

Name:

Gene:

Task 4: *Analysis of Microarray Data*

In this assignment, you will be looking at gene expression data, i.e., the results of DNA microarray experiments. In particular, you will be searching for genes which appear to be similarly expressed (transcriptionally) as evinced by the microarray data. We will be using clustering algorithms (such as k -means clustering and hierarchical clustering) to find such similarly expressed groups of genes.

For this assignment, we will be using 2 freely available programs (both programs work on Macs and Windows PCs). The first program is called "Cluster 3.0" and, as the name implies, this program will cluster microarray data. Cluster 3.0 is available here:

<http://www.wellesley.edu/CS/courses/CS-BiSc303/milestones/Cluster.app.zip> (Mac)
<http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm#ctv> (Windows)

The second program is called Java TreeView, and this program allows you to view the clustering results (obtained from Cluster 3.0) graphically. Java TreeView is available for download here:

<http://genetics.stanford.edu/~alok/TreeView/>

For Macs we recommend downloading TreeView-1.0.4-osx.dmg, and for Windows PCs we recommend downloading TreeView-1.0.4-bin.zip.

Step 1: Once you have the 2 programs Cluster 3.0 and Java TreeView working on your computer, you will need to download a file containing the results of microarray experiments (gene expression data). The following file contains expression data from a series of microarray experiments performed on our beloved yeast organism:

<http://www.wellesley.edu/CS/courses/CS-BiSc303/milestones/yeast.txt>

After starting the Cluster 3.0 program, you should open this file of yeast expression data. You can confirm that the Cluster program correctly opened the yeast file by checking if the middle of the Cluster window says "2467 Rows" and "79 columns". These numbers indicate that the yeast file contains information on the expression of 2467 yeast genes as measured in 79 different microarray experiments. A summary of the 79 microarray experiments can be found here:

<http://genome-www.stanford.edu/clustering/YeastCols.html>

Are you ready to cluster? I can't hear you. ARE YOU READY TO CLUSTER? Ok, then. Let's click on the "Hierarchical" tab in the Cluster window, select the "Cluster" box in the "Genes" section on the left and also select the "Cluster" box in the "Arrays" section on the right. By selecting both boxes, we will be finding groups (clusters) of similarly expressed genes, as well as finding groups (clusters) of similar experiments. Finally, you can click the "Average linkage" button at the bottom of the Cluster window to perform the clustering. At

the very bottom of the Cluster window you should see a message such as "Performing average linkage hierarchical clustering" while the program is executing. After a minute or two, the program should finish and you should see the message "Done Clustering" at the very bottom of the window. The Cluster program should have generated 2 or 3 new files as a result of the clustering. One of the file names should look something like "yeast.cdt".

To view the results of the clustering, open up the Java TreeView program. Using TreeView, open the file with name ending in ".cdt" which you created with the Cluster program. The TreeView program should show 4 vertical columns. The first column should contain a tree dendrogram. The second column should contain a lot of green and red spots. Try selecting individual genes by clicking on rows of the clustered data (the green and red spots). Do you see the gene name and its description in the fourth column? Now try dragging the mouse over a region of the clustered data to select a set of genes. You can also select groups of *genes* by highlighting branches of the tree in the first column. You can select groups of *experiments* by highlighting branches of the tree at the top of the second column.

Let's search for your gene. Using the "Analysis" menu at the top of the screen, we will "Find" you gene. Type in the name of your gene and press the "Search" button. If you press the "All" button, your gene should be highlighted in the clustered data. Try selecting a group of genes in the data which neighbor your gene (i.e., genes which cluster with your gene are genes which are similarly expressed in the 79 microarray experiments). Please list approximately 20 genes (and their function) which are similarly expressed to your gene. Are you surprised by this list? What are the functions (in the fourth column) of the genes which cluster with your gene? Are the functions related to the function of your gene? Are there yeast genes which have a function similar to your gene's function but which do not appear to be similarly expressed to your gene in the 79 experiments? In what type of experiments does your gene (and those which are similarly expressed) appear to be more highly expressed (red) than control and less expressed (green) than control?

Now let's return to the Cluster program. This time, select the "k-Means" tab. We'll check both the "Organize genes" box in the "Genes" section on the left and the "Organize arrays" box in the "Arrays" section on the right. For our number of clusters, let's fill in 20 (for the Genes section) and 10 (for the Arrays) section, and for the number of runs let's fill in 10 (for the Genes section) and 5 (for the Arrays). Now "Execute" the clustering. When the program finishes clustering the data, you should see something like "Solution was found X times" at the very bottom of the window.

If you return to the TreeView program, you can view the *k*-means clustering results by opening the new ".cdt" file (probably named something like "yeast_K_G20_A10.cdt"). The data (in the second column) should now have a bunch of white lines running through it. These indicate the various groups (clusters) of similarly expressed genes (rows) and similar experiments (columns). You should confirm that there are 20 clusters of genes and 10 clusters of experiments. Again, find your yeast gene in the data and look at the genes which cluster with your gene. Do the same genes cluster with your gene in this case (using the *k*-means clustering algorithm) as in the case when you clustered the data using a different approach (the average linkage hierarchical clustering algorithm)? Try returning to the Cluster program and clustering the data with different parameters or methods (e.g., use a different number of clusters in the *k*-means algorithm, or perform hierarchical clustering using "Centroid linkage" or "Single linkage" instead of "Average linkage"). Do the genes which cluster with your yeast gene change? How confident are you in the clustered results?

For further details on clustering this data see:

Eisen, Spellman, Brown, and Botstein, "Cluster analysis and display of genome-wide expression patterns", *Proc. Natl. Acad. Sci. USA* **95**, page 14863, 1998.

Step 2: One of the many challenges in diagnosing and treating cancers is that cancers which appear clinically similar can be genetically heterogeneous. For instance, prostate cancers which appear similar may be caused by different, independent gene defects. The different gene defects can have different implications for prognosis and treatment of the cancer. For this part of the assignment, we will be dealing with two different forms of acute leukemia, namely acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). The two leukemias appear very similar morphologically. However, because the chemotherapy regimens differ for AML and ALL patients, the ability to distinguish between them is critical for successful treatment.

You will be analyzing microarray data from experiments based on 38 patients with either AML or ALL. The microarray experiments were performed by extracting RNA samples from bone marrow cells of the patients and hybridizing the RNA to a microarray chip. You can retrieve the microarray data from:

<http://www.wellesley.edu/CS/courses/CS-BiSc303/milestones/ALL-AML.txt>

The data corresponds to the measured gene expression of approximately 7000 human genes in 38 microarray experiments (one experiment for each patient). Using the Cluster program, you should cluster this data using the k -means clustering algorithm. Try clustering the genes into 20 clusters (10 runs) and the arrays (i.e., experiments) into 2 clusters (20 runs). Now view the results in the TreeView program. Select a few genes from the data (second column). At the top of the third column, you should look at how the clustering algorithm grouped the 38 experiments into 2 clusters. Do the AML patients predominantly cluster together in one of the groups and the ALL patients predominantly cluster together in the second group? Re-run the k -means clustering algorithm a couple more times and see if the results change (i.e., do the same patients cluster together in the 2 groups). Do your results indicate that microarray experiments can be used to distinguish between different forms of acute leukemia? If a new patient was diagnosed with acute leukemia, and if a microarray experiment was performed on that patient's bone marrow RNA, how might the results of the new microarray experiment be used to help guide the patient's diagnosis? Based on your work in this milestone and on what you know of microarray experiments, how confident would you be in diagnoses made on the basis of microarray data?

Finally, re-cluster you data using the k -means clustering algorithm, but cluster the arrays (experiments) into 3 groups. Do one or two of your clusters correspond predominantly to AML? Do one or two of your clusters correspond predominantly to ALL? The researchers who first performed these experiments (they clustered the data just as we have been doing in this milestone) found that ALL experiments tended to cluster into 2 of the 3 different groups. After examining the groups more closely based on immunophenotype data, they found that the 2 ALL clusters corresponded to patients with "T-lineage" ALL and patients with "B-lineage" ALL (cells which express different levels of particular antigens).

For further details on this research and the microarray experiments see: Golub, Slonim, Tamayo, Huard, Gaasenbeek, Mesirov, Coller, Loh, Downing, Caligiuri, Bloomfield, and Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring", *Science* **286**, page 531, 1999.

Step 3: In Step 2, we clustered microarray data from approximately 7000 human genes in an attempt to distinguish between AML and ALL. Do you expect that 7000 genes are implicated in acute leukemia? The expression of the vast majority of these genes is unrelated to the leukemia. In essence, these unrelated genes are just causing noise in our clustering which we may be better served without. By using a subset of more informative genes, our results may improve (although they may not).

- Open the leukemia data set in the Cluster program.
- Select the "Filter Data" tab and check the "SD (Gene Vector)" box
- Fill in the value 900 and click "Apply Filter" followed by "Accept"

We've now used an extremely crude approach to whittle our data set into one tenth the number of genes we were using before (a set of genes whose values deviate significantly in the 38 experiments).

- Now cluster this reduced set of data using k-means (try clustering the genes into 20 clusters with 20 runs, and cluster the arrays into 3 clusters with 200 runs). Note that we can use more runs now because our data set is so much smaller.
- Open the results in TreeView and check how well the clustering distinguished the leukemia. Have the clusters improved as compared to the 7000 gene clusters?

Are there particular genes which you see which seem highly expressed in AML patients and less expressed in ALL patients or vice versa? Try searching for the gene "adipsin" and for the gene "TCL1". Is the expression of these genes informative in determining different classes of cancer?

The researchers who originally performed this study had the computer choose only 50 genes which appeared informative, and they then clustered the expression data for these 50 genes. With these 50 genes (less than one hundredth of the number we used in Step 2), they made no mis-classifications. Of the 50 genes chosen by the computer, most turned out to be closely related to the particular type of leukemia. For example, some of the genes are known oncogenes (c-MYB, E2A, HOXA9). Also, some genes (CD11c, Cd33, and MB-1) encode cell-surface proteins for which antibodies have been demonstrated to be useful in distinguishing lymphoid from myeloid lineage cells.

Due date: Wednesday, March 9th 2005

Please e-mail to: btjaden@wellesley.edu and dwebb@wellesley.edu