MIC653: Advanced Molecular Genetics Bioinformatics & Computational Genomics

Project #4: Analysis of gene expression data (uses CyanoBIKE)

About 20 years ago microarrays were introduced to measure simultaneously the expression of thousands of genes. With the drastic decrease in cost of sequencing, RNAseq has largely supplanted microarrays as a tool to measure gene expression. Despite the shift in methodology, certain problems remain, for example the issue of reproducibility and the problem of how to compare different experiments performed under different conditions. There are many programs available to help researchers work through these problems and analyze gene expression data. However, what can you do if you want to analyze the data in a way that was not anticipated by those who wrote those programs? For that, you need to be able to program the computer yourself.

A. Introduction to computational analysis of microarray data in BioBIKE

A tour entitled Combining pathways with microarrays* takes you through a specific problem of gene expression analysis, using BioBIKE functions. Go through this tour, replicating each step within CyanoBIKE.

B. How to compare the results of different microarray experiments?

The main purpose of measuring gene expression is usually to determine a ratio of some sort, the change of expression at some time point or condition relative to another. It is also useful at times to know absolute levels of expression – perhaps expression went up by a factor of two, but did it rise from a high level to a higher level or a very low level to a still lower level? Attempts to make sense of microarray data are bedeviled by the variability in expression measurements, which is only partially addressed by replicate measurements. While measurements of gene expression via RNAseq generally have less variability than measurements via microarrays, the issue still remains, and considering how it can be handled with microarrays may offer general insight into the problem.

Ehira et al (2006) [Mol Microbiol 59:1892-1703] measured the expression of genes in the cyanobacterium Anabaena PCC 7120 over the course of differentiation initiated by nitrogen deprivation. Never mind the biology, let's look at the variability in the measurements.

1. Look at the raw expression data from the Ehira experiment, i.e. the fluorescence intensity measurements before statistical manipulation. To do this, bring down INSIDE-MICROARRAY, and specify ehira_2006 as the experiment by mousing over the **Data** tab, then **Microarrays**, then **Anabaena.** Click the *Gene* box and bring down CODING-GENES-OF from the alphabetical list of functions, and type `A7120` as the *entity.* for the gene (all the genes measured by the microarray). Doing this specifies that you want values for all genes in the experiment. Click the *condition* box and type 2 (the condition for 8 hrs after initiation, but that's not important for our purposes). Choose the options RAW and +CONTROL-VALUES, getting the following:

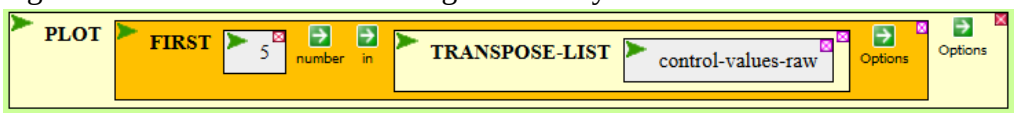
* Available by clicking the link or by going to the BioBIKE portal (http://biobike.csbc.vcu.edu), clicking the link to the guided tours, and choosing the tour called *Combining Pathways with Microarray Data*.

Execute the function. Each line of the results gives you the six replicate measurements for one gene. What is the range of values from amongst replicates of one gene? What is the range of values between different genes?

2. Of course there is variation from one replicate to the next, but are the replicates systematically different from one another? One way of assessing this is to look at the distribution of values for each replicate. Ideally, the distributions should be the same. To determine this, we need to create six sets of values, one for each of the six replicates.. To start off, DEFINE a variable (maybe call it `control-values-raw`), setting its value to the numbers returned by INSIDE-MICROARRY (Step 1) -- drag the above function into the *value* box. Before executing, choose the –LABELS option of INDIDE-MICROARRAY in addition to the others. This will suppress listing of the names of the genes. You'll get just the numbers, which will simplify matters. Execute the function and verify that the result (in the Result Pane) has the same numbers as before.

3. Now to plot the distribution of values for each replicate. Right now, the numbers are grouped horizontally by gene. We want the same numbers grouped horizontally by replicate. You can swap vertical for horizontal by using the TRANSPOSE-LIST function. Bring that function down into the workspace and put *control-values-raw* into the *list* box. Execute the function, and examine the result to verify that the transposition has taken place. The first set of numbers should consist of the first replicate value for the first gene, then the second gene, and so forth.

4. The second step in plotting the distributions is to bring down the PLOT function into the workspace. Unfortunately, this function can display no more than five distributions at one time, so we'll use just the first five replicates. In the *list-or-table* box bring down the FIRST function, mouse over the *number* icon and click *number*, and enter 5 for *n*. Then drag into the *entity* box the TRANSPOSE-LIST function you made in Step 3. The resulting function should look something like what you see below:



5. If you execute this function, you'll get a meaningless plot, where the x-axis is gene 1 through gene 5336, and the y-axis is expression level for the 5 replicates. We want a distribution plot, where the x-axis is the expression level and the y-axis is the number of genes with that level. Do achieve this, go back to the PLOT function and choose the BIN-INTERVAL. If you enter 1 as the BIN-INTERVAL *value*, then the function will count how many values lie between 0 and 1, 1 and 2, 2 and 3, etc. Execute the function…

6. Still not very helpful, because almost all of the values are scrunched near 1. To spread out the distributions, go back to the PLOT function and choose the options MAX, specifying 300 as the maximum value on the X axis. Now when you execute the function, you should see 5 distributions. Do you see them? Are they the same?

7. That was the raw values. Now repeat Steps 1 through 6, replacing RAW in Step 1 with NORMALIZED-BY-MEDIAN. What do you make of the plotted distributions?

If you've made it this far, then you have gone through the process of normalizing a microarray, a procedure required in gene expression analysis whether you use microarrays or RNAseq.