CENG 465 – Introduction to Bioinformatics
Spring 2006-2007

Assignment #3 (Programming Assignment)
K-means clustering of microarray data
Due Date: June 04, 2007

Programming Assignment about Clustering Microarray Data

In this assignment, you are going to implement the K-means clustering algorithm to cluster *C. elegans* genes based on their gene expression profiles. For this, you will analyze *C. elegans* microarray data available at http://www.ceng.metu.edu.tr/~tcan/ceng465/microarrayData.zip . The file contains expression ratios as floating point numbers for 1000 *C. elegans* genes in 553 microarray experiments. The first line in the file describes the columns. Each line following the first line contains tab separated gene expression ratios followed by the gene identifier.

Using the Euclidean distance as the distance measure between pairs of gene expression profiles, cluster the 1000 genes using K-means clustering algorithm. Experiment with the following values of **K** $\in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200\}$. Evaluate the quality of a certain clustering of genes using the following quality measure:

$$\text{Cluster Quality} = \frac{\sum_{i=1}^{K}\sum_{j=i+1}^{K} \text{distance between centroids of clusters } i \text{ and } j}{(K^2 - K)/2} - \frac{\sum_{i=1}^{K} \text{average intra cluster distance of cluster } i}{K}$$

where the average intra cluster distance is the average distance between all pairs of points in a cluster. In other words, for a cluster of *m* genes:

$$\text{Average intra cluster distance} = \frac{\sum_{x=1}^{m}\sum_{y=x+1}^{m} \text{distance between gene x and gene y}}{(m^2 - m)/2}$$

Important Notes: Sometimes a gene may have the same distance to more than one cluster centroid; you can arbitrarily assign the gene to any of such clusters. When a gene has 0 (zero) distance to more than one centroid; such situations may cause some clusters to have 0 (zero) members at the end of cluster reassignment. You should ignore clusters with 0 members (i.e., decrease K by the number of such clusters) for such cases. Also you should

ignore clusters of size 1 (one) when computing the "average intra cluster distances", (i.e., do not count such clusters in the right part of first Cluster Quality equation; therefore you should not divide by K, it should be a smaller number).

a) Which value of **K** produces the best clustering of the genes, i.e., maximize the cluster quality measure given above? Does the best K value change on different runs of your program? Why/Why not? Run your program 10 different times to answer this question. Which value of K produces the best average quality considering the average of the quality values you obtained in these 10 runs?

b) We expect that genes that work in the same biological process will have similar expression profiles. Validate this hypothesis using the following sets of genes, some of which are involved in the same biological process and others which are random sets.

Gene Set	Genes in the set	Biological Process Name
1	B0272.3, T02G5.8, T05G5.6	Fatty acid biosynthesis
2	F58B3.5, K02F2.2, R03D7.1	Methionine metabolism
3	F23B12.5, R05F9.6, T03F1.3	Glycolysis / Gluconeogenesis
4	W03F8.3, ZK673.7, C40H5.6	Random set
5	M106.1, F28F8.6, JC8.3	Random set

Describe whether genes in the gene sets given above are clustered together or not. We want the genes in the random sets to be in different clusters and genes in a specific biological process to be in the same cluster. Instead of the inter/intra cluster distance quality measure used in part (a), use the following quality measure to find the best value of **K** (again **K** $\in \{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200\}$):

$$\text{Cluster Quality} = \begin{array}{l} \text{number of clusters spanned by gene set 4} + \\ + \text{number of cluster spanned by gene set 5} \\ - \text{number of cluster spanned by gene set 1} - \\ - \text{number of cluster spanned by gene set 2} - \\ - \text{number of cluster spanned by gene set 3} \end{array}$$

Again, you will try to maximize the cluster quality when searching for the best value of **K**. For example, for the best case in which all the genes in the random sets are in different clusters and all the genes in a specific biological process are in the same cluster, the quality will evaluate to $3 + 3 - 1 - 1 - 1 = 3$. Does the best K value change on different runs of your program? Why/Why not? Run your program 10 different times to answer this question. Which value of K produces the best average quality considering the average of the quality values you obtained in these 10 runs?

Deliverables:

- The source code of your program(s). You may use any programming language of your choice.
- A short report containing your answers to parts (a) and (b) given above.

Submission:

Send the deliverables as a zip bundle or as a tarball to umuero@ceng.metu.edu.tr by the due date. Use the subject [CENG465] Assignment #3 <Your_Student_ID> in you e-mail submission.