CENG 465 – Introduction to Bioinformatics
Spring 2008-2009

Assignment #4 (Data Analysis Assignment)
Analysis of Microarray Data

Due Date: June 1, 2009, 11:59PM

Data Analysis Assignment about Microarrays

In this assignment, you are going to analyze a microarray dataset of eight samples. The experiment is performed on the human genome and contains signal intensities (calculated using the RMA technique) for 11714 human genes. The experiments are performed on two different tissues. In other words *m* samples are taken from a normal **x** tissue and *n* samples are taken from a normal **y** tissue. $m+n = 8$, and *m* may or may not be equal to *n*. Here the tissue types **x** and **y** are also not known. You have two main goals in this assignment:

Goal 1: Determine which sample is taken from which tissue. What is *m*? What is *n*? Provide a matching for each sample to one of the two tissue types. The labels of the tissues can be chosen arbitrarily. Here, the goal is to correctly cluster the samples.

Goal 2: Find 10 genes each that are expressed specifically in each of the tissues. In other words find 10 genes that are highly expressed in the **x** tissue and not expressed in the **y** tissue. Similarly find 10 genes that are highly expressed in the **y** tissue and not expressed in the **x** tissue. You should list a total of 20 genes.

Non Credit Goal: (You grade from this assignment will not be effected from your answer to this goal) Take a wild guess about the real types of these tissues. (Sample guess: liver tissue and kidney tissue).

You are free to use any of the techniques we have learned in class, or techniques you already know, or techniques you have just invented. You may write a program or use existing tools. You should clearly describe the methods you have used to accomplish the goals given above.

You may use your own judgment for any issue that is not specified clearly in this text.

The Dataset:

- The complete dataset can be downloaded as a single tab separated text file from the following address:
  - http://www.ceng.metu.edu.tr/~tcan/ceng465_s0809/Assignments/dataset.txt

Deliverables:

- A short report which contains a step by step description of the tasks that you have performed, the grouping of the samples into the two tissues, and the list of 10+10 genes specifically expressed in each tissue.

Submission:
Submit the deliverable as a .txt document using the COW system.

Late Submission Policy:
Your final assignment grade will be penalized 20 points per late day.

CHECK THE NEWSGROUP REGULARLY FOR POSSIBLE UPDATES ON THE ASSIGNMENT.