

GCB 535 / CIS 535  
Fall 2004

## Homework 5

Due in class, Monday, November 22, 2004

1. (6 pts) The following is an example of results from a spotted microarray shown in class:

| Normal Breast Tissue |       |       |       |
|----------------------|-------|-------|-------|
| Gene                 | Cy3   | Cy5   | Ratio |
| A                    | 12500 | 10000 | 1.25  |
| B                    | 5000  | 5000  | 1.0   |
| C                    | 950   | 850   | 1.1   |
| D                    | 600   | 700   | 0.87  |

| Breast Tumor |       |      |       |
|--------------|-------|------|-------|
| Gene         | Cy3   | Cy5  | Ratio |
| A            | 7500  | 6300 | 1.19  |
| B            | 4329  | 4300 | 1.01  |
| C            | 1800  | 900  | 2.0   |
| D            | 10000 | 700  | 14.29 |

Cy3 = experimental sample, Cy5 = reference sample

This experiment used a common reference to be able to make comparisons between the two conditions.

a) How would the data look if you had performed a loop-style microarray experiment – that is, ran both the normal breast tissue and breast tumor on the same array? Provide a table similar to the ones above, clearly indicating which sample is which and your expected expression ratios.

b) What is the main disadvantage of using the loop-style experiment rather than the common reference?

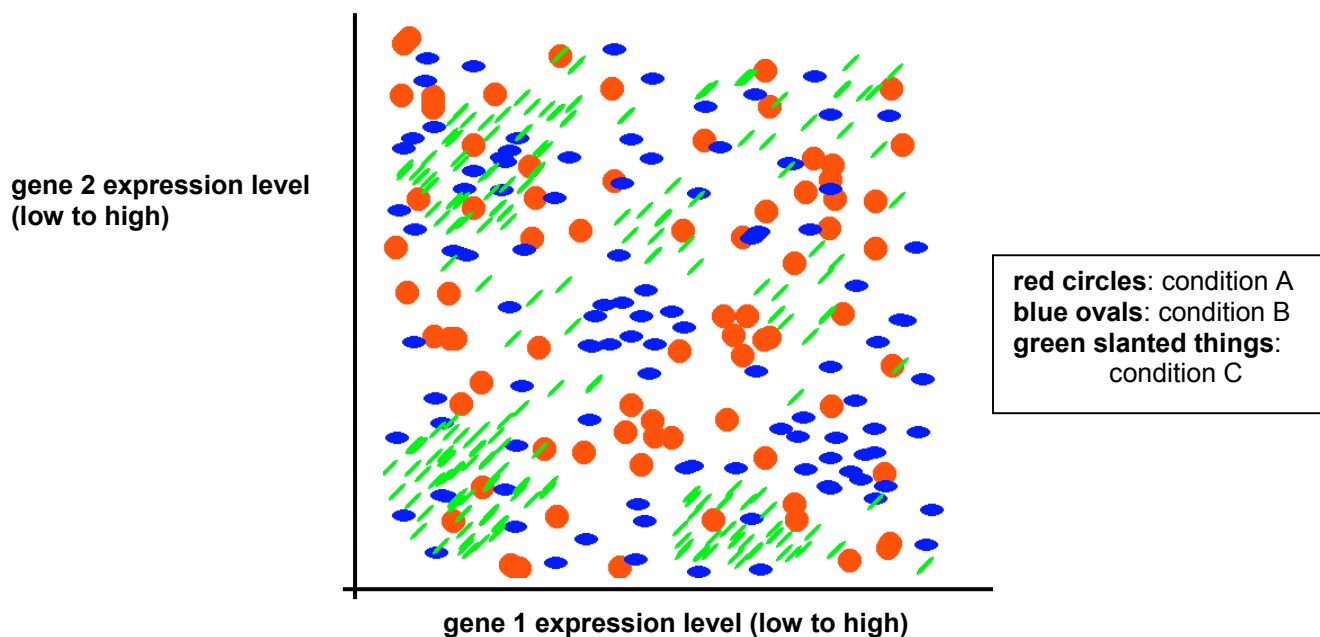
2. (9 pts) Using any of the tools we've discussed so far in class, describe how you would go about deciding on a set of oligonucleotide probes for a particular gene that could be used for an Affymetrix-style photolithographic microarray chip.

3. (12 pts) Let's say we did a set of microarray experiments in an attempt to distinguish between three conditions – A, B, and C. The hope was we would be able to cluster the results based on the expression levels of the genes on the microarray, and then be able to classify new samples as belonging to condition A, B, or C based on their expression levels.

Unfortunately, because the grad student running the experiments fell asleep during the runs (he was up too late studying for his CIS 535 exam), most of the data was ruined, with the exception of expression data for two genes, over several replicates of the A, B, and C conditions. Rather

than repeating the experiments, we want to see if there is any way we can still use the results to build our classifier.

Below is a graphical representation of the data. Each dimension represents the expression level of a gene (2 genes total). Each point corresponds to one microarray experiment for one of the conditions (as labeled).



Evaluate each of the following techniques in terms of how well they would do in this situation. Specifically, for hierarchical and k-means, describe what sort of clusters you would end up with, with this data. For all three methods, describe how you would go about classifying a new point (i.e., a new pair of expression levels for the two genes that represent an unknown condition) as belonging to condition A, B, or C, and how accurate you expect the classification to be.

- Hierarchical clustering using single-linkage based on Euclidean (straight-line) distances
- K-means clustering using 3 means ( $k=3$ ), based on Euclidean distances
- K-nearest neighbors using 7 neighbors ( $k=7$ ) (the nearest neighbor of any point is the closest other point to it according to Euclidean distance)

4. (8 points) The following data file contains the expression data of 11 yeast genes under 80 conditions: [http://www.seas.upenn.edu/~cis535/Lab/yeast\\_small.txt](http://www.seas.upenn.edu/~cis535/Lab/yeast_small.txt)

Use Gene Cluster 3.0: <http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm> and TreeView: <http://jtreeview.sourceforge.net/> to analyze the data and answer the following question. These data are already normalized, you do not need to filter or normalize again in Cluster 3.0.

a) Conduct hierarchical clustering of the genes, using correlation (uncentered) for similarity metric and average linkage for clustering method. View your result in TreeView. Save your tree image together with the data matrix with the leaves labeled by YORFs using Export -> Save Tree Image, select Gene Tree, Data matrix and use YORF as Gene Headers . Paste your tree image with your answer. Based on the shape of the tree, pick out a cluster formed by 5 genes. Mark them on your tree. What is the color coding in the image of expression data, i.e. what does red, green, black and gray each represents? For the 5 genes picked above, report one condition under which all five of them are repressed and one condition under which all five are over-expressed.

b) Conduct K-means clustering of the genes, using  $k = 3$  and correlation (uncentered) for similarity metric. Report the ORFs for the genes in each cluster. Use GO term finder (<http://jtreeview.sourceforge.net/>) to annotate each cluster, report the top 2 GO term for cell process for each cluster together with their P-values. Based on these, do you think the result from k-means clustering makes biological sense?