Introduction to Bioinformatics (autumn 2005)

Excercise 7

Group	time	place
Riikka Kaven	Tuesday 29.11 at 12.15–14.00	BK106

1. Consider the following 6×3 intensity matrix M produced by a DNA array experiment:

	Time X	Time Y	Time Z
Gene 1	2.0	11.0	0.5
Gene 2	12.0	0.1	10.0
Gene 3	5.0	5.0	10.0
Gene 4	8.0	4.0	2.0
Gene 5	2.0	1.0	0.5
Gene 6	0.5	1.0	2.0

The values in the intensity matrix M represent the amount of mRNA in the experiment data divided by the amount of mRNA in the control. Before data analysis, the following two preprocessing steps are executed:

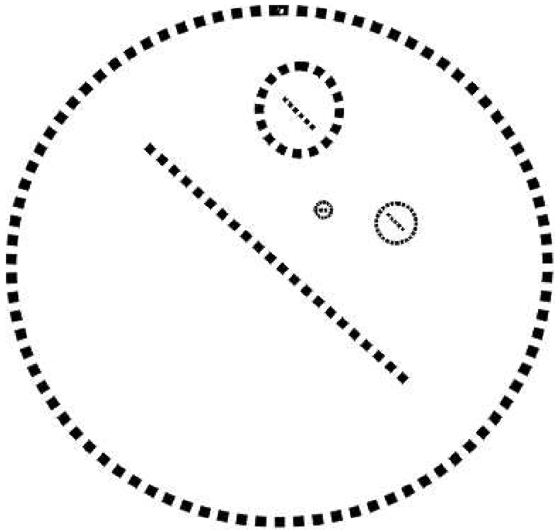
i) The matrix M is converted into matrix M_{log} where logarithm is taken from the numbers: $M_{log}[i,j] = \log M[i,j]$.

ii) The matrix M_{log} is converted into a matrix $M_{logdif}[1 \ldots 6, 2 \ldots 3]$ where only the differences between values in consecutive columns is stored: $M_{logdif}[i,j] = M_{log}[i,j] - M_{log}[i,j-1]$.

Why step (i) is necessary? How do you think step (ii) would help in data analysis?

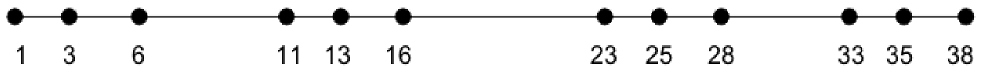
2. Reveal at least 3 possible regulation dependencies from the intensity matrix M of previous assignment. Which of those you think clustering would find?

3. How would you cluster the point set below?



Do you find a clustering criteria that would automatically produce your solution? What is the home-take message of this example?

4. The one-dimensional point set below has obviously 4 clusters.



Give a starting configuration for Lloyd's algorithm (for 4-means clustering) such that

a) the algorithm converges to the correct solution.

b) the algorithm fails to find the correct solution.

5. Visualize the hierarchical clustering for the example in the previous assignment.