

# Microarray Analysis Demonstration

In this microarray analysis demonstration you will be using data from the Affymetrix *E. coli* 2.0 microarray chip. This array represents approximately 14,000 well-characterized *E. coli* genes. All of the probe set sequences used on the GeneChip *E. coli* Expression Array *E. coli* were taken from GenBank, dbEST (Expressed Sequence tag), and RefSeq.

Today's tutorial will take you through a very simple microarray analysis. We will start out with raw data as returned from the microarray processing facility. In order to save time the data has already been normalized (background fluorescence removed) and annotated for you. Annotation simply means the probe numbers on the microarray have been replaced with gene names and descriptions to make the tutorial easier to understand. The data set we will be working with is a set of stressed *E. coli* data. *E. coli* cells were grown and exposed to a DNA damage agent at time zero. This is the experimental culture of *E. coli* cells. Control *E. coli* cells were grown at the same time without exposure to the DNA damage agent. Samples from both control and experimental cultures were taken at 2 minute intervals up to a total time of 10 minutes. The mRNA was extracted, converted to cDNA and hybridized to two separate chips (experimental and control) at each time point. The control mRNA samples were labeled with Cy3 (green) and the experimental mRNA samples were labeled with Cy5 (red). The two chips were compared for each time point resulting in a green color if a gene is expressed at lower levels in the experimental cells and a red color if a gene is expressed at a higher level in the experimental cells. This will produce a profile of the time course response of the cell to stress for each gene in its genome.

Generally, microarray analysis consists of three separate parts.

1. Normalization and Annotation (associating gene names to probes)
2. Clustering or Grouping of genes with similar expression patterns in the experiment (up-regulated or down-regulated).
3. Applying a Test (searching the clusters for genes with similar biological function)
- 4.

## Loading Normalized *E. coli* Microarray Data into MeV (Multiple Experiment Viewer)

*You can download a copy of this program by going to <http://www.tm4.org/mev/>*

1. Double click on the MeV application to open the program.
2. To load data into MEV, pull down the "File" tab in the upper left hand corner of the "Multiple Array Viewer". Then Click on "Load Data".
3. Click on 'Browse' and find the DemoData.txt file in the MeV folder on the desktop. Highlight this file and click on 'open'.
4. The data file will appear in the Expression Table port. This is our data. The first column is the Gene Name and the second column is the Description of the Gene. You will need to tell the program where the gene expression data is located in this file. Click on the upper leftmost expression value. This will be the expression data in the cell just below the header labeled "Time0". Press "Load" to load the data.

5. Adjust the Color Scheme of the array by pulling down the Display tab in the upper left hand corner of the Multiple Array Viewer. Then click on “Set Color Scale limits”. Set the lower limit to -1.0, the midpoint to -0.2, and the upper limit to 1.0. This will allow us to easily visualize the over and under expressed genes in the experimental cells as compared to the control cells.
6. Display gene descriptions by once again pulling down the display menu, highlight ‘Gene/Row Labels’ and click on ‘Label by Description’.
7. What you see in front of you is a series of 6 different microarray experiments. Each column represents the data from a single microarray experiment at a given time (0 min., 2 min., 4 min., 6 min., 8 min., and 10 min.). Each row represents the time course expression for a single gene. We can actually watch a single gene response to stress. Clicking on an individual square will bring up additional data about the gene including the gene name, description, and expression ratio. By looking at this data you should be able to determine which genes are up-regulated (red) and which genes are down-regulated (green) in the stressed cells (experimental) compared to the control cells.

$$M = \log (\text{Red}/\text{Green}) = \log \text{Red} - \log \text{Green}$$

**If M>0**

More Cy5 (red) has hybridized to the microarray chip than Cy3 (green). Remember, RNA from the stressed cells was hybridized to Cy5 (Red) and RNA from the control cells was hybridized to Cy3 (green). Therefore this gene is up-regulated in the stressed cells as compared to the control cells.

**If M=0** (or really close to 0)

Stressed and Control are equally expressed

**M<0**

More Cy3 (green) has hybridized to the microarray chip than Cy5 (red). Therefore this gene is down-regulated in the stressed cells as compared to the control cells.

All of the green squares represent genes that are under expressed in the stressed bacteria compared to the control bacteria. The red squares represent genes that are over expressed in the stressed bacteria compared to the control bacteria .

# Clustering Genes According to Similar Expression Patterns

8. We will group the genes that have similar expression levels throughout the time course of the experiment. To do this we will use a method called K-means clustering. K-means clustering is a very simple way of grouping similar data together. Pull down the 'Clustering' icon in the upper left corner of the multiple array viewer and click on k-Means/Medians Clustering (KMC). A new window called KMC should appear. Set the number of clusters to 50 and press 'OK'.
9. To view the clustering results click on the KMC tab in the left hand side of the multiple array viewer. Then click on "Expression graphs", then "all clusters". This is shown below.
  - a. Click on the Table view icon to view gene names and descriptions for each cluster.
  - b. Click on Expression images for a visualization of the actual microarray chip.
10. Each graph represents a cluster of genes that are expressed in a similar fashion across all time points. These graphs represent gene expression ratios (ratio  $< 0$  = down-regulated genes in the stressed cells; ratio  $> 0$  = up-regulated genes in the stressed cells both compared to non-stressed cells). We are looking at an overall gene response to stress (up-regulation or down-regulation) as well as how the group of genes responds over the time course of the experiment. As the gene responses are turned on and off, the expression ratios change as indicated by fluctuations in the graphs.
11. Return to the original array by clicking on "Original Data" in the navigation bar at the left hand side of the multiple array viewer.
12. Now we will construct a Hierarchical tree of gene relationships based on their expression levels. Pull down the 'Clustering' icon in the upper left corner of the multiple array viewer and click on Hierarchical Clustering (HCL). A new window called HCL should appear. Press 'OK'.
13. This will construct a tree based on gene expression relationships. Genes with the same time course expression will be grouped with their closest relative, much in the same way phylogenetic trees are constructed. To view the tree, click on the HCL button in the navigator at the left hand side, then on the HCL tree.

\*Try many of the other ways to view and analyze this data on your own.