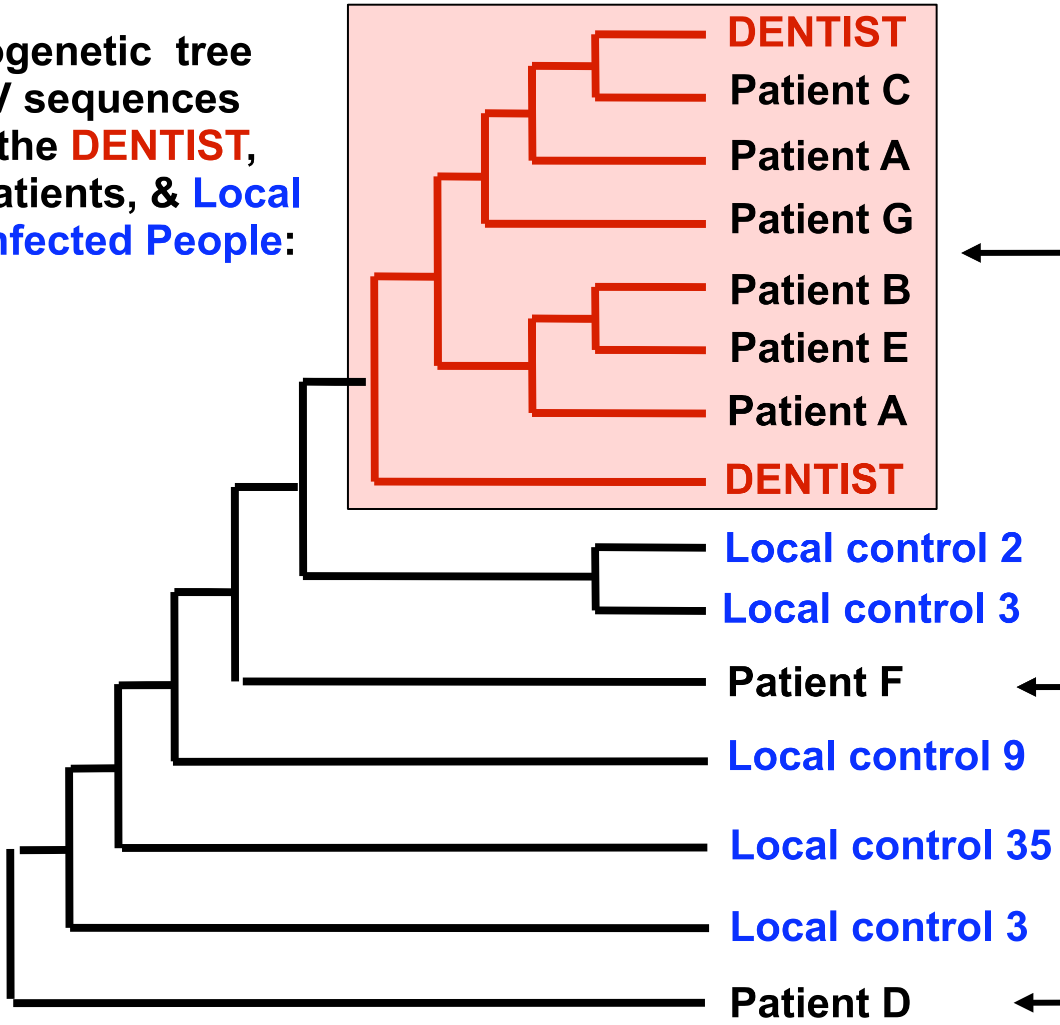


(D:5,(A:1,(C:1,B:6):1):3)

parenthesis notation can have both labels and distances.

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

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



Yes:
The HIV sequences from these patients fall within the clade of HIV sequences found in the dentist.

No

No

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	TTGACCAGTTCTCTAGTTCG

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 distance: count the changes, divide by the length.

```
Species A      ATGGCTATTCTTATAGTACG
Species B      ATCGCTAGTCTTATATTACA
Species C      TTCACTAGACCTGTGGTCCA
Species D      TTGACCAGACCTGTGGTCCG
Species E      TTGACCAGTTCTCTAGTTCG
```

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	----

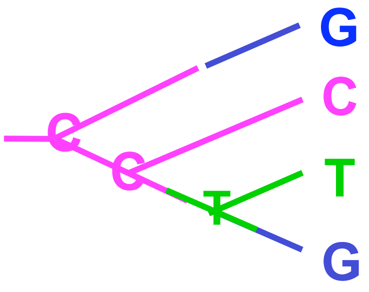
Jukes-Cantor correction:

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

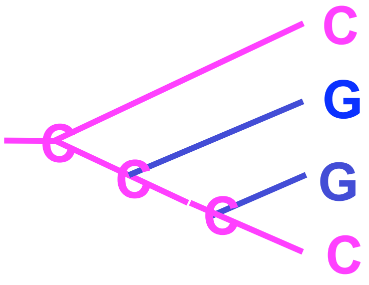
Homoplasy

Independent evolution of the same character.

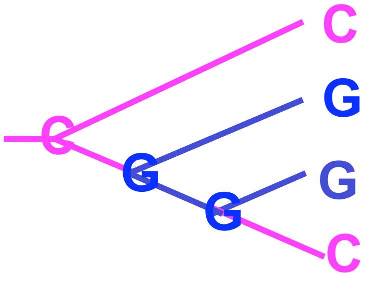
(1) Convergent events (in either related on unrelated entities),
(2) Parallel events (in related entities)
(3) Reversals (in related entities)



(1)



(2)



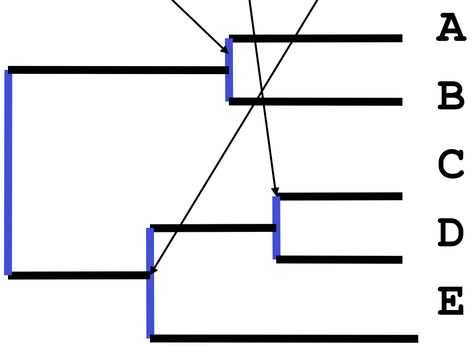
(3)

The **Jukes-Kantor 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, ad so on.

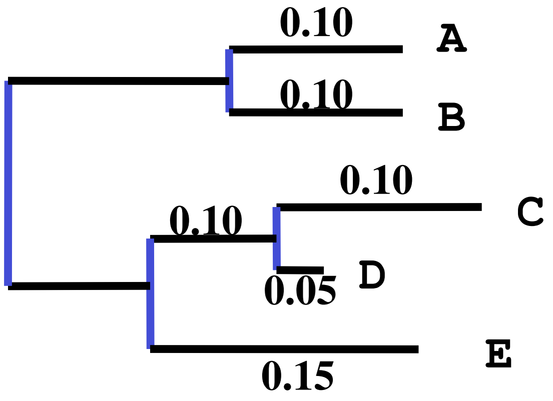
	A	B	C	D	E
Species A	----	0.20	0.50	0.45	0.40
Species B	0.23	----	0.40	0.55	0.50
Species C	0.87	0.59	----	0.15	0.40
Species D	0.73	1.12	0.17	----	0.25
Species E	0.59	0.89	0.61	0.31	



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	----



In class: create a rooted phylogram with 4 taxa

A	TTGACCAGACCTGTGGTCCG
B	TTGAACAGACCTGCGGTCGG
C	TAGAAAAGACCTGTCGTAGG
D	GTGCAAAGTCCTGTGTATCG

Directions:

- Make a distance matrix
- Use Neighbor-joining to make a tree.
- Adjust branch lengths using Fitch-Margoliash.
- Choose the root using the Midpoint method.

Which method do I use?

Sequence similarity	Method to use
strong	parsimony
weak	distance
very weak	maximum likelihood

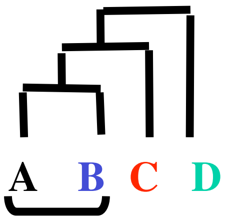
Geneious exercise

- Search NCBI-->Protein for any gene (example: mitochondrial peptidase)
- Run a **blastp** search on that gene.
- Select 8 sequences with e-values between 0.1 and 0.00001
- Align the sequences using Geneious Align. Rename them A,B,C,D,E,F,G and H for simplicity.
- Trim the alignment.
- Extract a 20 aa block. Save it separately.
- Make a UPGMA tree using distances. Draw the cladogram with taxa.
- Repeat with a different 20 aa block, until you have 5 trees.

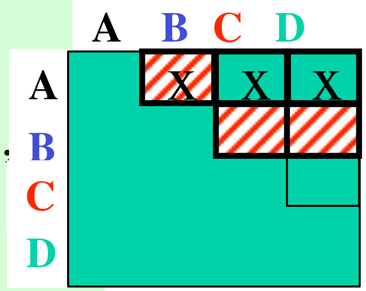
- Compare the five trees. Choose a tree as the “consensus” tree.
- For each node in the consensus tree, count how many trees have the equivalent branch point, or node (identical sub-clade content).
- Write this number (1-5) at the node position on the consensus tree.

If you did this exercise 100 times and counted the times you get a certain branch point, then you have computed a "*bootstrapping*" value for that branch point.

Fitch-Margoliash algorithm for calculating the branch lengths

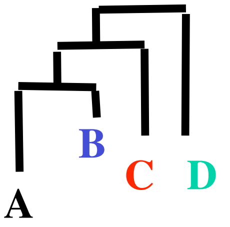


1. Find the most closely-related pair of sequences, **A** and **B**



2. Calculate the average distance from **A** to all other sequences, $\boxed{x}$ 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 D(A,B).

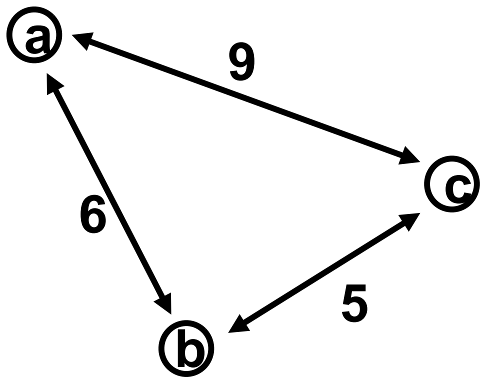


NOTE: the difference between the averages may be greater than D(A,B), making step 3 *impossible*.

Distance metrics

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

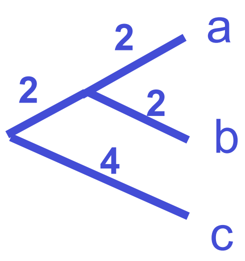
Property 1:	**d (a, b) ≥ 0**	**Non-negativity**
Property 2:	**d (a, b) = d (b, a)**	**Symmetry**
Property 3:	**d (a, b) = 0 if and only if a = b**	**Distinctness**
Property 4:	**d (a, c) ≤ d (a, b) + d (b, c)**	**Triangle inequality**



ULTRAMETRIC DISTANCES

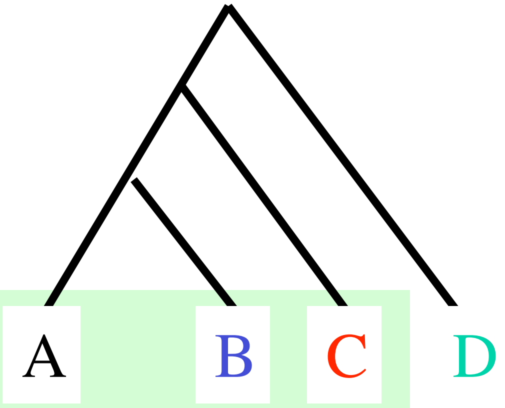
must satisfy the previous four conditions, plus:

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.

Additivity



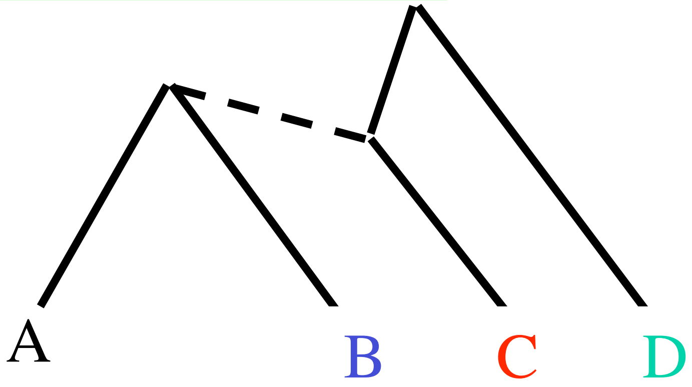
ADDITIVE DISTANCES:

Property 6: Example: if (a,b) are nearest neighbors,

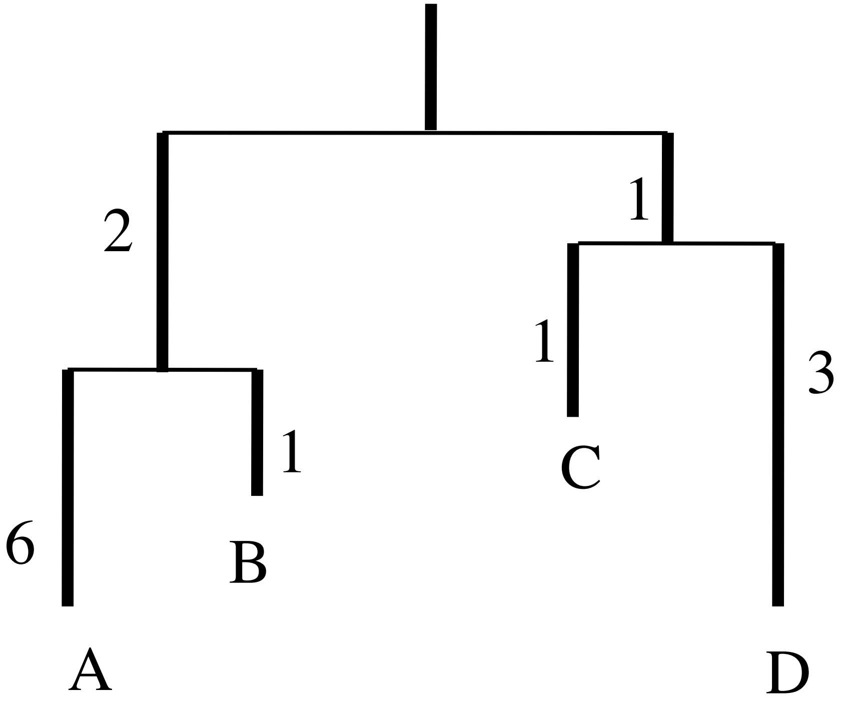
$$d\,(a, b) + d\,(c, d) \le \text{maximum}\,[d\,(a, c) + d\,(b, d),\ d\,(a, d) + d\,(b, c)]$$

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 this tree?



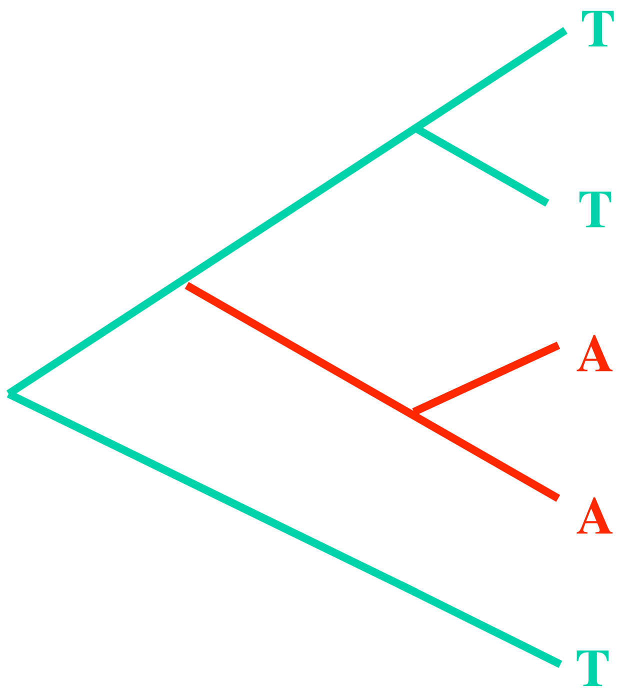
What's wrong with these distances?

	A	B	C	D
A		3	5	7
B			1	4
C				9
D				

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.

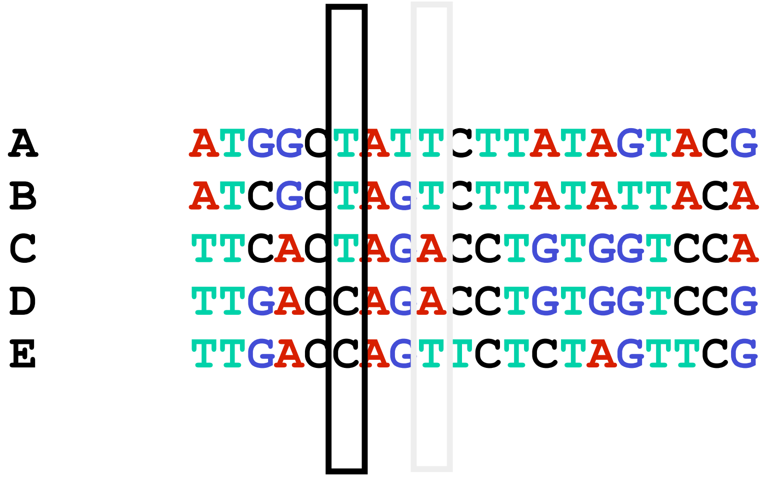
```
A        ATGGCTATTCTTATAGTACG
B        ATCGCTAGTCTTATATTACA
C        TTCACTAGACCTGTGGTCCA
D        TTGACCAGACCTGTGGTCCG
E        TTGACCAGTTCTCTAGTTCG
```



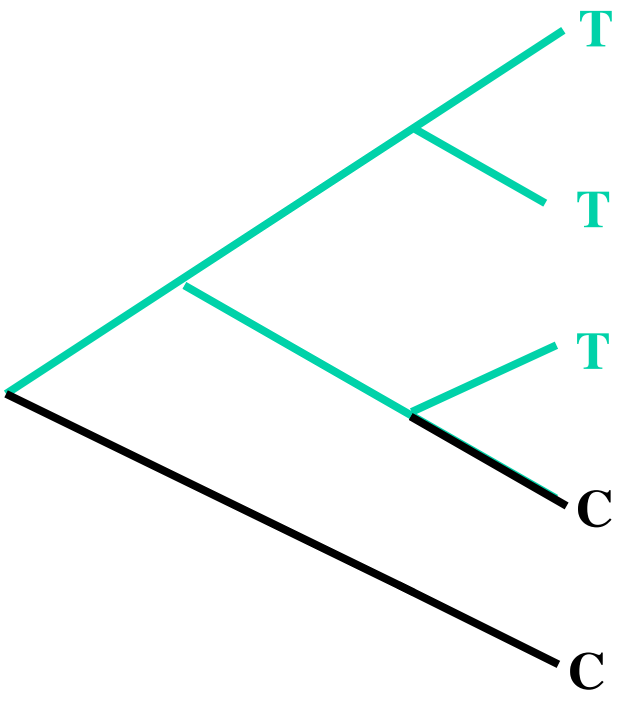
For this column, and this tree, one mutation event is required.

character-based tree-building

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

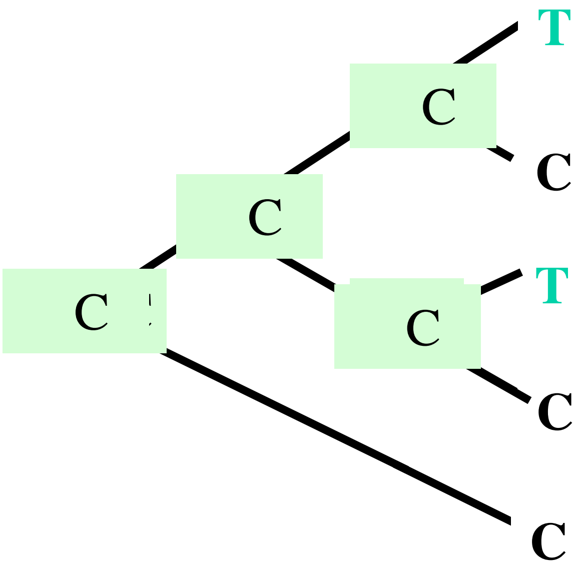


```
A        ATGGCTATTCTTATAGTACG
B        ATCGCTAGTCTTATATTACA
C        TTCACTAGACCTGTGGTCCA
D        TTGACCAGACCTGTGGTCCG
E        TTGACCAGTTCTCTAGTTCG
```



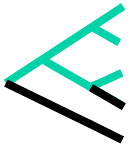
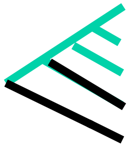
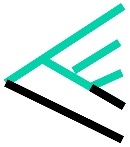
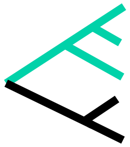
Minimum number of mutations

Given a tree and a set of taxa, one-letter each, choose optional characters for each ancestor, starting from the most recent. Choose the most popular character at the root, then choose not to mutate if possible.



Max Parsimony: Trying all trees

```
     ....|....0....|....0
A    ATGGCTATTCTTATAGTACG
B    ATCGCTAGTCTTATATTACA
C    TTCACTAGACCTGTGGTCCA
D    TTGACCAGACCTGTGGTCCG
E    TTGACCAGTTCTCTAGTTCG
```

																					TOTALS
	1	0																			
	1	0																			
	2	0																			
	1	0																			