Phylogenetic Trees

Genome 373

Genomic Informatics

Elhanan Borenstein

A quick review

Global alignment algorithm:

Needleman-Wunsch.

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Local alignment algorithm:

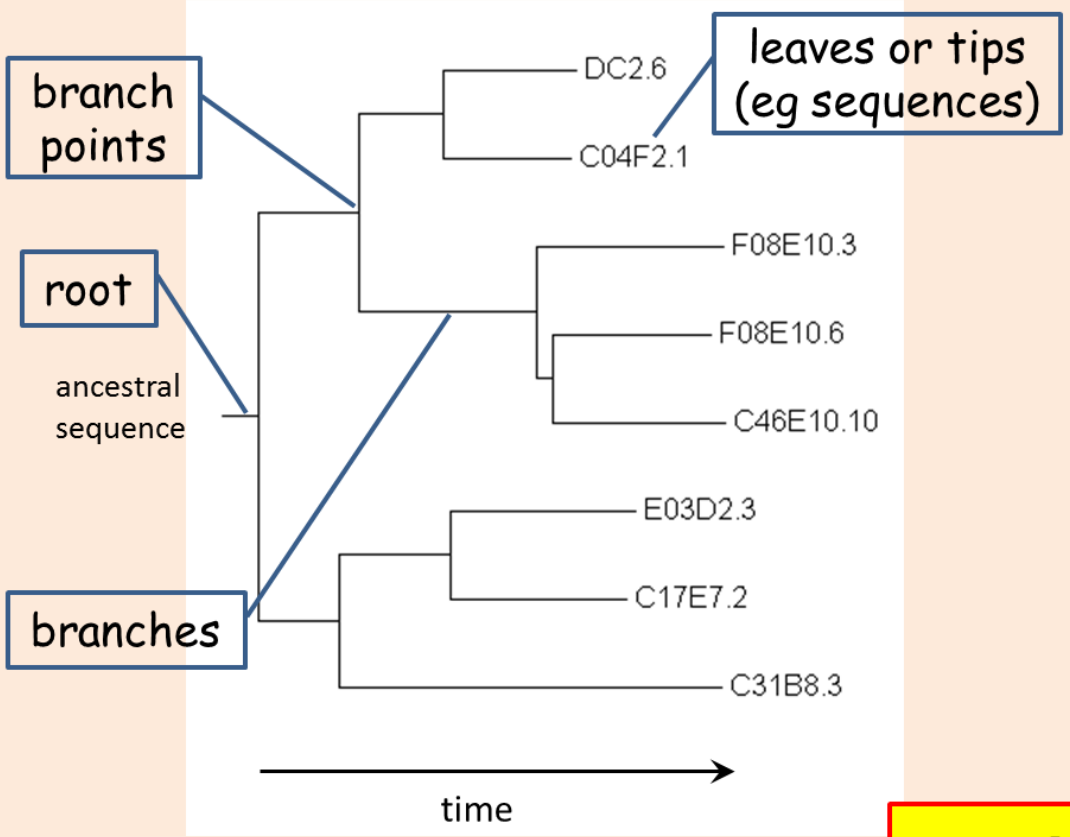
Smith-Waterman.

		A	A	G
	0	0	0	0
G	0	0	0	2
A	0	2	2	0
A	0	2	4	0
G	0	0	0	6
G	0	0	0	2
C	0	0	0	0

- Significance of similarity scores (P-values)
  - Empirical null score distribution
  - Extreme value distribution
- Multiple-testing correction (Bonferroni) and E-values

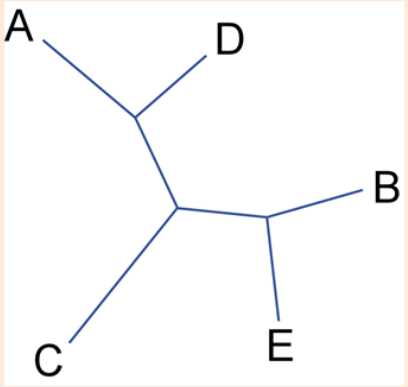
Trees

rooted tree (all real trees are rooted):



… sequence divergence is proportional to (horizontal) branch lengths

unrooted tree:
(used when the root isn't known):



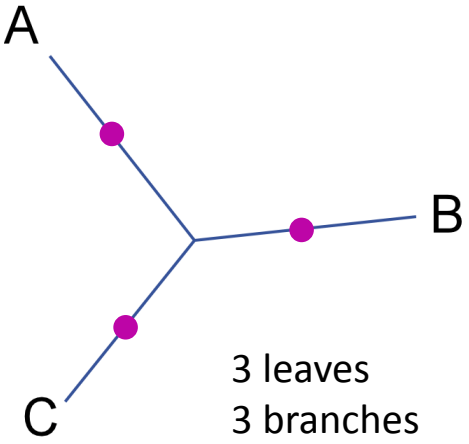
time radiates out from somewhere (probably near the center)

Note: A tree has topology and distances

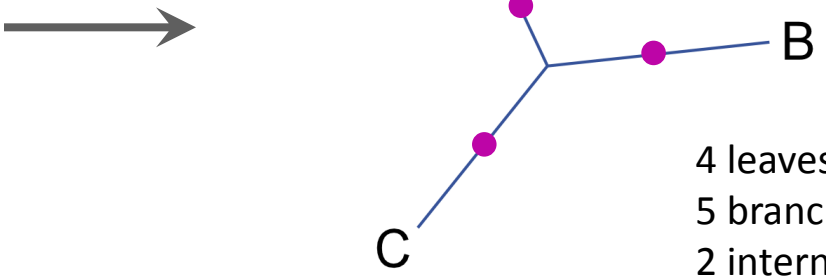
Why is inferring phylogeny a hard problem?

How many different tree topologies for linking N species?

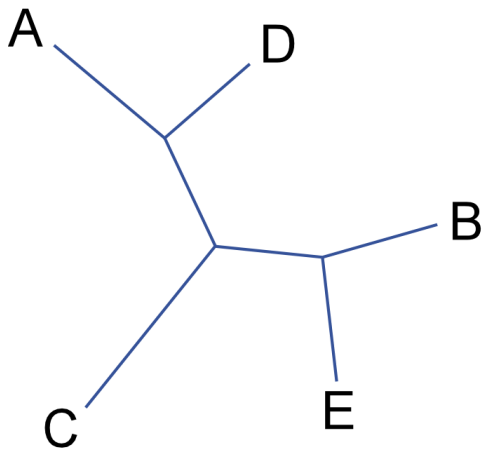
How many different tree topologies for linking N species?



3 leaves
3 branches
1 internal node
1 topology
(3 insertions)



4 leaves
5 branches
2 internal nodes
3 topologies (x3)
(5 insertions)

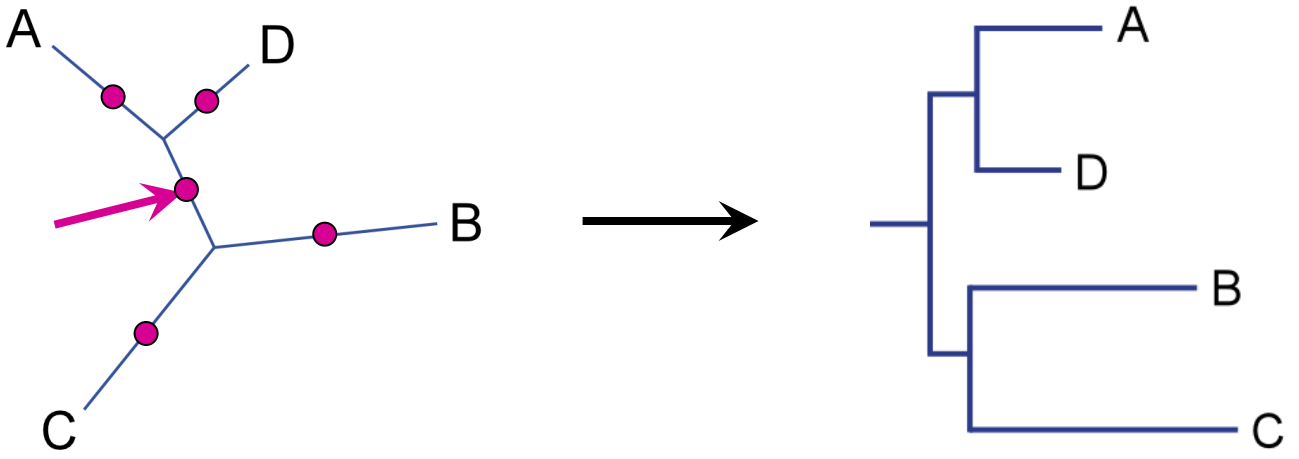


5 leaves
7 branches
3 internal nodes
15 topologies (x5)
(7 insertions)

In general, an unrooted tree with **N** leaves has:
2N - 3 total branches
N leaf branches
N - 3 internal branches
N - 2 internal nodes
3*5*7*…*(2N-5) ~O(N!) topologies

There are many rooted trees for each unrooted tree

For each unrooted tree, there are 2N - 3 times as many rooted trees, where N is the number of leaves (# branches = 2N – 3).



The number of tree topologies grows extremely fast

20 leaves - 564,480,989,588,730,591,336,960,000,000 topologies

How can you find the “right” tree?

- Many methods available, we will talk about:
  - Distance trees
  - Parsimony trees
- Others include:
  - Maximum-likelihood trees
  - Bayesian trees

Distance matrix methods

- Methods based on a set of pairwise distances typically from a multiple alignment.

```
                1 2 3 4 5 6
human           a g t c t c
chimp           a g a g t c
gorilla         c g g c a g
orangutan       c g g g a c
```

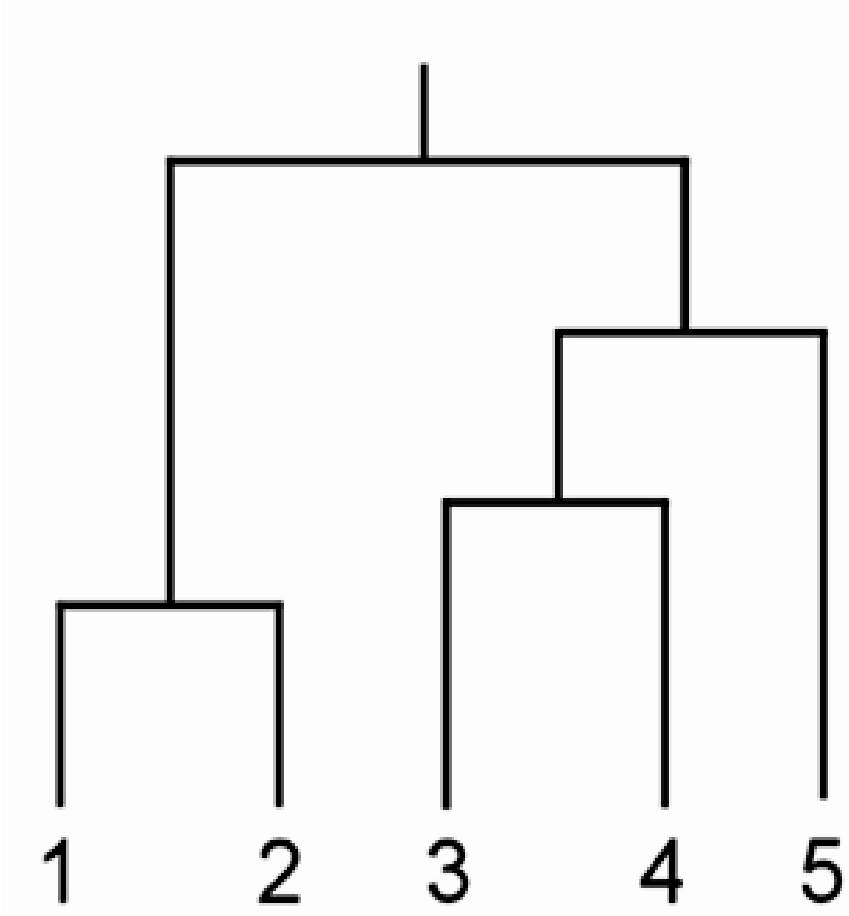
human - chimp has 2 changes out of 6 sites
human - orang has 4 changes of out 6 sites
etc.

	human	chimp	gorilla	orang
human	0	2/6	4/6	4/6
chimp		0	5/6	3/6
gorilla			0	2/6
orang				0

(symmetrical, lower left not filled in)

- Many different metrics can be used !!
- Try to build the tree whose distances best match the real distances.

Trees and distances



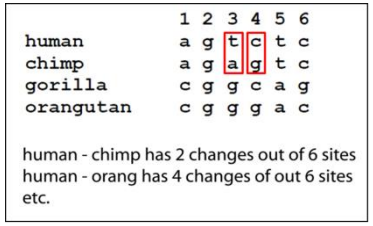
- Try to build the tree whose distances best match the real distances.

Best Match?

- "Best match" based on **least squares** of real pairwise distances compared to the tree distances:

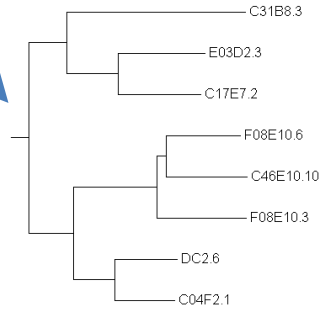
Let D_m be the measured distances.
Let D_t be the tree distances.
Find the tree that minimizes:

$$\sum_{i=1}^{N} \left(D_t - D_m \right)^2$$



	human	chimp	gorilla	orang
human	0	2/6	4/6	4/6
chimp		0	5/6	3/6
gorilla			0	2/6
orang				0

(symmetrical, lower left not filled in)



Enumerate and score all trees?

- **How about the following algorithm:**

1) *Construct all possible trees*
2) Fit least-squares best distances for each topology
3) *Pick the tree with the best score*

- Not used for distance trees - there is a much faster way to get very close to correct.

The UPGMA algorithm

1) generate a table of pairwise sequence distances and assign each sequence to a list of N tree nodes.

2) look through current list of nodes (initially these are all leaf nodes) for the pair with the smallest distance.

3) merge the closest pair, remove the pair of nodes from the list and add the merged node to the list.

4) repeat until only one node left in list - it is the root.

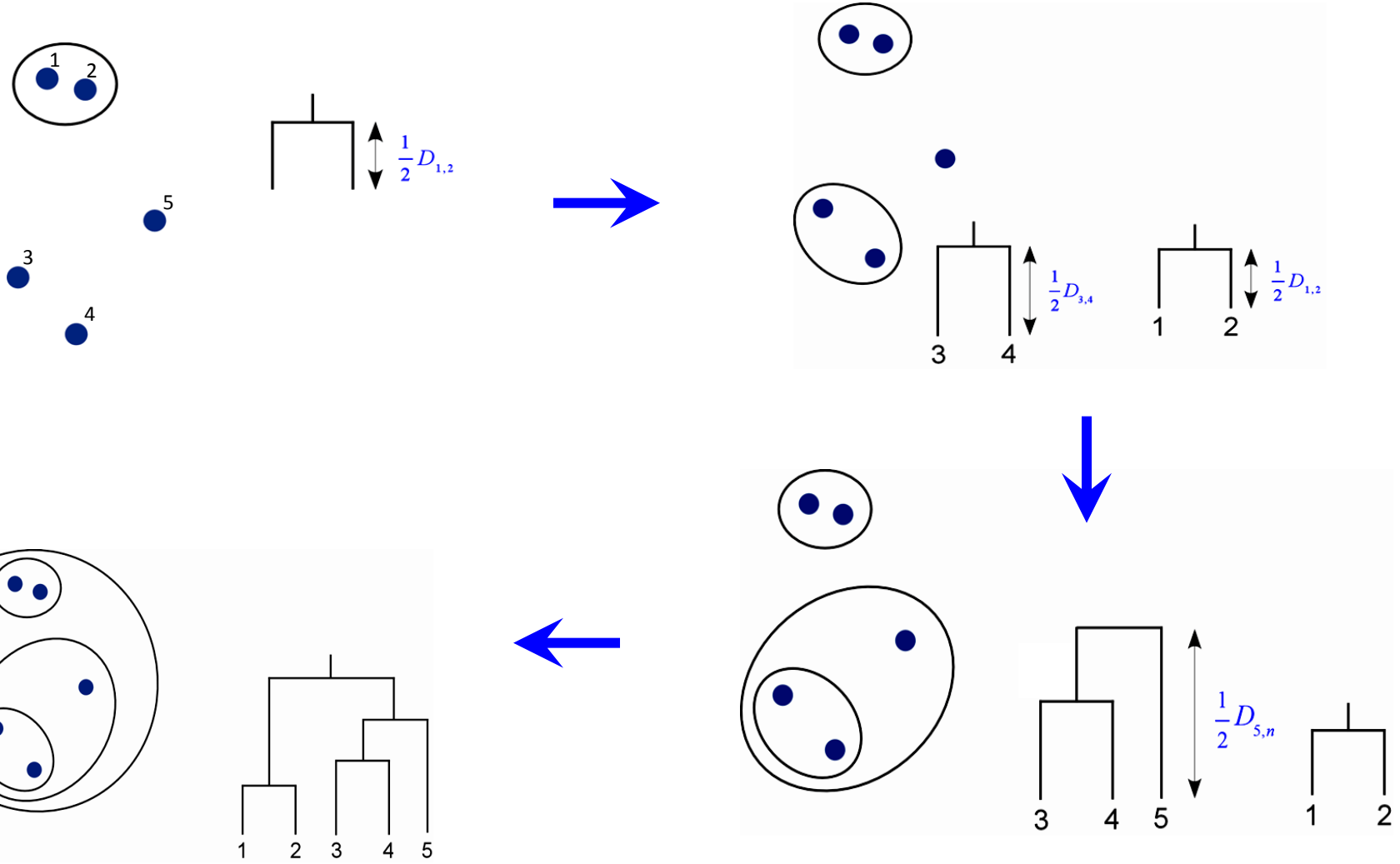
$$D_{n1,n2} = \frac{1}{N} \sum_i \sum_j d_{ij}$$

where i is each leaf of $n1$ (node1), j is each leaf of $n2$ (node2), and N is the number of distances summed

definition of distance

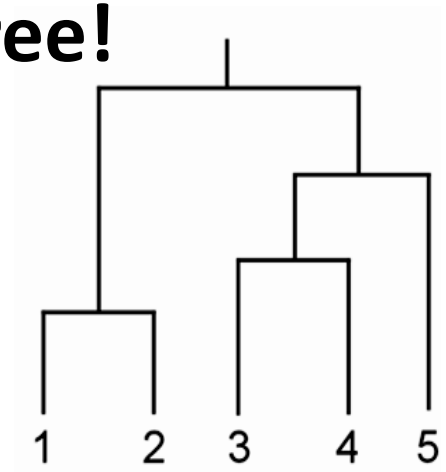
(in words, this is just the arithmetic average of the distances between all the leaves in one node and all the leaves in the other node)

UPGMA (**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean)

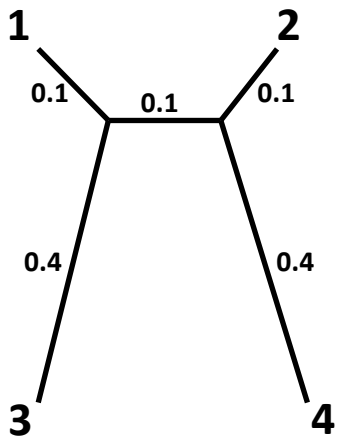


The Molecular Clock

- **UPGMA assumes a constant rate of the molecular clock across the entire tree!**
  - The sum of times down a path to any leaf is the same



- This assumption may not be correct … and will lead to incorrect tree reconstruction.



Neighbor-Joining (NJ) Algorithm

- Essentially similar to UPGMA, but correction for distance to other leaves is made.

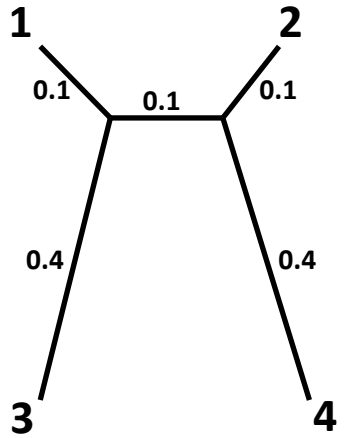
- Specifically, for sets of leaves *i* and *j*, we denote the set of all **other** leaves as *L*, and the size of that set as $|L|$, and we compute the corrected distance D_{ij} as:

$$D_{ij} = d_{ij} - (r_i + r_j)$$

where

$$r_i = \frac{1}{|L|} \sum_{k \in L} d_{ik}$$

(the mean distance from i to all 'other' leaves)

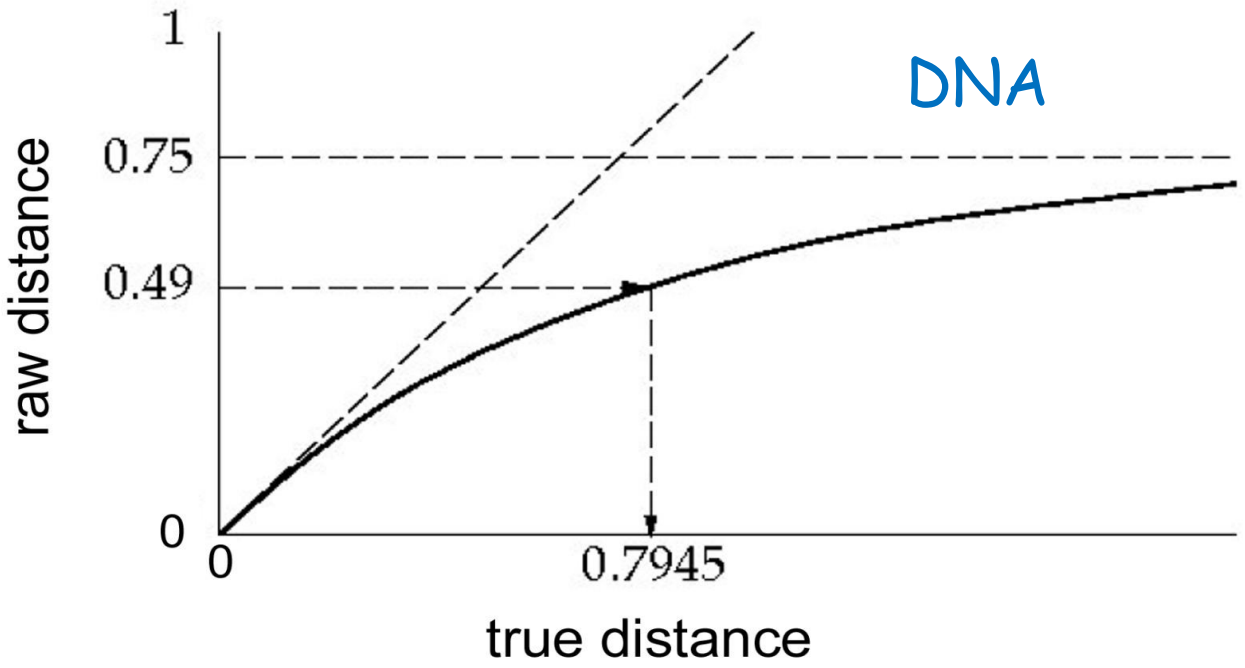


Raw distance correction

- As two DNA sequences diverge, it is easy to see that their maximum raw distance is ~0.75 (assuming equal nt frequencies, $\frac{1}{4}$ of residues will be identical even if unrelated sequences).

- We would like to use the "true" distance, rather than raw distance.

- This graph shows evolutionary distance related to raw distance:



Jukes-Cantor model

Jukes-Cantor model:

$$D = -\frac{3}{4}\ln\left(1 - \frac{4}{3}D_{raw}\right)$$

D_{raw} is the raw distance (what we directly measure)

D is the corrected distance (what we want)

Distance trees - summary

- Convert each pairwise raw distance to a corrected distance.

- Build tree as before (UPGMA algorithm).

- Notice that these methods don't need to consider all tree topologies - they are very fast, even for large trees.

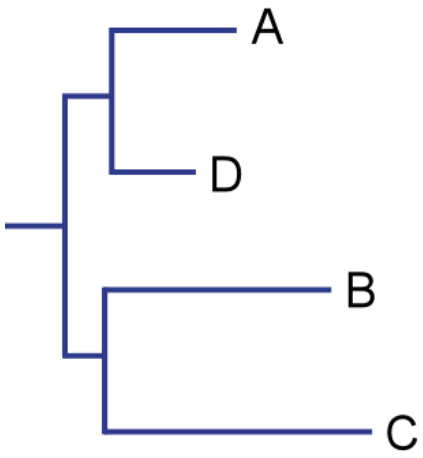
Trees and Python

Representing a tree in Python

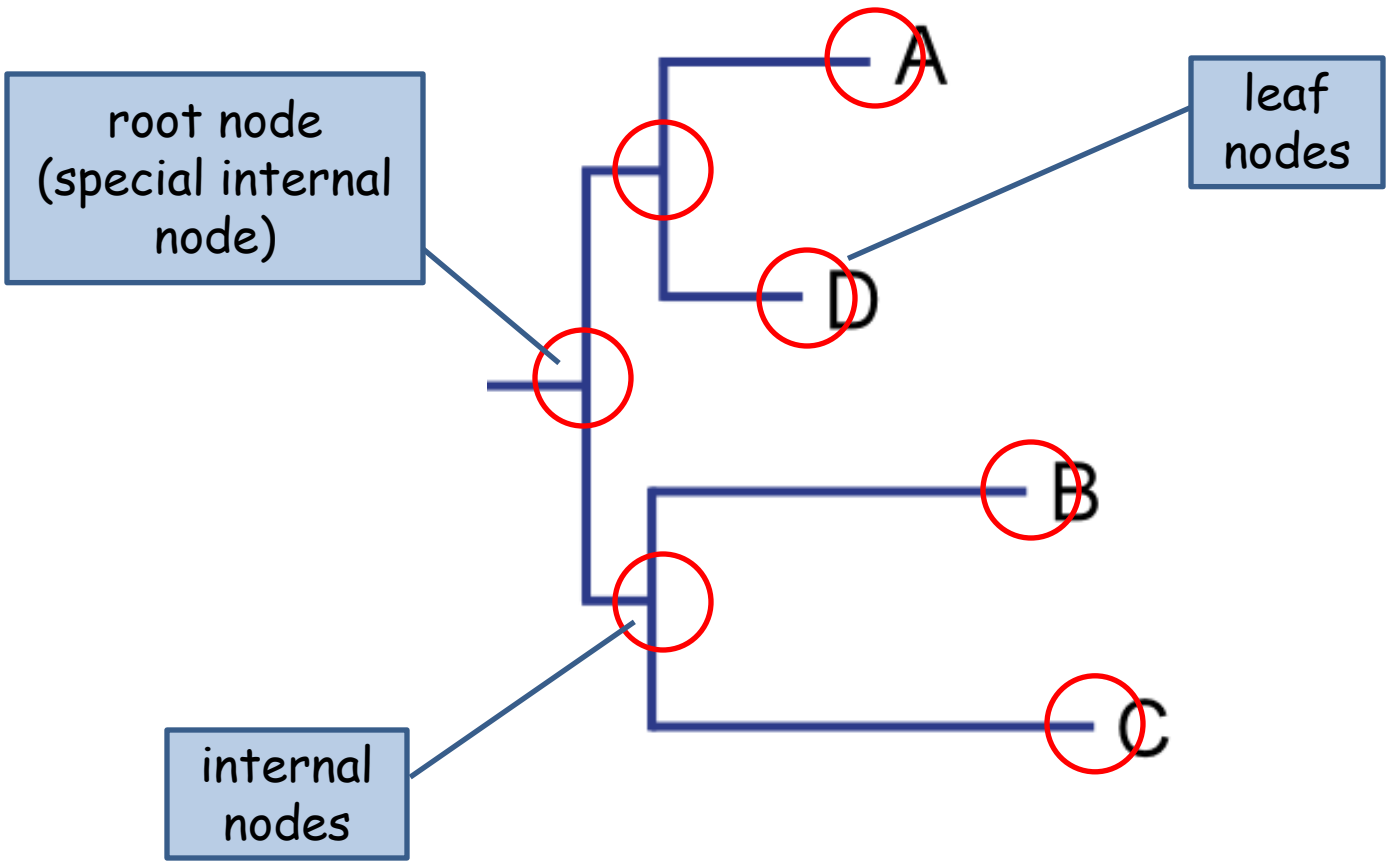
Some bioinformatic entities are easy to represent with standard Python types, e.g. :

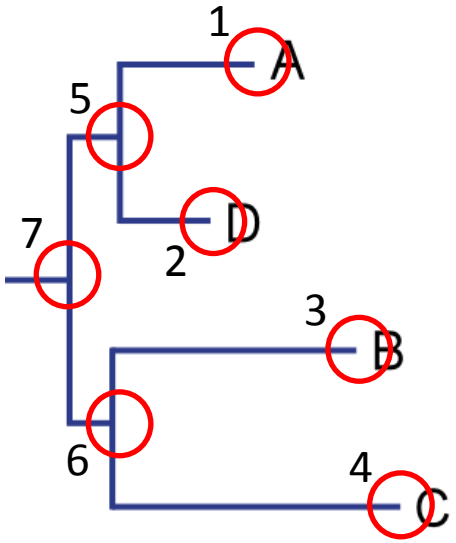
- Protein or DNA sequence
- Alignment score
- Sequence names paired with scores (or other things)

How would you represent a tree??



Natural approach - represent tree nodes





tree nodes numbered for reference

What kinds of information should we associate with nodes?

1) A sequence name (for leaf nodes)
2) A distance to parent (except for the root)
3) Connections to other nodes