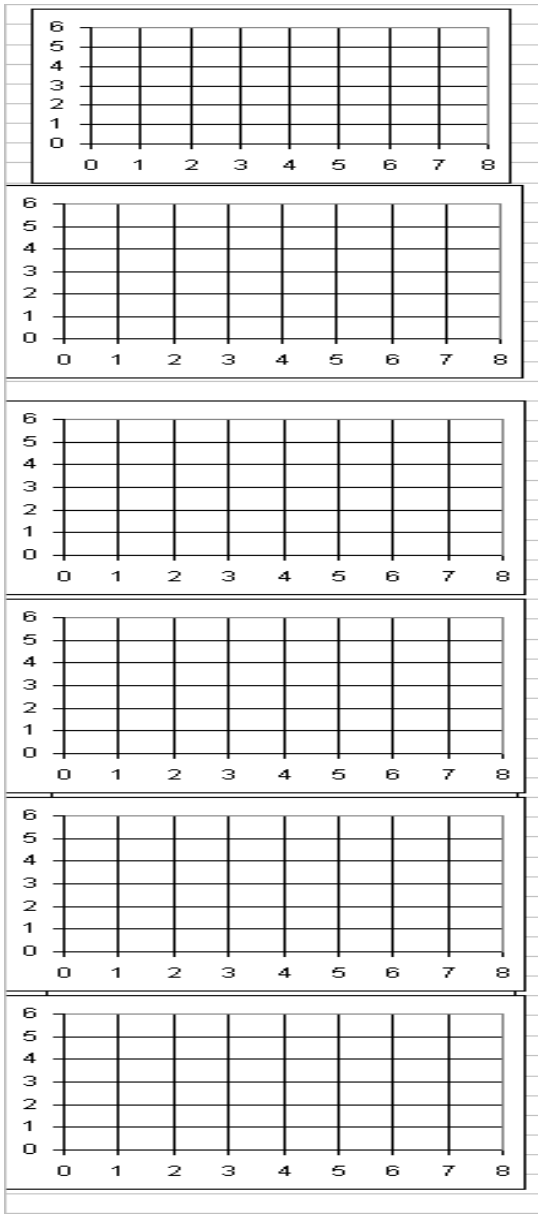
LAST NAME : 8th week, mars 2012
FIRST NAME : Group : .

Mathematics for Biology (2nd semestre) : Answer Sheet 7
Hierarchical Cluster Analysis

Exercice 1. : We want to perform a clustering of the following points using Hierarchical Cluster Analysis and single linkage between clusters : $M_1 = (1\,;4)$, $M_2 = (4\,;4)$, $M_3 = (5\,;3)$, $M_4 = (2\,;2)$, $M_5 = (2\,;5)$, $M_6 = (7\,;2)$ et $M_7 = (2\,;1)$.



	M1	M2	M3	M4	M5	M6	M7
M1							
M2							
M3							
M4							
M5							
M6							
M7							

1. Compute the square of the Euclidian distance of M_1 and M_2 and of M_1 à M_3.

2. Complete the first table as the distance matrix of the seven points and place the points in the 5 figures at the left.

3. Agregate on the second figure the two nearest points, in cercling them with a curve, to form the first cluster of two points and complete the second distance matrix using the single linkage.
4. Continue to agregate the next nearest clusters and compute the new distance matrix.
5. Drow the dendrogram that summarize the process.

Exercice 2. : *(Exercice inspired by a paper of John Hartshorne from the Journal of the British Ecological Society)*

A researcher in Ecology is studying the micro animal species leaving in ponds or rivers. She decide to perform several water samplings in 9 specific sites, 6 in some river sites denoted $R1$, $R2$, $R3$, $R4$, $R5$ et $R6$, and 3 in some small lakes denoted $E1$, $E2$ et $E3$. Repeated samplings allow her to set up the list of species present in each site and to notice the one that are present in several sites among the one studied. The following matrix contains, for each site A and B, the number of species present in these two sites. For example, 11 species are present in site $R1$ but only 7 in site $R1$ and together $R2$.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	11	7	4	6	6	7	4	4	3
$R2$	7	15	8	8	9	6	3	3	2
$R3$	4	8	13	7	7	4	2	3	2
$R4$	6	8	7	15	7	6	6	8	6
$R5$	6	9	7	7	12	4	3	5	4
$R6$	7	6	4	6	4	10	6	5	5
$E1$	4	3	2	6	3	6	13	10	9
$E2$	4	3	3	8	5	5	10	15	11
$E3$	3	2	2	6	4	5	9	11	12

We want to groupe the 9 sites en 3 or 4 clusters containing sites having about the same species. Let's do this clustering using the following distance, between site A and site B :

$$d(A,B) = \frac{n_A + n_B - 2n_{AB}}{n_A + n_B}$$

where n_A (resp. n_B) denote the number of species present in site A (resp. au site B) and n_{AB} the number of species present together in sites A and B.

1. Compute $d(R_1, R_3)$ and then $d(R_2, E_1)$.

2. What is the distance of two sites having no species in common and the distance of two sites having exactly the same species.

3. Assume that the following matrix is the distance matrix of the 9 sites. Complete the empty column without computation. Explain.

	$R1$	$R2$	$R3$	$R4$	$R5$	$R6$	$E1$	$E2$	$E3$
$R1$	0	$0,462$	.....	$0,538$	$0,478$	$0,334$	$0,666$	$0,692$	$0,74$
$R2$	$0,462$	0	.....	$O,466$	$0,334$	$0,52$	$0,786$	$0,8$	$0,852$
$R3$	$0,666$	$0,428$	.....	$0,5$	$0,44$	$0,652$	$0,846$	$0,786$	$0,84$
$R4$	$0,538$	$0,466$	.....	0	0.481	0.52	$0,571$	$0,466$	$0,556$
$R5$	$0,478$	$0,334$	.....	$0,481$	0	$0,636$	0.76	$0,63$	$0,666$
$R6$	$0,334$	$0,52$	.....	$0,52$	$0,636$	0	0,478	0,6	$0,546$
$E1$	$0,666$	$0,786$	.....	$0,572$	$0,76$	$0,478$	0	O,285	$0,28$
$E2$	$0,692$	$0,8$	.....	$0,466$	$0,63$	$0,6$	0,285	0	$0,185$
$E3$	$0,74$	$0,852$	.....	$0,556$	$0,666$	$0,546$	$0,28$	$0,185$	0

4. The hierarchical cluster analysis produce the following dendrogram.

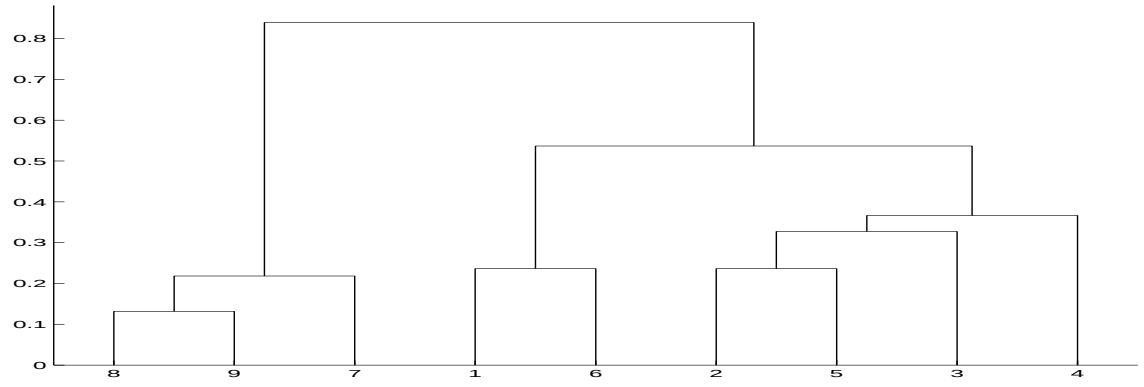


FIG. 1 – Clustering of the 9 sites

Describe the content of the different clusters when cutting the dendrogram at the hight 0.7.

5. Same question when cutting at 0.3.

6. Among the two previous partitions, which one corresponds to the maximal jump ? Is it the more apropriate according to you ?

7. Imagine a realistic situation where such a clustering of ponds and rivers sites could be usefull.