Introduction to Bioinformatics – AS 250.265
Laboratory Assignment 2

Last week, we discussed several high-throughput methods for the analysis of gene expression in cells. Of those methods, microarray technologies are quickly becoming the technique of choice. In this lab, we will follow the steps necessary to normalize and process microarray data using a sample dataset. Your book recommends visiting the Stanford Microarray Database (SMD, http://www.dnachip.org/), but because of the sophistication of the site and the rather confusing nature of its data retrieval system, we will be using a simple text file with spot intensities. You are, however, invited to visit the SMD website yourself and browse the recent publications list to find out how researchers are using this technology.

The dataset that your books uses—and the dataset that we will use here—is taken from a paper by Chu and DeRisi, et. al., investigating how gene expression changes during yeast sporulation. In addition to performing microarray analysis, they also confirm some of their results using Northern Blots, which you will investigate in your homework assignment for this week.

Yeast (*Saccharomyces cerevisiae*) is an organism capable of undergoing both asexual and sexual reproduction. Instead of being classified as male and female, yeast mating types are either a or α. Both of these mating types are haploid, that is, they only possess one copy of their chromosomes. Once a and α mating types have mated, they conjugate (merge) form a new mating type, a/α, which is diploid—possessing one set of chromosomes from each of the yeast's "parents." This new mating type (and only this new mating type) is capable of sporulation: During starvation conditions, this mating type can undergo meiosis to produce four haploid spores, two of which are of mating type a and two of which are α.

What genes are expressed that could be causing sporulation? This is the question these authors attempted to address with microarrays. They compared sporulating cells with non-sporulating cells at different time points in an attempt to address which genes are turned on at which time. Here, we will look at their original microarray data, normalize it, and examine it to see if there are any interesting results. The original paper by Chu and DeRisi, et. al., is available on the course web site as a supplemental file for lab assignment two.

This lab will be a large part of your assignment for the coming week. You will submit your processed microarray dataset as well as the answers to the questions from this lab (which are highlighted in gray). Your dataset should be uploaded to the course web server under assignment five, but you may submit the answers to the lab questions either digitally or by hand at the start of class on March 17. Your answers from this lab, as well as the construction of your spreadsheet, will count as 50 points of assignment number five.

Part 1: Setup and Global Normalization

Downloading and Opening the Dataset

The dataset we will use for this lab is available on the laboratory page of the course website. Download this file (`spospread.txt`) and save it somewhere on your local account. You will notice that it contains lots of text in tabular format. For convenience, we will work with this file in Microsoft Excel.

Open Excel (a green X in your Macintosh dock), and use it to load the spreadsheet data. When you select the spreadsheet file, you will be prompted with several options. You may either simply click "Finish" to use the defaults, or you may scroll through the next several dialog boxes to customize how Excel will load the file.

When all is finished, you will be looking at a spreadsheet containing the raw, unprocessed microarray data. The columns store the expression data for different time points: green represents vegetative (stationary) cells, and red represents sporulating cells. The time points are in hours, so `t0.5` is the time at ½ hour, whereas `t5` is at five hours after the experiment was started. The rows contain the identifiers for nearly all (97%) of the open reading frames (genes) in the yeast genome. Each of these rows correspond to a single spot on the microarray.

1. How many open reading frames are examined in this experiment? You may assume that there are no duplicates in the list.

Normalization of Expression Data

If you look carefully, you will notice that on average the green intensity values are somewhat higher than the red values. This is because the data in the spreadsheet are not *normalized*. Normalization is a technique that is performed to allow differing datasets to be compared. Basically, the assumption is made that *on average* the genes are being expressed at equivalent levels, and therefore any systematic variation in the intensities are an experimental artifact. In fact, it is known that the labeling efficiency of the red fluorescent dye is lower than that of the green dye.

Another experimental artifact arises in microarray data, having to do with nonspecific binding of the RNA to the chip itself. This nonspecific binding results in a "background" level of intensity that has nothing to do with the gene expression. Another part of the normalization procedure is to subtract the background intensity from the spot intensity.

Because of the inconsistencies in the manufacture of each chip, or in the experimental method, background intensities vary along with the *x, y* position on the chip. Thus, after the chip is scanned and converted into a computer image, the software that analyzes the image (and produces the nice Excel worksheet) must also quantify the background "noise" at each location on the chip. You'll notice that for each column there is a "bkg" column, which corresponds to the background signal at that position.

Our goal in this lab is to leave the `spospread.txt` worksheet untouched and use another "safe" worksheet to contain our data. We will also use a third worksheet to store things like normalization values, regression data, etc. Additional worksheets will contain plots, etc. All of these worksheets should be stored in one Excel (.xls) file. If you are not familiar with saving multiple worksheets in Excel, we'll cover this as we go along.

First, let's create a new worksheet that will store the modified data from the two hour time point. Right-click on "spospread.txt" tab at the bottom of your workbook and select "insert." Find the entry to insert a worksheet, and click "Open." After doing this, you may have to select "View→Normal" to view the worksheet without physical page borders. Rename your new worksheet "2hr data." Then, copy the list of ORF's and the data from the two hour microarray dataset to your new worksheet by cutting and pasting the appropriate columns (ignore the "flag" column).

At this point, stop for a discussion of how to use equations and functions in Microsoft Excel.

Subtract the red and green background columns from their respective signal columns. Do this by inserting new columns, naming them appropriately, and using Excel syntax to subtract one column from the other. When you are finished you should have two new columns that contain your background subtracted values. We'll call these the "signal" columns.

Next, create a new worksheet called "2hr params." This worksheet will store normalization (and other) constants for your data so they can be easily accessed. Stretch column A so that it is wide enough to store descriptive labels of your parameters. Column B will store the parameters themselves, next the to appropriate label. Calculate the average values of the raw minus background columns you made above, and store those average values in the parameters worksheet. You can do this using the "AVERAGE" function in Excel.

Now, normalize your data. Insert a new column, next to your columns with the signal data. Then, create an Excel formula that fills in the values of this new column with those from the old column divided by the cell containing your normalization constant. When you specify the location of the cell containing the average value, use dollar signs to prevent Excel from automatically advancing the normalization value as you fill the equation down the column. For example, if your normalization value is stored in cell "B1," specifying "B1" in Excel will prevent that value from changing when you drag your equations elsewhere.

2. Check your work by re-averaging the newly normalized columns of data. Store the new average values (for red and green) in additional rows in the parameter worksheet. What is the average value of both normalized columns?

Plotting the Results of Initial Normalization

Now that you have performed some basic normalization, let's compare the original data with your normalized data. We will do this by plotting the expression levels of vegetative versus sporulating yeast.

Insert a new plot by selecting the raw data columns for the vegetative and sporulating expression intensities and clicking on the graph toward the top of your screen. You can use the apple key to select multiple, discontinuous columns of data. Create a new scatter plot, and follow the directions, formatting your plot and giving it appropriate axis labels. At the final dialog box, insert the plot as a new worksheet named “2hr raw plot.” Repeat this process for the normalized data, inserting it as a new worksheet named “2hr normalized plot.”

3. What is the utility of these plots? Specifically, on the normalized graph, why are the genes that fall significantly above the line y = x important? Why are genes that fall significantly below that line important?

It is often the case for expression data that the dataset is clearer when log-transformed. Insert new columns that contain the base ten logarithm of each of your red and green expression data (the Excel function for $\log_{10}$ is LOG10), and plot that in a new worksheet. Name the new worksheet “2hr log plot.”

As an aside, notice how the x- and y-axes of this plot fall underneath the data and are thus not visible. Microsoft provides no easy way of changing this placement. This is one of the reasons that very few real researchers use Excel for publication-quality graphs.

Checkpoint: At this point you should have five columns for each of the red and green datasets on your “2hr data” worksheet. These are: raw intensity, background intensity, signal (difference), normalized data, and log data. You should also have three charts: non-normalized, normalized, and logarithmic.

4. Compare the plot you have just made of log-normalized microarray data to analogous plot of t = 0 hr data in figure 7.3 of your text (page 195). How do these two plots differ? Why does it make sense that they should differ?

While the plot you just made is useful for viewing gross differences, one final transformation is typically performed on the data. This transformation takes advantage of the fact that most often we are interested in the *ratio* of translation products. For example, if a gene is highly expressed in a given cell, but is roughly the same for both vegetative and sporulating conditions, it will appear in the far right corner as an outlier but will be relatively unimportant. By plotting the log ratio of the two intensities on the y-axis, genes will only be outliers if their ratio of sporulating to vegetative expression is far from unity (recall that log(1) = 0). On the x-axis, a logarithmic mean of the two intensities is plotted, and therefore segregating genes with high expression to the right and genes with low expression to the left.

If I is the standard normalized intensity, these transformation variables are:

$$X = \tfrac{1}{2}\log_{10}(I_{sporulating}) + \tfrac{1}{2}\log_{10}(I_{vegetative})$$

$$Y = \log_{10}\left(\frac{I_{sporulating}}{I_{vegetative}}\right)$$

Append two new columns to your dataset for the mean log intensity and the log intensity ratio. Fill in these columns with the transformations given above. Then, create another plot worksheet called “2hr log transformed” plotting these two values (X, the mean of the log intensities, vs. Y, the log of the intensity ratio).

Part 2: Local Normalization

The plot that you just created shows that there are several genes that are overexpressed by almost a factor of 10 ($\log_{10}(10) = 1$) in your dataset. However, close inspection also shows that there is a slight skew to the data: some of the points at the left hand side of the plot fall below the y-axis, and it looks like the data overall may slant upward. While this is small, it is an artifact that may be important for some datasets. Another advantage of this type of plot is that it makes visualizing this systematic error easier than simply looking at a log-log plot.

Perform a linear regression on your data by right-clicking a data point and selecting "Add Trendline." Under "Options," display the equation of best-fit as well as the regression constant (R^2). When you select okay, you'll see a line has been added to your plot, and you'll be able to drag the equation/R^2 text box to a sensible position on the graph. Copy the value for the slope and intercept into new entries on your "2hr params" worksheet. We will now be correcting the data so that this line is simply $y = 0$.

5. Given that the mean log intensities (the X axis) will remain constant, how should use the linear equation of best fit to transform the Y variables so that the systematic skew will be removed?

Apply your answer from question five to a new column of data: corrected log ratios. Remember that you will need to use dollar signs when specifying the cells containing your slope and intercept. Plot these new Y values along with the original X values and store the plot in a worksheet named "2hr log corrected." Perform another linear regression of this data, displaying the equation of best fit, and show that your transformation was sufficient to remove the tilt to the graph. (Your slope and intercept at this point should be miniscule.)

Now that we have corrected for the most obvious systematic error, we can begin to observe some of the trends in our data.

6. By clicking on a data point and holding your mouse there, Excel will tell you the X and Y values for this data point. Using this, combined with the Excel search functionality, identify the most overexpressed and underexpressed ORF identifiers in this experiment.

Generally, the NCBI website keeps "aliases" of genes that correspond to the ORF identifiers on these DNA chips. You can search for these identifiers using Entrez Gene.

7. What are the RefSeq protein accession numbers and gene names (where applicable) for the proteins that are encoded by these genes? For your overexpressed answer, why does it make sense that this gene may be up-regulated during sporulation?

Part 3: Statistical Significance

While there are methods for determining the significance of microarray data, many of these assume some tangible level of statistical training that would be inaccessible for us. However, it is possible to apply some of the same techniques we used for global alignments to determine a rough estimate for how significant the two outliers we identified above may be.

Recall that for global alignments we defined something called a Z-score: that is, given the variance σ we can determine the number of "standard deviations" our outlier is from the mean μ. In brief,

$$Z = \frac{|x - \mu|}{\sigma}$$

where *x* is the corrected log-ratio value of our outlier. The problem with this methodology is that (1) it assumes a normal distribution for the dataset we are looking at, and (2), it only incorporates one value for *x*, which itself may have a degree of experimental uncertainty. Let's take a look at (1) and see what our distribution of log ratios looks like.

Constructing a Distribution of Log-Ratios

Create a new worksheet in your Excel workbook, and call it "2hr histogram data." We will use some obscure functionality of Microsoft Excel to create a histogram of the distribution of log ratios. Looking at the distribution of data on our plot of corrected, log-transformed data, we see that we'll need bins running from about -1.5 to 1.5, and given the number of data points we have, we can probably get away with a bin size of about 0.05.

On your new worksheet, create a column of numbers containing {-1.45, -1.40, …, 1.5} (be sure to leave room for a column title!). You can do this by typing the first two entries (-1.45 and -1.4), selecting both cells, and dragging the little box that appears in the lower right of your selection down until Excel displays 1.5. These cells will be the bins of our histogram.

To fill in the column containing the number of observations, we need to use an *array formula*. This is simply a formula that operates on multiple cells and returns the result in multiple cells. First, highlight the cells next to your histogram bin sizes. Then, begin typing

```
=FREQUENCY(
```

At this point, select the cells from your corrected, log-transformed data that correspond to your log ratios. When you have selected the cells, type a comma (,). Now go back to your histogram data worksheet and select your bin sizes. Finally, type a closing parenthesis. (Don't hit enter yet!) When I do this on my worksheet, the final formula looks like this:

```
=FREQUENCY('2hr data'!N6:N6123,'2hr histogram data'!A5:A64)
```

When you have entered your formula, instead of pressing enter, use ⌘-Enter instead (on Windows, use Control-Shift-Enter). This tells Excel to evaluate the result as an array formula instead of a normal formula. When you are done, you should see that the observation frequencies have been filled in.

Estimating the Gaussian Fit

The next step in creating our histogram is to make a column of true normal (Gaussian) values for comparison. To do this, create two new entries on your parameters worksheet: one should contain the mean of your log-corrected ratios, and the other should contain the standard deviation (variance). You can calculate these using the AVERAGE and STDEV Excel functions, respectively. Add another variable to the parameter list, called "Normal Distribution Scaling Factor" that you will use to adjust the height of your Gaussian curve. Set the adjacent cell to be 100 for now. We will adjust this later.

With these three values set, fill in the column next to your observed histogram with the following formula:

$$f(x) = Ae^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Here, A, μ, and σ are the scaling factor, mean, and variance you determined above, respectively. The variable x should be the bin size for that particular value. The Excel function EXP will calculate the Euler number e raised to a given power. For some reason, I had to make the negative sign explicit in the function by multiplying by -1. Before you apply this formula to the entire column, double check your first entry to make sure you have entered the right formula: The value f(-1.45) should be quite small.

Create a new plot on a worksheet called "2hr distribution" that contains all three columns—the bin size, the observed histogram, and the Gaussian fit. Your plot should still be a scatter plot, but select a scatter plot with smoothed lines between the data points. Because this plot is not a bar chart, it's not strictly a histogram—instead, it's a distribution. In the first dialog of the plot setup, name your series data series appropriately so you can distinguish between the observed histogram and the Gaussian fit. (When it asks you about "Source Data," click the series button and you will be able to edit the series name. On my worksheet, the actual data is blue and the normalized estimate is red.)

Once your plot has been created, you can adjust the Gaussian scaling factor on your parameters worksheet so that the maxima are approximately equal.

8. Calculate the Z-score for both your highest and lowest outlier. Looking at the distribution plot, would it be a good approximation to call the corrected, transformed log-ratios "normal" or "Gaussian distributed?" Why do you think the two curves are so different?

Part 4: Gene Clustering

The final analysis that we will perform on our microarray data is clustering. On your Macintosh, in the applications directory, you should find two applications: Cluster and Java TreeView. Both of the applications are available for PC's as well, and can be found on the following website:

```
http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/index.html
```

The clustering techniques we will perform here will organize the genes into artificial classes based on their expression characteristics. Clustering in this regard takes some highly-complex problem and attempts to simplify it using computer algorithms. One of the limitations of our cognitive abilities is the visualization of highly-dimensional data: we can't imagine a 6,000 by 7 dimensional space. Clustering tools (and the other tools discussed in your book) are designed to fix this problem by taking the expression data of thousands of genes over several hours and grouping those genes that behave similarly into clusters.

Download the file `spo100.txt` from the lab website. This file contains the corrected, globally normalized expression data for 100 of the open reading frames, chosen at random from the original data set. Each row corresponds to one yeast open reading frame (as before), and each column corresponds to the normalized $\log_{10}$ ratio for each time point.

K-means Clustering

The first clustering application that we will perform uses a method called "*k*-means" clustering. In this application, the computer program attempts to sort your data into *k* clusters, where you specify the number of cluster *k* that you would like. Here, we have three classes that we would like to impose on our data set: genes that are up-regulated during sporulation, genes that are down-regulated, and genes that remain constant. Thus we will set *k =3* in our clustering program.

Open up the program Cluster and load the normalized sporulation dataset (use the File → Open Data to select the `spo100.txt` file). Name your job "spo100_kmeans," select "k-means" clustering, and check "Cluster Genes." Set the number of clusters to be three, and then click "Execute." You'll notice that not much has happened, but if you check in the directory where you first stored `spo100.txt`, data files containing your clustering result will now be located there.

Open the Java TreeView application, and open the .CDT file from your clustering result. Make sure that "Kmeans" is selected as the style. When you are presented with your clustering data, you will notice three blocks in the middle of the window—these are the clusters. Each block is actually a matrix, with rows corresponding to the yeast genes and columns corresponding to the time points of the experiment. Each cell in the matrix is colored: red means the genes are up-regulated during sporulation (with a $\log_{10}$ ratio that is positive), and green means the genes are down-regulated (with a negative $\log_{10}$ ratio). Cells that are black are relatively un-regulated.

Clicking on a particular row in the matrix will zoom in on that row at right, and you will be able to move your mouse over the cells at the right to obtain actual values for the $\log_{10}$ ratios. You can also click on the name of the ORF at the far right to get a helpful web page that will describe what that gene does, if known. Try this out with a few genes.

9. How do the three clusters look? Did the *k*-means algorithm cluster the genes as we would have hoped (with up-regulated, unaffected, and down-regulated genes in different clusters)? What happens if you repeat your clustering, now with a *k* of four?

Hierarchic Clustering

In hierarchic clustering, the number of clusters isn't known: instead, comparisons are performed between every pair of genes and a hierarchic tree is formed. If two genes are similar in their expression patterns (more similar than any other pair of genes), they will be paired. Then, that pair of genes is compared with all other genes in the algorithm: as a result, a hierarchic tree of similar pairs is created. We will explore exactly what this means using the Cluster program.

Using the same dataset, construct a hierarchical cluster: In the Cluster program, select "Hierarchical," and tell the program to cluster genes. Change the name of your job to "spo100-hierarchic" or something similar. Leaving all the other options as they were, click "Centroid linkage." Now you can open up this new file in TreeView (use the "Auto" style)—you'll notice that at the left is now a hierarchic tree corresponding to the clustering result. Clicking on the branches of the tree will highlight and "zoom in" on the members of that branch.

10. What regulation characteristics do the top two-level branches of your tree correspond to (roughly)? The top two branches of the tree can be selected by clicking on the longest two horizontal lines at the far left. The length of these lines roughly corresponds to the "difference" of the two sub-trees. Because the horizontal lines are relatively long, that means that the branches beneath this particular tree node are rather different.

It's important to note that, regardless of the clustering method used, there are no good statistics for determining whether a clustering tree is significant. Thus, if you use clustering in your analysis of microarrays, you will have to perform checks to make sure your tree makes sense: One question you might ask in this case is whether the up-regulated genes identified by clustering are actually part of reasonable (and known) pathways for sporulation. In their original paper, Chu and DeRisi, et. al. confirmed this by examining the behavior of known genes from their dataset.

Part 5: Cleaning Up and Submitting Your Work

For this lab, you will be submitting your answers to the questions above as well as your Excel worksheet. Thirty points of your assignment will come from your answers to the ten questions, and the remaining 20 points of this lab will be given based on your completion of the worksheet. In addition to being graded on the completeness of the worksheet, you will also be graded on style.

First, delete the worksheet containing the original data: spospread.txt. As all of your data has been entered on a new worksheet, there's really no need to keep the original anymore. You can delete this worksheet by right-clicking on its tab and selecting "Delete."

Next, make sure your data worksheets are stylistically pleasing: Is everything well-labeled? Do all of your charts have a title? Are your numerical values formatted with a reasonable number of decimal places?

Finally, make sure your plots are in order. Again, are there reasonable titles for all plots? Are all the axes labeled?

When you are satisfied with your work, submit the entire Excel worksheet to the course web server for assignment 5.