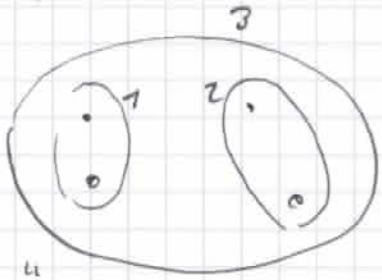
2.3.4 UPGMA algorithm

UPGMA unweighted pair group method using arith. averages, Sokal & Michener 1958

Idea: successively combine the closest sequences (groups of s.) until one large group is formed *



Given ℓ sequences to be grouped

Algorithm:

Initialization:

1) Assign each s to a "one-sequence" group
2) Calculate the distance matrix $D^{(0)}$ for all pairs of groups

Iteration(k) while $n_{group} > 1$

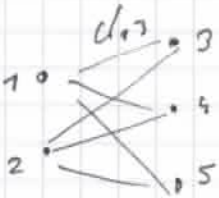
1) Find closest pair of groups: $\min(D^{(k-1)}) := p, q$
2) Join p and q to form new group pq
3) Add a node on distance level $d_{pq}/2$
4) Re-calculate distance matrix for "pq" group using the recursion formula

$$d_{pq,i} = \frac{1}{n_p + n_q}(n_p d_{p,i} + n_q d_{q,i})$$

Works if ultrametricity condition is fulfilled (2013)

* define the distance of two groups as average distance of all pairs of elements.

$$d_{\alpha\beta} = \frac{1}{n_\alpha n_\beta} \sum_{i \in \alpha} \sum_{j \in \beta} d_{i,j}$$



UPGMA: calculate $D^{(0)}$

		1	2	3	4
s_1	GGGAA...A	0	2	6	6
s_2	GGAGA...A	2	0	6	6
s_3	AAAAGGGAA	6	6	0	4
s_4	AAAAGAAGG	6	6	4	0

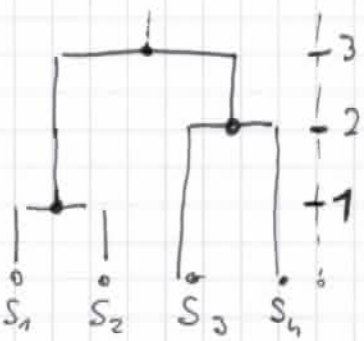
$D^{(0)}$

iteration 1:

closest pair is $p=1$, $q=2$

new group is $(12) = \{s_1, s_2\}$

distance level is $d_{1,2} = 2$



new distances:

$$d_{12,3} = \frac{1}{n_1+n_2}(n_1 d_{1,3} + n_2 d_{2,3})$$

$$= \frac{1}{2}(6+6) = 6$$

$$d_{12,4} = 6$$

$D^{(1)}$

	(12)	3	4
(12)	0	6	6
3	6	0	4
4	6	4	0

iteration 2:

closest pair is $p=3$, $q=4$

new group: $(34) = \{s_3, s_4\}$

distance level is $d_{3,4} = 4$

new distances:

$$d_{12,34} = \frac{1}{n_3+n_4}(n_3 d_{12,3} + n_4 d_{12,4})$$

$$= \frac{1}{2}(6+6) = 6$$

$D^{(2)}$

	12	34
12	0	6
34	6	0

iteration 3:

closest pair is $p=(12)$ $q=(34)$, $d=6$