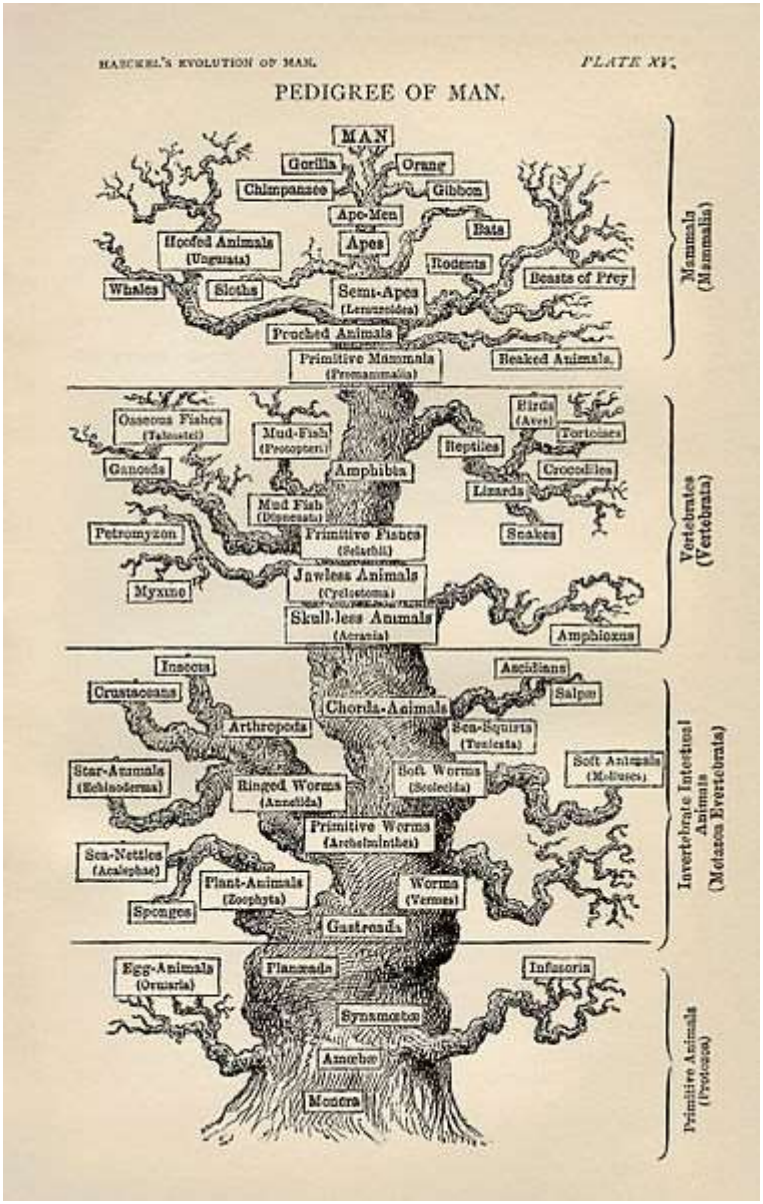
Phylogenetics

Pradipta Ray,

BIOL 6385 / BMEN 6389,

The University of Texas at Dallas

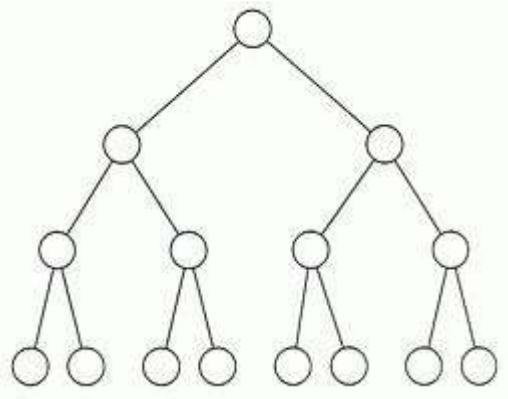
(some material based on content by PR in Eric Xing's 10-810 Carnegie Mellon class)

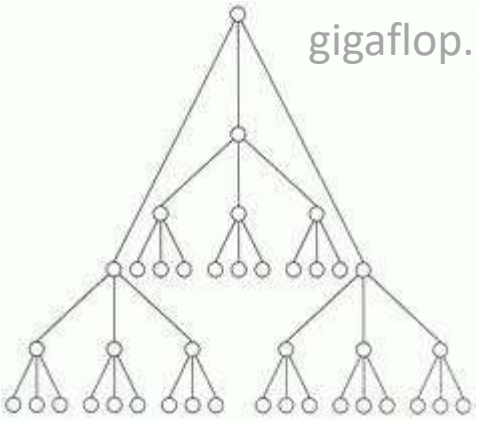


Ernst Haeckel, Tree of life, 1879

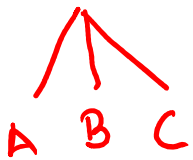
Some nomenclature on trees

- Binary (bifurcating) or multiway (multifurcating)



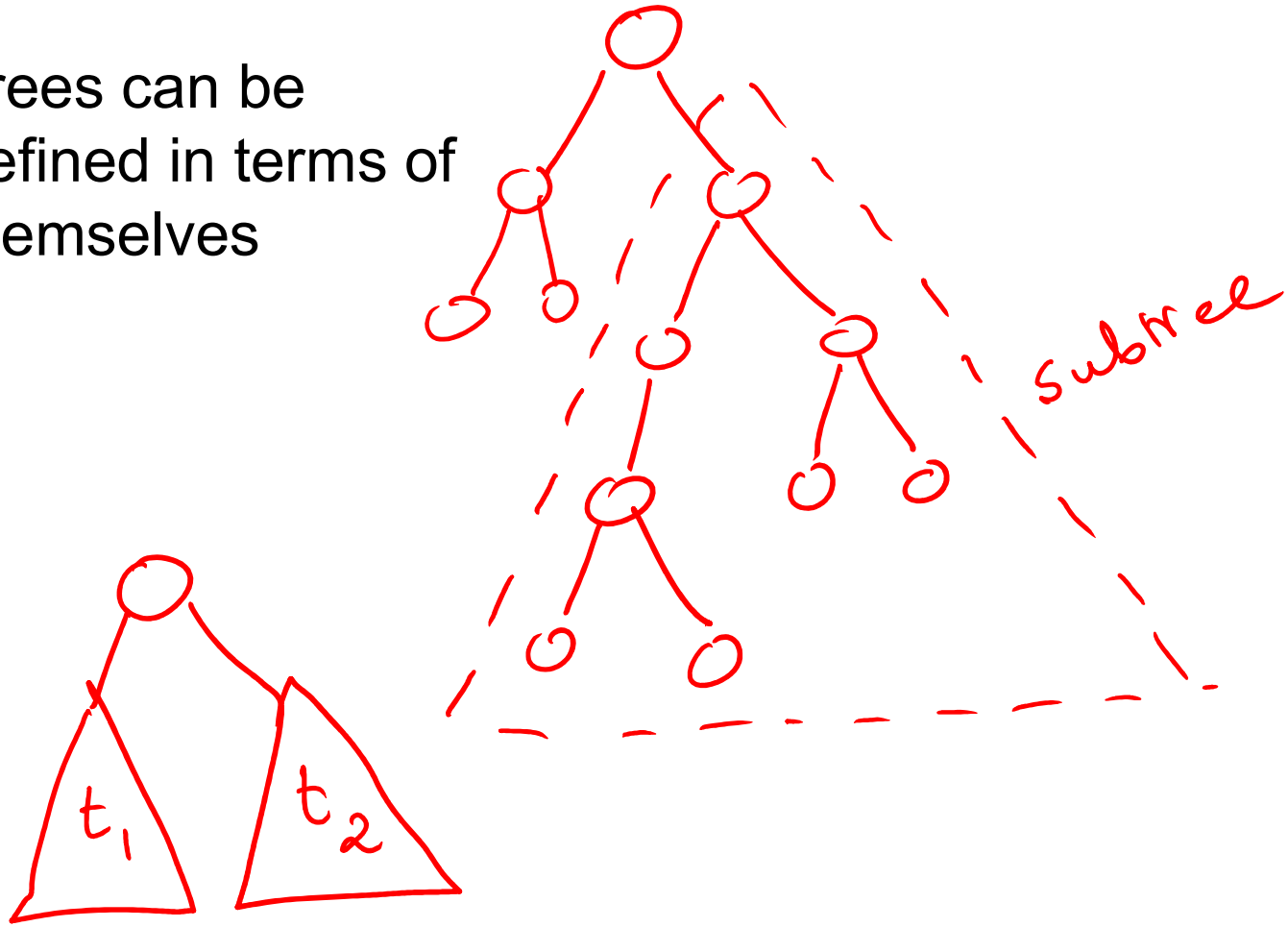


- A multiway split can be envisioned as a series of binary splits

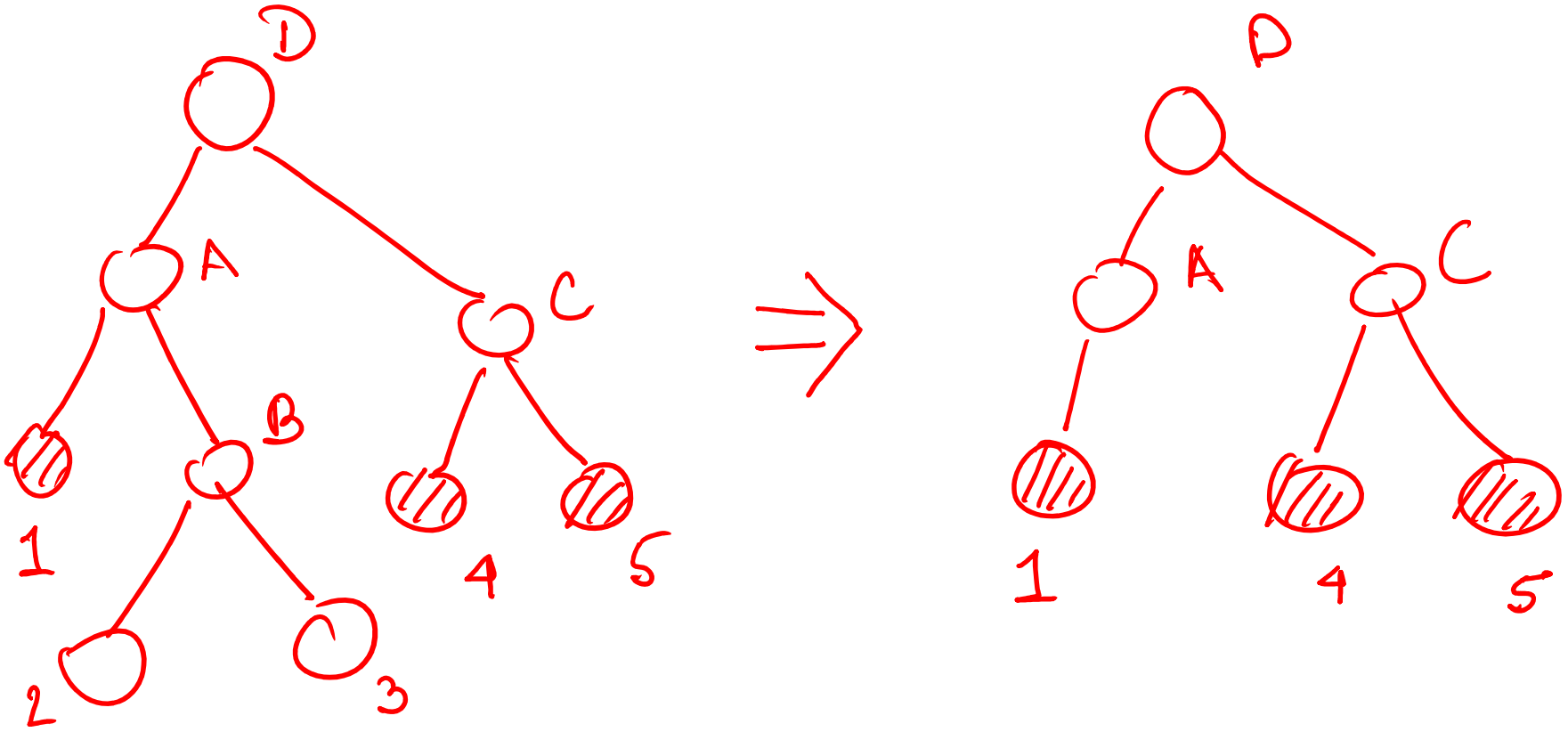


Recursively defining trees : subtrees

Trees can be defined in terms of themselves



Induced subtree



- For a subset of nodes, the induced subtree is the minimal set of edges and nodes to connect them together, taken from the original tree

Parameterizing a tree

- Topology : connectivity of the tree
  - path between any two nodes is unique !
  - topology unchanged by “squishing” a drawing of the tree
- Length of edges : the geometry of the tree
  - notion of “distance” between neighboring nodes
- Labels of nodes (?) : which nodes were not born equal : identifying the ones we are interested in

Rooted trees

- "Rooting" : a notion of direction of flow
  - in case of phylogenies, flow of time
- Root : a privileged node
- Direction flows outward from the root
- Rooted tree = Directed edges away from root

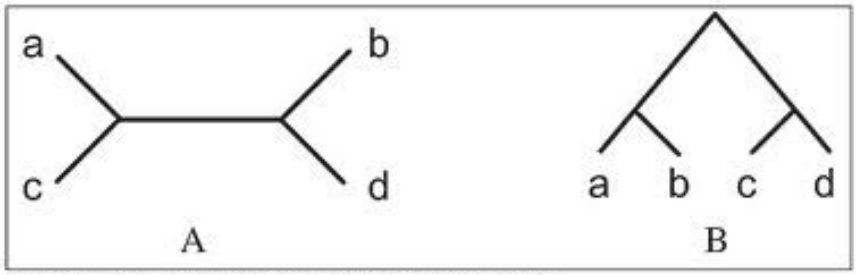


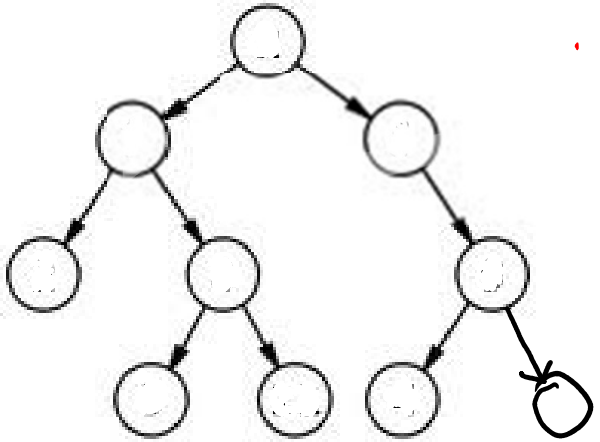
Figure 1. Topologies of phylogenetic trees: A. unrooted tree, B. rooted tree.

Perrero & Lopez, Gen Mol Res, Sep 2005

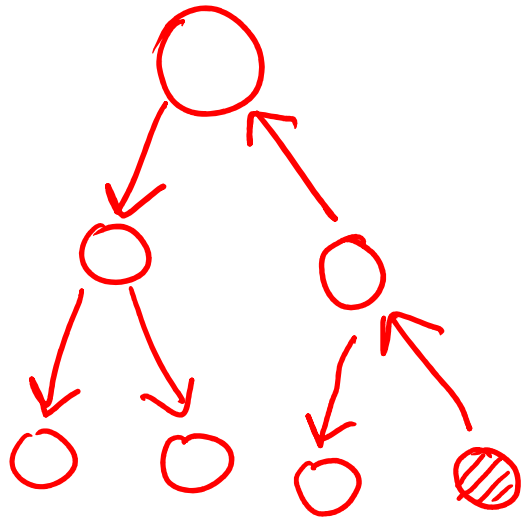
Directed edges

- Directed edges away from root

  = Rooting

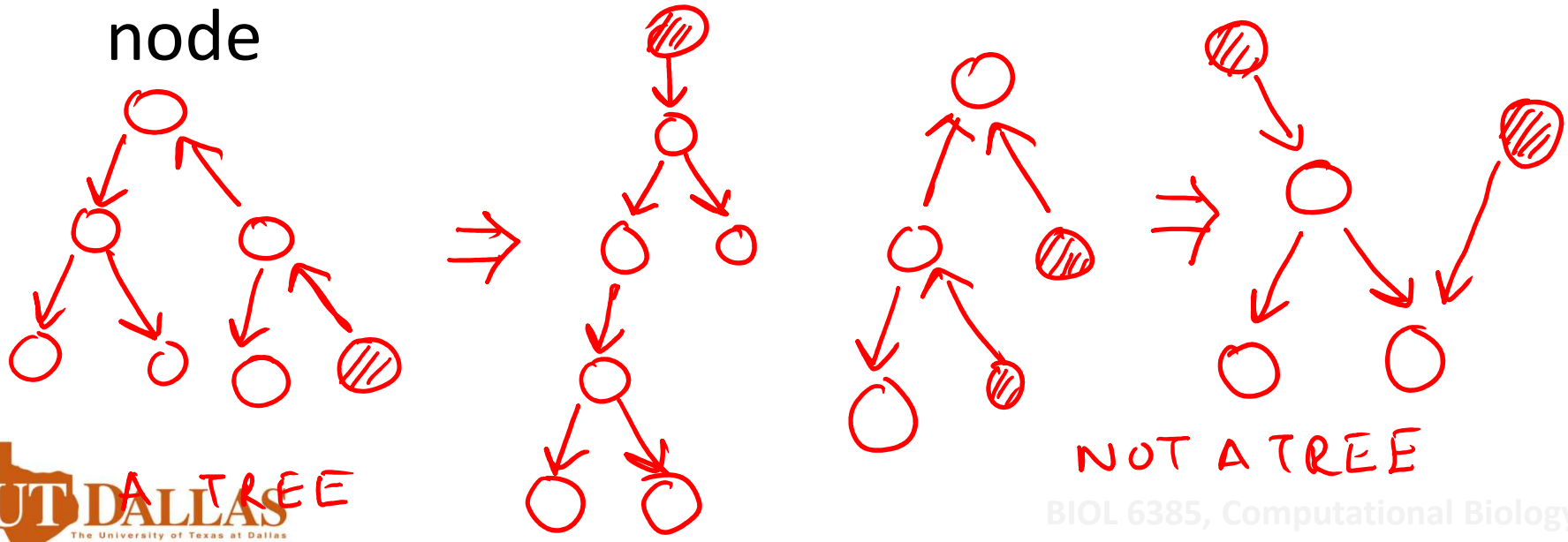


ouwarovite

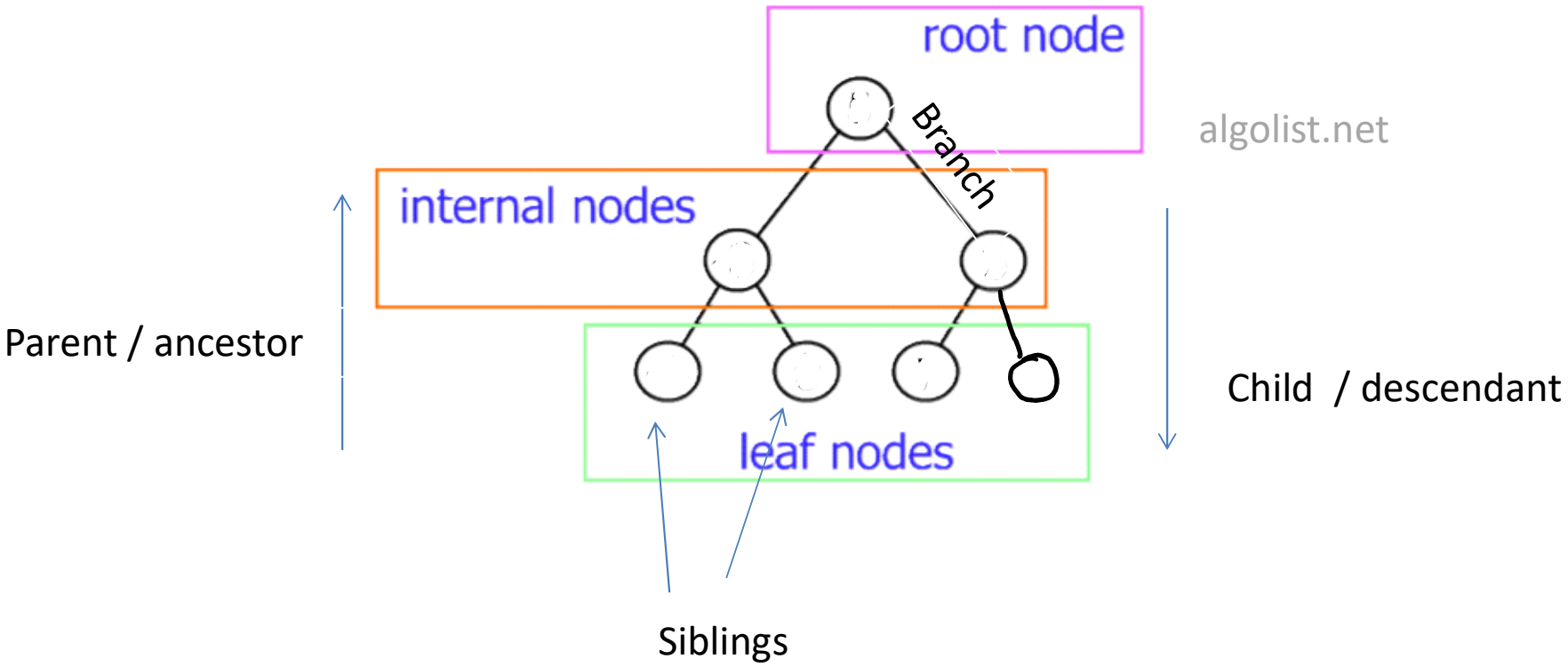


Is this graph a rooted tree?

- Do a topological sorting on it, is there a unique root ?
- Can we order all the nodes in the direction of the edges, and be left with a single topmost node

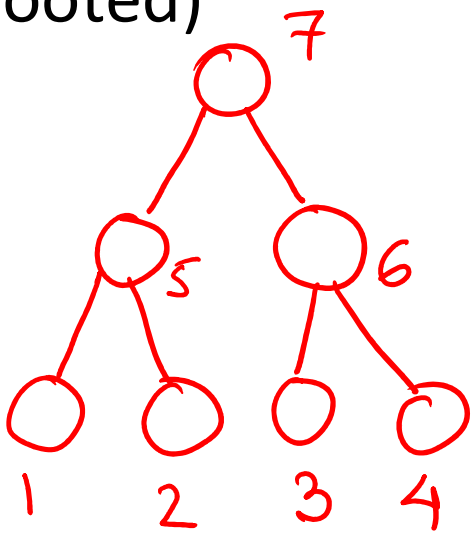


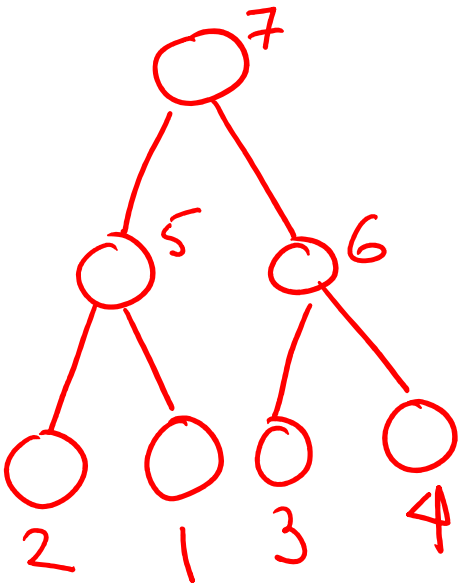
Rooted tree nomenclature



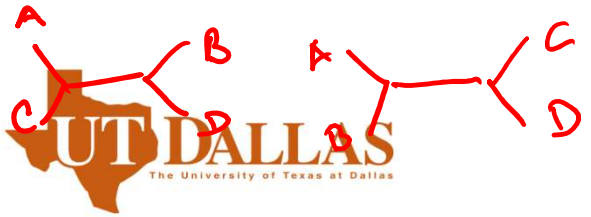
Ordered vs unordered branches

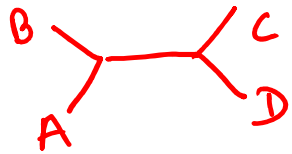
- Is the order in which we represent the siblings important ? 2^n ways to draw for n interior nodes (rooted)

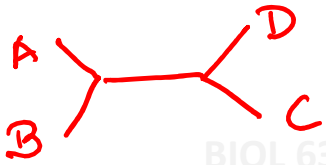


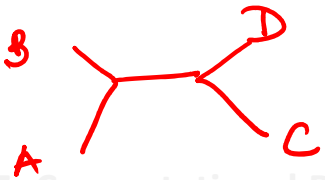


- Which is different ?



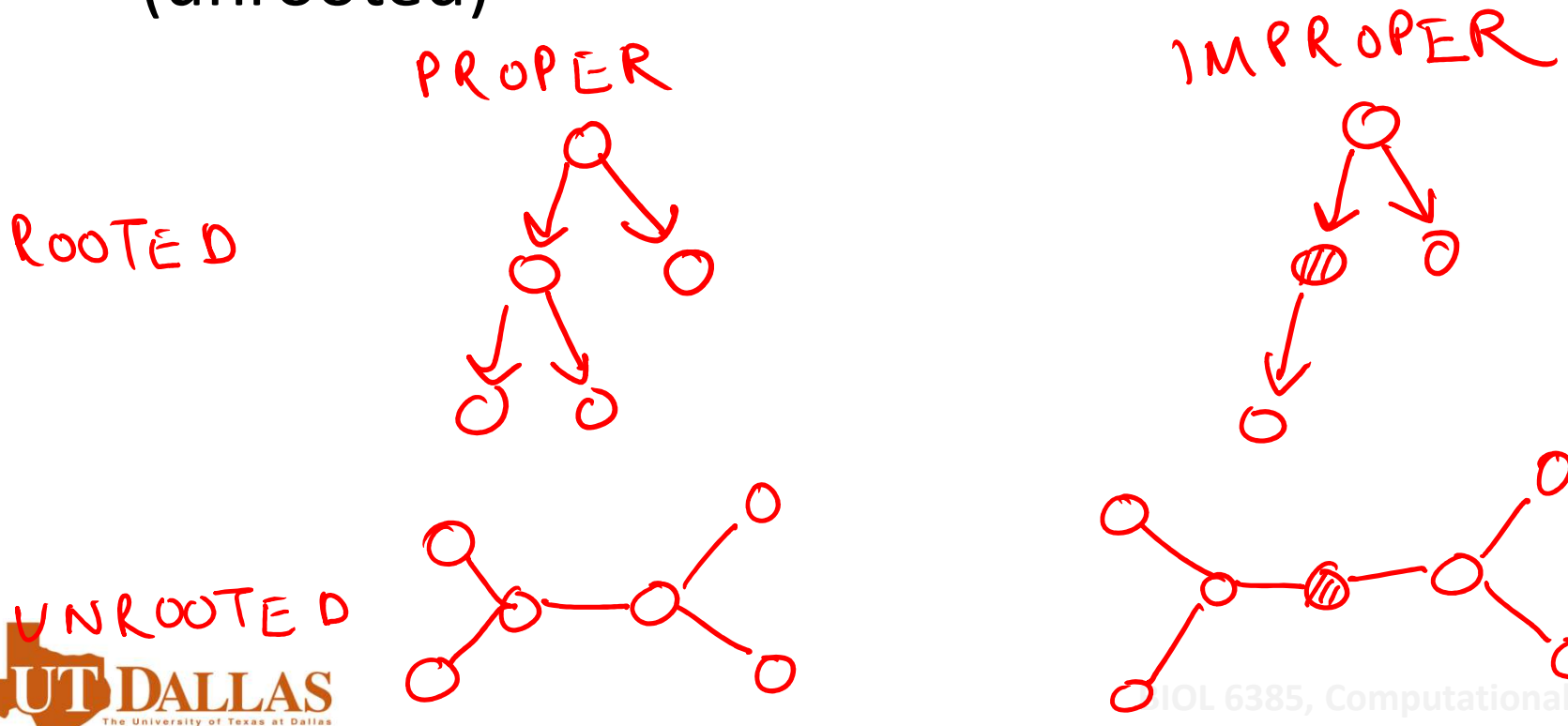






Proper vs improper trees

- Proper : each node has 0 or 2 children (rooted), each node has 1 or 3 neighbors (unrooted)

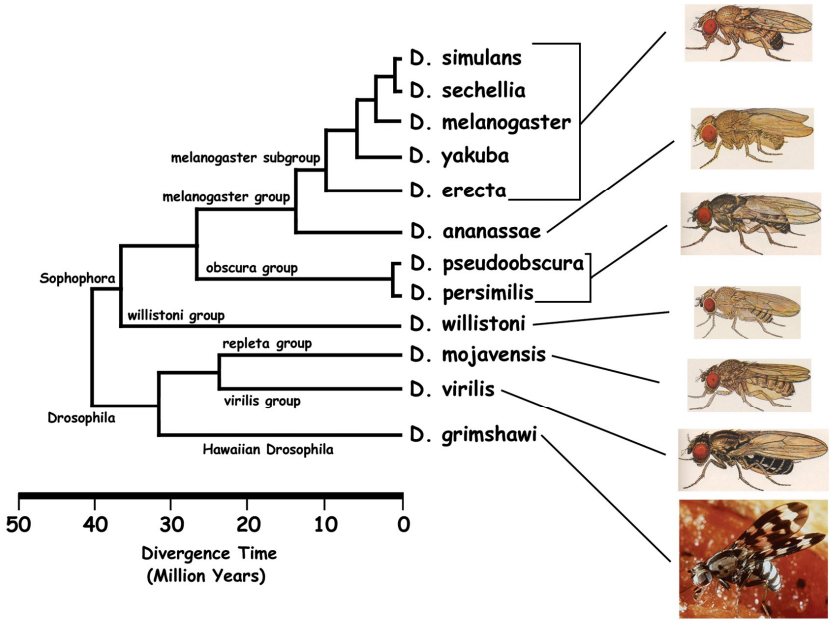


Phylogeny / phylogenetic tree

- Taxon/taxa/operational taxonomic unit (OTU)
  - unit of classification : species, subspecies, individual, etc

- Phylogeny = evolutionary tree
  - Hypothesis concerning evolutionary history of taxa

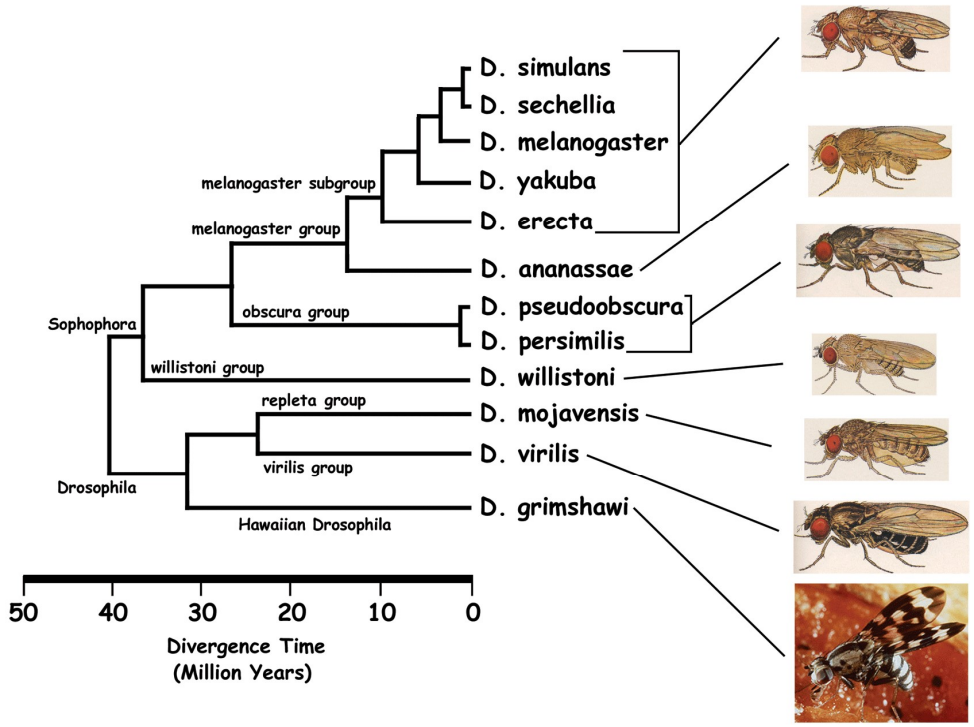
insect.eugenes.org



UT DALLAS
The University of Texas at Dallas

Molecular Phylogeny

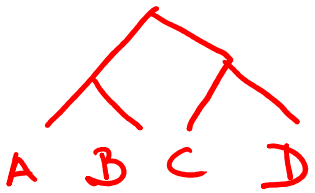
- Molecular phylogeny = based on models (or distances based on models) of molecular evolution



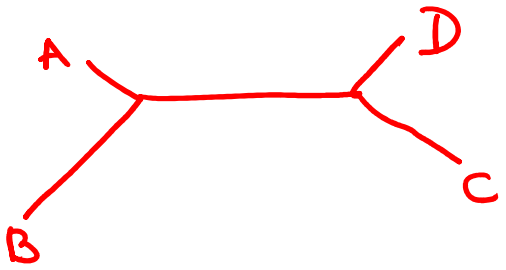
insect.eugenes.org

Phylogenetic Trees

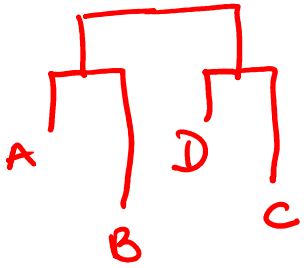
- Edges = branches
- Interior nodes : ancestral taxa
- Leaf / exterior nodes : contemporary taxa
- Topology : relation between species
- Branch length = edge weights: amt of change
- Labelled leaves, unlabelled ancestors



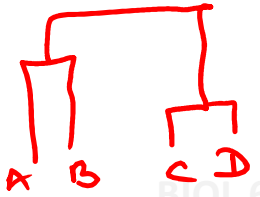
NO BRANCH LENGTH



BRANCH LENGTH, NO ROOT

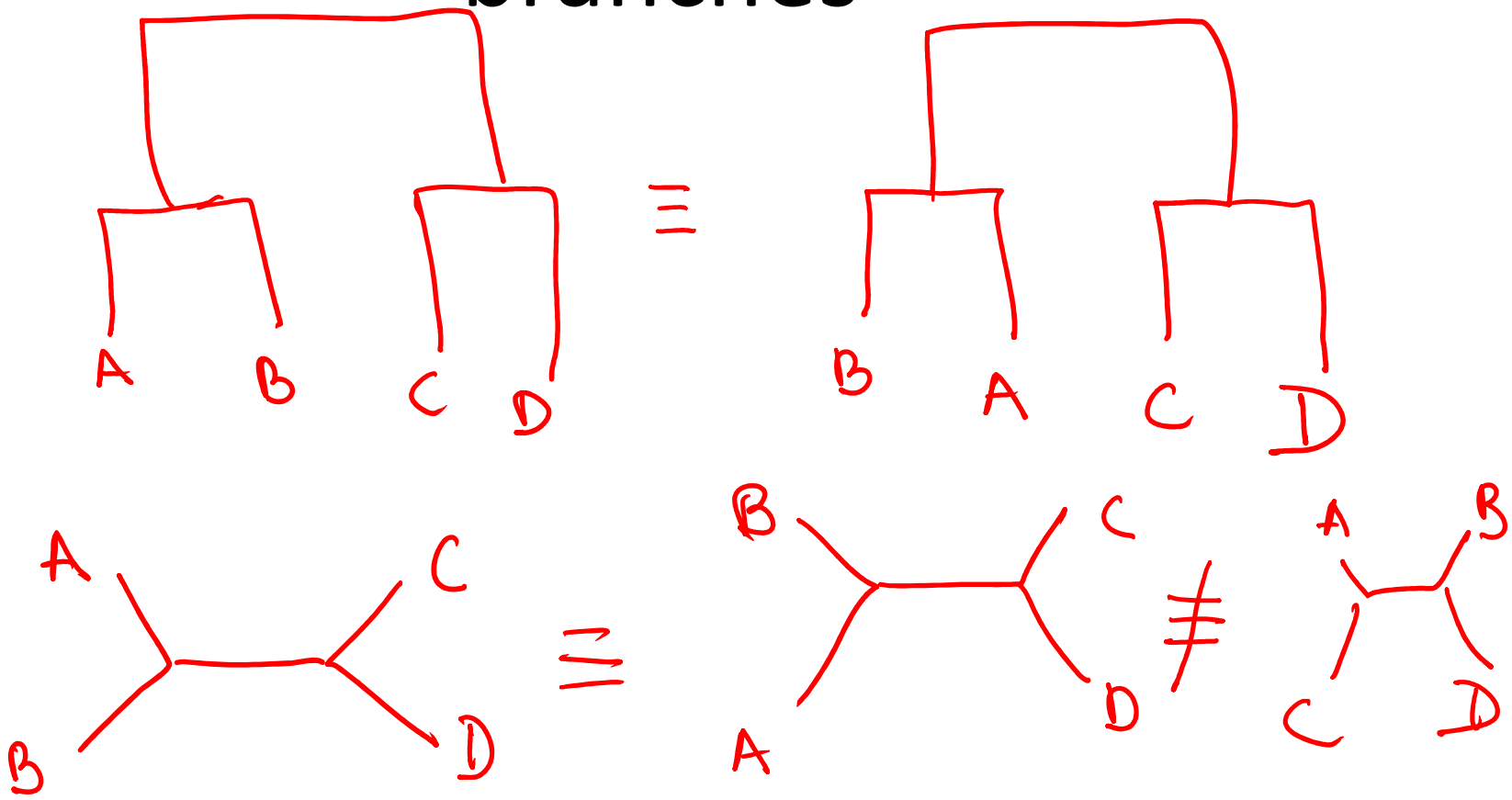


ROOT, NO CLOCK



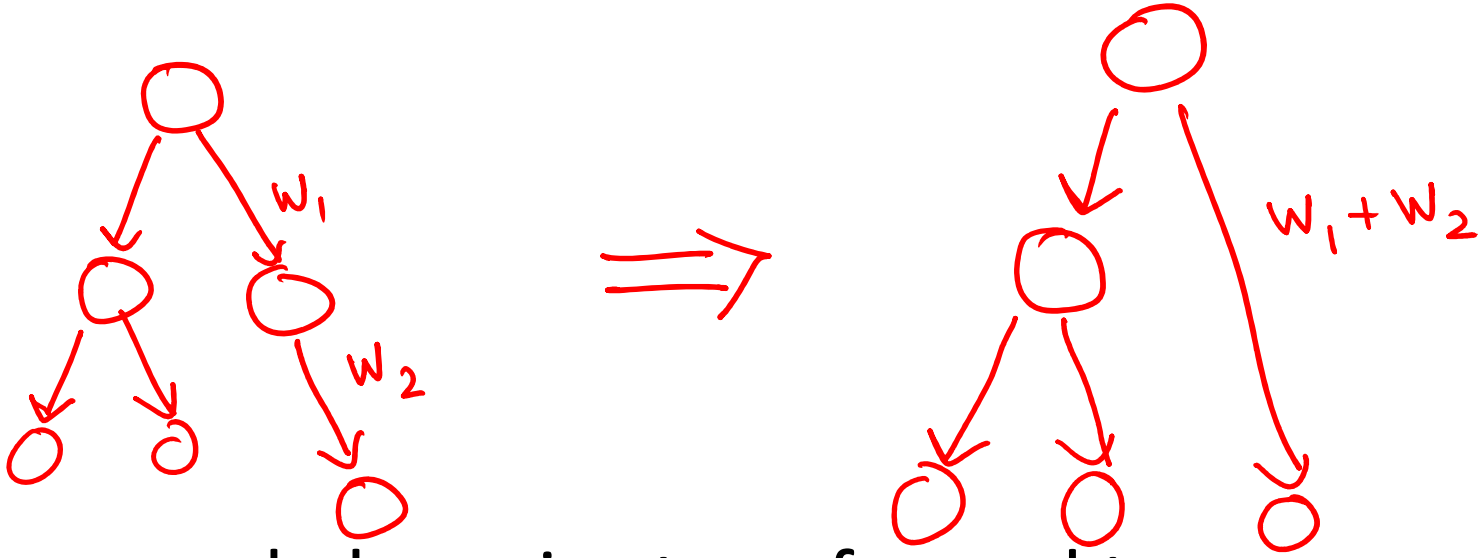
ROOT, WITH CLOCK

Phylogenies have unordered branches



Ordering has no evolutionary connotation

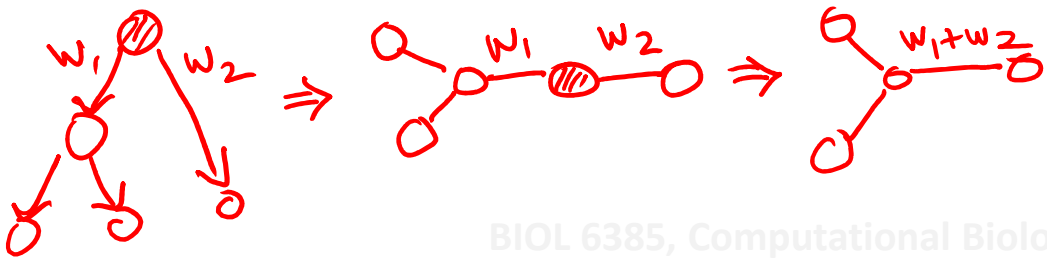
Phylogenetic trees are proper trees



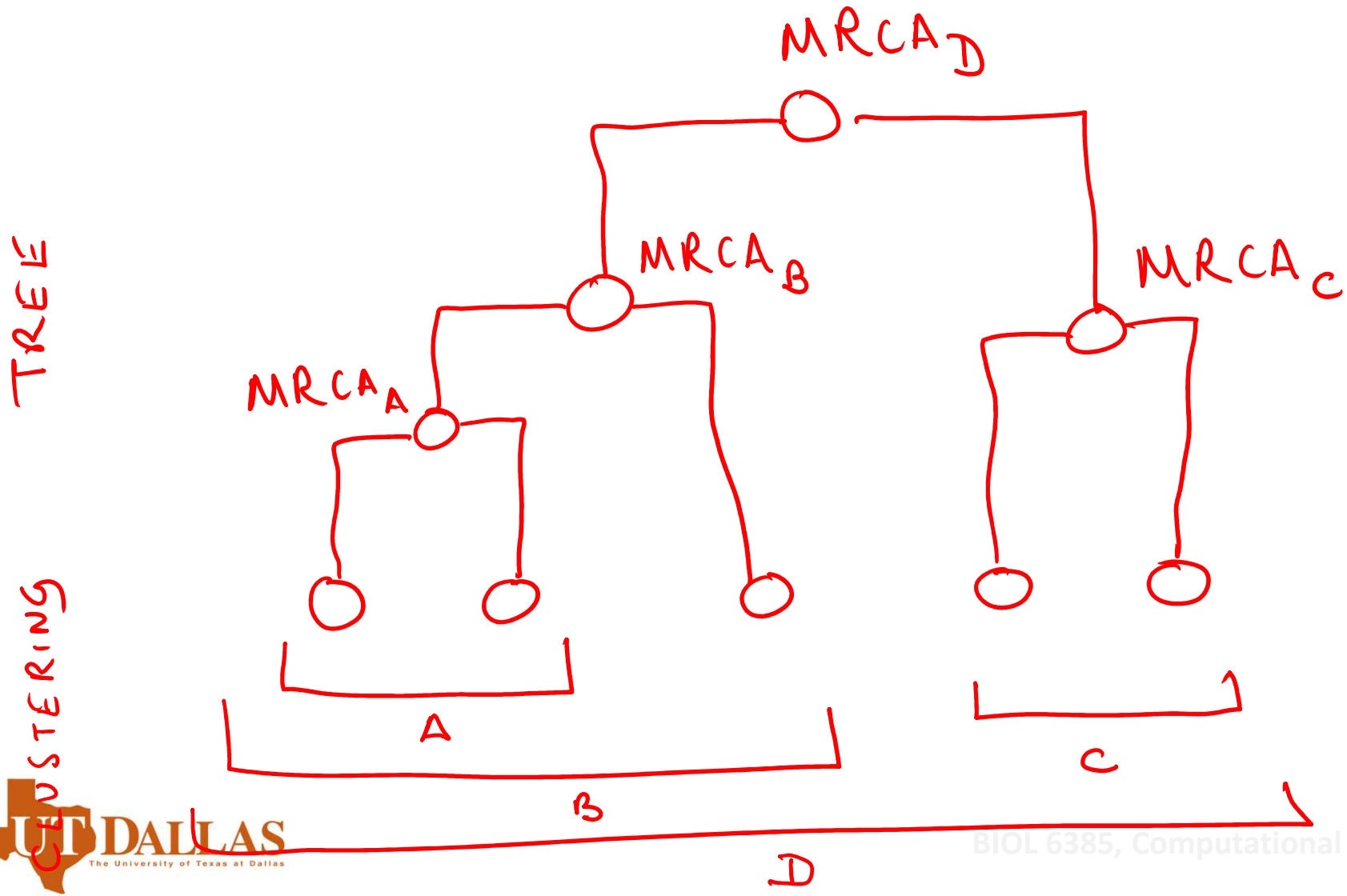
Improper phylogenies transformed to proper ones

- Each node = Common ancestor of subset of species

– Unrooting



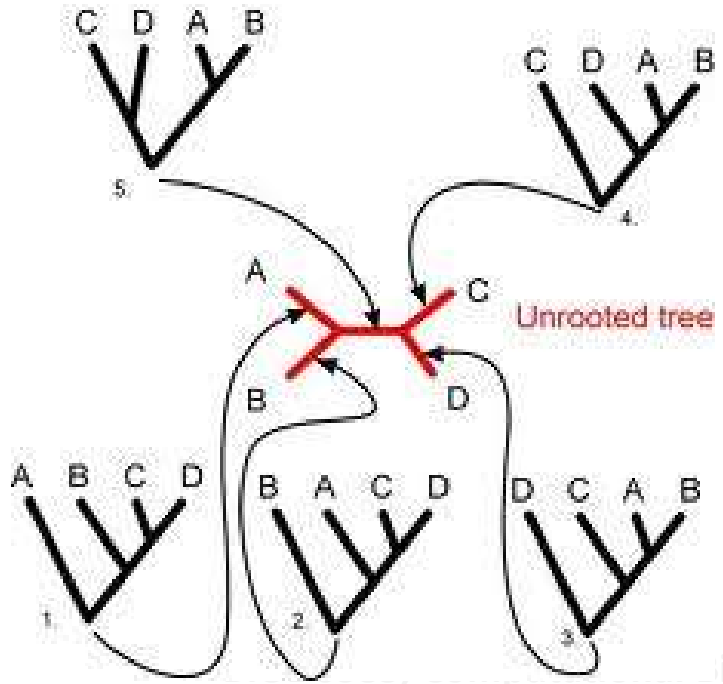
Most recent common ancestor



Rooting an unrooted tree

- Root at any branch

- Sometimes, we may not know (or may not care) where the root is



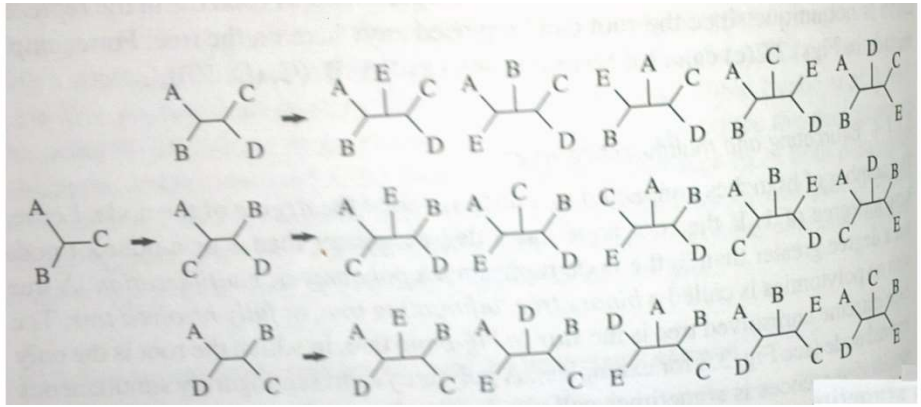
Counting edges

- For an unrooted tree on n leaves, we have
  - n-2 internal nodes
    - (proof by induction, tree on 3 leaves has 1 internal node, every additional leaf incorporated into a tree adds one leaf and one internal node)
  - 2n-3 edges [no of nodes minus one]
    - (think of shrinking the tree one node and edge at a time)

embl.de

Counting labelled leaf, unlabelled ancestor phylogenetic trees

- Topologically equivalent = a tree changed to another by flipping neighbors, w/o breaking branches
- For a tree of n - 1 leaves, 2n – 5 branches : add nth leaf to a branch : T(n) = T(n-1) X (2n – 5)
  - T(3) = 1



Z Yang

- If rooted, rooting can happen at any branch : additional factor of (2n - 5)

No of unrooted trees : T(k)

$$T(k) = \frac{(2k-5)!}{2^{k-3}(k-3)!}$$

Grows fast

- Felsenstein, Counting trees

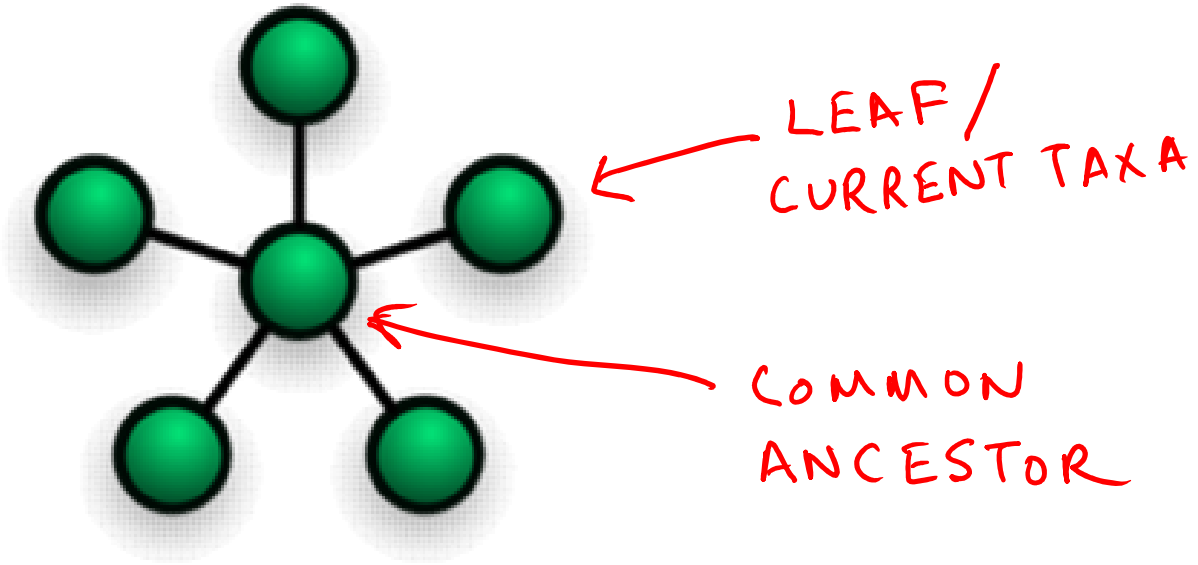
TABLE 1. THE NUMBERS OF ROOTED TREES WITH n LABELLED TIPS AND WITH UNLABELLED INTERIOR NODES. THE LEFT COLUMN COUNTS ALL TREES, THE RIGHT COLUMN ONLY BIFURCATING TREES.

n	All trees	Bifurcating trees
1	1	1
2	1	1
3	4	3
4	26	15
5	236	105
6	2,752	945
7	39,208	10,395
8	660,032	135,135
9	12,818,912	2,027,025
10	282,137,824	34,459,425
11	6,939,897,856	654,729,075
12	188,666,182,784	13,749,310,575
13	5,617,349,020,544	316,234,143,225
14	181,790,703,209,728	7,905,853,580,625
15	6,353,726,042,486,112	213,458,046,676,875
16	238,513,970,965,250,048	6,190,283,353,629,375
17	9,571,020,586,418,569,216	191,898,783,962,510,625
18	408,837,905,660,430,516,224	6,332,659,870,762,850,625
19	18,522,305,410,364,568,764,416	221,643,095,476,699,771,875
20	887,094,711,304,094,583,095,296	8,200,794,532,637,891,559,375
21	44,782,218,857,751,551,087,214,592	319,830,986,772,877,770,815,625
22	2,376,613,641,928,796,906,249,519,104	13,113,070,457,687,988,603,440,625

- No of trees on 500 taxa ~ 1 X 10^1074
- No of atoms in observable universe ~ 10^80

Minimum evolutionary hypothesis

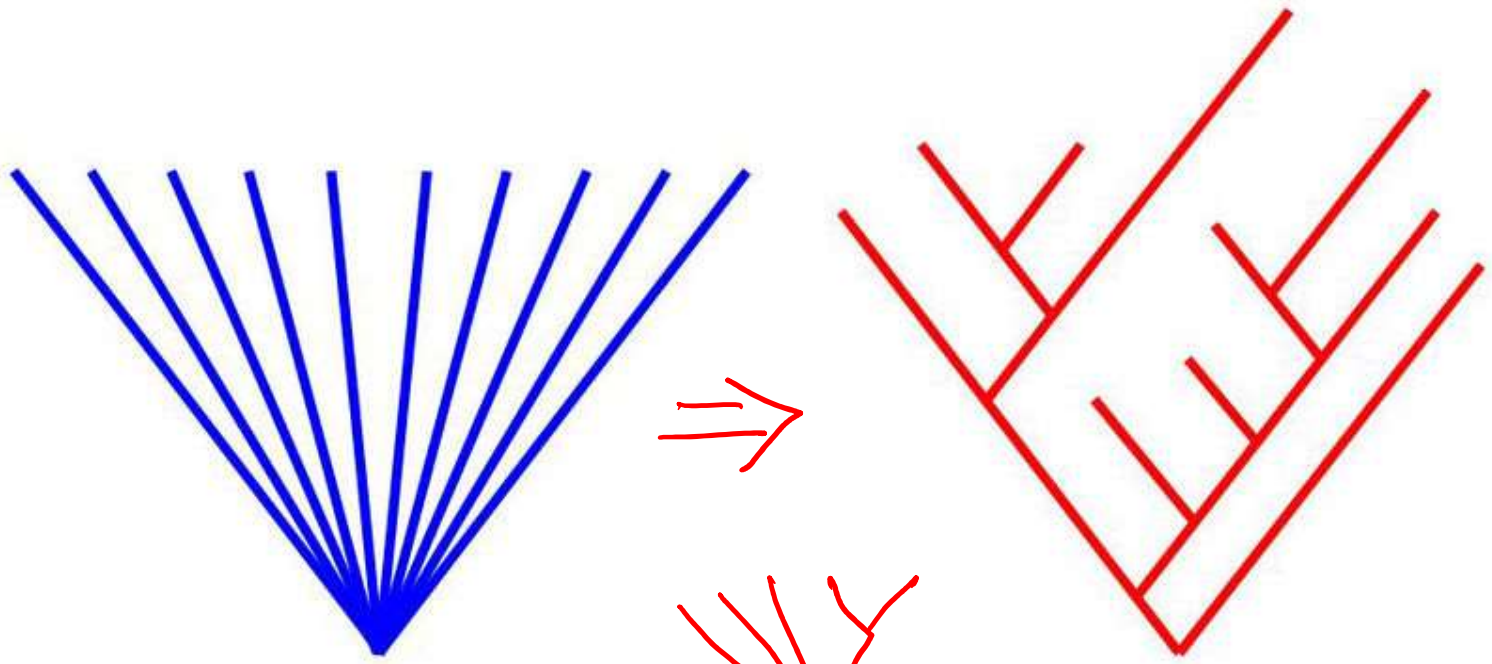
- In terms of topology : the star topology (not a binary tree)
  - only assumes a most recent common ancestor



wikipedia

Resolution

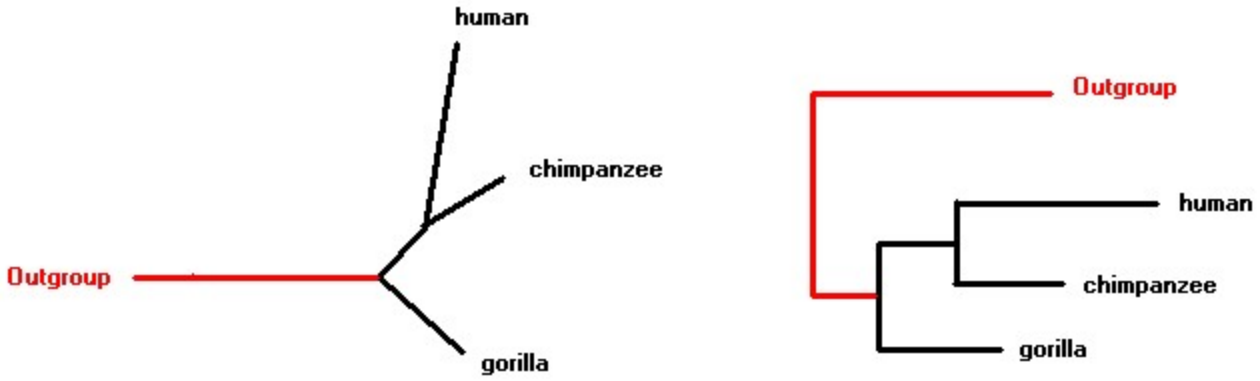
- Resolution = process of figuring out topology = generating more complicated evolutionary hypotheses
- Partial resolution = intermediate stage



wikipedia

Outgroups

- When studying a group, we may want a control which is outside that group



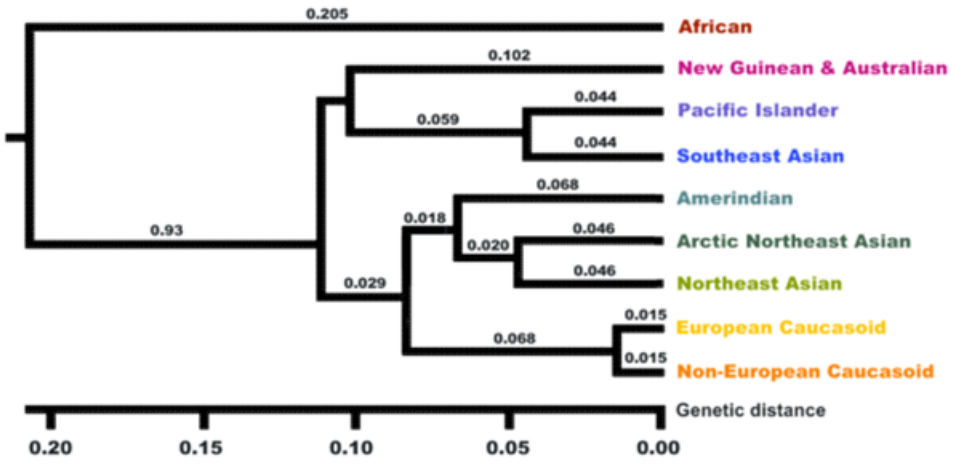
bioinf. manchester.ac,uk

- One way to root a tree

2 schools of phylogeny reconstruction

- Distance based methods

Linkage tree for 9 population clusters showing genetic distances (F_{ST})
(Cavalli-Sforza *et al.*, 1994:80)



F_{ST} distance matrix for the 9 clusters shown above
(x10,000 with standard errors obtained by bootstrap analysis)

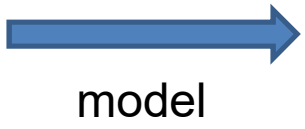
	AFR	NEC	EUC	NEA	ANE	AME	SEA	PAI	NGA
African	0.0								
Non-European Caucasian	1340.0 ± 301	0.0							
European Caucasian	1655.6 ± 416	154.7 ± 29	0.0						
Northeast Asian	1979.1 ± 452	640.4 ± 134	938.2 ± 217	0.0					
Arctic North-east Asian	2008.5 ± 387	708.2 ± 160	746.7 ± 210	459.7 ± 98	0.0				
Amerindian	2261.4 ± 434	955.5 ± 204	1038.2 ± 276	746.5 ± 183	577.4 ± 89	0.0			
Southeast Asian	2206.3 ± 529	939.6 ± 262	1240.4 ± 339	630.5 ± 299	1039.4 ± 326	1341.7 ± 418	0.0		
Pacific Islander	2505.4 ± 648	953.7 ± 230	1344.7 ± 354	723.8 ± 262	1181.2 ± 331	1740.7 ± 544	436.7 ± 87	0.0	
New Guinean and Australian	2472.0 ± 536	1179.1 ± 189	1345.7 ± 231	734.4 ± 118	1012.5 ± 257	1457.9 ± 283	1237.9 ± 277	808.7 ± 264	0.0

metric

2 schools of phylogeny reconstruction

- Character based methods

A B C D E
A X C D E
A X Y D E



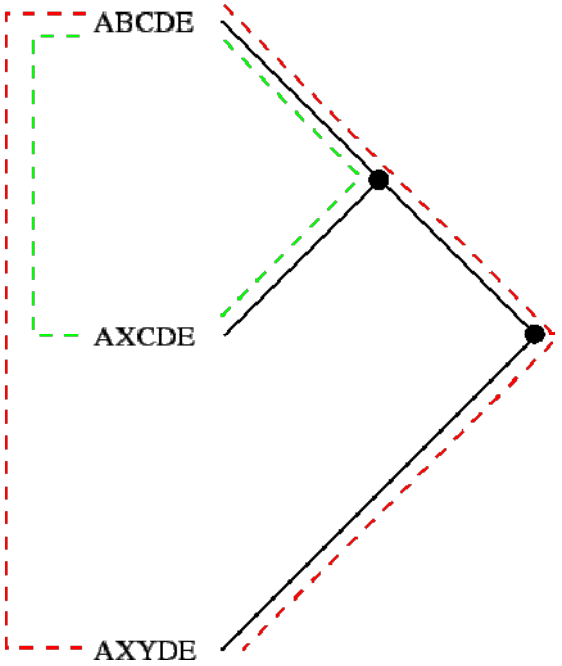


Figure: A Phylogenetic Tree

molgen.mpg.de

UT DALLAS
The University of Texas at Dallas

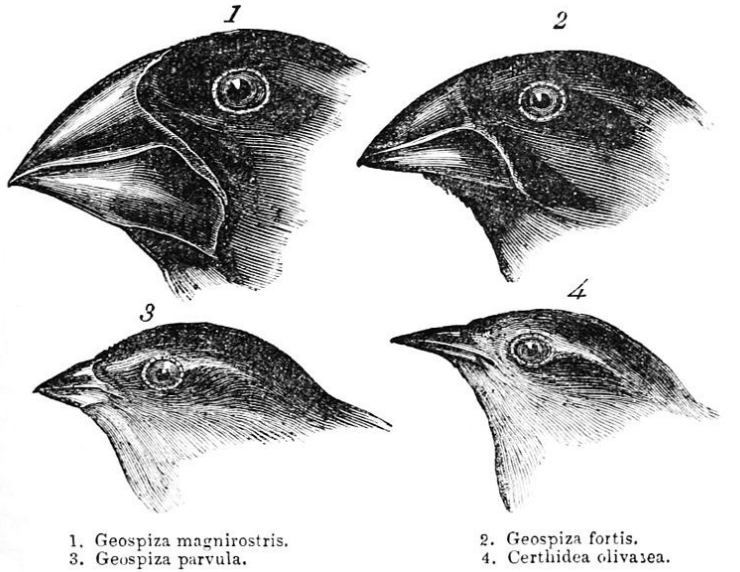
What kind of data can we use ?

- Both genetic and phenotype based models of evolution can be either character based or distance based.

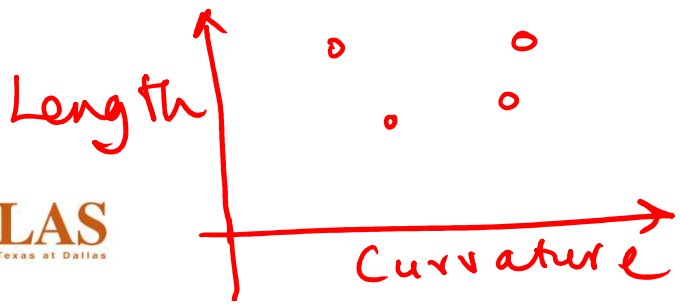
Distance based methods

- How to infer evolutionary relationships on the basis of some similarity measure ?

- Notion of a similarity measure
  - eg. Curvature and length
  - 2D Euclidean distance



1. Geospiza magnirostris.
3. Geospiza parvula.
2. Geospiza fortis.
4. Certhidea olivacea.



Genetic distance

- Similar to alignment score
  - how far away are two orthologous sequences in sequence space ?
  - alignment usually required (compare apple to apple)
  - one option : pairwise entropy measure
  - similar to DOBS = fraction of sites with substitutions in pairwise alignment

Model based genetic distances

- Typically such simple measures cant capture complicated evolutionary models (which model selection)
- However, in JC 69, all mutations are equally likely : can we use D_{OBS} as distance measure ?

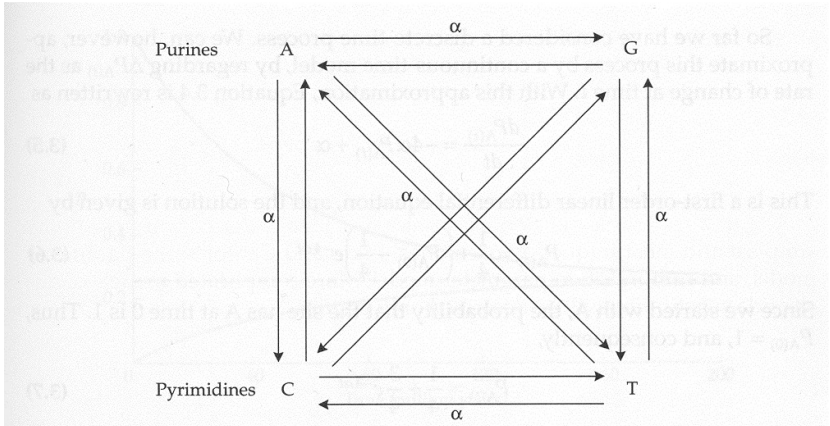
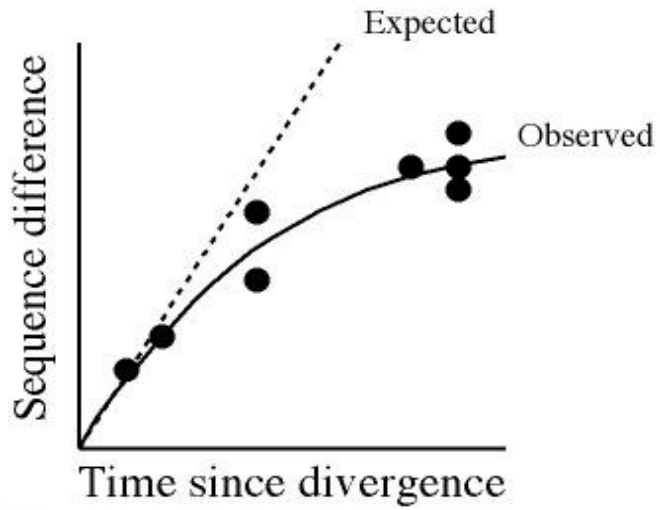


FIGURE 3.1 One-parameter model of nucleotide substitution. The rate of substitution in each direction is α.

Durbin

Correction for multiple substitutions

- A → T → C (multiple mutations when we see one substitution)
- A → T → A (multiple mutations when we see no substitution)



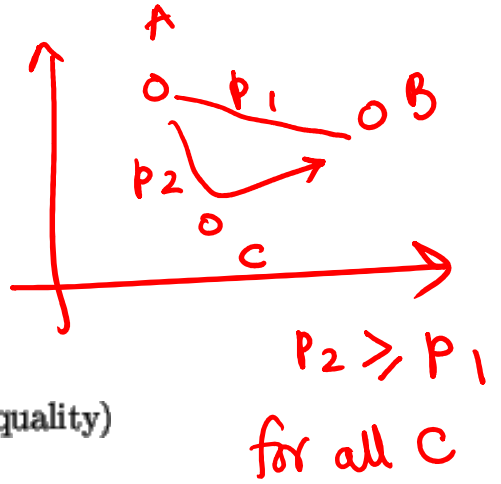
CORRECT FOR:
P(2 or more mutation at each site)
= infinite series [HW]

$$D_{JC} = -\frac{3}{4}\ln\left[1 - \frac{4}{3}D_{OBS}\right]$$

Metric

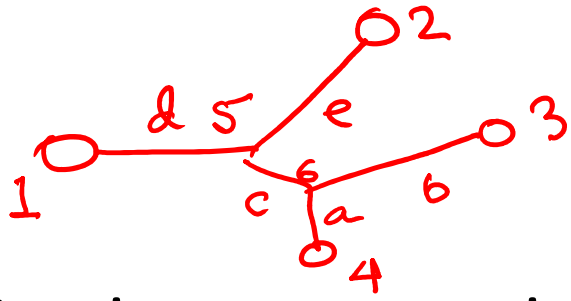
- Mathematical notion of “distance”
- Intuitive properties

$$
\begin{aligned}
d(x,y) &> 0 \qquad \text{for } x \neq y \\
d(x,y) &= 0 \qquad \text{for } x = y \\
d(x,y) &= d(y,x) \qquad \forall\, x,y \\
d(x,y) &\leq d(x,z) + d(y,z) \qquad \forall\, x,y,z \quad \text{(triangle inequality)}
\end{aligned}
$$

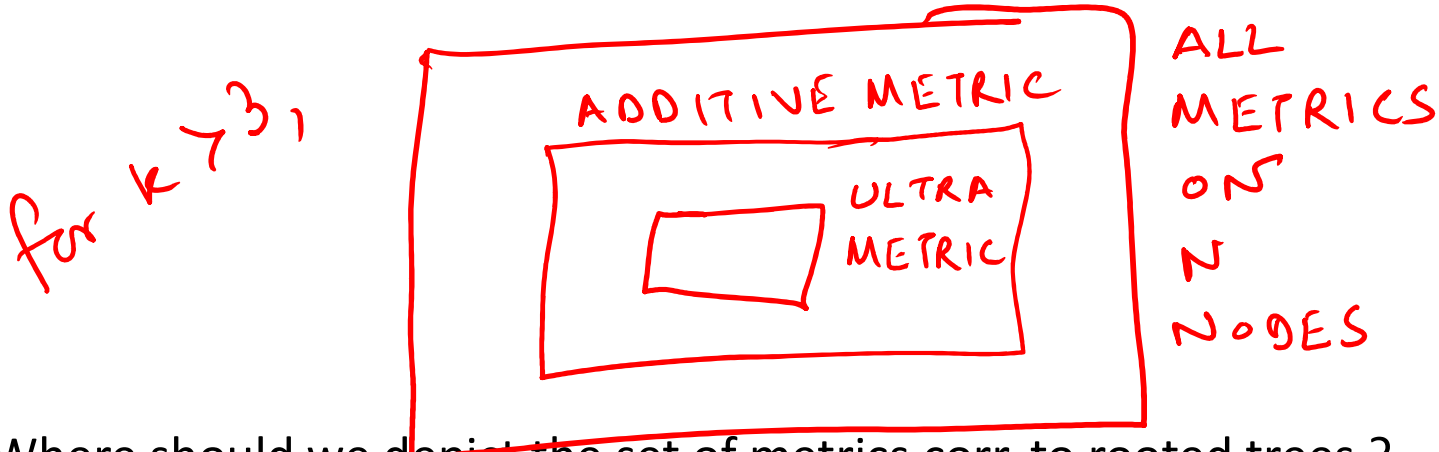


- Space in which x, y, z lives + metric definition
  - Metric space

Additive metrics & ultrametrics

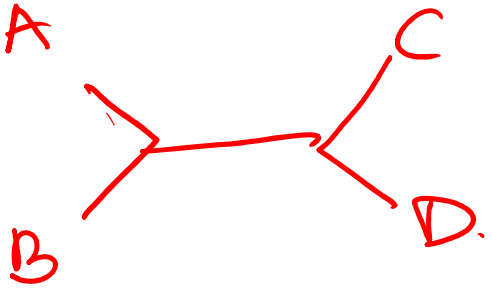


- Additive / tree metric
  - remember, route between 2 points on a tree is unique
- Ultrametric : $d(x,y) \leq max\,(d(x,z),\ d(y,z))$



Where should we depict the set of metrics corr. to rooted trees ?

Tree metrics : 4 point condition



- Consider every quartet of leaves in the tree

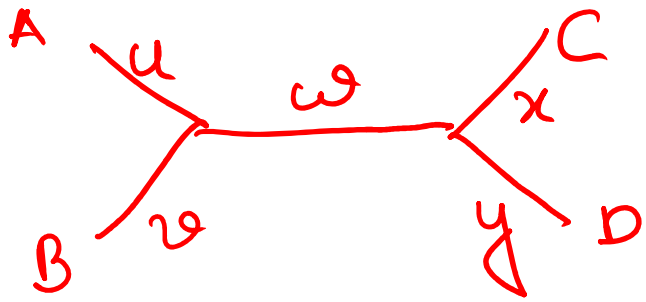
- 3 ways to split into two subsets of 2 nodes :
  - AB & CD, AC & BD, AD & BC

- $d(AB) + d(CD) <= \max(d(AC) + d(BD), d(AD) + d(BC))$
- $d(AC) + d(BD) <= \max(d(AB) + d(CD), d(AD) + d(BC))$
- $d(AD) + d(BC) <= \max(d(AC) + d(BD), d(AB) + d(CD))$

4PC iff additivity

$\binom{n}{2}$ equations

$2n-3$ variables



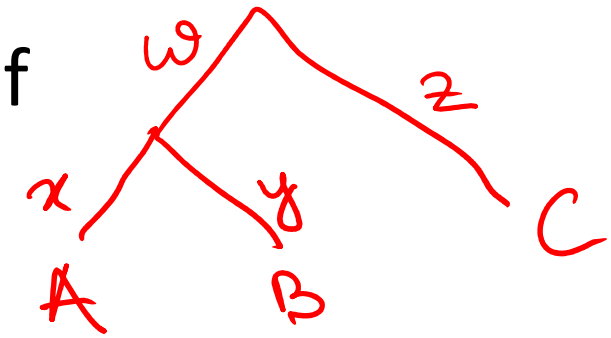
$$u+v = d(AB)$$
$$u+w+x = d(AC)$$
$$u+w+y = d(AD)$$
$$v+w+x = d(BC)$$
$$v+w+y = d(BD)$$
$$x+y = d(CD)$$

have $\Rightarrow$ unique soln.

Ultrametric = 3 point condition

$$d(x, y) \leq max\ (d(x, z),\ d(y, z))$$

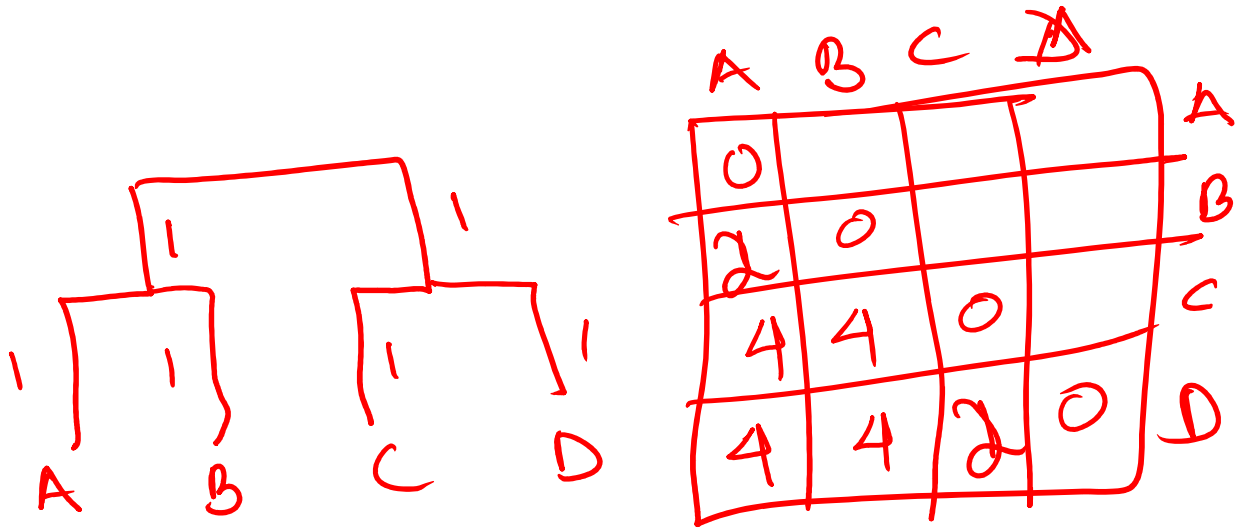
- Consider every triplet of leaves A, B, C



- d(AB) < = max (d(AC), d(BC))
- d(BC) < = max (d(AB), d(AC))
- d(AC) < = max (d(AB), d(BC))

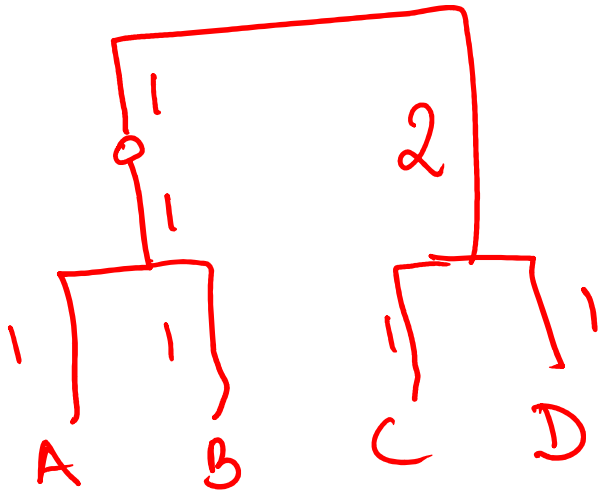
$x + y$
$< (x + w + z$
$= y + w + z)$

3 PC satisfied (hence 4 PC satisfied)



Claded tree

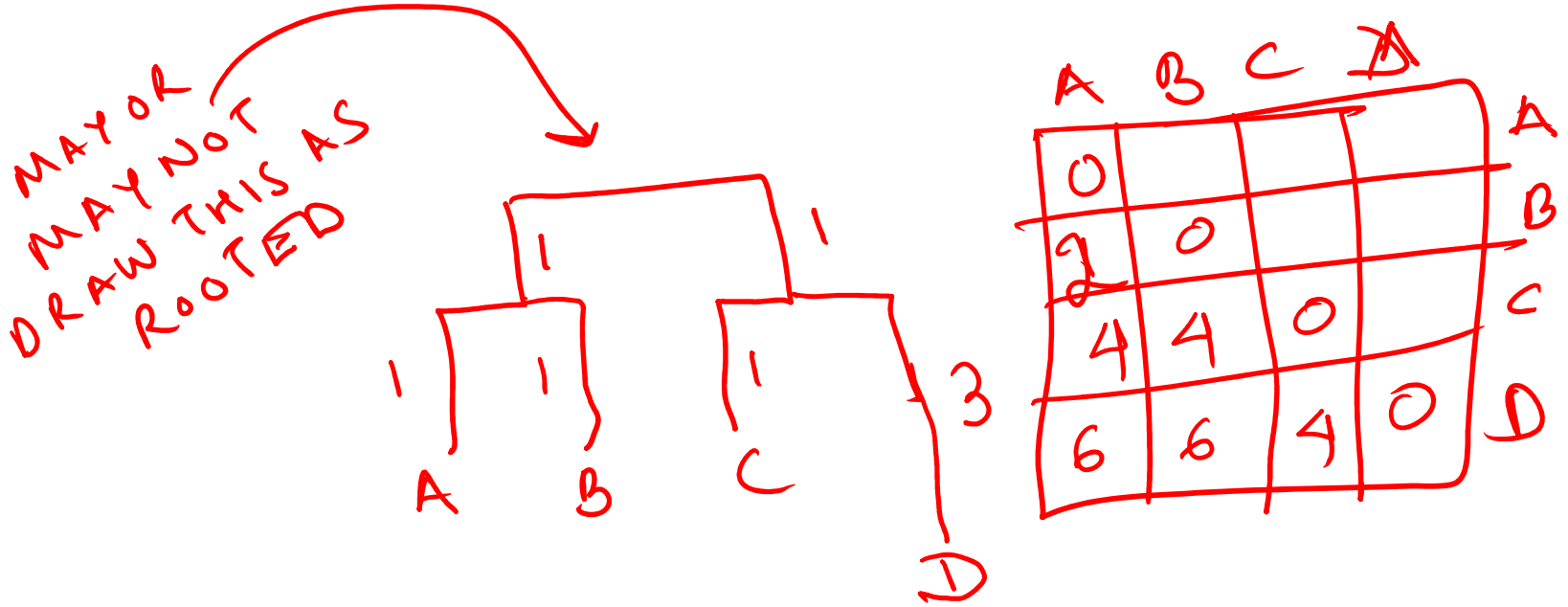
A few changes to 3 PC matrix may still be 3 PC matrix !



	A	B	C	D
A	0			
B	2	0		
C	6	6	0	
D	6	6	2	0

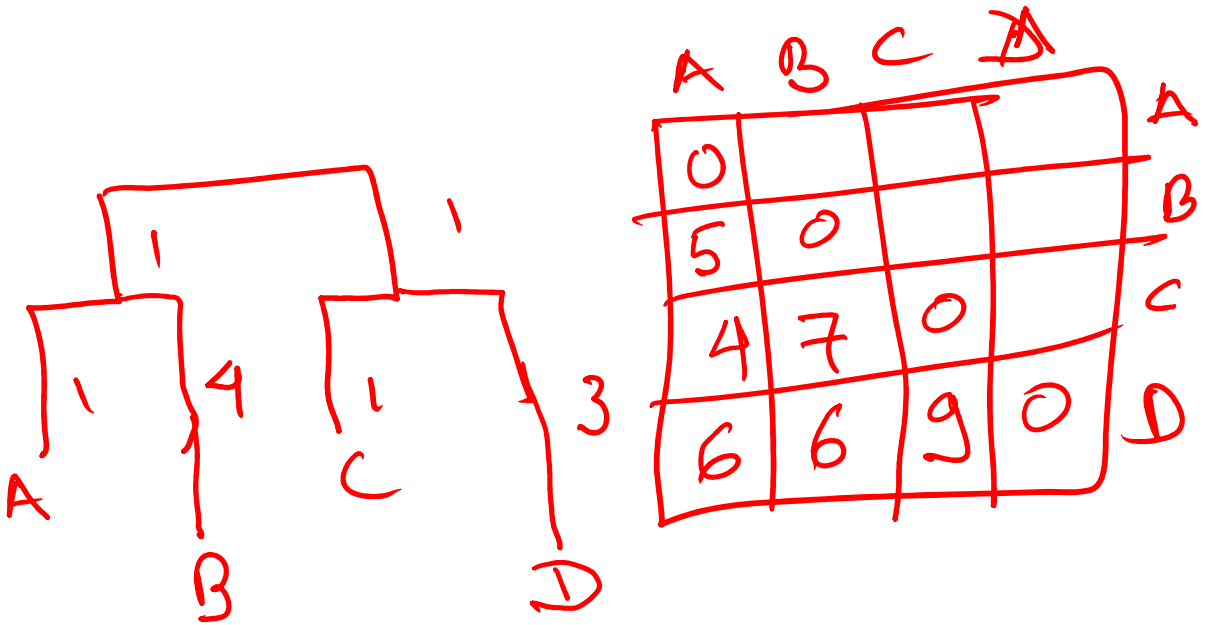
Some changes may still keep the clock

Some changes to 3 PC matrix may only satisfy 4 PC, not 3 PC



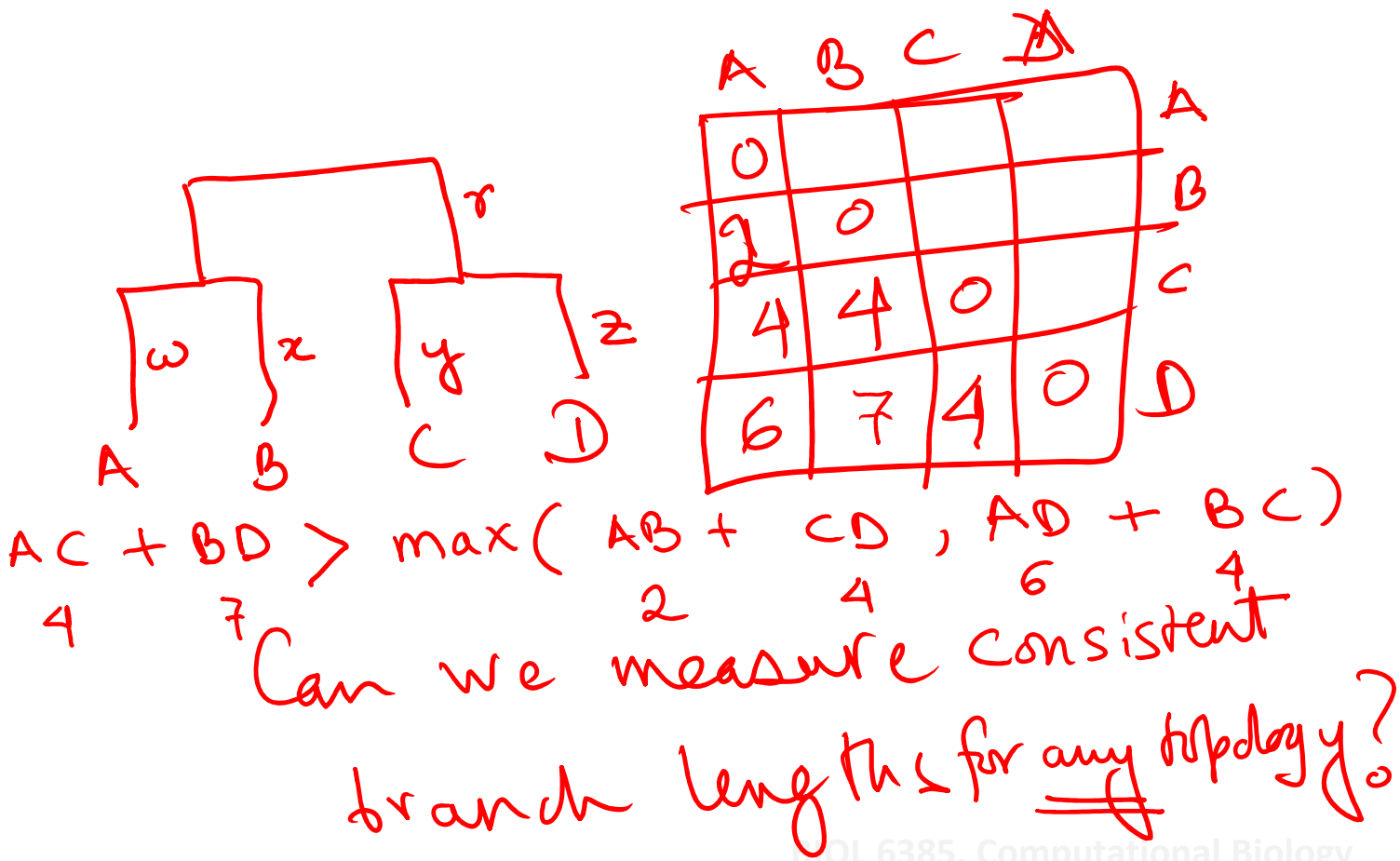
Is it docked?
Subtrees may still be docked

Some changes to 4 PC matrix can preserve 4 PC condition

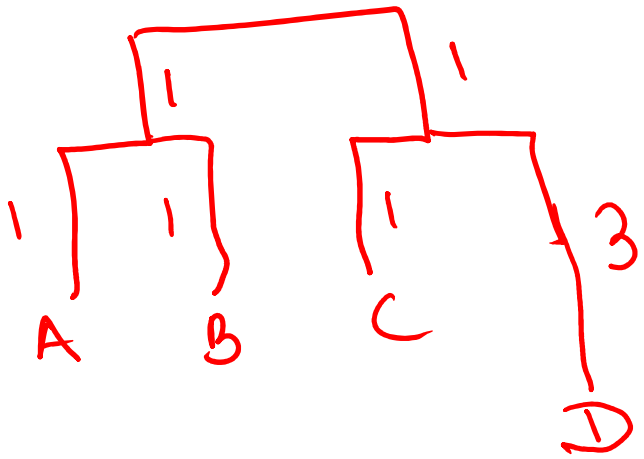


Some changes can preserve additivity

Some changes to 4 PC : only triangle inequality satisfied

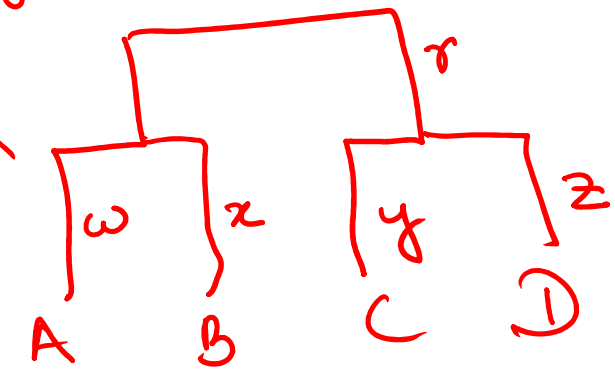


So, what changed ?



	A	B	C	D
A	0			
B	2	0		
C	4	4	0	
D	6	6	4	0

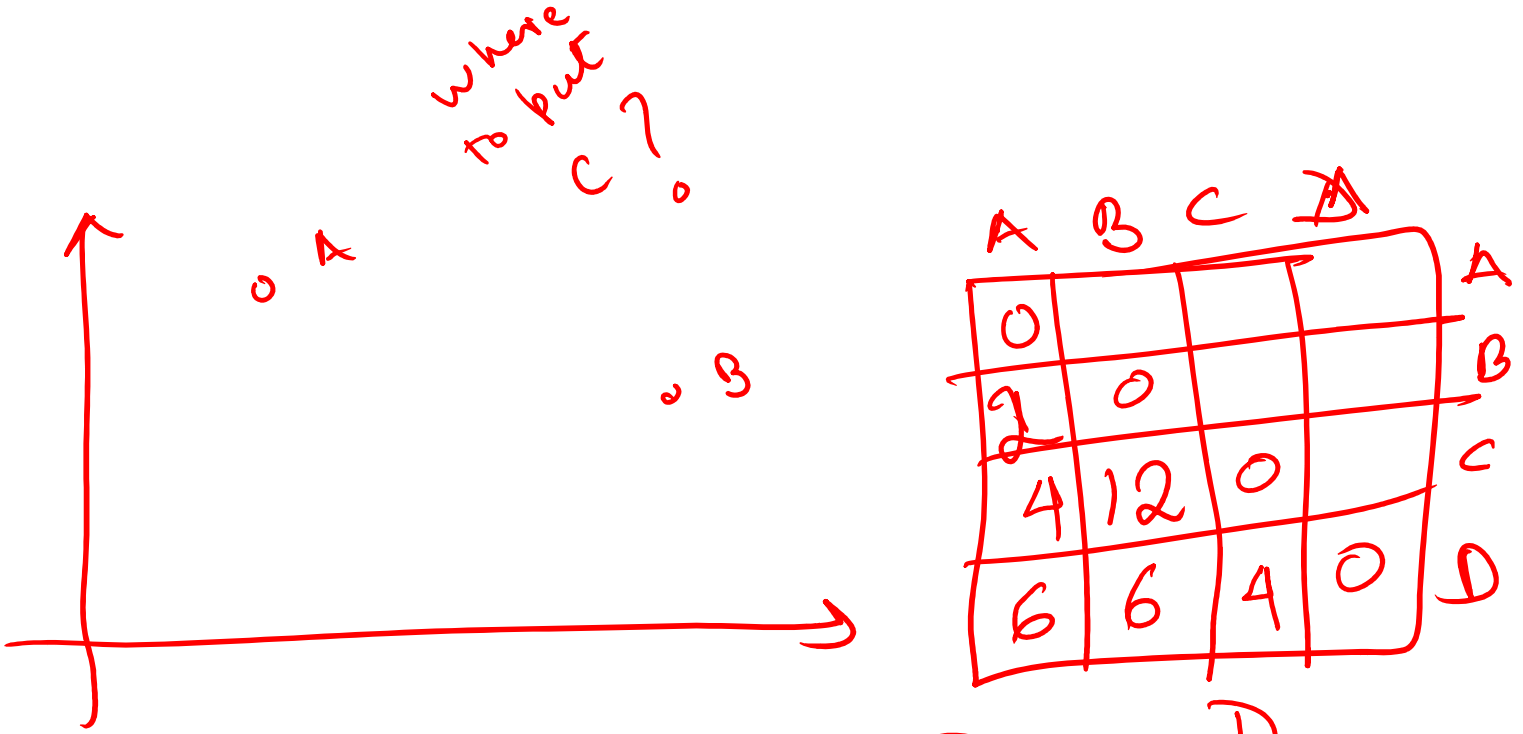
WE NEED TO INCREASE γ, x OR z BUT OTHER ENTRIES WILL BECOME INCONSISTENT



	A	B	C	D
A	0			
B	2	0		
C	4	4	0	
D	6	7	4	0

BUT, WE CAN PLOT THEM IN A METRIC SPACE!

Arbitrary changes to the distance matrix



$$D_{BC} > D_{BA} + D_{AC}$$

We can no longer plot the points in any metric space

Comparison

Distance metrics

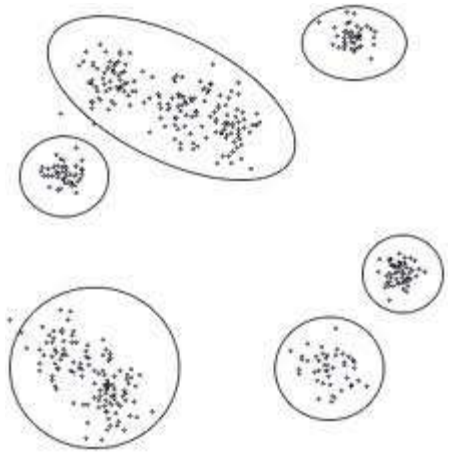
Stricter condition, set of matrices satisying this shrinks

Triangle inequality / distance metric	**4 point condition / additive metric**	**3 point condition / ultrametric**
Topology recovered, branch lengths may be inconsistent	Fits unique unrooted (or one of its many equivalent rooted) tree with unique branch lengths	Fits unique rooted, clocked tree with branch lengths
Points may still be plotted in metric space	Tree may be drawn	Tree drawn, one point on tree (root) is equidistant from all leaves

Polynomial time greedy algorithm to reconstruct tree for all

Clustering

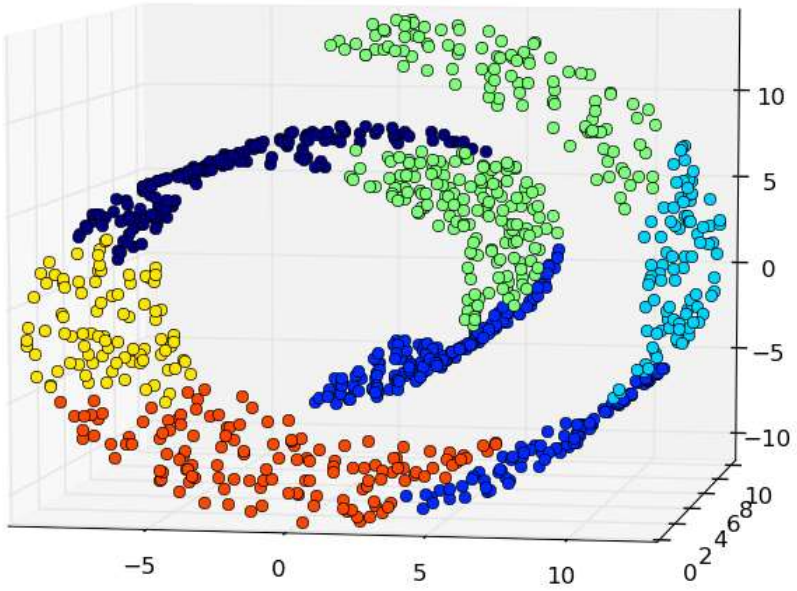
- Partitioning some data points (species or individuals in our case) based on some metric or distance measure



Synaptic, Peltarion

Clustering is hard

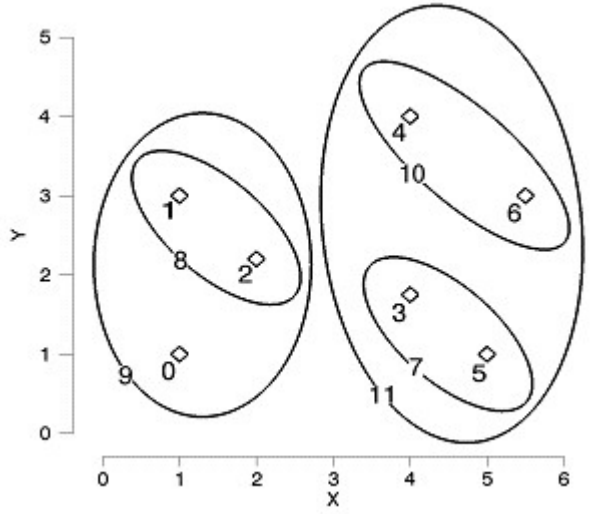
- k-means is NP hard



scikit-learn.org

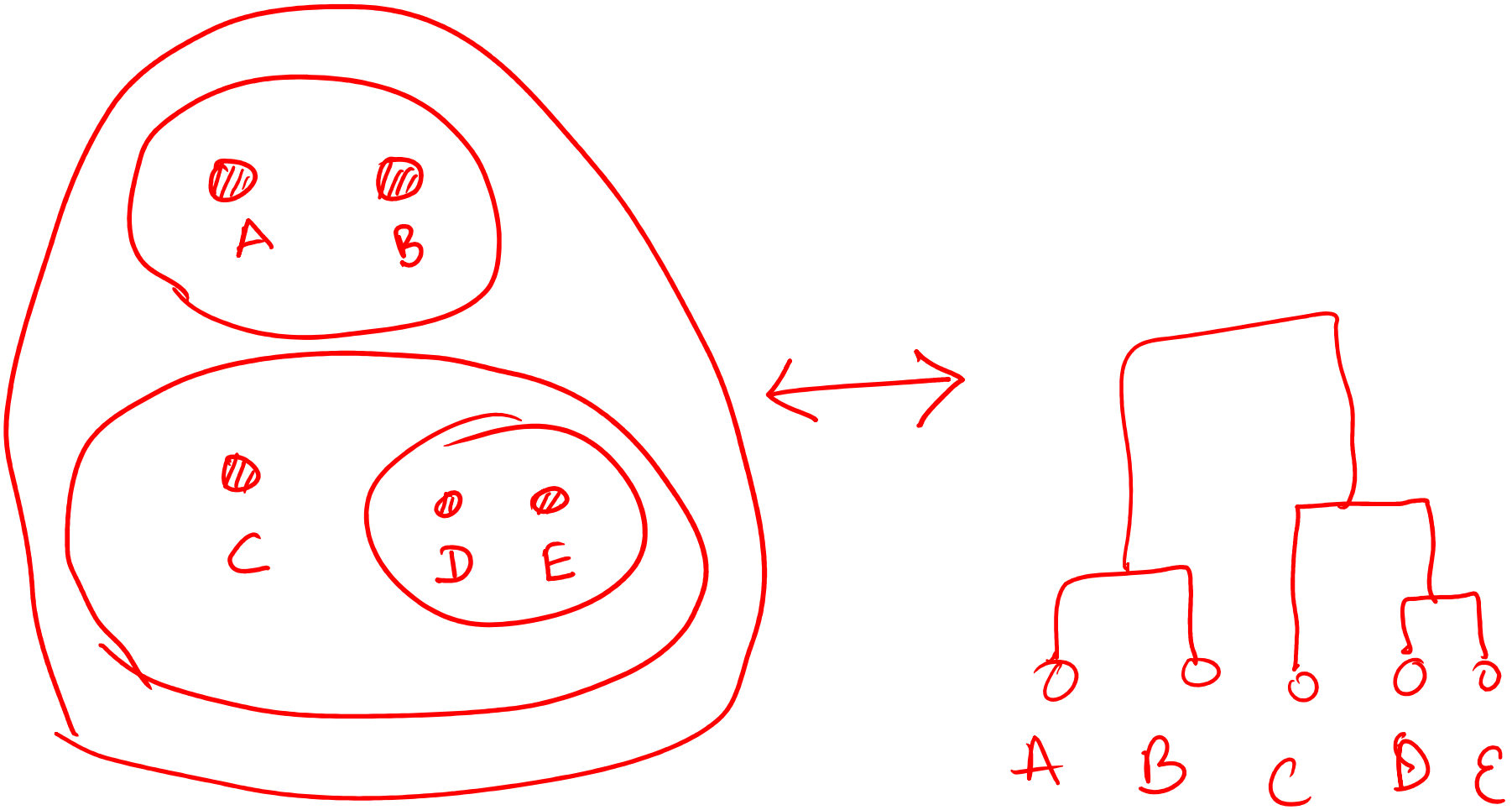
Hierarchical clustering

- We may want to find such clusters at different levels : all the way from the whole data set to individual data points
- Notion of distance between clusters, between data points, and inbetween both categories

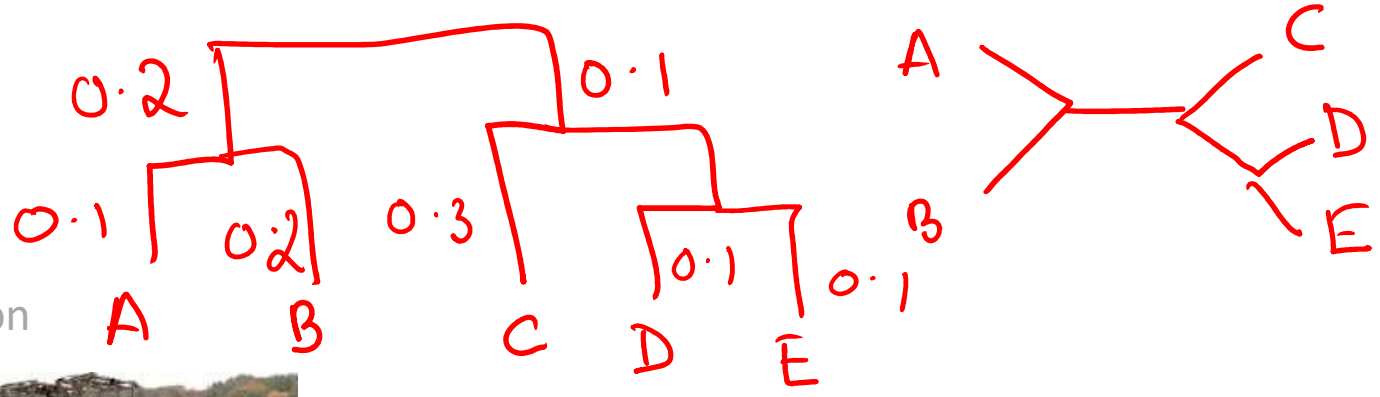


nasa.gov

Clustering ←→ Tree topology



Tree notation : the Newick format



Susan Hillson

ARTHUR CAYLEY
1857

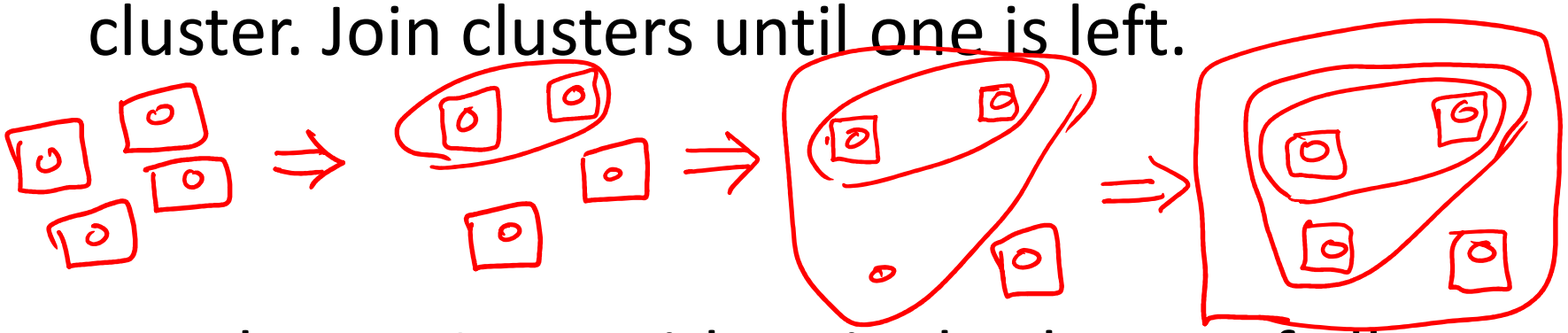
ROOTED: ((A:0.1, B:0.2):0.2, (C:0.3, (D:0.1, E:0.1)):0.1)

UNROOTED: ((A, B), C, (D, E))

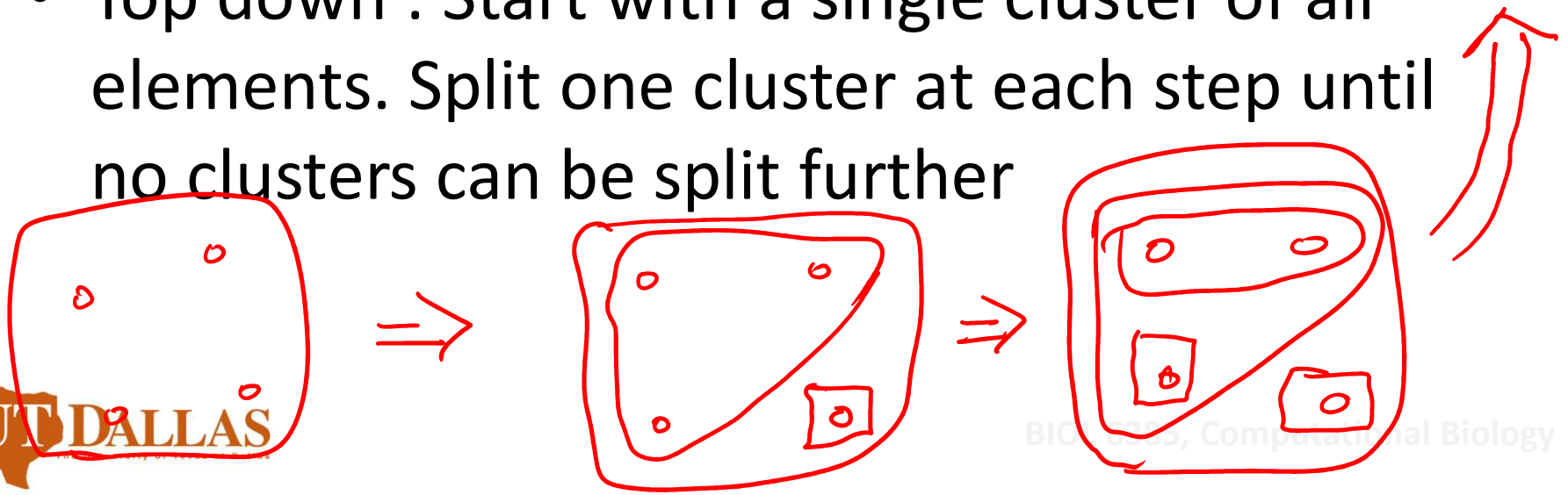
Each comma corresponds to an internal node, hence unrooted tree has one ternary node

Top down & bottom up

- Bottom up : Start with each element in its own cluster. Join clusters until one is left.



- Top down : Start with a single cluster of all elements. Split one cluster at each step until no clusters can be split further

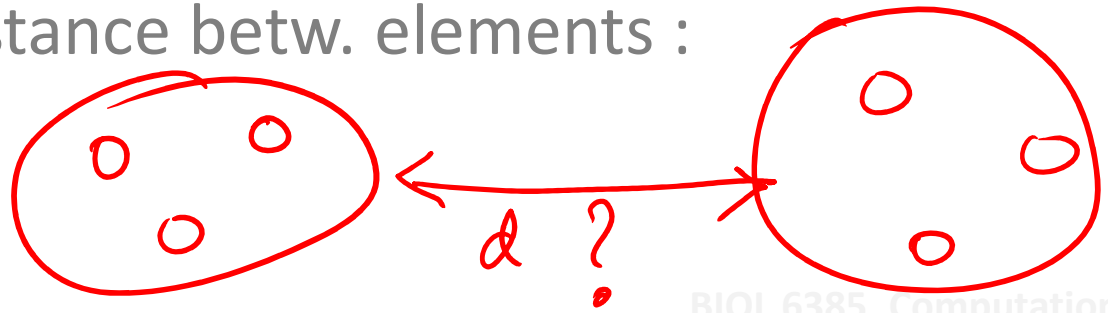


Simple bottom-up hierarchical clustering

- Start with every taxa in its own cluster
- Join the two nearest clusters at each step
- Recompute distance matrix
- Repeat until only one cluster is left

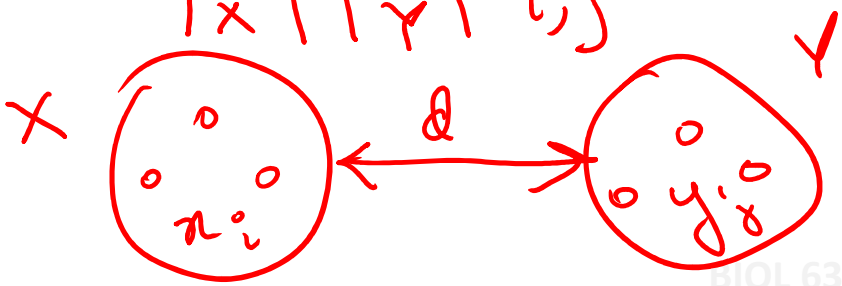
Distance between 2 clusters, cluster and an element based on distance betw. elements :

- Max – link
- Min – link
- Avg – link



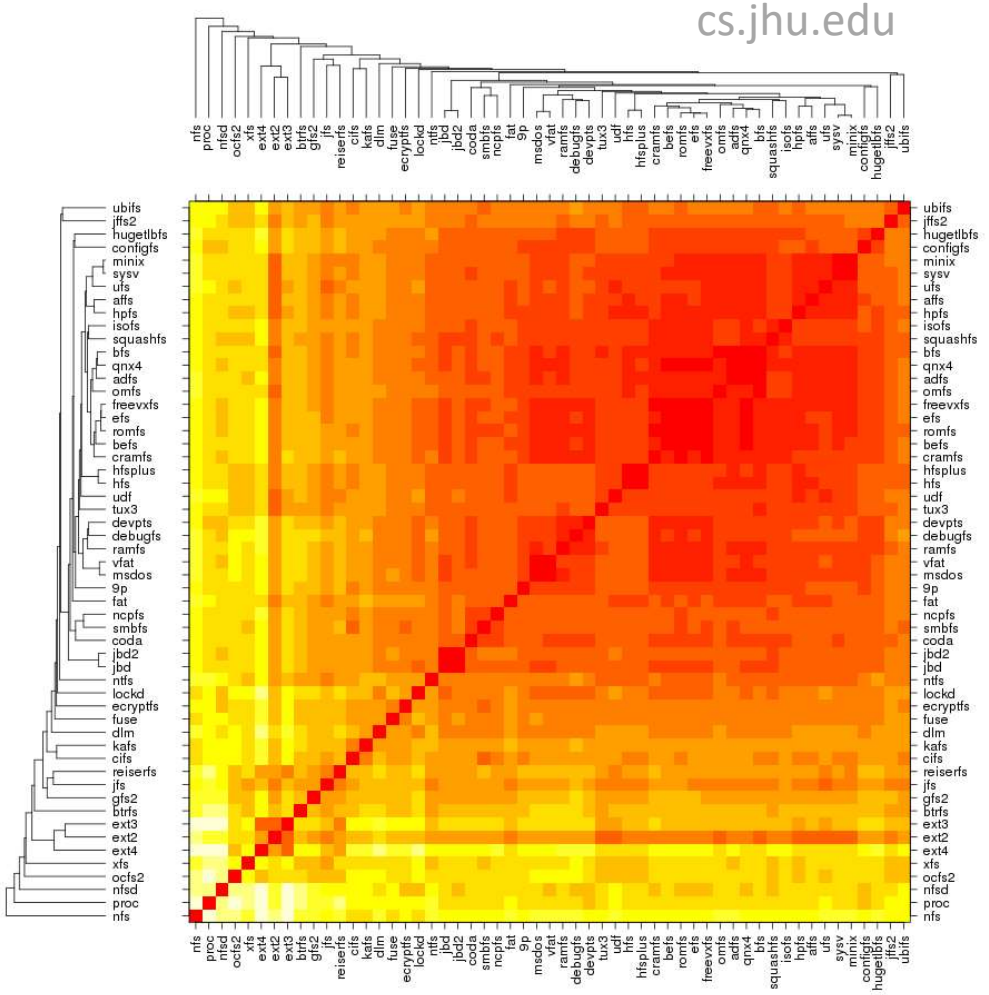
Distance between clusters

- For single elements, consider them to be in their own cluster, then compute :
  - max link $\max_{i,j} d(x_i, y_j)$
  - min link $\min_{i,j} d(x_i, y_j)$
  - avg link $\frac{1}{|X||Y|} \sum_{i,j} d(x_i, y_j)$



Input / output of hierarchical clustering

- Input : a metric, the data points / heatmap
- Output : tree relation of the clusters, if possible branch lengths of tree (if further possible a root)



Lets build some trees …

UPGMA

- Unweighted pair-group method with arithmetic mean

- In our language : bottom up hierarchical clustering with average-link distance between clusters

- Gives a clocked (hence rooted) tree ! (how ?)

UPGMA

– See a beautiful handtrace here :

http://www.southampton.ac.uk/~re1u06/teaching/upgma/

– However, UPGMA is clocked : lineage specific evolutionary rates (if 3 PC is violated) cant be handled

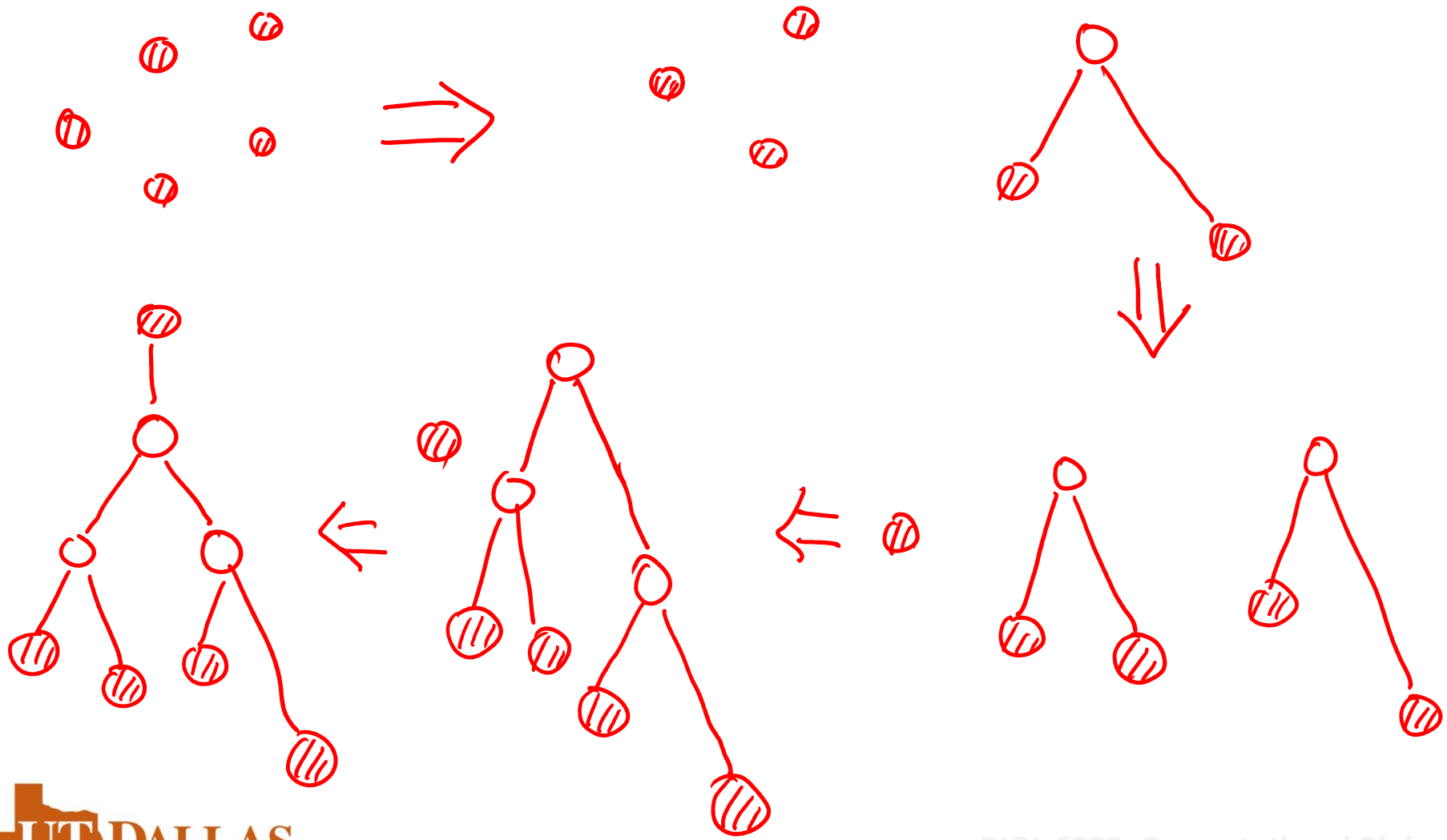
Neighbor joining

Make each taxa a tree of one species

- At each step identify 2 trees which are the most similar to each other AND further apart from the rest
- Make a new ancestral node and connect these 2 trees to it with **different branch lengths**
  - Accounts for lineage specific evolutionary rates : makes an unclocked tree : so rooting is arbitrary if performed
- Update distance matrix

Repeat until two trees are left, then connect them

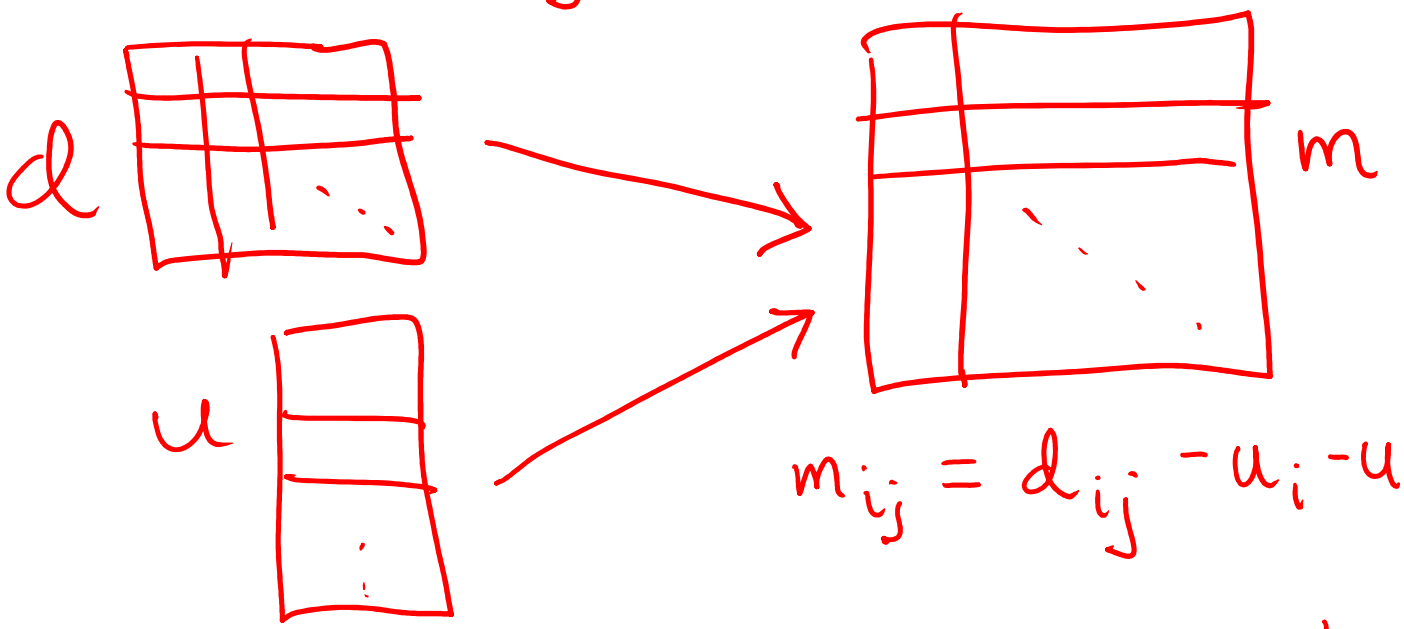
Neighbor joining



Neighbor joining

- Notion of how different from the rest of the leaves one element is :

$$u_i = \sum_j d_{ij} / (n-2)$$

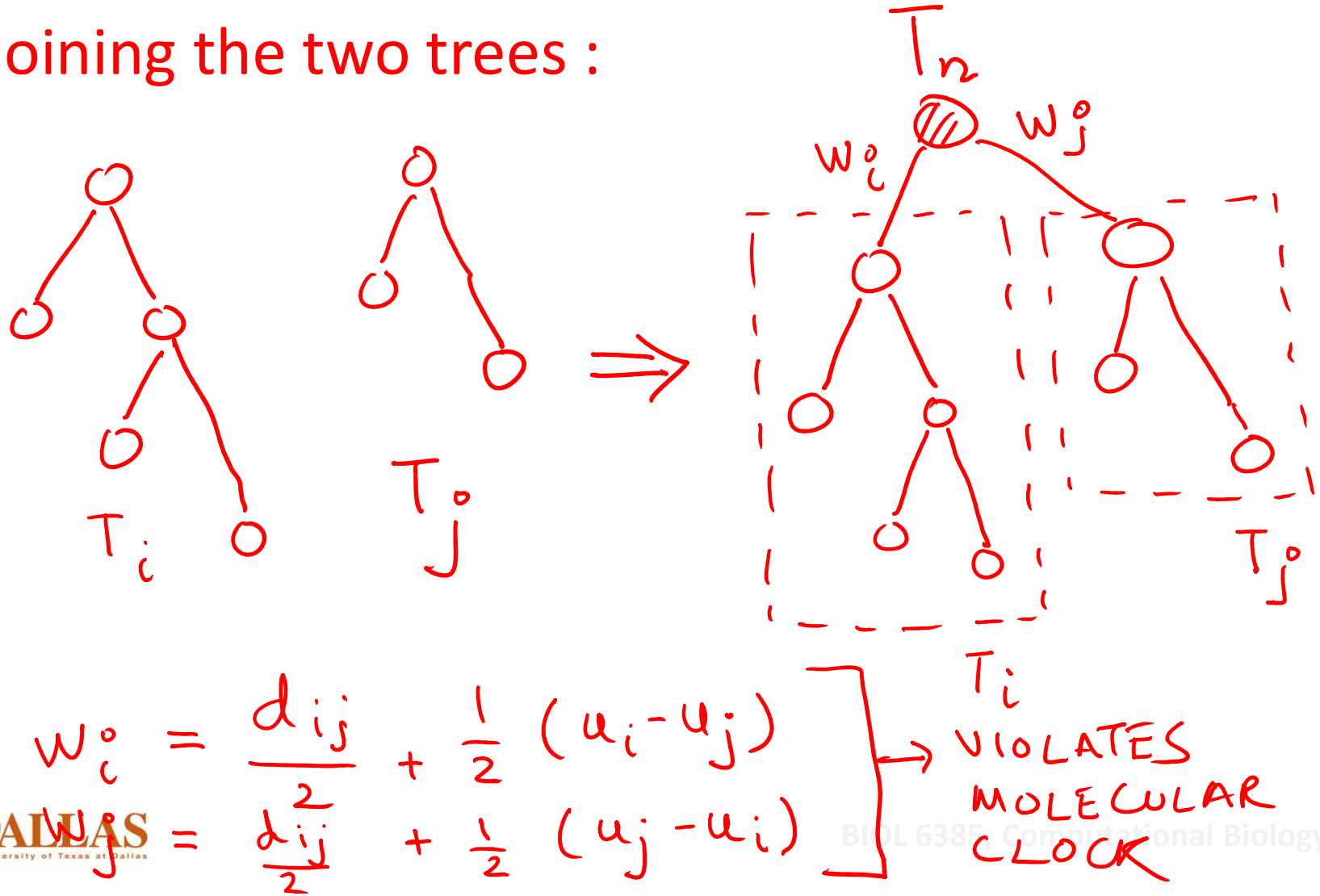


$$m_{ij} = d_{ij} - u_i - u_j$$

Pick i & j based on lowest m_{ij}

Neighbor joining

- Joining the two trees :



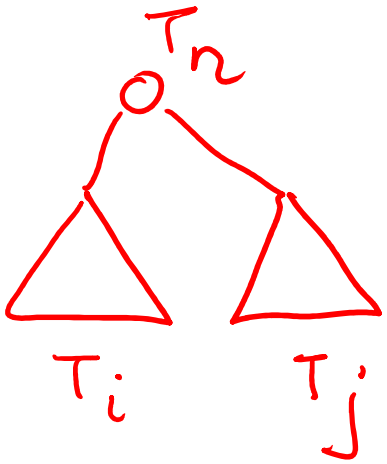
$$w_i^o = \frac{d_{ij}}{2} + \frac{1}{2}(u_i - u_j)$$

$$w_j^o = \frac{d_{ij}}{2} + \frac{1}{2}(u_j - u_i)$$

→ VIOLATES MOLECULAR CLOCK

Neighbor joining

- If i th and j th trees were joined, then the distance from the new ancestor to any other remaining tree k is updated, and the original 2 trees are removed from the distance matrix :



$$d_{n,k} = (d_{ik} + d_{jk} - d_{ij})/2$$

- Originally, when each tree has single element, the distances matrix are just distances betw taxa

Neighbor joining

- Another beautiful handtrace :
  - http://www.cbs.dtu.dk/dtucourse/cookbooks/gorm/27615/molevol.powerpoints/MolEvolClass05.ppt

Bottom up hierarchical clustering

- Decisions :
  - how to identify neighbors and merge them together into a single tree / cluster
  - how to assign branch lengths
  - how to update the distance matrix once merging two clusters is performed

In practice …

- Most distance matrices don't follow 3PC or 4PC
  - Converting an arbitrary distance matrix to 3PC or 4PC is difficult and may change the nature of relationships unfaithfully
  - Forcing a hierarchical clustering on a non – conforming matrix leads to wrong branch lengths and / or topology
    - Squared error = sum of squares of difference between distance matrix and distance on tree for each taxon pair

Then, what ?

- Pick any tree (!!!) – you may start with NJ tree
- Calculate squared error between the original distance matrix and distances on the tree
- Change the tree slightly : if the squared error decreases, keep the changes
  - Keep doing this step till no more improvement

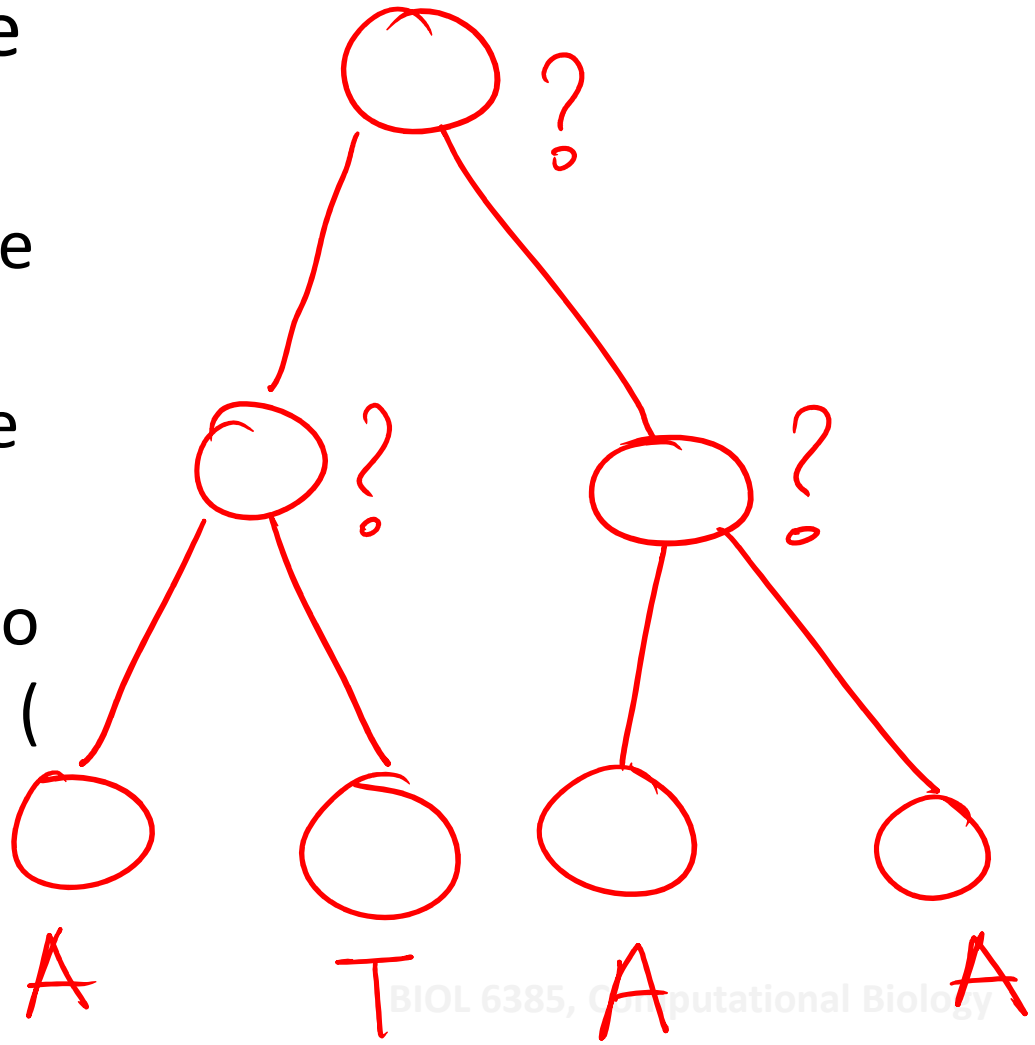
How to change a tree "slightly" ? (coming up next class)

Notion of a “better” tree

- Distance methods do not provide a rational notion of whether one tree is better than the other
  - Most distance methods produce one tree
  - Comparing trees across methods is difficult
- Minimum evolution : minimal sum of branch lengths is better
  - Premise may not be true for distant species

Figuring out the ancestral state

- How to figure out the ancestral state ?
  - going from phenotype to distance is relatively easy (define a metric)
  - going from distance to phenotype is difficult (may be ambiguous)



To the rescue …

- Character based methods

A B C D E
A X C D E
A X Y D E

model

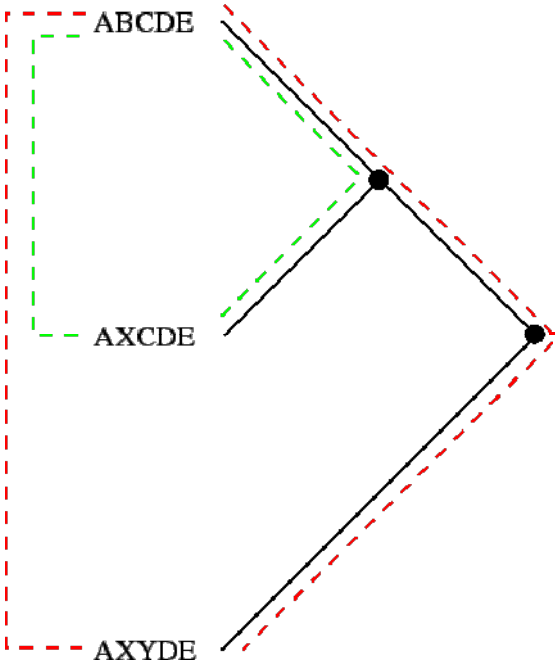


Figure: A Phylogenetic Tree

molgen.mpg.de

Acknowledgements

- Eric Xing
- Dannie Durand
- R Ravi