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

VIEWPOINT

The Proposal to Lower *P* Value Thresholds to .005

John P. A. Ioannidis, MD, DSc
Stanford Prevention Research Center, Meta-Research Innovation Center at Stanford, Departments of Medicine, Health Research and Policy, Biomedical Data Science, and Statistics, Stanford University, Stanford, California.

P values and accompanying methods of statistical significance testing are creating challenges in biomedical science and other disciplines. The vast majority (96%) of articles that report *P* values in the abstract, full text, or both include some values of .05 or less.[1] However, many of the claims that these reports highlight are likely false.[2] Recognizing the major importance of the statistical significance conundrum, the American Statistical Association (ASA) published[3] a statement on *P* values in 2016. The status quo is widely believed to be problematic, but how exactly to fix the problem is far more contentious. The contributors to the ASA statement also wrote 20 independent, accompanying commentaries focusing on different aspects and prioritizing different solutions. Another large coalition of 72 methodologists recently proposed[4] a specific, simple move: lowering the routine *P* value threshold for claiming statistical significance from .05 to .005 for new discoveries. The proposal met with strong endorsement in some circles and concerns in others.

P values are misinterpreted, overtrusted, and misused. The language of the ASA statement enables the dissection of these 3 problems. Multiple misinterpretations of *P* values exist, but the most common one is that they represent the "probability that the studied hypothesis is true."[3] A *P* value of .02 (2%) is wrongly considered to mean that the null hypothesis (eg, the drug is as effective as placebo) is 2% likely to be true and the alternative (eg, the drug is more effective than placebo) is 98% likely to be correct. Overtrust ensues when it is forgotten that "proper inference requires full reporting and transparency."[3] Better-looking (smaller) *P* values alone do not guarantee full reporting and transparency. In fact, smaller *P* values may hint to selective reporting and nontransparency. The most common misuse of the *P* value is to make "scientific conclusions and business or policy decisions" based on "whether a *P* value passes a specific threshold" even though "a *P* value, or statistical significance, does not measure the size of an effect or the importance of a result," and "by itself, a *P* value does not provide a good measure of evidence."[3]

These 3 major problems mean that passing a statistical significance threshold (traditionally $P = .05$) is wrongly equated with a finding or an outcome (eg, an association or a treatment effect) being true, valid, and worth acting on. These misconceptions affect researchers, journals, readers, and users of research articles, and even media and the public who consume scientific information. Most claims supported with *P* values slightly below .05 are probably false (ie, the claimed associations and treatment effects do not exist). Even among those claims that are true, few are worth acting on in medicine and health care.

Lowering the threshold for claiming statistical significance is an old idea. Several scientific fields have carefully considered how low a *P* value should be for a research finding to have a sufficiently high chance of being true. For example, adoption of genome-wide significance thresholds ($P < 5 \times 10^{-8}$) in population genomics has made discovered associations highly replicable and these associations also appear consistently when tested in new populations. The human genome is very complex, but the extent of multiplicity of significance testing involved is known, the analyses are systematic and transparent, and a requirement for $P < 5 \times 10^{-8}$ can be cogently arrived at.

However, for most other types of biomedical research, the multiplicity involved is unclear and the analyses are nonsystematic and nontransparent. For most observational exploratory research that lacks preregistered protocols and analysis plans, it is unclear how many analyses were performed and what various analytic paths were explored. Hidden multiplicity, nonsystematic exploration, and selective reporting may affect even experimental research and randomized trials. Even though it is now more common to have a preexisting protocol and statistical analysis plan and preregistration of the trial posted on a public database, there are still substantial degrees of freedom regarding how to analyze data and outcomes and what exactly to present. In addition, many studies in contemporary clinical investigation focus on smaller benefits or risks; therefore, the risk of various biases affecting the results increases.

Moving the *P* value threshold from .05 to .005 will shift about one-third of the statistically significant results of past biomedical literature to the category of just "suggestive."[1] This shift is essential for those who believe (perhaps crudely) in black and white, significant or nonsignificant categorizations. For the vast majority of past observational research, this recategorization would be welcome. For example, mendelian randomization studies show that only few past claims from observational studies with $P < .05$ represent causal relationships.[5] Thus, the proposed reduction in the level for declaring statistical significance may dismiss mostly noise with relatively little loss of valuable information. For randomized trials, the proportion of true effects that emerge with *P* values in the window from .005 to .05 will be higher, perhaps the majority in several fields. However, most findings would not represent treatment effects that are large enough for outcomes that are serious enough to make them worthy of further action. Thus, the reduction in the *P* value threshold may largely do more good than harm, despite also removing an occasional true and useful treatment effect from the coveted significance zone. Regardless, the need for also focusing on the magnitude of all treatment effects and their uncertainty (such as with confidence intervals) cannot be overstated.

Lowering the threshold of statistical significance is a temporizing measure. It would work as a dam that could

Corresponding Author: John P. A. Ioannidis, MD, DSc, Stanford Prevention Research Center, 1265 Welch Rd, Medical School Office Building, Room X306, Stanford, CA 94305 (jioannid@stanford.edu).

FROM THE ABSTRAT:

P values and accompanying methods … are creating challenges in biomedical science … ***However, many of the claims that these reports highlight are likely false.*** *Recognizing the major importance of the statistical significance conundrum, the American Statistical Association (ASA) published a statement on P values in 2016.* ***The status quo is widely believed to be problematic, but how exactly to fix the problem is far more contentious.***

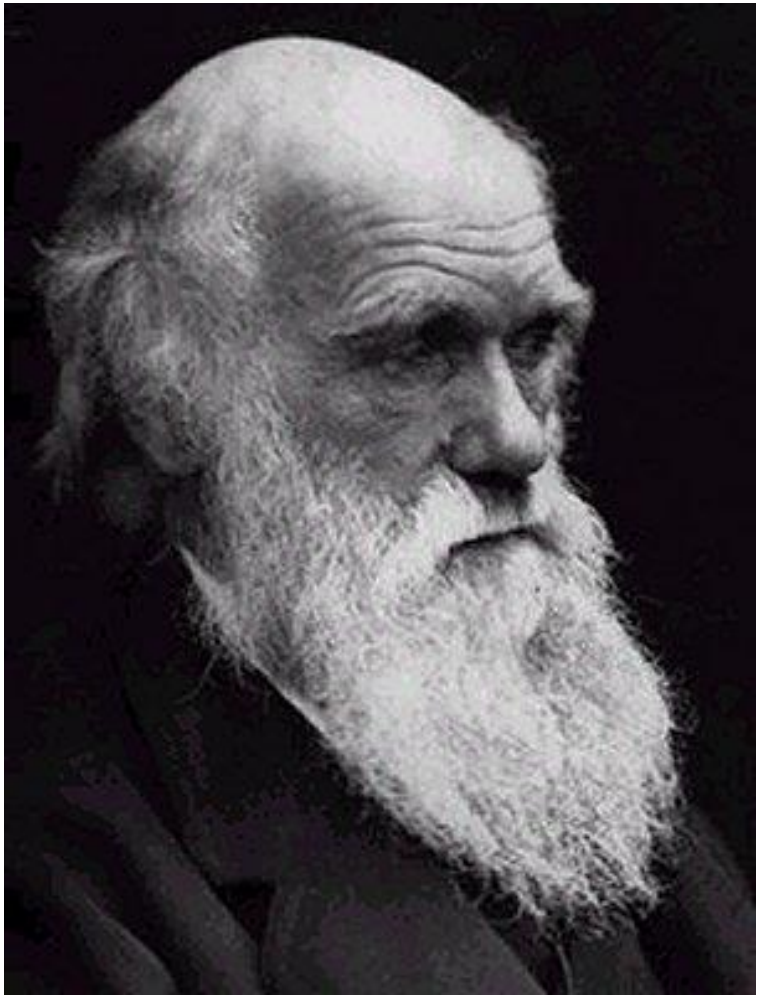
```
Human: CGAAT-CGA-TTCA
Chimp: C-A-TACGAGT-CA
Gorilla: --A-TGCGT-TGCA
Orangutan: --A-TGCGT-GGCA
```

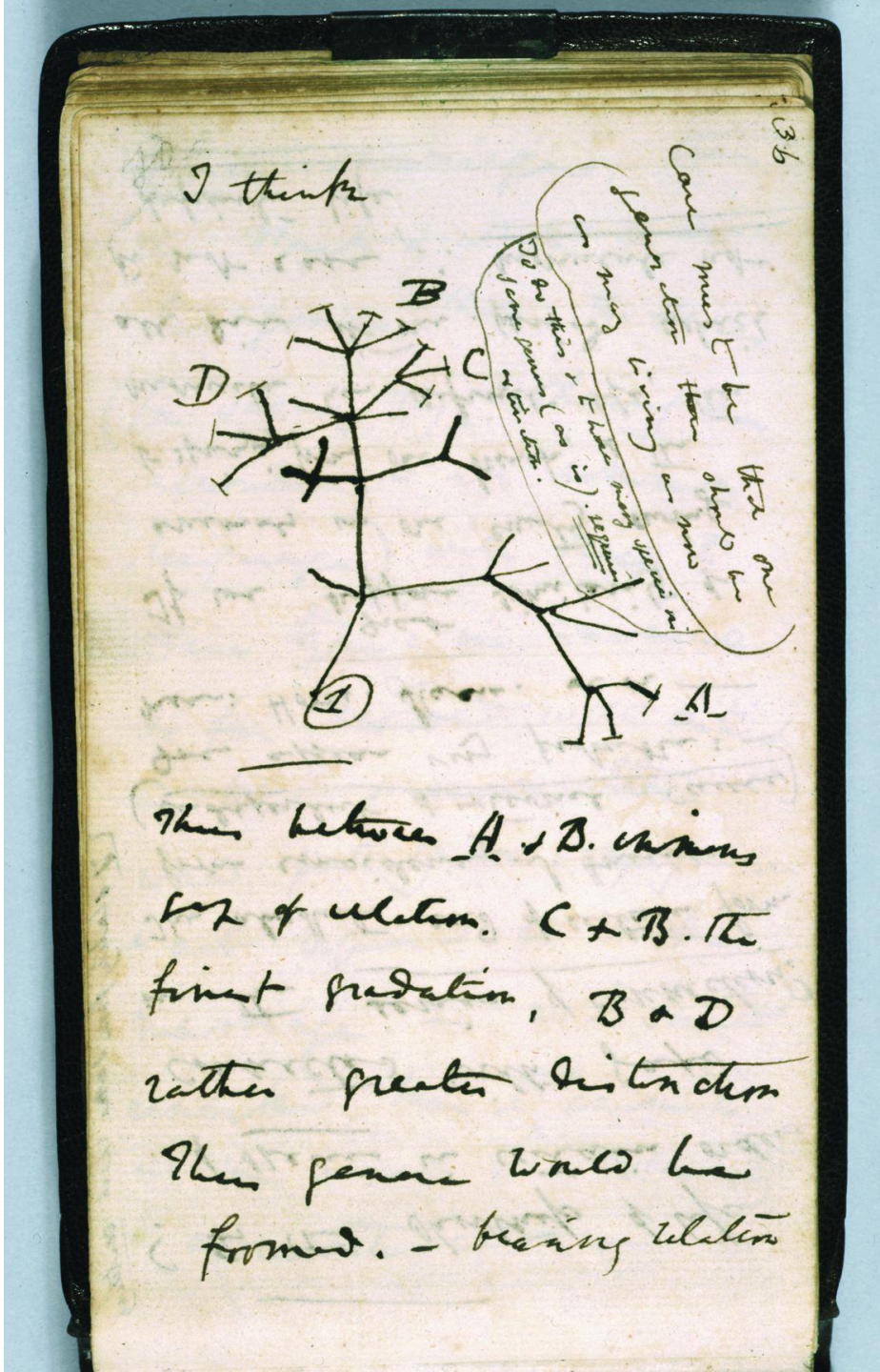
Phylogenetic Trees

Genome 373

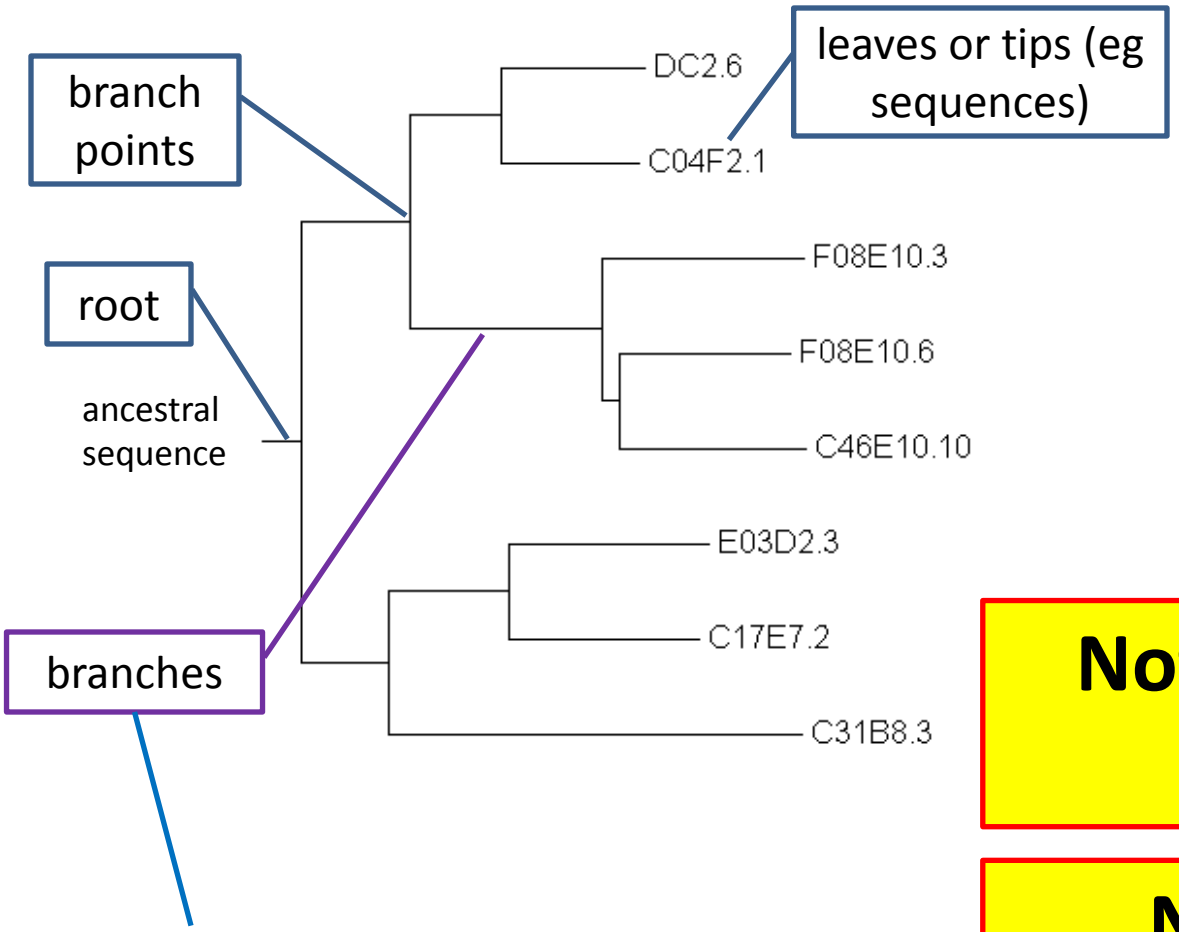
Genomic Informatics

Elhanan Borenstein




36
I think
B
C
D
A
I

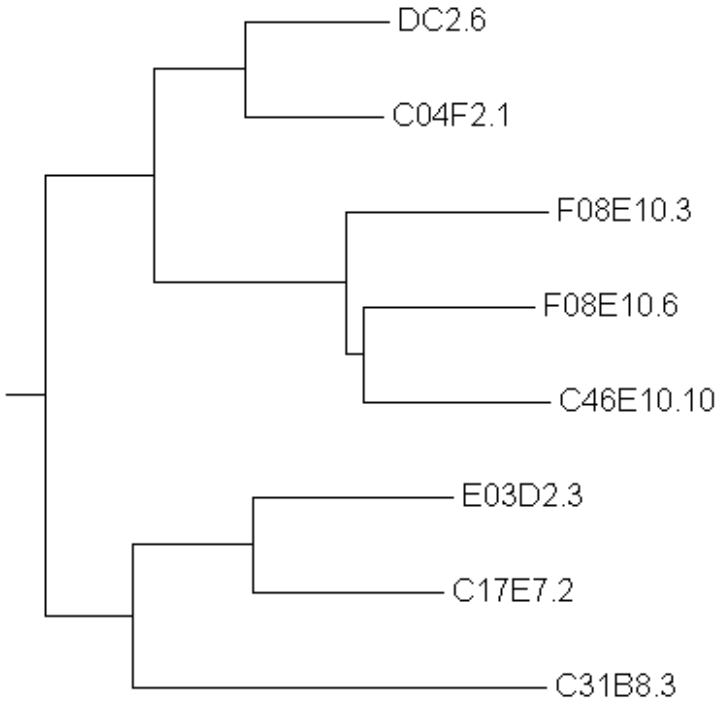
Defining what a “tree” means

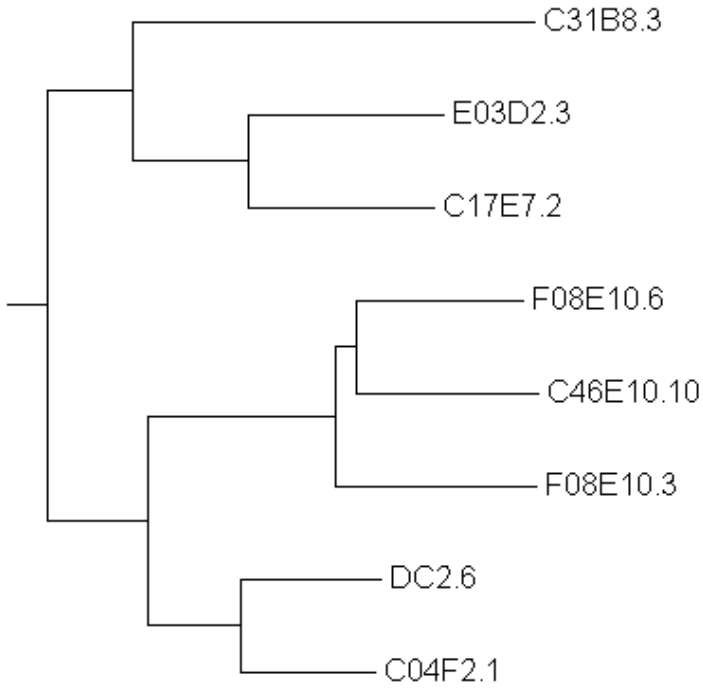


Note: Many drawing practices exist

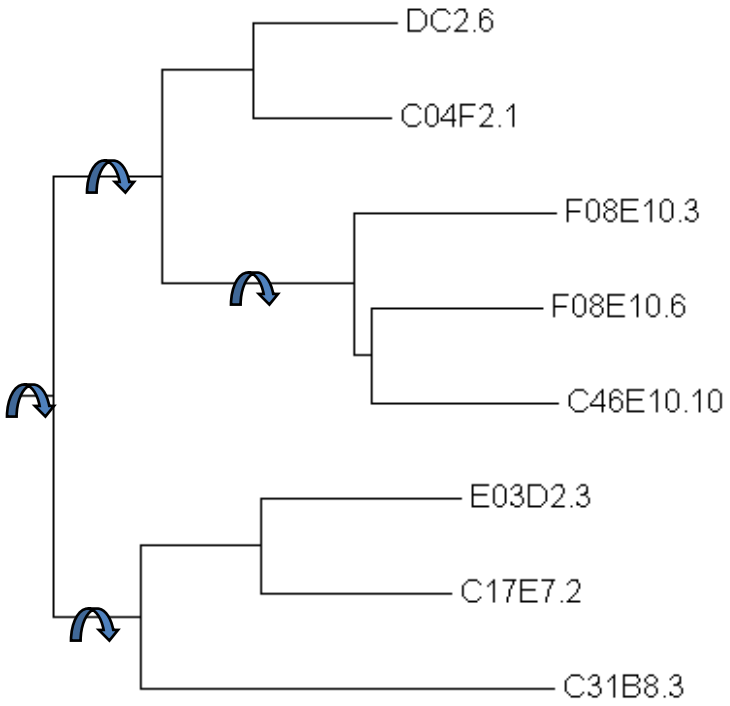
Note: A tree has topology and distances

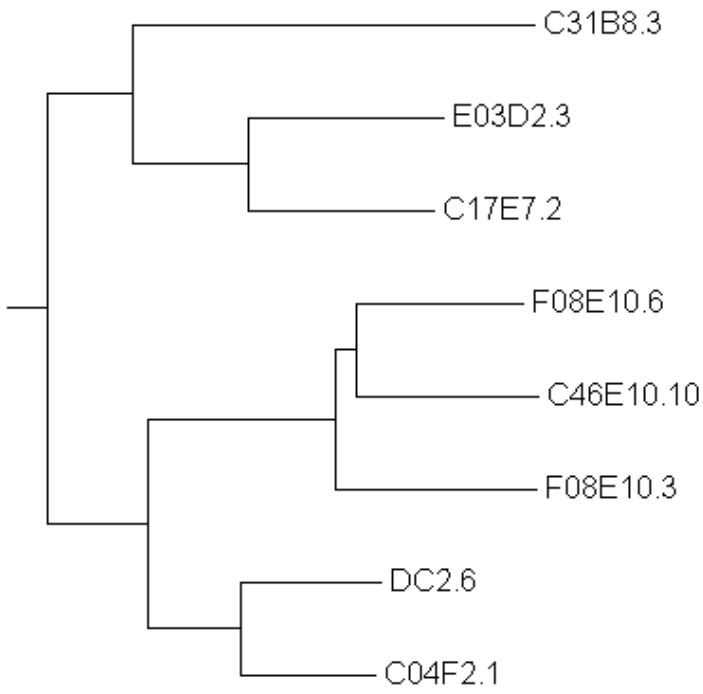
Are these topologically different trees?





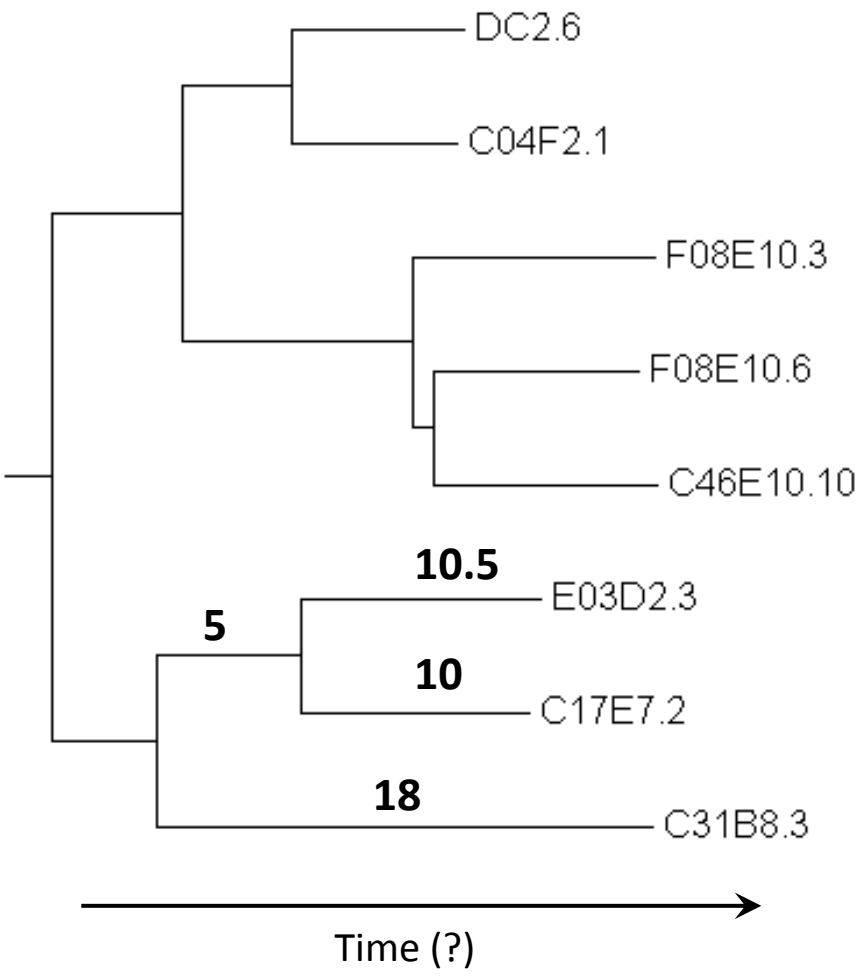
Are these topologically different trees?





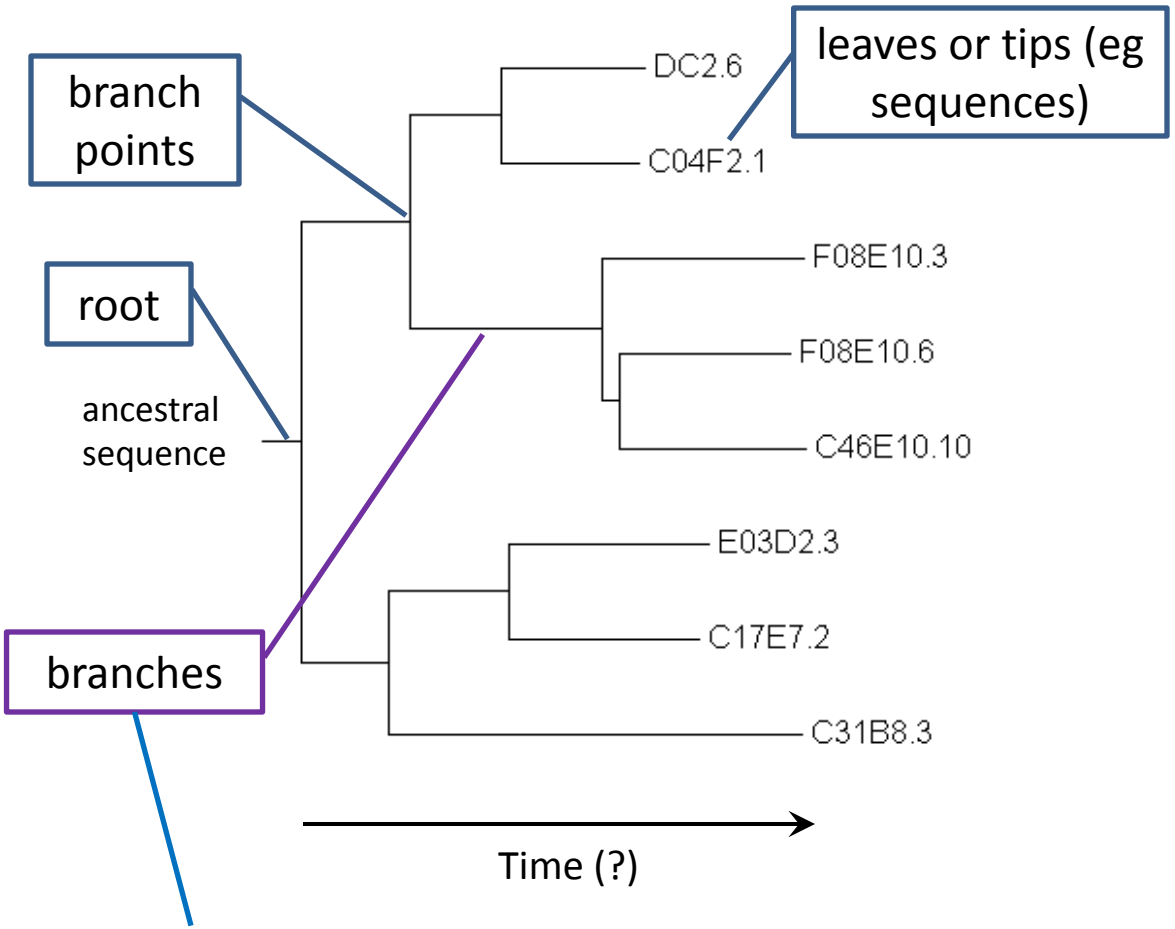
Topologically, these are the SAME tree. In general, two trees are the same if they can be inter-converted by branch rotations.

Branch lengths and evolutionary divergence time



Rooted and unrooted trees

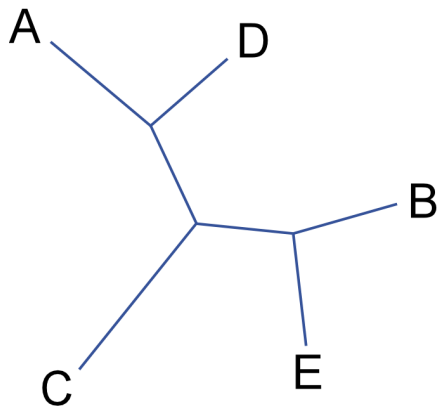
Rooted tree (all real trees are rooted):



… (horizontal) branch lengths sequence is proportional to divergence

Unrooted tree

(used when the root isn't known):

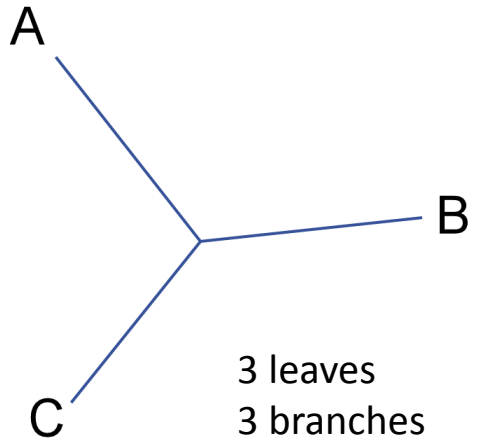


time radiates out from somewhere (probably near the center)

Why is inferring phylogeny a hard problem?

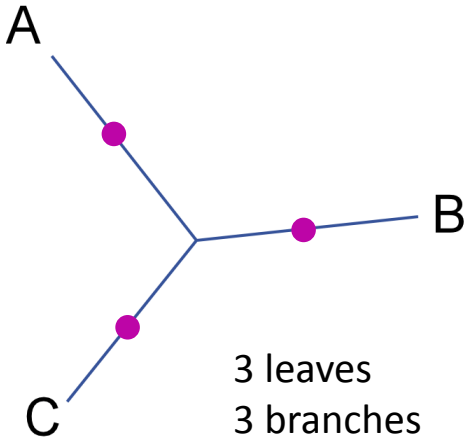
(assume, for example, we are trying to infer the phylogenetic tree for 20 primate species)

The number of tree topologies grows extremely fast

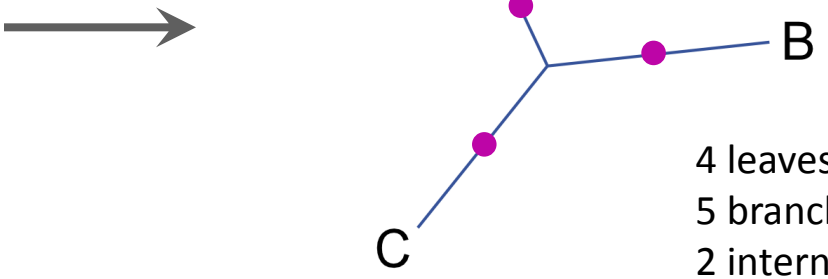


3 leaves
3 branches
1 internal node
1 topology
(3 insertions)

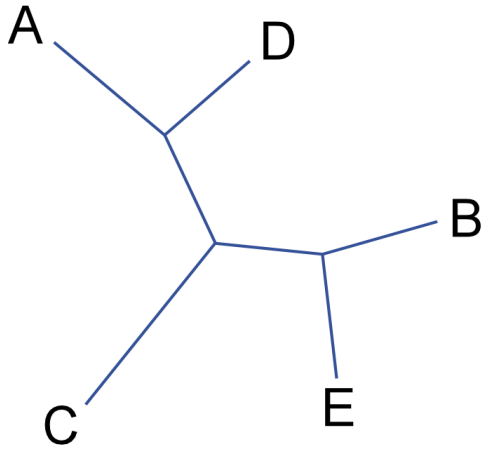
The number of tree topologies grows extremely fast



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

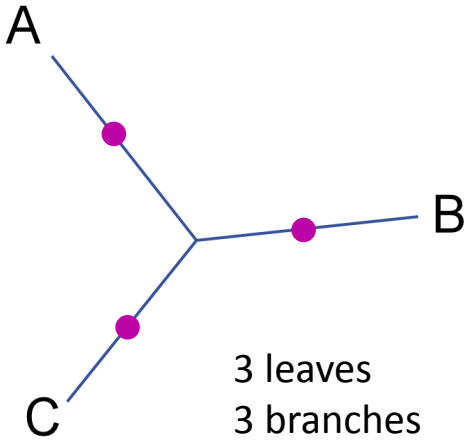


4 leaves
5 branches
2 internal nodes
3 topologies (x3)
(5 insertions)

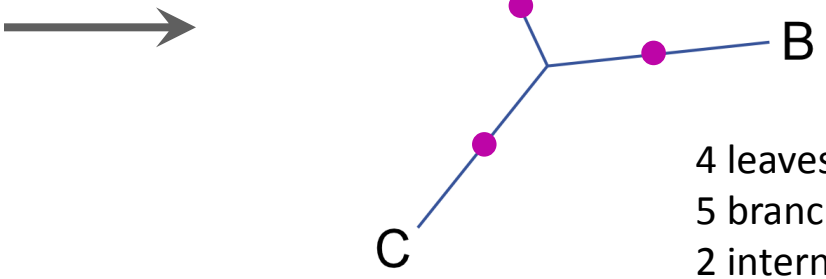


5 leaves
7 branches
3 internal nodes
15 topologies (x5)
(7 insertions)

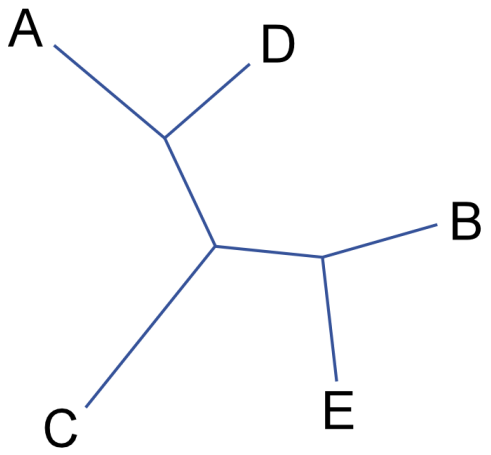
The number of tree topologies grows extremely fast



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



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

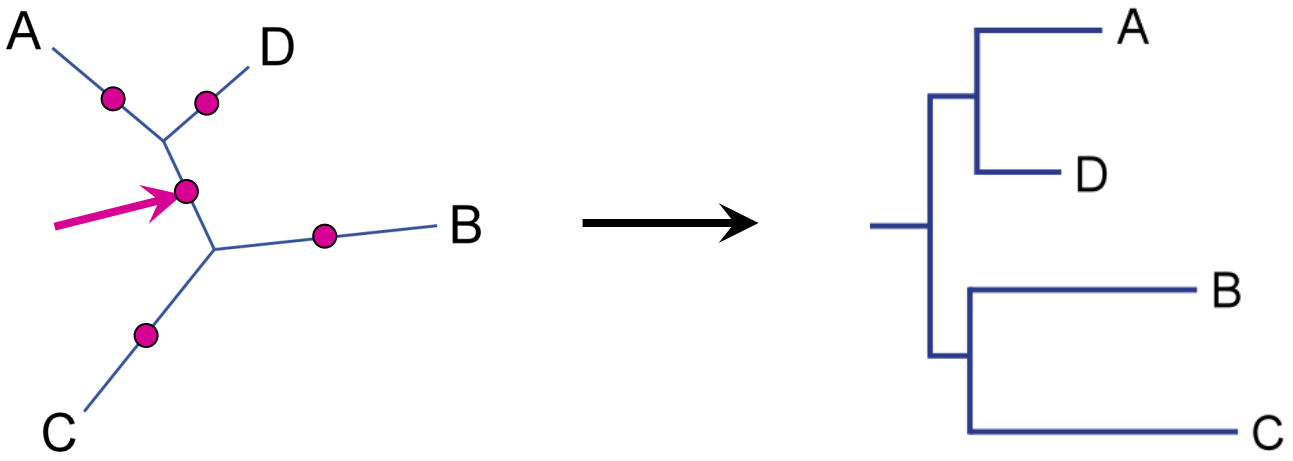


5 leaves
7 branches
3 internal nodes
15 topologies (x5)
(7 insertions)

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

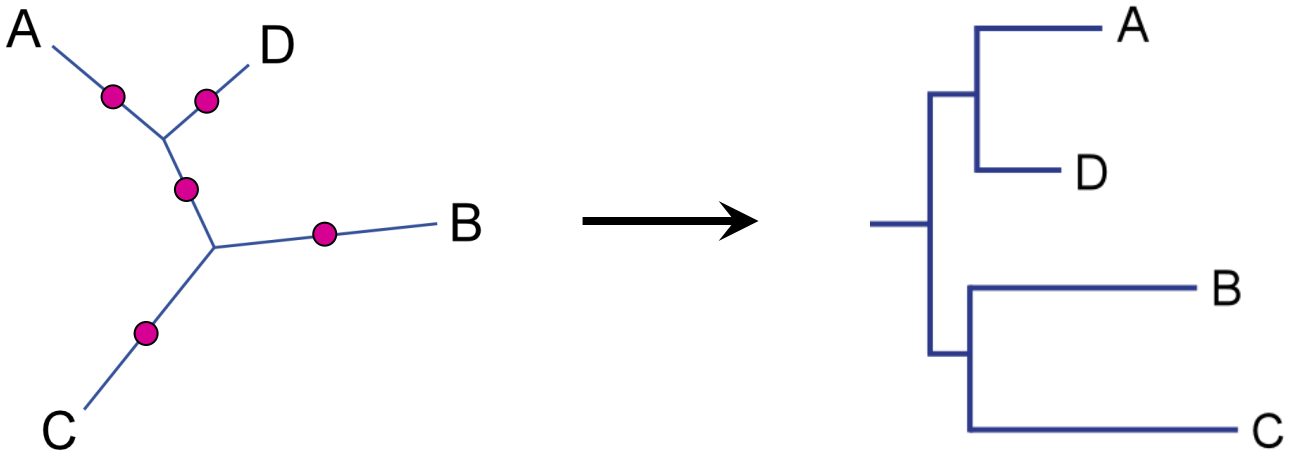
For each unrooted tree, there are 2N - 3 times as many rooted trees, where N is the number of leaves (# branches = 2N – 3).



The number of tree topologies grows extremely fast

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 infer a 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

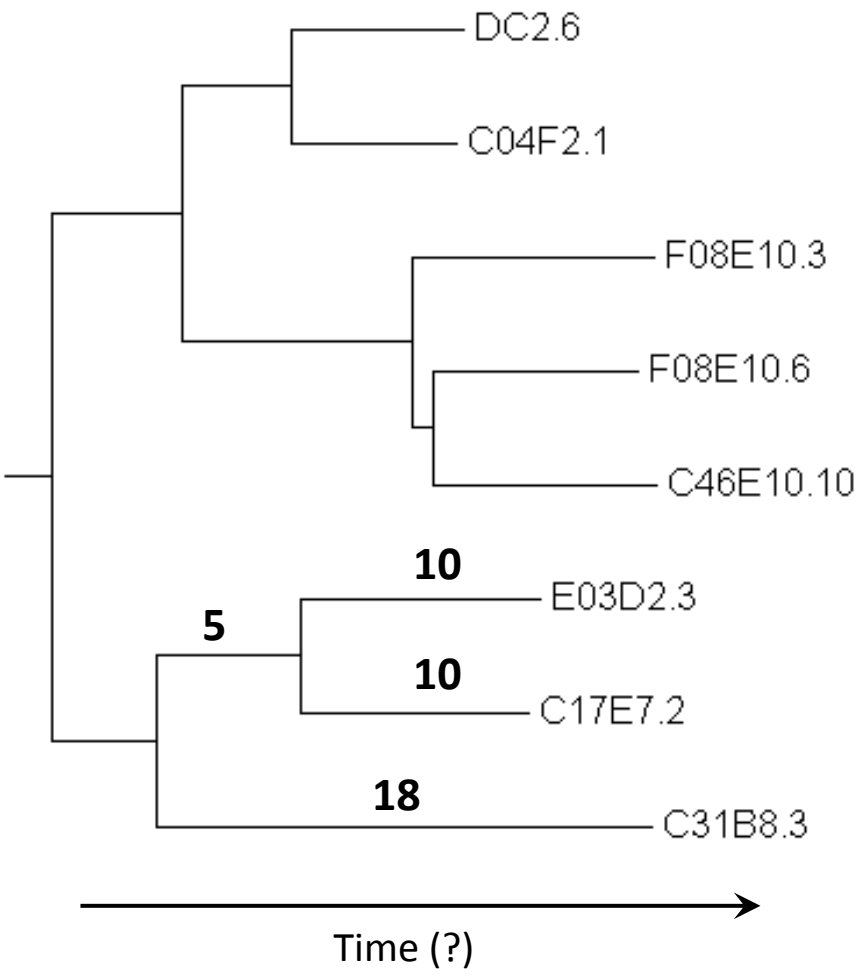
(symmetrical, lower left not filled in)

- **Many different metrics can be used !!**

Approach:

Try to build the tree whose **distances best match** the real distances

Trees and distances



	E03D2.3	C17E7.2	C31B8.3	…
E03D2.3	0	20	33	.
C17E7.2		0	33	.
C31B8.3			0	.
…				0

Best Match?

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

Let D_m be the measured distances.

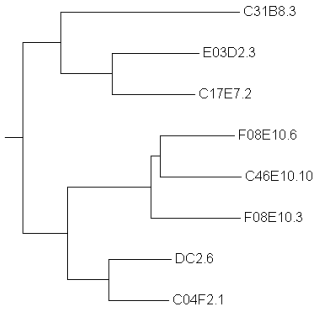
	1	2	3	4	5	6
human	a	g	t	c	t	c
chimp	a	g	a	g	t	c
gorilla	c	g	g	c	a	g
orangutan	c	g	g	g	a	c

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)

Let D_t be the tree distances.



Find the tree that minimizes:

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

Why not enumerate and score all trees?

The UPGMA algorithm

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

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.

The UPGMA algorithm

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

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.

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

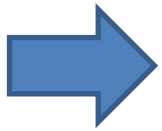
where i is each leaf of $n1$ (node1), j is each leaf of $n2$ (node2), and N is the number of distances summed

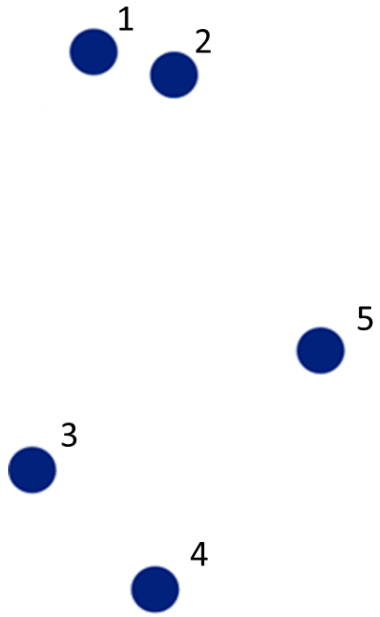
definition of distance

(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)

The UPGMA algorithm

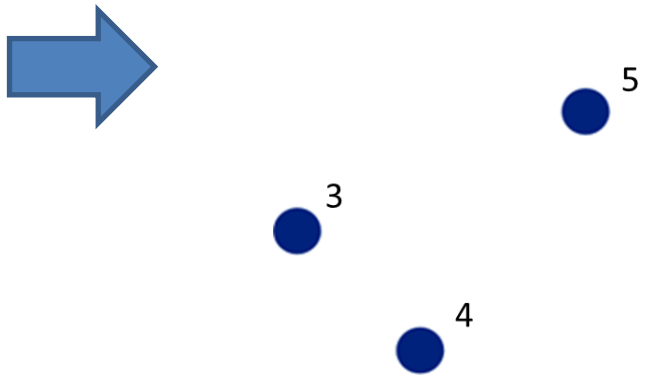
	1	2	3	4	5
1	0	5	18	22	17
2		0	20	24	15
3			0	10	12
4				0	12
5					0





The UPGMA algorithm

	1,2	**3,4,5**
1,2	0	19.33
3,4,5		0

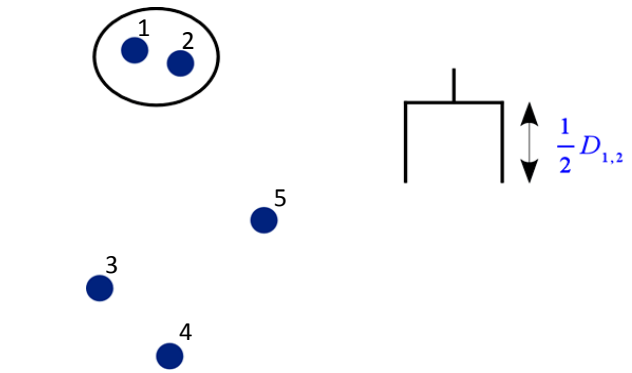


UPGMA

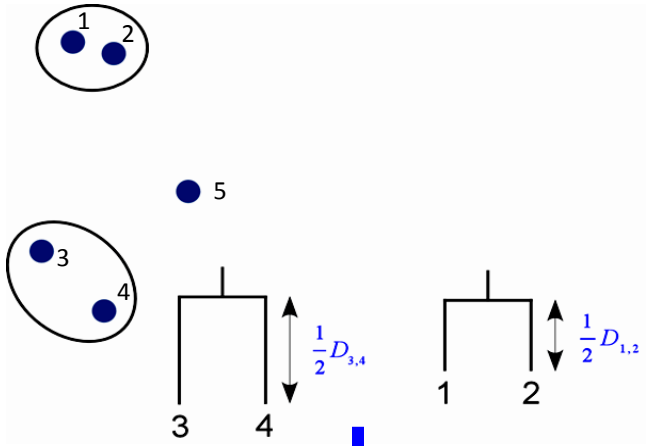
(**U**nweighted **P**air **G**roup **M**ethod with **A**rithmetic Mean)

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, <u>remove</u> the pair of nodes from the list and <u>add</u> the merged node to the list.
4) repeat until only one node left in list - it is the root.

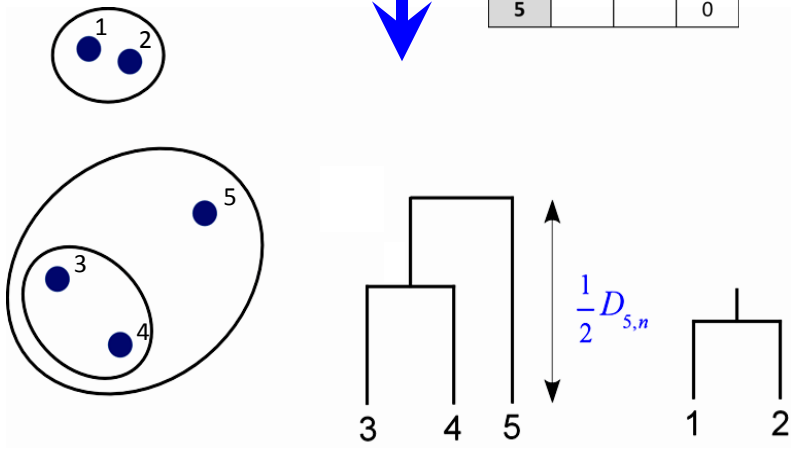
	1	2	3	4	5
1	0	5	18	22	17
2		0	20	24	15
3			0	10	12
4				0	12
5					0



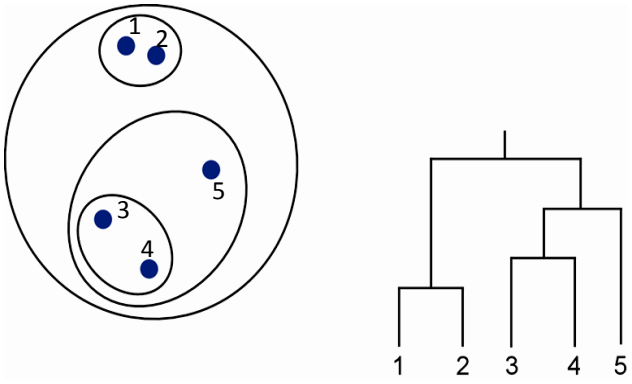
	1,2	3	4	5
1,2	0	19	23	16
3		0	10	12
4			0	12
5				0



	1,2	3,4	5
1,2	0	21	16
3,4		0	12
5			0

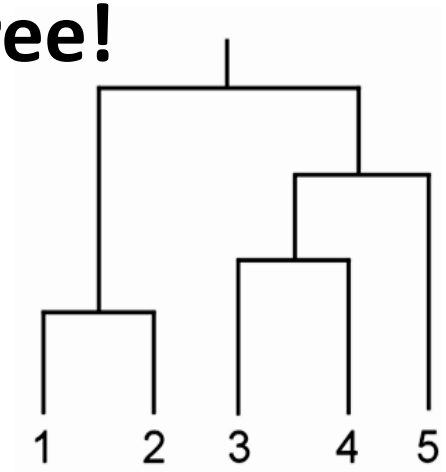


	1,2	3,4,5
1,2	0	19.33
3,4,5		0

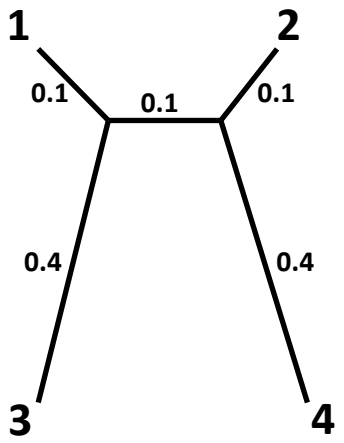


The Molecular Clock

- **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.

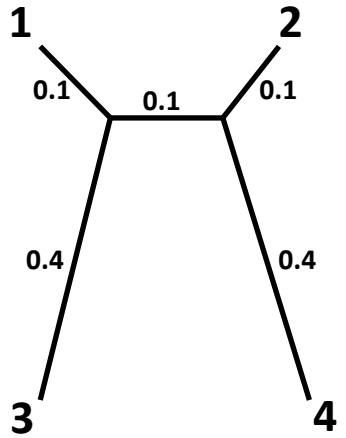
- Specifically, for sets of leaves *i* and *j*, we denote the set of all **other** leaves as *L*, and the size of that set as $|L|$, and we compute the corrected distance D_{ij} as:

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

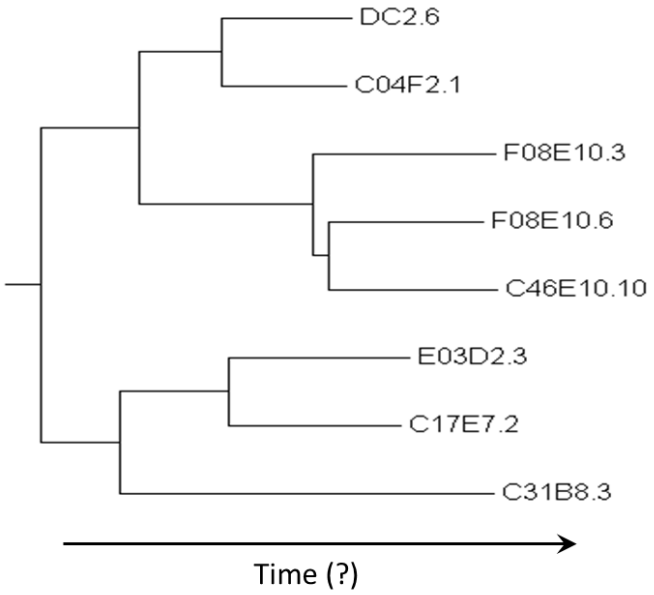
where

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

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



But wait, there's one more problem



VS

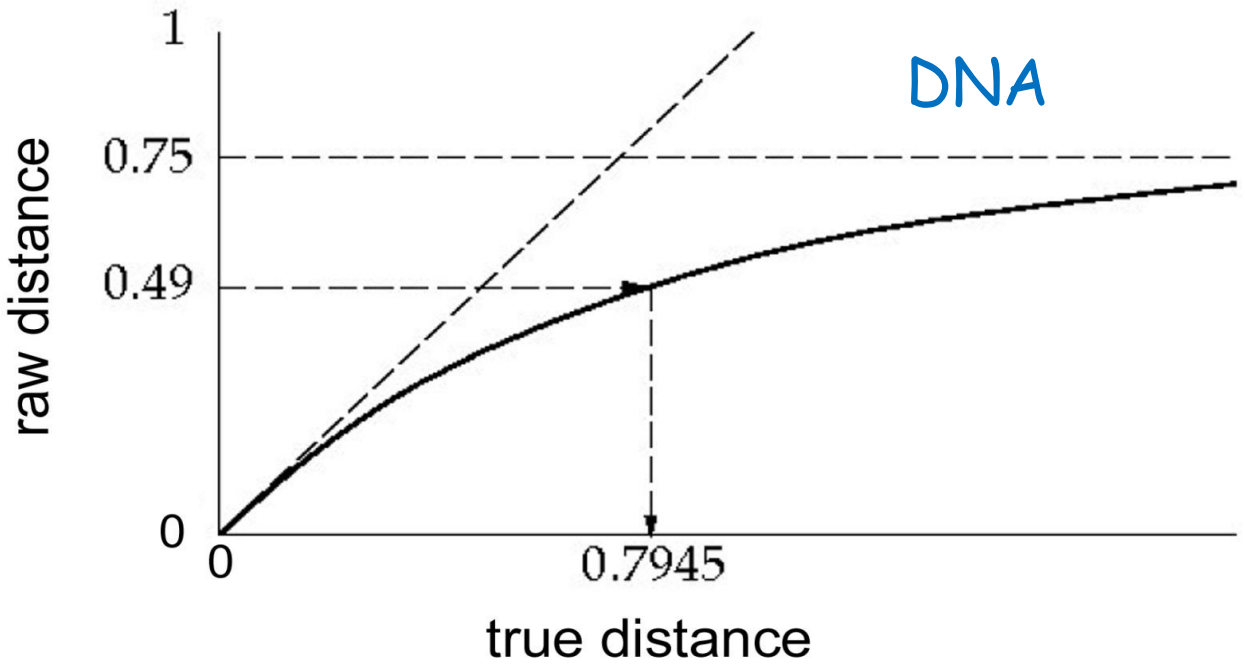
	E03D2.3	C17E7.2	C31B8.3	…
E03D2.3	0	20	33	.
C17E7.2		0	33	.
C31B8.3			0	.
…				0

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, ¼ 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.