**CS423 Lab 9: Microarrays and Clustering**

**Due:** Wednesday, Nov 18 [beginning of class]

**Purpose:** This lab gives you practice clustering microarray data using Cluster 3.0 and viewing the results in Java Treeview (free programs for you to download if you want to work on your own computer; available on the computers in Shiley; see links on Moodle). There is no programming in this lab, although the programs you will use have implemented the clustering algorithms we discussed in class.

**Directions:** Choose a partner so that one person has more experience in biology and one person has more experience in computer science, since this lab focuses on biology and clustering algorithms. You may work with someone with whom you have already worked with for a prior lab. All submitted work should be typed and include your name(s). Submit your write-up *electronically* to Moodle (one per pair).

**Background Information:** Review your class notes on k-means clustering and hierarchical clustering.

**Lab Exercises and Write-up (85 points):**

In this lab you will use gene expression data from microarray experiments to find genes that are similarly expressed. For each exercise below that has *(Put in write-up),* include your answer in the write-up and use the same number as the lab exercise in your write-up.

1. Download the two microarray datasets from Moodle. These are: yeast.txt and ALL-AML.txt

**Analyzing yeast expression data:**

2. Launch the Cluster 3.0 program (should be under Start->All Programs->Courseware->Computer Science). You will perform clustering algorithms on the yeast dataset first. This dataset is compiled from several microarray experiments on all yeast genes. In the Cluster 3.0 software, go to File->Open data file and open yeast.txt.

3. There are several kinds of clustering, including those we discussed in class, that Cluster 3.0 can perform for you. Click on the hierarchical tab to perform the hierarchical clustering (tree-making) technique. The left panel allows you to cluster by genes and the right panel allows you to cluster by experiments. (Clustering by experiments will produce clusters using each column as a data item (vector) to cluster.) Select the “Cluster” checkboxes in both panels. Select Euclidean distance as the metric in both panels. Click on the “Average Linkage” button to perform the clustering. It might take a few minutes to complete, but you should see a message “Done Clustering” at the bottom when it is finished.

4. Cluster 3.0 should create several new files (located in the same folder as your input file). Among them are named yeast.cdt, yeast.atr, and yeast.gtr. The yeast.cdt file is a tab-delimited text file with the genes and arrays shuffled into their clusters. Open yeast.cdt in Excel, so you can see its contents easily. The yeast.atr and yeast.gtr files are text files that show the linkages for the tree cluster for the experiments/arrays (yeast.atr) and the genes (yeast.gtr). Now that the clusters have been formed, you will use another program to visualize the clusters.

5. Launch Java Treeview (Start->All Programs->Courseware->Computer Science). Click do not register, if prompted to register the software. Go to File->Open to open the yeast.cdt file.

Be sure the box on the right-hand side of the open dialog box has the word “linked” for style. You should see four panels. The left panel is the tree cluster for the genes. The second panel shows a heat map of the expression levels of all genes in all experiments. Click on a row in the second panel to select a gene. You should see its name and description in the fourth column. You can select groups of genes by highlighting a rectangular region in the second panel. You can select a merged group of genes by clicking on a clade in the tree on the left panel. You can select a group of experiments by highlighting branches at the top of the second panel.

6. Search for your partner’s or your yeast gene from lab 2. Use the Analysis->Find Genes menu option. Type in your gene name and press search. You should get a list of matches and clicking on the name will show you its position in the second panel. Try selecting a group of genes that neighbor your gene in the tree cluster.

7. (Put in write-up)

a. (1 pt) Write down the gene you or your partner selected in lab 2.

**CYS3 (YAL012W)**

b. (3 pts) List approximately 10 genes and their functions (you’ll see this in the fourth panel) which are similarly expressed to your gene. You can select a window of about 10 genes surrounding your gene in the heat map window.

1. **SAM2 – Methionine Biosynthesis**
2. **ADE5, 7 – Purine Biosynthesis**
3. **ADE12 – Purine Biosynthesis**
4. **PDC6 – Glycolysis**
5. **ADH2 – Glycolysis**
6. **PFK2 – Glycolysis**
7. **HIG1 – Glycolysis**
8. **PDR13 – Drug Resistance**
9. **ADH5 – Glycolysis**
10. **HPT1 – Purine Biosynthesis**

c. (3 pts) How many in this list of 10 perform a different function than your yeast gene?

**CYS3 – Methionine Biosynthesis. 9 are different.**

d. (2 pt) Go to: <http://genome-www.stanford.edu/clustering/YeastCols.html>

In what experiments (you can get info on the experiments at the URL below) is your gene highly expressed (red)?

1. **diau e**
2. **diau f**
3. **diau g**

3. (2 pt) In what experiments is your gene not expressed (green)?

1. **heat 40**
2. **heat 160**
3. **cold 40**
4. **cold 160**
5. **spo 5**
6. **spo 7**
7. **spo 9**
8. **spo 11**
9. **spo- early**
10. **spo- mid**

8. Go back to Cluster 3.0. This time, cluster the yeast genes and experiments using k-means (use the k-means tab). Check the Organize genes and Organize arrays in the two panels. Find 20 clusters for genes and find 10 clusters for arrays. The number of runs for each is the number of times k-means will execute on a different set of random initial cluster assignments. Do 10 runs for each. Execute the clustering using Euclidean distance. When the program finishes, you should see “Solution was found X times” at the bottom, which indicates how many times the given solution was found given the 10 different initial cluster assignments.

9. Go to the Java Treeview program to open the file yeast\_K\_G20\_A10.cdt. When opening, be sure to check the style option as Kmeans. Cluster 3.0 created three files again (but now the other two are .kgg and .kag which show the cluster assignments of the genes and experiments, respectively). You should see a heat map of the genes and experiments in the second column and white lines separate the clusters. Confirm there are 20 gene clusters and 10 experiment clusters.

10. (10 points) (Put in write-up) Find the yeast gene you listed in exercise 7 in this heat map and see which genes cluster with it using k-means.

a. (4 pts) How many of the 10 or so genes you found closest in the hierarchical clustering are in the same cluster as your gene when using k-means?

**2 genes are in the same cluster:**

1. **ADE5, 7 – Purine Biosynthesis**
2. **PDC6 – Glycolysis**

b. (3 pts) Which are in k-means but not in the hierarchical cluster?

**Probably about 40 genes.**

c. (3 pts) Which are in hierarchical but not included in the k-means cluster?

1. **SAM2 – Methionine Biosynthesis**
2. **ADE12 – Purine Biosynthesis**
3. **ADH2 – Glycolysis**
4. **PFK2 – Glycolysis**
5. **HIG1 – Glycolysis**
6. **PDR13 – Drug Resistance**
7. **ADH5 – Glycolysis**
8. **HPT1 – Purine Biosynthesis**

11. (8 points) (Put in write-up) Try clustering using k-means again with at least two different number of clusters (value for K). Are you confident in the clustering results of your yeast gene based on the results you get for different numbers of clusters? Why or why not?

**No. We ran with higher K values and got a lot of different clusters. We even used the same K value a couple times and the clusters were pretty different.**

**Analyzing cancer data:**

12. Different gene defects in cancer cells need different treatments, so it is important to determine cancer type in cells. For this part of the lab, you will look at two forms of leukemia: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). It is important to differentiate between the two for successful treatment of the patients. The dataset contains 38 patients with ALL or AML (each in own column in dataset) and approximately 7000 human genes.

Open the text file ALL-AML.txt in excel to see the raw data.

13. (10 points) (Put in write-up) Cluster this dataset using **k-means** (Euclidian distance) into 20 gene clusters and 2 patient clusters. Do 20 runs for each. View the results in Java Tree View. Be sure to open with style “kmeans”.

a. Do the AML and ALL patients tend to cluster into two independent groups? (Look at the patients in each group to the left and to the right of the vertical line)

**No they don’t. About half of the AML patients were in each cluster.**

b. List the patients put into one of the patient clusters.

1. **ALL2**
2. **ALL9**
3. **ALL15**
4. **ALL17**
5. **ALL20**
6. **AML28**
7. **AML30**
8. **AML32**
9. **AML33**
10. **AML36**
11. **AML37**

c. List the patients put into the other patient cluster.

1. **ALL1**
2. **ALL3**
3. **ALL4**
4. **ALL5**
5. **ALL6**
6. **ALL7**
7. **ALL8**
8. **ALL10**
9. **ALL11**
10. **ALL12**
11. **ALL13**
12. **ALL14**
13. **ALL16**
14. **ALL18**
15. **ALL19**
16. **ALL21**
17. **ALL22**
18. **ALL23**
19. **ALL24**
20. **ALL25**
21. **ALL26**
22. **ALL27**
23. **AML29**
24. **AML31**
25. **AML34**
26. **AML35**
27. **AML38**

Re-run the k-means two more times with the same parameters (20 gene clusters and 2 patient clusters) to see if the patients consistently cluster into the same two groups.

b. Were the results consistent across the next two runs? If not, list the clusters produced for each of the two new runs.

**Yes, the results were consistent.**

14. (8 points) (Put in write-up) Do your results indicate that a microarray experiment can differentiate between the two types of leukemia using k-means clustering with 2 clusters? Why or why not?

**No, a microarray could not differentiate between the two types of leukemia using k-means clustering with 2 clusters. The AML patients are split between the two clusters with about**

**1/3 – 1/2 in one cluster and 1/2 – 2/3 in the other cluster.**

15. (6 points) (Put in write-up) Re-cluster the ALL/AML data using k-means (Euclidean distance) with **3** clusters for the patients and 20 clusters for the genes.

a. Do one or two of your patient clusters generally correspond to AML?

**Not quite. There is one cluster with 6 AML and 3 ALL, but that’s not a strong correspondence.**

b. Do one or two of your patient clusters generally correspond to ALL?

**Yes. There is a small cluster that is all ALL patients. There is another large cluster with a lot of ALL patients, but it also contains some AML patients.**

c. List the members for each cluster.

**Cluster 1**

1. **ALL2**
2. **ALL17**
3. **ALL20**
4. **AML28**
5. **AML30**
6. **AML32**
7. **AML33**
8. **AML36**
9. **AML37**

**Cluster 2**

1. **ALL3**
2. **ALL6**
3. **ALL9**
4. **ALL10**
5. **ALL11**
6. **ALL23**

**Cluster 3**

1. **ALL1**
2. **ALL4**
3. **ALL5**
4. **ALL7**
5. **ALL8**
6. **ALL12**
7. **ALL13**
8. **ALL14**
9. **ALL15**
10. **ALL16**
11. **ALL18**
12. **ALL19**
13. **ALL21**
14. **ALL22**
15. **ALL24**
16. **ALL25**
17. **ALL26**
18. **ALL27**
19. **AML29**
20. **AML31**
21. **AML34**
22. **AML35**
23. **AML38**

16. (8 points) (Put in write-up) You just looked at clustering all 7000 genes to distinguish between ALL and AML. But, it’s not likely that all 7000 genes are implicated in leukemia, as many have expression levels and functions unrelated to leukemia. So, these unrelated genes are just causing noise in clustering the data. A subset of genes might prove more informative. Use Cluster 3.0 to open the leukemia dataset. Select the Filter Data tab and check the SD (Gene Vector) box. Fill in the value 700 and click “Apply Filter” followed by “Accept Filter” when the filtering is complete. This filters the dataset to about 1/10th of the size. Cluster this filtered dataset using k-means (20 gene clusters, 3 patient clusters, and 100 runs each with Euclidean distance). Note we can do a lot more runs now with a smaller dataset.

View the results in Java Tree View and see how well the clustering distinguished the leukemia types. a. What 3 patient clusters do you get?

**Cluster 1**

1. **AML28**
2. **AML29**
3. **AML30**
4. **AML31**
5. **AML32**
6. **AML33**
7. **AML34**
8. **AML36**
9. **AML37**
10. **AML38**

**Cluster 2**

1. **ALL2**
2. **ALL3**
3. **ALL6**
4. **ALL9**
5. **ALL10**
6. **ALL11**
7. **ALL14**
8. **ALL23**

**Cluster 3**

1. **ALL1**
2. **ALL4**
3. **ALL5**
4. **ALL7**
5. **ALL8**
6. **ALL12**
7. **ALL13**
8. **ALL15**
9. **ALL16**
10. **ALL17**
11. **ALL18**
12. **ALL19**
13. **ALL20**
14. **ALL21**
15. **ALL22**
16. **ALL24**
17. **ALL25**
18. **ALL26**
19. **ALL27**
20. **ALL35**

b. Have the results improved from the previous exercise when using all the genes?

**Definitely. One cluster was exclusively AML, one exclusively ALL, and the other was ALL except for one AML patient.**

17. (5 points) (Put in write-up) If we wanted to determine which genes give us the most information, we could see which genes are acting most differently in AML versus ALL patients. Which genes seem highly expressed in AML patients and lowly expressed in ALL patients? (Look for genes that are primarily green for the ALL clusters and red for the AML cluster.)

18. (5 points) (Put in write-up) Which genes seem highly expressed in ALL patients and lowly expressed in AML patients?

19. (6 points) (Put in write-up)

a. Look for the gene adipsin. Is this informative for determining the type of leukemia?

b. Look for the gene TCL1. Is this informative for determining the type of leukemia?

c. Look for the gene ANT2. Is this informative for determining the type of leukemia?

16. (8 points) (Put in write-up)

a. If a new patient is diagnosed with leukemia and the microarray experiment was performed on his/her bone marrow RNA, how might the results guide a doctor in making a diagnosis?

b. Based on this lab and your knowledge of microarray experiments, how confident would you be in making a diagnosis based on microarray data only?

**Appendix A: Authorship (please include statement in your write-up)**

The write-up submitted for this lab was authored by the named person(s) on this lab report. All external sources to BIO/CS423 are cited properly.