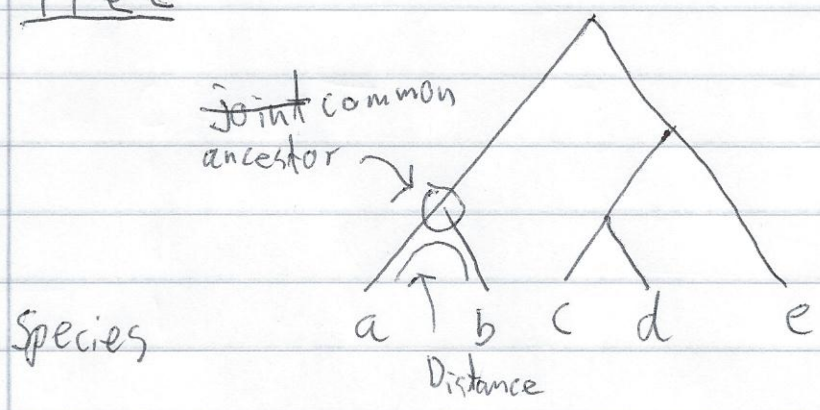
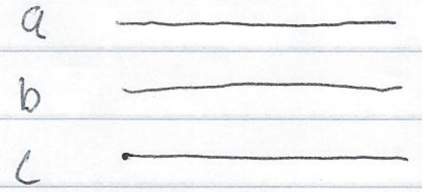
1 Tree reconstruction

Tree



Reconstruction

From sequence data



Multiple String Alligment

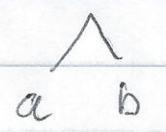
Gives us distances between species

	a	b	c
a			
b			
c			

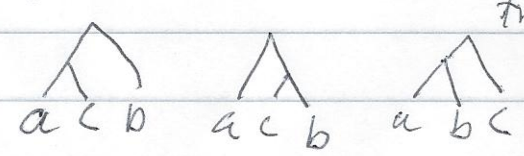
Bruteforce

Can we bruteforce tree topology? How many binary rooted trees?

#2



#3



#edges = $2n-2$

$$R(n) = R(n-1)(2(n-1)-2+1)$$
$$= R(n-1)(2n-3)$$
= exponential!

$$Q(T) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(D_{ij} - d_{ij})^2$$

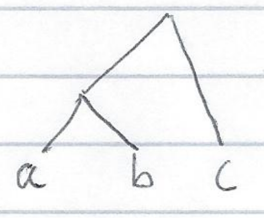
D_{ij} = distance induced by T
d_{ij} = distance matrix
w_{ij} = confidence (eg: 1)

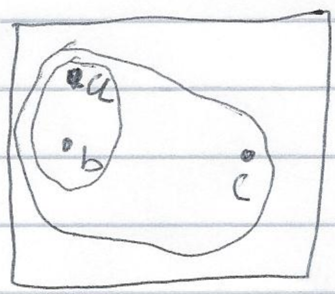
IF T is:
unknown: NPC
known : solveable using LSM $O(n^3)$

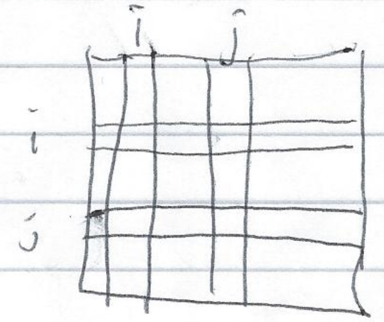
Hierachical clustering

1) Make n individual clusters $O(n)$
2) Join C_i, C_j into C_K where D_{ij} is minimum $O(n^2)$
3) Update D $O(n)$
4) Loop until only one cluster remains

$O(n^3)$ (steps 2–4)

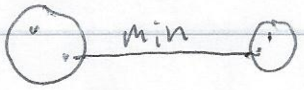




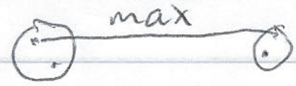


What distances to use?

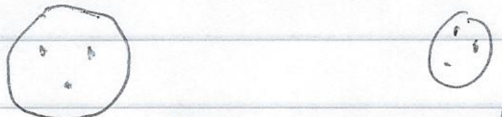
Single linkage



Full linkage



Average linking



$$D(C_i, C_j) = \frac{1}{|C_i| \cdot |C_j|} \cdot \sum_{\substack{p \in C_i \\ q \in C_j}} d(p,q)$$

Can $C_K = C_i \cup C_j$ and C^*'s distance be easy calculated?

$$D(C_K, C^*) \overset{\text{def}}{=} \frac{1}{|C_K| \cdot |C^*|} \cdot \sum_{\substack{p \in C_K \\ q \in C^*}} d(p,q)$$

$$\overset{C_i \cap C_j = \emptyset}{=} \frac{1}{(|C_i| + |C_j|) \cdot |C^*|} \cdot \left(\sum_{\substack{p \in C_i \\ q \in C^*}} d(p,q) + \sum_{\substack{p \in C_j \\ q \in C^*}} d(p,q) \right)$$

$$= \text{—''—} \cdot \left(D(C_i, C^*) \cdot |C_i| \cdot |C^*| + D(C_j, C^*) \cdot |C_j| \cdot |C^*| \right)$$

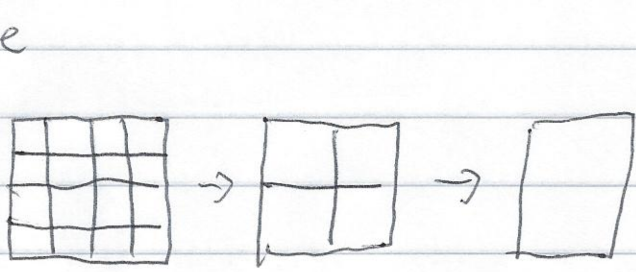
$$= \frac{D(C_i, C^*) \cdot |C_i| + D(C_j, C^*) \cdot |C_j|}{|C_i| + |C_j|}$$

Speed up using Quad Tree

Construction: $O(n^2)$

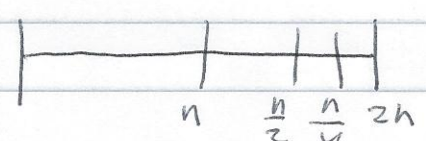
Find Minimum: $O(1)$

Update rows: $O(n)$



Update: $\cancel{4n + 4\frac{n}{2} + 4\frac{n}{4}} + \dots$

$$4\left(n + \frac{n}{2} + \dots + 1\right) = 4(2n) = 8n = O(n)$$

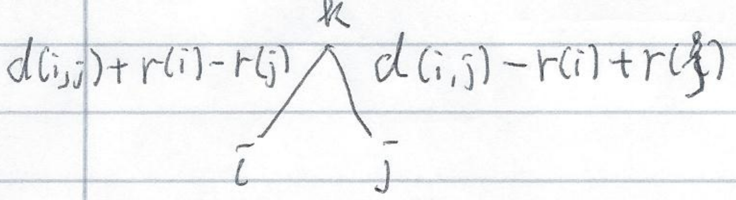


Neighbour joining

Join clusters that are close but far away from others

$$D(i,j) = d(i,j) - r(i) - r(j)$$

$$r(i) = \frac{1}{\#\text{clusters} - 2} \cdot \sum_{p \in \text{clusters}} d(i,p)$$



$$\cancel{D(k, C^*) =} \quad d(C_k, C^*) = \frac{d(i, C^*) + d(j, C^*) - d(i,j)}{2}$$