

# Spring 2008

## 10-810 / MSCBIO2070: Computational Genomics

### Problem Set 3: Normalization, Expression Analysis and Clustering

**Due date: Feb. 28, 2008, before class**

The contact TA for this assignment is Aabid Shariff (aabid@cmu.edu). Aabid's office hours are at Mellon Institute 409D, on Fridays, 2.30 – 3.30 pm, or by appointment.

Where implementation is required, supply source code in Python, Java, Matlab, C, C++, or R. Code should be portable; it should execute on a UNIX/Linux platform. Please include any auxiliary files and short description of how to execute your code. Submission of this source code may be by email to the contact TA.

Collaboration is permitted, but solutions must be completed individually. Include a list of your collaborators. Refer to the course website for complete policies.

---

### 1. Normalization

In this question you will implement compare several 'between array' normalization methods.

(a) Download the microarray expression dataset file "PS3\_dataset.txt" from the course webpage. The file is a tab-delimited text file. You can simply type "matrix = textread('PS3\_datatset.txt')" in Matlab to load the file. The rows are genes and the columns are 10 different arrays.

(b) Implement the mean and variance normalization method and the rank based normalization method we discussed in class. Plot the distributions (x values ranked order for first microarray and y for second) for the first and second array values following each normalization. Submit your code to <aabid@cmu.edu> with subject "PS3 Code <yourname>".

(c) Next we will implement the dChip normalization method. Read the Methods section (starting with "Normalization of arrays based on an 'invariant set'") of the following paper: **Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application**. Genome Biol. 2001;2(8).

This explains the high level ideas of dChip. There are two design issues that need to be addressed: Which PRD to use for different values and how to construct the piecewise linear normalization curve. These two are linked since our goal is to have enough points to construct the curve in each segment. Assume we are using segments of length 1000 (that is, we have a linear curve for each 1000 genes in the ranking) and assume we need

at least 50 points for each segment to construct the curve. How would you compute the PRD for each segment (not that the PRD should be monotonically non decreasing)?

(d) Write down the equations for calculating the values for the piecewise linear curve. Use least squares to find the best curve. Remember that we want the set of curves to be continuous, that is the curve for the second segment should start where the first curve ended etc. It is fine to start with the first segment (ranks 1-500) and go on to compute the other segments sequentially.

(e) Implement the method you described above and normalize the first and second arrays. List the PRDs you used for each segment when normalizing the first and second arrays. Plot the distributions for the first and second array following the normalization.

## 2. Differential expression analysis

Statistical tests were performed for identifying differentially expressed genes by some method. Assume that we know the true results of the test and we wish to compare it to the method we used. The following table contains the results of this comparison.

Test Result $\Rightarrow$	Not differentially expressed	Differentially Expressed	Total
True Result $\Downarrow$			
Not differentially expressed	P = 9020	Q = 480	U = 9500
Differentially expressed	R = 105	S = 395	V = 500
Total	M = 9125	N = 875	10000

(a) The false positive rate (FPR) is the proportion of negative instances that were erroneously reported as being positive. Compute the FPR for the above tests. How does decreasing the true number of not-differentially-expressed genes affect the false positive rate? Based on the FPR, is the method used for testing a good method? Why?

(b) In multiple hypothesis testing, the False Discovery Rate (FDR) is defined as the **expected** proportion of incorrectly rejected null hypotheses (or expected proportion of false positives among the declared significant results). As discussed in class, we see that the FDR is a useful measure to control. For the above tests, compute the FDR, sensitivity and specificity.

(c) Controlling false positives is an important issue in the identification of differentially expressed genes. FDR is a useful measure to control, in order to control the false positives. In multiple hypothesis testing, another error measure is the family wise error rate (FWER): it is defined as the probability of making one or more false discoveries. (1) Show that the FWER equals the FDR when all the null hypotheses are true. (2) Show that any procedure that controls the FWER also controls the FDR. [*Hint*:  $E(X) = E(X|A=0)P(A=0) + E(X|A \geq 1)P(A \geq 1)$  ]

(d) What is a more useful error measure to control between the FDR and FWER, in order to control the false positives? Why?

### 3. Clustering

In this problem you will see a method of biclustering of gene expression data. The goal of this method is to mine a large matrix (see below) for sub-matrices that represent biclusters.

	Condition 1	...	Condition $j$	...	Condition $m$
Gene 1	$a_{11}$	...	$a_{1j}$	...	$a_{1m}$
Gene ...	...	...	...	...	...
Gene $i$	$a_{i1}$	...	$a_{ij}$	...	$a_{im}$
Gene ...	...	...	...	...	...
Gene $n$	$a_{n1}$	...	$a_{nj}$	...	$a_{nm}$

(a) Some submatrices (describing a bicluster) in an expression matrix will be such that the values in each row or column can be generated by shifting the values of rows and columns by a common offset. See example below.

1	3	5	2
2	4	6	3
4	6	8	5
5	7	9	6

For such matrices, write an expression for the value of an element  $a_{ij}$  as a function of its column mean, row mean and bicluster mean.

(b) Often, finding matrices with such behavior is difficult due to noise in measurement. Hence, we can define  $r(\mathbf{a}_{ij}) = \text{realValue}(\mathbf{a}_{ij}) - \text{expectedValue}(\mathbf{a}_{ij})$ . Using this metric, we can compute the overall quality of a bicluster using a score  $S$  defined by,

$$S(I, J) = \frac{1}{|I||J|} \sum_{i \in I, j \in J} r(a_{ij})^2$$

Compute the scores  $S$  for the following matrices. What do the scores mean biologically?

2	2	2	2
2	2	2	2
2	2	2	2
2	2	2	2

1	3	5	2
2	4	6	3
4	6	8	5
5	7	9	6

1	2	0.5	1.5
2	4	1	3
4	8	2	6
3	6	1.5	4.5

(c) Assume you have an algorithm that identifies biclusters one at a time. Suppose you identified a bicluster. How will you find another bicluster without identifying the same one as before? Do you see any problems with your approach?

(d) Now we would like to develop a greedy iterative search algorithm that can find for us  $K$  biclusters. Assume that you had a method of node removal, node addition on some condition being satisfied every iteration of the step. Describe an algorithm that uses these addition, removal algorithms to identify  $K$  biclusters in your graph.